

MCCMB 2025

Сборник тезисов

12-й Московской
конференции по
вычислительной
молекулярной биологии
(MCCMB)



Тезисы устных докладов

**Ремоделирование структуры хроматина в клетках
периферической крови у пациентов с тяжелой формы Delta
COVID-19**

В.Э. Акимов¹, Д.И. Тычинин¹, О.А. Антонова¹,
А.М. Шаймарданов¹, М.Д. Воронина¹, К.А. Дейниченко¹,
О.Д. Фатеев¹, В.С. Юдин¹, С.М. Юдин¹, В.Е. Мухин¹,
С.В. Романова¹, А.И. Некрасова¹, А.С. Жданова¹, А.В. Цыпкина¹,
И.С. Владимиров¹, А.В. Махотенко¹, А.А. Кескинов¹,
С.А. Краевой¹, Е.А. Снигирь¹, Д.В. Светличный¹, В.И. Скворцова²

¹Федеральное государственное бюджетное учреждение
«Центр стратегического планирования и управления медико-
биологическими рисками здоровью» Федерального медико-
биологического агентства (ФГБУ «ЦСП» ФМБА России). 119121,
Россия, г. Москва, ул. Погодинская, д. 10, стр. 1.

²Федеральное медико-биологическое агентство (ФМБА
России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.
VAkimov@cspfmba.ru

Инфекционные заболевания, вызванные разными штаммами вируса SARS-CoV-2, вызывают глубокие изменения в иммунной системе, приводя к провоспалительным процессам и дисфункциональным нарушениям. Для всестороннего анализа этих изменений проведена интеграция данных секвенирования транскриптома единичных клеток (scRNA-seq) и профилирования открытости хроматина единичных клеток (scATAC-seq) на 112 249 ядрах мононуклеарных клеток периферической крови пациентов со средним и тяжелым течением COVID-19, опосредованного штаммом дельта. Этот подход позволяет соединить информацию об экспрессии генов и регуляции их активности на уровне отдельных клеток, что открывает новые возможности для понимания молекулярных механизмов, определяющих тяжесть течения инфекции.

Результаты исследования показали, что тяжелая форма дельта-COVID-19 характеризуется появлением специфической

субпопуляции моноцитов Mon IFI30, демонстрирующей повышенную активность провоспалительных генов, таких как CCL3, IL1RN и TNF.

Ключевыми активаторами экспрессии генов в этой субпопуляции являются факторы транскрипции EGR2, MITF, SPI1, MAFB, BACH1 и NFE2L2, тогда как KLF2, FOS и ETS1 проявляют репрессорную активность. Уровни активности KLF2 снижены у пациентов с тяжелой формой дельта-COVID-19, что согласуется с его противовоспалительной ролью. ATF3 и ATF5, специфичные для Mon IFI30, регулируют маркерные гены, включая CLEC5A, который способствует цитокиновому шторму.

Особое внимание уделено ключевым сигнальным путям, таким как TGF- β и SPP1, которые играют важную роль в развитии воспалительных реакций. Установлено, что роль модуля активности TGF- β постепенно возрастает вместе с тяжестью заболевания, в то время как активность сигнального пути SPP1 достигает максимума при тяжелой форме дельта-COVID-19.

Таким образом, исследование расширяет понимание молекулярных механизмов, лежащих в основе тяжелых форм дельта-COVID-19, и подчеркивает важность анализа открытости хроматина для изучения регуляции иммунного ответа.

Множественное блочное выравнивание биологических последовательностей

А.В. Алексеевский¹, Б.Э. Нагаев, И.М. Левин², Д.В. Самборский³,
И.С. Русинов¹

¹НИИ ФХБ МГУ имени М.В. Ломоносова Москва, Ленинские Горы
дом 1, строение 40)

²ФББ МГУ имени М.В. Ломоносова Москва, Ленинские Горы дом
1, строение 73)

³НИЦ «Курчатовский институт» - НИИСИ, Москва,
Нахимовский пр-кт, д.36 к.1
aba@belozersky.msu.ru

Блочное выравнивание последовательностей ДНК или белков представляет входные последовательности в виде блоков, плюс блоков и минус блоков. Каждая буква входных последовательностей входит однократно в один блок.

Плюс блок - это достоверное выравнивание фрагментов входных последовательностей по выбранному критерию. Плюс блоки делятся на стабильные (s-блоки) по одному фрагменту из каждой входной последовательности, полустабильные (h-блоки) по одному фрагменту, но не из каждой входной последовательности, (от hemi stable), блоки с повторами (r-блоки) содержат более одного фрагмента из хотя бы одной последовательности, уникальные последовательности (u-блоки). Минус блок (m-блок) также состоит из фрагментов входных последовательностей в которых невозможно найти ни одного плюс блока см. рис.1В.

Программа NPGe блочного выравнивания прокариотических геномов (<https://github.com/npge/npge>) была разработана нами ранее и использовалась в разных работах [1,2], Результат работы NPGe называется нуклеотидным пангеномом (NPG). Для рисунка 1 используется NPG шести штаммов бруцелл пяти видов. Каждый геном представлен двумя хромосомами

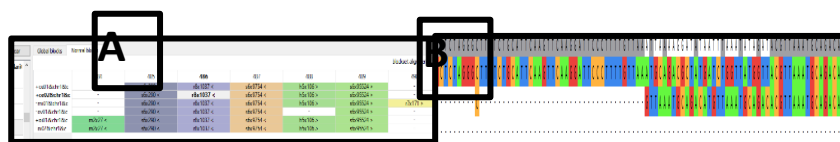


Рис.1. NPG Bru 6 **A.** Все 6 последовательностей показаны как последовательности блоков:m2x27, s6x9754, h5x106, s6x95524, r7x17. Первый символ тип блока, число до разделителя «x» – число последовательностей в блоке, число после x – длина блока **B.**Выравнивание (точнее его отсутствие) в минус блоке m4x93. Идентичных колонок 8.17% от длины блока.

В эволюции белков крупных изменений меньше, чем в геномах, но они есть. Актуальна в разрабатываемом множественном блочном выравнивании последовательностей белков допустимость МИНУС блоков (рис.2А), в которых нет

выравнивания

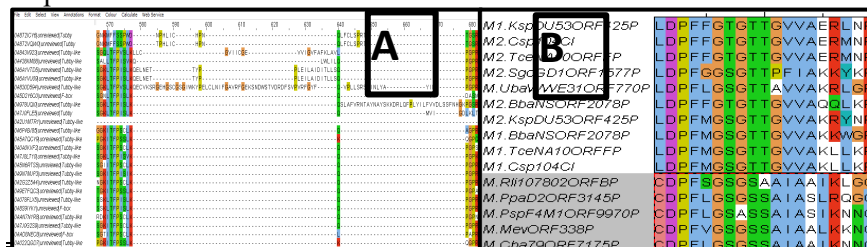


Рис 2. А. Полноразмерное выравнивание 672 последовательностей семейства Tubby-like F-box protein [3]. N-terminal F-box domain and а C-terminal Tubby domain В >600 последовательностях не показанных одна буква в колонке 640 и гэпов столько только же, сколько в последних. В. Блоки, поддерживающие ветвь дерева 15 ДНК метилтрансфераз (Мтаз) делящую дерево на клады из с 10 Мтазами и 5ю Мтазами. Все последовательности содержат известный SAM-связывающий мотив FX[GA]X[GA]

В полноразмерных выравниваниях белков из одного семейства с одинаковой доменной архитектурой, потребность в блочном выравнивании зачастую очевидна (рис.2.А) На сегодня частично реализован первый этап проекта. Во входном множественном выравнивании и филогенетическом дереве найти поддержку ветвей блоками достоверного выравнивания в подвыравнивании «листьев» ветви, не продолжающимися на дополнительное подмножество последовательностей (рис.2В). Такая программа может иметь ценность как наглядное подтверждение адекватности дерева входному выравниванию.

Следующие этапы проекта нахождение блоков в выравнивании и реконструкция филогении по ним; построение блочного выравнивания de novo.

Исследование выполнено в рамках государственного задания МГУ имени М.В. Ломоносова и по программе частичной при частичной поддержке НИЦ «Курчатовский институт» - НИИСИ (части 1 и 3), проект задуман в рамках гранта РФФИ 21-14-00135.

[1] Goryunov D.V., Nagaev B.E., Nikolaev M.Yu, Alexeevski A.V.,

- Troitsky A.V. (2015) Moss Phylogeny Reconstruction Using Nucleotide Pangenome of Complete Mitogenome Sequences Biochemistry (Moscow), том 80, № 11, с. 1522-1527
- [2] Panova V.V., Dolinnaya N.G., Novoselov K.A., Savitskaya V.Y., Chernykh I.S., Kubareva E.A., Alexeevski A.V., Zvereva M.I. (2023) Conserved G-quadruplex-forming sequences in mammalian TERT promoters and their effect on mutation frequency Life, MDPI (Basel, Switzerland), том 13, № 7, с. 1478
- [3] Bano N, Aalam S, Bag SK. (2022)Tubby-like proteins (TLPs) transcription factor in different regulatory mechanism in plants: a review. Plant Mol Biol. Dec;110(6):455-468.

Геномная карта клеточных моделей

В.А. Арзуманян¹, М.П. Королев², Е.В. Поверенная¹

¹*Федеральное государственное бюджетное научное учреждение
«Научно-исследовательский институт биомедицинской химии
имени В.Н. Ореховича», Москва, k.poverennaya@gmail.com*

²*Федеральное государственное автономное образовательное
учреждение высшего образования «Московский физико-
технический институт (национальный исследовательский
университет)», Москва*

Клеточные линии по-прежнему являются незаменимым инструментом в биомедицинских исследованиях благодаря их доступности и относительной молекулярной стабильности. Последние работы по анализу массива образцов одной клеточной линии опровергли постулат об их единообразии. Выполненное в 2018 году масштабное тестирование противоопухолевых препаратов на 27 образцах клеточной линии MCF-7 показал наличие как минимум двух подтипов клеток, по разному реагирующих на воздействие лекарств [1]. Годом позже было

проведено широкомасштабное исследование образцов клеточной линии HeLa из 14 лабораторий, которое показало, что образцы одной клеточной линии могут существенно отличаться между собой, формируя три подтипа клеток [2].

В данном исследовании была проведена оценка хромосомной дивергенции десяти наиболее часто упоминаемых в базе PubMed клеточных линий: HeLa, HEK293, MCF-7, A-549, HepG2, MDA-MB-231, HCT116, THP-1, HT-29 и PC3. Поиск данных был проведен по следующим критериям: технология Illumina, данные не старше 10 лет, >20 млн прочтений, длина чтений >60 п.н, метод RNA-seq.

Сравнительный анализ цитогенетических и полногеномных данных позволил определить степень гетерогенности этих линий, в том числе частоту появления специфических хромосомных aberrаций. Наибольший уровень геномной нестабильности был выявлен у линии HeLa (рак шейки матки) — во всех проанализированных образцах отсутствовало стабильное состояние хотя бы одной хромосомы. В то же время клеточная линия HCT116 (колоректальный рак) продемонстрировала наименьшее количество уникальных хромосомных перестроек и общую стабильность геномной структуры.

Дополнительно была проведена мультиомиксная характеристика клеточной линии HepG2, которая показала стабильность на уровне генома. Результаты анализа демонстрируют, что геномные вариации могут компенсироваться на этапах постгеномной регуляции. В дальнейшем планируется создание открытой базы данных, включающей мультиомиксные профили клеточных линий. Такой ресурс будет способствовать более точной интерпретации экспериментальных данных и позволит снизить влияние биологической изменчивости *in vitro* на результаты исследований.

-
1. Ben-David U, Siranosian B, Ha G, Tang H, Oren Y, Hinohara K, Strathdee CA, Dempster J, Lyons NJ, Burns R, Nag A, Kugener G, Cimini B, Tsvetkov P, Maruvka YE, O'Rourke R, Garrity A, Tubelli AA, Bandopadhyay P, Tsherniak A,

- Vazquez F, Wong B, Birger C, Ghandi M, Thorner AR, Bittker JA, Meyerson M, Getz G, Beroukhi R, Golub TR. (2018) Genetic and transcriptional evolution alters cancer cell line drug response. *Nature*. 560(7718):325-330.
2. Liu, Y., Mi, Y., Mueller, T. et al. (2019) Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nat Biotechnol*, 37:314–322
-

Молекулярные механизмы регуляции ферроптоза: анализ омиксных данных транскриптома и протеома

М.В. Баранникова^{1,2}, В.К. Сулягин³, В.С. Ледяева³, Д.А.
Корженевский², М.А. Сорокина⁴, А.Г. Шохина³, О.М. Кудряшова²

¹*Московский физико-технический институт (национальный
исследовательский университет), Долгопрудный, Институтский
пер., д. 9*

²*ФГБУ Федеральный центр мозга и нейротехнологий ФМБА
России, Москва, улица Островитянова, 1, стр. 10,
barannikova.mv@phystech.edu*

³*Лаборатория экспериментальной онкологии НИИ
трансляционной медицины ФГАОУ ВО РНИМУ им. Н.И. Пирогова
Минздрава России (Пироговский Университет), г. Москва,
Рахмановский пер, д. 3*

⁴*ФГАОУ ВО «Российский национальный исследовательский
медицинский университет им. Н.И. Пирогова» Минздрава России,
Москва*

Ферроптоз - открытый в 2012 году железозависимый тип
клеточной гибели. Ферроптоз обусловлен перекисным окислением
липидов и является одним из механизмов, лежащих в основе

нейродегенераций [1]. Для изучения и индукции ферроптоза широко используются реагенты, воздействующие на разные элементы антиоксидантной xC-/GSH/Gpx4 оси клетки: Erastin (эрастин, блокирует антипортерную xC- систему цистина-глутамата), BSO (L-бутионинсульфоксимин, останавливает синтез глутатиона GSH), ML210 (ингибирует Gpx4 - глутатионпероксидазу-4). Однако сходства и различия молекулярных механизмов запуска ферроптоза разными индукторами до сих пор остаются неясными.

Данное исследование представляет собой комплексный анализ данных РНК-секвенирования и антаргетной shotgun масс-спектрометрии клеточной линии Pfa1 для поиска ключевых регуляторов ферроптотической гибели при выключении различных элементов антиоксидантной xC-/GSH/Gpx4 оси клетки. Анализ проводился в двух временных точках: 24 и 48 часов после индукции ферроптоза (ранний и поздний ответы). Использовались три химических индуктора: Erastin, BSO, ML210; также производился тамоксифен-индуцируемый нокаут Gpx4. Были проведены анализ дифференциальной экспрессии (limma), анализ насыщения по термам геной онтологии (clusterProfiler) и анализ активности программ транскрипционных факторов (decoupler). Дополнительно исследовались публичные данные РНК-секвенирования датасетов GSE72991 и GSE240106.

По результатам дифференциальной экспрессии (ДЭ) как на транскриптомном, так и на протеомном уровнях, поздний ответ был охарактеризован как более общий ответ клетки на индукцию ферроптоза с большим количеством общих генов и белков между разными индукторами. Для каждого индуктора также выделились и уникальные гены, описанные в литературе как драйверы или супрессоры ферроптоза. В том числе, при индукции Erastin и BSO высоко экспрессировалась гемооксигеназа1 (Hmox1), которая может играть двойственную роль в окислительном стрессе.

Одной из интересных находок оказалось разделение индукторов на две большие группы по анализу активности программ транскрипционных факторов (ТФ). При воздействии Erastin и BSO высокую активность имели ТФ Atf4 и Ppargc1a; при воздействии на Gpx4 (как химическим реагентом ML210, так и с помощью нокаута Gpx4) повышали активность ТФ Srebf1, Srebf2.

Atf4 - многофункциональный ТФ, но в совокупности с анализом генных онтологий по данным транскриптома и протеома, мы полагаем, что основной его функцией при индукции ферроптоза является активация трансмембранного транспорта клетки с такими игроками как Slc7a11 (транспортная субъединица антипортера цистина-глутамата), Slc40a1 (ферропортин). Ppargc1a отвечает за метаболизм жирных кислот (ЖК) в клетке, в зависимости от соотношения моно- и полиненасыщенных форм которых определяется судьба клетки [2]. Активация этих ТФ подтвердилась и при анализе активности ТФ публичных данных с нокаутом антипортера системы xC-.

Srebf1 и Srebf2, в свою очередь, являются регуляторами мевалонатного пути, дальнейший анализ ssGSEA сигнатуры генов мевалонатного пути показал увеличение экспрессии генов синтеза холестерина (Fdft1, Sqle, Sc5d) при индукции ферроптоза через воздействие на Gpx4.

Различия клеточного ответа, выявленные в ходе комплексного анализа, и сравнение двух временных точек могут помочь в понимании молекулярных механизмов ферроптоза и нейродегенераций.

1. Dixon, S.J., Lemberg, K.M., Lamprecht, M.R., Skouta, R., Zaitsev, E.M., Gleason, C.E., Patel, D.N., Bauer, A.J., Cantley, A.M., Yang, W.S., Morrison, B., Stockwell, B.R., 2012. Ferroptosis: an iron-dependent form of nonapoptotic cell death. *Cell* 149, 1060–1072.
2. Yamane, D., Hayashi, Y., Matsumoto, M., Nakanishi, H., Imagawa, H., Kohara, M., Lemon, S.M., Ichi, I., 2022. FADS2-dependent fatty acid desaturation dictates cellular sensitivity to ferroptosis and permissiveness for hepatitis C virus replication. *Cell Chem. Biol.* 29, 799-810.e4.

Совместный тканеспецифичный механизм действия мастер-регуляторов и G-квадруплексов, открытый с помощью моделей глубокого обучения

Артём Башкатов¹, Андрей Андеасян¹, Дмитрий Коновалов¹, Алан Герберт^{1,2}, Мария Попцова¹

¹Международная лаборатория биоинформатики, НИУ ВШЭ,
Москва, Покровский бульвар, 11 mpoptsova@hse.ru

²InsideOutBio, Чарльзтаун, Массачусетс, США

G-квадруплексы (GQ) — это вторичные структуры ДНК, формируемые G-флипами, которые играют роль в регуляции генов и организации хроматина. В данной работе исследуется роль G-флипов в тканевой специфичности. Нами был разработан метод глубокого обучения на основе омиксных данных для предсказания G-флипов на уровне всего генома в 14 типах тканей человека. Модель обучалась на экспериментальных картах GQ-формирующих последовательностей, участках открытого хроматина, данных о локализации РНК-полимеразы, гистонов и сайтов связывания транскрипционных факторов. Обучающая выборка основывалась на аннотированных последовательностях GQ уровня 4–6 базы данных Endoquad, а предсказания проверялись на полном наборе Endoquad уровня 1–6. Было обучено 14 моделей для 14 типов тканей, используя тканеспецифичный набор омиксных признаков, после чего было сгенерировано 14 полногеномных карт расположения GQ. Для выявления тканеспецифичных регуляторных паттернов предсказанные GQ в промоторах были разделены на «общие» и «тканеспецифичные». Было обнаружено статистически значимое обогащение тканеспецифичных GQ в сайтах связывания мастер-регуляторов (MRG), в тканеспецифичных участках открытого хроматина, а так же в областях связывания белков, модулирующих формирование R-петель.

В целом, фреймворк DeepGQ представляет собой инструмент для изучения тканеспецифичного распределения G-квадруплексов (GQ) и их регуляторного влияния. Исследуя взаимодействие между GQ, промоторами генов и мастер-регуляторами (MRG), настоящее исследование расширяет понимание геномной регуляции и открывает новые перспективы для изучения функциональной роли GQ при нормальной работе и заболеваниях. В будущем этот метод может быть адаптирован для других организмов или интегрирован в стратегии персонализированной медицины для таргетного воздействия на

тканеспецифичные регуляторные элементы.

**Различия в дорсальном и вентральном отделах
гиппокампа при прогрессировании болезни Альцгеймера**

Н.А. Бугаев-Макаровский¹, С.Р. Мареева¹, П.А. Фортигина^{1,2}, О.А. Антонова¹, В.Э. Акимов¹, О.Д. Фатеев¹, А.А. Бородинова², С.М. Юдин¹, А.В. Махотенко¹, А.А. Кескинов¹, В.С. Юдин¹, Е.А. Снигирь¹, О.С. Ганчарова¹, П.М. Балабан², Д.В. Светличный¹, В.И. Скворцова³

¹Федеральное государственное бюджетное учреждение
«Центр стратегического планирования и управления медико-
биологическими рисками здоровью» Федерального медико-
биологического агентства (ФГБУ «ЦСП» ФМБА России). 119121,
Россия, г. Москва, ул. Погодинская, д. 10, стр. 1.
nbugaev@cspfmba.ru

²Федеральное государственное бюджетное учреждение
науки Институт высшей нервной деятельности и
нейрофизиологии Российской академии наук (ФГБУН ИВНДиНФ
РАН). 117485, Россия, г. Москва, ул. Бутлерова, д. 5А.

³Федеральное медико-биологическое агентство (ФМБА
России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.

Болезнь Альцгеймера (БА) — это тяжелое нейродегенеративное заболевание, молекулярный механизм развития которого до конца не определен, а терапевтические подходы в настоящее время представлены лишь симптоматическим лечением. В данном исследовании использовалась животная модель — линия мышей 5xFAD, содержащая 5 генетических вариантов, которые ассоциированы с наследственной формой БА. Это позволяет не только моделировать само заболевание, но и наиболее полно исследовать механизм его развития и прогрессирования с течением времени, в том числе на молекулярно-

генетическом уровне.

В данной работе выполнено одноклеточное секвенирование транскриптома (scRNAseq) 32 образцов вентрального и дорсального гиппокампа, полученных из 16 мышей линии 5xFAD и WT в 3 возрастных группах (8, 16 и 22 месяца). Вычислительный анализ выявил отличия клеточной композиции между зонами гиппокампа. В частности, выявлено увеличение ExN.CA1.2, ExN.CA3, ГАМКергических нейронов и уменьшение ExN.CA1.1, ExN.DG и олигодендроцитов в вентральном гиппокампе для WT и 5xFAD групп. Исследование изменений генной экспрессии позволило выявить, что к 8 месяцам клеточные популяции имеют значительные отличия между WT и 5xFAD группами. Анализ данных одноклеточного секвенирования транскриптома позволил выявить подтипы микроглиальных и астроцитарных клеток, которые сопровождают нейродегенеративный процесс в модельной линии 5xFAD. Вычислительный анализ выявил генные модули, которые коэкспрессируются в подтипах глиальных клеток, ассоциированных с нейродегенерацией. В частности, выявлены генные модули, которые связаны с изменением нейропластичности, формированием синапсов, окислительным стрессом, а также иммунной активацией микроглии.

Выполненные транскриптомные исследования дополнены иммуногистохимическим определением амилоида и установлено, что в вентральной зоне более выражено накопление амилоида. Более того, именно в вентральной зоне снижена активность экспрессионных сигнатур нейродегенеративной микроглиальной трансформации. Для определения ключевых факторов, которые контролируют активность генов дисфункциональной микроглии, выполнена вычислительная реконструкция генной сети и построена модель для оценки роли пертурбации экспрессии транскрипционных факторов. Идентифицированы основные регуляторы, координирующие экспрессию генных модулей активированной микроглии. В частности, выявлено, что транскрипционные факторы Maf, Mitf, Zfhx3 являются наиболее значимыми регуляторами экспрессии маркерных генов дисфункциональной микроглии.

В данной работе на модели 5xFAD охарактеризованы молекулярные изменения в области гиппокампа при развитии БА,

выявлены временные отличия клеточной композиции и молекулярной активности в дорсальном и вентральном гиппокампе, а также исследованы регуляторные факторы, которые влияют на выраженность активности генных модулей, определяющих фенотипическое состояние микроглии. Полученные результаты могут быть использованы для определения предиктивных биомаркеров БА, а также предполагают возможные мишени для воздействия посредством генной терапии.

**Перспективное клиническое исследование
биоинформатической платформы анализа данных РНК-
секвенирования OncoBox для персонализированного
назначения таргетных препаратов онкобольным**

А.А. Буздин¹, М.И. Сорокин², А.В. Гаража², М.В. Сунцова², В.С. Ткачев², Е.В. Поддубская¹, Н.М. Гайфуллин³, Т.В. Сушинская⁴,
Д.С. Ланцов⁵, В.И. Борисов⁶, Д.Р. Насхлеташвили⁷, К.А. Ильин⁸,
А.П. Серяков⁸, А.А. Глускер¹, А.А. Моисеев¹

¹Сеченовский университет, Москва, ул. Трубецкая, д.8,
bu3din@mail.ru

²ООО «Онкобокс», Долгопрудный, Лихачевский проезд, д.6 стр.1

³МГУ им. М.В. Ломоносова, 119991, Москва

⁴МНИОИ им. П.А. Герцена, 125284, Москва

⁵ГБУЗ КО «Калужский областной клинический онкологический
диспансер», 242002, г. Калуга, ул. Вишневского, 2

⁶Московская I-я онкологическая больница, Москва, 105005

⁷Российский онкологический научный центр им. Н.Н. Блохина,
Москва, 115478, Россия

⁸Медицинский холдинг «СМ-Клиника», 105120 Москва, Россия

Изучение экспрессии генов в опухоли позволяет выявить активированные и подавленные молекулярные мишени противоопухолевых препаратов. Здесь мы сообщаем о результатах проспективного клинического исследования использования биоинформатического анализа данных РНК-секвенирования для персонализированной терапии рака. Транскриптомные профили анализировались с помощью аналитической платформы OncoBox, которая выявляет изменения экспрессии генов и молекулярных путей -мишеней терапии и формирует персонализированный рейтинг таргетных препаратов.

В исследовании в качестве контроля использовались профили РНК секвенирования, полученные для тканей здоровых доноров, погибших в результате дорожно-транспортных происшествий [1], с использованием тех же протоколов, оборудования и реагентов. Таким образом, OncoBox оперировал несколькими нормальными образцами по сравнению с каждой опухолевой тканью, что теоретически может уменьшить экспериментальный шум за счет подсчета статистики и лучшей дискриминации дифференциально экспрессированных генов; в протоколе исследования OncoBox в качестве биоматериала опухоли использовались блоки парафинизированной ткани.

Сравнивая профили экспрессии опухоли и нормы, рассчитывали соотношение экспрессии отдельных молекулярных мишеней препаратов к норме, а также дифференциальную активацию молекулярных путей, на которые направлено действие лекарств [2]. Каждый образец экспрессии опухолевых генов нормировали на профили экспрессии генов тех же нормальных тканей, например, для образцов колоректального рака в качестве нормы использовали нормальную толстую кишку, для образцов рака легкого - нормальные легкие и т. д. На основании полученных молекулярных данных каждому таргетному препарату присваивался индекс BES (balanced efficiency score), отражающий его потенциальную эффективность для лечения отдельной опухоли. BES рассчитывали с помощью алгоритма, который суммировал две составляющие: первая отражает вклад молекулярных путей, на которые нацелен конкретный препарат (DESMP), вторая - вклад отдельных молекулярных генов-мишеней (DESTG). Для различных классов лекарств используются различные весовые коэффициенты

для DESMP и DESTG, варьирующие от -1 до 1,5. Препараты были разделены на функциональные группы в зависимости от способа действия и молекулярной специфичности, как подробно описано в работе Ткачева и других [2]. В тесте Oncobox 137 противоопухолевых таргетных препаратов были ранжированы в соответствии с их прогнозируемой эффективностью, отраженной в значениях BES. Препарат со значением BES выше 0,1 считался потенциально эффективным, в то время как препарат со значением BES, равным или ниже 0,1, считался неэффективным. Эта информация была предоставлена лечащим врачам всех пациентов, включенных в исследование.

Всего в исследование было включено 239 взрослых пациентов с солидными злокачественными опухолями продвинутых стадий: 135 получали лекарственную терапию, остальные проходили паллиативное лечение или лучевую терапию, либо умерли до начала терапии. Препараты, рекомендованные Oncobox, были назначены в 59 % случаев терапии. В остальных случаях пациенты получали нетаргетную химиотерапию или же таргетную терапию, которую платформа Oncobox отнесла к неэффективной (контроль). Пациенты в группе Oncobox имели значительно большее число линий предыдущей терапии по сравнению с контрольной группой, однако мы наблюдали более длительную беспрогрессивную выживаемость (PFS) в группе Oncobox. Кроме того, post-hoc анализ показал, что продолжительность временного интервала между взятием биопсии и профилированием опухоли существенно влияет на прогностическую способность платформы, оптимально – не более 7 месяцев от взятия материала до секвенирования. Мы зафиксировали статистически значимую разницу в PFS между группой Oncobox и контрольной группой с коэффициентом опасности 0,45 ($p\text{-value}=0,023$). Кроме того, медианная выживаемость без прогрессирования в группе Oncobox была на 10 месяцев больше, чем в контрольной группе.

Доля объективных ответов в группе Oncobox составила 24% против 4% в контрольной группе. Контроль над заболеванием составил 86% против 50% в группе сравнения [3].

Эти результаты позволяют предположить, что транскриптомное профилирование позволяет получать клинически

значимые терапевтические рекомендации по лечению онкобольных и может улучшить показатели контроля над заболеванием при солидных опухолях.

Список литературы:

1. M. Suntsova et al. (2019) Atlas of RNA sequencing profiles for normal human tissues, *Sci Data*, **6**: 36.
2. V. Tkachev et al. (2020) Oncobox Method for Scoring Efficiencies of Anticancer Drugs Based on Gene Expression Data, *Methods Mol Biol*, **2063**: 235-255.
3. M. Sorokin et al. (2025) Prospective trial of the Oncobox platform RNA sequencing bioinformatic analysis for personalized prescription of targeted drugs, *Comput Biol Med*, **187**: 109716.

Экспериментальный поиск ассоциации генетических вариантов с эндокринными заболеваниями, проведенный на 4 302 пациентах российской популяции с использованием полноэкзомного и панельного секвенирования

А.А.Буздин¹, М.А.Золотовская², Д.В.Луппов³, Д.Н.Штокало²,
П.А.Пугачева², А.А.Слепухина², О.О.Голоунина²,
Г.А.Мельниченко², Н.Г.Мокрышева², В.П.Чехонин², И.И.Дедов²

¹ФГАОУ ВО Первый МГМУ им. И. М. Сеченова Минздрава России,
119991, г. Москва, ул. Трубецкая, д. 8, buzdin@oncobox.com

²ФГБУ "НМИЦ Эндокринологии" Минздрава России,
117292, г. Москва, ул. Дмитрия Ульянова, д. 11

³МФТИ, 141701, Московская область, г. Долгопрудный,
Институтский переулок, д.9

Эндокринные патологии, включая такие заболевания, как диабет и дисфункции эндокринных желез, часто связаны с генетической предрасположенностью. Оценка соответствующих генетических факторов важна для определения риска развития

эндокринных патологий у отдельного пациента до их манифестации.

В данном масштабном экспериментальном исследовании изучена ассоциация эндокринных заболеваний с генетическими вариантами, вариациями числа копий (CNV) и мутационной нагрузкой молекулярных путей у 4302 эндокринологических пациентов с 409 диагнозами по МКБ-10, прошедших ДНК-тестирование с использованием секвенирования нового поколения в НМИЦ Эндокринологии (Москва) с ноября 2017 года по январь 2024 года. Мы проанализировали редкие генетические варианты, изменяющие структуру кодируемых белков, используя три контрольные когорты (gnomAD3, RUSeq healthy, наша экспериментальная группа). В результате мы выявили 143 ассоциированных варианта для сахарного диабета и 188 генетических вариантов для других 18 различных групп диагнозов по МКБ-10, включая 25 % и 30 % ранее не описанных вариантов, соответственно. Кроме того, мы исследовали агрегацию генетических вариантов по отдельным генам и их функциональным ансамблям (молекулярным путям) и выявили 105 и 101 ассоциацию с диагнозами МКБ-10, соответственно. Кроме того, мы выявили 35 патогенных и 91 вероятно патогенный CNV у 925 пациентов с профилями полноэкзомного секвенирования. Среди них 10 и 43 CNV, соответственно, не были описаны ранее. В общей сложности мы обнаружили статистически значимые ассоциации между CNV и эндокринными патологиями для 168 генов.

Мы также изучили взаимосвязь между генами лейкоцитарного антигена человека (HLA) и 13 эндокринными заболеваниями с помощью экспериментальных профилей полноэкзомного секвенирования, полученных для 895 пациентов. Кроме того, была проведена оценка неравновесного сцепления выявленных аллелей в контексте соответствующих диагнозов. Выявлено 45 статистически значимых ассоциаций между аллелями HLA и конкретными диагнозами эндокринной патологии. Среди них 33 были описаны впервые, а 12 в предыдущих сообщениях уже связывали с диабетом 1 типа. В общей сложности, в нашем исследовании 17 аллелей были связаны с диабетом 1-го типа и четыре - с другими формами диабета. Кроме того, впервые HLA аллели удалось ассоциировать и с другими эндокринными

заболеваниями. Так, три аллеля были связаны с ожирением, пять - с адреногенитальными заболеваниями, три - с гипогликемией и три - с преждевременным половым созреванием. Единичные аллели оказались связаны с врожденным гипотиреозом без зоба, гиперфункцией гипофиза, гиперфункцией адреномедуллярной системы и низкорослостью вследствие эндокринного заболевания.

Проведенное нами исследование показывает, что раннее HLA-типирование и более широкая генетическая диагностика могут выявлять генетические факторы риска многих эндокринных заболеваний, для некоторых заболеваний это было доказано впервые. Полученные результаты расширяют понимание механизмов патогенеза эндокринной системы и указывают на новые потенциальные терапевтические мишени.

По результатам исследования была также создана самая крупная на сегодня клинически аннотированная публичная база эндокринных патологий с доступными генетическими профилями.

Кроме того, создан оригинальный пайплайн аннотации первичных генетических данных, который может быть использован для быстрой автоматизированной обработки профилей ДНК пациентов с широким кругом патологий и здоровых индивидуумов.

Мы благодарны коллегам, проводившим медицинское и генетическое консультирование пациентов и их лечение, аннотацию клинических профилей, а также организовывавшим и проводившим генетическое профилирование и отдельные этапы статистического и биоинформатического анализа, чей вклад был важен для проведения этой работы: С.А.Румянцева, Е.А.Трошину, М.В.Шестакову, О.Б.Безлепкину, В.А.Петеркову, М.Помазному, С.Баранова, А.Г.Емельянову, А.О.Алексееву, А.А.Емельянову, А.Л.Новоселова, А.А.Матросову, М.В.Сунцову, С.В.Попова, Е.В.Плаксину, В.М.Петрова, Н.Гусельникову, В.В.Захарову, Ж.Е.Белую, Е.В.Бондаренко, М.В.Воронцову, В.С.Ткачева, М.И.Сорокина.

An atlas of transcribed gene regulatory elements reveals the diverse molecular phenotypes of human skeletal muscles

Andrey I. Buyan¹, Vladimir A. Nozdrin², Guzel R. Gazizova³,
Georgy A. Meshcheryakov¹, Elena I. Shagimardanova³,
Oleg A. Gusev³, Ivan V. Kulakovskiy¹

¹*Institute of Protein Research, Russian Academy of Sciences,
Pushchino, Russia, ivan.kulakovskiy@gmail.com*

²*Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov
Moscow State University, Moscow, Russia*

³*Research Laboratory «Regulatory Genomics», Institute of
Fundamental Medicine and Biology, Kazan Federal University, Kazan,
Russia*

More than 600 unique types of skeletal muscles constitute up to 40% of the total weight of the human body [1]. The skeletal muscles differ in anatomical position, morphology, origin, and function, although the diversity of their gene expression profiles, the molecular phenotypes, remain poorly explored. Depending on the tissue ontology, functional needs, and fiber type composition, the gene expression profile can change significantly, affecting muscle contraction speed, stiffness, resting tension, or even insulin resistance and susceptibility to various myopathies [2]. Recent studies revealed striking differences between skeletal muscle transcriptomes in mice [3] and livestock species such as pigs [4], sheep [5], horses [6], and chickens [7], yet, in humans, the gene expression patterns were explored in selected leg muscles only [8-9].

In this study we report the results of FANTOMUS, functional annotation of muscle promoters and enhancers, a collaborative project where we performed cap analysis of gene expression (CAGE-Seq) using 222 autopsy samples of 75 skeletal muscles from 4 individuals. We identified more than 30 thousand transcribed regulatory elements (TREs), including nearly a thousand bidirectionally transcribed enhancers. More than 60% of TREs exhibited tissue-specific activity,

with the most distinct expression profiles found in the skeletal muscles of the eye, tongue, and diaphragm. Comparing muscles of different body parts and fiber composition, the activity of promoters of Hox genes and genes encoding myosin light and heavy chains agreed with the muscle origin and functional roles, respectively. Considering individual genes, we detected upregulated cardiac troponin T TNNT2 and downregulated muscle aging marker CFAP61 expression in eye samples, suggesting specific Ca^{2+} -dependent regulation and slower aging of these muscles, respectively [10]. Moreover, we observed a significant differential expression of genes CAPN3 and FKRP, which mutations cause calpainopathy and dystroglycanopathy, respectively, between muscles of different susceptibility to the corresponding diseases. Besides changes in gene expression, we detected differential promoter usage across skeletal muscles, revealing eye-specific prevalence of SEPTIN9 short isoform. This gene encodes a cytoskeleton protein that contains a microtubule-associated domain in most but not the extraocular muscles [11].

To identify key transcription factors that drive the observed gene expression diversity, we performed motif activity response analysis. As a result, we found about 4 hundred transcriptional regulators orchestrating tissue-specific gene expression in muscles, with myocyte enhancer factors MEF2, YY1, and steroid hormone receptors being among the most significant.

Finally, we called more than 16 thousand single-nucleotide variants revealing more than 6 thousand allele-specific variants with significant allelic imbalance in at least one skeletal muscle type. More than half of these variants were annotated as expression quantitative trait loci from GTEx [12] or allele-specific transcription factor binding sites from ADAstra [13]. Using stratified linkage disequilibrium score regression [14], we found significant enrichment for the normalized muscle volume [15], suggesting that this trait has an extensive regulatory genetic basis.

Taking together, we constructed an atlas of transcribed regulatory elements in diverse human skeletal muscles and performed the allele-specific analysis of promoters and enhancers activities. We believe that FANTOMUS results will be useful for studying gene expression patterns and variant effects in healthy muscles and in myopathies.

The study is supported by the Ministry of Science and Higher Education of the Russian Federation (No. 075-15-2021-601).

1. Frontera, W.R., Ochala, J. (2014) Skeletal Muscle: A Brief Review of Structure and Function, *Calcif Tissue Int*, **96**:183–195.
2. Hoh, J.F.Y. (2023) Developmental, physiologic and phylogenetic perspectives on the expression and regulation of myosin heavy chains in mammalian skeletal muscles, *J Comp Physiol B*, **193**:355–382.
3. Terry E.E., Zhang X., Hoffmann C. et al. (2018) Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues, *eLife*, **7**:e34613.
4. Jin, L., Tang, Q., Hu, S. et al. (2021) A pig BodyMap transcriptome reveals diverse tissue physiologies and evolutionary dynamics of transcription, *Nat Commun*, **12**:3715.
5. Leng, D., Huang, Z., Bai, X. et al. (2024) Gene expression profiles in specific skeletal muscles and meat quality characteristics of sheep and goats, *Sci Data*, **11**:1390.
6. Lee H.Y., Kim J.Y., Kim K.H. et al. (2020) Gene Expression Profile in Similar Tissues Using Transcriptome Sequencing Data of Whole-Body Horse Skeletal Muscle, *Genes*, **11**:1359.
7. Kui, H., Ran, B., Yang, M. et al. (2022) Gene expression profiles of specific chicken skeletal muscles., *Sci Data*, **9**:552.
8. Abbassi-Daloii T., El Abdellaoui S., Voortman L.M. et al. (2023) A transcriptome atlas of leg muscles from healthy human volunteers reveals molecular and cellular signatures associated with muscle location, *eLife*, **12**:e80500.
9. Ruonan T., Ziwei X., Dengfeng R. et al. (2023) MSdb: An integrated expression atlas of human musculoskeletal system, *iScience*, **26**:106933.
10. Tumasian R.A., Harish A., Kundu G. et al. (2021) Skeletal muscle transcriptome in healthy aging, *Nat Commun*, **12**:2014.
11. Kuzmić M., Linares G.C., Fialová J.L. et al. (2022) Septin-microtubule association via a motif unique to isoform 1 of septin 9 tunes stress fibers, *J Cell Sci*, **135**:jcs258850.
12. Lonsdale J., Thomas J., Salvatore M. et al. (2013) The Genotype-Tissue Expression (GTEx) project, *Nat Genet*, **45**:580-585.
13. Abramov S., Boytsov A., Bykova D. et al. (2021) Landscape of allele-specific transcription factor binding in the human genome, *Nat Commun*, **12**:2751.

14. Finucane H., Bulik-Sullivan B., Gusev A. et al. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics, *Nat Genet*, **47**:1228-1235.
15. van der Meer D., Gurholt T.P., Sønderby I.E. et al. (2022) The link between liver fat and cardiometabolic diseases is highlighted by genome-wide association study of MRI-derived measures of body composition, *Commun Biol*, **5**:1271.

**Поиск подписей отрицательного отбора на переходы между
стоп-кодонами в данных полиморфизма *Homo sapiens* и
*Drosophila melanogaster***

О.А. Вахрушева

*Сколковский институт науки и технологий, Территория
Инновационного Центра «Сколково», Большой бульвар д. 30, стр.
1, Москва 121205, Россия, O.Vakhrusheva@skoltech.ru*

У большинства организмов для терминации трансляции используется три стоп-кодона (TAA, TAG и TGA). Поскольку распознавание любого из трех стоп-кодонов в норме должно приводить к терминации трансляции, обычно их рассматривают как функционально эквивалентные. Однако экспериментальные данные указывают на то, что стоп-кодоны отличаются по точности [1, 2]. В некоторых случаях, вместо терминации трансляции, рибосома может прочитывать стоп-кодон как кодирующий и затем продолжать чтение 3'-нетранслируемой области [3]. Иногда сквозное прочтение стоп-кодонов имеет функциональное значение: так, ряд вирусов использует регулируемое прочтение стоп-кодонов для расширения белкового репертуара [4, 5]. Однако, по всей видимости, большая часть событий сквозного прочтения стоп-кодонов является результатом ошибок [6]. В ряде экспериментальных работ было показано, что частота спонтанного сквозного прочтения отличается для разных стоп-кодонов: сквозное прочтение наиболее вероятно для TGA и наименее вероятно для TAA [1, 2]. При этом очевидной связи между точностью разных стоп-кодонов и частотой их встречаемости в геномах разных видов нет. Так, в геноме *D. melanogaster* чаще всего используется самый

«аккуратный» стоп-кодон TAA [7]. В то же время в геноме человека и других млекопитающих чаще всего встречается стоп-кодон TGA, наиболее подверженный ошибочному прочтению [7].

Мутационные события, в частности однонуклеотидные замены, могут приводить к переходам между стоп-кодонами. Всего за счет однонуклеотидных замен возможно четыре типа таких переходов: TAA $\Leftrightarrow$ TGA и TAA $\Leftrightarrow$ TAG. Поскольку «смысл» кодона в результате таких замен не меняется, их можно рассматривать как специальный случай синонимических замен. Однако в связи с тем, что для трех стоп-кодонов характерна разная частота ошибочного прочтения, представляется вероятным, что мутации, приводящие к переходам из стоп-кодона в стоп-кодон, будут сильнее влиять на приспособленность по сравнению с обычными синонимическими и другими нейтральными типами мутаций. В данной работе мы с использованием данных полногеномного полиморфизма исследуем вопрос о том, находятся ли под отбором в популяциях *H. sapiens* и *D. melanogaster* мутации, приводящие к переходам между разными стоп-кодонами.

Действие отрицательного отбора на мутации определенного типа может, с одной стороны, выражаться в более низкой вероятности увидеть варианты данного типа в популяции [8]. С другой стороны, ожидается, что в случае отрицательного отбора для тех вариантов, которые присутствуют в популяции, будут характерны более низкие популяционные частоты по сравнению с нейтральным контролем [8, 9]. Наши результаты указывают на то, что мутации, приводящие к переходам между стоп-кодонами, находятся под отрицательным отбором как в популяции *H. sapiens*, так и *D. melanogaster*: в данных полногеномного полиморфизма для этих видов присутствуют оба типа подписей отрицательного отбора на исследуемый класс вариантов. Кроме того, сигнал отрицательного отбора на переходы между стоп-кодонами у *H. sapiens* был выявлен при анализе возрастов аллелей разных типов и при проведении других видов эволюционно-генетического анализа.

Поскольку в результате ошибочного прочтения стоп-кодонов могут возникать нефункциональные и токсичные для клетки белковые продукты, можно предположить, что под действием отрицательного отбора должны, в первую очередь,

находиться мутации, повышающие вероятность ошибочного прочтения (TAA $\Rightarrow$ TGA и TAA $\Rightarrow$ TAG). Действительно, в данных как для *H. sapiens*, так и *D. melanogaster* присутствует сигнал отбора на варианты, предположительно понижающие точность терминации трансляции и повышающие вероятность ошибочного прочтения стоп-кодона. Но интересно, что вопреки тому, что ожидалось бы в том случае, если бы отбор на переходы между стоп-кодонами был связан исключительно с изменением вероятности ошибочного прочтения, в части случаев сигнал отрицательного отбора присутствует и для мутаций, которые должны повышать точность терминации трансляции (TGA $\Rightarrow$ TAA и TAG $\Rightarrow$ TAA). Эти наблюдения могут указывать на то, что отбор на использование стоп-кодонов, вероятно, связан не только с вероятностью ошибочного прочтения, но и с другими факторами.

1. G. Loughran et al. (2014) Evidence of efficient stop codon readthrough in four mammalian genes, *Nucleic Acids Res.* **42**:8928–8938.
2. A.G. Cridge et al. (2018) Eukaryotic translational termination efficiency is influenced by the 3' nucleotides within the ribosomal mRNA channel, *Nucleic Acids Res.* **46**:1927–1944.
3. H. Beier, M. Grimm (2001) Misreading of termination codons in eukaryotes by natural nonsense suppressor tRNAs, *Nucleic Acids Res.* **29**:4767–4782.
4. J.M. Skuzeski et al. (1991) The signal for a leaky UAG stop codon in several plant viruses includes the two downstream codons, *J. Mol. Biol.* **218**:365–373.
5. T.W. Dreher, W.A. Miller (2006) Translational control in positive strand RNA plant viruses, *Virology* **344**:185–197.
6. C. Li, J. Zhang (2019) Stop-codon read-through arises largely from molecular errors and is generally nonadaptive, *PLOS Genet.* **15**:e1008141
7. J. Sun et al. (2005) Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes, *J. Mol. Evol.* **61**:437–444.
8. R. Savisaar, L.D. Hurst (2018) Exonic splice regulation imposes strong selection at synonymous sites, *Genome Res.* **28**:1442–1454.

9. Y.C.G. Lee, J.A. Reinhardt (2012) Widespread polymorphism in the positions of stop codons in *Drosophila melanogaster*, *Genome Biol. Evol.* 4:533–549.

**Быстрая эволюция митохондриального генома
трипаносоматид в лабораторных условиях**

Е.С. Герасимов¹, Д.А. Афонин², В.Ю. Юрченко³, С.Л. Циммер⁴,
А.А. Колесников¹

¹Биологический факультет Московского государственного
университета им. М.В. Ломоносова, Москва, Россия, Ленинские
горы 1с12 119991, jalgard@gmail.com

²Fox Chase Cancer Center, Temple University Hospital, Филадельфия,
США, PA 19111-2497

³Life Science Research Centre, Faculty of Science, University of
Ostrava, Острава 710 00, Чехия, Vyacheslav.Yurchenko@osu.cz

⁴Department of Biomedical Sciences, University of Minnesota Medical
School, Дулут, США, MN 55812, szimmer3@d.umn.edu

Семейство Trypanosomatidae включает в себя жгутиковых простейших, которые являются облигатными паразитами. Для абсолютного большинства трипаносоматид хозяином является насекомое, при этом представители данного семейства – наиболее часто выявляемые в насекомых паразиты. Среди трипаносоматид встречаются диксенные виды, проходящие более сложный жизненный цикл, включающий смену хозяина. В таком случае в качестве второго (называемого в данном случае основным) хозяина обычно выступает теплокровное животное, в более редких случаях хозяевами могут быть представители совершенно разных таксонов, например растения для рода *Phytomonas*. Представители наиболее изученных родов *Leishmania* и *Trypanosoma* вызывают тяжелые

заболевания человека и домашних животных, что в свое время послужило стимулом к изучению простейших данной группы, однако и на сегодняшний день многие аспекты их биологии остаются малоизученными.

Дело в том, что трипаносоматиды – представители рано ответившейся на общем древе эукариот группы, которые сохранили или приобрели ряд молекулярных механизмов. Одной из уникальных особенностей группы является организация единственной в клетке крупной митохондрии [1, 2]. Митохондриальная ДНК в ней собрана в дискообразную структуру, называемую кинетопластом (определившим название отряда Kinetoplastida, к которому относится семейство), который представляет собой дискообразную структуру из компактно уложенных переплетающихся кольцевых молекул разных типов [3, 4]. Кольцевые молекулы разделяются на два типа по структурно-функциональной организации: максикольца и миникольца. В типичной клетке около 50 копий максикольца размером 20 - 50 т.п.н. На данных молекулах располагаются гены митохондриального генома. Часть генов максикольцевого генома для образования функциональной мРНК нуждается в дополнительном процессинге – уридиловом редактировании РНК [5]. Это уникальный для группы процесс, при котором с помощью белкового комплекса «эдитосомы» в пре-мРНК вставляются или удаляются уридиловые основания, формируя зрелую мРНК. Позиции вставки или удаления оснований определяются путем взаимодействия редактируемого локуса пре-мРНК с короткими молекулами гидовых РНК (гРНК). Молекулы гРНК кодируются минимольцами, которых в типичной клетке около 10 тысяч копий, а их размеры сильно варьируют у разных видов. У миникольца выделяются классы по последовательности, миникольца разных классов обычно кодируют разные гРНК, так у типичной лейшмании около 125 классов миникольца размером ~700 п.н.

По современным представлениям сложная структура митохондрии трипаносоматид является результатом адаптации к паразитическому образу жизни. Система РНК редактирования за счет взаимодействия с разными гРНК способна генерировать самый сложный транскриптом [6], представленный многообразием изоформ. Она также представляет собой дополнительный уровень

регуляции экспрессии генов.

Митохондриальный геном трипаносоматид всегда был в фокусе исследований. Последние данные говорят о том, что он может быть чрезвычайно пластичен. Так, например, у лабораторного штамма *Leishmania tarentolae* репертуар миниколец оказался более бедным по числу классов, чем у выделенного из природы штамма LEM125, при этом система редактирования ряда генов также оказалась нарушенной из-за утраты соответствующих им гРНК [7].

В наших исследованиях мы детально изучили структуру генома и транскриптома митохондрии различных видов трипаносоматид, в том числе, видов и штаммов с разной историей культивирования. В результате было показано, что эволюция митохондриального генома может происходить достаточно стремительно при культивировании в лабораторных условиях. Наиболее ранние изменения выражаются в утрате классов миниколец, сопровождающейся утратой редактирования части транскриптов. Так, например, в сравнении с относительно недавно культивируемым в лаборатории видом *Vickermania spadyakhi*, близкий вид *V. ingenoplastis* имеет сильно редуцированный по набору гРНК репертуар миниколец [8]. Схожий паттерн эволюции наблюдается при сравнении двух представителей рода *Leptomonas*, также отличающихся длительностью культивирования [6]. Сравнительный анализ показал, что первый этап редукции редактирования обычно связан с генами, кодирующими субъединицы I комплекса ЭТЦ митохондрии, а ген RPS12, наоборот, никогда не подвергается редукции редактирования [9]. Другое наблюдение связано с тем, что несмотря на наблюдаемую у ряда видов делецию локусов максикольца, кодирующих гены III и/или IV комплексов ЭТЦ (например, роды *Phytomonas* или *Vickermania*), аналогичные события в ходе лабораторного культивирования не выявляются, при этом у части длительно культивируемых видов (и штаммов) наблюдается редукция редактирования транскриптов этих генов. Это говорит о чрезвычайной лабильности системы миниколец, что также подтверждается сравнительным анализом их структуры у близких видов и штаммов [6], который выявляет крайне высокий уровень вариабельности последовательностей миниколец.

Наши данные указывают на то, что большинство так называемых «модельных» трипаносоматид, культуры которых велись в лабораторных условиях длительное время, вероятнее всего не являются хорошими объектами для изучения процессов, связанных с функцией митохондрии. Также они позволяют существенно пересмотреть наши взгляды на данные, полученные с использованием «модельных» трипаносоматид и лучше планировать подобные опыты в будущем.

1. L. Simpson (1987) The mitochondrial genome of kinetoplastid protozoa: Genomic organization, transcription, replication, and evolution, *Annu. Rev. Microbiol.* **41**:363–382.
2. D. A. Maslov et al. (2019) Recent advances in trypanosomatid research: Genome organization, expression, metabolism, taxonomy and evolution, *Parasitology* **146**:1–27.
3. R. E. Jensen, P. T. Englund (2012) Network news: The replication of kinetoplast DNA, *Annu. Rev. Microbiol.* **66**:473–491.
4. P. Borst, J. H. Hoeijmakers, S. Hajduk (1981) Structure, function and evolution of kinetoplast DNA, *Parasitology* **82**:81–93.
5. I. Aphasizheva et al. (2020) Lexis and grammar of mitochondrial RNA processing in trypanosomes, *Trends Parasitol.* **36**:337–355.
6. E. S. Gerasimov et al. (2021) Complete minicircle genome of *Leptomonas pyrrhocoris* reveals sources of its non-canonical mitochondrial RNA editing events, *Nucleic Acids Res.* **49**:3354–3370.
7. L. Simpson et al. (2015) Comparison of the mitochondrial genomes and steady state transcriptomes of two strains of the trypanosomatid parasite, *Leishmania tarentolae*, *PLoS Negl. Trop. Dis.* **9**:e0003841.
8. E. S. Gerasimov et al. (2025) Evolutionary divergent kinetoplast genome structure and RNA editing patterns in the trypanosomatid *Vickermania*, *Proc. Natl. Acad. Sci. U.S.A.* **122**:e2426887122.
9. D. A. Afonin et al. (2024) *Blastocrithidia nonstop* mitochondrial genome and its expression are remarkably insulated from nuclear codon reassignment, *Nucleic Acids Res.* **52**:3870–3885.

**Исследование иммунного компонента опухолей разных
подтипов HER2-отрицательного рака молочной железы с
использованием методов транскриптомного анализа**

Р.А. Глушак¹, С.А. Кондратьева¹, О.А. Ракитина¹, К.С. Гречухина²,
Н.С. Карнаузов², К.С. Корчагина², И.В. Алексеенко¹, Л.Г. Жукова²,
Д.А. Дидыч¹

¹*Государственный научный центр Федеральное государственное
бюджетное учреждение науки Институт биоорганической химии
им. академиков М. М. Шемякина и Ю. А. Овчинникова РАН,
Москва, Россия, roma272731@mail.ru*

²*Московский клинический научный центр имени А. С. Логинова,
Москва, Россия*

Рак молочной железы (РМЖ) — лидирующее по смертности онкологическое заболевание у женщин, отличающееся высокой гетерогенностью по молекулярным, гистологическим характеристикам и чувствительности к терапии. Среди ранних форм гормон-положительного HER2-отрицательного люминального В РМЖ выделяется агрессивный подтип, условно названный «люминальный С», с прогрессией в течение года после неoadъювантной или адъювантной терапии. Он характеризуется высоким индексом Ki-67 и высокой гистологической степенью злокачественности, напоминая по клинической картине тройной негативный РМЖ. Люминальный С остается малоизученным на молекулярном уровне, что осложняет диагностику и подбор терапии, включая иммунотерапию.

Целью работы было исследовать иммунный компонент опухолей ТНРМЖ, люминального В и люминального С РМЖ на молекулярном уровне с использованием транскриптомики (RNA-seq).

Из образцов опухолевой ткани получали суспензии клеток, которые окрашивали панелью флуоресцентно-меченых антител против поверхностного маркера иммунных опухолевых популяций клеток (CD45) и витальным красителем DAPI. С помощью сортировки FACS были собраны популяции живых иммунных

клеток, которые лизировали в буфере для сохранения целостности РНК. Далее из лизатов клеток выделяли тотальные РНК, из которых были приготовлены библиотеки кДНК с двойной индексацией, предназначенные для секвенирования нового поколения. Далее проводили тримминг прочтений с помощью инструмента Cutadapt (версия 4.6), после чего картировали с использованием hisat2 (версия 2.2.2) и подсчитывали число каунтов с помощью featureCounts пакета Subread (версия 2.0.3). С помощью инструмента DESeq2 (версия 1.44.0) были определены гены с дифференциальной экспрессией между исследуемыми вариантами РМЖ, на основе этих генов было предложено разделение изучаемых транскриптомов на два кластера. Первый кластер похож на иммунный компонент ТНРМЖ, который является более агрессивным и для которого сейчас одобрена иммунотерапия ингибирования иммунных контрольных точек. Второй кластер подобен иммунному компоненту нормальной ткани молочной железы и люминального В (люмВ) подтипа. Функциональный анализ обогащения генов по сигнальным путям GSEA позволил выявить преобладание процессов клеточного деления для ТНРМЖ-подобной группы, в то время как для ЛюмВ-подобной было показано обогащение по процессам фагоцитоза и регуляции воспаления. Чтобы оценить представленность различных популяций иммунных клеток в исследуемых образцах была проведена деконволюция транскриптомов с помощью инструмента MuSiC (PMID: 30670690). Для ТНРМЖ-подобной группы было показано преобладание дендритных и В-клеток, в то время как для ЛюмВ-подобной - моноцитов и макрофагов.

В ходе работы были найдены молекулярные и клеточные различия между иммунными популяциями разных подтипов РМЖ. Была предложена новая кластеризация подтипов согласно особенностям их иммунного микроокружения. Полученные данные станут основой для создания тест-системы, позволяющей отличить люминальный С от двух других подтипов РМЖ. Планируется также валидация результатов методами иммуногистохимии и изучение влияния иммунных популяций на прогрессию опухолей на моделях *in vivo*.

Исследование выполнено при поддержке гранта Правительства Москвы (проект №0810-8/25).

Гены, лежащие в основе инновационности и решения проблем у птиц

А.А. Денисова¹, J.N.Audet²

¹ г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73, Факультет биоинженерии и биоинформатики, alexandraa.denisova@gmail.com

²Rockefeller University Field Research Center 495 Tyrrel Road,
Millbrook, NY 12545,
jaudet@rockefeller.edu

Если ещё полвека назад оскорбление “птичий мозг” (bird brain) неплохо отражало отношение человека к когнитивным способностям птиц, в последнее время воцарилось понимание сложности и превосходства птичьего интеллекта. Основной причиной непризнания за птицами способности к сложным когнитивным операциям послужило то, что мозг птиц долго считался примитивным из-за отсутствия черт, присущих мозгу когнитивно сложных млекопитающих. Со временем стало понятно, что птицы развивались параллельно, и несколько иначе развивая пришедшие от общего предка структуры, они также получили мозг, способный к выполнению сложных задач.

Для измерения уровня когнитивных способностей можно придумать множество метрик. Мы сосредоточились на двух из них: на инновационности, способности решать новые задачи или находить новые решения для уже известных задач в естественной среде, и на решении проблем, аналоге инновационности в контролируемых условиях.

Наличие базы данных, в которой для восьми тысяч видов птиц подсчитаны уровни инновационности, а также появление все большего количества полных геномов птиц хорошего качества, позволяет проводить разные сравнительные анализы. Начав с простого построения филогенетического дерева (Рис. 1) с соседствующим с ним уровнем инновационности мы заметили, что часто близкие к друг другу виды имеют очень далёкие уровни

инновационности. Это наблюдение позволило нам предположить, что птицы с высоким уровнем инновационности проходили конвергентную эволюцию, приобретая изменения в пересекающихся наборах генов и их регуляторов.

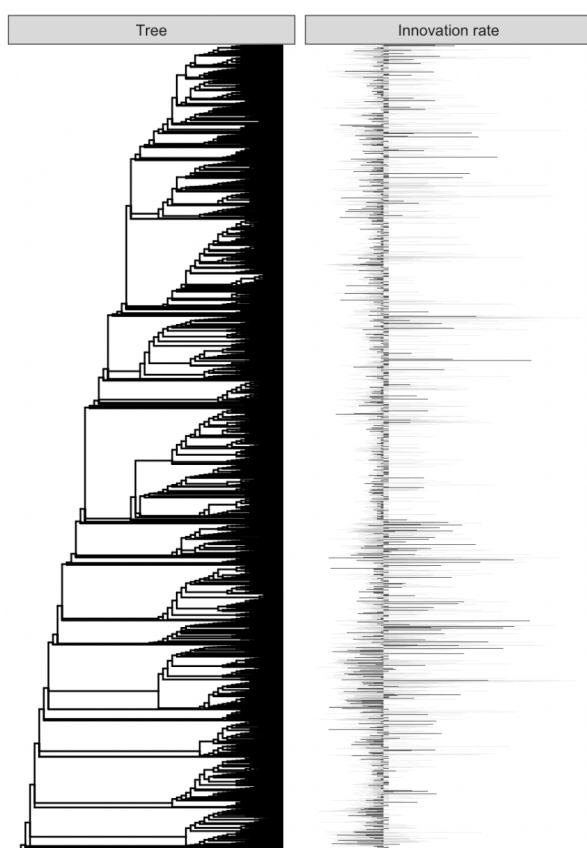


Рис 1. Распределение уровня инновационности для всех птиц, на данный момент присутствующих в базе данных инноваций.

Для начала мы посмотрели, в какие участки попадают позиции, нуклеотиды в которых отличаются для четырёх пар

близкородственных видов птиц с далёким уровнем инновационности, но при этом совпадают между эволюционно далёкими птицами с близким уровнем инновационности. Мы увидели, что эти позиции чаще, чем случайно расположены на микрохромосомах, что интересно, так как именно их предполагают основными кандидатами для размещения генов, участвующих в локальной адаптации [1]. Нам также удалось показать, что генами, в которых или рядом с которыми находятся найденные позиции, обогащены процессы ассоциированные с развитием и функционированием нервной системы. Среди наших находок были позиции, находящиеся внутри кодирующей части генов, в том числе и приводящие к миссенс-вариантам. Миссенс-варианты были найдены в гене DENND4A, кодирующем белок, участвующий в регуляции транспорта везикул и синаптической передачи в возбуждающих нейронах и олигодендроцитах. А также в FOXD2, кодирующем транскрипционный фактор, который взаимодействует с генами, контролирующими клеточную миграцию и дифференцировку в том числе в нейронах и мезодермальных производных.

Основная же часть найденных позиций находилась в некодирующих участках: в частности в интронах и апстрим регионах важных для функционирования и развития нервной системы генов. Мы задались вопросом, будет ли относительная скорость эволюции сайтов связывания транскрипционных факторов (TFBS) для найденных апстрим регионов скоррелирована с когнитивными чертами. Хотя для уровня инновационности значимых корреляций не нашлось, было показано, что относительная скорость эволюции сайта связывания транскрипционного фактора VSX2 в апстрим регионе CACUL1 положительно коррелирует с размером мозга ($r = 0.61$, $p\text{-value} = 0.035$).

Генетические основы способности к решению проблем мы исследовали на примере Белошейного воробья (*Zonotrichia albicollis*). Эти птицы примечательны тем, что имеют внутри популяции два варианта второй хромосомы: ZAL2 и ZAL2m, вторая из которых была получена в результате двойной инверсии. Примечательно, что популяция этих птиц делится на две морфы: белую (ZAL2 ZAL2m) и тёмную (ZAL2 ZAL2). Птицы отличаются

не только по окраске, но и по поведению: белые более промискуитетный, и как было показано нами, в целом хуже решают задачи.

Для исследования этих птиц мы использовали два подхода: поиск полиморфизмов между хорошо решающими задачи, но при этом белыми птицами, и плохо решающими задачу тёмными. Наше предположение было в том, что позиции, принципиально важные для формирования более высоких когнитивных способностей, наверняка будут отличаться именно у птиц, которые хорошо решают задачи, даже если они белые и наоборот. Такие позиции находились в upstream регионах CLTC и LOC113459755, в некодирующих участках генов MEGF11, PTCD3, PTRN2 и SP3, а также в downstream SP3 и VMP1.

Также мы провели GWAS, используя морфу в качестве ковариатов. Среди генов, в которые попадали позиции, имеющие $p\text{-value} < 1e-5$ были гены, которые мы уже находили ранее, в том числе FOXD2 и DENND4A, а позиции с $p\text{-value} < 5e-8$ находились рядом с генами PPP1R26 и NOP10.

Сравнение с работами предшественников обнаружило пересечения между найденными нами генами и генами, для которых ранее было показано участие в когнитивных процессах. Также нам удалось найти пересечения между нашим списком генов и генами, уровень экспрессии которых сильно отличается в областях мозга, относительный размер которых коррелирует с уровнем инновационности [2].

Таким образом, в своей работе мы провели многоуровневый анализ генов, связанных с когнитивными способностями у птиц. Впереди много работы по выявлению закономерностей и поиску регуляторных путей.

1. Campagna, L. and Toews, D.P.L. (2022) 'The genomics of adaptation in birds', *Current Biology*, 32(20). doi:10.1016/j.cub.2022.07.076.
2. Jean-Nicolas Audet et al., Divergence in problem-solving skills is associated with differential expression of glutamate receptors in wild finches. *Sci.Adv.* 4, eaao6369(2018). DOI:10.1126/sciadv.aao6369

**Молекулярная эволюция митохондриальной ДНК человека:
переосмысление в контексте новых знаний об асимметричном
мутационном спектре**

Б.Э. Ефименко¹, К.Ю. Попадьян¹, К.В. Гунбин²

¹*Балтийский федеральный университет им. И. Канта, ул.
Александра Невского, 14, Калининград, 236041, Россия
bogdan-efimenko@list.ru*

²*Институт молекулярной и клеточной биологии СО РАН, просп.
акад. Лаврентьева, 8/2, Новосибирск, 630090, Россия
genkvg@gmail.com*

Недавно стало хорошо известным, неоспоримым фактом, что митохондриальная ДНК (мтДНК) человека и животных имеет особый асимметричный спектр комплементарных мутаций ($G \rightarrow A/C \rightarrow T$, $T \rightarrow C/A \rightarrow G$) [1]. Революцию во взглядах на изменчивость мтДНК человека положили 5 направлений современных исследований: (1) детальные популяционно-генетические исследования [2, 3], которые отчетливо показали как асимметрию частот комплементарных замен, так и негомогенность их распределения вдоль мтДНК, (2) полногеномные исследования ассоциаций генотип мтДНК - фенотип (например, [4]) статистически и клинически обосновавшие функциональную значимость отдельных, асимметрично мутирующих сайтов мтДНК, (3) новейшие методы в исследовании динамики трансляции в митохондриях [5, 6], позволившие вскрыть селективное преимущество синонимических кодонов взаимодействующих с антикодонами тРНК по Уотсон-Крик правилам [6], (4) данные массового параллельного генотипирования мтДНК отдельных клеток и тканей [7, 8], позволившие привести к комплексному пониманию происхождения, динамики и функциональных последствий изменчивости мтДНК в соматических клетках человека, а также (5) детальные исследования отдельных генеалогий мтДНК человека, позволившие получить тонкую и

точную оценку скоростей фиксации мутаций [9]. Полученные на уровне популяционных данных факты об асимметрии комплементарных нуклеотидных замен в мтДНК человека были полностью независимо подтверждены в современных исследованиях на раковых данных [10] и в детализированных исследованиях повреждающих изменений ДНК в одноцепочечных и двуцепочечных состояниях [11].

Таким образом, стало очевидно, что крайне важно взвешенно подходить как к наблюдаемой схожести мтДНК, так и к различию нуклеотидных последовательностей мтДНК полученных после секвенирования, т.к. частоты возникновения мутаций разного типа, разного нуклеотидного контекста и разной локализации внутри мтДНК человека различаются в десятки и даже в тысячи раз. Однако до сих пор факт этих колоссальных различий частот разных типов/позиций мутаций в мтДНК не учитывается ни в судебной медицине, в криминалистике, ни в установлении родства, например, в археологии.

В связи с этим, целью настоящей работы являлась оценка вклада каждой мутации на каждом нуклеотидном сайте человеческой мтДНК при помощи биоинформатических методов, используя новейшие данные об особенностях вариативности мтДНК. Используя необратимые во времени марковские модели нуклеотидных замен [12], а также данные о структурной и функциональной организации мтДНК, были восстановлены вероятности предковых состояний для каждого сайта мтДНК во внутренних узлах наиболее полного и уточненного глобального филогенетического дерева мтДНК человека (включающего свыше 49 000 уникальных современных последовательностей мтДНК (GenBank), а также все доступные высококачественные древние образцы мтДНК [13, 14]).

В результате мы реконструировали эволюционно-датированный, детализированный (192-компонентный) мутационный спектр для каждого структурно-функционального района мтДНК человека в каждой кладе уточненного филогенетического дерева мтДНК.

Ключевые результаты: (i) выявлены значительные различия между спектрами фиксации мутаций в несинонимических и синонимических сайтах, особенно в случае мутаций T>C/A>G; (ii)

выявлены значительные отличия частот нуклеотидных замен в районах гена ND6 и контрольного региона по сравнению с районами остальных генов мтДНК, (iii) спектры мутаций в синонимических сайтах на внутренних ветвях дерева демонстрируют меньшую асимметрию комплементарных замен по сравнению с терминальными ветвями; (iv) обнаружена гетерогенность спектров синонимических мутаций среди разных филогенетических суб-клад U-гаплогруппы, контрастирующая с гомогенностью в гаплогруппах T, J, B, F, H, N и I; (v) выявлена неоднородность спектров мутаций в разных сайтах контрольного региона мтДНК, с выраженным их сходством между HVS2 и областью ориджина репликации H-цепи, а также между всем контрольным регионом и HVS1, и 7S-подобной мтДНК.

Работа поддержана проектом РНФ 21-75-20143.

1. A.G. Mikhailova et al. (2022) A mitochondria-specific mutational signature of aging: increased rate of A > G substitutions on the heavy strand, *Nucleic Acids Res.*, **50**: 10264-10277.
2. K.M. Laricchia et al. (2022) Mitochondrial DNA variation across 56,434 individuals in gnomAD, *Genome Res.*, **32**: 569-582.
3. A. Bolze et al. (2020) A catalog of homoplasmic and heteroplasmic mitochondrial DNA variants in humans. *bioRxiv.*, **doi**: 10.1101/798264.
4. E. Yonova-Doing et al. (2021) An atlas of mitochondrial DNA genotype–phenotype associations in the UK Biobank, *Nat Genet.*, **53**: 982-993.
5. S. H.-J. Li et al. (2021) Monitoring mammalian mitochondrial translation with MitoRiboSeq, *Nat Protoc.*, **16**: 2802-2825.
6. T. Wakigawa et al. (2024) The complexity and dynamics of in organello translation assessed by high-resolution mitochondrial ribosome profiling, *bioRxiv.*, **doi**: 10.1101/2023.07.19.549812v2.
7. C. A. Lareau et al. (2023) Single-cell multi-omics of mitochondrial DNA disorders reveals dynamics of purifying selection across human immune cells, *Nat Genet.*, **55**: 1198–1209.
8. J. An et al. (2024) Mitochondrial DNA mosaicism in normal human somatic cells, *Nat Genet.*, **56**: 1665–1677.
9. E. R. Arnadottir et al. (2024) The rate and nature of mitochondrial DNA mutations in human pedigrees, *Cell*, **187**: 3904-3918.e8.
10. Y. Yuan et al. (2020) Comprehensive molecular characterization of

mitochondrial genomes in human cancers, *Nat Genet.*, **52**: 342-352.

11. M. H. Liu et al. (2024) DNA mismatch and damage patterns revealed by single-molecule sequencing, *Nature*, **630**: 752-761.

12. M. D. Woodhams et al. (2015) A new hierarchy of phylogenetic models consistent with heterogeneous substitution rates, *Syst Biol.*, **64**: 638-650.

13. S. Mallick et al. (2024) The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes, *Sci Data*, **11**: 182.

14. E. Ehler et al. (2018) AmtDB: a database of ancient human mitochondrial genomes, *Nucleic Acids Res.*, **47**: D29–D32.

Мутационный спектр объясняет свыше 60% наблюдаемых аминокислотных замен РНК-содержащих вирусов и митохондриальной ДНК

Б.Э. Ефименко¹, К.В. Гунбин², К.Ю. Попадьин¹

¹*Балтийский федеральный университет им. И. Канта, ул. Александра Невского, 14, Калининград, 236041, Россия
bogdan-efimenko@list.ru*

²*Институт молекулярной и клеточной биологии СО РАН, просп. акад. Лаврентьева, 8/2, Новосибирск, 630090, Россия*

Введение. Ранее было показано, что аминокислотные замены в белках различных организмов происходят с выраженной асимметрией: одни аминокислоты со временем накапливаются ("победители", gainers), тогда как другие – вытесняются ("проигравшие", losers) [1]. С момента обнаружения этого феномена появилось множество интерпретаций, объясняющих направленность наблюдаемых изменений. Одной из наиболее убедительных является гипотеза о неравномерности мутагенеза, при котором смещение возникает на уровне нуклеотидов и передается на уровень аминокислотного состава [2].

В данной работе мы оцениваем вклад мутагенеза как предположительно ключевого механизма, формирующего

выявленные смещения. Для анализа были выбраны два объекта, обладающих высокой частотой изменений и хорошо изученным мутационным спектром: (1) РНК-содержащие вирусы, в том числе коронавирусы, и (2) митохондриальный геном позвоночных, чьи последовательности представлены в GenBank в наибольшем объеме среди всех белок кодирующих генов. Эти объекты позволяют напрямую проверить, каков процент аминокислотных замен определяется особенностями мутагенеза.

Построение нейтральной модели. Мы получили нейтральную модель аминокислотных замен на основе (i) мутационного спектра – двенадцати частот однонуклеотидных замен, оцененных по синонимичным мутациям, как наиболее нейтральной категории замен, – и (ii) таблицы генетического кода. Для каждого кодона, кодирующего определенную аминокислоту, были вычислены вероятности его трансформации в другие кодоны путем только однонуклеотидных замен, как в работе [3] (например, для стандартного кода учитываются всего 150 типов замен, прямые и обратные замены рассматриваются отдельно). Затем частоты кодонных замен, приводящих к одному типу аминокислотных замен, суммировали, что позволило получить матрицу нейтральных частот аминокислотных замен. На основе этой матрицы мы рассчитали равновесные частоты аминокислот, а также определили аминокислоты-"победители", увеличивающие свою долю в равновесном состоянии, и аминокислоты-"проигравшие", демонстрирующие противоположную динамику.

Нейтральная модель на митохондриальных белках. Сравнение предсказанных нейтральной моделью частот аминокислотных замен с эмпирическими данными по заменам в белках, кодируемых в мтДНК, выявило высокую согласованность, что подтверждает хорошую предсказательную способность модели. Во-первых, анализ коллекции из 3382 соматических несинонимичных мутаций в митохондриальных генах, выявленных в раковых опухолях человека (Pan-Cancer Analysis of Whole Genomes Consortium, 2658 образцов, 38 типов рака [3]), показал, что нейтральная модель объясняет 83% наблюдаемых замен. Иными словами, большинство соматических аминокислотных замен в митохондриальной ДНК опухолей согласуется с нейтральной эволюцией, т.е. определяется мутационным спектром и структурой

генетического кода. Во-вторых, проанализировав 20160 несинонимичных замен в генах мтДНК, реконструированных из митохондриального дерева человека, точность модели составила 61%. В-третьих, проанализировав 67419 несинонимичных полиморфных замен мтДНК, наблюдаемых у трех тысяч видов хордовых [5], мы получили точность модели 63%, т.е. значение, сравнимое с предыдущим набором мутаций. Наконец, с помощью ранее разработанного NeMu-пайплайна [6] мы получили набор несинонимичных мутаций на филогенетическом дереве на основе последовательностей гена *Cytb* шести тысяч хордовых животных (сравнительно-видовой уровень), и оценив качество модели, получили точность 52%. Важно отметить, что снижение процента предсказанных замен аминокислот от соматических мутаций в раках (83%) до полиморфизмов человека и других хордовых (61 и 63%) и мутаций на сравнительно-видовом уровне (52%) отражает ожидаемо возрастающую роль очищающего отбора на более длительных эволюционных масштабах. Для оценки адекватности предсказаний нейтральной модели мы применили модель, основанную на случайных спектрах, к нашим наборам мутаций и получили значимое уменьшение точности до 30-40%.

Нейтральная модель на белках РНК-содержащих вирусов. Далее мы применили нашу модель на вирусах. На наборах аминокислотных замен в нескольких кладах коронавируса SARS-CoV-2, реконструированных филогенетическим методом за период пандемии [7], мы обнаружили, что около 67-72% всех аминокислотных замен объясняются нейтральной моделью, причем анализируя только консервативные позиции генома, на которых предположительно действует сильный очищающий отбор, наблюдается значимое уменьшение точности модели до 30-40%. Для других РНК-содержащих вирусов мы также получили мутационные спектры и наборы аминокислотных замен с помощью NeMu-пайплайна [6] и, применив нейтральную модель, получили точность 55-60%. Мы предполагаем, что такая высокая точность на SARS-CoV-2 (около 70%) по сравнению с другими вирусами может быть объяснена высоким уровнем генетического дрейфа и/или крайне высокой степенью детализации филогенетических деревьев клад, которые были получены на основе сотен тысяч геномных последовательностей, в то время как для других вирусов мы

использовали не больше тысячи. Кроме того, на вирусах с положительной и отрицательной цепью мы показали значимую корреляцию между ожидаемыми на основе нейтральной модели и наблюдаемыми отличиями в аминокислотных составах.

Выводы. Таким образом, на разных наборах данных и эволюционных масштабах мы однозначно подтвердили сильное влияние мутационного спектра на частоты аминокислотных замен и аминокислотный состав. Наши находки позволяют: (1) предсказывать частоты аминокислотных замен в раковых заболеваниях человека и геномах РНК-содержащих вирусов; (2) переоценить потенциально патогенные и полезные варианты мтДНК человека; (3) более точно предсказывать направление эволюции вирусов.

Работа поддержана проектом РНФ 21-75-20143.

1. Jordan I.K. et al. (2005) A universal trend of amino acid gain and loss in protein evolution, *Nature*, Vol. 433, № 7026. P. 633–638.
2. Hurst L.D., Feil E.J., Rocha E.P. (2006) Protein evolution: causes of trends in amino-acid gain and loss, *Nature*, Vol. 442, № 7105.
3. Tang H., Wu C.-I. (2006) A new method for estimating nonsynonymous substitutions and its applications to detecting positive selection, *Mol Biol Evol.*, Vol. 23, № 2. P. 372–379.
4. Yuan Y. et al. (2020) Comprehensive molecular characterization of mitochondrial genomes in human cancers, *Nature Genetics*, Vol. 52, № 3. P. 342–352.
5. Iliushchenko D. et al. (2025) Deciphering the Foundations of Mitochondrial Mutational Spectra: Replication-Driven and Damage-Induced Signatures Across Chordate Classes, *Mol Biol Evol.*, Vol. 42, № 2.
6. Efimenko B., Popadin K., Gunbin K. (2024) NeMu: a comprehensive pipeline for accurate reconstruction of neutral mutation spectra from evolutionary data, *Nucleic Acids Res.*, Vol. 52, № W1. P. W108–W115.
7. Bloom J.D. et al. (2023) Evolution of the SARS-CoV-2 Mutational Spectrum, *Mol Biol Evol.*, Vol. 40, № 4.

Анализ метилирования геномных инсерций разных типов

О. В. Заякина¹, П.А. Гребнев¹, О.С. Шингалиева¹, В.С. Юдин¹,
С.М. Юдин¹, А.А. Кескинов¹, Е.А. Снигирь¹, Д.В. Светличный¹,
В.И. Скворцова²

¹*Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства (ФГБУ «ЦСП» ФМБА России). 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1.
info@cspfmba.ru*

²*Федеральное медико-биологическое агентство (ФМБА России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.*

В настоящей работе выполнен анализ метилирования инсерций разных типов с использованием данных Oxford Nanopore и исследовано влияние покрытия, а также характеристик самой инсерции и ее геномного контекста на достоверность обнаружения инсерции и точность оценки эпигенетических модификаций ДНК. Также мы исследовали возможности анализа в зависимости от способа получения консенсусной последовательности вставки, для сравнения использовали высокоточную сборку методом hifiasm из Hi-Fi прочтений.

Тип инсерции и ее геномного окружения определяет точность производимых с ней операций - секвенирования, картирования и сборки консенсусной последовательности. Так, для tandemных повторов и сателлитов само существование инсерции в заданной позиции недостоверно, длина и положение инсерции часто не воспроизводятся после сборки и повторного картирования. Уникальные последовательности и диспергированные повторы успешно анализируются в большинстве случаев, результаты меньше зависят от способа получения консенсуса и использованного покрытия.

Исследована корреляция типа вставки и ее ближайшего геномного окружения. Tandemные повторы и сателлиты располагаются в контексте того же типа в соответствии с механизмом их образования. Для диспергированных повторов

можно ожидать случайного распределения в геноме, однако, такое распределение было обнаружено только для повторов SINE. Мобильные элементы других типов оказались ассоциированы с контекстами того же типа, однако речь идет не о перемещении полноразмерных МГЭ, а об утрате-приобретении фрагментов уже существующей копии мобильного элемента, возможно, путем рекомбинации между разными копиями.

Важная характеристика сборки, напрямую влияющая на анализ метилирования, - ее локальная точность. Если в прочтениях фиксируется сигнал метилирования, в консенсусе ожидается последовательность CG. Это может быть не так по объективным причинам, но также может быть следствием ошибок секвенирования и сборки консенсуса.

Анализ степени метилирования инсерций показал, что МГЭ метилированы сильнее и стабильнее, чем, например, уникальные участки и простые повторы. Это согласуется с тем, что метилирование мобильных элементов, подавляющее их активность, контролируется специальным механизмом, развившимся для обеспечения стабильности генома. Минимальную степень метилирования показали участки низкой сложности и сателлиты. Слабо метилированные исключения для мобильных элементов могут объясняться разными причинами. Относительно слабо метилированные полноразмерные SINE в половине случаев расположены в области промоторов и первого интрона и, вероятно, именно положение относительно генов определяет их статус метилирования. Однако, для слабо метилированных элементов LINE такой зависимости не наблюдается.

PARADE: A computational framework for the evaluation and design of mRNA untranslated regions with enhanced tissue-specific activity

Arsenii O. Zinkevich¹, Elizaveta O. Aristova¹, Dmitry D. Penzar²
Ivan V. Kulakovskiy²

¹*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow*

State University, Russia, 119992, Moscow,
arsenii.z.work+mccmb@gmail.com

² Institute of Protein Research, Russian Academy of Sciences, Russia,
142290, Pushchino

Untranslated regions (UTRs) of mRNA significantly contribute to the post-transcriptional regulation of gene expression. Specific segments or structural motifs within UTRs can influence translation efficiency and mRNA stability [1]. For developing mRNA vaccines, highly essential are those UTRs that ensure high stability and expression levels of the target protein [2]. Unfortunately, depending on the protein, such highly active mRNA may exhibit toxicity in non-target tissues [3]. A potential solution lies in designing UTRs that enable not constitutively high but tissue-specific activity of the target protein. Therefore, the rational computational design of such UTRs represents both an intriguing theoretical challenge and a critical practical objective.

Massively parallel reporter assays (MPRAs) enable the simultaneous quantification of the activity of thousands to millions of distinct genetic sequences, thereby providing a high-throughput platform for the systematic evaluation of regulatory elements. This capability facilitates the efficient training of computational models aimed at predicting the functional characteristics of regulatory sequences, as well as the rational design of novel sequences with desired regulatory properties.

In collaboration with Hani Goodarzi's laboratory at the University of California, San Francisco, we performed a computational analysis of MPRA data. These experiments measured the regulatory activity of tens of thousands of 5' and 3' UTRs across six cell types, assessing their influence on reporter protein expression. Using this dataset, we developed **PARADE** (Prediction And RAtional DEsign of mRNA UTRs) — a computational platform for designing mRNA UTRs with high activity, stability, and cell-type-specific activity of target proteins [4]. In it, several methods have been adopted for generating mRNA UTR sequences, such as diffusion model, genetic algorithm, motif insertion, and wide-scale *in silico* screening of pseudorandom sequences.

Using PARADE, we managed to obtain millions of *de novo* generated mRNA UTR sequences, of which 15,800 were tested in five cell lines. Sequences demonstrating higher specificity and activity than those used in existing RNA therapies were identified. Finally, the tissue specificity of the designed UTRs was validated, and the therapeutic efficacy of PARADE-generated UTRs in mRNA constructs was demonstrated in mouse models.

Overall, we propose that PARADE serves as a starting point for developing safer, more precise, and stable mRNA vaccines with optimized 5' and 3' UTR sequences.

1. Floor, S. N. & Doudna, J. A. (2016) *Elife*, 2016. 5.
2. Rohner, E., et al. (2022) *Nat. Biotechnol.*, 40, 1586–1600.
3. Metkar, M., et al., (2024) *Nat. Rev. Drug Discov.*, 23, 67-83.
4. Khoroshkin M., et al., (2024) *bioRxiv*, 2024.12.31.630783

Acknowledgements. We thank Hani Goodarzi's laboratory at the University of California, San Francisco, and its members (Matvey Khoroshkin, Sean Lee, Tabea Mittmann, Hassan Yousefi) for providing experimental data that made this work possible.

Улучшение предсказания активности цис-регуляторных последовательностей при помощи алгоритмов машинного обучения

Е.А. Зубова¹, А.О. Зинкевич¹, Д.Д. Пензар²

¹*Факультет биоинженерии и биоинформатики МГУ*

²*Институт общей генетики им. Н.И.Вавилова РАН
dmitrypenzar1996@gmail.com*

Данная работа посвящена изучению регуляторных последовательностей человека. С мутациями в них ассоциировано множество клинически значимых заболеваний, таких как диабет, шизофрения и болезнь Паркинсона, в связи с чем изучение, выявление, исследование нуклеотидных вариаций в таких

последовательностях является актуальной задачей.

В работе [1] было проведено измерение активности всех потенциальных цис-регуляторных последовательностей, выявленных на основе чувствительности к ДНКазе I, в клеточных линиях HepG2, K562 и WTC11 при помощи лентивирусных векторов (lentivirus-based massively parallel reporter assays, lentiMPRA). Эксперимент заключался в интеграции потенциальных цис-регуляторных последовательностей в геном, после чего измерялась сила их влияния на экспрессию репортерного гена.

Наилучшее качество в оригинальной работе показала модель LegNet [3], обученная исключительно на нуклеотидной последовательности, ей уступила модель Enformer [2] предобученная на задаче предсказания биохимических признаков по последовательности и дообученная на предсказание регуляторной активности. В то же время, на данных малого размера, LegNet уступала предобученной модели Enformer. Для того чтобы выяснить, вносят ли дополнительную информацию биохимические признаки, выучиваемые Enformer и можно ли объединить два подхода, нивелировав их недостатки, мы модифицировали архитектуру LegNet таким образом, чтобы она принимала дополнительные признаки, предсказанные Enformer. Эти признаки предварительно пропускаются через линейный слой для понижения размерности, что позволяет выделить признаки информативные для обучения и предотвратить переобучение. Конкатенация признаков Enformer производилась к слою global average pooling модели LegNet, при этом другие подходы опять же приводили к переобучению. Финальный вариант архитектуры мы назвали Legformer.

Полученная сеть демонстрирует наилучшее качество по сравнению с LegNet и Enformer на данных lentiMPRA, в том числе на маленьком датасете WTC11 (рис. 1). Кроме того мы протестировали Legformer на наборе из трех меньших наборов данных, полученных для тех же клеточных линий. На этих данных новый подход также продемонстрировал лучшее качество, что опять же свидетельствует о том, что Legformer позволяет эффективно работать с меньшими наборами данных.

При валидации на независимых данных о *in silico* насыщающем мутагенезе из статьи Kircher et al. [4], наша модель

показала лучшее качество на 3 из 4 цис-регуляторных элементах при предсказании эффектов мутаций, по сравнению с моделями LegNet и Enformer.

Также полученная архитектура была протестирована на данных об аллель-специфическом связывании транскрипционных факторов. Для этого мы использовали базы данных ADAstra и UDACHA [5]. Качество оказалось сопоставимым с полученным ранее для модели LegNet, в частности для экспериментов DNase-seq для клеточной линии K562, число совпадающих по предпочтению вариантов оказалось 73%, а для ATAC-seq — 67%. Для сравнения, соответствующие показатели для LegNet оказались 71% и 69%.

Таким образом предложенный подход действительно позволяет работать с меньшими наборами данных, эффективно используя дополнительные знания, выученные сетями наподобие Enformer.

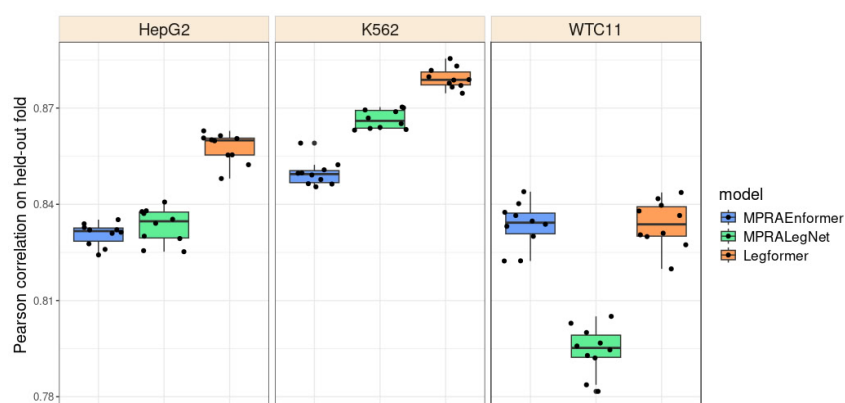


Рис. 1. Корреляция Пирсона, полученная для моделей Enformer, LegNet и Legformer (LegNet с добавлением признаков из модели Enformer) на тестовом фолде для трех клеточных линий.

1. Agarwal, V., Inoue, F., Schubach, M. et al. Massively parallel characterization of transcriptional regulatory elements. Nature (2025).
2. Avsec, Ž., Agarwal, V., Visentin, D. et al. Effective gene expression

prediction from sequence by integrating long-range interactions. Nat Methods 18, 1196–1203 (2021).

3. Dmitry Penzar, Daria Nogina et al., LegNet: a best-in-class deep learning model for short DNA regulatory regions, Bioinformatics (2023).

4. Kircher, M. et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. 10, 3583 (2019).

5. Abramov, S., Boytsov, A., Bykova, D. et al. Landscape of allele-specific transcription factor binding in the human genome. Nat Commun 12, 2751 (2021).

**Выявление генов человека, связанных с риском развития
большого депрессивного расстройства, на основе сочетанного
анализа сигнальных сетей, геномных и транскриптомных
данных**

С.М. Иванов, В.С. Сухачёв, О.А. Тарасова, В.В. Поройков

*ФГБНУ «НИИ биомедицинской химии имени В.Н. Ореховича»,
Погодинская ул., 10с8, г. Москва, 119121, Российская Федерация,
vladimir.poroikov@ibmc.msk.ru*

Большое депрессивное расстройство (БДР) — одно из наиболее распространённых заболеваний, которым страдают миллионы людей во всем мире. Использование существующих антидепрессантов во многих случаях не позволяет добиться устойчивой ремиссии, вероятно, из-за недостаточного понимания этиопатогенеза заболевания. Целью данной работы являлось определение ключевых генов человека, изменение функции которых может повышать риск развития БДР. Для этого был выполнен комплексный анализ геномных и транскриптомных данных с использованием методов анализа сигнальных сетей.

Сводная статистика по связям между однонуклеотидными полиморфизмами и БДР была получена из исследования D.M. Howard et al. [1], включавшего сотни тысяч пациентов и здоровых людей. Анализ сводной статистики с использованием инструментов

Variant	Effect	Predictor

(<https://www.ensembl.org/info/docs/tools/vep/index.html>) и MAGMA [2] позволил выявить 843 гена человека, ассоциированных с БДР, и установить, что полиморфизмы преимущественно картируются в интроны генов, а функции и экспрессия генов связаны с различными структурами мозга. Для оценки того, как полиморфизмы влияют на транскрипцию отдельных генов в тканях мозга, нами был выполнен полнотранскриптомный поиск ассоциаций (англ. transcriptome-wide association study, TWAS) [3] с использованием программы S-PrediXcan [4] и данных по локусам количественных признаков экспрессии из Genotype-Tissue Expression (GTEx) (<https://gtexportal.org/>). В результате нами были получены оценки направления изменения транскрипции 684 генов в разных структурах мозга под влиянием полиморфизмов, ассоциированных с повышенным риском развития БДР.

Анализ функций выявленных генов позволил установить, что риск развития БДР ассоциирован с изменением регуляции транскрипции генов, сплайсинга, модификации и транспорта РНК; нарушением регуляции трансляции, фолдинга, транспорта белков; процессов репарации ДНК. Выявленные гены также были связаны с изменением энергетического метаболизма клеток мозга, включая пути гликолиза, окислительное фосфорилирование в митохондриях; изменением процессов синтеза и транспорта нейромедиаторов.

Для верификации результатов TWAS нами был выполнен анализ данных по транскрипции генов в разных структурах мозга при БДР и в норме. Процессы, характерные для изменения транскрипции генов, в большинстве случаев, совпали с функциями генов, найденных нами в ходе TWAS.

При помощи алгоритмов, реализованных в пакетах R GIGSEA, VIPER и CausalR, нами были установлены транскрипционные факторы, которые предположительно вызывают изменение транскрипции выявленных 684 генов. Далее были выявлены мастер-регуляторы – белки на вершине сигнальной сети, ответственные за изменение активности транскрипционных факторов, включая толл-подобные рецепторы, рецептор гормона роста и фактор роста / дифференцировки 9 (GDF9). Эти белки потенциально играют ключевую роль в развитии БДР и могут рассматриваться как перспективные мишени для лекарственной

терапии.

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 24-25-00453).

1. D.M. Howard, M.J. Adams, T.K. Clarke, et al. (2019) Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions, *Nat. Neurosci.*, 22:343–352.
2. C.A. de Leeuw, J.M. Mooij, T. Heskes, D. Posthuma. (2015) MAGMA: generalized gene-set analysis of GWAS data, *PLoS Comput Biol.*, 11: e1004219.
3. B. Li, M.D. Ritchie. (2021) From GWAS to Gene: Transcriptome-Wide Association Studies and Other Methods to Functionally Understand GWAS Discoveries, *Front. Genet.*, 12:713230.
4. A.N. Barbeira, S.P. Dickinson, R. Bonazzola, J. Zheng, H.E. Wheeler, J.M. Torres, et al. (2018) Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics, *Nat. Commun.*, 9:1825.

Филогенетическая реконструкция трёх древних геномов *Lautropia mirabilis*.

Иванова А.О.^{1*}, Павлова А.В.¹, Боцманов Е.И.¹, Цедилина Т.Р.¹,
Перевозчикова А.А.², Березина Н.Я.², Крицкий А.А.¹, Климук
Е.И.¹, Северинов К.В.¹

¹ ООО «Биотехнологический камус», Москва, Россия

² Научно-исследовательский институт и музей антропологии им.
Д.Н. Анучина

* aivanova@biotc.ru

Мотивация и цель: *Lautropia mirabilis* — граммотрицательная, подвижная бактерия, факультативно-анаэробный кокк. Является компонентом микробиоты верхних дыхательных путей и ротовой полости.[1] Отмечено увеличение его доли в микробиоме у лиц с иммунодефицитами и метаболическими нарушениями.[2] На данный момент известно 42 полных генома

данного вида. В ходе палеогеномных исследований человеческих останков нами реконструированы три древних генома *L. mirabilis*, возрастом 6012 ± 28 , 4447 ± 22 лет и 500 лет, и проведён их функциональный и филогенетический анализ.

Методы и алгоритмы: Образцы - постоянные зубы людей различных временных эпох, из коллекции Научно-исследовательского института и музея антропологии им. Д.Н. Анучина. Образец LM_6000 - 6012 ± 28 лет ВР, мужчина погребённый в курганном могильнике Прогресс 2 (Россия, Ставропольский край); LM_4400 - 4447 ± 22 лет ВР, мужчина погребённый в курганном могильнике Расшеватский 1 (Россия, Ставропольский край); и LM_500 - 500–600 лет — подросток погребённый в грунтовом могильнике Переславля-Залесского (Россия, Ярославская область). Древняя ДНК была выделена из порошка зубного цемента в “чистых” помещениях ООО “Биотехнологический Кампус”, методом сорбции на магнитных частицах по Rohland et al., 2018. Биоинформатическая обработка включала использование программных пакетов FastP, SPAdes, MetaBAT, PhyloPhlAn, CheckM, Roary, Prokka, BLAST, CRISPRDetect, ResFinder. Аутентификация древней ДНК осуществлялась с помощью MapDamage2. Филогенетический анализ проводился с использованием IQ-TREE. Для визуализации применялись библиотеки seaborn (Python) и iTOL.

Результаты: Реконструированы и проанализированы три древних полных генома *L. mirabilis*: LM_6000, LM_4000 и LM_500. *L. mirabilis* является слабо изученным видом. Нами проведено построение филогенетического дерева известных современных геномов, с включением трех древних. Среди современных геномов выявлено 9 филогенетических клад с короткими генетическими дистанциями. Все три древних генома не входят в филогенетические клады современных штаммов *L. mirabilis*, однако два из них (LM_6000, LM_4000) образуют отдельную кладу с близкой генетической дистанцией. Из 116 известных для *L. mirabilis* факторов патогенности, у LM_4000 отсутствуют 14 генов, включая гены оперона жгутика, колонизации и пилей; у LM_500 отсутствуют отдельные гены жгутика и пилей, а у LM_6000 — гены, связанные с колонизацией и пилями.

Дополнительно была оценена степень рекомбинации и выявлен умеренный уровень горизонтального переноса генов. Анализ защитных систем у современных *L. mirabilis* показал, что 11 из 42 современных геномов этого вида содержат CRISPR-Cas системы типа I-E. Среди древних геномов, такую защитную систему удалось идентифицировать в LM_6000.

По результатам анализа антибиотикорезистентности у четырёх современных штаммов были выявлены гены устойчивости к стрептомицину, тетрациклину, хлорамфениколу и сульфаметоксазолу. В древних геномах признаки антибиотикорезистентности не обнаружены.

Выводы: Реконструированы и охарактеризованы три древних полных генома *L. mirabilis*, продемонстрированы значимые отличия от современных штаммов, включая потенциальную эволюционную утрату и/или приобретение генов, ассоциированных с патогенностью и устойчивостью к антибиотикам.

Благодарности: Исследование поддержано ООО «Биотехнологический кампус»

Список литературы:

1. **P. Gerner-Smidt et al.**, *Lautropia mirabilis* gen. nov., sp. nov., a Gram-negative motile coccus with unusual morphology isolated from the human mouth, **Microbiology**, Volume 140, Issue 7, 1994.
2. **S. N. Rossmann et al.**, Isolation of *Lautropia mirabilis* from Oral Cavities of Human Immunodeficiency Virus-Infected Children, **Journal of Clinical Microbiology**, Volume 36, Issue 6, 1998.

Мутационный анализ аминокислотных остатков протонных полуканалов бактериальной F_0F_1 -АТФсинтазы

Л.А. Ивонцин¹, Е.В. Машковцева^{1,2}, Я.Р. Нарциссов^{1,3}

¹НИИ цитохимии и молекулярной фармакологии, Москва, Россия,
ivontsin@icmph.ru

2 ФГАОУ ВО "Российский Национальный Исследовательский
Медицинский Университет им. Н.И. Пирогова" МЗ РФ, Москва

³Группа биомедицинских исследований, БиДиФарма ГмбХ, Зик,
Германия

Перенос протонов через мембранный фактор F_o АТФсинтазы является одним из важнейших этапов в процессе сопряжения энергии градиента ионов водорода и энергии образования макроэргической связи АТФ. Однако, в отличие от самой реакции синтеза, точные траектории движения протонов до сих пор остаются недостаточно исследованными. Несмотря на высокую скорость данного процесса, транспорт протонов оказывается чрезвычайно чувствительным к различным мутациям в области протонных полуканалов, приводящим к блокировке работы всего фермента.

В настоящем исследовании для того, чтобы выявить функционально важные аминокислотные остатки в описанной нами ранее последовательности полярных аминокислот и молекул воды, формирующих путь переноса протона в полуканалах бактериальной АТФсинтазы [1], был проведен мутационный анализ с использованием молекулярно-динамического (МД) моделирования. За основу была взята модельная система, которая представляла собой мембранный фактор F_o АТФсинтазы дикого типа [PDBID: 6VWK], встроенный в нативный липидный бислой и водное окружение. На основе построенной молекулярной модели мембранной части фермента дикого типа были сконструированы мутантные системы путем следующих замен во входном полуканале: *aArg210Ala*, *aArg210Lys*, *aAsn214Leu*, *aAsn214His*, *aGlu219Gly*, *aGlu219Gln*, *aHis245Gly*, *aHis245Ser*, *aHis245Tyr*, *aGln252Leu* и ревертант *aGlu219His/aHis245Glu*. В выходном полуканале шесть полярных аминокислотных остатков, которые находились в непосредственной близости от *cAsp61*, были одновременно заменены на аланин (*aGlu196*, *aSer199*, *aSer202*, *aLys203*, *aSer206*, *aTyr263*). При анализе МД траекторий особое внимание было уделено структурной динамике боковых групп мутированного остатка и близлежащих к нему аминокислот, а

также гидратации белка. Важность остатка для процесса протонного транспорта оценивалась по сохранению возможного пути переноса.

Прежде всего была проведена оценка стабильности структуры α -субъединицы всех рассмотренных мутантов. Рассчитанный энергетический ландшафт $\Delta\Delta G$ показал, что структура белка устойчива к мутациям, поскольку замены существенных остатков не приводили к сильным дестабилизирующим эффектам. При исследовании структур мутантной α -субъединицы, предсказанных с помощью AlphaFold, а также анализе результатов МД моделирования не было выявлено разрушительного воздействия аминокислотных замен на конформационную стабильность. Однако, в случае замен $\alpha\text{His245Tyr}$ и $\alpha\text{Gln252Leu}$ наблюдались конформационные изменения во взаимном расположении $\alpha\text{TMC4-6}$ (что также отражено в более высоких значениях RMSD), которые оказывали влияние на цепь переноса протона. Тем не менее, мутации могут нарушать активность белка, не меняя его стабильности, например, путем разрыва цепи переноса протонов или изменяя гидратацию полуканалов [2].

Мутационный анализ аминокислот, входящих в цепь переноса протона во входном полуканале, показал, что остатки могут функционально по-разному влиять на транспорт: участвовать в прямом переносе протонов, способствовать гидратации или обеспечивать стабильность самой структуры полуканала. Замены консервативного αArg210 приводили к формированию длинного водного канала, по которому протоны могли пересекать мембрану, не вызывая вращения ротора, что в конечном итоге привело бы к дисфункции синтеза АТФ. Остатки αHis245 и αAsn214 также являлись критическими для процесса протонного транспорта. Они принимали непосредственное участие в переносе H^+ , их замены приводили к разрыву предполагаемого пути переноса, а молекулы воды или боковые группы соседних полярных остатков не могли заменить их. В то же время, αGlu219 , αGln252 и остатки выходного полуканала непосредственно не играли решающей роли в переносе протонов, поскольку существовал альтернативный путь с участием цепочки молекул воды [2].

Таким образом, МД моделирование позволило получить

данные для прогнозирования того, как цепь переноса протона во входном полуканале бактериальной АТФсинтазы будет реагировать на такие изменения, как мутации. Рассмотренные аминокислотные замены подтвердили описанный нами ранее возможный путь переноса протона и подчеркнули значимость аминокислотных остатков [3]. Дальнейшие исследования, направленные на расшифровку молекулярного механизма переноса протонов с учетом различных влияющих факторов, позволят выявить особенности энергетического обеспечения клетки как в норме, так и при патологии.

1. L.A. Ivontsin, E.V. Mashkovtseva, Y.R. Nartsissov (2022) Insights on the proton translocation pathways in F_oF_1 -ATP synthase using molecular dynamics simulations, *Archives of Biochemistry and Biophysics*, **717**:109135.
2. L.A. Ivontsin, E.V. Mashkovtseva, Y.R. Nartsissov (2024) Molecular dynamics simulations of the mutated proton-transferring a subunit of *E. coli* F_oF_1 -ATP synthase, *International Journal of Molecular Sciences*, **25**:5143.
3. L.A. Ivontsin, E.V. Mashkovtseva, Y.R. Nartsissov (2023) Membrane lipid composition influences the hydration of proton half channels in F_oF_1 -ATP synthase, *Life*, **13**:1816.

Эволюция и функции нуклеоплазмينا в поддержании структуры ядрышка

И.С. Ильницкий^{1,2}, А.А. Жарикова^{1,3}, Д.А. Зуев¹, В.В. Вяльцев¹,
Е.В. Шеваль⁴, А.А. Миронов^{1,2}

¹Факультет биоинженерии и биоинформатики, Московский государственный университет имени М.В. Ломоносова, Ленинские горы, Москва, Россия
mironov@bioinf.fbb.msu.ru

²Институт общей генетики имени Н.И. Вавилова, Российская академия наук, Москва, Россия

³ФГБУ "Национальный медицинский исследовательский центр
терапии и профилактической медицины" Минздрава России.
Москва

⁴Институт физико-химической биологии имени А.Н. Белозерского,
Московский государственный университет имени М.В.
Ломоносова, Ленинские горы, Москва, Россия

Нуклеоплазмин (NPM) — многофункциональный белок, критически важный для организации ядрышка и биогенеза рибосом, преимущественно локализованный в гранулярном компоненте (ГК) ядрышка. NPM способствует образованию биомолекулярных конденсатов посредством жидкостно-жидкостного фазового разделения (ФР), концентрируя молекулы, необходимые для сборки рибосом [1]. В данном исследовании изучалась эволюционная динамика доменов NPM у эукариот, с акцентом на роль его внутренне неупорядоченной области (intrinsically disordered regions, IDRs) в ФР и сборке ядрышка, а также исследовались компенсаторные механизмы у видов, лишенных NPM. NPM регулирует ФР через два механизма: гетеротипные взаимодействия с рРНК и белками, содержащими аргинин-богатые R-мотивы (например, SURF6), и гомотипные взаимодействия, опосредованные A3- и B2-трактами в его IDR [2]. Полиамфолитная структура IDR с чередующимися кислыми (А-тракты) и основными (В-тракты) областями обеспечивает мультивалентные взаимодействия, критически важные для ФР. У человека NPM1 содержит три кислых тракта: A1 (остатки 38–53), A2 (120–132) и A3 (161–188), которые взаимодействуют с R-мотивами в ядрышковых белках, таких как нуклеолин и фибрилларин, и основной B2-тракт (примыкающий к С-концевому домену, CTD), который связывается с рРНК [3]. N-концевой домен (NTD) формирует пентамерную β -сэндвич структуру, способствующую олигомеризации и связыванию с гистонами, тогда как CTD с его α -спиралями взаимодействует с рРНК и G-квадруплексной ДНК, стабилизируя ассоциацию NPM с ядрышковым хроматином [4].

Мы проанализировали последовательности NPM у 3000 видов эукариот с использованием мультимодального подхода, комбинирующего методы на основе последовательностей (BLASTp, tBLASTn, HMMER), структурные методы (AlphaFoldDB, FoldSeek) и транскриптомные методы (StringTie). Было установлено, что NTD высоко консервативен, тогда как CTD, присутствующий у общего предка хоанофлагеллят и многоклеточных животных (например, *Salpingoeca rosetta*), был утрачен в нескольких линиях, включая *Spiralia* (*Ciona intestinalis*) и *Ecdysozoa* (*Drosophila melanogaster*, *Caenorhabditis elegans*). Примечательно, что нематоды (*C. elegans*) и некоторые другие группы (например, тихоходки, головоногие) полностью лишены NPM, что подтверждено отсутствием находок с помощью всех методов. Филогенетический анализ NTD показал, что CTD возник до появления многоклеточности, а последующие потери, вероятно, были обусловлены генетическим дрейфом или селективным давлением, связанным с изменением функций ядрышка.

Роль IDR в ФР была охарактеризована с использованием MobiDB и CIDER, что позволило проаннотировать полиамфолитную структуру с чередующимися заряженными остатками. У позвоночных области IDR длиннее и сложнее, тогда как у беспозвоночных, таких как *Trichoplax adhaerens*, они короче, но сохраняют чередование зарядов, поддерживая ФР. А-тракты способствуют взаимодействиям с R-мотивами, выступая в роли сигналов локализации в ядрышке (NoLS), тогда как B2-тракт усиливает связывание рРНК, концентрируя ее для сборки рибосом. Посттрансляционные модификации, такие как фосфорилирование остатков B2-тракта во время митоза, изменяют заряд и конформацию IDR, снижая ФР и способствуя разборке ядрышка [3].

Потеря CTD в ряде линий животных поднимает вопрос о функциональной значимости этого домена. Для выявления роли CTD мы провели эксперименты с нокдауном и восстановлением экспрессии NPM1 человека (knockdown-rescue experiment), используя shРНК для подавления эндогенного NPM1 с последующей экспрессией дикого типа или мутантного NPM1 с делецией CTD (NPM1N240). Уровень экспрессии белков контролировали с помощью вестерн-блоттинга, структура ядрышек

изучалась с использованием трансмиссионной электронной микроскопии. Нокдаун NPM1 хотя и не приводил к разрушению ядрышек, но существенно диспергировал ГК. Экспрессия полноразмерного NPM1 восстанавливала структуру ГК, однако экспрессия NPM1N240 такого эффекта не имела. Таким образом, CTD необходим для поддержания целостности ГК, по крайней мере, в клетках человека. Это ставит вопрос о существовании механизмов, способных компенсировать эффекты потери CTD или белка целиком в отдельных группах многоклеточных животных. Уровни экспрессии ортологов NPM и других ядрышковых белков (нуклеолин, фибрилларин, TCOF1) были проанализированы с использованием данных RNA-seq по Metazoa. У видов, лишенных NPM (например, *C. elegans*), наблюдалась повышенная экспрессия нуклеолина и фибрилларина, что указывает на компенсаторные механизмы. Анализ главных компонент (PCA) реплик RNA-seq обеспечил согласованность транскриптома, а проценти́ли экспрессии (TPM) были визуализированы в виде тепловой карты, подчеркивая отсутствие NPM в определенных линиях. Наши результаты переопределяют эволюционную историю NPM, показывая, что CTD является древней особенностью, утраченной в некоторых линиях, тогда как полиамфолитная IDR консервативна и критически важна для ФР.

Гетерогенность заряда IDR, вероятно, развивалась для взаимодействия с разнообразными молекулярными партнерами в ядрышках эукариот, повышая адаптивность. Полученные данные о роли NPM в фазовом разделении ядрышка имеют значение для понимания биомолекулярных конденсатов в здоровье и болезнях.

Благодарности: Работа была поддержана Российским научным фондом 24-24-00168.

1. Feric M., et al. (2016) Coexisting liquid phases underlie nucleolar subcompartments, *Cell*, 165:1686–1697.
2. Mitrea D.M., et al. (2016) Nucleophosmin integrates within the nucleolus via multi-modal interactions with proteins displaying R-rich linear motifs and rRNA, *eLife*, 5:e13571.
3. Mitrea D.M., et al. (2018) Self-interaction of NPM1 modulates multiple mechanisms of liquid–liquid phase separation, *Nature Communications*, 9:842.

4. Gallo A., et al. (2012) Structure of nucleophosmin DNA-binding domain and analysis of its complex with a G-quadruplex sequence from the c-MYC promoter, *The Journal of Biological Chemistry*, 287:26539–26548.
5. Banani S.F., et al. (2017) Biomolecular condensates: organizers of cellular biochemistry, *Nature Reviews Molecular Cell Biology*, 18:285–298.

Реконструкция полноразмерных изоформ на основе длинных прочтений

А.М. Касьянова^{1,2}, А.В. Клепикова², М.Д. Логачева^{1,2}, А.А. Пенин²

¹*Сколковский институт науки и технологий, Москва, Россия*

²*Институт общей генетики им. Н.И. Вавилова РАН, Москва, Россия*

alekseypenin@gmail.com

Сплайсинг играет ключевую роль в формировании транскриптомного разнообразия, регулируя процессы клеточной дифференцировки, развития, адаптации и иммунного ответа у животных и растений. Многообразие транскриптов, возникающее вследствие альтернативного сплайсинга и использования различных сайтов начала транскрипции и полиаденилирования, значительно усложняет задачу точной реконструкции полноразмерных изоформ. При этом существующие инструменты имеют ограниченные возможности для точной реконструкции изоформ.

Для решения этой проблемы мы разработали программу FLIC (Full-Length Isoform Constructor), которая объединяет преимущества длинных прочтений для обеспечения точного определения границ изоформ и коротких — для корректного определения сайтов сплайсинга (<https://github.com/albidgy/FLIC>). FLIC имеет ряд ключевых особенностей: (1) используется согласованность между биологическими повторностями для устранения технических артефактов и искажений данных; (2) сайты начала транскрипции и полиаденилирования определяются как

протяжённые области, а не точечные позиции в геноме; (3) обеспечивается надёжная реконструкция изоформ даже при отсутствии или низком качестве референсных аннотаций.

Мы оценили эффективность FLIC на реальных и смоделированных данных секвенирования кДНК *Arabidopsis thaliana*. Полученные результаты свидетельствуют о том, что FLIC эффективно реконструирует как известные, так и новые изоформы, превосходя другие инструменты, особенно в условиях отсутствия эталонных аннотаций. Сравнение с CAGE показало, что FLIC обеспечивает сопоставимую точность определения сайтов начала транскрипции, при этом значительно превосходя CAGE по скорости обработки данных.

Таким образом, FLIC представляет собой мощный инструмент для анализа изоформ у широкого круга видов, особенно в исследованиях немодельных организмов или при работе с неточными аннотациями. Кроме того, у разработанного инструмента есть потенциал для создания аннотаций *de novo* для немодельных видов.

Исследование выполнено при поддержке Института общей генетики им. Н.И. Вавилова РАН, проект № FFRW-2024-0003.

Стохастическое происхождение эффекта Мпембы в чистой воде: экспериментальное и теоретическое решение парадокса

Андрей А. Климов, Алексей В. Финкельштейн
*Институт белка РАН, 142290 Пущино, Московская область,
Россия, afinkel@vega.protres.ru*

«Эффект Мпембы» — название утверждения, что горячая вода замерзает быстрее холодной¹ или, в современной и более общей форме, что система, изначально более далекая от своего равновесного состояния, приходит к этому состоянию раньше². Это контр-интуитивное утверждение по видимости нарушает фундаментальные термодинамические и кинетические законы; но многочисленные эксперименты с классическими и квантовыми системами³⁻⁵ продемонстрировали этот парадоксальный эффект Мпембы, что привело к его обширному обсуждению в ведущих

научных журналах^{2,5,6}. Однако фундаментальные физические механизмы, лежащие в основе этого эффекта, всё ещё остаются неуловимыми⁶. В этой работе мы показываем, что эффект Мпембы в воде возникает из-за стохастической природы зародышеобразования льда. Мы повторили эксперименты по замораживанию воды в тщательно контролируемых условиях и обнаружили, что эффект Мпембы возникает только тогда, когда температура морозильника была близка к температуре нуклеации льда. В этом случае диапазон времени замерзания как горячей, так и холодной воды настолько велик, что он превышает задержку, связанную с более медленным охлаждением изначально более горячей жидкости, которая иногда замерзает раньше изначально холодной. Наш теоретический анализ показывает, что «случайность» касающегося замерзания воды парадокса Мпембы коренится в стохастической природе нуклеации льда, типичной для фазовых переходов первого рода, хотя в ряде других систем эффект Мпембы происходит совершенно закономерно⁷. Мы ожидаем, что наш анализ станет отправной точкой для пересмотра известного парадокса Мпембы в воде и других системах, претерпевающих подобные фазовые переходы.

Мы благодарны Б.С. Мельнику за помощь в экспериментальной и теоретической работе, В.В. Филимонову и С.А. Гарбузинскому за дискуссии, и РНФ (грант № 25-14-00260) за поддержку.

1. E.B. Mpemba, D.G. Osborne (1969) Cool? Phys. Educ. 4:172–175.
2. L.K Joshi et al. (2024) Observing the quantum Mpemba effect in quantum simulations. Phys. Rev. Lett. 133:010402.
3. D. Auerbach. Supercooling and the Mpemba effect: when hot water freezes quicker than cold. (1995) Am. J. Phys. 63:882–885.
4. A. Kumar, J. Bechhoefer (2020) Exponentially faster cooling in a colloidal system. Nature 584:64-68.
5. J. Zhang et al. (2025) Observation of quantum strong Mpemba effect. Nature Communications 6:301.
6. A. Santos (2024) Mpemba meets Newton: Exploring the Mpemba and Kovacs effects in the time-delayed cooling law. Phys. Rev. E 109:044149.
7. A.V. Finkelstein (2025) Analogue of the Mpemba effect in classical

Analysis of Oral Microbiome Diversity and the Lack of Distinct Orotypes in Metagenomic Profiles

Aleksandra Kozlova^{1,2}, Aleksandra Denisova^{1,3}, Michil Trofimov¹,
Alexander Rakitko^{1,2}

¹ *Genotek Ltd., Moscow, Russia*

² *Higher school of economics, Analysis in biology and medicine*

³ *Lomonosov Moscow State University, Moscow, Russia*
Moscow, Russia, amandwho@gmail.com

The oral saliva microbiome is one of the most taxonomically and functionally diverse microbial ecosystems in the human body. It is characterised with 700 species representation in metagenomes¹. In the last 5 years, >7000 articles have been published devoted to research on the salivary microbiome. In some of these studies it is mentioned orotypes^{2,3}, which are recognised as stable clusters of microorganisms formed based on their functional similarity analogous to enterotypes in the gut⁴. But there was no complex large-scale research about the existence of stable clusters in salivary metagenomes.

In our study, we analyzed the genetic data of salivary samples from 2671 individuals with 55% women and with mean age 36 years ($\sigma=10.3$) from the Genotek database, the Russian consumer genetics and research company (www.genotek.ru). Genotek clients included in our analyses provided informed consent for their data to be used for research purposes and took an online questionnaire.

Sequencing was performed on the Illumina NovaSeq 6000 platform following manufacturer's protocols. Sequenced samples were pre-processed using a Genotek pipeline. Shortly, metagenomic reads of low quality (quality score <Q20), fragmented short reads (<30 bp), and reads with more than 2 ambiguous nucleotides were removed using Cutadapt⁵ V4.4. Contaminant and host DNA was identified using BWA⁶ version 0.7.18 with the hg19 human genome release. Metagenomic

samples with less than 1.5 million unmapped reads after the preprocessing were discarded.

Species-level genome (SGB) metagenomic profiling was performed using MetaPhlAn4⁷ with the Jan21 markers database using default parameters. Alpha and beta diversity metrics were calculated on the SGB-level taxonomic profiles using the Vegan⁸ package in R. Beta diversity was calculated from the Euclidean distances based on centered log ratio transformation (equal Aitchison distances⁹).

Dimensional reduction and clustering methods were repeated based on the code from the article Bulygin I, et al¹⁰. It applied various non-linear dimensionality reduction methods, including t-SNE and UMAP, to project high-dimensional microbiome data into lower-dimensional spaces. To compare the dimensionality reduction algorithms, points of original space were reconstructed obtained from low-dimensional manifolds and count the median of the absolute error (MAE). For the embeddings with lowest MAE employed clustering algorithms to detect potential groupings within the data, after different dimensionality reduction methods.

The datasets were analyzed at the Order, Family, and Genus taxonomic levels. Organisms found in less than 1% of the samples or with a standard deviation less than 0.001 each taxon level and samples with one extremely dominant genera >70% abundance were removed from the analysis.

We obtained significant dimensionality reduction with minuscule information loss by using projection on relatively many (12 through 23, dependent on the taxonomy level) principal components with MAE from 0.105 to 0.11. Dimensionality dPCA of the PCA projections was defined as the number of first principal components that meet the selected threshold of 99% of the cumulative explained variance. It seems, at the Genus level *Prevotella*, *Neisseria* and *Vellioneella* contribute the most to the variance and the resulting PCA components. At the Family and Order level the parent eponymous taxonomy groups were found.

For manifold learning, counting of the intrinsic dimension of the project was made with PCA results via the maximum likelihood estimation (MLE). It showed that there were from 6 to 8 intrinsic dimensions in the data. Subsequent non-linear dimensionality reduction from dPCA to dMLE, is performed by different manifold learning algorithms. For each taxonomic level of different manifold learning

algorithms MAE had low variety, for order level it was 0.15, for family level MAE was equal 0.18 and for genus level it varied from 0.2461 to 0.255.

We applied several clustering methods using the Euclidean metric - Spectral Clustering¹¹, PAM and HDBSCAN¹²-for each embedding provided by the dimensionality reduction algorithms. Based on Silhouette score and Davies–Bouldin Index there were no significant stable clusters approved.

As a result, we do not observe distinct or stable clusters that would indicate the presence of orotypes. This finding remains consistent across various dimensionality reduction methods and clustering algorithms. Although the data lacks distinct and stable clusters, our analysis shows that human salivary microbial communities vary smoothly along a low-dimensional manifold.

The absence of orotypes in the oral microbiome suggests a unique community assembly process characterized by resilience, adaptability, and individual-specific functional configurations. These findings can have implications for disease diagnostics and for foundational microbial ecology, taxonomy, and future efforts to the human oral metagenome analysis.

1. Kilian, M. *et al.* The oral microbiome – an update for oral healthcare professionals. *Br. Dent. J.* **221**, 657–666 (2016).
2. Lee, S. H., Lee, H., You, H. S., Sung, H. & Hyun, S. H. Metabolic pathway prediction of core microbiome based on enterotype and orotype. *Front. Cell. Infect. Microbiol.* **13**, 1173085 (2023).
3. Goyal, A. & Maslov, S. Diversity, stability, and reproducibility in stochastically assembled microbial ecosystems. Preprint at <https://doi.org/10.1101/213199> (2017).
4. MetaHIT Consortium (additional members) *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
5. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
6. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
7. Blanco-Míguez, A. *et al.* Extending and improving metagenomic

- taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
8. Dixon, P. VEGAN, A Package of R Functions for Community Ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
 9. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **44**, 139–160 (1982).
 10. Bulygin, I. *et al.* Absence of enterotypes in the human gut microbiomes reanalyzed with non-linear dimensionality reduction methods. *PeerJ* **11**, e15838 (2023).
 11. von Luxburg, U. A Tutorial on Spectral Clustering. (2007) doi:10.48550/ARXIV.0711.0189.
 12. McInnes, L. & Healy, J. Accelerated Hierarchical Density Based Clustering. in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* 33–42 (IEEE, New Orleans, LA, 2017). doi:10.1109/ICDMW.2017.12.

**Объединение данных дипольной ЭПР-спектроскопии с
молекулярным моделированием для структурных
исследований биологических комплексов**

М. И. Колоколов^{1,2}, Н. И. Санникова^{1,2}, С. А. Дементьев^{1,2}, Р. А.
Подаров^{1,2}, К. А. Жданова³, Н. А. Брагина³, А. С. Чубаров⁴,
М. В. Федин¹, О. А. Крумкачева¹

¹ *Международный Томографический Центр СО РАН, 630090,
Россия, г. Новосибирск, ул. Институтская, д. 3а*

² *Новосибирский Государственный Университет, 630090, Россия,
г. Новосибирск, ул. Пирогова, д.1
m.kolokolov@tomo.nsc.ru*

³ *МИРЭА — Российский технологический университет, Институт
тонких химических технологий, г. Москва, пр. Вернадского, д. 86*

⁴ *Свободный Университет Берлина, Германия*

Понимание специфики связывания малых молекул с белками имеет важное значение для оптимизации фармакологических свойств новых лекарств и предсказания потенциальных побочных эффектов. Однако при наличии нескольких сайтов связывания или при нестабильном взаимодействии лиганда с белковой поверхностью традиционные методы структурной биологии часто оказываются недостаточно эффективными, из-за чего часто прибегают к компьютерному моделированию. Так, сегодня активно используется метод «слепого докинга», который позволяет определить возможные участки связывания лиганда на белке. Однако этот метод имеет недостатки: неполное покрытие конформационного пространства и упрощенные методы подсчёта энергии приводят к большой неоднозначности результатов, особенно в случаях нескольких сайтов.

В то же время метод дипольного электронного парамагнитного резонанса (ЭПР) находит все более широкое применение в структурной биологии благодаря своей способности предоставлять уникальную информацию о конформационном разнообразии систем. ЭПР позволяет измерять не только средние расстояния, но и их распределения, выявляя все одновременно существующие конформации. Тем не менее, определение точной взаимной ориентации участников комплекса лишь с помощью ЭПР затруднительно.

В данной работе представлен новый интегративный подход, который объединяет предсказательные возможности молекулярного моделирования со структурными данными ЭПР для построения экспериментально подтвержденных комплексов биомолекул. Так, предложенный алгоритм поиска сайтов связывания использует единственное экспериментальное распределение расстояний ЭПР для фильтрации и валидации результатов слепого докинга. Предсказанные положения лиганда кластеризуются, фильтруются согласно данным ЭПР и затем уточняются с помощью фокусированного докинга и молекулярной динамики.

Для демонстрации нашего подхода мы изучили связывание человеческого сывороточного альбумина (ЧСА) с фотосенсибилизаторами на основе порфирина, используемыми в

фотодинамической терапии. С помощью лазерно-индуцированных экспериментов ЭПР были получены распределения расстояний между возбужденными состояниями фотосенсибилизаторов и спиновой меткой MTSL, селективно введенной в остаток Cys34. Анализ этих распределений в сочетании с молекулярным моделированием позволил идентифицировать вероятные участки связывания для ряда фотосенсибилизаторов с разными зарядами и гидрофобностью. Было показано, что фотосенсибилизаторы могут связываться вне стандартных сайтов альбумина и зачастую в нескольких сайтах одновременно. Было продемонстрировано влияние заряда, размера и гидрофобности фотосенсибилизаторов на то, какие сайты будут заселены и их относительные заселенности.

Помимо этого, нами был разработан и применен мультицентровый подход ЭПР для построения подробной картины лабильного комплекса альбумина со связанными фотосенсибилизатором Хлорин еб (Ce6) и ионом меди Cu^{2+} . Выполнив лазерно-индуцированные эксперименты ЭПР, мы получили набор данных о распределениях расстояний между парами Ce6-MTSL, Ce6- Cu^{2+} , Cu^{2+} -MTSL и даже между одновременно связанными Ce6-Ce6.

Используя полученные распределения расстояний как ограничения для молекулярного докинга и молекулярной динамики, мы построили первую экспериментально подтвержденную структурную модель комплекса альбумин-Ce6- Cu^{2+} . Итоговая модель демонстрирует, что Хлорин еб занимает три различных сайта связывания на альбумине, причем на одном белке два из них заселены одновременно. Помимо этого, было показано, что N-конец альбумина, с которым связан ион меди Cu^{2+} , не вращается свободно, а имеет ограниченный набор конформаций с небольшой подвижностью.

Таким образом, наш алгоритм представляет собой универсальный инструмент для исследования связывания ЭПР-активных лигандов, разрешая неоднозначности как в молекулярном моделировании, так и в данных ЭПР. При этом наш подход не содержит ограничений, характерных для флуоресцентных методов, которые подвержены ошибкам интерпретации из-за отсутствия однозначных отношений донор-акцептор при связывании в нескольких сайтах. Также предложенный подход применим ко всем

ЭПР-активным лигандам, таким как фотоактивные молекулы или металлолиганды.

Программный код для построения теоретических распределений межспиновых расстояний и сравнения с экспериментом доступен по ссылке <https://github.com/KrumkachevaLab/EPR-BindingSite>.

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации (№ 075-15-2021-580).

1. Mikhail Kolokolov et al. (2025) Enhanced Binding Site Identification in Protein–Ligand Complexes with a Combined Blind Docking and Dipolar Electron Paramagnetic Resonance Approach, *Journal of the American Chemical Society*, Article ASAP, 10.1021/jacs.5c01274

Виртуальная клетка - агентная модель

Ф.А. Колпаков¹, И.Н. Киселев², С.Н. Новиков¹, И.Р. Акбердин¹,
М.А. Куляшов¹

¹Научно-технический университет «Сириус», Сириус, Россия
kolpakov.fa@talantiuspeh.ru

²ООО "Биософт.Ру", Новосибирск, Россия

Основная цель проекта - итеративное построение виртуальной клетки, которая с хорошей точностью может воспроизводить экспериментальные омиксные данные для заданного типа клеток (клеточных линий) человека:

- уровень экспрессии генов (транскриптомика);
- количество белков (протеомика);
- количество метаболитов (метаболомика).

Архитектура проекта представлена на рисунке 1. Ее основные идеи следующие:

- Агентный подход - основные клеточные процессы (транскрипция, трансляция, метаболизм и т.д.) представлены агентами.
- Разные агенты используют разные подходы для моделирования их динамики. Например, нейронная сеть используется для предсказания уровня экспрессии генов, используя информацию из базы данных GTRD и FANTOM5. Атлас метаболизма человека предоставляет геномную модель потоков метаболитов. Обыкновенные дифференциальные уравнения используются для моделирования клеточных путей передачи сигналов и энергетического метаболизма.
- Определены 3 основных пула общих молекул: РНК (RNAs), белки и их комплексы (proteins), а также другие малые молекулы и метаболиты (metabolites). Они служат входом и выходом для всех агентов. Каждый агент определяет, какие молекулы он принимает в качестве входных или выходных.
- Итерационное развитие модели позволяет постепенно уточнять ее и добавлять новые агенты.
- Существует несколько сборок (builds) виртуальной клетки, соответствующих различным типам клеток (линиям), а также различным итерациям проекта.

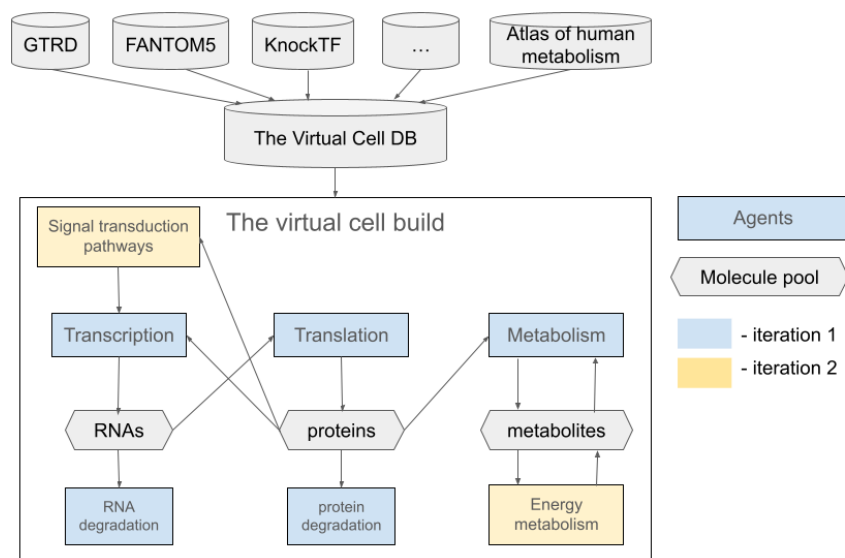


Рисунок 1. Архитектура проекта.

Таким образом, в архитектуре виртуальной клетки определены следующие базовые понятия:

- Pool - коллекция биологических объектов одного типа (RNAs, Proteins, Metabolites), для каждого объекта задается начальное количество молекул или концентрация для заданного типа клеток на основе информации из базы данных (The Virtual Cell DB, рис. 1).
- Process - агент, который обеспечивает мета-информацию о входных и выходных параметрах, связанных пулах молекул и производит расчеты - изменение концентрации или количества соответствующих молекул в пуле.

Процессы могут быть 3 типов:

1. CodeProcess – код агента может задаваться в виде кода. При настройке такого процесса пользователь может определять набор используемых параметров для его настройки. Например для настройки процесса трансляции пользователь может из базы данных выбрать набор скоростей трансляции. В версии

1.0 определены следующие процессы:

- TranslationProcess - рассчитывает изменения количества белка на основе RNA и Parameters set из базы данных.
 - ProteinDegradationProcess - рассчитывает изменения количества белка на основе текущего количества белка и Parameters set из базы данных.
 - PopulationProcess - пересчитывает количество молекул белка в одной клетке без учета деления в среднее количество молекул белка в популяции клеток.
2. ModuleProcess - обертка вокруг модуля платформы BioUML, чтобы можно было поместить в архитектуру виртуальной клетки:
- обычные дифференциальные уравнения, наиболее часто используемые для моделирования путей передачи сигнала в клетке;
 - flux balance analysis модули, используемые для моделирования метаболизма;
 - модульные модели путей передачи сигнала в клетке и метаболических путей.
3. DockerProcess - запускает Docker образ для работы агента. Взаимодействие с агентом идет через специально разработанных REST API.

Для визуального представления, создания и редактирования конкретных сборок виртуальной клетки в платформе BioUML разработан новый тип диаграммы "Виртуальная клетка" (рис. 2). Основные элементы этого типа диаграмм это описанные выше объекты типов пул и процесс и соединяющие их ребра.

На рисунке 2 представлена модель виртуальной клетки версии 1.0 для клеточных линий K562 и HepG2. В ней процессы реализованы следующим образом:

- Transcriptin - предсказание количества мРНК на основе нейронной сети, обученной по информации из баз данных GTRD и FANTOM5;
- TRanslation, Protein degradation - CodeProcess (см. выше);

- Population - осуществляет расчет концентрации белков при переходе от одной клетки к клеточной культуре в экспонентенциальной фазе роста с учетом длительности клеточного цикла;
- Metabolism - потоковая модель метаболизма.

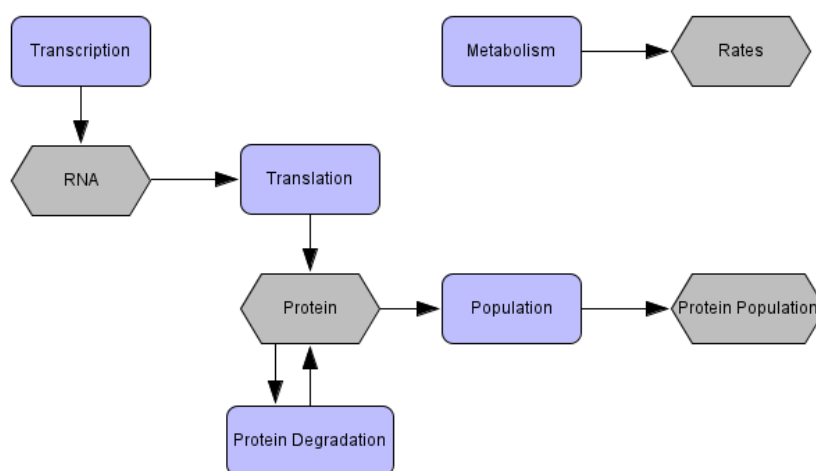


Рисунок 2. Диаграмма модели виртуальной клетки версии 1.0 в платформе BioUML.

Разработана система тестирования, позволяющая сравнить результаты численных расчетов виртуальной клетки и соответствующих экспериментальных омиксных данных, используя Jupyter ноутбуки.

Исследование выполнено за счет гранта Российского научного фонда № 24-14-20031, <https://rscf.ru/project/24-14-20031/>.

Анализ трёхмерной организации хроматина с использованием полимерных моделей: от параметров к $P(s)$ и экспериментальным данным

В.В. Константинов^{1,2}, Т.А. Лагунов^{1,2}

¹ *Институт цитологии и генетики СО РАН, 630090, Россия, г. Новосибирск, пр. ак. Лаврентьева, д. 10*

² *Новосибирский Государственный Университет, 630090, Россия, г. Новосибирск, ул. Пирогова, д. 1
v.konstantinov.work@gmail.com*

Верификация результатов полимерного моделирования в крупнозернистом приближении с экспериментальными данными является важной задачей для изучения трёхмерной организации хроматина. Полимерные модели позволяют воспроизводить пространственные конформации ДНК, однако их согласование с экспериментальными данными, такими как карты контактов Hi-C, сталкивается с рядом методологических и технических трудностей. Одной из ключевых проблем является неоднозначность параметров моделирования. Различные наборы параметров, включая потенциалы взаимодействий или плотность посадки экструдеров, могут приводить к схожим результатам, что затрудняет выбор оптимальных условий для моделирования. Более того, использование внешних ограничений на полимере (вытягивание цепи, ограничение доступного пространства) для формирования структур, наблюдаемых *in vivo*, может существенно влиять на получаемые результаты.

Экспериментальные данные также накладывают ограничения на верификацию моделей. Карты Hi-C представляют собой усреднённые данные по популяции клеток, тогда как полимерные модели часто фокусируются на отдельных конформациях. Это приводит к необходимости учитывать шум и артефакты данных, а также намеренно вводить в модель искусственные шумы при сравнении результатов с экспериментом.

Для проверки качества моделей важно использовать подходы, которые учитывают динамические изменения структуры хроматина. Например, активная экструзия хроматина экструдерами без барьерных белков может формировать временные домены, не привязанные к последовательности ДНК. Такие механизмы требуют тщательной настройки параметров модели и их проверки на экспериментальных данных.

В данной работе рассматривается влияние различных параметров моделирования и способов обработки полученных конформаций на воспроизведение экспериментальных карт контактов хроматина за счёт сравнения графиков $P(s)$ (вероятности контакта от геномного расстояния). В работе использовались данные экспериментальные данные по клеточной культуре клеток курицы DT40 с деградацией SMC-белкового комплекса конденсин II [1, 2]. Предварительные результаты показывают возможность улучшения соответствия моделей и данных Hi-C путём оптимизации параметров и учета методов обработки модельных данных. Это открывает перспективы для более точного воспроизведения трёхмерной структуры хроматина и её функционального анализа.

Работа выполнена при финансовой поддержке Новосибирского государственного университета (НГУ) и бюджетного проекта (№FWNR-2022-0019).

1. Gibcus, J.H. et al. (2018) A pathway for mitotic chromosome formation, *Science*, 359:eaao6135.
2. Samejima, K. et al. (2025) Rules of engagement for condensins and cohesins guide mitotic chromosome formation, *Science*, **388**: adq1709.

Идентификация нового потенциального ингибитора стволовой области гемагглютинина H1 вируса гриппа А

Ф.О. Константинов, В.А. Урбан, В. Г. Вересов
*Институт биофизики и клеточной инженерии НАН Беларуси,
Минск., Беларусь
fk_smmmb@ibce.by*

Введение. Вирус гриппа А является причиной ежегодных вспышек и эпидемий гриппа. Исследования выявили мутационно устойчивую стволовую область [1] гемагглютинина – белка, определяющего слияние вирусной частицы с клеткой хозяина, что позволяет рассматривать эту область в качестве перспективной

мишени для разработки универсальных вакцин и терапевтических средств.

Цель работы – идентификация методами молекулярного моделирования низкомолекулярных ингибиторов стволовой области гемагглютинаина H1 вируса гриппа А.

Задачи работы: подготовка библиотеки лигандов и структуры гемагглютинаина, виртуальный скрининг методом молекулярного докинга, ранжирование решений с использованием нейросети по значению оценочной функции свободной энергии связывания, оценка параметров ADMET, молекулярно-динамическое моделирование комплекса гемагглютинин-лиганд.

Методы исследования. Виртуальный скрининг проводили в программе Maestro [2]. Библиотека лигандов включала молекулы из PubChem и содержала более миллиона соединений. Скрининг включал три этапа молекулярного докинга повышающейся точности с отбором 10 процентов наилучших соединений на каждом этапе. Комплекс соединения JNJ4796 с гемагглютинином (PDB ID: 6CF7) [3] использовали как референсный, из него же использовали структуру гемагглютинаина. После докинга лиганды повторно ранжировали по значению оценочной функции ΔG скоринг-функций, основанных на машинном обучении: KDeep [4], AGL [5], также оценивали параметры ADMET с помощью сервиса SwissADME [6], а также осуществляли молекулярно-динамическое моделирование в программе GROMACS [7] в интервале 200 нс.

Результаты и их обсуждение. Обнаружено, что лиганд PubChem CID 137381686 обладает низким значением оценочной функции свободной энергии связывания, которое сопоставимо со значением оценочной функции свободной энергии связывания для референсного соединения JNJ4796 (KDeep Score = -9,20 ккал/моль для идентифицированной молекулы и -9.49 ккал/моль для JNJ4796, AGL Score = -8.6 ккал/моль и -9,7 ккал/моль соответственно). Лиганд имеет молекулярную массу $M_r = 399,4$, потенциально не проходит через ГЭБ, соответствует правилам подобия на лекарство Липински, Гёзе, Вебера, Эгана, Мюнге, обладает приемлемой оценкой синтетической доступности: SAScore = 2.94. Анализ RMSD при моделировании методом молекулярной динамики показал стабильность комплекса на интервале 200 нс.

Работа выполнена по заданию 1.32 подпрограммы «Молекулярные и клеточные биотехнологии-2» Государственной программы научных исследований Республики Беларусь «Биотехнологии-2», 2021-2025 гг.

1. C. Dreyfus, et al. (2012) Highly conserved protective epitopes on influenza B viruses, *Science*, 337(6100): 1343–1348.
2. , R. A. Friesner, et al. (2006) Extra precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes, *Journal of Medicinal Chemistry*, 49: 6177–6196
3. M. J. P. van Dongen, et al. (2019) A small-molecule fusion inhibitor of influenza virus is orally active in mice, *Science*, 363(6431): aar6221.
4. J. Jiménez, et al. (2018) KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks, *Journal of Chemical Information and Modeling*, 58(8): 287-296
5. D. D. Nguyen G.-W. Wei (2019) AGL-Score: Algebraic Graph Learning Score for Protein-Ligand Binding Scoring, Ranking, Docking, and Screening, *Journal of Chemical Information and Modeling*, 59, 3291-3304.
6. A.Daina, O.Michielin, V. Zoete (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules *Scientific Reports*, 7, 42717.
7. M.J. Abraham, et al. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, 1–2: 19–25.

**Скрытые технологические особенности секвенирования
нового поколения на примере полногеномного и панельного
подходов**

Н.А. Кулемин

ФГБУ ФНКЦ ФХМ им.Ю.М.Лопухина ФМБА России

Zenome.io LTD
drkulemin@gmail.com

Секвенирование нового поколения (NGS) полноценно заняло свое место среди клинических и научных методов. Технологии активно применяются как в диагностике различных заболеваний человека, так и для целей селекции растений и животных. Однако, как и любое другое технологическое решение метод секвенирования короткими прочтениями имеет систематические и случайные ошибки, которые, в дальнейшем приводят к ошибочным диагнозам (в случае клинического применения), так и к неверным статистическим выводам (в случае ассоциативной генетики).

При применении технологии в коммерческих целях организации стремятся сократить свои расходы и сосредоточиться на одной платформе и одном наборе для пробоподготовки. Такой подход практически всегда приводит к возникновению систематических ошибок, которые невозможно определить прямым контролем качества секвенирования.

На базе совместного коллектива ФГБУ ФНКЦ ФХМ им.Ю.М.Лопухина ФМБА России и стартапа Zenome.io был проведен агрегационный технический анализ данных полногеномного, полноэкзомного и панельного секвенирования человека, кукурузы и свиньи домашней короткими прочтениями с различным покрытием, предоставляемых в качестве результатов основными коммерческими лабораториями в России и странах мира. В качестве технологической платформы была использована уникальная платформа анализа и хранения NGS данных Enteromics собственной разработки, а количество различных исследованных образцов превысило 5 тысяч тестов. Проанализированные данные получены в рамках научных проектов коллектива и являются доступными для обезличенных агрегационных исследований.

В работе оценивались основные технические характеристики: глубина покрытия различных участков генома и соотношение ее со средним и среднемедианным покрытием, возможность определения генетических вариантов и среднее количество по типам (для тестов генетики человека).

В результате исследования было получено значительное количество выводов о влиянии технологического протокола на итоговый результат определения мутационного профиля.

Например, при некоторых разбиениях полногеномного тестирования на ячейки, дорожки и баркоды существенно искажается равномерность покрытия итоговых результатов после суммирования технических повторов между собой.

Равномерность также значительно влияет на результаты дальнейших ассоциативных исследований, проводимых на базе низкопокрытого секвенирования множества образцов. В случае, если производится ассоциативное исследование двух групп, каждая из которых выполнена по однотипному протоколу, но методы даже незначительно не совпадают между собой, то результаты будут отличаться по своему профилю, что свидетельствует о наличии и воспроизводимости систематических ошибок протокола, которые в последствии могут стать результатами ассоциативного сравнения и быть неверно интерпретированы.

При полноэкзомном секвенировании генома человека на итоговый мутационный профиль кроме непосредственно коммерческих наборов для обогащения также оказывает существенное влияние использование различных подходов к снижению стоимости секвенирования, а также стратегии по совместному обогащению образцов в пуле и компоновке при непосредственном нанесении на секвенирующие ячейки. Выводы, которые базируются на моделях сравнения покрытия в различных точках (например, вариации числа копий) в некоторых случаях могут быть поставлены под сомнение.

К сожалению, при проведении большинства вычислительных экспериментов авторам не удалось получить полную информацию обо всех этапах пробоподготовки, так как они, зачастую, являются коммерческой тайной лабораторий. Однако, полученные результаты позволяют иначе оценить удельный вклад систематических и случайных ошибок в больших когортных проектах и доступных базах данных, данные в которых суммируются без разделения по технологическим протоколам.

Часть данных, проанализированных в данной работе получены по гранту Российского Научного Фонда (РНФ) №22-75-

10122 «Оценка влияния эндогенных и экзогенных факторов на развитие различных типов ожирения».

Type 1 Diabetes Associated TCRs: Insights from Bulk TCR sequencing and Functional assays

Kseniia Lupyr^{1,2,3}, Anna Basova², Artemiy Dakhnovets¹, Irina Shagina^{2,3}, Kseniya Korneva^{4,5}, Alexandra Shabalkina^{2,3,7}, Daniil Luppov^{2,6,7}, Mikhail Shugay^{2,3}, Mikhail Pogorelyy⁸, Olga Britanova^{3,7}, Georgy Sharonov^{2,3,4}, Vladimir Zagainov^{4,9}, Dmitry Chudakov^{1,2,3,10}

¹ Skolkovo Institute of Science and Technology, Moscow, Russian Federation

² Pirogov Russian National Research Medical University, Institute of Translational Medicine, Moscow, Russian Federation

³ Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Department of Genomics of Adaptive Immunity, Moscow, Russian Federation

⁴ Privolzhsky Research Medical University, Nizhny Novgorod, Russia

⁵ City clinical hospital 13, Nizhny Novgorod, Russia

⁶ Institute of Personalized Oncology, I.M. Sechenov First Moscow State Medical University, Moscow, Russia

⁷ Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow, Russia

⁸ St. Jude Children's Research Hospital, Memphis, Tennessee, United States

⁹ Regional Clinical Cancer Hospital, Nizhny Novgorod, Russia

¹⁰ Abu Dhabi Stem Cells Center, United Arab Emirates

In Type 1 diabetes (T1D), the immune system destroys insulin-producing pancreatic β -cells. While insulin therapy remains the current

standard treatment, it fails to halt disease progression or prevent severe chronic complications. T1D is associated with HLA haplotypes and genetic variations linked to CD4 T cell function and thymic development, suggesting that T cell-targeted therapies could be feasible. Since T cell specificity is defined by the T-cell receptor (TCR), our study focuses on identifying T1D-associated TCRs.

We employed three experimental strategies: comparative analysis of bulk T-cell repertoires from T1D patients, their healthy siblings, and unrelated healthy cohorts; functional assays using peripheral blood mononuclear cells from T1D patients to detect TCRs specific to diabetes-associated peptides; phenotypic characterization of T cells of interest in healthy individuals.

To evaluate TCR identification methods, we framed the task as a feature selection problem and trained machine learning models using three feature sets: features selected by neighborhood enrichment approach, neural network-derived features, statistically selected features.

The statistical approach outperformed others in predictive accuracy. Notably, models trained on data from T1D patients and unrelated healthy donors achieved superior performance compared to those using data from patients and their siblings.

For functional assay analysis, we developed a statistical pipeline within the TCRgrapher R package. These experiments identified TCRs specific to T1D-associated peptides across multiple patients, distinct from bulk repertoire findings. Cells expressing these TCRs were enriched in the CD4⁺ conventional T cell subset. Conversely, single-cell data from healthy individuals revealed that analogous β TCRs correlate with regulatory T cell phenotypes.

In summary, our work identifies potential TCR targets for T1D therapy and emphasizes the critical role of genetically matched cohorts in such analyses. We demonstrate that functional assays and bulk TCR repertoire analysis yield divergent results, highlighting the need for comprehensive studies of T1D T cells that integrate receptor specificity, phenotypic differences, and functional validation. This methodological framework advances the foundation for future translational research in T1D.

Revealing the unique features and age-related changes of B cell immunity in the long-lived rodent *Spalax galili*

Aleksei A. Mazalov¹, Kseniia R. Lupyr¹, Alexey N. Davydov², Maria E. Metsger², Irina A. Shagina³, Pavel V. Shelyakin⁴, Imad Shams⁵, Dmitriy Chudakov⁶, Olga V. Britanova³

¹*Skolkovo Institute of Science and Technology, Life Sciences, Moscow, Russian Federation, Aleksei.Mazalov@skoltech.ru*

²*Central European Institute of Technology, Brno, Czech Republic*

³*Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Department of Genomics of Adaptive Immunity, Moscow, Russian Federation*

⁴*Abu Dhabi Stem Cells Center, 4600, United Arab Emirates*

⁵*Institute of Evolution & Department of Evolutionary and Environmental Biology, University of Haifa, Haifa 3498838, Israel*

⁶*Pirogov Russian National Research Medical University, Institute of Translational Medicine, Moscow, Russian Federation*

Spalax rodents (aka blind mole-rats) display exceptional longevity, hypoxia tolerance, cancer resistance and unique immunity features. Spalax B cell immunity is a broad and largely unstudied scientific field.

The aim of this work is to study spalax B cell immunity and the changes it exhibits with age. We explore both the transcriptome and the B cell receptor (BCR) repertoire features of splenic B cells at the single cell level. We compare the spalax B cell population breakdown, the expression of immunoglobulin isotype genes, somatic hypermutation (SHM) rates and repertoire diversity metrics with the results from freely available mouse single-cell sequencing data. We identify B cell populations unique for spalax and describe these populations dynamics

with aging.

Our analysis shows that, compared to mice, spalax germinal center and marginal zone B cell clusters are intriguingly enriched in the young and old age groups respectively. Pro-B cells, detected in young spalax, are thought to be rare in murine secondary lymphoid organs.

Furthermore, spalax BCR repertoire exhibits increased presence of IgG compared to mice, largely due to the memory cell subtype, which is enriched with age. Like other subpopulations growing with age, this cell subpopulation shows no signs of the senescent-associated phenotype commonly observed in aged mouse cells. The relatively high level of SHM in both spalax age groups, together with the high proportion of class-switched B cells, hint towards the existence of particularities in the expression pattern of the activation-induced cytidine deaminase (AID) enzyme, which is the subject of our ongoing research.

To the best of our knowledge, this study pioneers research on spalax B cell immunity and unveils the distinct features of B cell subpopulations in the context of aging. It contributes to the body of knowledge that, in the long run, can help to understand the mechanisms of healthy longevity and avoid premature aging.

Predicting Single-Cell Gene Expression Responses to Small Molecules: A Comparative Analysis of Deep Learning Methods

E.E. Markelova¹, V.M. Smirnov², D.V. Antonets¹, D.N. Shtokalo¹,
Y.V. Vyatkin¹, Y.A. Medvedeva³, V.E. Ramensky¹

¹*MSU Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia, markelova.ke@gmail.com*

²*Lomonosov Moscow State University, Moscow, Russia,*

³*Federal Research Centre «Fundamentals of Biotechnology» of the Russian Academy of Sciences, Moscow, Russia,*

Modern advances in multiplex single-cell RNA sequencing (scRNA-Seq) enable increasingly extensive studies of drug and genetic perturbations in cells. However, conducting experiments with large numbers of samples, tissues, and cells involves numerous challenges and high costs. Exhaustive exploration of the perturbation space is difficult to achieve experimentally. Therefore, computational methods are needed to predict, interpret, and prioritize perturbations.

We compare approaches for predicting gene expression profiles in single cells exposed to drug treatments. In this study, we focus on perturbations by small chemical molecules, as the most data is currently available for such experiments. We test several models, namely chemCPA[1], biolord[2], and a latent additive (LA) model based on multilayer perceptrons[3]. An important criterion for inclusion of a model in the comparison is its ability to utilize data about chemical compounds in the form of chemical embeddings. In such embeddings, each molecule can be represented as a high-dimensional vector describing its properties. Compared to one-hot encodings of compounds, this representation improves overall prediction quality and enables the prediction of expression profiles for molecules not present in the training set (out-of-distribution). Several approaches exist for generating chemical embeddings, and we test a subset of them in combination with the models, including Morgan molecular fingerprints and embeddings from transformer models MolFormer[4] and SelfFormer[5]. As a baseline for comparison, we use gene expression from a random sample of cells in the dataset.

We trained the models on the sciPlex 3 dataset, which includes approximately 650,000 cells from three cell lines treated with 188 different compounds[6]. Cells treated with 9 specific compounds were excluded from the training set and used to evaluate prediction quality metrics (mean R^2 coefficient, mean RMSE). We built the pipeline using Nextflow workflow system to increase the reproducibility and

portability.

The results show that chemCPA utilizing MolFormer embeddings is the most effective model among those tested, achieving a mean RMSE of 0.043 ± 0.002 when predicting gene expression profiles for unseen compounds. A key finding is that transformer-generated chemical embeddings, including both MolFormer and Selfformer, consistently yield more accurate predictions compared to traditional Morgan fingerprints. During training, models can produce the perturbation space matrix: a latent representation of compound effects with a vector for each compound. This matrix proves valuable for analyzing how compounds influence single-cell gene expression, enabling clustering by Mechanisms of Action and identification of drugs with similar biological effects. For instance, applying this approach to interleukin IL-1 blocker data confirmed its clustering with other compounds known to have anti-inflammatory and tyrosine kinase signaling inhibition properties.

These findings highlight how computational approaches can effectively predict and enable interpretation of cellular responses to chemical perturbations, offering a cost-effective alternative to extensive experimental screens. The superior performance of transformer-based embeddings suggests that advanced representation learning methods are particularly well-suited for modeling chemical perturbations in single-cell data. Future research directions could involve expanding the diversity of cell types and compounds in the training data, as well as exploring ways to incorporate additional biological context to further refine predictions. The perturbation space analysis methodology shows particular promise for drug discovery applications, where understanding compound mechanisms and identifying novel drug candidates are critical objectives.

1. Lotfollahi, M. et al. (2023). Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6), e11517. <https://doi.org/10.15252/msb.202211517>
2. Theodoris, C. V. et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), Article 7965. <https://doi.org/10.1038/s41586-023-06139-9>
3. Wu, Y. et al. (2024). PerturBench: Benchmarking Machine Learning Models for Cellular Perturbation Analysis (No. arXiv:2408.10609).

- arXiv. <https://doi.org/10.48550/arXiv.2408.10609>
4. Wu, F. et al.(2023). Molformer: Motif-Based Transformer on 3D Heterogeneous Molecular Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4), Article 4. <https://doi.org/10.1609/aaai.v37i4.25662>
5. Yüksel, A., Ulusoy, E., Ünlü, A., & Doğan, T. (2023). *SELMFormer*: Molecular Representation Learning via SELFIES Language Models (No. arXiv:2304.04662). arXiv. <http://arxiv.org/abs/2304.04662>
6. Srivatsan, S. R. et al. (2020). Massively multiplex chemical transcriptomics at single-cell resolution. *Science*. <https://doi.org/10.1126/science.aax6234>

Impaired tissue T-cell immune surveillance is linked to senescent cell accumulation during aging

Kseniia S. Matveeva, Semyon K. Kolmykov, Daniil V. Shevyrev

Sirius University of Science and Technology, 354340, Russian Federation, Krasnodar region, Sirius Federal Territory, Olympic Ave., 1., ksum.miha@gmail.com

Senescent cells are aged, dysfunctional cells that accumulate with age and contribute to the development of various age-related pathologies. Experimental studies have shown that eliminating these cells improves tissue function, restores metabolic balance, and extends lifespan¹. Under normal conditions, senescent cells are cleared by components of the innate immune system; however, the efficiency of this process declines with age². While the role of innate immunity in senescent cell clearance has been relatively well characterized, the contribution of the adaptive immune system — particularly T cells — remains poorly understood. To address this gap, we set out to investigate whether the age-related accumulation of senescent cells in tissues is associated with changes in the T-cell compartment of the immune system.

To perform our analysis, we utilized publicly available single-cell RNA-seq data from skin biopsies of 22 donors aged between 23 and 90

years, obtained from the GEO NCBI database. Several computational tools were evaluated for cell type annotation and senescent cell identification. Based on their performance, CellTypist and SenCID were selected for cell type classification and senescent cell detection, respectively. Following cell type annotation, we isolated clusters containing senescent cells and T lymphocytes for further in-depth analysis. Particular focus was placed on assessing age-related changes in the proportions of CD4⁺, CD8⁺, and Treg lymphocyte populations within the tissue, and their potential association with the abundance of senescent cells.

Our findings revealed age-associated shifts in the subpopulation composition of tissue-resident T cells, which correlated with an increased presence of senescent cells. These results suggest that age-related alterations in tissue T cell immunity may contribute to the accumulation of senescent cells over time.

This work supported by Russian Scientific Foundation, project № 24-15-20003.

1. Zhang, L., Pitcher, L. E., Prahalad, V., Niedernhofer, L. J. & Robbins, P. D. (2023) Targeting cellular senescence with senotherapeutics: senolytics and senomorphics. *FEBS J* **290**, 1362–1383.
2. Matveeva, K., Vasilieva, M., Minskaia, E., Rybtsov, S. & Shevyrev, D. (2024) T-cell immunity against senescence: potential role and perspectives. *Front. Immunol.* **15**.

Haplotype-Resolved Genomes of a Russian Male and Female: A Resource for Structural Variant Discovery.

Dmitry Meleshko, Platon Bykadorov, Dmitry Pustoshilov, Ivan Tsalkovsky, Evgeny Klimuk, Konstantin Severinov

Biotech Campus, Moscow, DMeleshko@biotc.ru

Haplotype-resolved human genome assembly remains a complex task, typically requiring the integration of multiple sequencing technologies, trio information or Hi-C data, and substantial computational resources. As part of the Platinum Genome project at

Biotech Campus, we conducted Oxford Nanopore Technologies (ONT) and MGI sequencing of a married couple along with their parents and six children. The man is of Russian descent, while the woman has Georgian ancestry. As a preliminary step toward genome assembly, we constructed MGI-based k-mer databases of the parents using meryl, and corrected the couple's ONT reads with Herro. Haplotype-resolved assemblies were generated using the latest version of Verkko, incorporating both corrected and raw ONT reads, along with trio information. The resulting draft assemblies were further polished using a combination of automated tools and manual curation to close gaps and improve accuracy. Although the absence of PacBio HiFi reads limited our ability to produce complete telomere-to-telomere (T2T) assemblies, our results effectively reach telomere-to-centromere (T2C) continuity. More specifically, for the male genome we achieved complete T2T assembly of chromosomes 2, 9, 10, 11, 17, 18, and 20, while for the female genome we assembled chromosomes 1, 9, and 11 in T2T fashion. The maternal haplotype of the female genome has an N50 of 90 Mbp, and the paternal haplotype 66 Mbp; for the male genome, the N50s are 91 Mbp and 76.5 Mbp, respectively. To the best of our knowledge, this is the first high-quality haplotype-resolved genome assembly of individuals of Russian and Georgian descent, and it can serve as a valuable reference in future population and medical genomics studies.

In our project, we aim to establish a benchmark set of “platinum” structural variants that can be used to evaluate and certify bioinformatics SV-calling pipelines. Existing benchmarks, such as those from the Genome in a Bottle (GIAB) consortium, are limited in scope, often biased toward well-characterized populations and focused primarily on simple variant types, leaving complex SVs and diverse ancestries underrepresented. One of the reasons, is that this benchmark is mostly constructed using read alignments to the reference. On the other hand, haplotype-resolved assemblies are a highly sensitive source of structural variants (SVs), often revealing more SVs than the combined output of short-read and long-read SV callers. However, the reasons behind this discrepancy have not been systematically assessed. To call SVs from haplotype-resolved assemblies, we aligned the assemblies to the GRCh38 reference genome and used the hapdiff package, which separately calls SVs for each haplotype and then merges the results. In the male sample, we identified 39,312 SVs longer than 50 bp; in the

female sample, 39,484 such events were detected. In contrast, calling SVs directly from ONT reads using Sniffles yielded only 30,259 and 29,209 events, respectively. These discrepancies cannot be attributed to minor variations between SV callers. To investigate the discrepancies, we manually examined several cases where assembly-based and read-based SV calling results did not align. In our study, we identified multiple categories of structural variants where assembly-based methods clearly outperform state-of-the-art read-based approaches. One such category includes genomic regions where long-read SV calls violate Mendelian inheritance rules. By comparing SV calls across different methods, we observed that read-based tools often fail to correctly characterize insertions present on both haplotypes when they differ in length or genomic position. For example, tools like Sniffles frequently report such events as a single homozygous insertion of arbitrary size, disregarding the heterozygous nature of the underlying haplotypes. These misclassified insertions are abundant in the sequencing data and can significantly affect downstream analyses if not properly resolved. Another common source of errors arises in repetitive genomic regions, where individual long reads often produce inconsistent SV signatures that differ in size, quantity, and breakpoint positions. This is particularly problematic inside tandem repeat regions. Read-based SV callers typically struggle to cluster these variable signals into a single coherent event. As a result, such tools may generate multiple spurious SV calls within these regions or, conversely, fail to report the event altogether in the final VCF file. While some assembly-based SV calling methods can partially resolve issues in repetitive regions, haplotype-resolved genome assembly offers a more comprehensive solution, as variant clustering is inherently performed during the assembly process. In this study, we identified several hundred events falling into the described categories and demonstrated that haplotype-resolved assemblies consistently outperform current read-based methods in these challenging regions.

Using our haplotype-resolved assemblies, we were also able to assess the presence of complex structural variants—events that cannot be classified as simple insertions, deletions, or inversions, but instead represent combinations of multiple types. We identified several complex events that can be described as inversion combined with deletion (INVDEL), inversion with duplication (INVDUP), or inversion with insertion (INVINS). The inclusion of such complex events in our

benchmark set of structural variants will help expand the scope of SV benchmarking and enable more comprehensive evaluation of SV-calling tools, particularly those capable of detecting compound rearrangements.

Our results demonstrate that haplotype-resolved genome assembly is not only feasible using current long-read technologies and trio information, but also essential for accurate and comprehensive structural variant discovery. By comparing assembly-based and read-based SV detection methods, we reveal specific classes of variants—particularly in repetitive and compound rearrangement regions—where traditional tools fall short. The resulting “platinum” SV benchmark set, enriched with complex and population-specific variants, represents a significant step toward more inclusive and accurate evaluation of SV-calling pipelines. This work lays the foundation for improved population-scale SV analysis and better characterization of genomic variation in underrepresented groups, ultimately supporting precision medicine and large-scale genomics efforts.

Применение интерпретируемых моделей машинного обучения в метаболомике для анализа возрастных изменений

Р.Д. Мухаметшина, А.Д. Мухаметшина

*119234, г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73,
Факультет биоинженерии и биоинформатики,
1709mrd@gmail.com*

С возрастом метаболический профиль человека меняется, отражая как естественные процессы старения, так и возможные метаболические нарушения. Выявление биомаркеров, связанных с возрастными изменениями, может стать основой для персонализированных подходов к замедлению старения. Методы машинного обучения (ML) всё чаще применяются в биомедицине, особенно при анализе метаболомных данных, где важно учитывать большое число взаимосвязанных признаков. В контексте старения организма это особенно актуально, поскольку отражает как физиологическое старение, так и индивидуальные биохимические изменения, связанные с возрастом. В данной работе применены

интерпретируемые модели машинного обучения (XGBoost + SHAP-анализ) к данным метаболомики плазмы из KarMeN study, с целью предсказания возрастной группы и выявления ключевых метаболитов, изменяющихся с возрастом [1]. Таким образом, целью исследования является выявление биомаркеров старения на основе данных метаболомики плазмы крови из KarMeN study с использованием интерпретируемых методов машинного обучения (XGBoost и SHAP-анализ), а также сравнительный анализ с результатами оригинального исследования. Для достижения цели были поставлены следующие задачи:

1. Построение модели классификации возрастной группы (<37 лет vs. ≥60 лет) отдельно для мужчин и женщин с использованием алгоритма XGBoost.
2. Применение SHAP-анализ для оценки вклада каждого метаболита в модели и интерпретации значимости признаков.
3. Проведение сравнения полученных ключевых метаболитов с результатами оригинальной публикации (Rist et al., 2017) для валидации биомаркеров старения.

В исследовании проанализированы 150+ метаболитов из плазмы крови участников KarMeN study [1]. После разделения данных по возрасту и полу в модели было включено ~120 образцов. Основная задача — классификация образцов на две возрастные группы (<37 и ≥60 лет) с использованием XGBoost и интерпретацией результата через SHAP.

Были рассчитаны показатели качества моделей, включая общую точность (Accuracy) и значение площади под ROC-кривой (AUC). Модель показала высокие значения.

Для интерпретации вклада признаков применялся SHAP-анализ (SHapley Additive exPlanations), позволивший не только оценить значимость метаболитов, но и визуализировать направление их влияния (увеличивает или снижает вероятность старения). Были построены SHAP summary plot и bar-графики средней важности признаков по группам («молодые» и «пожилые»).

В оригинальной работе KarMeN использовались методы SVM, GLMnet, и PLS, а SHAP и XGBoost не применялись. Интерпретация метаболитов осуществлялась на основе средних коэффициентов

моделей, без локальной интерпретации по участникам [1].

Сравнение с оригинальной статьей (Rist et al., 2017, KarMeN study)

Параметр	KarMeN study	Настоящее исследование
Модели	SVM (linear), GLMnet, PLS	XGBoost
Оценка точности	R^2 до 0.77 у мужчин (PLS), до 0.60 у женщин	Accuracy до 0.84, AUC до 0.94
Интерпретация	Среднее ранжирование метаболитов	SHAP (локальный и глобальный вклад)

Для визуализации вклада метаболитов в классификацию возрастных групп был построен bar plot среднего абсолютного значения SHAP (SHapley Additive Explanations) для каждой группы (молодые/пожилые) (рис.1).

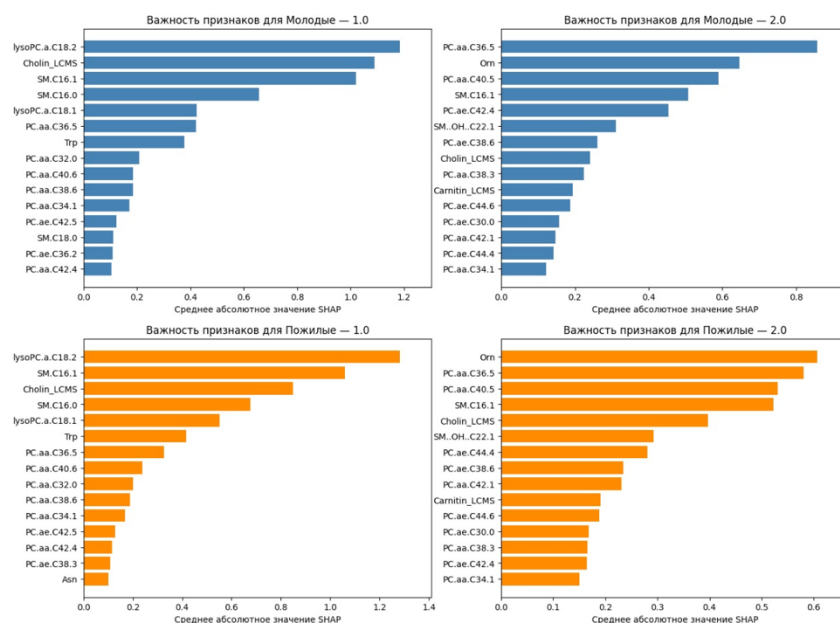
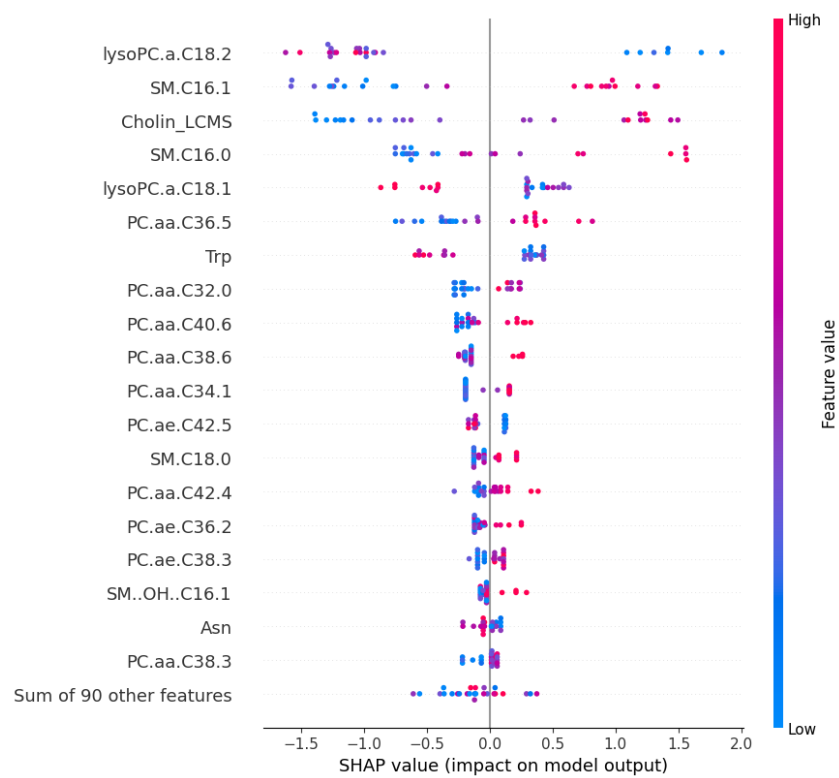


Рис.1 bar plot среднего абсолютного значения SHAP

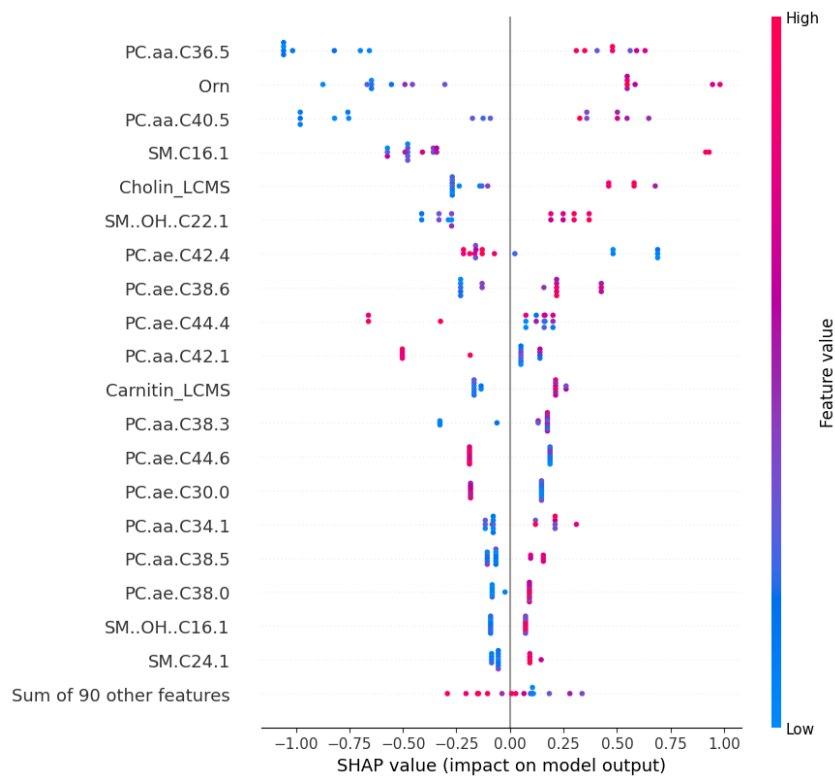
Дополнительно для интерпретации влияния метаболитов был построен SHAP summary plot (beeswarm) — визуализация, отображающая направленность и силу вклада каждого признака в предсказание модели (рис.2).



Пол: 1.0 (мужчины)

Accuracy: 0.84

ROC AUC: 0.97



Пол: 2.0 (женщины)
 Accuracy: 0.81
 ROC AUC: 0.94

Рис.2 SHAP-анализ и классификация

Результаты SHAP-анализов подтвердили важность ряда метаболитов, описанных в оригинальной работе (фенилаланин, орнитин, холин), и дополнили их новыми биомаркерами, такими как глутамин, глутамат, аминокислоты с разветвлённой цепью (BCAA) и фосфолипиды.

Совпадения и расхождения в значимых метаболитах

Метаболит	В KarMeN	В данном исследовании	Совпадает?
Фенилаланин	Да	Да	Да
Орнитин	Да	Да	Да
Аспартат	Да	Да	Да
Холин/холестерин	Да	Да	Да
Псевдоуридин	Да	Нет	Отсутствует
Седогептулоза	Да	Нет	Отсутствует
Глутамин	Нет	Да	Новый результат
Глутамат	Нет	Да	Новый результат
ВСАА (лейцин, валин)	Нет	Да	Новый результат
Фосфатидилхолины	Нет	Да	Новый результат

В настоящем исследовании был реализован альтернативный подход к предсказанию возрастной принадлежности на основе

метаболических данных плазмы — с использованием XGBoostClassifier и SHAP-анализа — к метаболическим данным плазмы из KarMeN study. Основной задачей было предсказание возрастной группы (молодые <37 лет, пожилые ≥60 лет) на основе метаболического профиля. Было проведено разделение данных по полу, построены модели для мужчин и женщин, а также выполнено сравнение полученных метаболитов с результатами оригинальной публикации KarMeN (Rist et al., 2017). Что позволило сравнить надёжность выявленных метаболитов, а также выявить потенциальные мишени для нутрицевтических и поведенческих интервенций с целью «обращения» метаболических признаков старения.

Полученные результаты позволили выявить ряд метаболитов, ассоциированных с возрастными изменениями, включая фенилаланин, холин, глутамат и аминокислоты с разветвлённой цепью (BCAA), уровни которых потенциально поддаются коррекции. Выделение таких биомаркеров формирует основу для разработки целевых вмешательств, направленных на поддержание метаболического здоровья и замедление процессов старения.

Литература

1. Rist MJ, et al. (2017) Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. PLoS ONE 12(8): e0183228.
<https://doi.org/10.1371/journal.pone.0183228>

Transcription Initiation in Pericentromeric Regions of the Human Genome

Pavel Nikitin¹, Anastasia A. Zharikova¹, Victor Spangenberg^{2,3},
Sviatoslav Sidorov^{4,5}

¹*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia,
nikitinp@fbb.msu.ru*

²*Vavilov Institute of General Genetics, Russian Academy of Sciences,
Moscow, Russia,*

³*Scientific Center of Zoology and Hydroecology, Yerevan, Armenia,*

⁴*MRC Laboratory of Medical Sciences, Hammersmith Hospital
Campus, London, United Kingdom*

⁵*Institute of Clinical Sciences, Faculty of Medicine, Imperial College
London, Hammersmith Hospital Campus, London, United Kingdom*

Centromeres play a crucial and evolutionarily conserved role in ensuring proper chromosome segregation during cell division. Despite the conserved function of centromeres, centromeric DNA sequences differ greatly across species, and the formation of centromeres is determined by epigenetic mechanisms. Although the epigenetic landscape of centromeres has been a subject of investigation for many years, only recently we gained the ability to use the entire toolkit of bioinformatics for this analysis, as the Telomere-to-Telomere (T2T) Consortium published a gapless human genome assembly, T2T-CHM13. For the first time, this assembly revealed the entire pericentromeric tandem repeat arrays.

Using Oxford Nanopore long reads, as well as short reads that map uniquely or to unique reference k-mers, the T2T Consortium detected transcription and protein-DNA binding in pericentromeric regions. On the other hand, there is a large volume of publicly available short-read data from RNA-seq, ChIP-seq, Cap Analysis of Gene Expression (CAGE) and other experiments which may contain functional signals from pericentromeric regions but cannot be uniquely mapped to tandem repeat arrays. In this study, we aim to develop a biologically meaningful method of mapping such reads to highly repetitive pericentromeric regions. This method will significantly expand our ability to detect transcriptional and protein-DNA interaction signals in pericentromeres.

The most common repetitive DNA sequence found in human

centromeres is alpha satellite (α Sat) DNA which represents ~2.8% of the genome. α Sat monomers, ~171 bp long, have different types and primarily form various tandemly repeated combinations known as higher-order repeats (HORs). Alpha satellites are transcribed, and their transcripts play a key role in the normal course of mitosis. However, some fundamental characteristics of α Sat RNAs – such as transcription start sites (TSSs), the typical length and stability – remain unexplored. Studying these properties of α Sat transcripts will provide deeper insights into the mechanisms underlying their functional roles. To investigate these features of α Sat RNAs, we mapped CAGE and native elongating transcript-CAGE (NET-CAGE) sequencing data to pericentromeric regions and analysed the signal of transcription initiation in α Sat arrays. For our analysis, we selected 27 human CAGE samples taken from keratinocyte progenitors, erythroid progenitors, hematopoietic progenitors, myeloid progenitors, dermal fibroblasts, placental villous mesenchyme, osteoblasts, myosatellite cells (primary cell types) and NET-CAGE samples (profiling nascent RNA) from GM12878, H9hESC, HeLa-S3, HepG2, K562 and MCF-7 cell lines. We mapped the selected CAGE and NET-CAGE libraries to the T2T-CHM13 reference genome with STAR and used the result for genome-wide quality control and for further selection of putative pericentromeric reads. For the latter, we selected unmapped reads and reads that mapped to pericentromeric regions. Then, we remapped the selected reads to α Sat HORs, using STAR, HISAT2 and bowtie2 with different maximum allowed numbers of alignments to assess the suitability of these tools for the mapping of highly repetitive regions and to reach alignment saturation. Unexpectedly, we found that with the increasing maximum number of alignments, STAR reclassifies as "unmapped" a substantial number of reads that were already mapped with a lower maximum number of alignments. Also, we found that HISAT2 cannot map reads on the negative strand with 5' end soft-clipping which is a key part of our mapping strategy. In contrast to STAR and HISAT2, bowtie2 did not have mapping artifacts and, we used it, allowing up to 100,000 mapping loci per read, to find all possible α Sat alignments for each read.

After filtering reads mapped to α Sat arrays by the proportion of mapped nucleotides and by the stability of the mapped 5' dinucleotide, monomer type, strand orientation and the location of the 5' end of the read within the monomer, we obtained 872 reads, out of ~684 million

reads, that show a reliable transcription initiation signal in alpha satellites. After stratifying mapped α Sat monomers by length, we found positions with the highest, clearly dominating, CAGE signal in the D1, J1 and J2 monomers. These positions had canonical, pyrimidine-purine, initiator dinucleotides, which suggested that the found positions are true TSSs. These positions were different across the α Sat types, which could be explained either by the irrelevance of the particular initiation position, as long as it is efficient, or by the preservation of the properties of transcripts produced from different monomer sequences. Additionally, we showed that the transcription of the α Sat monomers of these three types predominantly initiates from the strand on which they are annotated. Finally, based on the fact that we detected the pericentromeric transcription initiation signal mainly from the NET-CAGE, but not CAGE, data, we hypothesise that α Sat transcripts are unstable.

Next, we plan to expand our analysis by using all available NET-CAGE replicates from the same cell types. It should allow us to investigate transcription initiation in other types of alpha satellites. Next, we aim to use long-read CAGE-like data from ENCODE and recently published "true length of diverse capped RNA sequencing" (TLDR-seq) data to verify detected α Sat TSSs and estimate the typical lengths of α Sat transcripts. After confirming the α Sat TSSs, we plan to compare sequences around them with known promoter architectures, as well as nucleosome positioning and RNA polymerase II pausing signals, and predict putative transcription factor binding sites. Finally, using publicly available CAGE data from the same cell lines, we will estimate the stability of α Sat transcripts more precisely, measuring the change in the transcription initiation signal in the same positions between the NET-CAGE data (profiling nascent transcription) and CAGE data (profiling the total accumulated pull of transcripts).

Ре-инженерия лиганд связывающих белков

А.С. Николаев¹, Я.А. Орлов¹, Д.А. Лунегова², Н.Н. Случанко²,
И.Ю. Гущин¹

¹*Московский физико-технический институт,
nikolaev.as@phystech.edu,*

²*Федеральный исследовательский центр "Фундаментальные
основы биотехнологии",*

Недавние достижения в области методов машинного обучения привели к разработке множества новых подходов для инженерии и дизайна белков. Один из таких подходов, ProteinMPNN [1], предсказывает аминокислотную последовательность, которая наиболее вероятно сложится в заданную пользователем структуру белка. Этот алгоритм может быть применен для поиска новых аминокислотных последовательностей - как для уже существующих в природе белковых конформаций (реинжиниринг), так и для тех, которые были созданы искусственно (de novo белковый дизайн). Особо интригующим остаётся вопрос, можно ли использовать ProteinMPNN для получения термостабильных, растворимых и легко экспрессируемых аналогов природных лиганд-связывающих белков посредством реинжиниринга.

Для этого мы применили ProteinMPNN для реинжиниринга трех разных лиганд-связывающих белков - растворимого, периферического и трансмембранного. Во-первых, нам удалось получить три аналога флавин-связывающего флуоресцентного белка CagFbFP. Несмотря на то, что они были лишь на 55%-66% идентичны исходному белку, они были функциональны, показали высокую термостабильность и имели фотофизические свойства, аналогичные оригинальному белку. Во-вторых, нам удалось создать растворимый аналог мембранного ретиналь-связывающего белка бактериородопсина, который связывает ретиналь и имеет характерный спектр поглощения, аналогичный исходному белку. Наконец, мы обнаружили, что ProteinMPNN имеет ограниченную применимость для белков, структуры которых были определены с помощью метода ЯМР. Чтобы устранить это ограничение, мы разработали специальный протокол применения ProteinMPNN к

таким структурам и протестировали его на примере белка AstaPo1, участвующего в транспорте каротиноидов. Один из полученных нами аналогов был на 51 аминокислоту короче и имел всего лишь 38% идентичности исходному белку. Тем не менее, данный аналог был способен извлекать каротиноиды из мембраны и переносить их к другим растворимым белкам-реципиентам, подобно исходному белку.

Также мы предложили способ применения ProteinMPNN, при котором итоговый белок имеет заданный заряд при нейтральном pH и заданное количество потенциально заряженных аминокислот. В случае с AstaPo1, такой подход может помочь оптимизировать уровень экстракции каротиноидов из заряженных мембран.

Наши результаты показывают, что ProteinMPNN может быть использован для создания разнообразных функциональных аналогов природных лиганд-связывающих белков с высокой вероятностью успеха, а также демонстрируют возможность применения ProteinMPNN к белкам, структура которых получена с помощью ЯМР-спектроскопии.

Исследование выполнено при поддержке Министерства науки и высшего образования РФ (номер темы FSMG-2025-0003, соглашение 075-03-2025-662)

1. Dauparas J. et al. Robust deep learning-based protein sequence design using ProteinMPNN // Science. 2022. Vol. 378, № 6615. P. 49–56.

Динамика изменения экспрессии генов в процессе полного превращения у двух филогенетически удалённых видов *Drosophila*

А.М. Озерова¹, М.С. Гельфанд¹

¹*Центр молекулярной и клеточной биологии Сколковского института науки и технологий, Большой бульвар. 30, Сколково, Московская область, 143025, Россия, alexapogorelskaya@gmail.com*

Полное превращение у насекомых — это сложный биологический процесс, сопровождающийся масштабными

морфологическими, физиологическими и транскрипционными перестройками. Чтобы изучить динамику экспрессии генов на этом ключевом этапе развития, мы провели сравнительный анализ транскриптомов на разных стадиях метаморфоза у двух видов: *Drosophila melanogaster* и *Drosophila virilis*.

Мы подтвердили частичное воспроизведение эмбриональной транскрипционной программы у куколок¹. Однако вместо традиционной для эмбриогенеза модели «песочных часов», предполагающей максимальную консервативность экспрессии на средних эмбриональных стадиях, в развитии куколки мы наблюдаем более сложную картину чередования низкого и высокого уровня разнообразия. Этот процесс напоминает обращенную модель «песочных часов», или «веретено».

Также примечательно, что недавно возникшие гены, специфичные для насекомых, демонстрируют ярко выраженные пики экспрессии на средних стадиях развития куколки. Это указывает на их вероятную роль в регуляции переходов между различными фазами развития.

1. Ozerova AM, Gelfand MS, Scientific Reports 2022, 12 (1): 17570.

Выявление биомаркеров метилирования для определения возраст-зависимых заболеваний

В.Г. Палагина¹, А.И. Ступников², Д. О. Крюков¹, Е. Г. Ефимов¹,
Е. Е. Храмева¹

¹Сколковский институт науки и технологий, Москва, Россия
Viktoriiia.Palagina@skoltech.ru

²Институт общей генетики имени Н. И. Вавилова РАН, Москва,
Россия

Старение представляет собой постепенное снижение

физиологических функций, что увеличивает риск развития возраст-зависимых заболеваний, таких как рак, деменция, диабет и сердечно-сосудистые расстройства [Franceschi *et al.*, 2018]. В последние десятилетия продолжительность жизни человека значительно увеличилась, что сопровождалось ростом распространенности хронических заболеваний. Это подчеркивает необходимость разработки новых технологий, способных улучшить качество жизни пожилых людей [Pignolo, 2019].

Одним из перспективных направлений решения этой задачи является создание методов выявления ускоренного старения, что позволит подбирать правильные терапевтические стратегии. Для этого важно определить биологический возраст — показатель, отражающий функциональное состояние организма [Pyrkov, 2019]. Исследования показали, что уровни метилирования ДНК коррелируют с возрастом и могут служить надёжными биомаркерами биологического возраста [Christiansen *et al.*, 2016]. Такие модели, основанные на метилировании, называются эпигенетическими часами.

Несмотря на их широкое применение, до сих пор не существует общепринятого золотого стандарта оценки точности и надёжности эпигенетических часов. Отсутствие универсальной метрики затрудняет интерпретацию и сравнение результатов разных моделей [Kriukov *et al.*, 2024]. Более того, многие существующие часы предсказывают хронологический возраст с высокой точностью, но не всегда способны достоверно определять биологический возраст и выявлять возраст-ускоренные заболевания [Rollandi *et al.*, 2019].

Для повышения точности эпигенетических часов необходимо оптимизировать сигнатуры метилирования, ориентированные на распознавание признаков заболеваний. Вместо оценки возраста как такового, такие сигнатуры должны фокусироваться на изменениях метилирования, характерных для патологических состояний, развивающихся в следствие старения. Это позволит построить модели, чувствительные к ухудшению здоровья и отражающие биологический возраст более адекватно.

Поэтому целью исследования стала разработка универсальной сигнатуры метилирования, способной надёжно распознавать возраст-ускоряющие состояния. Для этого были

решены следующие задачи: (1) выявлен подход к извлечению эффекта хронологического возраста из данных; (2) оценены 13 опубликованных эпигенетических часов по их способности различать здоровых и больных индивидуумов на основе метрики H-score; (3) проведен отбор и аннотация набора CpG сайтов с высоким предсказательным потенциалом; (4) построена новая модель часов старения.

Поскольку возраст оказывает сильное воздействие на уровни метилирования необходимо исключить этот эффект, чтобы сделать сигнал, отражающий заболевание более заметным. Для этого протестированы различные модели регрессии. Наиболее эффективной оказалась множественная линейная регрессия.

Было проведено сравнение сигнатур эпигенетических часов на предмет их способности распознавать заболевания, ускоряющие старение. GrimAgeV1 и GrimAgeV2 показали наилучший результат, различая 26 из 41 наборов данных.

Для создания новой сигнатуры был проведен дифференциальный анализ метилирования между здоровыми и больными. Для каждой болезни определялись наиболее значимые CpG-сайты. Наилучший результат показала сигнатура из 59 CpG-сайтов, успешно различая больных и здоровых в 31 из 39 наборов данных, что превышает результаты существующих часов. Проведена аннотация отобранных сайтов: 82% сайтов расположены в генах, преимущественно в промотерах. EWAS-анализ показал ассоциации сайтов с различными возраст-ассоциированными заболеваниями.

Используя выбранную сигнатуру, была построена модель эпигенетических часов с помощью ElasticNet регрессии. Бенчмаркинг моделей с помощью пайплайна ComputAgeBench [Kriukov *et al.*, 2024] показал, что построенные нами часы имеют устойчивый результат в тестах на выявление заболеваний, а также получают общую оценку в 23 балла по шкале BenchScore, уступая только модели часов второго поколения PhenoAgeV2.

Наша модель часов продемонстрировала высокую чувствительность к возраст-ускоряющим заболеваниям несмотря на то, что для ее построения не использовались данные по продолжительности жизни, что характерно для часов второго поколения. Данная модель обладает диагностической ценностью и

может использоваться как универсальный биомаркер функционального состояния организма.

1. C. Franceschi et al. (2018) Inflammaging: a new immune–metabolic viewpoint for age-related diseases, *Nat Rev Endocrinol*, **14**:576–590.
2. R.J. Pignolo (2019) Exceptional human longevity, *Mayo Clinic Proceedings*, **94**(1):110–124.
3. T.V. Pyrkov et al. (2019) Biological age is a universal marker of aging, stress, and frailty, In: *Biomarkers of Human Aging*, 23–36.
4. L. Christiansen et al. (2016) Longitudinal changes in the genetic and environmental determinants of biological age, *Biol Sci Med Sci*, **71**(1):45–52.
5. D.V. Kriukov et al. (2024) ComputAgeBench: Epigenetic Aging Clocks Benchmark, *bioRxiv*.
6. G.A. Rollandi et al. (2019) Biological age versus chronological age in the prevention of age-associated diseases, *OBM Geriatrics*, **3**(2): article 51.

Роль пионерных факторов в воспроизводимости АТАС-seq экспериментов

С.А. Панишева¹ Д.Р. Салимов², Т.С. Соколова², М.А. Куляшов²,
С.К. Колмыков²

¹Новосибирский государственный университет, 630090,
Новосибирск, ул. Пирогова, 1, s.panisheva@g.nsu.ru

²Научно-технологический университет «Сириус», 354340,
Российская Федерация, Краснодарский край, федеральная
территория «Сириус», Олимпийский пр., д.1

Введение

Структурная организация хроматина обеспечивает избирательный доступ регуляторных белков к ДНК и играет ключевую роль в транскрипционной пластичности, позволяя клеткам быстро реагировать на изменения окружающей среды и внутренние сигналы (A.Balsalobre, J.Drouin, 2022). Одним из механизмов этой регуляции является динамическое изменение плотности упаковки хроматина, которое нередко инициируется пионерными транскрипционными факторами (пТФ) (R.Lemma et al., 2022).

В данной работе были картированы события проявления пионерной активности ТФ и исследовано, как активность пТФ влияет на воспроизводимость районов открытого хроматина (РОХ) в bulk- и single-cell ATAC-seq экспериментах.

Результаты

На первом этапе на основе bulk- и single-cell- ATAC-seq данных из БД GTRD была проанализирована частота встречаемости (воспроизводимость) РОХ в схожих по условиям проведения экспериментах. Было показано, что воспроизводимость РОХ значительно варьирует между биологическими репликами. Данный эффект может быть связан с гетерогенностью открытого хроматина на уровне клеточных субпопуляций. Для исследования данной взаимосвязи была подсчитана воспроизводимость РОХ внутри каждого scATAC-seq эксперимента, и сопоставлена с воспроизводимостью данных РОХ между биологическими репликами (см. Рисунок 1). Данный подход позволил выявить районы конститутивно открытого хроматина, а также продемонстрировать корреляцию между воспроизводимостью РОХ на разных уровнях.

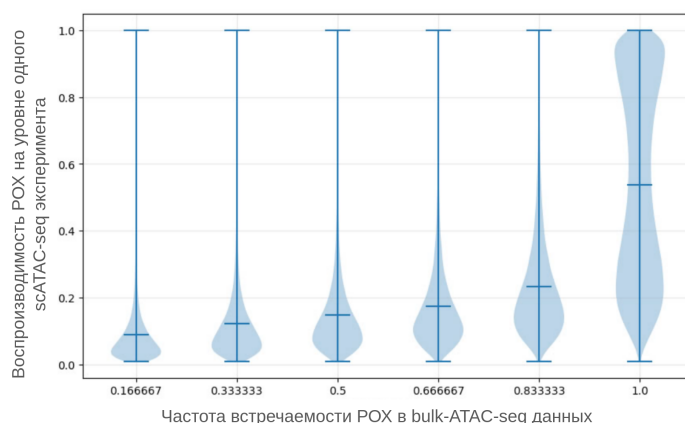


Рисунок 1. Сопоставление воспроизводимости районов открытого хроматина (POX) в каждом scATAC-seq эксперименте с частотой встречаемости POX в схожих по условиям проведения bulk-ATAC-seq экспериментах для НЕК293Т.

Далее POX были сопоставлены с районами связывания ТФ (РСТФ), основанными на ChIP-seq данных из GTRD. Наличие ассоциации ТФ с ремоделингом хроматина определялась на основе результатов работы Lemma с соавт. (R.Lemma et al., 2022). Наблюдалось статически значимое обогащение потенциальных пТФ в менее воспроизводимых POX, в частности, в POX, в небольших субпопуляциях клеток.

Чтобы подтвердить взаимосвязь между наличием РСТФ и инициацией процессов ремоделинга хроматина, были использованы БД: UDACHA (A.Buyan et al., 2025) и ADAstra (S.Abramov et al., 2021). В БД UDACHA хранится информация, отражающая аллель-специфичный уровень упаковки хроматина (ASV) в различных тканях и клеточных типах, а БД ADAstra описывает аллельный дисбаланс связывания ТФ в ChIP-seq данных. Присутствие SNV одновременно в двух этих базах данных в одном и том же клеточном типе говорит о наличии взаимосвязи между эффективностью связывания ТФ и степенью упаковки хроматина. Также было оценено влияние SNV, связанных с ASV, на точность распознавания мотива связывания ТФ при помощи позиционно-весовых матриц из БД NOCOMOCO (I.Vorontsov et al., 2024). Таким образом были получены список пТФ и картированы районы, в

которых демонстрируется их активность, связанная с ремоделингом хроматина. Полученные данные также подтверждают взаимосвязь РСТФ пТФ с РОХ, демонстрирующими низкую воспроизводимость, в частности, за счет клеточных субпопуляций.

Заключение

Благодаря интеграции bulk- и single-cell ATAC-seq данных вместе с аннотациями из БД: GTRD, UDACHA и ADAstra была продемонстрирована взаимосвязь между воспроизводимостью РОХ в ATAC-seq данных и присутствием РСТФ пТФ. Были идентифицированы РСТФ, ассоциированные с инициацией процессов ремоделинга хроматина и показана их ассоциация с РОХ в малых субпопуляциях клеток.

Исследование выполнено за счет гранта Российского научного фонда № 24-24-20108, <https://rscf.ru/project/24-24-20108/>.

Список литературы

1. A.Balsalobre, J.Drouin (2022) Pioneer factors as master regulators of the epigenome and cell fate, *Nature Reviews Molecular Cell Biology*, **23**:449-464.
2. R.Lemma et al. (2022) Pioneer transcription factors are associated with the modulation of DNA methylation patterns across cancers, *Epigenetics & Chromatin*, **1**:1-19.
3. A.Buyan et al. (2025) Statistical framework for calling allelic imbalance in high-throughput sequencing data, *Nature Communications*, **16**:1739.
4. S.Abramov et al. (2021) Landscape of allele-specific transcription factor binding in the human genome, *Nature Communications*, **12**:2751.
5. I.Vorontsov et al. (2024) HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors, *Nucleic Acids Research*, **52**:D1.

**Открытый конкурс по предсказанию специфичности
связывания факторов транскрипции человека на различных
типов экспериментальных данных**

Д. Д. Пензар¹, Н.С. Грызунов², С. Е. Бушуев²,
В.В. Стрекаловских², А.О. Зинкевич², И.В. Кулаковский^{1,3}

¹*Институт общей генетики им. Н.И. Вавилова РАН*
dmitrypenzar1996@gmail.com

²*Факультет биоинженерии и биоинформатики МГУ*

³*Институт белка РАН*

Взаимодействия транскрипционных факторов (ТФ) с ДНК являются важным компонентом регуляции транскрипции, ответственным за скоординированную экспрессию генов в сетях регуляции генов. На сегодняшний день разработано множество алгоритмов, позволяющих предсказывать связывание ТФ с последовательностью и влияние на него замен в сайтах связывания. Среди классических методов, наиболее популярным являются занимают подходы, основанные на позиционно-весовых матрицах (ПВМ). В связи с этим разработано множество вычислительных инструментов для распознавания образов для построения ПВМ, описывающих специфичность связывания факторов транскрипции. Одновременно с этим продолжают предлагаться более сложные модели, многие из которых основаны на методах машинного обучения, включая глубокие нейронные сети.

Основным недостатком ПВМ является явное предположение о независимости вкладов каждой позиции сайта связывания ТФ в итоговое взаимодействие. Однако несмотря на это ограничение, ПВМ до сих пор широко используются, и в части исследований показывают качество не хуже или даже лучше более современных методов, в том числе основанных на методах машинного обучения¹.

В то же время, даже только для ПВМ имеется лишь малое число исследований, надежно сравнивающих разные подходы для их построения. Недавно нами было проведено систематическое

сравнение различных методов для построения ПБМ в конкурсе МЕХ, проводившемся в пределах консорциума GRECO-BIT².

В рамках данного исследования мы подготовили провели открытый конкурс IBIS (Inferring Binding Specificities, <https://ibis.autosome.org/>), в котором участникам предлагалось участие в нескольких номинациях, основными из которых являлись номинация A2G – обучение моделей на данных *in vitro* экспериментов (ПБМ, SMiLE-seq и HT-SELEX) о связывании ТФ и тестирование на данных *in vivo* (GHT-SELEX и ChIP-seq) – и номинация G2A, где обучение моделей осуществлялось на *in vivo* данных, а тестирование – на данных *in vitro*.

В ходе анализа конкурса мы выявили наиболее успешные решения, как среди основанных на ПБМ, так и среди использующих методы машинного обучения. Нами было показано, что хоть решения на основе ПБМ действительно демонстрируют высокое качество предсказаний, методы на основе нейронных сетей при правильной настройке превосходят их во всех дисциплинах. При этом модели победителей обеих номинаций оказались основаны на современных сверточных нейронных сетях, в то время как подходы на основе трансформерных и рекуррентных архитектур показывали стабильно более низкие результаты.

Далее мы проверили, насколько модели, обученные участниками, способны события аллель-специфичного связывания³. Оказалось, что и в этом случае модели на основе методов машинного обучения показывают себя лучше, нежели как модели ПБМ из конкурса IBIS, так и модели ПБМ, отобранные в ходе внутреннего соревнования МЕХ.

Помимо этого, в силу того, что участники конкурса не предоставили решений, основанных на популярных в наше время днк-языковых моделях, мы самостоятельно оценили качество таких моделей как GENE-LM⁴, DNABERT2⁵ и Nucleotide Transformer⁶. В соответствии с результатами недавних работ⁷, эти модели показали крайне низкое качество в конкурсе, уступая в том числе и решениям, основанным на позиционно-весовых матрицах.

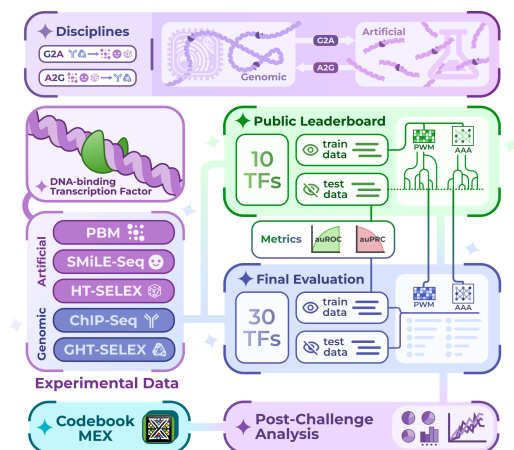


Рисунок 1. Общая схема конкурса IBIS

Работа была выполнена в рамках Госзадания FFRW-2025-010.

Список литературы

1. Boytsov, A., Abramov, S., Makeev, V. J. & Kulakovskiy, I. V. Positional weight matrices have sufficient prediction power for analysis of noncoding variants. *F1000Res*. **11**, 33 (2022).
2. Vorontsov, I. E. *et al.* Cross-platform DNA motif discovery and benchmarking to explore binding specificities of poorly studied human transcription factors. *Bioinformatics* (2024).
3. Jolma, A. *et al.* Perspectives on Codebook: sequence specificity of uncharacterized human transcription factors. *Genomics* (2024).
4. Fishman, V. *et al.* GENA-LM: a family of open-source foundational DNA language models for long sequences. *Nucleic Acids Res.* **53**, (2025).
5. Zhou, Z. *et al.* DNABERT-2: Efficient foundation model and benchmark for multi-species genome. *arXiv [q-bio.GN]* (2023).
6. Dalla-Torre, H. *et al.* Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nat. Methods* 1–11 (2024).
7. Xu, Z. *et al.* Specialized foundation models struggle to beat supervised baselines. *arXiv [cs.LG]* (2024).

Создание стандартизованного набора результатов массовых параллельных репортерных экспериментов для предсказания регуляторной активности коротких последовательностей с помощью машинного обучения

Пензин Н.С.¹, Зинкевич А.О.¹, Кулаковский И.В.^{2,3}, Пензар Д.Д.³

¹*ФББ МГУ им. М. В. Ломоносова, Москва;
nios.583@gmail.com*

²*Институт белка РАН, Пуцино*

³*Институт общей генетики им. Н.И. Вавилова РАН, Москва*

Массовый параллельный репортерный анализ (МПРА, англ. Massively Parallel Reporter Assay — MPRA) — высокопроизводительный метод исследования функциональной активности регуляторных элементов ДНК (промоторов, энхансеров, сайленсеров) вне их нативного геномного контекста[1]. Благодаря возможности одновременного тестирования от десятков тысяч до десятков миллионов последовательностей, MPRA стал ключевым высокопроизводительным инструментом в функциональной геномике. С помощью методов классической биоинформатики и машинного обучения он позволяет моделировать регуляторную грамматику клетки, в частности, предсказывать результаты взаимодействия транскрипционных факторов, эффекты регуляторных вариантов и проектировать последовательности с клеточно-специфической активностью.

Однако отсутствие стандартных практик работы с этими данными осложняет сравнение моделей машинного обучения (МО) из разных работ даже в случае, когда модели обучаются и тестируются на одних и тех же выборках. Все это приводит к высокому порогу вхождения для исследователей, не обладающих специализированными навыками и затрудняет разработку

обучающих материалов по машинному обучению в регуляторной геномике.

Для решения данной проблемы нами был разработан пакет MPRA-MNIST, содержащий информацию о регуляторной активности более 75М последовательностей из 4 организмов, объединенных в 13 предобработанных в унифицированный формат наборов данных. К каждому набору данных приложен пример обучения глубокой нейронной сети на этих данных. Эти примеры основаны на рекомендациях авторов оригинальных работ и призваны еще больше облегчить усилия, необходимые для начала работы с данными MPRA(Рис 1).

В силу описанных особенностей, MPRA-MNIST — по аналогии с наборами данных MNIST (Modified National Institute of Standards and Technology), представляющим собой обширную базу рукописных цифр, и MedMNIST[2] (коллекцию медицинских снимков разной природы) — значительно снизит сложность обучения применению глубоких нейронных сетей в задачах регуляторной геномики, повысит воспроизводимость исследований и ускорит разработку новых алгоритмов предсказания активности регуляторных элементов.



Рис 1. Схематичное изображение пакета MPRA-MNIST. Более 75М последовательностей объединены в 13 наборов данных. Для каждого набора приложен пример использования и обучения глубокой нейронной сети на основе оригинальных исследований.

1. Romanov S.E., Laktionov P.P. Practical application of massively parallel reporter assay in biotechnology and medicine // Клиническая практика.– ECO-Vector LLC, 2023.– Vol. 13, № 4.– P. 74–87.
2. Yang, J., Shi, R., Wei, D. et al. MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci Data 10, 41 (2023). <https://doi.org/10.1038/s41597-022-01721-8>

Работа была выполнена в рамках Госзадания FFRW-2025-010.

Пост-транскрипционная регуляция экспрессии генов по механизму непродуктивного сплайсинга

Д.Д. Первушин

*Сколковский институт науки и технологий, Москва, территория
инновационного центра «Сколково», Большой бульвар, д. 30,
стр. 1, d.pervouchine@skotech.ru*

Экспрессия генов эукариот контролируется большим числом факторов, которые поддерживают баланс между синтезом и деградацией мРНК. Появление нонсенс-мутаций и сдвигающих рамку считывания ошибок сплайсинга приводит к возникновению изоформ мРНК, содержащих преждевременные стоп-кодоны. У эукариот существует система селективной деградации таких транскриптов, называемая нонсенс-опосредованным распадом (nonsense mediated decay, NMD).

Ранее считалось, что основная функция NMD состоит в предотвращении трансляции усеченных, и поэтому вредоносных белков. Однако в последнее время появляется все больше свидетельств тому, что NMD повсеместно используется для регуляции уровня экспрессии генов. Непродуктивный сплайсинг - это посттранскрипционный механизм контроля экспрессии генов эукариот, при котором в результате регулируемого альтернативного сплайсинга в белок-кодирующих транскриптах возникают преждевременные стоп кодоны, что приводит к их деградации по пути NMD. Этот механизм в особенности характерен для РНК-связывающих белков, которые контролируют уровни экспрессии друг друга и других генов с помощью множественных авто- и кросс-регуляторных каскадов. Нарушения непродуктивного сплайсинга приводят к развитию различных заболеваний, в том числе опухолей, и потенциально могут служить терапевтическими мишенями.

В докладе будет дан обзор механизмов функционирования системы NMD и известных из литературы случаев регуляции экспрессии генов по механизму непродуктивного сплайсинга [1]. Основной акцент будет сделан на разработку биоинформатических методов для нахождения таких событий по данным высокопроизводительного секвенирования РНК. В частности,

будет рассказано о предсказании ауторегуляторных [2] и кросс-регуляторных [3] сетей непродуктивного сплайсинга. Также будет рассказано о регуляции непродуктивного сплайсинга эволюционно консервативными РНК-структурами, независимо возникшими у паралофов из семейства ВЕТ белков [4], и об обнаружении значительного числа криптических ядовитых экзонов в интронах генов человека, и их особенных тканеспецифических и регуляторных свойствах [5].

Работа была выполнена при поддержке грантов Российского Научного Фонда №21-64-00006 и №22-14-00330.

1. L.G. Zavileyskiy, D.D. Pervouchine. Post-transcriptional Regulation of Gene Expression via Unproductive Splicing. (2024) *Acta Naturae*. **16**(1):4-13.
2. D. Pervouchine D, Y. Popov, A. Berry, B. Borsari, A. Frankish, R. Guigó. (2019) Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. *Nucleic Acids Res.* **47**(10):5293-5306.
3. A. Mironov, M. Petrova, S. Margasyuk, V. Vlasenok, A.A. Mironov, D. Skvortsov, D.D. Pervouchine. Tissue-specific regulation of gene expression via unproductive splicing. (2023) *Nucleic Acids Res.* **51**(7):3055-3066.
4. M. Petrova, S. Margasyuk, M. Vorobeve, D. Skvortsov, O.A. Dontsova, D.D. Pervouchine. (2024) BRD2 and BRD3 genes independently evolved RNA structures to control unproductive splicing. *NAR Genom Bioinform.* **6**(1):lqad113.
5. S. Margasyuk, A. Kuznetsova, L. Zavileyskiy, M. Vlasenok, D. Skvortsov, D.D. Pervouchine. (2024) Human introns contain conserved tissue-specific cryptic poison exons. *NAR Genom Bioinform.* **6**(4):lqae163.

Phylogenetic analysis and expression analysis of processed human pseudogenes in glioblastoma

N.A.Pogodin¹, T.G. Chikadze¹, A.A. Makashov²

¹*SBEI HCS 225, St.Petersburg, Russia, nikita.al.pogodin@gmail.com*

²*Peter the Great St.Petersburg Polytechnic University, St.Petersburg,*

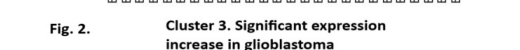
In this study, we performed a phylogenetic analysis of human processed pseudogenes and analyzed their expression level in glioblastoma and normal human brain cells.

Sequences and coordinates of pseudogenes were obtained from the Ensembl database. SyntenyPortal was used to search for syntenic regions. GDC Data Portal (TCGA project) and GTEx databases were used to analyze the expression of pseudogenes in glioblastoma and normal tissues respectively.

Homologs were searched by the HMMER program package in automatic mode using the author's program code. The search was carried out in the genomes of the following species: *Pan troglodytes*, *Papio anubis*, *Callithrix jacchus*, *Canis lupus*, *Mus musculus* and *Monodelphis domestica*.

We performed the phylogenetic analysis the following processed human pseudogenes: FAM3C2P, GPX1P1, COX20P1, TTC3P1, SNRPGP15, EEF1A1P11, CALM2P2, YWHAZP3, EIF4HP1, PEBP1P2, EEF1A1P35, RNPS1P1. The COX20P1 gene is specific only to humans; the FAM3C2P, GPX1P1, TTC3P1, CALM2P2, EIF4HP1, SNRPGP15, YWHAZP3, PEBP1P2, RTN3P1 and RNPS1P1 genes originated within primates; and the EEF1A1P11 and EEF1A1P35 genes were found in Boreoeutheria.

By analyzing the expression of pseudogenes in normal and glioblastoma, 4 clusters were identified (fig. 1). Cluster 0 - genes are highly correlated with each other in terms of expression level, many genes start to be expressed in the tumor or increase expression in the tumor. Cluster 1 - correlation is weak or none at all, single genes slightly increase expression in tumor, Cluster 2 - no significant changes in expression level. Cluster 3 - expression of some genes decreases in the tumor (others are not expressed at all), strong positive correlation of



121

S. R. Prokopchuk ^{1,2}, A.V. Klepikova ², A.M. Kasianova ^{1,2}, V.M. Zubarev ³, D.D. Sokoloff ⁴, M.V. Remizova ³, M.D. Logacheva ^{1,2},
A.A. Penin ²

¹ *Skolkovo Institute of Science and Technology, Center of Molecular and Cellular Biology, Moscow, Russia*
psr1653@gmail.com

² *Vavilov Institute of General Genetics, Laboratory of Plant Genomics, Moscow, Russia*

³ *Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia*

⁴ *School of Plant Sciences and Food Security, Faculty of Life Sciences, Tel Aviv University, Tel Aviv 6997801, Israel*

The core transcriptome of an organ which determines its identity is presumably conserved in the evolution. The analysis of gene expression patterns can be used to identify orthologous genes with expression restricted to specific tissues of an organism (tissue-specific) in a wide variety of species. Such genes are the main part of core organ transcriptome and the analysis of conserved transcription factor motifs can result in identification of common regulation of tissue-specific genes. This approach is particularly relevant for angiosperms, a morphologically diverse and species-rich group of plants with relatively few well-characterized model organisms. The discovery of such regulatory elements holds significant potential for practical applications. For example, anther-specific regulatory elements may allow the development of crop varieties with induced male sterility, thereby facilitating the production of hybrid seeds.

In this study, we assembled transcriptomes of 47 species of flowering plants and two species of gymnosperms using Trans2express software [1] on short Illumina reads and long Oxford Nanopore reads. The resulting transcriptome assemblies consisted of longest isoform for each gene expressed in studied organs: anther, leaf, ovule, and seed.

We then used OrthoFinder to identify orthologous genes across all assembled transcriptomes together with four well-characterized

species (*Zea mays*, *Fagopyrum esculentum*, *Arabidopsis thaliana*, and *Solanum lycopersicum*). The resulting groups of orthologous genes (orthogroups) were searched for orthogroups containing genes with expression patterns restricted to a single plant organ. The genes belonging to such orthogroups were analyzed in terms of gene and promoter characteristics thus allowing to infer the common regulatory patterns of tissue-specific genes across wide variety of plant species.

1. A. M. Kasianova et al. (2024) Trans2express—de novo transcriptome assembly pipeline optimized for gene expression analysis, *Plant Methods*, **20**:128

The study was supported by Vavilov Institute of General Genetics, project # FFRW-2024-0003.

The concept of cell assemblability

L. Yu. Rusin

*Institute for Information Transmission Problems (Kharkevich Institute),
Russian Academy of Sciences, Bolshoy Karetny per., 19 build. 1,
127051 Moscow, Russian Federation, rusin@iitp.ru*

Since the advent of cell theory in middle 19th century, the diversity of cell kinds has naturally been classed into “cell types” based on their outward (morphological) and/or inner (physiological) properties. While any rigorous phenotypic definition of cell type is inherently lacking, the observable, intuitively distinct cell types, such as animal muscle, neuron, epithelial or stem cells, have traditionally been interpreted as representing elementary organismal units that are directly comparable in evolution and development, and that form according to the same basic logic that underlies the organismal tree of life. Nearly two centuries after, the doctrinal view of cell identity is called into question due to the discovery of key molecular signatures at the level of gene expression in single cells that incontrovertibly demonstrate the presence of backbone similarities that permeate the cells regardless of their structure, form, function or developmental origin, both between-species

and within a same organism. The demand to place this major new fact of life within a theoretical framework is answered by the emerging “cell type theory”, which imputes a primary role to gene regulation in defining the biological nature of cell type. It rethinks the cell type as a population of cells separated from other cells in a multicelled organism via regulation-determined, selective access to common genomic information. Cell type identity is thereby defined as a core set of active molecular regulators of gene transcription (transcription factors), whereas evolutionary change in this core drives the birth and divergence of cell types, determines their genealogy and, hence, innate homology. The emerging regulatory paradigm of cell identity has fundamental implications by providing objective criteria to establish and interpret evolutionary links at the cell level, explaining why and how related cell types can diverge beyond recognition, and demonstrating that cell phylogeny is not equivalent to ontogenetic cell lineage. Although transformative to current biology, this concept still operates with cell types as static and discrete states to describe individual cells, both in evolution and development, which is at odds with the heterogeneity of single-cell data observed at various resolutions and sequencing depths of a single sample or time series, and is not explicative of the cell’s remarkable plasticity to blend and swap functions and flow between states in a cell lineage.

The concept of *cell assemblability* explained here is introduced to provide a more realistic theoretical framework, detailing the sources and mechanisms that shape cell identity in two dimensions, genotypic (in evolution) and phenotypic (in development), via assembly of regulatory circuits as individual “building units”. It rethinks the living cells as *avatars*, transient or terminal cell states deployed in a continuum of states by the developmental programme of one and the same omnipotent cell of the germline (Rusin, 2023), encapsulating the baseline definition of cell type (Arendt et al., 2016) in terms of the basic mechanism for the origin and divergence of novel types via regulatory isolation, whereby a cell type is only equivalent to a molecular cell identity acquired with a newly established (quasi)autonomous regulatory circuit. Meanwhile, it explicitly decouples cell identity from the observable living cell, treating the cell as a dynamic entity that can swap and mix identities, essentially distilling the classical, intuitive cell types (such as muscle, neuron or epithelia) into purely conventional, transient entities, with the

recognition that “cell evolution” and “fate commitment” actually refer to the histories of individual regulatory module transformations and firings, respectively, in the context of a common developmental programme. The concept explains how this logic of cell defines the re-use and redundancy of its “building units” in both phylo- and ontogenesis, enabling the mosaic nature of cell identities and lineages, respectively, governs cell reprogramming and fate transitions (highlighting implications for applied biology and medicine; Zhang et al., 2025), and renders mosaic cell evolution fundamentally different from the vertical evolution of species.

The main statements are as follows:

- 1) The cell is formed with assembly in two dimensions: its identities are irreversibly assembled during evolution at the genomic level from regulatory modules of diverse origins to form a scaffold of ontogenesis, where the cell phenotype is assembled reversibly by swapping, mixing and modifying evolutionary identities to form a versatile diversity of virtually unique cells, possibly including bespoke, non-physiological states that never existed in evolution or normal development.
- 2) The cell is relativistic in terms of an identity with respect to its current regulatory state. This identity can be reassembled seamlessly over large evolutionary distances and a dramatic range of morphology and function, whereby individual cells act as biological avatars operated by regulatory networks of one and the same founder cell, the omnipotent cell of the germline, who interactively morphs to change age, guise and function as the networks are rewired.
- 3) The cell’s assemblability defines the hybrid, mosaic and heterochronic nature of cell identities and lineages:
 - (a) many identities share the same portions of the genome for deployment and are therefore horizontally related in a convoluted and inherently network-like phylogeny – the mode that differs principally from vertical divergence, making cell evolution fundamentally different from the evolution of species;
 - (b) newly invented regulatory modules are usually linked to small cellular functions and can be re-used in multiple contexts during both evolution (in identities with distinct ages and origins) and ontogenesis (in those deployed at distinct times or lineages), producing a diversity of cell types spaced in evolution and development but related by parts of the common genome;

(c) the age of cell identity is decoupled from that of its supporting genes, setting its lower bound at the origin of the youngest genes, while leaving the upper bound open;

(d) the order of identity recall in a cell lineage may differ from their evolutionary succession, whereby older identities may succeed the younger ones.

4) Cell fate transitions are governed by cell transcriptional competency unfolded by pioneer transcription factors:

(a) the competencies are conserved and delineate cell differentiation trajectories into robust developmental tunnels; within-tunnel transitions proceed mechanistically and lineage-autonomously to facilitate developmental redundancy and robustness at the primal cellular level, whereas those across tunnels require specific triggers that induce turnpoint conditions stitching the competencies and their tunnels into a fabric of development;

(b) a turnpoint is theoretically a starting mixture of cell identities induced by a joint action of pioneer factors, which needs not occur in a stem cell;

(c) cell fate transitivity and convertibility depend on the amount of genomic regulatory sites shared between the competencies, thus reflecting the amount of evolutionary distance between them and the cell diversity they support, and defining the type of fate transition, from seamless direct conversion to reprogramming via pluripotency;

(d) the ages of the tunnels and their supported developmental domains are approximated by the ages of the pioneer transcription factors that unfold the underlying competencies.

5) A cell recalls ancestral identities as it differentiates within a competency tunnel, thereby broadly following the recapitulation law of classical biology at the cell lineage level, which states that ontogeny recapitulates phylogeny.

6) The genome stores at least as many instructions for the deployment of ancestral cell types as there are intermediate identities realised in a cell lineage, while some become overridden by evolutionarily newer identities and fossilise as silenced cell types; many fossil instructions are expected to remain, forming the “dark diversity” of cell types that can potentially be resurrected by inactivation of the overriding module regulator gene(s).

The work was supported by the Russian Science Foundation (24-44-00099; <https://rscf.ru/en/project/24-44-00099/>).

1. L.Yu.Rusin (2023) *Journal of Morphology*, **284**: e21569.
2. D.Arendt et al. (2016) *Nature Reviews Genetics*, **17**: 744–757.
3. X.Zhang et al. (2025) *Advanced Science*, accepted (10.1002/advs.202406208)

Effect of placental serotonin stimulation via maternal signaling on hypothalamic developmental plasticity and offspring's behavior

Marat Sabirov¹, Anastasiia Gainullina¹, Victoria Melnikova¹, Evgeniia Chikina¹, Zakhar Starinnov¹, Elena Shagimardanova², Oleg Gusev², Roman Romanov³

¹*Koltzov Institute of Developmental Biology of the Russian Academy of Sciences, Moscow, Russia, m.sabirov@idbras.ru*

²*Life Improvement by Future Technologies (LIFT) Center, Moscow, Russia*

³*Department of Molecular Neurosciences, Center of Brain Research, Medical University of Vienna, Vienna, Austria
Roman.Romanov@meduniwien.ac.at*

Neurodevelopment is shaped by both genetic and environmental factors, with maternal conditions such as stress and infection impacting fetal brain development through epigenetic mechanisms (Hoek et al., 1998; Brown, 2012; Toth, 2015). Serotonin, beyond its role as a neurotransmitter, serves as an epigenetic modulator during early brain formation (Bonnin et al., 2011; Farrelly et al., 2019). Recent studies suggest that maternal serotonin may act as a transgenerational signal influencing offspring brain organization and social behavior.

To test this hypothesis, we induced a transient physiological elevation of maternal serotonin levels through oral 5-HTP administration in pregnant rats during embryonic days 11 to 14. We assessed offspring outcomes through behavioral assays and performed single-cell RNA

sequencing (scRNA-seq) on the hypothalamus at embryonic day 20 and postnatal day 28. Additional analyses included HPLC measurement of serotonin levels, regulatory network reconstruction using SCAFE and Cicero, spatial transcriptomics analyses, and cell communication assessment via CellChat. Furthermore, we utilized GAM-clustering to explore metabolic pathways at the single-cell level.

We confirmed that placental and fetal brain serotonin levels are elevated during critical neurogenic stages (Bonnin et al., 2011; Lopez-Rodriguez et al., 2023). Behavioral analysis revealed that offspring exposed to elevated prenatal serotonin exhibited reduced aggression and enhanced social interaction patterns.

Single-cell RNA-seq uncovered compositional and transcriptional shifts in GABAergic and glutamatergic neuronal populations within the hypothalamus. By integrating scRNA-seq data with spatial transcriptomics data, we successfully mapped the affected neuronal populations to distinct hypothalamic nuclei, revealing region-specific susceptibility to prenatal serotonin elevation. Neuronal activity was changed in socially-relevant regions such as the lateral preoptic area, dorsomedial hypothalamus, and zona incerta (Dawson et al., 2023; Froemke and Young, 2021).

Accelerated oligodendrocyte maturation was associated with lipid metabolism shifts, especially a transition from phosphatidylcholine to phosphatidylethanolamine dominance (Naffaa et al., 2022). Using GAM-clustering, we predicted metabolic modules linked to oligodendrocyte differentiation, highlighting perturbed sphingomyelin and phospholipid pathways critical for myelination and validated this by lipidome mass spectrometry. Region-specific alterations in lipid profiles further emphasized the functional impact on neuronal circuitry stability and plasticity.

At the molecular level, prenatal serotonin elevation resulted in the downregulation of key neurodevelopmental transcription factors *Nfia*, *Nfib*, and *Nfix* (Deneen et al., 2006; Fraser et al., 2020), which are pivotal for gliogenesis and neurogenesis. Regulatory network analysis revealed a broad remodeling of regulatory landscape. Using SCAFE and Cicero, we counted the co-activity of regulatory modules and demonstrated that prenatal serotonin exposure leads to decreased co-activity of regulatory landscape, especially in progenitor cell populations transitioning towards neuronal fates. Notably, Wnt signaling pathway

activity was upregulated, supporting enhanced progenitor proliferation and differentiation (Clevers et al., 2014). These regulatory changes were consistent with the altered expression of synaptic and axonogenesis gene programs.

In conclusion, maternal serotonin acts as a non-genetic modulator of fetal brain development, affecting neuronal, glial, and metabolic pathways that ultimately shape social behavior in adulthood. Our findings highlight the power of integrating spatial transcriptomics, single-cell gene expression, lipidomics, and regulatory elements analyses to unravel complex neurodevelopmental adaptations. These insights shed light the potential for prenatal interventions targeting serotonergic signaling to enhance brain plasticity and resilience against neurodevelopmental challenges.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2021-1344).

1. Hoek, H.W., Brown, A.S., and Susser, E. (1998). The Dutch famine and schizophrenia spectrum disorders. *Social psychiatry and psychiatric epidemiology*, **33**:373-379.
2. Brown, A.S. (2012). Epidemiologic studies of exposure to prenatal infection and risk of schizophrenia and autism. *Developmental neurobiology*, **72**:1272-1276.
3. Toth, M. (2015). Mechanisms of non-genetic inheritance and psychiatric disorders. *Neuropsychopharmacology*, **40**:129-140.
4. Bonnin, A., Goeden, N., Chen, K., Wilson, M.L., King, J., Shih, J.C., Blakely, R.D., Deneris, E.S., and Levitt, P. (2011). A transient placental source of serotonin for the fetal forebrain. *Nature*, **472**:347-350.
5. Farrelly, L.A., Thompson, R.E., Zhao, S., Lepack, A.E., Lyu, Y., Bhanu, N.V., Zhang, B., Loh, Y.E., Ramakrishnan, A., Vadodaria, K.C., et al. (2019). Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3. *Nature*, **567**:535-539.
6. Lopez-Rodriguez, D., Rohrbach, A., Lanzillo, M., Gervais, M., Croizier, S., and Langlet, F. (2023). Ontogeny of ependymogial cells lining the third ventricle in mice. *Frontiers in Endocrinology*, **13**: 1073759.
7. Dawson, M., Terstege, D.J., Jamani, N., Tsutsui, M., Pavlov, D., Bugescu, R., Epp, J.R., Leininger, G.M., and Sargin, D. (2023).

Hypocretin/orexin neurons encode social discrimination and exhibit a sex-dependent necessity for social interaction. *Cell reports* **42**.

8. Froemke, R.C., and Young, L.J. (2021). Oxytocin, neural plasticity, and social behavior. *Annual Review of Neuroscience*, **44**:359-381.

9. Naffaa, V., Magny, R., Regazzetti, A., Van Steenwinckel, J., Gressens, P., Laprevote, O., Auzeil, N., and Schang, A.L. (2022). Shift in phospholipid and fatty acid contents accompanies brain myelination. *Biochimie*, **203**:20-31.

10. Deneen, B., Ho, R., Lukaszewicz, A., Hochstim, C.J., Gronostajski, R.M., and Anderson, D.J. (2006). The transcription factor NFIA controls the onset of gliogenesis in the developing spinal cord. *Neuron*, **52**:953-968.

11. Fraser, J., Essebier, A., Brown, A.S., Davila, R.A., Harkins, D., Zalucki, O., Shapiro, L.P., Penzes, P., Wainwright, B.J., Scott, M.P., et al. (2020). Common Regulatory Targets of NFIA, NFIX and NFIB during Postnatal Cerebellar Development. *Cerebellum*, **19**:89-101.

12. Clevers, H., Loh, K.M., and Nusse, R. (2014). Stem cell signaling. An integral program for tissue renewal and regeneration: Wnt signaling and stem cell control. *Science*, **346**:1248012.

Облигатные аэробные бактерии демонстрируют более высокую частоту мутаций А>Г по сравнению с анаэробами

А.В. Скуднов¹, Э.А. Бадамшин¹, Б.Э. Ефименко¹, К.В. Гунбин², К.Ю. Попадьян¹, С.В. Денисов³

¹Балтийский федеральный университет им. И. Канта,
Калининград, Россия, alexey.skudnov@yandex.ru

²Институт молекулярной и клеточной биологии СО РАН,
Новосибирск, Россия

³Манчестерский университет, Манчестер,
Великобритания, stepadenisov@gmail.com

Понимание мутационного спектра (MutSpec) дает представление о факторах, вызывающих мутации. В то время как процессы, лежащие в основе специфических мутационных подписей, были широко изучены в раковых и нормальных тканях [1], бактериальным видам уделялось сравнительно мало внимания [2].

Мы представляем BacNeMu - биоинформатический конвейер, предназначенный для реконструкции нейтральных бактериальных мутационных спектров. BacNeMu адаптирован из недавно опубликованного пайплайна NeMu [3]. BacNeMu уникален своей способностью анализировать все доступные бактериальные данные с точностью и легкостью благодаря глубокой интеграции с ведущими прокариотическими базами данных AnnoTree [4] и GTDB [5].

Спектры мутаций, полученные с помощью BacNeMu, демонстрируют хорошее соответствие со спектрами, измеренными непосредственно в экспериментах по накоплению мутаций у нескольких видов бактерий из разных частей филогенетического древа. Это подтверждает, что мы можем строить нейтральные мутационные спектры нашим подходом и позволяет проводить крупномасштабные исследования факторов, формирующих нейтральный бактериальный MutSpec.

В качестве первого анализа мы проверили гипотезу о том, что аэробные и анаэробные виды бактерий демонстрируют различные мутационные спектры. Наш предварительный анализ сделанный на основе 6-компонентных мутационных спектров подтвердил, что мутация A>G (и её комплементарная мутация T>C), которая ранее была связана с реакциями окислительного повреждения в митохондриях [6], значительно чаще встречается у облигатных аэробов по сравнению с облигатными анаэробами. Наблюдаемые переходы также соотносятся с ранее описанными отличиями в GC составе между облигатными аэробами и облигатными анаэробами [7]. Кроме того, мы проанализировали четыре клады близкородственных аэробных и анаэробных бактерий и обнаружили, что мутационный спектр в целом сохраняется из-за филогенетической инерции, что согласуется с данными работы [8]. Также нами показано, что на больших филогенетических расстояниях он в первую очередь определяется аэробным статусом

бактериальной клады [8].

В дальнейшем мы собираемся использовать пайплайн VacNeMu для всех секвенированных/аннотированных видов бактерий, чтобы выявить влияние экологических факторов (например, температуры) и внутриклеточных механизмов (например, репарации ДНК, свойств репликации) на эволюцию мутационных спектров.

Исследование выполнено при финансовой поддержке Российского Научного Фонда (RSF № 21-75-20143).

1. Alexandrov, L. B. et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463), 415–421.
2. Ruis, C. et al. (2023). Mutational spectra are associated with bacterial niche. *Nature communications*, 14(1), 7091.
3. Efimenko, B. et al. (2024). NeMu: a comprehensive pipeline for accurate reconstruction of neutral mutation spectra from evolutionary data. *Nucleic acids research*, 52(W1), W108–W115.
4. Mendler, K. et al. (2019). AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic acids research*, 47(9), 4442–4448.
5. Parks, D. H. et al. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research*, 50(D1), D785–D794.
6. Mikhailova, A. G. et al. (2022). A mitochondria-specific mutational signature of aging: increased rate of A > G substitutions on the heavy strand. *Nucleic acids research*, 50(18), 10264–10277.
7. Aslam, S. et al. (2019). Aerobic prokaryotes do not have higher GC contents than anaerobic prokaryotes, but obligate aerobic prokaryotes have. *BMC evolutionary biology*, 19(1), 35.
8. Davin, A. A. et al. (2025). A geological timescale for bacterial evolution and oxygen adaptation. *Science*, 388(6742):eadp1853.

**Reconstructing N-Glycomes of Individual Glycoproteins from
Total Plasma N-Glycome: Synthetic GWAS of Immunoglobulin G
N-Glycosylation**

A. Soplenkova¹, D. E. Maslov¹, A. Timoshchuk¹, T. Spector²,
M. Georges³, S. Sharapov¹, G. Lauc⁴, Y. S. Aulchenko^{1,5}

¹*MSU Institute for Artificial Intelligence, Lomonosov Moscow State
University, Moscow, Russia, anna.mail10@mail.ru*

²*Department of Twin Research and Genetic Epidemiology, School of
Life Course Sciences, King's College London, St Thomas' Campus,
Lambeth Palace Road, London, SE1 7EH, UK*

³*Unit of Animal Genomics, WELBIO, GIGA-R and Faculty of
Veterinary Medicine, University of Liège, (B34) 1 Avenue de l'Hôpital,
Liège 4000, Belgium*

⁴*Genos Glycoscience Research Laboratory, Borongajska cesta 83H,
10000 Zagreb, Croatia University of Zagreb Faculty of Pharmacy and
Biochemistry, Zagreb, Croatia*

⁵*Institute of Cytology and Genetics, Novosibirsk, 630090, Russia,*

N-glycosylation is a common post- and co-translational modification of proteins that impacts their physical properties, as well as biological functions. The total blood plasma N-glycome is a mixture of the N-glycomes from individual circulating glycoproteins, with B-cell-derived immunoglobulin G (IgG) being one of the major contributors. While the biochemical network of human N-glycome biosynthesis is well characterized, the genetic regulation of these processes—particularly at the protein and tissue levels—remains poorly understood. By reconstructing individual protein N-glycan concentrations from total plasma glycan profiles, existing datasets can be repurposed for genetic analyses, such as genome-wide association studies (GWAS), effectively increasing the available sample sizes. The latter, in turn, enables a more

comprehensive investigation of the genetic regulation underlying protein-specific N-glycosylation

We propose a general workflow for reconstructing unmeasured, but biologically linked phenotypes from a set of correlated, readily available variables. As a case study, we focus on predicting the N-glycome of IgG using the total plasma N-glycome, using the Twins UK (N = 2063) and CEDAR (N = 188) for the training and performance evaluation respectively. To model the relationship between the two N-glycome profiles, we evaluated performance for a range of statistical and machine learning predictive methods, including linear regression models (with and without regularization), random forest and boosting. After identifying the most suitable model, we applied it to the GWAS summary statistics for total plasma N-glycosylation, which enabled us to infer the corresponding GWAS results for IgG N-glycosylation.

For predicting IgG N-glycan concentrations from the total blood plasma N-glycome, a simple regularization-free linear regression showed the best performance, as measured by R^2 . Comparison of the resultant GWAS summary statistics with the previously published GWAS for IgG N-glycosylation by Klarić et al. [1] demonstrated a strong consistency between the synthetic and real data, with substantial overlap between the genome-wide significant loci.

The proposed approach enables the reconstruction of most of the IgG N-glycosylation spectrum. This has allowed us to conduct a meta-analysis between synthetic and real IgG N-glycosylation GWAS, increasing the sample size from 11,000 to 17,000 individuals. We also showed the possibility of obtaining GWAS results for an individual glycoprotein from the results for the total blood plasma N-glycome. This work demonstrates a transferable strategy for deriving new molecular phenotypes from existing data, expanding the analytical potential of large-scale biobanks and omics datasets.

The work of AS, AT, DM, SSh, YSA was supported by the Research Program at the MSU Institute for Artificial Intelligence. The work of AS was partially supported by Non-commercial Foundation for the advancement of Science and Education “INTELLECT”.

1. Klarić L et al. (2020) Glycosylation of immunoglobulin G is regulated by a large network of genes pleiotropic with inflammatory diseases. *Sci Adv.* 6(8):eaax0301.

Филогенетические бенчмарки из аминокислотных и нуклеотидных выравниваний

С.А. Спирин¹, Д.Д. Латорцева²

¹*НИИ ФХБ имени А.Н. Белозерского, Московский государственный университет имени М.В.Ломоносова, Москва, 119992, Россия, sas@belozersky.msu.ru*

²*Факультет биоинженерии и биоинформатики, Московский государственный университет имени М.В.Ломоносова, Москва, 119992, Россия*

В статье [1] описан филогенетический бенчмарк, состоящий из выравниваний последовательностей ортологических групп эволюционных доменов белков из организмов различных таксонов, всего более 10000 выравниваний. Референсными деревьями служат деревья соответствующих видов. В той же работе было проведено сравнение нескольких филогенетических программ на основе предложенного бенчмарка. Сравнение показало, что на данном материале методы филогенетической реконструкции, основанные на принципе максимального правдоподобия (МП), в среднем проигрывают в точности дистанционным методам, таким, как Neighbor-joining. В то же время показано, что если ограничиться только выравниваниями с достаточно большим числом позиций, то МП начинает в среднем выигрывать, а для наборов ортологичных белков, референсное дерево которых имеет глубокое ветвление, то есть значительная часть видообразования происходила ближе к корню, разница между методом МП и дистанционными методами становится недостоверным. Метод максимальной экономии (парсимонии) показывает худшие среди всех методов результаты, особенно сильно отставая на «глубоко ветвящихся» наборах последовательностей.

В работе [2] был представлен бенчмарк из выравниваний двухдоменных белков и проведено исследование влияния фильтрации выравниваний на качество реконструкции. Выяснено, что отбрасывание междоменных линкеров однозначно

отрицательно влияет на качество филогенетической реконструкции. Что касается фильтрации программой NOISY [3], то она может быть полезна только для выравниваний эукариотических белков и только при определённых условиях на количество последовательностей в выравнивании и его длину.

К настоящему времени нами создан бенчмарк из нуклеотидных выравниваний, а именно из выравниваний последовательностей генов ортологичных белков млекопитающих. Значительная доля позиций белковых последовательностей консервативна в пределах класса млекопитающих, поэтому основной филогенетический сигнал в таких выравниваниях сосредоточен в синонимических позициях кодонов. Первые результаты сравнений программ показали, что на таких нуклеотидных выравниваниях МП превосходит дистанционные методы по точности, то есть имеет место ситуация, обратная к тому, что наблюдается для последовательностей ортологичных белков относительно удалённых организмов, использованных в работах [1,2].

Во всех случаях разброс качества реконструкции между различными выравниваниями существенно превосходит разброс качества между различными методами, применёнными к одному и тому же выравниванию. Нами получены определённые результаты по предсказанию качества филогенетической реконструкции по свойствам выравнивания и реконструированного дерева методами машинного обучения. Ещё одним направлением работы с бенчмарками является оценка предсказательной силы различных мер достоверности ветвей дерева, таких как бутстреп-поддержка.

1. S. Spirin, A. Sigorskikh, A. Efremov, D. Penzar, A. Karyagina (2024). PhyloBench: A benchmark for evaluating phylogenetic programs. *Molecular Biology and Evolution*, **41**(6):msae084.2.
2. А. И. Сигорских, Д. Д. Латорцева, А. С. Карягина, С. А. Спирин (2022). Насколько часто фильтрация колонок выравнивания улучшает реконструкцию филогении двухдоменных белков? *Биохимия*, **87**(12):2078–2088
3. A.W. Dress, C. Flamm, G. Fritzsche, S. Grünwald, M. Kruspe, S.J. Prohaska, P.F. Stadler (2008) Noisy: identification of problematic columns in multiple sequence alignments, *Algorithms Mol. Biol.*, **3**:7.

Interplay between cutoff radius and loop extrusion in interphase chromosomes: An analytical model suggests a monomeric mechanism for loop-extruding cohesin complexes

Dmitry Starkov¹, Kirill Polovnikov^{1,2}

¹*Skolkovo Institute of Science and Technology, Moscow, Russia*

²*Institute for Physics and Astronomy, University of Potsdam, Potsdam-Golm, Germany*

Interphase chromosomes exhibit a complex three-dimensional organization. At large genomic scales, chromosomes segregate into A/B compartments [1], while at the mesoscopic scale of hundreds of kilobases, topologically associating domains (TADs) emerge [2]. One of the principal mechanisms proposed to underlie TAD formation is the active extrusion of DNA loops by the cohesin complex [3,4]. Despite direct evidence for such loop extrusion [4], the underlying molecular details remain elusive. In particular, there is debate as to how cohesin performs loop extrusion: by a single complex or by two complexes acting together, that is, via a monomeric or a dimeric mechanism [5]. To address this issue, here we introduce an analytical framework that describes how the contact probability $P(s)$ depends on genomic distance s in a fractal globule model with frozen disorder of random loops, where the loop sizes and spacings follow exponential distributions of means λ and d , respectively [6]. In addition, the model takes into account the cutoff radius a in Hi-C experiments, a parameter determined by physical and chemical factors (such as crosslinking efficiency, restriction enzyme digestion, and proximity ligation) that effectively set the spatial range for detecting DNA contacts [7].

In the absence of loops, the fractal globule yields a power-law dependence for $P(s)$. Once loops are introduced, they result in characteristic modifications to this dependence. To identify these features, it is convenient to examine the logarithmic derivative, which represents the local slope of the $P(s)$ curve on log-log axes. Within our model, this logarithmic derivative behaves non-monotonically: it reaches a minimum (a “dip”) at some genomic distance s^* . The position of this minimum is governed by two parameters: the effective contact radius a

and the inverse loop density, $\lambda + d$.

Based on comparing the position of this minimum with experimental Hi-C data, we derive a quantitative estimate for the loop density in human chromatin (6 ± 1 loops per megabase). Then, using mass spectrometry measurements of DNA-bound cohesin [8], we show that the observed number of extruded loops is consistent with a *monomeric* mode of cohesin action, whereas a dimeric mechanism would lead to a substantial discrepancy in the required number of complexes. Hence, our approach provides a quantitative biophysical argument in favor of monomeric loop extrusion by cohesin in interphase.

1. Lieberman-Aiden E., van Berkum N.L., Williams L., Imakaev M., Ragoczy T., Telling A. et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome, *Science*, **326**: 289–293.
2. Rao S.S.P., Huntley M.H., Durand N.C., Stamenova E.K., Bochkov I.D., Robinson J.T. et al. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, *Cell*, **159**: 1665–1680.
3. Fudenberg G., Imakaev M., Lu C., Goloborodko A., Abdennur N., Mirny L.A. (2016) Formation of Chromosomal Domains by Loop Extrusion, *Cell Reports*, **15**: 2038–2049.
4. Davidson I.F., Bauer B., Goetz D., Tang W., Wutz G., Peters J.-M. (2019) DNA loop extrusion by human cohesin, *Science*, **366**: 1338–1345.
5. Barth R., Davidson I.F., van der Torre J., Taschner M., Gruber S., Peters J.-M., Dekker C. (2025) SMC motor proteins extrude DNA asymmetrically and can switch directions, *Cell*, **188**(3) 749–763.e21.
6. Polovnikov K.E., Brandão H.B., Belan S., Slavov B., Imakaev M., Mirny L.A. (2023) Crumpled polymer with loops recapitulates key features of chromosome organization, *Phys. Rev. X*, **13**(4) 041029.
7. Belaghzal H., Dekker J., Gibcus J.H. (2017) Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation, *Methods*, **123**: 56–65.
8. Holzmamm J., Politi A.Z., Nagasaka K., Hantsche-Grininger M., Walther N., Koch B. et al. (2019) Absolute quantification of cohesin, CTCF and their regulators in human cells, *eLife*, **8**: e46269.

Идентификация энхансерных РНК, связывающихся с андрогеновым рецептором и связанных с прогрессированием рака предстательной железы

Р.И. Султанов, А.С. Суздаленко, П.В. Шнайдер, О.А. Зубкова,
Г.П. Арапиди

Москва, ул. Малая Пироговская 1с3, rhenium112@gmail.com

Энхансерные РНК (эРНК) - это длинные некодирующие РНК, экспрессирующиеся с транскрипционно активных энхансеров. В большинстве своем эРНК - это транскрипционный шум, но были обнаружены некоторые эРНК, регулирующие транскрипцию на разных уровнях. Известно, что нарушения транскрипционной программы андрогенового рецептора (AR) при раке предстательной железы (РПЖ) приводит к изменению хроматинового ландшафта, и как следствие к экспрессии эРНК, которые непосредственно участвуют в онкогенезе РПЖ. Однако до сих пор не известно, может ли AR связываться с эРНК и какую роль в регуляции транскрипции этот комплекс может играть. В этой работе мы решили изучить этот вопрос.

Проанализировав открытые данные РНК-секвенирования, иммунопреципитации РНК за AR (AR RIP-seq), профилирования РНК, связанных с рибосомами (RIBO-seq) для клеточной линии РПЖ LNCaP, а также данные Cancer Cell Line Encyclopedia (CCLE) и Genotype-Tissue Expression (GTEx) мы обнаружили, что всего одна стабильно экспрессирующаяся эРНК связывается с AR (назовем ее ARA-elncRNA1). Мы подтвердили связывание этой РНК с AR при помощи RIP-qPCR и эксперимента по электрофоретической задержке в геле. В данных The Cancer Genome Atlas мы обнаружили, что экспрессия ARA-elncRNA1 ассоциирована с лучшим прогнозом для пациентов. Также мы показали, что нокдаун ARA-elncRNA1 повышает пролиферацию клеток и увеличивает размер единичных колоний в клоногенном тесте.

Для AR было показано, что lncRNA могут направлять его на определенные геномные локусы. Проанализировав открытые данные по иммунопреципитации хроматина на AR для 52 пациентов с РПЖ и соответствующие данные секвенирования РНК, мы обнаружили 655 локусов, представленность AR на которых ассоциирована с экспрессией ARA-elncRNA1 (ARBS). Эти локусы лежат в промоторах AR-зависимых генов и обогащены сайтами связывания РНК-связывающих транскрипционных факторов (YY1, CTCF, KLF4, EZH2), но лишены ДНК-мотива связывания AR. Методом ChIP-qPCR мы показали, что как нокдаун, так и сверхэкспрессия ARA-elncRNA1 приводят к изменению представленности AR на промоторах некоторых выбранных нами генов, а также влияют на экспрессию самих этих генов. Кроме того, методом иммунопреципитации хроматина за РНК мы показали, что ARA-elncRNA1 связывается с хроматином в тех же локусах, что и AR. Анализ открытых данных ChIA-PET для клеточных линий РПЖ показал, что ARBS соединены хроматиновой петлей с сильным ChIP-seq пиком AR, при этом оба конца этой петли закорены в ChIP-seq пиках YY1. При помощи метода 3C (chromatin conformation capture) мы подтвердили, что нокдаун ARA-elncRNA1 снижает формирование этих петель. Также было обнаружено, что ARA-elncRNA1 не только связывается с AR, но регулирует его представленность как на уровне РНК, так и на уровне белка.

Подводя итоги, в этой работе мы обнаружили новую энхансерную РНК ARA-elncRNA1, которая связывается с AR и способна регулировать его представленность в клетке и направлять на промоторы определенных AR-зависимых генов. Мы показали, что снижение экспрессии ARA-elncRNA1 имеет онкогенный эффект и приводит к повышению пролиферации клеток, а в данных TCGA сниженная экспрессия этой РНК ассоциирована с высоким баллом Глиссона. Таким образом в этом исследовании на примере AR мы подтвердили, что эРНК может быть важным партнером транскрипционных факторов в регуляции экспрессии генов.

DrugForm-DTA: модель аффинности

А.С. Тащилова¹, И.С. Хохлов¹, Н.А. Бугаев-Макаровский¹, О.В. Глушкова¹, В.С. Юдин¹, А.А. Кескинов¹, С.М. Юдин¹, Д.В. Светличный¹, В.И. Скворцова²

¹Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства (ФГБУ «ЦСП» ФМБА России); 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1.
atashchilova@cspfmba.ru

²Федеральное медико-биологическое агентство (ФМБА России); 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.

Компьютерные методы все чаще применяются в области разработки лекарств, в том числе подходы с использованием машинного обучения и нейронных сетей. Одной из ключевых задач в этой сфере является определение аффинности малой молекулы к целевому белку (задача DTA). На сегодняшний день наиболее успешные методы решения данной задачи используют нейросетевые подходы с использованием обработки естественного языка (NLP). Целью данной работы было создание новой эффективной DTA-модели, способной работать на реальных данных с высокой точностью. Разработанная модель рассчитывает константы аффинности K_i и IC_{50} для произвольных белка и лиганда, представленных соответственно первичной аминокислотной последовательностью в формате однобуквенного кода, и SMILES-строкой.

Модель DrugForm-DTA использует достаточно простую архитектуру нейронной сети на основе Transformer [1], сочетая в себе несколько известных подходов для кодирования белковой последовательности и лиганда (ESM [2] и Chemformer [3]) посредством применения 1D-свертки для выравнивания размерностей численных представлений белка и лиганда. Модель DrugForm-DTA была протестирована на стандартных бенчмарках - Davis [4] и KIBA [5], продемонстрировав лучший в сравнении

существующими методами результат на KIBA ($MSE=0.117$) и второй на Davis ($MSE=0.152$) по состоянию на 2024 год [6].

Одним из крупнейших общедоступных источников экспериментальных значений констант аффинности является база данных BindingDB [7], которая содержит более 2,8 млн записей. BindingDB не является набором данных, готовым для обучения моделей машинного обучения, поэтому разработана серия эвристических вычислительных коррекций с целью трансформации и фильтрации данных. В последующем это позволило также сформировать один из наиболее крупных публичных наборов данных, применимых для обучения нейросетевых моделей.

На подготовленном датасете была обучена модель DrugForm-DTA методом кросс-валидации. Тестовая часть датасета была выделена по принципу cold-target и drug-scaffold. Обученная модель показала высокую эффективность на тестовой части датасета, сопоставимую с экспериментальными измерениями. В частности, точность по RMSE для тестового множества данных из BindingDB составила 0,66 (pIC50) и 0,67 (pKi).

Было также проведено дополнительное исследование на тестовой подвыборке с помощью методов молекулярного моделирования, такими как молекулярный докинг, молекулярная механика и полуэмпирические методы квантовой химии. Модель DrugForm-DTA показала существенно более высокий уровень корреляции по Спирмену ($CDTA=0.76$) с экспериментальными значениями, чем методы молекулярного моделирования ($CMod=0.20$). Кроме того, модель не требует знания пространственной структуры белка, а только первичной аминокислотной последовательности.

Таким образом, DrugForm-DTA модель показала высокую эффективность и имеет потенциал для разработки новых терапевтических молекул для широкого спектра заболеваний.

1. A. Vaswani et al. (2017) Attention Is All You Need, Advances in neural information processing systems, 30.
2. Z. Lin et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model, Science, 379: 1123-1130.
3. R. Irwin et al. (2022) Chemformer: a pre-trained transformer

- for computational chemistry, Machine Learning: Science and Technology, 3: 015022.
4. M.I. Davis et al. (2011) Comprehensive analysis of kinase inhibitor selectivity, Nature Biotechnology, 29: 1046-1051.
 5. J. Tang et al. (2014) Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis, Journal of Chemical Information and Modeling, 54: 735-743.
 6. X. Xiao et al. (2024) HGTDP-DTA: Hybrid Graph-Transformer with Dynamic Prompt for Drug-Target Binding Affinity Prediction, arXiv preprint, arXiv:2406.17697.
 7. M. Gilson et al. (2015) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic acids research, 44: D1045-D1053.

Enhancing the Reverse-Transcriptase Function in Taq Polymerase via AI-Driven Multiparametric Rational Design

Yulia E. Tomilova¹, Nikolay E. Russkikh², Igor M. Yi²,
 Elizaveta V. Shaburova³, Viktor N. Tomilov⁴, Galina B. Pyrinova¹,
 Svetlana O. Brezhneva¹, Olga S. Tikhonyuk¹, Nadezhda S. Gololobova¹,
 Dmitriy V. Popichenko¹, Maxim O. Arkhipov¹, Leonid O. Bryzgalov¹,
 Evgeniy V. Brenner¹, Anastasia A. Artyukh¹, Dmitry N. Shtokalo^{2,3,5},
 Denis V. Antonets^{3*}, Mikhail K. Ivanov^{1,6}

¹AO Vector-Best, Novosibirsk; ²AcademGene LLC, Novosibirsk; ³MSU
 Institute for Artificial Intelligence, Lomonosov Moscow State
 University, Moscow; ⁴SibEnzyme Ltd., Novosibirsk; ⁵Institute of
 Informatics Systems SB RAS, Novosibirsk; ⁶Institute of Molecular and
 Cellular Biology SB RAS, Novosibirsk, Russia.

*Correspondence: antonec@yandex.ru

Introduction. Modification of natural enzymes to introduce new properties and enhance existing ones is a central challenge in bioengineering. *Thermus aquaticus* DNA polymerase I (Taq pol) is a widely used thermostable enzyme in molecular biology, but its intrinsic reverse transcriptase (RTase) activity is very low. High-temperature

RTases are attractive for single-enzyme RT-qPCR workflows. Artificial intelligence (AI) methods based on protein language models (PLMs) offer a powerful route to navigate the immense mutational landscape of enzymes when multi-trait optimization is required. For example, it was demonstrated that even such a simple PLM as UniRep simulated a fitness landscape accurately enough for engineering applications with using only 24 functionally characterized proteins bearing amino acid substitutions [1]. Here, we report a multiparametric, AI-guided strategy for enhancing the RTase activity of Taq pol without sacrificing other critical properties such as fidelity, thermal stability, 5'-3' exonuclease activity, and the ability to incorporate dUTP and tolerance to LNA-containing substrates [2].

Methods. We first selected 18 single mutants targeting six important sites (E507, S515, D578, H639, F667, M747) implicated in primer, template or dNTP binding [3]. Each variant, together with the wild type (WT), was characterized in a ten-parameter experimental panel comprising: (i) RTase activity on four RNA templates; (ii) DNA-dependent polymerase rate and dissociation constant; (iii) PCR efficiency and probe-cleavage performance; (iv) error rate; (v) dTTP→dUTP substitution tolerance; (vi) extension speed on three LNA-modified hairpins; (vii) antibody and aptamer hot-start compatibility. Sequence embeddings were obtained with a ProtT5-XL encoder fine-tuned on 91K Taq pol homologues. Ridge regression models were trained on the 19-sequence dataset to predict each parameter; uncertainty-aware Gaussian process (GP) models replaced Ridge after a second data collection round. The GP ensemble was used for *in silico* screening of 18.3 M variants within a Hamming radius ≤ 3 from the WT. Two successive cycles prioritized 29 additional single, double and triple mutants for synthesis and testing, expanding the empirical dataset to 47 variants. Top candidates were evaluated in single-tube, 60 °C TaqMan RT-qPCR targeting synthetic HPIV2 and hMpV RNAs.

Results. Out of 47 tested mutants, 18 showed markedly improved RTase activity, with some achieving up to two orders of magnitude increase compared to wild-type Taq pol. Enhanced RTase activity was associated with lower K_d values, modestly reduced fidelity (median 1 error/5 200 nt vs 1/6 000 nt for WT), and greater tolerance to LNA modifications. Twelve variants met multi-trait acceptance

thresholds ($k_{\text{eff}} \geq 0.2 \times \text{WT}$; PCR efficiency ≥ 1.85 ; dT/dU ratio ≤ 1.5 ; error rate $\leq 1/3\,000$) and all supported hot-start control. In single-enzyme RT-qPCR they achieved C_q values comparable to a commercial p66 RT/Taq mix, with clean fluorescence rising phases and probe-cleavage amplitudes. Notably, double mutant E507K-D732N and triple mutant E507Q-D578S-I614M showed $\Delta C_q \leq 0.5$ relative to the two-enzyme control, while maintaining probe digestion and inhibitor tolerance.

Conclusions. We demonstrate that PLM-based embeddings, combined with Gaussian process regression, enable efficient multiparametric optimization of enzyme function with a small experimental dataset. Our AI-guided approach successfully identified Taq pol mutants combining high RTase activity and desirable polymerase traits, providing valuable candidates for single-enzyme RT-qPCR diagnostic and research workflows. This methodology is transferable to other multifunctional enzyme where exhaustive combinatorial mutational scanning is impractical.

This research was partially supported by the MSU Institute for Artificial Intelligence. Wet-lab reagents were provided by AO Vector-Best. We thank the Federal Research Center for Information and Computational Technologies SB RAS for high-performance computing resources.

1. Biswas S. et al. (2021) Low-N protein engineering with data-efficient deep learning, *Nat. Methods*, 18:389–396.
2. Tomilova Y.E. et al. (2024) Enhancing the reverse-transcriptase function in Taq polymerase via AI-driven multiparametric rational design, *Front. Bioeng. Biotechnol.*, 12:1495267.
3. Raghunathan G., Marx A. (2019) Engineering DNA polymerases for improved reverse transcription, *Trends Biotechnol.*, 37:1027–1041.

Анализ транскриптомов свободноживущих *E. coli* K-12 MG1655 и клеток в составе биопленок

М.Н. Тутукина^{1,2,4}, Т.А. Бессонова^{2,3,4}, А.Д. Казнадзей^{2,4},
А.И. Дахновец¹, М.С. Гельфанд¹

¹*Сколковский институт науки и технологий, Большой бульвар
30с1, Москва, Россия*

²*Институт проблем передачи информации РАН, Большой
Каретный пер. 19с1, Москва, Россия* masha306@gmail.com

³*Институт биофизики клетки РАН - ФИЦ ПНЦБИ РАН, Пушкино,
Россия*

⁴*Институт общей генетики РАН, ул. Губкина 3, Москва, Россия*

Формирование бактериями биопленок может приводить к развитию многих хронических заболеваний, плохо поддающихся терапии. Например, уропатогенная *E. coli* является причиной развития более 80% всех хронических инфекций мочеполового тракта. Задачей данного исследования является поиск метаболических регуляторов, вовлеченных в переключение образа жизни бактерий от свободноживущего к биопленкам.

На первом этапе мы сравнили транскриптомы свободноживущих клеток *E. coli* K-12 MG1655 и клеток, находящихся в составе биопленок. Эксперимент был проведен на клетках, растущих в присутствии глюкозы в качестве основного источника углерода или же галактуроната - интермедиата пути Эшвелла, необходимого для успешной колонизации бактериями организма хозяина [1]. Спектр генов, экспрессия которых менялась при переходе к биопленкам, практически не зависел от источника углерода.

В клетках, находящихся в составе биопленок, были активированы гены, кодирующие шапероны IbpA и IbpB, отвечающие за адаптацию к стрессу, а также гены ферментов углеводного метаболизма. Отдельно стоит отметить активацию, даже при росте на глюкозе, генов, кодирующих ферменты и

транспортёры метаболизма гексуронатов – гена тагатуронат редуктазы *ixaB* и гексуронатного транспортера *exuT*. Это подтверждает важную роль метаболизма гексуронатов в процессе образования биопленок, что было показано нами и ранее [2], но также это может говорить о дефиците более выгодных питательных субстратов. Среди генов с повышенной экспрессией в клетках биоплёнок были также гены, кодирующие транскрипционные регуляторы (*sutR*, *ycdR*), регуляторы образования биоплёнок (*bssR*), а также гены малых РНК – сигнальной 4.5S РНК (*ffs*) и малой регуляторной 6S РНК (*ssrS*). Возможно, суперпродукция *ssrS* связана с тем, что большинство клеток находились на стационарной фазе роста, где *SsrS* модулирует активность РНК-полимеразы [3]. Тем не менее, с учетом сильнейшей активации в биопленках *SsrS* была использована в качестве одного из четырех генов для оценки экспрессии при выборе отдельных клонов для дальнейшего секвенирования транскриптомов и поиска факторов, приводящих к формированию биопленок и дисперсии данных.

С помощью qRT-PCR для отдельных клонов в планшете было показано, что в отсутствии *UxuR*, основного регулятора пути Эшвелла, наблюдалась сильная дисперсия в эффективности формирования биопленок и экспрессии маркерных генов. При удалении гена другого регулятора, *YjjM*, наблюдалось сильное ингибирование экспрессии малой РНК *CsrC* – это показано впервые, и, наряду с активацией *flhA*, указывает на роль *YjjM* в регуляции подвижности бактерий и формирования ими биопленок. При удалении гена *cAMP-CRP*, который ранее был предсказан как один из регуляторов процесса формирования биопленок, наблюдалась небольшая активация *csgD*, что, в целом, согласуется с полученными нами ранее данными [2], но на малые РНК он не влиял, и, в целом, эффекты были не такими яркими, как в случае делеции *yjjM*. Это было полностью подтверждено и физиологическими тестами.

Влияние *YjjM* на гены подвижности и биопленок было полностью подтверждено с помощью секвенирования транскриптомов. в транскриптомном анализе. При его удалении были существенно активированы гены подвижности – *flgB*, кодирующий один из четырех белков, составляющих основу тела флагеллы; *flhB*, кодирующий белок биосинтеза флагелл, и ген

флагеллярного белка FliL. Были также активированы гены основного белка хемотаксиса CheA и сигма фактора подвижности FliA, хотя они не вошли в 50 наиболее активируемых генов. Напротив, гены, отвечающие за формирование биопленок, фимбрий и пилей curly (*fimA*, *csrA*, *bhsA*, *bssS*, *bssR*, *ycgZ*), были ингибированы. Были также сильно ингибированы гены малых РНК CsrB-CsrC, что подтверждает роль YjjM в формировании биопленок, а присутствие SsrS в его регулоне свидетельствует в пользу участия SsrS в регуляции этого процесса.

В регулон YjjM также вошли ген белка Hfq и ген глобального регулятора cAMP-CRP. Это говорит о том, что YjjM может влиять и на малые РНК, в том числе за счет Hfq, а показанное ранее влияние CRP на формирование биопленок может быть опосредованным.

Работа поддержана РФФ 24-24-00435.

1. A.G. Jimenez et al (2020) Diet-derived galacturonic acid regulates virulence and intestinal colonization in enterohaemorrhagic *Escherichia coli* and *Citrobacter rodentium*, *Nat Microbiol*, **5**(2): 368-378.
2. T.A. Bessonova et al (2024) Connection between Hexuronate Metabolism and the Ability of *Escherichia coli* to Adhesion and Biofilm Formation, *Microbiology*, **93**: 497–501.
3. K.M. Wassarman and G. Storz (2000) 6S RNA regulates *E. coli* RNA polymerase activity, *Cell*, **101**(6): 613-623.

Судьба дублированных митохондриальных генов при двуродительском наследовании у двустворчатого моллюска *Yoldia hyperborea* (Bivalve, Protobranchia)

Д.А. Фёдоров¹, А.В. Безменова², М.В. Ежова³, Г.Д. Колбасова³,
Г.А. Базыкин, Т.В. Неретина³, Д.А. Кнорре⁴

¹Центр наук о жизни, Сколковский Институт Науки и Технологий, Москва, Россия, fedorov.kuleshov@gmail.com

²Институт проблем передачи информации им. А.А. Харкевича
РАН, Москва, Россия

³Беломорская Биологическая Станция им. Н.А. Перцова,
Московский Государственный Университет, Москва, Россия,

⁴Научно-исследовательский институт физико-химической
биологии им. А.Н.Белозерского, Московский Государственный
Университет, Москва, Россия, dmitry.knorre@gmail.com

Двойное однородительское наследование (DUI) митохондриальной ДНК (мтДНК) — это способ передачи мтДНК, при котором мужской митохондриальный геном наследуется преимущественно самцами, а женский — самками. Такой способ наследования уникален для двустворчатых моллюсков и был выявлен во всех основных крупных кладах, включая первичножаберных (Protobranchia) — наиболее базальную группу среди всех двустворчатых моллюсков¹. Хотя эта система развивалась у двустворчатых моллюсков независимо по крайней мере десятки раз, механизм разделения вариантов митохондриальных геномов плохо изучен². В работе изучаются процессы, происходящие с на ранней стадии расхождения мужских и женских митохондриальных геномов у протобранхии *Yoldia hyperborea*, которая характеризуется одним из самых низких известных уровней дивергенции между геномами, что предполагает недавнее происхождение этой системы. Несмотря на высокую степень сходства женских и мужских геномов, мы обнаружили, что они претерпели быструю дивергенцию как на уровне нуклеотидных последовательностей, так и на структурном уровне, включая возникновение новых генов в мужском и женском геномах. Кроме того, женский митогеном содержит дублированную кассету, включающую ген цитохрома b. Эти результаты демонстрируют быструю адаптацию женских и мужских митохондриальных геномов вскоре после их разделения при двойном однородительском наследовании подобной той, которая наблюдается в эволюции половых хромосом.

1. S. Breton et al. (2007) The unusual system of doubly uniparental

inheritance of mtDNA: isn't one enough? *Trends Genet.* **23**:465–474.
2. N. Sano et al. (2007) Quantitation of the male and female types of mitochondrial DNA in a blue mussel, *Mytilus galloprovincialis*, using real-time polymerase chain reaction assay. *Dev. Growth Differ.* **49**:67–72.

Роль вариантов в старт-кодонах и последовательностях Козак в развитии наследственных заболеваний человека

А.Ю. Филатова, Я.В. Шинкевич, М.Ю. Скоблов

ФГБНУ «Медико-генетический научный центр им. Н.П. Бочкова»,
115522, г. Москва, ул. Москворечье, д. 1
maacc@yandex.ru
mskoblov@gmail.com

В отличие от вариантов, локализованных в белок-кодирующих регионах генома, варианты в некодирующих, в частности регуляторных, областях в настоящее время гораздо реже выявляются как причина развития наследственных заболеваний. Объектами нашего исследования являются старт-кодона и Козак-последовательности — элементы, играющие центральную роль в инициации трансляции и, следовательно, в определении уровня синтезированного белка.

При анализе данных NGS-секвенирования ДНК пациентов с наследственными заболеваниями выявленные варианты, нарушающие старт-кодона соответствующих генов, могут рассматриваться как потенциально каузативные. В то же время патогенные варианты в последовательности Козак на сегодняшний день не выявляются. Целью настоящего исследования является описание роли вариантов в старт-кодонах и последовательности Козак как возможных причин развития наследственных заболеваний.

Для анализа были использованы гены, ассоциированные с развитием наследственных заболеваний из базы данных OMIM. Для них мы оценили встречаемость вариантов в старт-кодонах и последовательностях Козак (-7 - +6 позиции относительно старт-кодона, без учета самого старт-кодона) как в популяционной базе

данных gnomAD v4.1, так и в базах данных вариантов, выявленные у пациентов с наследственными заболеваниями (ClinVar и HGMD). Варианты из баз данных были предварительно отфильтрованы по качеству и аннотировались по MANE транскриптам.

Хотя наличие варианта в старт-кодоне подразумевает отсутствие трансляции исследуемого белка, в популяционной базе данных gnomAD было выявлено 4 556 вариантов в старт-кодонах генов, ответственных за развитие наследственных заболеваний. Это может быть объяснено в первую очередь аутосомно-рецессивным типом наследования многих заболеваний, при которых здоровый человек может быть носителем патогенного варианта. А во вторую, молекулярным механизмом развития заболевания (потеря или приобретение функции гена).

В ходе дальнейшего анализа мы выявили, что в генах с исключительно аутосомно-доминантным типом наследования и высокой чувствительностью к потере функции ($pLI \geq 0,9$) в популяции также обнаруживаются 258 вариантов, нарушающих старт-кодон. Для объяснения данного феномена мы предложили три гипотезы: (1) наличие альтернативного старт-кодона вблизи основного; (2) неполная пенетрантность, описанная для некоторых генов; (3) мягкость ассоциированного с дисфункцией гена клинического фенотипа.

Используя эти гипотезы, мы смогли объяснить наличие более 70% вариантов в старт-кодонах генов, ассоциированных с аутосомно-доминантными наследственными заболеваниями. Для группы вариантов с частотой более 0,001% предложенные объяснения охватывают до 90% случаев.

Далее мы применили предложенный подход для анализа вариантов, выявленных у пациентов с наследственными заболеваниями и представленных в базах данных HGMD и ClinVar. Это позволило идентифицировать вероятно ложные каузативные варианты, которые, несмотря на локализацию в старт-кодонах и наличие в указанных базах данных, вероятно не приводят к развитию описанной в базе данных патологии.

Следующим этапом нашего исследования стал анализ вариантов, локализованных в последовательности Козак (без затрагивания старт-кодона). Для оценки их влияния на эффективность трансляции мы использовали экспериментальные

данные, охватывающие все возможные варианты последовательности Козак [1].

В результате, в базе данных gnomAD v4.1 было выявлено 15 530 вариантов, расположенных в пределах последовательностей Козак. Мы проанализировали взаимосвязь между изменением эффективности трансляции, обусловленными этими вариантами, и их частотой встречаемости в популяции, а также типом нуклеотидных замен (включая различие между транзициями и трансверсиями и конкретные виды замен). Кроме того, нами была выявлена тенденция действия отбора на определённые позиции последовательности Козак, отражающая их функциональную значимость.

Последовательность Козак состоит из двух участков: первый располагается перед старт-кодоном находится в не кодирующей области гена, а второй – сразу после старт-кодона и уже входит в кодирующую последовательность. Варианты, находящиеся в кодирующей части последовательности Козак, гипотетически могут оказывать патогенное действие как за счёт изменения эффективности трансляции, так и за счёт изменения аминокислотной последовательности белка. Напротив, варианты, локализованные в не кодирующей части, могут влиять исключительно на регуляцию трансляции, не затрагивая аминокислотный состав белка.

В ходе нашего анализа было выявлено значительное различие в эффекте вариантов в кодирующей и не кодирующей частях последовательности Козак на эффективность трансляции. При этом связь между типом наследования и степенью изменения эффективности трансляции установлена не была.

При анализе вариантов, выявленных у пациентов с наследственными заболеваниями (из баз данных ClinVar и HGMD), мы выявили соответственно 2577 и 249 вариантов, расположенных в пределах последовательности Козак. Из них особое внимание было уделено вариантам, локализованным в не кодирующей части и аннотированным как патогенные (категории P/LP в ClinVar и DM в HGMD). Анализ этих вариантов показал, что ни один из них не демонстрирует значимого влияния на изменение эффективности трансляции, а значит их патогенность связана с какими-то другими механизмами. При этом варианты, расположенные в кодирующей

части последовательности Козак имеют гораздо более значимое влияние на эффективность трансляции. Это может свидетельствовать о том, что изменения в последовательности Козак, влияющие исключительно на инициацию трансляции, вероятно, недостаточны для развития грубых нарушений экспрессии гена, что необходимо для развития наследственных заболеваний. Данное наблюдение определило дальнейшее направление нашей работы — поиск гипоморфных аллелей, частично нарушающих функцию гена и способных тем самым вносить вклад в развитие патологии.

1. Noderer WL et al. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol Syst Biol*, 10(8):748

Genome annotation and gene expression prediction using DNA Language Models

V. Fishman^{1,2,3}, A. Shmelev^{1,4}, A. Shadskii^{1,2}, A. Aspidova¹, Y. Kuratov¹, M. Burtsev⁵, O. Kardymon¹

¹*AIRI*

²*Sirius University*

³*Institute of Cytology and Genetics SB RAS, minja-f@ya.ru*

⁴*HSE*

⁵*London Institute of Mathematical Sciences*

Transcriptional regulation, DNA replication, repair, and other core molecular processes rely on the interaction between DNA-binding proteins and specific genomic sequences. The scale and complexity of these regulatory networks present a major challenge for decoding the underlying genomic grammar. However, the abundance of high-throughput data enables the application of AI-driven methods to identify sequence determinants of functional elements across genomes.

In our recent work [1], we present GENA-LMs, a family of open-source DNA language models for genomic analysis. Here, we thoroughly

investigate their performance in two fundamental genomics tasks: gene annotation and prediction of cell type-specific gene expression.

For gene annotation, we formulate the task as the identification and classification of genomic regions into exons, introns, untranslated regions, and protein-coding sequences. We systematically explore a broad range of model architectures, training datasets, and regimes to optimize performance. Our experiments highlight the importance of several factors:

- increased input context length for capturing long-range dependencies,
- training on multispecies genomic datasets,
- hybrid architectures that integrate transformer-based encoder with U-Net or decoder-style components for base-level resolution,
- reverse-complement and other data augmentation.

Integrating these findings, we introduce GENATATOR, a genome annotation tool that effectively segments both protein-coding and non-coding (lncRNA) genes. Extensive benchmarking demonstrates that GENATATOR outperforms existing methods in the total number of correctly segmented genes and in the annotation of lncRNA genes, while ranking second in performance for protein-coding gene annotation.

The second task addressed is gene expression prediction. We propose a hybrid architecture that combines GENA-LMs with general-purpose large language models (LLMs), enabling joint processing of various epigenetic assays, transcriptomic data, and structured textual descriptions of experimental conditions. While this architecture achieves high accuracy in predicting overall gene expression profiles across diverse species and tissues, it, like previous approaches, struggles with accurately predicting expression levels of a small subset of highly cell type-specific genes.

Our results demonstrate the versatility of DNA language models in both structural and functional genomic tasks, and provide a foundation for further integration of genomic and multi-modal data in deep learning frameworks.

Acknowledgments: development of datasets and validation of GENATATOR was partially supported by the grant of the state program

of the «Sirius» Federal Territory «Scientific and technological development of the «Sirius» Federal Territory» (Agreement №26-03, 27/09/2024).

1. V. Fishman, Y. Kuratov, A. Shmelev, M. Petrov, D. Penzar, D. Shepelin, N. Chekanov, O. Kardymon, M. Burtsev (2025) GENA-LM: a family of open-source foundational DNA language models for long sequences, *Nucleic Acids Research*, **53**:2.

Исследование плейотропных генов моногенных заболеваний человека

Е.А.Фоменко¹, В.Е.Раменский^{1,2,3}

¹*Факультет Биоинженерии и биоинформатики, Московский государственный университет им. М.В. Ломоносова. 19234, г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73, lizzafomenko@gmail.com*

²*ФГБУ «Национальный медицинский исследовательский центр Терапии и профилактической медицины» Минздрава России. 101990 г. Москва, Петроверигский пер., д. 10, стр.3*

³*Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем, Московский государственный университет им. М.В. Ломоносова. 119192 г. Москва, Ломоносовский просп., 27 корпус 1*

Неполная пенетрантность и изменчивая экспрессивность болезнетворных вариантов, ассоциированных с моногенными заболеваниями, весьма распространены [1]. На фенотипические проявления варианта влияют различные факторы, в частности, полигенный фон и внешние факторы риска [2]. В случае плейотропных генов, ассоциированных более чем с одним моногенным заболеванием [3], интерес представляют как свойства самих генов, так и факторы, определяющие «выбор» того или иного

заболевания болезнетворными вариантами. Целью данной работы было исследование свойств плеiotропных генов, ассоциированных более чем с одним моногенным заболеванием, и болезнетворных вариантов в этих генах. Исследование болезнетворных вариантов генома и содержащих их генов позволяет лучше понять механизмы патогенеза, повысить эффективность интерпретации результатов клинического секвенирования.

Использовалась выборка из 994 плеiotропных болезнетворных генов человека по данным базы OMIM. Был разработан метод для оценки сходства пар моногенных заболеваний, не зависящий от свойств генов и учитывающий свойства фенотипов HPO (Human Phenotype Ontology). Для описания свойств генов использовались базы данных GeneOntology, InterPro, GTEh и прочие; болезнетворные варианты и число их подач для заболевания извлекались из XML-версии базы ClinVar [4].

Из 2,931 пар моногенных заболеваний, ассоциированных с 994 генами человека, одинаковый тип наследования наблюдается в 1,491 паре заболеваний. Из них только 42 пары в 33 генах существенно различаются по своим фенотипам. При этом остается открытым вопрос о возможных неочевидных функциональных связях между этими фенотипами. Как можно ожидать [3,5], для плеiotропных генов характерно, в частности, большее число функциональных доменов, GO-терминов и белок-белковых взаимодействий, чем для прочих болезнетворных генов. Однако функциональное разнообразие генов напрямую не влияет на разнообразие заболеваний.

Значимая ассоциация типа болезнетворных вариантов (укорачивающие белок или несинонимичные) наблюдается примерно в 9% всех рассмотренных пар, что согласуется с полученными ранее результатами [6]. При этом значимые ассоциации с определенной болезнью в плеiotропном гене для индивидуальных вариантов или их предпочтительная локализация в определенных функциональных доменах гена встречаются крайне редко (1-3% случаев) и не характерны для случая существенно разных болезней.

Таким образом, фенотипические проявления и «выбор» заболевания болезнетворными вариантами в плеiotропных генах в

большинстве случаев определяется совокупностью действия различных факторов.

- [1] R.Kingdom R, C.F.Wright (2022) Incomplete Penetrance and Variable Expressivity: From Clinical Studies to Population Cohorts. Front Genet 13:920390.
- [2] J.K.Goodrich et al. (2021) Determinants of penetrance and variable expressivity in monogenic metabolic conditions across 77,184 exomes. Nat Commun 12:3505.
- [3] S. Ittisoponpisan et al. (2017) Landscape of Pleiotropic Proteins Causing Human Disease: Structural and System Biology Insights. Human Mutation 38:289–296.
- [4] M.J.Landrum et al. (2016) ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44:D862-868.
- [5] Y.A.Barbitoff et al. (2024) Functional determinants and evolutionary consequences of pleiotropy in complex and Mendelian traits. bioRxiv 2024.10.28.620594
- [6] W.Lu et al. (2024) Effect heterogeneity reveals complex pleiotropic effects of rare coding variants. bioRxiv 2024.10.01.614806

**От биоинформатического предсказания к валидации
нарушения трансляции: нарушение структуры uORF как
причина наследственной патологии человека**

Д.В. Харитонов, М.Ю. Скоблов, А.Ю. Филатова

*Медико-генетический научный центр им. Н.П. Бочкова,
Москворечье ул., 1, haritonovdaniil16@gmail.com*

Не смотря на то, что внедрение технологий NGS в клиническую практику позволило увеличить эффективность молекулярно-генетической диагностики пациентов с

наследственными заболеваниями [1-3], значительная часть подобных случаев остается нераскрытой. Это связано, во-первых, с тем, что интерпретация генетических вариантов в настоящее время всё ещё фокусируется на кодирующих частях генов, тогда как выявление вариантов в некодирующей части генома затруднено из-за недостаточной характеристики механизмов молекулярного патогенеза. Во-вторых, в случае обнаружения варианта, вероятно приводящего к заболеванию, необходимо проводить функциональные исследования, так как даже при наличии специализированных биоинформатических подходов [4] и рекомендаций [5] их применимость для точной и комплексной клинической интерпретации некодирующих вариантов остается под вопросом.

Часть вариантов с неопределенным клиническим значением расположены в некодирующих частях генома, в частности в 5'нетранслируемой области (5'UTR). Эта часть молекулы мРНК отвечает за регуляцию процессов трансляции посредством различных регуляторных элементов, в частности малых открытых рамок считывания (upstream open reading frame, uORF) [6]. И хотя более половины генов человека содержат uORF [4,7], связанные с заболеваниями варианты в них выявлены только в единичных случаях. Именно поэтому разработка подходов к исследованию механизмов патогенности вариантов в 5'UTR, в частности в uORF, представляет особый интерес для исследовательской деятельности. В нашей работе нами были идентифицированы и классифицированы 83 некодирующих варианта, изменяющих структуру uORF, из открытых баз данных и было проанализировано влияние на инициацию трансляции для вариантов в некоторых из генов.

Биоинформатический анализ выявил 83 вероятно патогенных варианта, которые изменяют структуру uORF

Для поиска некодирующих вариантов, изменяющих длину uORF, среди общедоступных баз данных ClinVar и HGMD был использован ранее разработанный в нашей лаборатории инструмент uORF_Annotator [4]. В результате было обнаружено 83 варианта, классифицированных как патогенные или неизвестного клинического значения, изменяющих структуру uORF за счет их удлинения или укорочения. Из них 12 вариантов уже были

Далее мы распределили выявленные варианты по типу затрагиваемых uORF. Существует два типа uORF: перекрывающиеся и неперекрывающиеся. Неперекрывающиеся uORF - это такие uORF, старт и стоп кодон которых расположены в 5'UTR; в случае, если стоп кодон uORF расположен за старто-кодоном белок-кодирующей последовательности (но не в её рамке), то такие uORF называются перекрывающимися. Оказалось, что 57 вариантов воздействуют на неперекрывающиеся uORF, тогда как 14 влияют на uORF, изначально перекрывающиеся с кодирующей последовательностью (CDS) (Рисунок 1А).

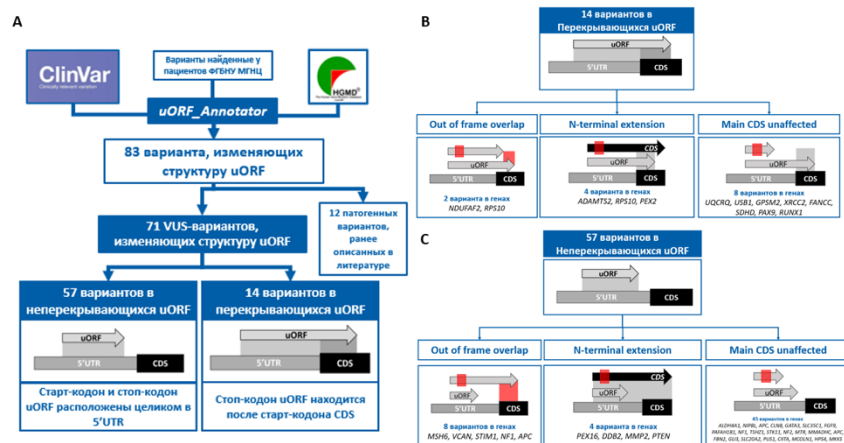


Рисунок 1. Биоинформатический анализ вариантов в uORF. (А) схема биоинформатического анализа; Классификация вариантов по типу затронутого uORF: в перекрывающихся (В) и неперекрывающихся (С)

На следующем этапе исследования мы проанализировали возможное влияние выявленных вариантов на процесс трансляции. Варианты в непересекающихся uORF были разделены на три категории в зависимости от их эффектов: (1) удлинение uORF с перекрытием кодирующей последовательности; (2) удлинение uORF с образованием удлиненного N-конца белка; (3) изменение длины uORF без влияния на кодирующую последовательность

(Рисунок 1С).

Среди вариантов, затрагивающих перекрывающиеся uORF, наблюдались следующие эффекты: (1) уменьшение области перекрытия с кодирующей последовательностью; (2) сдвиг рамки считывания, приводящий к N-концевому удлинению белка; (3) полная потеря перекрытия с белок-кодирующей областью (Рисунок 1В).

Функциональный анализ вариантов в uORF выявил нарушения трансляции белка

Для верификации данных нашего биоинформатического анализа были отобраны шесть вариантов в генах MCOLN1, XRCC2, DDB2, PAX9 и MMADHC. Для экспериментов были клонированы 5'UTR отобранных для анализа генов в вектор psiCHECK-2, таким образом, чтобы старт-кодон проверяемого гена замещал старт-кодон люциферазы. Далее проводили сайт-направленный мутагенез, в ходе которого вносили отобранные варианты в конструкции дикого типа. Трансфекцию плазмид дикого типа и содержащей анализируемый вариант проводили в клеточной культуре HEK293T. Результаты двойного люциферазного теста получали через 48 часов после трансфекции.

По результатам работы было выявлено, что варианты в перекрывающихся uORF приводят к снятию перекрытия с CDS. Так например, для гена PAX9 показано существование одной uORF, при этом вариант с.-119A>T в этом гене, приводящий к образованию преждевременного стоп-кодона с последующим укорочением uORF и нарушением перекрытия с белок-кодирующей последовательности, что и приводит к увеличению экспрессии белка в 1,5 раза (данные соотносятся с тем, что было описано ранее нами [4]). Схожий эффект продемонстрировал и вариант с.-2C>T в гене XRCC2, приводящий к увеличению экспрессии белка в 1,5 раза через снятие перекрытия за счет укорочения uORF (Рисунок 2А). Увеличение эффективности трансляции можно объяснить тем, что нарушение перекрытия нивелирует конкуренцию рамки с белок-кодирующей последовательностью за связывание с рибосомой.

В случае неперекрывающихся uORF, эффект варианта зависит от изменения длины и геномного окружения uORF. Так например, в случае двух вариантов с.-88G>T в гене MCOLN1 и с.-

62G>A в гене MMADHC не было выявлено изменений уровня хемилюминесцентного сигнала. Оба варианта приводят к появлению преждевременного стоп-кодона, что укорачивает uORF на 18 и 6 пар нуклеотидов соответственно.

Более интересным является случай с геном DDB2, для которого показано существование двух uORF, в каждой из которых нами был найден вариант с неизвестным клиническим значением. Оба варианта приводят к количественно схожему изменению уровня трансляции через разные механизмы молекулярного патогенеза. Так, оба варианта приводят к увеличению экспрессии гена более чем в два раза, однако вариант с.-120G>C приводит к данному событию через удлинение uORF1, в то время как другой вариант - с.-31G>C к N-концевому удлинению белка (Рисунок 2В).

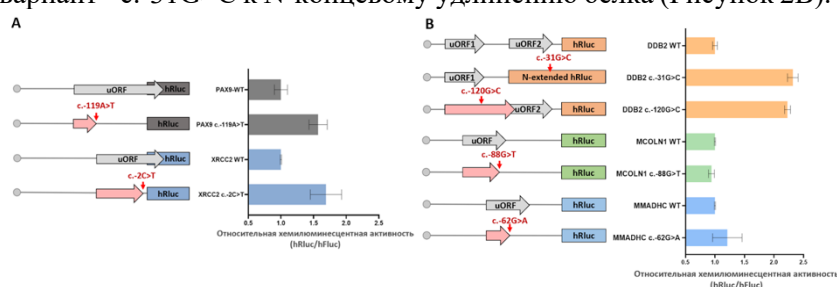


Рисунок 2. Результаты люциферазного теста отобранных вариантов. А. Варианты в генах с перекрывающимися uORF. В. Варианты в генах с неперекрывающимися uORF

Закключение

Поиск механизмов патогенности вариантов в не кодирующей части генома - важная и нетривиальная задача современной медицинской генетики. В данной работе мы провели поиск вариантов, которые разрушают структуру uORF, среди общедоступных баз данных и верифицировали влияние части вариантов на трансляцию основного белка. Проведенный нами функциональный анализ помог подтвердить существование альтернирующих событий трансляции. Данная работа продолжается для большего числа генов.

Список литературы

1. 100,000 Genomes Project Pilot Investigators et al. (2021) 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report, *N Engl J Med.*, 20: 1868–1880.
2. Wright C.F., FitzPatrick D.R. (2018) Paediatric genomics: diagnosing rare disease in children, *Nat Rev Genet.*, 5: 253–268.
3. Carss K.J., Arno G. (2017) Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease, *Am J Hum Genet.*, 1: 75–90.
4. Filatova A., Reveguk I. (2023) Annotation of uORFs in the OMIM genes allows to reveal pathogenic variants in 5'UTRs, *Nucleic Acids Res.*, 3: 1229–1244.
5. Ellingford J.M., Ahn J.W. (2022) Recommendations for clinical interpretation of variants found in non-coding regions of the genome, *Genome Med.*, 1: 73.
6. Whiffin N., Karczewski K.J., (2020) Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals, *Nat Commun.*, 1: 839.
7. Scholz A., Eggenhofer F., (2019) uORF-Tools-Workflow for the determination of translation-regulatory upstream open reading frames, *PLoS One.*, 9: e0222459.

**Биоинформатика в поиске генов сложных заболеваний:
оригинальные методики**

Г.В. Хворых, А.В. Хрунин, С.А. Лимборская

Национальный исследовательский центр «Курчатовский институт», 123182, Россия, Москва, пл. Академика Курчатова, д. 2, khvoryh_gv@nrcki.ru

Широко распространенным способом обнаружения генов многофакторных заболеваний (МФЗ) – являются полногеномные ассоциативные исследования (GWAS). Их отличительные черты: большие панели генотипов однонуклеотидных полиморфных локусов (SNPs), статистически проверяемых на ассоциацию с заболеванием у индивидов тестовой и контрольной выборок.

Первое GWAS провели 20 лет назад. С тех пор появились новые подходы к поиску генов МФЗ. Некоторые из них также нуждаются в индивидуальных генотипах, но есть и те, что используют другие данные. В этой работе представлены четыре методики идентификации генов МФЗ, альтернативные традиционному GWAS, примененные нами для поиска генов ишемического инсульта (ИИ).

В биоинформатике активно используют машинное обучение (МО), в том числе для поиска генов МФЗ. Один из подходов с использованием МО включает обучение классификатора для предсказания статуса индивидов «больной/здоров», а затем ранжирование признаков по важности вклада в фенотип. Признаки отвечают геномным локусам. Однако, важность признака – это не тоже самое, что p -value в статистической модели. Одной только важности, по всей видимости, не достаточно для идентификации локусов, ассоциированных с фенотипом. Развивая методы поиска генов на основе МО, мы пришли к выводу, что более продуктивна концепция, в которой статистическая модель и МО дополняют друг друга [1].

Если тестировать группы SNPs, а не отдельные полиморфизмы, как в традиционном GWAS, то мощность тестов выше. Однако, тогда нужно решить задачу разбиения SNPs на группы. Для этого потенциально подходят алгоритмы кластеризации из МО. Мы применили HDBSCAN. Мерой схожести SNPs была метрика неравновесия по сцеплению r^2 . Алгоритм создает кластеры любой формы и размера. В результате блоки SNPs получились мозаичными в отличие от монолитных блоков, которые давали другие методы, что позволило идентифицировать новые гены предрасположенности инсульта (гены протокадеринов гамма и др.) [2].

Разбивая SNPs на группы, возникает необходимость оценить их связанность в группе. Ранее считали среднее значение r^2 всех пар SNPs в группе. Мы предложили другую метрику – определитель матрицы неравновесия по сцеплению и разработали биоинформатический подход, основанный на разбиении матрицы неравновесия по сцеплению на сингулярные подматрицы. Обосновав метод на искусственных данных, мы применили его к реальным данным больных ИИ, сопоставили с результатами

применения других подходов и продемонстрировали его валидность для поиска генов риска ИИ [3].

Гены, ассоциированные с МФЗ, можно искать и без генотип-фенотипических данных, например, используя генные сети. Вначале их нужно преобразовать в объектно-признаковую матрицу с вещественными числами, используя алгоритмы обучения представлений. Затем добавить целевую переменную, например, с меткой принадлежности к известным генам заболевания. И, наконец, решить задачу поиска новизны методами МО. Применяв данный подход к генной сети и спискам генов, ассоциированных с ИИ, мы обнаружили несколько новых генов-кандидатов, сверхпредставленных в биологических процессах иммунной системы, функционирования внеклеточного матрикса, а также связанных с активацией металлопротеиназы. Все вместе напрямую связаны с ИИ, так как нейровоспаление и перестройка внеклеточного матрикса – два интегрированных компонента данного заболевания [4].

GWAS сравнивают с ген-кандидатным подходом, подчеркивая его независимость от априорных суждений. Вместе с тем, ген-кандидатный метод по-прежнему актуален благодаря новым знаниям о болезни, в том числе, на молекулярном уровне и доступности данных нового качества, например, полных транскриптомов животных моделей болезни. Разработав биоинформатический протокол, мы показали, что возможен перенос данных с уровня транскриптома крысы в условиях ишемии мозга на геном человека. В результате были идентифицированы новые гены предрасположенности (SOCS3, PTX3 и PDPN) и тяжести протекания ишемического инсульта (LGALS3, GPNMB, LSP1 и TAGLN) [5].

Таким образом, биоинформатика гибко реагирует на новые данные и методы, адаптируя их для поиска генов сложных заболеваний. В рассмотренных примерах прослеживается тренд на интеграцию одних методов и данных в другие. Можно ожидать, что он продолжится и далее, открывая новые перспективы в изучении генетической природы многофакторных заболеваний.

Исследование проведено в рамках государственного задания НИЦ «Курчатовский институт» и при поддержке Российского научного фонда (грант № 23-14-00131).

1. S.Nikolić, D.I.Ignatov, G.V.Khvorykh, S.A.Limborska, A.V.Khrunin (2024) Genome-wide association studies of ischemic stroke based on interpretable machine learning, *PeerJ Computer Science*, 10:e2454.
2. G.V.Khvorykh, N.A.Sapozhnikov, S.A.Limborska, A.V. Khrunin (2023) Evaluation of Density-Based Spatial Clustering for Identifying Genomic Loci Associated with Ischemic Stroke in Genome-Wide Data, *Int. J. Mol. Sci.* 24:15355.
3. G.Khvorykh, A.Khrunin (2025) Determinant-based grouping of SNPs and its application for detecting disease-associated genomic loci. *NAR Genomics and Bioinformatics*, 7(1):lqaf024.
4. Г.В.Хворых, К.С.Шафалович, С.А.Лимборская, А.В.Хрунин (2024) Поиск генов-кандидатов ишемического инсульта путем обучения представления геной сети. *Биотехнология*, 40(7):58.
5. A.V.Khrunin, G.V.Khvorykh, A.S.Arapova, A.E.Kulinskaya, E.A.Koltsova, E.A.Petrova, E.I.Kimelfeld, S.A.Limborska (2023) The Study of the Association of Polymorphisms in LSP1, GPNMB, PDPN, TAGLN, TSPO, and TUBB6 Genes with the Risk and Outcome of Ischemic Stroke in the Russian Population. *Int. J. Mol. Sci.* 24:6831.

Simple physical models of active extrusion describe loopy architecture of chromosomes: emergence of a new length-scale

A. Chervinskaya¹, E. Osintseva¹, K. E. Polovnikov^{1,2}

¹*Skolkovo Institute of Science and Technology, Moscow, Russia,
anastasia.chervinskaya@skoltech.ru*

²*Institute of Physics and Astronomy, Potsdam University, Potsdam,
Germany*

Loop extrusion by SMC (Structural Maintenance of Chromosomes) complexes, particularly cohesins and condensins, has emerged as a fundamental mechanism underlying 3D genome

organization [1, 2]. This process is responsible for generating chromatin loops of various sizes, shaping topologically associating domains (TADs), and influencing the statistical properties of chromatin contacts captured in Hi-C experiments [3]. Despite its central role, several basic aspects of loop extrusion remain poorly understood—particularly the statistical distribution of loop sizes and how it arises from the underlying dynamics of loop growth and constraint.

A key assumption in a recent successful polymer-based model of chromosome architecture is that the distribution of chromatin loop lengths is exponential [4]. Such an assumption has been shown to reproduce well the experimental results for contact probabilities. However, until now, a quantitative theoretical explanation for the emergence of this exponential distribution from mechanistic principles has been lacking.

In this work, we present a systematic analytical and numerical investigation of loop length distributions generated by stochastic extrusion processes. We show that under a broad class of biologically plausible assumptions—including finite cohesin lifetimes and random obstacles representing blocking complexes like CTCF—the loop length distribution is indeed exponential. Our models also account for both unidirectional and bidirectional extrusion regimes and predict how the characteristic loop length depends on key parameters such as cohesin processivity, loading density, and steric interference.

Importantly, we demonstrate that the effective loop length distribution arises not only from random dissociation but also from dynamic competition between loops in crowded chromatin environments. Our analysis reveals that the emergent exponential form is robust under various types of loop-loop interactions, providing a strong theoretical foundation for its use in large-scale simulations of chromatin.

Finally, we introduce a graph-theoretic framework to describe loop nesting hierarchies and show how loop dynamics lead to a steady-state distribution of nested structures.

In this study, we leverage an integrative approach to characterize the distribution of chromatin loop lengths using bulk and single-cell resolution experimental data. First, chromatin loops are identified from Hi-C contact maps using the HiCCUPS algorithm. Subsequently, single-cell Cleavage Under Targets and Tagmentation (scCUT&Tag) data provides cell-type-specific binding profiles for a cohesin subunit RAD21

[5]. The detected loops are then filtered based on the presence and intensity of RAD21 scCUT&Tag signal within the anchors at both ends of the loop. This methodology enables the characterization of cell-type-specific length distributions for putative architectural loops. Furthermore, by resolving cellular heterogeneity, it potentially overcomes limitations of bulk data and offers potential insights into loop dynamics. Additionally, we incorporate scCUT&Tag profiles for CTCF within the same cell types into the analysis to identify loops potentially anchored by CTCF-cohesin co-localization.

In summary, this work addresses an important open question in chromatin biology: what is the distribution of loop lengths during extrusion? Our results bridge the gap between polymer physics and genome biology, providing both mechanistic insight and practical tools for modeling genome architecture.

1. J. Dekker, L.A.Mirny (2024) The chromosome folding problem and how cells solve it, *Cell*, **187.23**: 6424-6450.
2. G.Fudenberg et al. (2016) Formation of chromosomal domains by loop extrusion, *Cell reports*, **15.9**: 2038-2049.
3. J.Nuebler et al. (2018) Chromatin organization by an interplay of loop extrusion and compartmental segregation, *Biophysical Journal*, **114.3**: 30a.
4. K.E. Polovnikov et al. (2023) Crumpled polymer with loops recapitulates key features of chromosome organization, *Phys. Rev. X*, **13**, 041029.
5. M. Bartosovic, M Kabbe, G. Castelo-Branco (2021) Single-cell CUT&Tag profiles histone modifications and transcriptopn factors in compex tissues, *Nat Biotechnol*, **39**: 825-835

Региональный паттерн альтернативного сплайсинга в нейрональных и глиальных клетках здорового головного мозга

М.Н. Черняева¹, О.И. Ефимова¹, А.С.Бинкевич², П.В. Мазин²,
Ф.Е. Хайтович¹

¹Центр нейробиологии и нейрореабилитации имени Владимира

Зельмана,
Сколковский институт науки и технологий, Москва,
mariia.cherniaeva@skoltech.ru

²*Институт Сенгера, Информационный центр клеточной
генетики, Хинкстон, Великобритания,*

Альтернативный сплайсинг (АС), будучи одним из важнейших механизмов увеличения белкового многообразия в организме, играет весьма важную роль в клеточной дифференцировке, развитие ткани, в особенности в нейрогенезе и развитие головного мозга млекопитающих, включая формирование сложных нейронных сетей и установление тонких синапсов [1]. К примеру, переключение с РТВР1 на паралог РТВР2 весьма важно для перехода от стволовых клеток к нейронам.

В головном мозге человека установлен крайне высокий уровень альтернативно сплайсируемых генов установлен в головном мозге человека. Так в недавнем исследовании программы АС в ходе развития различных тканей нескольких видов млекопитающих установлен особый паттерн включения/исключения кассетных экзонов в изоформах мРНК в головном мозге [2]. Например, установлен высоко консервативный характер включения в транскрипт трех экзонов гена DLG3, кодирующего белок, ассоциированный с синапсом и играющего важную роль в памяти и обучении.

Поскольку головной мозг является высоко гетерогенной тканью, разделяемой функционально и структурно на регионы, состоящие из различных типов нейрональных клеток и различных подтипов нейроглии, исследование АС в головном мозге стоит проводить как на клеточно-специфичном, так и регион-специфичном уровне. Тем не менее, анализ АС с использованием транскриптомных данных ткани головного мозга и данных транскриптов на уровне одиночных клеток, ограничено небольшим числом исследуемых регионов головного мозга. По этой причине, в данной работе проведен анализ различий событий АС (кассетных экзонов) между 33 регионами здорового головного мозга, включая кору головного мозга, регионы лимбической системы и мозжечок.

Далее определены маркерные события сплайсинга в основных нейрональных и глиальных клеточных типах и сравнены с регион-специфичными кассетными экзонами. По результате исследования установлены регионы, отличающиеся наибольшим числом регион-специфичных кассетных экзонов, такие как регионы белого вещества головного мозга и серого вещества мозжечка, причем некоторые экзоны, например 3 и 4 экзон гена CD47, имеют противоположный характер включения в сером веществе мозжечка и белом веществе. Также определен схожий паттерн использования кассетных экзонов между регионами белого вещества и олигодендроцитами, серым веществом мозжечка и нейронами. Интересен факт, что наиболее регион-специфичные экзоны, имеющие одинаковый с клеточными типами паттерн, относятся к генам, участвующим в процессах регуляции актина, соответственно морфологии клеток, таким как CAPZB в белом веществе и олигодендроцитах, ADD3 в сером веществе мозжечка и нейронах, PPP1R12A. Однако, регуляция актиновых филаментов хорошо рассмотрена в нейрональных нежели в глиальных клетках [3]. Данное исследование дает более глубокое понимание программы AC в морфологии клеток и взаимосвязи различий основных клеточных типов центральной нервной системы со структурной организацией мозга.

Источник финансирования данной работы – грант РФФИ № 22-15-00474_П.

1. B. Raj, B. J. Blencowe (2015) Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles, *Neuron*, **87**:14–27.
2. P.V. Mazin, P. Khaitovich, M. Cardoso-Moreira et al (2021) Alternative splicing during mammalian organ development, *Nature genetics*, **53**:925–934.
3. E.J. Thomason, M. Escalante, D.J. Osterhout, et al (2019) The oligodendrocyte growth cone and its actin cytoskeleton: A fundamental element for progenitor cell migration and CNS myelination, *Glia*, **6**:1329–1346.

DBP-Finder: Utilizing Large Language Models for Sequence-based Prediction of DNA-binding Proteins

Elizaveta Shaburova, Alexander Gavrilenko, Denis Maslov,
Dmitry Shtokalo, Yuri Vyatkin, Vasily Ramensky, Denis Antonets

*Institute for Advanced Study of Artificial Intelligence and Intelligent
Systems,*

*M.V. Lomonosov Moscow State University, Moscow, 27 bldg. 1
Lomonosovsky Ave., e.shaburova@iai.msu.ru*

Introduction. DNA-binding proteins (DBPs) play crucial roles in numerous biological processes, including DNA replication, repair, transcription, recombination, and chromatin remodeling. The precise identification of DBPs is fundamental for advancing our understanding of cellular mechanisms and developing therapeutic interventions for diseases such as cancer. Traditional experimental methods for identifying DBPs are labor intensive, costly, and limited by experimental biases. Hence, there is a significant demand for accurate computational approaches capable of rapid, large-scale identification of DBPs from vast protein sequence data.

We introduce DBP-Finder, a sophisticated computational framework utilizing large pretrained protein language models (PLMs) combined with advanced transfer learning techniques to predict DNA-binding proteins directly from amino acid sequences. The model capitalizes on recent advances in natural language processing applied to protein sequences, providing contextually rich embeddings that significantly improve prediction performance.

Methods. To develop DBP-Finder, we first constructed balanced training datasets derived from the UniProtKB/Swiss-Prot database, explicitly annotated as DBPs (positive samples) or non-DBPs (negative samples) based on Gene Ontology (GO) annotations. Data processing included stringent criteria to exclude sequences outside the length range of 50 to 1024 amino acids and sequences with non-canonical amino acids, thus maintaining high dataset quality and computational efficiency. Furthermore, we carefully eliminated potential data leakage by excluding homologous sequences (>50%

identity) from the training datasets relative to independent testing datasets.

We used Ankh [1], a large protein language model with 1.15 billion parameters, to transform amino acid sequences into feature-rich embeddings. Each protein sequence was represented as a sequence length-dependent embedding with dimensions of 1536. To optimize computational resources, sequences within batches were sorted by length, thus minimizing necessary padding.

Embeddings generated by Ankh are submitted to the ConvBert-based model with approximately 13 million parameters. The model is a transformer followed by max pooling and additional linear layers for binary classification. Training involved the AdamW optimizer with a batch size of 64 over nine epochs, including adaptive learning rate adjustments based on validation performance to avoid overfitting. Robustness was enhanced using an ensemble approach consisting of five independently trained models employing grouped 5-fold cross-validation, with proteins from identical sequence clusters kept within the same fold to prevent bias. Each model's decision threshold was individually tuned based on validation set performance to optimize sensitivity and specificity.

Results. DBP-Finder demonstrated substantial improvements in predictive performance compared to existing methods across different benchmark datasets. Specifically:

- **PDB2272 dataset:** In order to prove superiority of the model in the class of existing sequence-based DBP predictive models, the established benchmark PDB2272 [2, 3] has been chosen. The model exhibited an accuracy of 82.5%, a specificity of 76.8%, and an MCC of 0.653, demonstrating clear improvements over previous state-of-the-art approaches like Target-DBPPred [2] and MsDBP [3], particularly in highly diverse datasets.
- **TrEMBL/Swiss dataset:** To provide a comprehensive evaluation, we compared DBP-Finder with two "general purpose" models that predict GO terms corresponding to molecular functions: ProteInfer [4] and SPROF-GO [5]. We compiled a test dataset from UniProtKB/TrEMBL and UniProtKB/Swiss-Prot, ensuring that all proteins in these

datasets exhibited no more than 50% sequence identity with those in the training sets of the aforementioned models. DBP-Finder attained an AUC of 97%, outperforming ProteInfer with 70.3% and SPROF-GO with 96.4%.

In addition to benchmark performance, DBP-Finder was evaluated across different taxonomic groups, including Metazoa, Bacteria, Viridiplantae, Fungi, Archaea and Viruses. The method consistently exhibited high predictive accuracy across these diverse groups, with better results for Metazoa and Bacteria. However, the model faced challenges with viral protein prediction, attributed to the inherent sequence variability and complexity of viral polyproteins.

We validated the practical utility of DBP-Finder with proteome-scale predictions for several model organisms. The method predicted thousands of putative DBPs with high confidence ($\geq 90\%$) across multiple model organisms, including 1142 proteins in *Homo sapiens*, 911 proteins in *Arabidopsis thaliana*, 120 proteins in *Escherichia coli*, and 79 proteins in *Saccharomyces cerevisiae* with no reported DNA-binding annotation from InterProScan.

Conclusion. DBP-Finder represents the successful integration of natural language processing methods and computational biology to address the need for accurate, efficient prediction of DNA-binding proteins from sequence data alone. Its robust performance, validated across challenging benchmarks and diverse biological kingdoms, underlines its potential as a powerful tool in biological research and drug discovery pipelines.

1. A. Elnaggar, H. Essam, et al. (2023). ♀: Optimized Protein Language Model Unlocks General-Purpose Modelling, *bioRxiv*, 2023.01.16.524265
2. F. Ali, H. Kumar, et al. (2022). Target-DBPPred: An intelligent model for prediction of DNA-binding proteins using discrete wavelet transform based compression and light eXtreme gradient boosting. *Computers in Biology and Medicine*, **145**:105533.
3. Du X., Diao Y., et al. (2019) MsDBP: Exploring DNA-Binding Proteins by Integrating Multi-

Scale Sequence Information via Chou's 5-Steps Rule, *Journal of Proteome Research*, **18**:3119–3132.

4. Sanderson, T., Bileschi, M. L., et al. (2023). ProteInfer, deep neural networks for protein functional inference. *eLife*, **12**:e80942.

5. Yuan, Q., Xie, J., et al. (2023). Fast and accurate protein function prediction from sequence through pretrained language model and homology-based label diffusion. *Briefings in Bioinformatics*, **24**:bbad117.

A genome-wide association study in 10,000 individuals links plasma N-glycome to liver disease and anti-inflammatory proteins

S. Sharapov¹, A. Timoshchuk¹, O. Zaytseva², D. Maslov¹, A. Soplenkova¹, E. E. Elgaeva^{3,4}, E. S. Tiys³, M. Mangino^{5,6}, C. Wittenbecher⁷, L. C. Karssen⁸, M. Timofeeva^{9,10}, A. Nostaeva^{3,8}, F. Vucković², I. Trbojević-Akmačić², T. Štambuk², S. Feoktistova³, N. A. Potapova¹¹, V. Voroshilova^{3,12}, F. Williams⁵, D. Primorac^{13,14}, J. V. Zundert^{15,16}, M. Georges¹⁷, K. Suhre¹⁸, M. Allegri¹⁹, N. Chaturvedi²⁰, M. Dunlop⁹, M. B. Schulze^{7,21,22}, T. Spector⁵, Y. A. Tsepilov^{3,23}, G. Lauc^{2,24}, Y. S. Aulchenko^{1,3}

¹*MSU Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow, Russia, timoschukanny@gmail.com*

²*Genos Glycoscience Research Laboratory, Borongajska cesta 83H, 10000 Zagreb, Croatia*

³*Institute of Cytology and Genetics, Novosibirsk, 630090, Russia*

⁴*Novosibirsk State University, Novosibirsk, 630090, Russia*

⁵*Department of Twin Research and Genetic Epidemiology, School of Life Course Sciences, King's College London, St Thomas' Campus,*

Lambeth Palace Road, London, SE1 7EH, UK

*⁶NIHR Biomedical Research Centre at Guy's and St Thomas'
Foundation Trust, London SE1 21 9RT, UK*

*⁷Department of Molecular Epidemiology, German Institute of Human
Nutrition Potsdam-Rehbruecke, Nuthetal, Germany*

⁸PolyOmica, 's-Hertogenbosch 5237 PA, Netherlands

*⁹Colon Cancer Genetics Group, Cancer Research UK Scotland Centre,
Institute of Genetics & Cancer, Western General Hospital, The
University of Edinburgh, Edinburgh EH4 2XU, UK*

*¹⁰D-IAS, Danish Institute for Advanced Study, Department of Public
Health, University of Southern Denmark, J.B. Winsløws Vej 9, DK-
5000 Odense C, Denmark*

*¹¹Engelhardt Institute of Molecular Biology, Russian Academy of
Sciences, 119991 Moscow, Russia*

*¹²Vavilov Institute of General Genetics Russian Academy of Sciences,
119991 Moscow, Russia*

*¹³St. Catherine Specialty Hospital, Ulica Kneza Branimira 71E,
Zagreb, Croatia; University of Split School of Medicine, Split 21 000,
Croatia*

¹⁴University of Osijek School of Medicine, Osijek 31 000, Croatia

*¹⁵Department of Anesthesiology and Multidisciplinary Paincentre, ZOL,
Genk/Lanaken, Belgium*

*¹⁶Department of Anesthesiology and Pain Medicine, Maastricht
University Medical Centre, P. Debyelaan 25, Maastricht, 6229 HX, The
Netherlands*

¹⁷*Unit of Animal Genomics, WELBIO, GIGA-R and Faculty of Veterinary Medicine, University of Liège, (B34) 1 Avenue de l'Hôpital, Liège 4000, Belgium*

¹⁸*Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Education City, P.O. Box 24144 Doha*

¹⁹*Centre Lemanique d'antalgie et neuromodulation – EHC - Morges - CH*

²⁰*MRC Unit for Lifelong Health & Ageing University College London, London, UK*

²¹*German Center for Diabetes Research (DZD), Neuherberg, 85764, Germany*

²²*Institute of Nutrition Science, University of Potsdam, Potsdam, Germany*

²³*Wellcome Sanger Institute, Cambridge, CB10 1RQ, the UK, yt4@sanger.ac.uk*

²⁴*Croatia University of Zagreb Faculty of Pharmacy and Biochemistry, Zagreb, Croatia*

Glycosylation is a post-translational protein modification that influences both the physicochemical properties and biological functions of proteins. While the biochemical network of human N-glycan biosynthesis is well understood [1], little is known about *in vivo* regulation of this process [2], including tissue- and protein-specific regulation. In this study, we aimed to find candidate genes involved in the regulation of N-glycosylation and explore the potential role of N-glycans in the control of complex traits and multifactorial diseases. We performed univariate and multivariate genome-wide association

meta-analyses on more than 10,000 individuals, examining 117 N-glycome traits. To identify causal genes for each replicated locus, we used eight in silico-based predictors. Moreover, we conducted a phenome-wide association study to investigate shared genetic control between variants associated with changes in N-glycosylation and pathological conditions. Finally, we performed a bidirectional Mendelian randomization to identify potential causal relationships between glycans and diseases.

This work represents the largest to date genome-wide association study on N-glycosylation of blood plasma proteins to date. We doubled the number of genetic loci known to be associated with blood N-glycosylation by identifying 16 novel loci and prioritizing 13 novel genes contributing to N-glycosylation. Among these were the *GCKR*, *TRIB1*, *HP*, *SERPINA1* and *CFH* genes. These genes are predominantly expressed in the liver and show a previously unknown genetic link between plasma protein N-glycosylation, metabolic and liver diseases, and inflammatory response. Furthermore, we linked the regulation of protein N-glycosylation with the regulation of over three hundred complex traits and diseases, such as diabetes (types 1 and 2), cardiovascular diseases and cardiometabolic traits, immune-mediated diseases, neuropsychiatric disorders and anthropometric traits. Moreover, we have found causal relationships between high-mannose N-glycosylation of proteins and asthma and lipoprotein metabolism disorders.

Notably, this study is the first to analyze and reveal strong tissue-specificity of the genetic regulation of population variability of human protein N-glycosylation. The two main sources of glycoproteins present in human blood are liver and lymphoid tissue. While a number of glycosyltransferases are expressed both in liver and lymphoid tissues, we demonstrate that, to the resolution allowed by the current study, the genetic variation regulating N-glycosylation in each of the two tissues is unique and does not overlap.

By integrating glycomics, proteomics, and genomics, we provide a resource that facilitates deeper exploration of disease pathogenesis and supports biomarker discovery.

The work of AS, AT, DM, SSh, YSA was supported by the Research Program at the MSU Institute for Artificial Intelligence.

1. A.Varki et al. (2015) Essentials of Glycobiology, *The Book*, (Cold Spring Harbor Laboratory).
2. A. Timoshchuk et al. (2023) Twelve Years of Genome-Wide Association Studies of Human Protein N-Glycosylation, *Engineering*.

Инструмент интерпретируемого предсказания карт Hi-C

А.С. Школиков^{1,2}, М.С. Гельфанд³

¹ФББ МГУ (Москва, Ленинские горы 1с73), a.shkolikov@gmail.com

²Институт общей генетики им. Н.И. Вавилова РАН (Москва, ул. Губкина 3),

³Сколтех, Центр наук о жизни (Москва, Большой бульвар, 30с1),

Трехмерная организация генома играет важную роль в регуляции экспрессии генов и функционировании. В последние годы был накоплен большой объем данных о пространственной укладке генома из экспериментов Hi-C и Micro-C для различных видов. Хотя наблюдаемые паттерны, такие как топологически ассоциированные домены и петли, сохраняются у разных видов, базовые биологические механизмы могут различаться. Как видоспецифичные архитектурные факторы, так и последовательности ДНК влияют на сворачивание хроматина, усложняя сравнительные исследования эволюции стоящих за этим механизмов. Мы использовали существующие данные Hi-C и машинное обучение для исследования видоспецифичных механизмов сворачивания трехмерного генома и предсказания структур хроматина по последовательностям ДНК. Мы представляем инструмент Chimaera – нейронную сеть, которая не только предсказывает карты Hi-C по последовательности ДНК, но и позволяет осуществлять поиск, количественную оценку и интерпретацию ассоциаций между последовательностями ДНК и паттернами укладки генома. Исследуя скрытые представления, порождаемые нейронной сетью, мы предлагаем инструмент для создания атласа структур хроматина, таких как домены, петли и фонтаны. Также мы проводим направленный поиск элементов

последовательности ДНК, связанных с определенными структурами хроматина, что расширяет наше понимание организации генома. Расширяя поиск элементов последовательности ДНК на несколько видов, мы подтверждаем роль CTCF в формировании инсуляции у позвоночных и BEAF-32 у *Drosophila*, а также выявляем мотивы, ранее не отмеченные у мыши и *Drosophila*. У амёбы *Dictyostelium Chimaera* демонстрирует важность расположения генов на цепи ДНК для образования петель, подтверждая гипотезу о влиянии конвергентного расположения генов на трехмерную организацию генома этого организма. В предсказаниях для других организмов также замечен выраженный, но отличающийся эффект расположения генов. Наконец, мы обучаем модель на данных одного вида, а затем применяем ее для предсказания по последовательностям других видов. Таким образом мы проверяем, могут ли модели укладки хроматина согласовываться между видами, и выявляем эволюционные сходства между механизмами, построив кластерное дерево видов на основе хроматина.

Работа проводится в рамках ГЗ FFRW-2024-0004 (ИОГен РАН).

Тезисы стендовых докладов

Analysis of Mpox RNA-seq Data Reveals Low Evidence of APOBEC3-Mediated RNA Editing

R.Kh. Abasov¹, A.O. Lyskova¹, E.V. Matveev¹, F.M. Kazanov², G.V. Ponomarev³, M.D. Kazanov^{3,4}

¹*Dmitry Rogachev National Medical Research Center Of Pediatric Hematology, Oncology and Immunology, Moscow, Russia, ruslan.abasov.2013@gmail.com*

²*Faculty of Economic Sciences, Higher School of Economics, Moscow, Russia*

³*Vavilov Institute of General Genetics, Moscow, Russia*

⁴*Skolkovo Institute of Science and Technology, Moscow, Russia*

Monkeypox disease, known for a multi-country outbreak in 2022, is caused by the monkeypox virus (MPXV) – a double-stranded DNA virus belonging to the Orthopoxvirus genus, Poxviridae family. The estimated mutation rate for orthopoxviruses is 1–2 substitutions per genome per year; however, the extent of divergence of the 2022 outbreak genomes from the related 2018–2019 viruses reached about 50 single-nucleotide substitutions. Analysis of these substitutions showed that most of them were C→T or G→A transitions occurring in tC or complementary Ga contexts, which represent the mutational signature of APOBEC3 cytidine deaminases.

In this study, we analyzed RNA-seq data of MPXV to determine whether the MPXV transcriptome is also affected by APOBEC3 enzymes. We considered six RNA-seq studies, detected substitutions in transcripts, and analyzed whether these substitutions occur at the DNA or RNA level using indirect evidence, including effects on amino acid changes, correlation with transcriptional features, and analysis of possible secondary structures known for mutation hotspots. We concluded that there is low evidence of RNA editing of MPXV transcripts by APOBEC3 enzymes.

This work is supported by FFRW-2024-0004 (VIGG RAS).

**Генетические маркеры предрасположенности к спорту
высоких достижений и их потенциальное влияние на
продолжительность жизни**

Г.А. Абушинова¹, М.В. Терехов¹, А.А. Мамчур¹, М.В. Брутан¹,
И.Х. Джуманиязова¹, М.В. Иванов¹, А.В. Жолинский²,
А.И. Кадыкова^{2,3}, Н.С. Гладышев², Р.В. Деев², Л.Р. Маткава¹,
М.Н. Пилипенко¹, С.И. Митрофанов¹, А.А. Ивашечкин¹,
Е.Д. Маралова¹, В.С. Юдин¹, В.В. Макаров¹, А.А. Кескинов¹,
Д.А. Каштанова¹, С.М. Юдин¹, В.И. Скворцова³

¹ФГБУ «ЦСП» ФМБА России, Москва, gabushinova@cspfmba.ru

²ФГБУ ФНКЦСМ ФМБА России, Москва

³Федеральное медико-биологическое агентство (ФМБА России),
Москва

Профессиональный спорт требует серьёзной физической подготовки и предполагает определённые физиологические и психологические различия между спортсменами разного уровня успешности. Кроме того, для успеха в разных видах спорта спортсменам необходимо неравномерно развивать различные физиологические характеристики, к совершенствованию которых может быть генетическая предрасположенность.

Для определения маркеров генетической предрасположенности к спорту высоких достижений мы провели исследование 3676 профессиональных спортсменов, занимающихся различными видами спорта, а именно: единоборства ($n = 875$), многоборье ($n = 102$), скоростно-силовые дисциплины ($n = 146$), сложнокоординационные дисциплины ($n = 630$), спортивные игры ($n = 1143$) и циклические дисциплины ($n = 780$). Для поиска генетических вариантов, ассоциированных с успехом в разных спортивных дисциплинах, было проведено шесть GWAS-исследований. В качестве контрольной группы для них выступала общепопуляционная выборка из 21516 участников, входящих в базу ФГБУ «ЦСП» ФМБА России. Анализ и сравнение результатов GWAS-исследований показали наличие специфичных и общих для

сходных по особенностям физических нагрузок спортивных групп генетических вариантов. Интересно, что все значимые полиморфизмы расположены в межгенных и интронных областях генома, то есть влияют не на структуру и функцию кодируемых белков, а на регуляцию их экспрессии.

По результатам GWAS на основе методов машинного обучения построены соответствующие полигенные шкалы, предсказывающих генетическую предрасположенность человека к определенной спортивной группе (ROC AUC = 0,67 для сложнокоординационных дисциплин, 0,75 для единоборств, 0,77 для скоростно-силовых дисциплин, 0,82 для многоборья, 0,68 для спортивных игр и 0,67 для циклических дисциплин). С учётом полученных результатов выдвинута идея индивидуального полигенного профиля, указывающего на предрасположенность к занятию тем или иным видом спорта на основе генетических данных. В перспективе такой полигенный профиль может быть использован тренерами и самими спортсменами в рамках оптимизации усилий в спорте высших достижений и снижения временных и материальных затрат при подготовке профессиональных спортсменов.

Анализ отдалённых эффектов обнаруженных генетических вариантов в когорте долгожителей (90+ лет, N=2914) показал ассоциацию долголетия с высокими значениями по полигенным шкалам, предсказывающим предрасположенность к сложнокоординационным и циклическим дисциплинам, а также к спортивным играм. В то же время предрасположенность к единоборствам для долгожителей оказалась низкой — природу этой ассоциации ещё предстоит изучить.

Таким образом, полученные результаты имеют значение как в практическом, так и фундаментальном аспектах спортивной генетики. В связи со многими этическими проблемами прямое внедрение генетического профилирования как исключительного параметра для выбора наиболее подходящего вида спортивной деятельности и тренировочных программ невозможно. Однако оно может стать одним из важных инструментов в комплексном подходе тренера к своевременной специализации и выбору оптимальной траектории развития, необходимой для раскрытия спортивного таланта каждого конкретного человека.

Проблема интерпретации данных SAMap

А.И. Агламазова, Г.А. Шиловский, В.А. Любецкий

Институт проблем передачи информации aglamazova.an@iitp.ru

Метод SAMap представляет собой современный инструмент для сопоставления данных одноядерной (snRNA-seq) и одноячейковой (scRNA-seq) транскриптомики, который активно используется для создания межвидовых клеточных атласов и изучения эволюционных процессов. Его применение позволяет объединять данные из различных экспериментов, сопоставлять клеточные популяции между видами и выявлять молекулярные механизмы, лежащие в основе клеточной гетерогенности. Однако интерпретация результатов, полученных с помощью SAMap, ограничена рядом трудностей, включая зависимость от качества данных, сложность аннотации кластеров, трудности работы с редкими клеточными типами, ограниченность межвидового анализа и значительные вычислительные затраты.

Одной из ключевых проблем является недостаточная точность визуализации и интерпретации результатов сопоставления. Перекрывание кластеров в многомерном пространстве затрудняет точное определение клеточных типов, особенно в условиях высокой гетерогенности данных. Кроме того, редкие клеточные популяции часто теряются при кластеризации или объединяются с более крупными группами, что приводит к потере биологически значимой информации. Эти проблемы усиливаются при анализе данных из эволюционно удалённых видов, где различия в геномных последовательностях и экспрессии генов могут снижать точность сопоставления клеточных типов.

Предлагаемая комбинация методов SAMap и PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding) способна значительно улучшить интерпретацию результатов. PHATE — это метод редукции размерности и визуализации, который сохраняет как локальные, так и глобальные структуры

данных, делая его особенно полезным для анализа траекторий клеточных состояний и сложных многомерных отношений между клетками. На основе сочетания SAMar и PHATE нами разработан pipeline, который:

PHATE создает латентное пространство, которое используется SAMar для сопоставления клеточных типов, что помогает смягчить влияние шума и улучшить точность анализа.

PHATE обеспечивает гладкое и интуитивно понятное представление сопоставленных клеточных типов, особенно в задачах, связанных с траекториями клеточных изменений.

PHATE выделяет локальные структуры данных, что позволяет эффективно работать с редкими клеточными популяциями, которые традиционно теряются при стандартных методах обработки.

Комбинация SAMar и PHATE повышает биологическую обоснованность и воспроизводимость результатов. Оптимизация алгоритма визуализации UMAP, включая использование PHATE, уменьшает перекрытие кластеров, что упрощает определение клеточных типов, в том числе в далеко отстоящих на эволюционном древе таксонах Metazoa.

Предложенная интеграция может улучшить анализ данных из эволюционно удалённых видов, поскольку PHATE сохраняет глобальные структуры, что делает результаты сопоставлений SAMar более интерпретируемыми.

Метод SAMar в сочетании с PHATE открывает новые возможности для изучения клеточных типов и их эволюции, улучшая как точность сопоставления, так и удобство интерпретации. Эти подходы частично решают проблему определения родственных клеточных типов в наборе кластеризаций разных видов.

Исследование выполнено за счёт гранта Российского научного фонда № 24-44-00099, <https://rscf.ru/project/24-44-00099/>

Список литературы

1. Tarashansky, Alexander J., et al. "Mapping single-cell atlases throughout Metazoa unravels cell type evolution." *Elife* 10 (2021)
2. Moon, Kevin R., et al. "Visualizing structure and transitions in

high-dimensional biological data." Nature Biotechnology" 37.12 (2019)

3. X. Zhang, Z. Li, J. Chen, W. Yang, X. He, P. Wu, F. Chen, Z. Zhou, Ch. Ren, Y. Shan, X. Wen, V.A. Lyubetsky, L.Yu. Rusin, X. Chen, J.-R. Yang. Stereotyped subclones revealed by high-density single-cell lineage tracing support robust development. Advanced Science, Apr 30 2025, Art. 2406208.

OmiXAI: ансамблевый XAI-пайплайн интерпретации моделей глубинного обучения для омиксных данных

Алаева А.И., Лаптева А., Михайловская Н.Н., Мальков В.В.,
Боревский А.О., Попцова М.С.

*Международная лаборатория Биоинформатики ИИИиЦН
ФКН НИУ ВШЭ, Россия, 101000, г. Москва, ул. Мясницкая, д. 20
a.ameli.ig@gmail.com
mpoptsova@hse.ru*

Методы глубинного обучения в настоящее время становятся основным инструментом для анализа геномных данных. Однако качество работы таких моделей напрямую зависит от объема и характера данных, доступных для обучения. В последние годы наблюдается тенденция к использованию мультиомиксных данных, включающих различные уровни биологической информации: геномные, транскриптомные, эпигеномные, протеомные, метаболомные и различные другие. При достижении высокой точности модели, обученной на подобных данных, возникает важный вопрос: какие именно признаки и биомолекулярные факторы лежат в основе её предсказательной способности?

Методы объяснимого искусственного интеллекта (XAI, Explainable AI) помогают выявить самые важные признаки. Все XAI-подходы можно условно разделить на две категории: зависящие от модели (model-aware) и независящие от модели (model-agnostic). Вторые обычно основаны на комбинаторных нарушениях входных признаков и требуют значительных

вычислительных ресурсов, что делает их применение к глубоким моделям крайне затратным.

Чтобы преодолеть эти ограничения, мы разработали OmiXAI — модульный и масштабируемый пайплайн, основанный на ансамбле методов интерпретации нейронных сетей, учитывающих внутреннюю структуру модели. В состав OmiXAI входят градиентные методы интерпретации, такие как Integrated Gradients, InputXGradients, Deconvolution, Guided Backpropagation (для сверточных и графовых нейросетей), а также Saliency Maps и GNNExplainer (специально для графовых моделей).

Для тестирования эффективности OmiXAI мы применили его к задаче предсказания участков Z-ДНК на основе мультиомиксных данных. Результаты показали высокую информативность метода: была проведена оценка важности признаков и сравнительный анализ различных XAI-алгоритмов. Особенно важно, что с помощью OmiXAI удалось провести отбор признаков — сократив количество важных с почти 2000 до всего 50, что значительно облегчает последующую биологическую интерпретацию и уменьшает требования к ресурсам для анализа и обучения нейронных сетей.

Гибкая структура OmiXAI позволяет легко интегрировать новые методы интерпретации, что делает его применимым не только в области исследования омиксных данных, но и для других задач, требующих интерпретации моделей глубинного обучения. В рамках работы мы также провели бенчмарк каждого отдельного XAI-метода, обсудили их преимущества и ограничения, что может быть полезно для будущих работ в сфере геномных исследований и за ее пределами.

Study of cytochrome genes variability in the Russian population

V.A. Andreev¹, Y.M. Suvorova²

¹*MIPT, Moscow, Russia, andreev.va@phystech.edu*

²*Biotechnological Campus LLC, Moscow, Russia,*

Studying the prevalence of pharmacogenetic alleles in populations is critical for personalized medicine, as the frequencies of genetic variants associated with drug response can vary significantly between ethnic and geographic groups. International databases such as PharmVar provide information on clinically relevant alleles, but often do not reflect their actual frequencies in populations not represented in these databases.

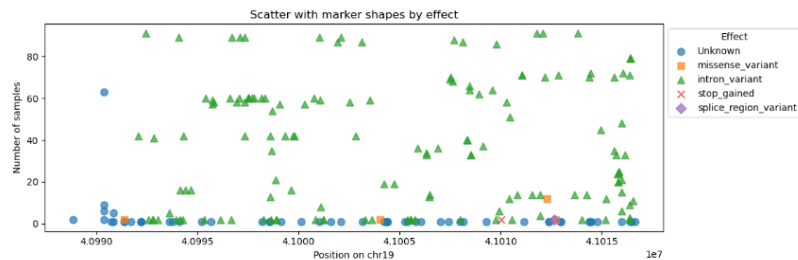
As part of the national initiative “100,000+ Me” implemented by Biotechnology Campus, we analyzed the full genomic data of more than 2,000 Russians. The main focus was on CYP2D6 and CYP2B6 genes - key enzymes involved in drug metabolism [1, 2]. Allele frequency was determined based on a consensus approach using three independent pharmacogenotyping algorithms: PyPGx, StellarPGx and Aldy. An allele was considered reliably predicted if it was detected by all three tools.

Comparison with other Russian studies showed discrepancies in the frequencies of some alleles, which may be due to both differences in genotyping methods and the lack of a common standard of data interpretation. We also provide data on the frequencies of a large number of alleles that are often not the focus of Russian pharmacogenomic studies. In a number of cases, we have identified potentially novel alleles. In order to classify a variant as potentially novel, the following criteria were used: it must be absent in the PharmVar database, belong to a specific background allele, and not be part of an already registered combination of variants.

The findings emphasize the need to consider ethnic specificity in pharmacogenetic studies, as well as the importance of using strict validity criteria when interpreting allelic predictions. Our results can be used to update national and international pharmacogenetic databases and to develop population-specific diagnostic tests.

As a result of our work, we found 11 new variants in the CYP2D6 gene, among which 1 variant is found in more than half of the samples, and 5 variants can potentially have an effect (missense variant, stop gained, frameshift). We also found 156 variants (figure 1) in the CYP2B6 gene not previously described in PharmVar, among which 4 variants have a specific background allele and may affect protein function.

Figure 1. New variants for CYP2B6



1. Maulana Y. et al. (2024) The variation landscape of CYP2D6 in a multi-ethnic Asian population, Scientific Reports, 14:16725
2. Langmia, I.M. et al. (2021) CYP2B6 Functional Variability in Drug Metabolism and Exposure Across Populations—Implication for Drug Safety, Dosing, and Individualized Therapy, Frontiers in Genetics, 12:692234

Optimization of whole genome sequencing depth and informativity profile

A. N. Antonenko, M. S. Belenikin

*LLC “Evogen” Moscow, 4-th Roshchinskij proezd, 20,
antonenko_a.n@mail.ru*

Advances in genetics and molecular diagnostic techniques have made sequencing a routine process that is used extensively in scientific and clinical purposes. As techniques have improved, depth of coverage in sequencing has become an important factor affecting data quality, variant search, and subsequent analysis. Understanding how to optimize and rank depth of coverage for specific targets is critical to improve efficiency, reduce costs, and ensure reliable, relevant, and reproducible

results. Extensive data have been collected in the literature on the application of different depths of coverage in sequencing and the problems solved by such a method. For example, ultra-low and low coverage whole genome sequencing in the range (<0.5 - $10\times$) is mainly used to perform genotyping and whole genome association search (GWAS), or a budget screening option. [1-3]. Medium to high coverage in the range (30 - $60\times$) is used to perform routine whole genome sequencing as well as for clinical samples and rare variant searches. In particular, this coverage significantly increases the possibilities of structural variant detection, but the result depends not only on the read depth, but also on the chosen technology, a set of approaches and processing methods [4-7]. Applications of ultra-high coverage sequencing ($>90\times$) are mainly focused on detection of somatic mutations in oncology or liquid biopsy, search for mosaicisms and rare genetics variants, mitochondrial sequencing or single-cell sequencing [8-10].

The criteria for evaluating and calculating the minimum sequencing depth are different for different applications. In our study, we utilize whole genome sequencing to detect rare diseases. Due to the diversity of sequencing platforms and bioinformatic pipelines, it is important not only to know the performance of each specific pipeline on the same amount of data, but also the variation in the results of counting with different amounts of analyzed data. This information is also important for evaluating the performance of pipelines within a single laboratory.

At this work, we processed different amounts of our own whole genome sequencing data with different read depths (in 1 - $5\times$ steps) to determine the informativeness profile of the two bioinformatic pipelines used in this work.

References:

1. Vylyny Chat et al. (2022) Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies, *Front Genet*, Feb 15;12:790445.

2. Martin, A. R. et al. (2021) Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations, *Am. J. Hum. Genet.* 108, 656–668.
3. Runyang Nicolas Lou et al. (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics, *Mol Ecol.* Dec;30(23):5966-5993.
4. Justin M. Zook et al. (2020) A robust benchmark for detection of germline large deletions and insertions, *Nature Biotechnology*, 38:1347–1355.
5. Justin Wagner et al. (2022) Curated variation benchmarks for challenging medically relevant autosomal genes, *Nature Biotechnology*, 40:672–680.
6. Madeline H Kowalski et al. (2019) Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations, *PLoS Genet*, 23;15(12):e1008500.
7. Marta Byrska-Bishop et al. (2022) High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios, *Cell*, 1;185(18):3426-3440.e19.
8. Alexandre Pellan Cheng et al. (2025) Error-corrected flow-based sequencing at whole-genome scale and its application to circulating cell-free DNA profiling, *Nature Methods*, 22:973–981.
9. Frederik Otzen Bagger et al. (2024) Whole genome sequencing in clinical practice, *BMC Medical Genomics*, 17:39.
10. Louise de Schaetzen van Brien et al. (2020) Comparative analysis of somatic variant calling on matched FF and FFPE WGS samples, *BMC Medical Genomics*, 13:94.

Системный анализ новой подгруппы бактериальных карго- белков для инкапсулинов типа 1

А.А. Анучина, П.Е. Шишкин, А.А. Ремеева, И.Ю. Гущин

¹ *Центр исследований молекулярных механизмов старения и возрастных заболеваний, Московский физико-технический институт, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 9. anichina.aa@mipt.ru*

Инкапсулины обладают способностью упаковывать различные типы карго, включая сульфурилазы, пероксидазы, оксидоредуктазы и ферритин-подобные белки (ferritin-like proteins, FLPs) [1–4]. В литературе имеются разрозненные сведения о возможном существовании ферритин-подобных карго белков, аннотируемых как бактериоферритины и включающих удвоенный ферроксидазный домен, однако системный анализ их многообразия на данный момент отсутствует [4–6].

В нашем исследовании мы провели анализ с помощью НММ-модели, а также моделирование структуры с помощью AlphaFold2 [7] для более, чем 200 белковых последовательностей, аннотированных в базе данных InterPro как ферритины с удвоенным функциональным доменом, построили филогенетическое дерево для известных и изучаемых групп ферритинов и определили эволюционное родство крупного кластера бактериальных целевых белков с группой бактериоферритинов. Для изучаемого кластера FLP-белков мы провели анализ геномного окружения и в ряде случаев определили наличие генов инкапсулинов, а также оперонную структуру, характерную для генов инкапсулинов 1 типа [6,8]. Проведение поиска мотивов связывания с инкапсулинами, а также моделирование комплексов инкапсулин-карго подтвердили наличие способности к связыванию для части исследуемых белков. Наконец, в группе исследуемых белков было выделено несколько подкластеров в зависимости от наличия сайта связывания с инкапсулинов и наличия соответствующей оперонной структуры. Проведенный анализ позволил выделить подкласс карго-белков для инкапсулинов типа 1 с новыми, уникальными характеристиками, и внести вклад в исследование возможностей упаковки карго инкапсулинами, что является важным для прикладных исследований, связанных с использованием мультимеров инкапсулинов типа 1 как наночастиц для таргетной доставки.

1. J.A.Jones, M.P.Andreas, T.W.Giessen (2014) Structural basis for peroxidase encapsulation inside the encapsulin from the Gram-negative pathogen *Klebsiella pneumoniae*, *Nat Commun*, **15**(1):2558.
2. C.-Y.Xing et al. (2020) A self-assembled nanocompartment in anammox bacteria for resisting intracellular hydroxylamine stress, *Science of The Total Environment*, **717**:137030.
3. E.Eren et al. (2022) Structural Characterization of the *Myxococcus xanthus* Encapsulin and Ferritin-Like Cargo System Gives Insight into its Iron Storage Mechanism, *Structure*, **30**(4): 551-563.
4. E.Eren et al. (2024) Encapsulated Ferritin-like Proteins: A Structural Perspective, *Biomolecules*, **14**(6):624.
5. T.W.Giessen, P.A.Silver (2017) Widespread distribution of encapsulin nanocompartments reveals functional diversity, *Nat Microbiol*, **2**(6):1–11.
6. M.P.Andreas., T.W.Giessen (2021) Large-scale computational discovery and analysis of virus-derived microbial nanocompartments, *Nat Commun*, **12**(1):4748.
7. J.Jumper et al. (2021) Highly accurate protein structure prediction with AlphaFold, *Nature*, **596**(7873): 583–589.
8. A.J.Jones, R. Benisch, T.W. Giessen (2023) Encapsulin cargo loading: progress and potential, *Journal of Materials Chemistry B*, **11**(20):4377–4388.

**Изменение соотношения 3p- и 5p-форм микроРНК во
внеклеточных везикулах, секретируемых мезенхимными
стромальными/стволовыми клетками человека,
культивируемыми в профибротических условиях**

М.С. Арбатский ¹, Д.Е. Баландин ¹, Н.А. Басалова ², М.А.
Виговский ², А.В. Чуров ¹, А.Ю. Ефименко ²

¹ ФГБУ ВО «РНИМУ имени Н.И. Пирогова» Минздрава России
(Пироговский Университет), ОСП «РГНКЦ», Институт изучения
старения, Москва, Россия
algenubi81@gmail.com

² МГУ имени М.В. Ломоносова, Медицинский научно-образовательный институт, Москва, Россия

Фиброз — это патологический процесс, возникающий из-за хронического повреждения тканей и сопровождающийся избыточным накоплением внеклеточного матрикса, что в итоге может приводить к нарушению функций органов. Вместе с этим одной из серьезных проблем в медицине является отсутствие эффективных методов лечения фиброз-ассоциированных патологий, что приводит к высокой заболеваемости и смертности по всему миру. С возрастом риск развития фиброза увеличивается, так как в тканях накапливаются повреждения, изменяется внутриклеточная регуляция генов и нарушается баланс между процессами клеточной регенерации и фиброзом, что может отражаться в изменении экспрессии микроРНК, так как они выступают основным механизмом тонкой регуляции экспрессии генов. В этом контексте изучение микроРНК при фиброзе полезно как со стороны понимания процесса развития заболевания, так и со стороны обнаружения потенциальных терапевтических мишеней. Ранее уже отмечались изменения в профиле экспрессии микроРНК при разных заболеваниях, включая и фибротические заболевания. Однако мало уделяется внимания степени вовлеченности 3p- и 5p-форм одной и той же микроРНК в развитие патологических процессов, опосредованных переносом этих (некодирующих РНК) нкРНК в составе внеклеточных везикул (ВВ) при межклеточной коммуникации. В связи с чем в данной работе мы сосредоточились на изучении изменений в регуляции биогенеза 3p- и 5p-форм микроРНК, секретируемых в составе ВВ мезенхимными стромальными/стволовыми клетками (МСК), культивируемыми в нормальных и профибротических условиях, с помощью секвенирования малых РНК (miRNA-seq) в ВВ-МСК. Мы также провели сопоставление полученных данных с транскриптомными профилями субпопуляций МСК, по-разному отвечающими на профибротические стимулы, выявленными по данным scRNA-seq.

Для определения профиля микроРНК было проведено miRNA-seq ВВ-МСК жировой ткани человека, культивируемых в стандартных и профибротических условиях, в результате которого

было определено 479 миРНК, среди которых были найдены 223 дифференциально экспрессирующихся микроРНК (ДЭмиРНК), разница экспрессии которых составляла более 1.75 раза. Для них было установлено, являются ли они 3p- или 5p-формами. Была обнаружена пониженная экспрессия miR-29b-1-5p в профибротических условиях, сниженное содержание которой вносит существенный вклад в развитие фиброза. Среди всех ДЭмиРНК были найдены 23 пары 3p и 5p-форм миРНК (let-7c, miR-9, miR-18a, miR-23b, miR-30b, miR-136, miR-154, miR-195, miR-299, miR-302a, miR-337, miR-365b, miR-370, miR-382, miR-505, miR-549a, miR-615, miR-664a, miR-671, miR-708, miR-766, miR-1287, miR-3613), экспрессия которых значимо различается в ВВ-МСК в стандартных и профибротических условиях.

Полученные пары миРНК можно разделить на четыре группы в зависимости от представленности отдельных форм 3p/5p: 1 Up/Up, 2 Up/Down, 3 Down/Up и 4 Down/Down, где Up означает повышенное содержание соответствующей формы в ВВ-МСК, культивированных в профибротических условиях по сравнению с контролем, а Down - пониженное. Исходя из характера регуляции каждой формы для конкретной микроРНК, их можно обозначить как дисрегулируемые - если 3p- и 5p-формы имеют различную представленность в профибротических условиях, к которым относятся микроРНК из 2 (miR-9, miR-30b, miR-299, miR-370, miR-549a) и 3 (let-7c, miR-18a, miR-382, miR-664a, miR-671, miR-766, miR-1287) групп, и со-регулируемые - если обе формы микроРНК совпадают по характеру изменений представленности в профибротических условиях, что характерно для 1 (miR-302a и miR-3613) и 4 (miR-23b, miR-136, miR-154, miR-195, miR-337, miR-365b, miR-505, miR-615, miR-708) групп.

Для каждой микроРНК были определены списки мишеней по базе данных MirNet, которые сопоставляли с транскриптомными профилями субпопуляций МСК, по-разному отвечающих на профибротические стимулы при культивировании в таких же условиях, выявленными по данным scRNA-seq. В результате микроРНК разделили на 3 группы в зависимости от характеристик совпавших генов-мишеней: ассоциированные с профибротическими генами (miR-18a-5p - MYLK, miR-18a-3p - COL4A1; miR-23b-3p - DSTN; miR-136-3p - MYLK; miR-382-5p -

TM4SF1; miR-505-3p - COL4A1; miR-708-3p - TM4SF1), ассоциированные с антифибротическими генами (let-7c-5p - MARCKSL1 и COLEC12; miR-299-5p - SOX4; miR-615-3p - MARCKSL1 и VIM; miR-708-5p - MARCKSL1; miR-1287-5p - VIM, miR-1287-3p - SOX4; miR-3613-3p - VIM и SOX4) и группа смешанного воздействия (miR-9-5p, miR-30b-5p, miR-30b-3p), мишени которых экспрессируются в разных субпопуляциях МСК. Для большинства микроРНК обе ее формы совпадают по репертуару мишеней, однако 3p-форма miR-708 связана с антифибротическим геном MARCKSL1, а ее 5p-форма с профибротическим TM4SF1, что говорит о сложном характере регуляции, связанном с соотношением 3p/5p форм одной микроРНК.

Важно отметить, что в профибротических условиях большая часть микроРНК, ассоциированных с генами факторов, препятствующих развитию фиброза, демонстрирует повышенную экспрессию относительно контрольных условий, а большинство микроРНК, подавляющих экспрессию профибротических генов, наоборот имеют пониженную экспрессию.

Исходя из полученных данных, видно, что ответ МСК на профибротические стимулы связан со сложными изменениями в профиле экспрессии 3p- и 5p-форм микроРНК, секретируемых этими клетками в составе ВВ. Так, наблюдается как согласованное изменение экспрессии обеих форм, так и дисрегуляция их биогенеза. При этом обнаружена закономерность увеличения экспрессии микроРНК, подавляющих экспрессию антифибротических генов, что обуславливается основным механизмом регуляции микроРНК, связанным с посттрансляционной репрессией мРНК целевого гена. В свою очередь, повышенное содержание таких микроРНК в составе ВВ может запускать механизм вовлечения соседних клеток в патологический процесс фиброзирования тканей, модулируя их фенотип в сторону профибротического. Определение характера действия отдельных форм микроРНК позволяет уточнить механизмы, лежащие в основе развития фиброза, и определить какие элементы потенциально могут выступать в качестве мишеней для терапии, направленной на лечение фиброза.

Исследование поддержано грантом РФФ 19-75-30007.

1. S. O'Reilly (2016) MicroRNAs in fibrosis: opportunities and challenges, *Arthritis Res Ther*, **18**.
2. O. Grigorieva et al. (2024) Modeling the profibrotic microenvironment in vitro: Model validation, *Biochem Biophys Res Commun*, **733**.
3. E. van Rooij et al. (2008) Dysregulation of microRNAs after myocardial infarction reveals a role of miR-29 in cardiac fibrosis, *Proc Natl Acad Sci*, **105**:13027–13032.

Диагональная интеграция мультиомиксных данных

М.С. Арбатский, Д.Е. Баландин

ФГБУ ВО «РНИМУ имени Н.И. Пирогова» Минздрава России
(Пироговский Университет), ОСП «РГНКЦ», Институт изучения
старения, Москва, Россия, algenubi81@gmail.com

С момента появления первых технологий получения омиксных данных прошло несколько десятилетий, за которые было разработано множество новых, более совершенных методов, и интерес к этой области научной деятельности продолжает расти, что привело к накоплению большого объема биологических данных разных модальностей, включая геномные, транскриптомные, протеомные и метаболомные.

Одной из причин постоянного роста объёма таких данных стала престижность использования собственных данных и указание получения новых данных, как одной из задач исследования при написании заявок на его финансирование.

Одновременно с этим начала формироваться идея о повышении эффективности использования уже существующих данных. Исследователи начали искать закономерности и корреляции между значениями параметров в разных типах данных. Начали появляться первые решения для объединения данных различных модальностей в различных форматах (онлайн-платформы, установочные программы, пакеты, алгоритмы и т.д.).

Вместе с этим требовалась и разработка новых математических методов и подходов для реализации этой задачи.

Основными математическими методами и подходами к интеграции разномодальных биологических данных являются метод слияния, метод модели и метод трансформации. В методе слияния происходит объединение матриц с омиксными данными в одну общую матрицу до момента создания модели. В методе модели для каждой матрицы с данными происходит создание промежуточных моделей, которые затем интегрируются в итоговую модель. В методе трансформации каждый датасет сначала трансформируется в графы или матрицы нулевого пространства, затем происходит их объединение перед созданием модели.

Современные методы интеграции можно рассматривать по-разному в зависимости от критериев классификации. По стратегии слияния данных, методы интеграции разделяются на ранние, промежуточные и поздние, также распространена классификация по внутренней природе мультиомиксных экспериментов - вертикальная, горизонтальная и диагональная.

При вертикальной интеграции "якорем", связывающим несколько экспериментов является источник биологического материала или образец. Применяется в экспериментальных схемах, где несколько модальностей данных одновременно анализируются в одних и тех же образцах (сопоставимые анализы). Примерами являются такие методики, как одноклеточное секвенирование метилома и транскриптома (scM&T-seq), одноклеточный анализ генотипа, экспрессии и метилирования (sc-GEM), клеточная индексация транскриптомов и эпитопов с помощью секвенирования (CITE-seq) и секвенирование доступности хроматина и экспрессии РНК на уровне отдельных ядер (SNARE-seq и SHARE-seq).

При горизонтальной интеграции "якорем" является тип биомолекул для которых получены данные. Используется в экспериментальных схемах, где одна и та же модальность данных профилируется для независимых образцов или групп клеток (несопоставимые анализы). Примером может быть секвенирование РНК одиночных клеток (scRNA-seq), где профиль клеток из одной ткани проводится для различных групп доноров или данные объединяются между

разными технологиями scRNA-seq, причем анализы связываются их общим набором генов.

Принципиально отличается от двух предыдущих подходов диагональная интеграция, при которой выбор “якоря” затруднен и может меняться в зависимости от целей исследования. Используется в экспериментальных дизайнах, когда источники материала и тип биомолекул, для которых были получены данные, различаются между экспериментами. Примером может быть применение scRNA-seq и анализа доступности хроматина в одиночных клетках с помощью секвенирования (scATAC-seq) для отдельных групп клеток.

Для диагональной интеграции из математических методов в основном используются выравнивание многообразий (метод слияния) и глубокие автоэнкодеры.

В диагональной интеграции, как и в других методах интеграции существует несколько существенных проблем. К ним относятся гетерогенность, недостаточность данных, выбросы, дисбаланс классов, переобучение и преобладания количества признаков над количеством данных.

Задачей нашего проекта является разработка нового подхода к диагональной интеграции данных. Преимуществом будет использование собственных данных, полученных в рамках различных исследований - WGS (585 vcf файлов от 200 долгожителей, 385 участников исследования Russ-Age), транскриптом (6 scRNA-seq, 2 miRNA-seq, 8 bulk RNA-seq), протеом (4 масс спекта), метаболом (53 человека старше 95 лет), липидом (818 человек исследования Russ-Age). Это позволит объяснить артефакты, возникающие в подобных исследованиях.

1. M. Picard et al. (2021) Integration strategies of multi-omics data for machine learning analysis, *Comput Struct Biotechnol J*, **22**:735–3746.
2. P.S. Reel et al. (2021) Using machine learning approaches for multi-omics data analysis: A review, *Biotechnol Adv*, **49**.

Generative modeling of mRNA untranslated regions with cold diffusion

E.O. Aristova¹, Arsenii Zinkevich¹, Ivan Kulakovskiy², Dmitry Penzar^{2,3,4}

¹*Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 119991, Moscow, Russia, elizavetanoskova@gmail.com*

²*Institute of Protein Research, 142290, Pushchino, Russia*

³*Vavilov Institute of General Genetics, 119991, Moscow, Russia*

⁴*Institute of Translational Medicine, Pirogov Russian National Research Medical University, Moscow, Russia,*

The 5' and 3' untranslated regions (UTRs) of mRNA influence translation efficiency and stability, with regulatory elements that may act in a cell type-specific manner. Rational design of UTRs is crucial for improving mRNA vaccines' efficacy and safety, as insufficient tissue specificity can lead to off-target effects and toxicity.

Massively parallel reporter assays (MPRAs) enable high-throughput quantification of regulatory sequence activity, allowing for training deep neural network (DNN) models of regulatory sequences. DNNs were successfully applied to designing UTRs to enhance mRNA stability, yet achieving cell type-specificity remains a challenge. We present a novel computational approach for engineering UTRs with tailored cell type-specificity, using DNN predictor and generator models, with the latter built upon a Cold Diffusion approach [1]. To train the models, we used MPRA results for 60,000 5' and 3' UTRs from 5 different human cell lines.

MPRA validation of ~16,000 generated sequences revealed a significant correlation between the desired and achieved expression levels. Further, diffusion-generated sequences achieved cell line-specific activity for most of the tested cell line pairs for both 3' and 5'UTRs. Notably, designed sequences covered a larger area of latent embedding space compared to the natural genomic sequences or sequences

generated with a genetic algorithm or motif seeding, highlighting the models' ability to explore sequence space.

We thank Matvei Khoroshkin, Sean B. Lee, Tabea Mittmann, Karoline Manegold, and Hani Goodarzi (UCSF) for orchestrating the study and for providing MPRA experimental data.

1. B. Arpit et. al. "Cold Diffusion: Inverting Arbitrary Image Transforms Without Noise." arXiv, 2022.

Коинфекция *AMFV* и *Nosema* sp. пчел *Apis mellifera*

А. Ф. Асадуллин¹, Г. А. Кашенко², А. Х. Тальдаев³, Д. В. Смутин⁴,
Л. С. Адонин⁴

¹Санкт-Петербургский государственный университет, Санкт-Петербург, Россия, asadullin@aptyr.ru

²Факультет почвоведения, геологии и ландшафтоведения, РГАУ-МСХА им. К. А. Тимирязева, Москва

³Институт биомедицинской химии имени В.Н. Ореховича, Москва

⁴Санкт-Петербургский государственный университет телекоммуникаций имени профессора М. А. Бонч-Бруевича, Санкт-Петербург, Россия

В последние десятилетия основными объектами интереса, направленного на изучение патогенов членистоногих, всё чаще становятся ДНК-вирусы. Недавние исследования показали, что филаментозный вирус медоносной пчелы – *Apis mellifera Filamentous Virus (AmFV)*, известный с середины XX века и являющийся наименее патогенным, часто ассоциирован с иными возбудителями, в особенности с представителями рода *Nosema*. *AmFV* может быть обнаружен во всех основных органах и тканях пчел, включая жировое тело, репродуктивную и пищеварительную

системы [1, 2]. Кроме того, этот вирус играет значительную роль в повышении уровня смертности ульев в зимний период. *AmFV* предположительно является частью комплекса возбудителей colony collapse disorder (CCD), которые в совокупности приводят к потере ~40% (80% от всех гибнущих семей) пчелиных ульев ежегодно [3]. Выявлены группы патогенов, чаще встречаемых вместе, чем по отдельности. Интересно, что вирус *AmFV* встречается практически независимо от других патогенов. С другой стороны, многие патогены встречаются чаще при наличии *AmFV* вируса, чем без него. Наибольшее распространение среди этих патогенов у микроспоридий рода *Nosema*, в сумме отмеченных в 24% образцов. Уровень заражения *AmFV* составил 56%, что значительно превосходит предыдущие оценки. При этом, коинфекция *AmFV* и *Nosema* наблюдается в 8 раз чаще с *N. ceranae* ($p < 10^{-16}$), с *N. apis* - в 2 раза ($p < 10^{-4}$).

Также в настоящем исследовании мы составили математическую модель инфицирования колонии пчел двумя патогенами на основе SIR [4]. С помощью нее удалось проверить возможность элиминирования *Nosema* за счет инфицирования дополнительного оппортуниста, который в совокупности с микроспоридиями давал синергичный эффект по уничтожению зараженных особей. Таким образом коинфицированные особи погибают быстрее, чем успевают значимо распространить патоген, за счет чего *Nosema* (в случае SIR-модели – сущности, больные *Nosema*) элиминируются, а *AmFV* увеличивает свою популяцию. Полученные результаты свидетельствуют о том, что патогенный комплекс *AmFV*–*Nosema* реализует двунаправленный эффект: резкое повышение летальности коинфицированных индивидов ограничивает распространение микроспоридий внутри улья, но одновременно создает условия для самоподдерживающейся циркуляции вируса и снижает устойчивость колонии к внешним стрессорам.

Применение математического моделирования и других биоинформатических инструментов позволяет предложить инновационный подход к биологическому контролю патогенов пчел: целенаправленное введение *AmFV* может ограничивать эпидемический потенциал *Nosema*, сохраняя при этом вирусный патоген на уровне, не угрожающем колонии. Экспериментальная

валидация синергической летальности и параметров конкурентного взаимодействия станет следующим шагом в разработке интегрированных стратегий защиты медоносных пчёл.

1. Hartmann U, Forsgren E, Charrière JD, Neumann P, Gauthier L. Dynamics of *Apis mellifera* Filamentous Virus (AmFV) Infections in Honey Bees and Relationships with Other Parasites. *Viruses*. 2017 7(5):2654-67.
2. Gauthier, L., Cornman, S., Hartmann, U., Cousserans, F., Evans, J.D., de Miranda, J.R. and Neumann, P. (2015) 'The *Apis mellifera* Filamentous Virus Genome', *Viruses*, 7(7), pp. 3798–3815.
3. Atanasov, A.Z., Georgiev, S.G. and Vulkov, L.G. (2021) 'Reconstruction analysis of honeybee colony collapse disorder modeling', *Optimization and Engineering*, 22(4), pp. 2481–2503.
4. Weiss, Howard Howie. "The SIR model and the foundations of public health." *Materials matematics* (2013): 0001-17.

Genomics-based reconstruction of inulin and glycogen degradation pathways in the human gut microbiome

German A. Ashniev^{1,2}, Dmitry A. Rodionov³, Mikhail S. Gelfand⁴

¹*Institute for Information Transmission Problems, Moscow, Russia*

²*Vavilov Institute of General Genetics, Moscow, Russia*

³*Sanford-Burnham-Precy Medical Discovery Institute, La Jolla, CA, USA*

⁴*Skoltech Center for Molecular and Cellular Biology, Moscow, Russia*

ashnievh@gmail.com, rodionov@burnham.org, m.gelfand@skoltech.ru

The human gut microbiome (HGM) represents a dynamic and complex ecosystem of microorganisms that plays a fundamental role in host physiology, influencing both health and disease states. A critical function of these microbial communities is their capacity to metabolize dietary and host-derived carbohydrates, a process that directly shapes microbial ecology and gastrointestinal homeostasis. Plant-based foods, particularly fruits and vegetables, serve as rich sources of fermentable

polysaccharides, which have demonstrated therapeutic potential in managing gut-related disorders such as inflammatory bowel disease (IBD), irritable bowel syndrome (IBS), and Crohn's disease. Among these carbohydrates, inulin and glycogen emerge as key substrates with distinct metabolic roles.

Inulin, a soluble dietary fiber abundant in plants like chicory root and garlic, functions as a prebiotic by selectively stimulating beneficial gut bacteria, including *Bifidobacterium* and *Lactobacillus* species. Its fermentation yields short-chain fatty acids (SCFAs), particularly butyrate, propionate, and acetate, which are crucial for maintaining gut barrier integrity, modulating immune responses, and suppressing inflammation. Glycogen, the primary glucose storage polymer in animals, is predominantly host-derived but also serves as an energy source for certain gut microbes, facilitating metabolic cross-feeding interactions that enhance microbial community stability and function.

To systematically investigate the genetic basis of these metabolic processes, we performed a genome-scale reconstruction of biochemical pathways involved in inulin and glycogen degradation across the gut microbiota. Leveraging a subsystem-based annotation approach, we mapped enzymatic functions to specific bacterial genomes, enabling the prediction of metabolic capabilities and identification of pathway gaps. Our analysis encompassed 3,000 reference genomes spanning over 800 known HGM species, providing a comprehensive view of carbohydrate utilization potential.

Beyond taxonomic profiling, we quantified the functional capacity of microbial communities using the Community Phenotype Index (CPI), a novel metric that integrates both the genetic potential of individual taxa (via Phenotype Indices, PI) and their relative abundances within a sample (Fig. 1).

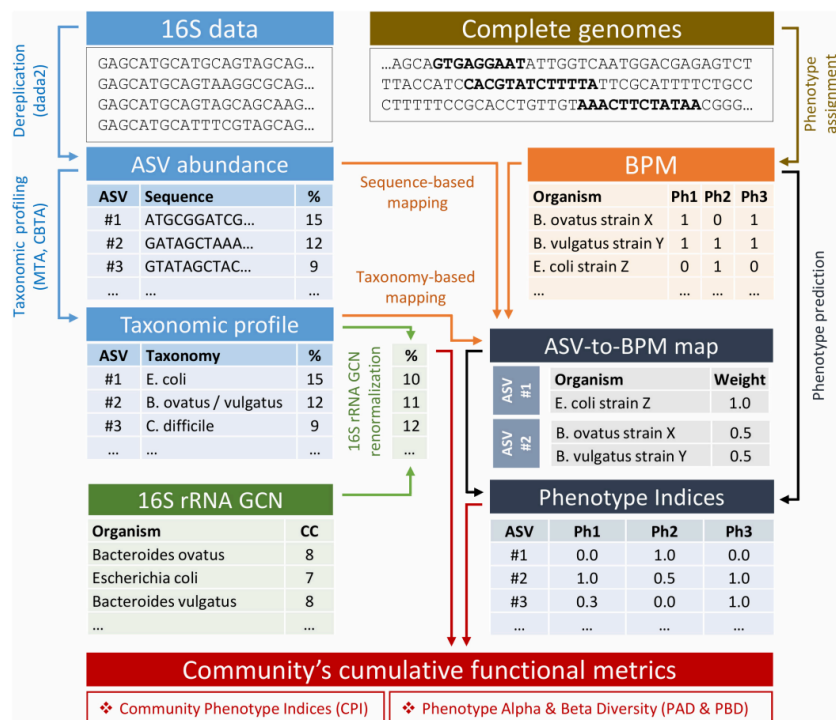


Figure 1. Pipeline for CPI accession (Iablokov et. al, 2019)

Applying this framework to 16S rRNA metagenomic data, we compared CPI values between healthy individuals and patients with gut disorders, revealing significant shifts in carbohydrate-utilizing bacteria. These findings not only highlight the mechanistic links between microbiome function and disease but also demonstrate the utility of CPI as a tool for bridging genomic potential with ecosystem-level functionality. By quantifying the collective metabolic traits of microbial communities, our approach provides a pathway-resolved perspective on microbiome-host interactions.

The work was partially supported by FFRW-2024-0004.

Реконструкция метаболических путей деградации полисахаридов бактериями кишечника рода *Prevotella*

Генрих Альфредович Ашниев¹, Михаил Сергеевич Гельфанд^{1,2,3},
Герман Альфредович Ашниев^{1,2,3,4}, Дмитрий Александрович
Родионов¹

¹*Институт проблем передачи информации имени А. А. Харкевича
РАН, Москва, Россия Ashnievhenry@gmail.com*

²*Национальный исследовательский университет «Высшая школа
экономики», Москва, Россия*

³*Сколковский институт науки и технологий, Москва, Россия*

⁴*Институт общей генетики им. Н.И. Вавилова РАН*

Полисахариды, такие как инулин, пектин, ксилан, ламинарин и хитин, входят в состав рациона и способны влиять на микробиоту человека, следовательно оказывают влияние на общее благосостояние организма носителя. Процесс ферментации полисахаридов комменсальными бактериями кишечника, сопровождается различными факторами, которые могут как поддерживать так и дестабилизировать иммунный и метаболический баланс в организме [1].

Изучение метаболических путей деградации полисахаридов бактериями кишечника человека является комплексной задачей, в которую входят анализ референсных коллекций полногеномных последовательностей бактерий, метаболомика на основании встречаемости необходимых для ферментации белков и регуляторных элементов, а также метаболомика с целью определения фенотипа бактерий референсной коллекции. Реконструкция путей деградации полисахаридов позволяет уточнить и расширить понимание механизмов утилизации и функциональной значимости рода *Prevotella* для деградации пищевых полисахаридов в толстой кишке человека [2].

Повышенная частота встречаемости бактерий рода *Prevotella* при анализе состава комменсальной микробиоты человека является биомаркером особого типа диеты с повышенным содержанием простых и сложных пищевых полисахаридов. Также отмечается положительная корреляция встречаемости бактерий

рода *Prevotella* при возникновении дисбиозов и других расстройств кишечника [3]. Цель исследования состоит в том, чтобы использовать сабсистемный подход для аннотации геномов представителей микробного сообщества кишечника для последующей реконструкции метаболических путей деградации различных полисахаридов таких, как инулин, пектин, ксилан, ламинарин и хитин.

В ходе данного исследования сабсистемный подход был применен к 314 полногеномным последовательностям бактериальных штаммов 12 видов бактерий рода *Prevotella* населяющих кишечник человека. Реконструированные пути деградации полисахаридов включают как одно-, так и многоступенчатые ферментативные реакции утилизации пищевых полисахаридов. В ходе анализа было установлено, что основным из рассматриваемых субстратом являются инулин и пектины. Таким образом, проведенная реконструкция метаболических путей деградации полисахаридов бактериями рода *Prevotella* предоставляет информацию для глубокого понимания механизма ферментации пищевых полисахаридов.

Результаты проведенного анализа могут найти применение в клинической практике, особенно в области профилактики заболеваний. Что позволит клиническим специалистам проводить разработку эффективных методов по оптимизации кишечной микробиоты и улучшению общего состояния здоровья пациентов.

Список литературы:

1. Gan L, et al., (2022) Polysaccharides influence human health via microbiota-dependent and -independent pathways. *Front. Nutr.* 9:1030063. doi: 10.3389/fnut.2022.1030063
2. Flint, H., et al. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol* 6, 121–131 (2008). doi.org/10.1038/nrmicro1817
3. G. Precup and D.-C. Vodnar, “Gut *Prevotella* as a possible biomarker of diet and its eubiotic versus dysbiotic roles: a comprehensive literature review,” *British Journal of Nutrition*, vol. 122, no. 2, pp. 131–140, 2019. doi:10.1017/S0007114519000680

Финансирование: Работа выполнена при поддержке гранта РФФ 24-14-00276

**Ведущая роль ассоциированного с репликацией
асимметричного мутагенеза в формировании GC-skew у
бактерий**

Э. А. Бадамшин¹, А. В. Скуднов¹, К. Ю. Попадьян¹, К. В. Гунбин²,
С. В. Денисов³

¹*Балтийский Федеральный Университет, ул.
Александра Невского, 14, Калининград, 236041, Россия
badamshin124@gmail.com*

²*Институт молекулярной и клеточной биологии СО РАН, просп.
акад. Лаврентьева, 8/2, Новосибирск, 630090, Россия*

³*Манчестерский университет, Манчестер, Великобритания*

В большинстве бактериальных геномов нуклеотидный состав лидирующей и отстающей цепей репликации не одинаков. На лидирующей цепи ДНК наблюдается повышенная доля гуанина по сравнению с тиминем (GC-skew) и в меньшей мере тимина относительно аденина (AT-skew). Это используется для компьютерных предсказаний точек начала и конца репликации [1].

Относительно причин GC-skew существуют разные точки зрения. Их наличие может объясняться как асимметричным мутагенезом, так и давлением отбора на функцию белка, особенно в первых и вторых позициях кодонов, в купе с неравномерным распределением генов на лидирующих и отстающих цепях репликации ДНК бактерий [2]. Чтобы определить степень влияния данных факторов на GC-skew, мы проанализировали экспериментальные данные по накоплению мутаций (MAE – mutation accumulation experiments) в 7 видах бактерий (всего 10 бактериальных линий дикого типа и 5 линий с дефицитом системы репарации неспаренных оснований) из различных таксонов [3–10]. С использованием специфичного для разных цепей ДНК 12-компонентного мутационного спектра нами была обнаружена асимметрия частоты накопления комплементарных мутаций внутри ДНК: GT-богатая лидирующая цепь характеризовалась

повышенной долей мутаций C>T и A>G по сравнению с комплементарными G>A и T>C. Данная закономерность наблюдалась как на основных хромосомах, так и на хромиде. Хотя сам дизайн МАЕ предполагает почти полное отсутствие отбора в любых участках генома, мы дополнительно удостоверились, что мутационный спектр в межгенных регионах и третьих позициях четырежды вырожденных кодонов не отличается от общегеномного. На основании этого было предположено, что данная асимметрия в распределении мутаций и есть главная причина GC-skew.

Для проверки данной гипотезы были проанализированы GC-skew в регионах, находящихся под разным давлением отбора (1-е, 2-е и 3-и позиции кодонов, межгенные участки, третьи позиции четырежды вырожденных кодонов), после чего было проведено их сравнение с равновесными частотами нуклеотидов, рассчитанных на основании 12-компонентных мутационных спектров. С уменьшением давления отбора (например, от 2 позиций кодонов к 3-им) наблюдалось увеличение GC-skew, который достигал максимума в гипотетическом состоянии мутационно-дрейфового равновесия (полное отсутствие отбора, расчетное состояние, которое не наблюдалось на природных ДНК), при этом также наблюдалось снижение CT-skew. Хотя асимметрия в накоплении мутаций была равномерно распределена вдоль бактериальной хромосомы, общая частота мутаций увеличивалась от точки начала к точке окончания репликации.

Сходство мутационных спектров, построенных по генам, лежащим на лидирующей и отстающей цепях, указывает на то, что связанный с репликацией асимметричный мутагенез, а не репарация, сопряженная с транскрипцией, является наиболее вероятной причиной асимметрии. Таким образом, наши результаты согласуются с гипотезой, объясняющей GC-skew разницей во времени, которое лидирующая и отстающая цепи ДНК проводят в одноцепочечном состоянии в ходе процесса репликации [1].

Исследование выполнено при финансовой поддержке Российского Научного Фонда (RSF № 21-75-20143).

1. Lobry JR, Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biol* 3:research0058.1. <https://doi.org/10.1186/gb-2002-3-10-research0058>
2. Karlin S (1999) Bacterial DNA strand compositional asymmetry. *Trends Microbiol* 7:305–308. [https://doi.org/10.1016/S0966-842X\(99\)01541-3](https://doi.org/10.1016/S0966-842X(99)01541-3)
3. Dillon MM, Sung W, Sebra R, et al (2017) Genome-Wide Biases in the Rate and Molecular Spectrum of Spontaneous Mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Mol Biol Evol* 34:93–109. <https://doi.org/10.1093/molbev/msw224>
4. Wei W, Xiong L, Ye Y-N, et al (2018) Mutation Landscape of Base Substitutions, Duplications, and Deletions in the Representative Current Cholera Pandemic Strain. *Genome Biol Evol* 10:2072–2085. <https://doi.org/10.1093/gbe/evy151>
5. Foster PL, Niccum BA, Popodi E, et al (2018) Determinants of Base-Pair Substitution Patterns Revealed by Whole-Genome Sequencing of DNA Mismatch Repair Defective *Escherichia coli*. *Genetics* 209:1029–1042. <https://doi.org/10.1534/genetics.118.301237>
6. Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci* 109:E2774–E2783. <https://doi.org/10.1073/pnas.1210309109>
7. Sung W, Ackerman MS, Gout J-F, et al (2015) Asymmetric Context-Dependent Mutation Patterns Revealed through Mutation–Accumulation Experiments. *Mol Biol Evol* 32:1672–1683. <https://doi.org/10.1093/molbev/msv055>
8. Dillon MM, Sung W, Lynch M, Cooper VS (2015) The Rate and Molecular Spectrum of Spontaneous Mutations in the GC-Rich Multichromosome Genome of *Burkholderia cenocepacia*. *Genetics* 200:935–946. <https://doi.org/10.1534/genetics.115.176834>
9. Long H, Kucukyildirim S, Sung W, et al (2015) Background Mutational Features of the Radiation-Resistant Bacterium *Deinococcus radiodurans*. *Mol Biol Evol* 32:2383–2392. <https://doi.org/10.1093/molbev/msv119>
10. Kucukyildirim S, Long H, Sung W, et al (2016) The Rate and Spectrum of Spontaneous Mutations in *Mycobacterium smegmatis*, a Bacterium Naturally Devoid of the Postreplicative Mismatch Repair

Мультиомный анализ молекулярных изменений при тройном негативном раке молочной железы

А. В. Барисёнок¹, А. Ю. Горбачёв²

¹*Московский физико-технический институт (национальный исследовательский университет) barisenok.av@phystech.edu*

²*ФНКЦ физико-химической медицины им. Ю. М. Лопухина*

Рак молочной железы остается одним из наиболее распространенных онкологических заболеваний у женщин, а его тройной негативный подтип характеризуется особенно агрессивным клиническим течением и ограниченными терапевтическими опциями. Мультиомные технологии, включая протеомику, транскриптомику и экзомное секвенирование, позволяют проводить комплексную молекулярную характеристику опухоли. Целью настоящего исследования является выявление устойчивых молекулярных изменений и биологических путей, вовлечённых в прогрессию опухолевого процесса.

В рамках исследования использовались данные, полученные из образцов пациентов с ТНРМЖ. В качестве нормальных образцов рассматривалась прилегающая к опухоли здоровая ткань. Для выделения значимо экспрессируемых генов использовалось ограничение в adjusted p-value < 0.05, для путей дополнительно накладывалось $|\log_2FC| > 1$.

Протеомные данные включали 30 опухолевых и 5 нормальных образцов, обработанных с использованием nfc-score/quantms [1, 2]. Идентифицировано 6524 уникальных белка. Для 5 пар «опухоль-норма» проведен статистический анализ в MSstats [7], в результате которого выявлено 16 белков с достоверной дифференциальной экспрессией. Анализ с использованием enrichR выявил 154 активированных пути. Дополнительно проведен попарный анализ для 5 пар образцов с помощью limma-voom [10],

выявлено 2573 белка, соответствующих 5280 сигнальным путям. Распределение путей позволило выделить один образец, демонстрирующий выраженное отличие от остальных.

РНК-секвенирование проанализировано для 25 опухолевых образцов с использованием пайплайна nf-core/rnaseq [1, 3]. В качестве контрольной группы применены экспрессионные профили 89 нормальных образцов из базы по TНРМЖ [9]. С помощью DESeq2 [8] выявлено около 14 тыс. генов с достоверными изменениями экспрессии. Сравнение каждого опухолевого образца с контрольной группой при помощи limma-voom позволило идентифицировать 1644 пути. По результатам pathway-профилей по образцам выделена группа из 5 опухолей, 4 из которых также продемонстрировали сходство на уровне экзома.

Анализ выполнен для 23 парных образцов с использованием пайплайна nf-core/sarek [1, 4, 5] и аннотирован с помощью regr [6]. Идентифицировано около 158 тыс. соматических мутаций в 15 тыс. генах. Для этих генов определены более 5 тыс. сигнальных путей. Проведена классификация мутаций по типу, признакам потери функции, онкогенности и принадлежности к базам CGC. На основе этих признаков выделены три образца, существенно отличающихся от остальных, один из которых также проявляет уникальность на протеомном уровне.

В рамках мультиомного исследования проведена интеграция результатов транскриптного, протеомного и мутационного анализа с целью выявления молекулярных событий, устойчиво проявляющихся на разных уровнях регуляции. Результаты следующие:

- Выявлено 45 генов и 65 сигнальных путей, демонстрирующих одновременную активность во всех типах данных на совокупности всех образцов.
- Интеграция путей на уровне отдельных пациентов:
 - Определено 203 сигнальных пути, активных на двух из трех уровней с учетом принадлежности к конкретным образцам.
 - Среди них:
 - Extracellular Matrix Organization — активен у 8 пациентов;

- Axon Guidance и Nervous System Development — в 7 образцах.
- Интеграция генов на уровне отдельных пациентов:
 - Найдено 253 гена, пересекающихся хотя бы на двух уровнях при анализе по образцам.
 - В одном из образцов выявлен ген ACTR3, одновременно детектируемый на всех трех уровнях.
 - Гены COL1A1, HRNR, LMNA и SFPQ оказались активированы на двух уровнях в двух разных пациентах.

Проведенное мультиомное исследование позволило получить целостную картину молекулярных изменений при тройном негативном раке молочной железы. Интеграция данных транскриптомики, протеомики и мутационного анализа позволила не только идентифицировать устойчивые биологические процессы, вовлеченные в прогрессию опухоли. Полученные результаты подчеркивают потенциал мультиуровневого анализа для формирования основ для разработки персонализированных подходов к терапии и прогнозированию течения заболевания.

Работа выполнена при поддержке гранта РФФИ #24-64-00028.

1. Ewels P. [et al.]. The nf-core framework for community-curated bioinformatics pipelines // Nat. Biotechnol. 2020. V. 38(3). P. 276–278.
2. Dai C. [et al.]. quantms: a cloud-based pipeline for quantitative proteomics enables the reanalysis of public proteomics data // Nat. Methods. 2024. V. 21. P. 1603–1607.
3. Patel H. [et al.]. nf-core/rnaseq v3.18.0 — Lithium Lynx [Электронный ресурс]. — 2024. — Оpubл. 20.12.2024, версия 3.18.0. — URL: <https://nf-co.re/rnaseq> (дата обращения: 03.03.2025).
4. Hanssen F. [et al.]. Scalable and efficient DNA sequencing analysis on different compute infrastructures aiding variant discovery // NAR Genom. Bioinform. 2024. V. 6(2).
5. Garcia M. [et al.]. Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants // F1000Research. 2020. V. 9. P. 63.
6. Nakken S. [et al.]. Personal Cancer Genome Reporter: variant interpretation report for precision oncology // Bioinformatics. 2017. V.

34(10). P. 1778–1780.

7. Kohler D. [et al.]. MSstats version 4.0: statistical analyses of quantitative mass spectrometry-based proteomic experiments with chromatography-based quantification at scale // J. Proteome Res. 2023. V. 22. P. 1466.

8. Love M. I. [et al.]. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // Genome Biol. 2014. V. 15. P. 550.

9. Национальный институт рака. The Cancer Genome Atlas (TCGA) [Электронный ресурс]. — URL: <https://www.cancer.gov/tcga> (дата обращения: 03.03.2025).

10. Chen A. и др. A comprehensive genomic and transcriptomic dataset of triple-negative breast cancers // Scientific Data. 2022. Т. 9. № 572.

Новые эндогенные ретровирусы в геномах китообразных

О. Д. Богомаз¹, А. А. Жарикова¹, Е. В. Шеваль²

¹119234, г. Москва, ГСП-1, Ленинские горы МГУ, дом 1, стр. 73,
МГУ им. Ломоносова, ФББ, olesjabogomaz@yandex.ru

²119992, Москва, Ленинские горы, дом 1, стр 40, МГУ им.
Ломоносова, НИИ ФХБ им. А. Н. Белозерского,

Эндогенные ретровирусы (ERV) служат важным источником информации об эволюции ретровирусов и их взаимодействии с хозяевами. Несмотря на обширные данные о лентивирусах и бетаретровирусах у наземных млекопитающих [1], их разнообразие у китообразных изучено недостаточно. Для выявления новых ERV проводили поиск гомологов среди млекопитающих через tBLASTn [2] (белки Gag/Pol лентивирусов) с дальнейшей фильтрацией. Находки аннотировали (LTR_Finder [3], Prokka [4]) и сравнивали с известными гамма-ERV китообразных. По результатам аннотации провели филогенетический анализ (PhyML [5]). Обнаружено 22 новых провируса у 4 видов китообразных: *Delphinus delphis* (6 вставок), *Globicephala melas* (1 вставка), *Phocoena phocoena* (7 вставок) и *Stenella coeruleoalba* (12 вставок). Вставки содержат гены gag, pol и env, ограничены LTR

длиной 319-1716 п.н. с высокой симметрией. Филогенетический анализ показал кластеризацию новых провирусов с бетаретровирусами и отличное от известных эндогенных гаммаретровирусов китообразных (KWERV [6], TTEV [7]) геномное окружение у вставок в геномы *Phocoena phocoena*, *Delphinus delphis* и *Stenella coeruleoalba*. Таким образом, мы впервые обнаружили эндогенные бетаретровирусы у китообразных.

Исследование выполнено в рамках государственного задания МГУ имени М. В. Ломоносова.

1. R.J. Gifford (2012) Viral evolution in deep time: lentiviruses and mammals, *Trends in Genetics*, 28(2):89-100.
2. G.M. Boratyn et al. (2013) BLAST: a more efficient report with usability improvements, *Nucleic Acids Research*, doi:10.1093/nar/gkt282.
3. Z. Xu, H. Wang (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons, *Nucleic Acids Research*, 35:W265-W268.
4. T. Seemann (2014) Prokka: rapid prokaryotic genome annotation, *Bioinformatics*, 30(14):2068-2069.
5. S. Guindon et al. (2010) New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0, *Systematic Biology*, 59(3):307-321.
6. Lamere S.A. et al. (2009) Molecular characterization of a novel gammaretrovirus in killer whales. *J Virol*, 83: 12956–12967.
7. L. Wang et al. (2013) Ancient invasion of an extinct gammaretrovirus in cetaceans, *Virology*, 441(1):66-69.

Изучение митохондриальной дисфункции в цибридных моделях методами биохимии, молекулярной биологии и транскриптомного анализа

Е. Е. Борисов¹, А. В. Омельченко²

¹*«Научно-исследовательский институт морфологии человека имени академика А.П. Авцына» Федерального государственного бюджетного научного учреждения "Российский научный центр хирургии имени академика Б.В. Петровского",
Москва, улица Цюрюпы 3
borisovevgenij5@gmail.com*

²*ФГБНУ Научно-исследовательский институт общей патологии и патофизиологии, Москва, Балтийская улица 8*

Мы изучаем вклад митохондриального генома в развитие атеросклероза [1], используя цибридные клеточные линии, содержащие различные митохондриальные геномы, выделенные у пациентов с диагностированным атеросклерозом, и с идентичными ядерными геномами [2]. Наша цель - исследовать влияние состояния мтДНК, предположительно, вызванного накоплением митохондриальных мутаций, на клеточные процессы. Ранее мы выявили дефект митофагии в ряде цибридных линий, в виде отсутствия ожидаемой реакции клетки на стимуляцию митофагии разобщителем дыхания FCCP.

Дефектность митофагии в линии TC-520 была установлена 2 независимыми методами – конфокальной микроскопией и ПЦР в реальном времени с генами-маркерами митофагии (*PINK1*, *PRKN*, *BNIP3*, *NIX*) и аутофагии (*BECN1*, *CTSD*, *LAMP1*, *LAMP2*, *MAP1LC3B*) [3].

В представленной работе мы сравниваем линию TC-520, ранее характеризованную как дефектную по митофагии, и линию TC-522, характеризованную как нормальную по митофагии. Транскриптомный анализ проводили методом РНК-секвенирования на платформе Illumina NovaSeq 6000. Данные обрабатывали с использованием пакетов *hisat2*, *samtools*, *htseq-count* и собственных скриптов на языке R. Уровень активных форм кислорода в

цитоплазме определяли с помощью проточной цитометрии на приборе CytoFLEX, используя специфический флуоресцентный зонд H2DCFDA. Уровень апоптоза анализировали двумя независимыми методами, количественной ПЦР в реальном времени для оценки уровня экспрессии генов-маркеров внутреннего пути апоптоза (*BID*, *BAX*, *BCL2*, *MCL1*) и проточной цитометрией с использованием коммерчески-доступного набора Annexin V-FITC/PI на приборе CytoFLEX для оценки количества апоптотически-гибнущих клеток.

В гибридной линии TC-520 с дефектной митофагией была выявлена пониженная экспрессия ряда генов по сравнению с линией TC-522 с нормальной митофагией, что, как мы предполагаем, может указывать на причину нарушений митофагии. Обе исследуемые линии демонстрируют повышенную экспрессию митохондриальных генов по сравнению с клетками THP-1, что, вероятно, является своего рода компенсаторной реакцией. Биохимические исследования показали, что дефект митофагии в линии TC-520 сопровождается повышенным уровнем АФК, нарушениями регуляции внутреннего пути апоптоза и снижением фагоцитарной активности.

Таким образом, дефект митофагии в линии TC-520, ассоциированный с нарушениями экспрессии митохондриальных генов и приводит к повышению уровня АФК и нарушениям в регуляции фагоцитоза и апоптоза. Полученные нами данные подчеркивают ключевую роль митохондриального генома в поддержании клеточного гомеостаза и выживаемости клеток, что имеет важное значение, как для понимания механизмов митохондриальной дисфункции, так и для разработки потенциальных терапевтических стратегий.

Работа выполнена при поддержке гранта РНФ № 24-65-00027

1. M. A. Sazonova, A. V. Zhelankin, et al. (2015) Mutations of mitochondrial genome in carotid atherosclerosis, *Frontiers in Genetics*, 6.
2. M.A. Sazonova, V.V. Sinyov, et al. (2019) Creation of Cybrid Cultures Containing mtDNA Mutations m.12315G>A and m.1555G>A, Associated with Atherosclerosis, *Biomolecules*, 499.

3. A.N. Orekhov, E. E. Borisov E.E., D.D. Gavrilova, et al. (2025). Defective Mitophagy is Associated with Mitochondrial DNA Mutations. *Current Medicinal Chemistry*, (in press).

**Оценка патогенности вариантов гена ACVRL1 на основе
эффективности связывания АТФ методом молекулярной
динамики**

Боровикова И.И.¹, Упоров И.В.¹, Замятин В.И.², Охрименко Г.С.²,
Попцова М.С.², Данковцева Е.Н.³, Затеицков Д.А.³

¹ФГОУ ВПО «Московский государственный университет имени
М.В.Ломоносова» Москва, Россия
borovikova_irina@bk.ru

²ФГАОУ ВО «Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

³ГБУЗ «ГКБ № 29 им. Н.Э. Баумана ДЗМ»

Для оценки патогенности генетических вариантов применяются критерии ACMG [1], учитывающие тип варианта, частоты встречаемости, положение варианта в последовательности гена. Использование критериев осложняется для заболеваний с низкой пенетрантностью или возникновением симптомов во взрослом возрасте из-за смещения частот встречаемости в контрольной группе здоровых людей.

Одним из возможных подходов к исследованию патогенности вариантов является биоинформатическая оценка влияния несинонимичных замен на взаимодействие белка с другими молекулами [2, 3].

Для исследования использовались данные полногеномного секвенирования 3 больных с клинически очевидной болезнью Рандю-Ослера-Вебера, у которых стандартными методами не было обнаружено патогенных вариантов в генах, ответственных за развитие заболевания (ACVRL1, ENG [4, 5]). Данные

секвенирования были очищены от адаптерных последовательностей и нуклеотидов низкого качества (Trimmomatic [6]), прочтения выравнены с референсным геномом (GRCh38, bwa 0.7.17), промаркированы дубликаты (GATK MarkDuplicatesSpark 4.3.0.0 [7]). Для нахождения однонуклеотидных вариантов использовалась программа DeepVariant 1.4.0 [8], проводилась фильтрация (PASS и quality > 20) и нормализация мультиаллельных вариантов, поиск вариантов, принадлежащих генам ACVRL1 и ENG. Оценка патогенности вариантов была проведена в соответствии с критериями ACMG инструментом InterVar [9]. Структура белка на основе последовательности с внесенной динуклеотидной делецией генерировалась с помощью AlphaFold2 [10]. Стабильность белковых структур исследовалась программным комплексом AMBER24 [11]. Для вычисления структурно-динамических свойств молекул белка полученные траектории обрабатывались с помощью программы CPPTRAJ [12, 13]. Вычисление свободной энергии комплексов белок-лиганд проводилось программой MMPBSA.py [14].

Из 155 обнаруженных в образцах вариантов (табл. 1) 142 оказались доброкачественными (Benign), еще 13 - вариантами неизвестной значимости (VUS). 2 VUS ACVRL1 принадлежат экзонным областям, в ENG экзонных вариантов не обнаружено. Распределение экзонных вариантов по образцам показано в таблице 2.

Таблица 1. Найденные в образцах пациентов варианты ключевых для развития болезни Рандю-Ослера-Вебера генов.

Ген	Benign	VUS	VUS (экзоны)
ACVRL1	73	5	2
ENG	69	8	0

Таблица 2. Обнаруженные в образцах экзонные VUS гена ACVRL1.

Образец	Варианты (ENST00000552678.2:p.HGVS)	Тип варианта
1	p.E224_S225del	Делеция
2	p.S284F	Несинонимичная замена
3	p.E224_S225del	Делеция

Были сгенерированы структуры ACVRL1 с делецией p.E224_S225del и заменой p.S284F. Для нативного и мутированного фрагмента структуры, соответствующему киназному домену, были рассчитаны площадь белка, доступная растворителю (методы LCPO и Конноли), а также параметр RMSF. Значимых отличий обнаружено не было.

Для дальнейшего анализа были созданы модели нативного и мутированного (p.S284F) белков в комплексе с молекулой АТФ. Положение АТФ было определено по результатам структурного выравнивания структуры киназного домена ACVRL1 (PDB:3MY0) с гомологичной серин-треониновой киназой (PDB: 3EQD), кристаллизованной в комплексе с АТФ (рис. 1).



Рисунок 1. Структурное выравнивание 3MY0 (зеленый) и 3EQB (розовый).

Были определены остатки ACVRL1 в области 5Å от молекулы АТФ, в базе данных UniProt был произведен поиск известных патогенных замен этих остатков.

Для подтверждения нами были созданы 6 моделей комплекса киназного домена ACVRL1 с АТФ с внесенными в белок известными патогенными заменами. Значение свободной энергии

связывания (модель Пуассона-Больцмана) для моделей представлены в таблице 3.

Таблица 3. Значения свободной энергии связывания молекулы АТФ (ккал/моль) нативного (дикий тип) и мутированных ACVRL1. Обнаруженная в образце замена выделена жирным.

Замена	Средняя величина (ккал/моль)	Стандартное отклонение	P-value
<i>Дикий тип</i>	<i>-6.3389</i>	<i>9.4414</i>	-
Ser284Phe	24.9278	11.3455	< 0.0001
Gly209Arg	0.1580	8.0989	< 0.0001
Leu257Pro	-15.4315	8.5293	< 0.0001
Lys229Met	-19.4445	14.3139	< 0.0001
Thr277Lys	-69.5185	7.1847	< 0.0001
His280Arg	-16.7511	9.9167	< 0.0001
Thr372Ile	-8.7883	9.4044	0.01

Статистическая значимость различий оценивалась с помощью критерия Стьюдента для независимых выборок с поправкой Уэлча (двусторонний тест). Свободная энергия связывания АТФ мутантных белков значимо отличается от энергии нативного белка. Замены Leu257Pro, Lys229Met, Thr277Lys и His280Arg значимо увеличивают эффективность связывания АТФ, что может мешать прохождению реакции переноса фосфата с АТФ на треонин белка. Эффект замены Thr372Ile оказался не выражен. Замены Ser284Phe и Gly209Arg снижают эффективность связывания АТФ, и, как следствие, негативно влияют на функцию белка, что указывает на патогенность данных вариантов.

Таким образом, с помощью методов молекулярной динамики было исследовано влияние несинонимичных замен на связывание с АТФ. В совокупности с другими критериями ACMG для оценки патогенности несинонимичных замен, варианты

Gly209Arg, Leu257Pro, Lys229Met, Thr277Lys, His280Arg и Ser284Phe могут считаться патогенными.

1. Richards S. et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology, *Genet Med*, **17**(5):405–424.
2. Agnew C. et al. (2021) Structural basis for ALK2/BMP2 receptor complex signaling through kinase domain oligomerization, *Nat Commun*, **12**(1):4950.
3. Chen Y.J., Li P.Y., Yang C.N. (2021) Molecular dynamics study of enhanced autophosphorylation by S904F mutation of the RET kinase domain, *J Struct Biol*, **213**(4):107799.
4. Mathavan A. et al. (2023) Pulmonary hypertension in hereditary hemorrhagic telangiectasia: A clinical review, *Pulm Circ*, **13**(4):e12301.
5. Eichstaedt C.A. et al. (2023) Genetic counselling and testing in pulmonary arterial hypertension: a consensus statement on behalf of the International Consortium for Genetic Studies in PAH, *Eur Respir J*, **61**(2).
6. Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**(15):2114–2120.
7. O'Connor V.d.A. (2020) Picard toolkit, URL: <http://broadinstitute.github.io/picard>.
8. Poplin R. et al. (2018) A universal SNP and small-indel variant caller using deep neural networks, *Nat Biotechnol*, **36**(10):983–987.
9. Li Q., Wang K. (2017) InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines, *Am J Hum Genet*, **100**(2):267–280.
10. Mirdita M. et al. (2022) ColabFold: making protein folding accessible to all, *Nat Methods*, **19**(6):679–682.
11. Case D.A. et al. (2023) AmberTools, *J Chem Inf Model*, **63**(20):6183–6191.
12. Roe D.R., Cheatham T.E. (2013) PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data, *J Chem Theory Comput*, **9**(7):3084–3095.
13. Roe D.R., Cheatham T.E. (2018) Parallelization of CPPTRAJ enables large scale analysis of molecular dynamics trajectory data, *J Comput Chem*, **39**(25):2110–2117.

14. Miller B.R. et al. (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations, *J Chem Theory Comput*, **8**(9):3314–3321.

Insights into atherogenesis: potential master regulators and epigenetic factors identified

D. Borodko, T. Kovyanova, A. Omelchenko, V. Sukhorukov

*Institute of General Pathology and Pathophysiology
8 Baltiyskaya str., Moscow
daria.borodko@gmail.com*

Atherosclerosis, as well as other chronic inflammatory diseases, poses significant Russian health burdens as almost every second citizen of middle age shows symptoms, and many others have not been diagnosed due to the latent form. Chronic inflammation, which atherosclerosis originates from [1], is based on dysregulated innate immune responses to endogenous stimuli, yet the early cellular and molecular drivers remain poorly characterised. This study aimed to identify potential biomarkers and master regulators (MRs) that govern the onset of chronic inflammation, intending to improve early diagnostics and precision medicine strategies.

We fused p0 THP-1 cells with thrombocytes of atherosclerotic patients, then stimulated the resulting cybrids with LPS and measured IL1b, IL6, IL8, and CCL2 levels. Cybrids were then categorised into tolerant, intolerant, and non-responder groups based on cytokine response profiles. RNA sequencing was performed afterwards, and analysis was conducted using STAR (v2.10b) for read mapping and counting and DESeq2 for differential gene expression ($|\log_2FC| > 0.6$, adjusted p-value < 0.05). A subsequent master regulator search was performed with the geneXplain platform and the TRANSFAC database. We studied DGE in a cytokine-based manner: each DESeq2 model was subset by cytokine and meant to assess stimulation effects. Two consistently downregulated genes, TREM2 and OSMR, emerged across all DEG sets. TREM2, known to suppress PI3K and NF-kappa-B

signalling and promote anti-inflammatory pathways, showed strong coverage and functional relevance, suggesting its role as a negative regulator of chronic inflammation. OSMR, part of the IL6 receptor family, also demonstrated consistent downregulation, supporting our previous findings identifying its ligand (OSM) as a master regulator [2].

In addition, transcriptomic profiling revealed novel candidate MRs. EIF2AK2 was identified as a master regulator associated with the tolerant immune response, exhibiting higher expression levels in tolerant cybrids compared to intolerant ones across both stimulation points. In contrast, GSK3A, PDE4D, and PASK were linked to the intolerant response, showing reduced expression in intolerant cybrids during the initial stimulation and increased expression during the second, an expression pattern not observed in tolerant cybrids. These genes act as transcription factors or interact with transcriptional regulators involved in modulating immune responses.

Additionally, a strong intolerant response in atherosclerotic monocytes suggests an epigenetic basis. We performed a search for epigenetic-related genes among DEGs, and discovered 8 of them, showing the same direction of expression change in both stimulations and 2 to 3 cytokine groups. MST1 and ZBTB16 were consistently upregulated in IL1 and IL6 groups with log2FC 0.7 and 2.4, respectively. Other genes (PCGF2, DPF3, PRDM6, IKZF3, NPM2 and TAF1L) were downregulated in various group combinations with mean log2FC close to -1. All of these genes, except NPM2 and TAF1L, are associated with inflammation (mostly through immune response modulation) [3-5].

In conclusion, we identified a network of transcriptional and epigenetic regulators implicated in early stages of chronic inflammation, such as TREM2, OSMR, EIF2AK2, and ZBTB16. These findings offer potential targets for future biomarker validation and therapeutic exploration aimed at modulating chronic inflammation in cardiovascular disease.

RSF supported this work (grant № 24-15-00217).

1. Roy, P., Orecchioni, M. & Ley, K. (2021) How the immune system shapes atherosclerosis: roles of innate and adaptive immunity. *Nat Rev Immunol*
2. Orekhov AN et al. (2024) Evidence for the Involvement of Gene Regulation of Inflammatory Molecules in the Accumulation of

Intracellular Cholesterol: The Mechanism of Foam Cell Formation in Atherosclerosis. *Cur Med Chem*

3. Šeda O, Šedová L, Včelák J, Vaňková M, Liška F, Bendlová B. (2017) ZBTB16 and metabolic syndrome: a network perspective. *Physiol Res*

4. Li T, Wen Y, Lu Q, Hua S, Hou Y, Du X, Zheng Y, Sun S. (2023) MST1/2 in inflammation and immunity. *Cell Adh Migr*.

5. Zhou Y, Liu Z, Lynch EC, He L, Cheng H, Liu L, Li Z, Li J, Lawless L, Zhang KK, Xie L. (2021) Osr1 regulates hepatic inflammation and cell survival in the progression of non-alcoholic fatty liver disease. *Lab Invest*.

Межгенные взаимодействия ДНК-метилаз/деметилаз с микроРНК и длинными некодирующими РНК при раке яичников и молочной железы

Бриль А. А.^{1,2}, Бурденный А. М.¹, Лукина С. С.¹, Логинов В. И.¹

¹ ФГБНУ НИИОПП, Москва, Россия

² ФГАОУ ВО «ПМГМУ им. И.М. Сеченова» (Сеченовский Университет) МЗ РФ, Москва, Россия

alexbril2005@gmail.com

Ключевые слова: ДНК-метилтрансферазы, деметилазы, некодирующие РНК (нРНК), рак яичников, рак молочной железы

Введение: Эпигенетические механизмы, в частности метилирование ДНК и взаимодействия между некодирующими РНК, играют ключевую роль в прогрессировании рака [1]. ДНК-метилтрансферазы, деметилазы, микроРНК и длинные некодирующие РНК могут регулировать экспрессию генов через взаимодействия друг с другом [2]. Рак молочной железы является одним из наиболее распространенных видов рака среди женщин, в то время как рак яичников характеризуется низкой выживаемостью ввиду позднего выявления [3][4]. Понимание взаимодействий между генами системы ДНК-метилаз/деметилаз и нРНК может

расширить представление о патогенезе РМЖ и РЯ и способствовать созданию новых таргетных методов терапии.

Цель исследования: Проанализировать взаимодействия между ДНК-метилазами, деметилазами и некодирующими РНК при раке яичников и молочной железы с использованием биоинформатических методов.

Методы и алгоритмы: Основными источниками данных для анализа и предсказания потенциальных взаимодействий некодирующих РНК с ферментами метилаз/деметилаз являлись miRWalk, miRCode и LncBook. Фильтрация данных происходила по значению score (при использовании miRWalk score>0.96), рассчитанном in silico на основании последовательностей РНК. Связи с метаболическими процессами были выявлены инструментом g:profiler, использовавшем базу Gene Ontology (GO) с применением скорректированного p-value (метод g:SCS threshold - порог значений с учетом структуры GO). Общие сведения, касающиеся активности этих эпигенетических регуляторов, получены из базы Pubmed:Gene.

Результаты: Анализ показал, что микроРНК miR-9-5p, miR-125b-5p и miR-17-5p участвуют в ингибировании деметилаз TET1–TET3, что может способствовать подавлению генов-супрессоров. Длинные некодирующие РНК класса SNHG (SNHG 1, 3, 6, 7, 8, 12, 14, 17, 20) образуют дуплексы с данными микроРНК, снижая их активность. Кроме того, SNHG 1, 6 и 7 ассоциированы с изменением активности DNMT1 и DNMT3B, что указывает на их роль в регуляции уровня ДНК-метилирования. Совокупность этих взаимодействий формирует ось «днРНК – микроРНК – TET/DNMT», которая может влиять на ключевые сигнальные пути (PI3K/Akt, MAPK, JAK-STAT) и играть роль в патогенезе рака яичников и молочной железы.

Финансирование: Исследование выполнено в рамках государственного задания № FGfU-2025-0010 Министерства науки и высшего образования РФ учреждению ФГБНУ «НИИОПП».

Список литературы:

1. Prabhakaran R, Thamarai R, Sivasamy S, Dhandayuthapani S, Batra J, Kamaraj C, Karthik K, Shah MA, Mallik S. Epigenetic frontiers: miRNAs, long non-coding RNAs and nanomaterials are pioneering to

- cancer therapy. Epigenetics Chromatin. 2024 Oct 16;17(1):31. doi: 10.1186/s13072-024-00554-6.
2. Yang Z, Xu F, Teschendorff AE, Zhao Y, Yao L, Li J, He Y. Insights into the role of long non-coding RNAs in DNA methylation mediated transcriptional regulation. Front Mol Biosci. 2022 Dec 2;9:1067406. doi: 10.3389/fmolb.2022.1067406.
 3. Łukasiewicz S, Czezelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. Cancers (Basel). 2021 Aug 25;13(17):4287. doi: 10.3390/cancers13174287.
 4. Liberto JM, Chen SY, Shih IM, Wang TH, Wang TL, Pisanic TR 2nd. Current and Emerging Methods for Ovarian Cancer Screening and Diagnostics: A Comprehensive Review. Cancers (Basel). 2022 Jun 11;14(12):2885. doi: 10.3390/cancers14122885.

**База данных EndoGene: генетические варианты при
эндокринной и метаболической патологии для 5926 пациентов
русской популяции**

А.А. Буздин^{1,2}, М.А. Золотовская^{1,2}, С.А. Румянцев², А.Г. Емельянова², О.О.
Голунина², П.А. Пугачева², Д.В. Лупшов², А.В. Кузьминых², А.О.
Алексеева², А.А. Емельянова², А.Л. Новоселов², А. Матросова², А.А.
Слепухина², С.В. Попов², Е.В. Плаксина², В.М. Петров², А.Р.
Гусельникова², М.В. Сунцова², В.В. Захарова², Ж.Е. Белая², М.В.
Воронцова², Г.А. Мельниченко², Н.Г. Мокрышева², В.П. Чехонин², И.И.
Дедов²

buzdin@oncobox.com

¹ФГАОУ ВО Первый МГМУ им. И.М. Сеченова Минздрава России
(Сеченовский Университет)

²ФГБУ «НМИЦ эндокринологии» Минздрава России

Заболевания эндокринной системы представляют собой серьезное бремя для общественного здравоохранения и могут быть вызваны генетическими вариантами в отдельных генах или

комбинированным воздействием нескольких вариантов, а также факторами окружающей среды и образа жизни. В базе данных EndoGene представлены результаты анализов секвенирования следующего поколения (NGS), использованных для генетического профилирования 5926 пациентов с диагнозом 450 эндокринных и сопутствующих заболеваний, которые проходили обследование и лечение в ФГБУ «НМИЦ эндокринологии» в период с ноября 2017 года по январь 2024 года. Среди них 494, 1785, 692 и 1941 пациента были профилированы с помощью четырех разработанных внутри организации генетических панелей, включающих 220, 250, 376 и 382 гена соответственно, отобранных на основе анализа литературы и клинических рекомендаций, а 1245 пациентов были профилированы с помощью полноэкзомного секвенирования, охватывающего 31969 генов.

Всего клиническими генетиками были отмечены 2711 генетических вариантов как значимые для пациентов. Для таргетного секвенирования такие варианты были найдены в 1882 из 4960 тестов. 1267 тестов содержали генетические варианты, классифицированные как «патогенные» и «вероятно патогенные», а 700 - как «варианты неопределенной значимости». Для полноэкзомного секвенирования значимые генетические варианты были выявлены в 448 из 1248 тестов, включая патогенные и вероятно патогенные варианты в 203 случаях и варианты неопределенной значимости в 284 случаях. В некоторых случаях (267 для таргетного и 129 для полноэкзомного секвенирования) было аннотировано и зарегистрировано более одного варианта для пациента.

Всего 1184 из 2073 (57%) уникальных генетических вариантов не были описаны на момент интерпретации в литературе, как релевантные для эндокринной или метаболической патологии.

Наиболее часто мутировавшие гены с преобладанием патогенных и вероятно патогенных вариантов, включали *GCK* и *HNF1A* для фенотипа сахарного диабета или нарушений метаболизма глюкозы (коды МКБ-10 E10, E11, E13, E14, E74, R73), *KCNJ11*, *ABCC8* и *GCK* для других нарушений внутренней секреции поджелудочной железы (E16), *AIRE* и *MEN1* при полигландулярной дисфункции (E31), *AR* и *PTPN11* при других эндокринных нарушениях, включая конституциональную

низкорослость (E34), *GNRHR* и *PROPI* при гипофункции и других нарушениях гипофиза (E23), *DICER1* при других формах нетоксического зоба (E04), *CYP24A1* и *PHEX* при нарушениях минерального обмена (E83).

В базе данных EndoGene для каждого пациента мы указали код диагноза МКБ -10, пол и возраст на момент тестирования, тип тестирования (таргетное или полноэкзомное секвенирование). Генетические варианты представлены в соответствии с номенклатурой Human Genome Variation Society (HGVS)(1), с указанием зиготности и патогенности по классификации ACMG/AMP(2). База данных EndoGene находится в открытом доступе (<https://doi.org/10.5281/zenodo.10894526>.) и будет полезна врачам и исследователям, работающим в области молекулярных основ эндокринных и метаболических заболеваний.

Благодарности:

Мы благодарим врачей ФГБУ «НМИЦ эндокринологии» Минздрава России, которые диагностировали и лечили пациентов, включенных в данное исследование. Генетическое профилирование было частично спонсировано благотворительным фондом «Альфа-Эндо». Благодарим Дмитрия Штокало за полезное обсуждение.

Список литературы:

1. D. Dunnen et al. (2016) HGVS recommendations for the description of sequence variants: 2016 update. *Hum.Mutat.* **25: 37:** 564-569.
2. S. Richards et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine*, **17:** 405–424.

Определение пространственной структуры N-концевого димеризующегося цинкового пальца белка CG1449 *Drosophila melanogaster* методом ЯМР в растворе

У. В. Булгакова¹, С. С. Марьясина¹, К. И. Балагуров², Э. В. Бочаров³, А. Н. Бончук², В. И. Польшаков¹

¹МГУ имени М.В. Ломоносова, Россия, г. Москва, Ленинские горы,
1с73, uliana-bulgakova@mail.ru

²Институт биологии гена РАН, Россия, г. Москва, ул. Вавилова,
34/5

³ИБХ РАН, Россия, г. Москва, Миклухо-Маклая, 16/10

Белок CG1449 *Drosophila melanogaster* относится к семейству транскрипционных факторов с множественными цинковыми пальцами. Генетические исследования показали, что этот белок критически важен для правильного развития конечностей и дифференцировки клеток кишечника у дрозофилы [1]. Однако структурные и молекулярные механизмы его функционирования остаются неизученными.

Особый интерес представляет N-концевой домен CG1449 (аминокислотные остатки 129–161), который, согласно биоинформатическому анализу, содержит характерный для цинковых пальцев C2H2-типа мотив. Нами было показано, что этот домен может участвовать в димеризации белка, что имеет важное значение для его ДНК-связывающей активности и регуляторных функций. Димеризация через цинковые пальцы описана для ряда транскрипционных факторов [2], но структурные основы этого процесса для CG1449 ранее не исследовались. В данной работе мы впервые определили пространственную структуру N-концевого домена CG1449 (129–161 а.о.) методом ЯМР-спектроскопии в растворе.

Изотопно-меченый (¹⁵N, ¹³C) белок CG1449¹²⁹⁻¹⁶¹ был экспрессирован в *E. coli* и очищен для ЯМР-исследований. Эксперименты ЯМР проводили при 25°C на спектрометре Bruker Avance 600 МГц с криодатчиком. Набор 3D-спектров (HNCA, HN(CO)CA, HNCACB, CBCA(CO)HN, HNCO, HN(CA)CO и HCCH-TOCSY) позволил выполнить отнесение сигналов для 91.9% амидных групп, 95.95% всех С и Н-атомов остова и 73% всех С- и Н-атомов белка. Анализ химических сдвигов с помощью программы TALOS-N выявил две β-цепи (остатки 129-132 и 141-142) и α-спираль (остатки 145-154), что соответствует канонической структуре цинкового пальца C2H2-типа.

Для определения пространственной структуры были получены ограничения на межатомные расстояния H...H. Для этого измерили 3D-ЯМР спектры $^{15}\text{N}, ^1\text{H}$ HSQC-NOESY и $^{13}\text{C}, ^1\text{H}$ HSQC-NOESY. Расчет структуры выполняли с использованием программы ARIA 2.3, которая позволила определить отнесение сигналов NOE и получить из этих данных ограничения на межатомные расстояния.

Важно отметить, что наши данные ЯМР подтверждают способность этого домена к димеризации. Структурный анализ позволяет предположить, что димеризация может происходить через взаимодействие N-концевой β -цепи, что дает новые данные для понимания функциональных механизмов этого семейства белков.

Дальнейшие исследования будут направлены на изучение ДНК-связывающих свойств домена и анализ влияния мутаций на его структуру и стабильность. Полученные результаты вносят значительный вклад в понимание структурных основ функционирования C2H2-цинковых пальцев и могут способствовать направленному дизайну белков этого класса для биотехнологических применений.

Работа выполнена при поддержке гранта РФФИ № 24-14-00081.

Литература

1. Kittelmann M. et al. (2009) Insertional mutagenesis screening identifies the zinc finger homeodomain 2 (zfh2) gene as a novel factor required for embryonic leg development in *Tribolium castaneum*, *Dev Genes Evol*, 219:399–407.
2. Bonchuk A.N., Georgiev P.G. (2024) C2H2 proteins: Evolutionary aspects of domain architecture and diversification, *Bioessays*, 46:e24000521.

Универсальные архитектуры для предсказания сайтов связывания транскрипционных факторов

С.Е. Бушуев¹, А.О. Зинкевич¹, Д.Д. Пензар²

¹Факультет биоинженерии и биоинформатики, МГУ имени М.В. Ломоносова, Москва, Россия
bushsebzz@gmail.com

²Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, Россия

Транскрипция – процесс синтеза матричной РНК гена РНК-полимеразой, который впоследствии приведет к появлению функционального продукта: РНК или белка. Транскрипция определяет фенотип клетки, поэтому ее регуляция, основанная на специализированных белках – транскрипционных факторах, является важным атрибутом живых систем, позволяя им адаптироваться к различным условиям среды. Геномные сайты узнавания, расположенные в цис-регуляторных элементах, обладают различной аффинностью к транскрипционным факторам, зависящей от их нуклеотидных последовательностей. Их мутации могут приводить к изменению фенотипа, в частности, к генетическим заболеваниям, что маркирует важность разработки подходов для предсказания их мест расположения в геноме и их аффинности к ДНК-связывающим белкам.

Исследование аффинных к транскрипционным факторам последовательностей основано на 2-х группах методов: *in vitro* и *in vivo*, в первую группу входят PBM, SMiLE-seq и HT-SELEX, во вторую GHT-SELEX и ChIP-seq. Фундаментальной разницей этих групп методов является источник отбираемых последовательностей: первая группа методов основана на искусственных синтезированных последовательностях, вторая на геномной библиотеке. Возможны обучение паттернам на результатах экспериментов *in vitro*, их применение на результатах экспериментов *in vivo* и обратное (A2G и G2A, ‘Artificial to Genomic’ и ‘Genomic to Artificial’, соответственно). Такой перенос знания может быть осложнен малым количеством геномных

интервалов, детектируемых *in vivo* методами и различной длиной геномных и искусственных синтезированных последовательностей. Данные задачи были представлены основными номинациями прошедшего в этом году соревнования The IBIS Challenge, разработанного в нашей лаборатории, на основании которого была проведена наша работа. Нашей целью было создание универсального подхода, основанного на глубинном обучении, опережающего решения победителей соревнования.

В качестве основы для архитектуры модели была использована нейронная сеть LegNet[1], имеющая EfficientNet-подобную архитектуру, занявшую 1-е место в предсказании регуляторной активности последовательностей в международном конкурсе DREAM-2022 [2].

Мы модифицировали данную архитектуру, внедрив следующие особенности: замена сложной «головы» на «global maximum pooling» и 3 полносвязных слоя, замена техники «global average pooling» на «global maximum pooling» в составе SE блоков, использование фильтров размеров [3, 6, 9, 12] в Stem блоке (Рис.1). «Global maximum pooling» и модификация SE блока мотивированы решением проблемы размытия сигнала: перенос знаний о паттернах связывания осложнен занижением предсказаний моделями, использующими слой «global average pooling», при их применении на последовательностях большей длины, чем у использованных для их обучения. Наш подход использует ансамблирование: для описания паттернов связывания каждого представленного в соревновании транскрипционного фактора мы обучили 5 моделей.

Для каждого белка были созданы наборы данных, включающих аффинные последовательности, случайно набранные последовательности, связывающиеся с другими транскрипционными факторами, имеющие схожее распределение GC-состава, и моонуклеотидно перемешанные аффинные последовательности в пропорции 1:1:1. В рамках номинации A2G модели были обучены с использованием данных HT-SELEX на задачу регрессии номера цикла. В номинации G2A мы использовали данные GHT-SELEX и ChIP-seq для обучения моделей на задачу бинарной классификации аффинных и несвязываемых последовательностей. Для получения предсказаний на геномных интервалах мы использовали метод, заключающийся

в применении моделей на 3-х центральных участках последовательностей размерами [51, 151, 301] для ChIP-seq, [51, 101, 301] для GHT-SELEX и усреднении получившихся предсказаний.

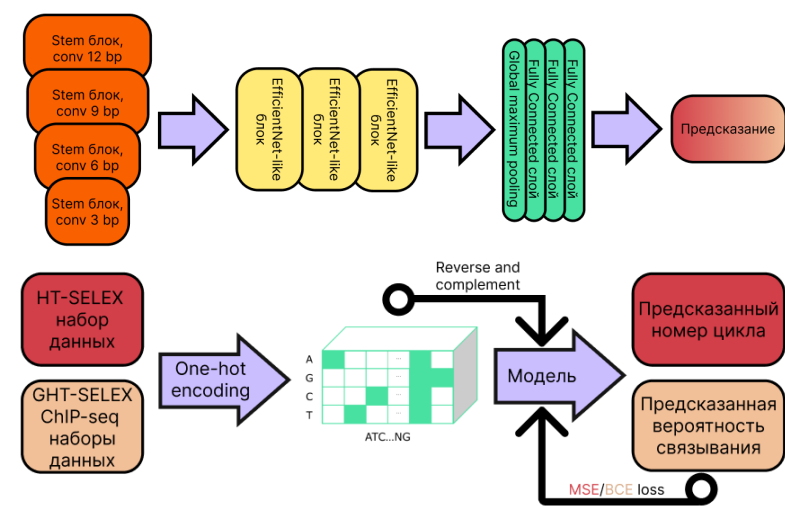


Рис.1 А) Архитектура LegNet-Max, полученная в ходе модифицирования LegNet. Б) Схема обучения моделей.

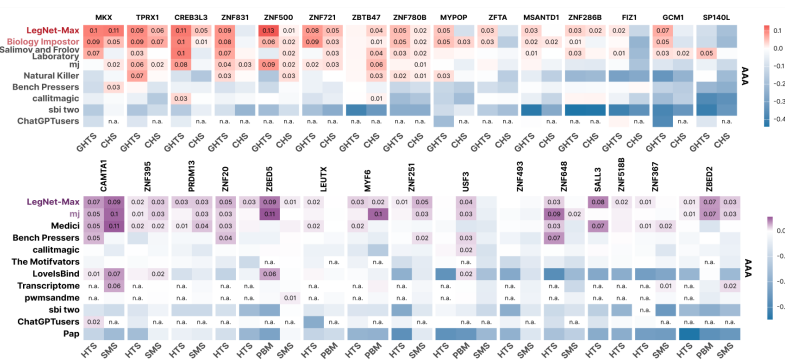


Рис.2 Сравнение подходов участников в рамках номинаций соревнования: A2G, G2A. Значения представляют из себя усредненные логарифмы отношений метрик решений участников The IBIS challenge к метрикам позиционно-весовых матриц.

В результате предложенная нами модель заняла первое место в пост-конкурсном анализе в номинациях A2G и G2A. Решения участников соревнования были сравнены с позиционно-весовыми матрицами, для которых была показана более качественная, чем у подходов, основанных на машинном обучении, генерализуемость между различными типами данных (Рис.2) [3].

Список литературы:

1. D.D. Penzar et al. (2023) LegNet: a best-in-class deep learning model for short DNA regulatory regions., *Nat Biotechnol*.
2. A.M. Rafi et al. (2024) A community effort to optimize sequence-based deep learning models of gene regulation, *Nat Biotechnol*.
3. R. Razavi, et al. (2024) Extensive binding of uncharacterized human transcription factors to genomic dark matter, bioRxiv

Comparative Single-Cell Analysis Reveals Primate-Like Zonation and Regenerative Signatures in the Adrenal Cortex of *Acomys cahirinus*

A.Bydanov¹, A.Bilyalov^{2,3}, D.Maksimov⁴, G.Gazizova¹,
E.Sharimardanova⁵, O.Gusev^{2,6}, O.Kozlova¹

¹*Kazan Federal University, Kazan, Russia, andrey.tokyo23@gmail.com*

²*Life Improvement by Future Technologies Center, Moscow, Russia*

³*The Loginov Moscow Clinical Scientific Center, Moscow, Russia*

⁴*Lomonosov Moscow State University, Moscow, Russia*

⁵*Skolkovo Institute of Science and Technology, Moscow, Russia*

⁶*Graduate School of Medicine, Juntendo University, Tokyo, Japan*

The African spiny mouse (*Acomys cahirinus*) is a valuable model for studying mammalian regeneration, exhibiting unique features such as scar-free wound healing, hair follicle regeneration, and rapid skin repair (1). Although its peripheral regenerative capacity is well studied, internal organs like the adrenal gland—a key regulator of stress and endocrine function—have not yet been examined at single-cell resolution. Because traditional rodent models have limited applicability to human adrenal biology, there is growing interest in alternative systems like *Acomys* for more relevant physiological insights.

Recent reviews highlight the potential of alternative models like *Acomys* to advance understanding of adrenal gland plasticity and disease by bridging regeneration and endocrine research (3). Single-cell transcriptomic studies have characterized adrenal cortex heterogeneity in humans and common laboratory mice. However, the detailed molecular and cellular composition of the *Acomys* adrenal gland remains uncharted. This gap limits insights into how *Acomys* adrenal biology may better reflect human physiology and regenerative capacity.

We applied single-cell RNA sequencing to adrenal cortex samples from *Acomys*, mouse, and human to identify both conserved and species-specific features of adrenal zonation, progenitor cell populations, and steroidogenic pathways. Our analysis uncovered unique cell states and regenerative gene expression signatures in *Acomys* that diverge from classical models of mammalian adrenal organization. These findings suggest an evolutionary integration of endocrine function with tissue regeneration, highlighting *Acomys* as a distinctive model for studying adrenal plasticity and regenerative endocrinology.

Methods and Algorithms: We performed single-cell RNA sequencing (scRNA-seq) of adrenal cortices from 10 adult *Acomys cahirinus* (5 males, 5 females) using the 10x Genomics Single Cell 5' v2 (R2-only) protocol. Raw sequencing data were aligned to the *Acomys cahirinus* genome using Cell Ranger (v6.1.2), yielding an average of ~88% total mapped reads and ~73% uniquely mapped reads per sample. Downstream analysis was conducted in Seurat v5 (3), using SCTransform normalization and dimensionality reduction (PCA, UMAP). To correct for batch effects and individual variation, we applied

multiple integration strategies, including Harmony (4), Canonical Correlation Analysis (CCA), Reciprocal PCA (RPCA), and scVI (5). Clustering was performed using Seurat algorithm at multiple resolutions to identify stable cell populations. Then we performed differential gene expression (DEG) analysis between clusters and across sexes. Comparative analyses included unpublished human and mouse adrenal scRNA-seq datasets generated using comparable protocols.

Cell type annotation was performed through a combination of manual marker-based curation and deep learning models, including scANVI, scBERT, and CellTypist. For advanced visualisation we used SCP, SeuratExtend packages.

Results: Following standard quality control metrics, single-cell transcriptomes from 10 *Acomys cahirinus* adrenal samples were retained for downstream analysis. Integration and clustering revealed distinct adrenal cell populations, including the zona glomerulosa (ZG), zona fasciculata (ZF), and zona reticularis (ZR), along with adrenal medullary cells, immune cells, endothelial cells, neuroglia, and mesenchymal and progenitor cells.

We integrated adrenal cortex single-cell datasets from *Acomys*, human, and mouse using multiple computational approaches. Among these, Harmony and scVI demonstrated the highest performance for cross-species data integration. Due to the inclusion of non-standard tissue types and inherent biological differences in *Acomys*, cell type annotation proved challenging. Consequently, we primarily relied on marker-based identification of cell subtypes, using deep learning and machine learning methods as complementary references to support our annotations. Further analysis was concentrated on adrenal cortex, mesenchyme and progenitor cells.

Cross-species integration with *Mus musculus* adrenal cortex data revealed a broadly conserved cellular landscape, yet with notable species-specific differences. *Acomys* displayed a well-defined zona reticularis population that was absent in the mouse dataset, consistent with previous reports highlighting its presence primarily in primates. Moreover, *Cyp17a1* — a gene essential for androgen synthesis and typically expressed in the human adrenal ZR — was highly expressed in

Acomys ZR-like cells but undetectable in mouse adrenals, suggesting Acomys recapitulates primate-specific steroidogenic features.

Multispecies embeddings (Acomys–Mouse–Human) revealed that Acomys clusters were transcriptionally distinct from both mouse and human counterparts, with several Acomys-specific clusters characterized by unique gene expression signatures. In particular, we observed unexpected spatial expressions of *Wnt4* and *Shh*. In contrast to their canonical zonation — with *Wnt4* enriched in the ZG and *Shh* localized to the subcapsular progenitor zone — Acomys adrenal cells showed altered zonation of both genes, suggesting a divergence in the spatial regulation of progenitor maintenance and steroidogenic differentiation.

Conclusion: Our scRNA-seq analysis reveals Acomys adrenal glands display unique features: atypical *Wnt4/Shh* expression, a primitive zona reticularis domain, and sexually dimorphic cell subtypes profile - distinct from both mice and primates. These findings redefine adrenal zonation paradigms and position Acomys as an evolutionary bridge for endocrine-regeneration studies. Ongoing spatial transcriptomics will resolve niche architecture and zonal boundaries to link molecular signatures with tissue organization in this regenerative model.

Funding: The study is supported by the Ministry of Science and Higher Education of the Russian Federation grant 075-15-2021-1344.

References:

1. A.W. Seifert, J.R. Monaghan, M.D. Smith, B. Pasch, M. Maden, S.R. Voss (2012) Skin shedding and tissue regeneration in African spiny mice (Acomys), *Nature*, 489:561–565.
2. A. Bilyalova, A. Bilyalov, N. Filatov, E. Shagimardanova, A. Kiyasov, M. Vorontsova, O. Gusev (2024) Non-classical animal models for studying adrenal diseases: advantages, limitations, and implications for research, *Lab Anim Res*, 40:25.
3. Y. Hao et al. (2021) Integrated analysis of multimodal single-cell data, *Cell*, 184:3573–3587.e29.
4. I. Korsunsky et al. (2019) Fast, sensitive, and accurate integration of single-cell data with Harmony, *Nat Methods*, 16:1289–1296.

5. A. Gayoso et al. (2022) scvi-tools: A library for deep probabilistic analysis of single-cell omics data, *Nat Methods*, 19:387–390.

OrgNet: Orientation-Gnostic Protein Stability Assessment Using 3D Convolutional Neural Networks

Ilya Buyanov, Anastasiia Sarycheva, Petr Popov

*Center for Molecular and Cellular Biology, Skolkovo Institute of
Science and Technology, Moscow, 121205, Russia,
ilia.buyanov@skoltech.ru*

Accurate prediction of the effects of single-point mutations on protein stability is crucial for understanding the molecular mechanisms underlying diseases and for advancements in protein engineering [1]. Current structure-based deep learning approaches employing three-dimensional convolutional neural networks (3D CNNs) have shown promise [2, 3]; however, they often suffer from orientation biases due to their sensitivity to the input protein orientation. To overcome this limitation, we introduce OrgNet—a novel, orientation-gnostic deep learning framework model designed to predict changes in protein thermostability ($\Delta\Delta G$) upon mutation that overcomes this limitation by applying a spatial standardization procedure.

Specifically, OrgNet represents protein structures as voxel grids that encode detailed atomic features (hydrophobic, aromatic, hydrogen bonding donor, hydrogen bond acceptor, positive ionizable, negative ionizable, and occupancy) at the mutation. Critically, OrgNet implements a spatial transformation strategy [4] by defining a residue-based coordinate frame and subsequently standardizing the input orientations. This process effectively mitigates the inherent SE(3) variance of conventional 3D CNN architectures, ensuring that predictions remain robust regardless of the original orientation of the protein structure.

Experimental evaluations were conducted on benchmark datasets (e.g., Ssym [5], S669 [6]) as well as cross-validation splits (using training sets such as Q1744 [7], Q3214 [7], and S2648 [8]) to test both direct and reverse mutation predictions. Results indicate that OrgNet not only reduces mean absolute error (MAE) and root-mean-square error (RMSE) but also enhances the Pearson correlation between experimental and predicted $\Delta\Delta G$ values compared to previous methods such as ThermoNet [7]. Moreover, OrgNet displayed balanced performance for both direct and reverse mutation predictions, underscoring the utility of eliminating orientation bias in 3D CNN-based approaches.

These findings highlight the potential of integrating spatial standardization within deep learning architectures for protein stability prediction. Additionally, the flexible design of OrgNet creates opportunities for future integration with sequence- and evolution-based features and for development of more comprehensive predictive frameworks in area of protein stability prediction. OrgNet is freely available at <https://github.com/i-Molecule/OrgNet>.

Acknowledgements

The authors acknowledge the support provided by RSF project No. 22-74-10098. The authors wish to thank the conference organizers.

1. Castrense Savojardo et al. (2016) Inps-md: a web server to predict stability of protein variants from sequence and structure, *Bioinformatics*, 32(16):2542–2544.
2. Ludovica Montanucciet al. (2019) Ddgun: an untrained method for the prediction of protein stability changes upon single and multiple point variations, *BMC bioinformatics*, 20:1–10.
3. S Benevenuta et al. (2021). An antisymmetric neural network to predict free energy changes in protein variants, *Journal of Physics D: Applied Physics*, 54(24):245403.
4. Guillaume Pag`es, et al. (2019) Protein model quality assessment using 3d oriented convolutional neural networks, *Bioinformatics*, 35(18):3313–3319.

5. Fabrizio et al. (2018) Quantification of biases in predictions of protein stability changes upon mutations, *Bioinformatics*, 34(21):3659–3665.
6. Corrado Pancotti et al. (2022) Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset, *Briefings in Bioinformatics*, 23(2):bbab555
7. Bian Li et al. (2020) Predicting changes in protein thermodynamic stability upon point mutation with deep 3d convolutional neural networks, *PLoS computational biology*, 16(11):e1008291
8. Yves Dehouck et al. (2011) Popmusic 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality, *BMC bioinformatics*, 12:1–12.

Белковые футпринты в РНК-хроматиновых интерактомах

А.А. Валяева¹, А.А. Жарикова^{2,3}, А.А. Миронов³

¹*НИИ ФХБ имени А.Н. Белозерского, МГУ имени М.В. Ломоносова, Ленинские горы, д. 1, стр. 40, 119992 Москва, Россия, valyaeva.ann@gmail.com*

²*Факультет биоинженерии и биоинформатики, МГУ имени М.В. Ломоносова, Ленинские горы, д. 1, стр. 73, 119234 Москва, Россия*

³*Институт общей генетики им. Н.И. Вавилова Российской академии наук, ул. Губкина, д. 3, Москва, 119991, Россия*

В полногеномных экспериментах, направленных на построение сетей взаимодействия РНК и локусов ДНК, часто проводится этап фиксации межмолекулярных взаимодействий с помощью формальдегида, который в первую очередь формирует сшивки между аминокислотными группами белков и нуклеиновых кислот. Вследствие этого подобные эксперименты детектируют преимущественно РНК-ДНК взаимодействия, опосредованные

различными белками, про которые, однако, напрямую из эксперимента обычно ничего неизвестно. Тем не менее, информация о белковых партнерах хроматин-ассоциированных РНК могла бы существенно помочь в исследовании функций данных РНК.

В этой работе мы поставили перед собой цель детектировать и охарактеризовать ДНК-связывающие белки, опосредующие РНК-ДНК контакты. Наш подход заключался в поиске и описании следов - футпринтов, которые оставляют белки, экранирующие участок ДНК, с которым они связываются. Поиск футпринтов мы проводили, анализируя особенности покрытия генома прочтениями ДНК-частей вблизи сайтов связывания белков с известными мотивами. Мы использовали ранее обработанные в нашей лаборатории данные полногеномных экспериментов по установлению РНК-хроматиновых взаимодействий - RedC [1] и RedChIP [2], которые были получены на клеточной линии K562. Позиции сайтов связывания белков мы определяли с помощью пересечения ChIP-seq пиков (БД ReMap2022 [3]) и предсказанных по всему геному сайтов ДНК-связывающих белков с известными мотивами связывания (БД HOCOMOCO v12 [4], HOMER [5]). Надежные РНК-ДНК контакты мы отбирали из значимых пиков, определенных алгоритмом BaRDIC [6]. Для каждого белка с известными сайтами связывания мы строили в окнах вокруг потенциальных сайтов связывания (± 400 п.н.) усредненное покрытие РНК-ДНК контактами.

В данной работе нам удалось собрать данные для 134 ДНК-связывающих белков и получить для них распределения покрытий вокруг сайтов их связывания в геноме. В первую очередь мы проанализировали покрытие ДНК-частями вокруг сайтов связывания белка CTCF - транскрипционного фактора, участвующего в поддержании структуры хроматина. Мы показали, что в самих сайтах связывания CTCF наблюдается провал в покрытии РНК-ДНК контактами, то есть весь пик обогащения РНК-ДНК контактами в области сайтов связывания белка разделяется на два - справа и слева от сайта. Это может говорить о том, что данные

локусы ДНК закрыты белком и недоступны для взаимодействия с РНК. Однако такая форма покрытия РНК-ДНК контактами прослеживается лишь для нескольких других ДНК-связывающих белков. Для подавляющего большинства проанализированных белков, в локусах связывания которых мы имеем достаточно сигнала РНК-ДНК взаимодействий, мы наблюдали лишь один пик в покрытии, что значит, что эти белки не экранируют собой ДНК полностью. Такая ситуация оказалась типичной для ДНК-связывающих белков с короткими мотивами связывания. Таким образом, анализ формы покрытия сайтов связывания ДНК-связывающих белков с известными мотивами прочтениями ДНК-частей позволил разделить белки, потенциально опосредующие РНК-хроматиновый контакт, на две группы.

С помощью проведенного анализа мы продемонстрировали, что в данных РНК-хроматиновых интерактомах возможна детекция футпринтов белков, опосредующих РНК-ДНК взаимодействия. Описанный подход позволяет определить ДНК-связывающие белки, за счет которых могут формироваться контакты. Определение предпочитаемых белковых партнеров в дальнейшем поможет предположить функции хроматин-ассоциированных РНК.

Работа финансировалась РФФ грант №23-14-00136.

1. A. A. Gavrilo et al. (2020) Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics, *Nucleic Acids Research*, 48: 6699–6714.
2. A. A. Gavrilo et al. (2021) RedChIP identifies noncoding RNAs associated with genomic sites occupied by Polycomb and CTCF proteins, *Proc. Natl. Acad. Sci. U.S.A.*, 119: e2116222119
3. F. Hammal, et al. (2021) ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments, *Nucleic Acids Research*, 50: D316–D325.

4. I. E. Vorontsov et al. (2023) HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors, *Nucleic Acids Research*, 52: D154–D163.
5. S. Heinz et al. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities, *Molecular Cell*, 38: 576–589.
6. D. E. Mylarshchikov et al. (2024) BaRDIC: robust peak calling for RNA–DNA interaction data, *NAR Genomics and Bioinformatics*, 6: lqae054.

Восстановление энергетических профилей олигомеризации пептидов по кинетике агрегации

Е.О. Василенко^{1,2,3}, Д.Н. Иванков¹

¹*Сколковский институт науки и технологий,*

²*Институт молекулярной биологии им. В.А. Энгельгардта РАН,*

³*Научно-исследовательский институт системной биологии и
медицины Роспотребнадзора,
Egor.Vasilenko@skoltech.ru*

Пептиды и белки могут собираться в крупные волокна, называемые фибриллами. Рост таких агрегатов при наличии зародышей может начинаться сразу, но без них обычно предваряется лаг-фазой, во время которой появляются олигомеры, разнообразие и неустойчивость которых пока что препятствуют пониманию их влияния на зарождение фибрилл [1]. Существует несколько моделей появления зародышей упорядоченных агрегатов: (1) частично денатурированные белки соединяются гидрофобными участками, образуя первичные димеры; (2) нативные белки побуждают друг друга развернуться и перестроиться в бета-слои; (3) наконец, в переходных олигомерах из 2-7 мономеров происходят перестройки, например, из альфа-спиралей в бета-слои [2]. Такие олигомеры и перестройки их

вторичных структур обнаружены у бета-амилоида, альфа-синуклеина и инсулина методами седиментации, хроматографии, динамического рассеяния света, масс-спектрометрии, кругового дихроизма и др.

Модели линейных агрегатов дают согласующиеся с экспериментами кинетические кривые и позволяют получать оценки размеров зародышей [3]. Чтобы добавить к ним представления о природе перехода от неустойчивых олигомеров к первым спонтанно растущим зародышам, мы развили капельные модели [4], которые позволяют получать термодинамически осмысленные оценки размеров критических олигомеров. Связь со скоростями была проведена через классическую модель образования критических (наименьших самопроизвольно растущих) зародышей и их роста по уравнению диффузии в пространстве размеров [5]. Сопоставив открытые данные о концентрациях олигомеров и кинетических константах, мы сделали предположения о содержащих олигомеры путях перехода от мономеров к быстро растущим фибриллам. Количество первичных зародышей оценивалось из соотношения между продолжительностями лаг-фазы и фазы роста. Вычисления проводились в предположении, что (1) лаг-фазу можно оценить через время созревания первых критических олигомеров до зародышей, (2) фибриллы могут расти только из критических олигомеров, образовавшихся в лаг-фазе, (3) продолжительность фазы роста определяется скоростью удлинения фибрилл до их равновесных размеров. Предложенный алгоритм численно подбирает параметры олигомеров, которые согласуются с экспериментальными концентрациями пептидов, размерами агрегатов и длительностями фаз агрегации.

Такой подход позволил исследовать возможность участия переходных олигомеров в агрегации инсулина, альфа-синуклеина, бета-амилоида и сравнить такие пути с моделью медленной сборки первичных бета-димеров. Различия между полученными распределениями оценок термодинамических параметров олигомеризации различных пептидов мы связали с их длинами и

первичной и вторичной структурами, которые влияют на водородные связи и потери степеней свободы в олигомерах. Мы надеемся, что разработанный алгоритм позволит продвинуться в понимании, какие олигомеры способствуют зарождению упорядоченных упаковок (on-pathway), а какие — лишь забирают в себя мономеры, не давая им собираться в упорядоченные зародыши (off-pathway), и как это связано с аминокислотной последовательностью пептида, а также pH, ионной силой и температурой.

Исследование выполнено за счёт гранта Российского научного фонда № 19-74-30007, <https://rscf.ru/project/19-74-30007/>.

1. M. Muschol et al. (2013) Amyloid oligomers as on-pathway precursors or off-pathway competitors of fibrils, *Frontiers in molecular biosciences*, 10:1120416.
2. T. Gurry et al. (2013) The dynamic structure of α -synuclein multimers, *Journal of the American Chemical Society*, 135.10:3865-3872.
3. N.V. Dovidchenko et al. (2014) How to determine the size of folding nuclei of protofibrils from the concentration dependence of the rate and lag-time of aggregation. I. Modeling the amyloid protofibril formation, *The Journal of Physical Chemistry B*, 118.5:1189-1197.
4. A. Šarić et al. (2016) Kinetics of spontaneous filament nucleation via oligomers: Insights from theory and simulation, *The Journal of chemical physics*, 145.21.
5. Я.Б. Зельдович (1942) К теории образования новой фазы. Кавитация, ЖЭТФ, 12.11-12:525-538.

Моделирование взаимодействия пластоцианина и фотосистемы 1 гороха

Вольхин И.А.¹, Фёдоров В.А.², Хрущёв С.С.², Коваленко И.Б.²

¹*Факультет биоинженерии и биоинформатики, МГУ им М.В. Ломоносова*
E-mail: ilyavolkhin2@gmail.com

²*Биологический факультет, МГУ им. М.В. Ломоносова*

Пластоцианин - белок-переносчик электронов в электрон-транспортной цепи цианобактерий и хлоропластов, осуществляющий перенос электронов от цитохрома *f* цитохром-*b6f*-комплекса к фотосистеме 1. В данной работе исследовалось формирование белок-белкового комплекса между пластоцианином и фотосистемой 1 гороха *Pisum sativum* с использованием методов молекулярного моделирования. Мы предполагаем, что на первом этапе белок-переносчик диффундирует под действием случайных сил и электростатических взаимодействий. Результатом диффузии является формирование столкновительного комплекса, который затем может превратиться в финальный комплекс, в котором возможен перенос электронов [1]. Мы моделировали формирование столкновительного комплекса с использованием метода броуновской динамики, в котором молекулы белков представлены как твердые тела, движущиеся в непрерывной среде растворителя [2]. Затем мы использовали структуры столкновительных комплексов в качестве начальных структур для моделирования формирования финального комплекса с использованием молекулярной динамики. На первом этапе было выделено два варианта столкновительных комплексов, отличающихся ориентацией пластоцианина относительно реакционного центра фотосистемы 1. Для каждой структуры столкновительных комплексов было проведено по три вычислительных эксперимента методом молекулярной динамики продолжительностью до 2 мкс

для моделирования образования финального комплекса. Показано, что столкновительный комплекс образуется за счет электростатического взаимодействия между пластоцианином и субъединицей PsaF фотосистемы 1, а финальный комплекс стабилизируется гидрофобными взаимодействиями с субъединицами PsaA и PsaB. Выявлены ключевые аминокислотные остатки, участвующие в этом процессе. Помимо формирования продуктивного финального комплекса также удалось наблюдать образование непродуктивного, а также субоптимального комплексов, вносящих вклад в кинетику переноса электрона по электронтранспортной цепи в ходе фотосинтеза.

1. Fedorov VA, Kovalenko IB, Khrushev SS, Ustinin DM, Antal TK, Riznichenko GY, et al. Comparative analysis of plastocyanin–cytochrome f complex formation in higher plants, green algae and cyanobacteria. *Physiol Plant* (2019), <https://doi.org/10.1111/ppl.12940>.
2. Kovalenko, I.B. et al. Computer simulation of protein-protein association in photosynthesis (2011), <https://doi.org/10.1051/mmnp/20116704>

Prediction of Post-Myocardial Infarction Cardiac Complications Considering CYP2C19 Genotypes Using Machine Learning

Vorobev A.S.¹, Burkin K.M.², Kirdeev A.V.², Nikolaev K.Yu³,
Poptsova M.S.²

¹*District Cardiology Dispensary "Center for Diagnostics and Cardiovascular Surgery", Surgut, Russia
a.s.vorobyov@gmail.com*

²*National Research University Higher School of Economics, Moscow, Russia*

³*Scientific Research Institute of Therapy and Preventive Medicine, SB
RAS, Novosibirsk, Russia*

Despite advances in therapy, major adverse cardiovascular events (MACEs) remain frequent after myocardial infarction (MI), occurring in approximately 20–30% of patients within the first year [1]. Improving risk stratification for these high-risk patients is critical. Machine learning (ML) models have recently emerged as promising tools for long-term prognostication in MI by integrating diverse clinical, imaging, and genetic factors [2, 3]. In parallel, genetic markers such as polymorphisms of the CYP2C19 enzyme are widely studied for their potential impact on post-MI outcomes [4,5]. In this study, we aimed to leverage ML algorithms to identify clinical and genetic predictors of long-term adverse cardiovascular complications in post-MI patients with different CYP2C19 genotypes.

We analyzed data from 218 patients with acute MI who underwent percutaneous coronary intervention and were followed for 9 years. Adverse outcomes (composite MACEs) tracked during follow-up included cardiovascular death, recurrent acute coronary syndromes, stent or bypass graft thrombosis, stroke, and repeat myocardial revascularization. Three tree-based ML models (CatBoost, LightGBM, and Random Forest) were trained and compared, alongside two feature selection methods (Shapley additive explanations, SHAP, and sequential selection). The CatBoost model with SHAP-based backward feature elimination achieved the best performance (area under the ROC curve AUC = 0.721). The most significant prognostic features identified were patient age, Charlson comorbidity index, severity of coronary artery lesions, type of P2Y12 inhibitor, type of coronary stent, and CYP2C19. Notably, the presence of CYP2C19 loss-of-function or gain-of-function allelic variants was associated with a higher risk of adverse events, whereas treatment with ticagrelor (a potent P2Y12 inhibitor) and the use of drug-eluting stents were linked to better long-term outcomes. In conclusion, the CatBoost-based ML model effectively identified key clinical (e.g. age, comorbidities), angiographic (significant coronary

lesions), and genetic (CYP2C19 polymorphisms affecting enzyme function) factors as significant predictors of late complications after MI. These results suggest that ML algorithms can enhance risk assessment in post-MI patients by pinpointing individuals at higher risk of long-term MACEs. At the same time, our findings highlight that modern interventions – specifically, drug-eluting stents and more potent antiplatelet therapy (ticagrelor) – are associated with improved prognosis in this high-risk population.

1. Bergmark B.A., et al. (2022) Acute coronary syndromes. *Lancet*. 399(10332):1347-1358
2. Khera R., et al. (2024) Transforming Cardiovascular Care with Artificial Intelligence: From Discovery to Practice. *J. Am. Coll. Cardiol.*, 84:97–114.
3. Johnson K.W., et al. (2018) Artificial Intelligence in Cardiology. *J Am Coll Cardiol*. 2018;71:2668-2679.
4. Mega J.L., et al. (2010) Reduced-function CYP2C19 Genotype and Risk of Adverse Outcomes in Clopidogrel-Treated Patients. *JAMA*, 304:1821–1830.
5. Martinez-Quintana E, et al. (2017) CYP2C19 Activity and Cardiovascular Risk Factors in Acute Coronary Syndrome. *Med. Clin. (Barc)*, 149:235–239.

Изменение профиля экспрессии генов в покоящихся яйцах *Daphnia magna*

Е. Д. Воробьева¹, М. Петрова³, Г. Р. Газизова¹, П. Ерохов², А. Тухбатуллин², Е. Воронежская², О. А. Гусев^{1,4,5}, Я. Р. Галимов²,
Е. И. Шагимарданова^{5,6}

¹*Казанский Федеральный университет, Казань, Россия*
elizavetavrbv@gmail.com

²*Институт Биологии Развития РАН им. Кольцова, Москва, Россия*

³*Биологический Факультет МГУ им. Ломоносова, Москва, Россия*

⁴Университет Джунтендо, Токио, Япония

⁵ООО «Лифт-Центр», 121205, Москва, Россия

⁶Сколковский Институт Науки и Технологий, Москва, Россия,

Адаптация к экстремальным условиям окружающей среды является ключевым фактором эволюционного успеха многих видов. Она включает в себя формирование состояний физиологического покоя, например, диапаузу — генетически запрограммированную форму приостановки роста и развития, характерную для многих позвоночных и беспозвоночных животных.

Представители рода *Daphnia* опираются на стратегию откладывания эфиппийных яиц (находящихся в эмбриональной спячке) для выживания популяции в неблагоприятный период. Известно, что для возобновления их развития необходим период инкубации перед стимуляцией. В случае стимуляции яиц сразу после откладки, их вылупления не происходит. Молекулярные процессы, протекающие в покоящемся яйце, обеспечивающим нечувствительность к стимуляции остаются недостаточно изученными.

В данной работе проводился анализ транскриптомных данных покоящихся яиц *Daphnia magna* в условиях холодной воды (+1°C) и темноты. Яйца инкубировали в течение 7ми, 30ти и 90 суток, после чего проводили выделение РНК, подготовку библиотек и секвенирование на платформе Illumina HiSeq 2500.

Полученные RNA-seq данные были триммированы и картированы на референсный геном *D. magna* GCA_030254905.1 [1]. Картирование выполняли с помощью программы STAR-2.7.11b. Дальнейшая обработка данных включала в себя анализ дифференциальной экспрессии генов в RStudio с помощью пакета DESeq2-v.1.46.0 с использованием LRT-модели. Была проведена дополнительная фильтрация по коэффициенту вариации для снижения внутригрупповой вариативности. Дальнейший анализ главных компонент (PCA) выявил четкое разделение образцов по временным точкам. При этом наблюдался двухфазный характер изменений: резкий сдвиг экспрессии на 7-й день и последующая

«стабилизация» транскриптома на более поздних сроках хранения (30 и 90 сутки).

Дифференциально экспрессируемые гены для дальнейшего расчета корреляций и анализа были отобраны по статистической значимости: $|\log_2\text{FoldChange}| > 1$ и скорректированное $p\text{-value} - p.\text{adjusted} < 0.05$. Данный порог прошло 9660 генов.

Для оценки паттернов экспрессии генов была рассчитана матрица корреляции экспрессии генов на основе коэффициента Спирмена. На основе полученной матрицы корреляций была проведена иерархическая кластеризация данных

На основе оценки кластерной структуры гены были распределены на шесть кластеров с различными паттернами изменения экспрессии в ходе диапаузы. Раннее снижение транскрипции затронуло гены, связанные с митохондриальным дыханием, окислительным фосфорилированием и синтезом АТФ, что может указывать на подавление энергетического обмена. Аналогичные метаболические изменения описаны в работе на копеподах *Calanus finmarchicus* [2]. Одновременно наблюдался кратковременный пик экспрессии генов рибосомного биогенеза и процессинга РНК на ранней стадии покоя. На более поздних этапах возрастала активность генов, участвующих в морфогенезе, нейромышечных и сенсорных процессах, что может быть связано с подготовкой организма к восстановлению активности.

Среди генов, демонстрирующих дифференциальную экспрессию, были обнаружены гены белков теплового шока, гены кодирующие антиоксиданты, метаболизма трегалозы и другие.

Например, ген белка теплового шока 70 кДа (Hsp70) возрастает к 7му дню инкубации и затем сохраняет высокий уровень экспрессии до 90го дня. Ген, кодирующий мультидоменный белок с аминоксил-тРНК-синтетазной активностью, имел максимальную экспрессию на 7ой день хранения.

Таким образом, показано, что при отсутствии морфологических изменений покоящихся яиц дафний при трехмесячной инкубации, на молекулярном уровне происходят

значительные транскрипционные изменения. Эти процессы, наряду с поддержанием состояния покоя, обеспечивают подготовку эмбрионов к возобновлению развития.

Работа выполнена в рамках Программы стратегического академического лидерства Казанского федерального университета (ПРИОРИТЕТ-2030).

1. A. Chaturvedi et al. (2023) The hologenome of *Daphnia magna* reveals possible DNA methylation and microbiome-mediated evolution of the host genome, *Nucleic Acids Research*, 51: 9785–9803.
2. L. Payton et al. (2024) Revealing the profound influence of diapause on gene expression: insights from the annual transcriptome of the copepod *Calanus finmarchicus*, *Molecular Ecology*, 33 (13).

**Пайплайн HLA-типирования при рецидиве лейкоза после
ТГСК и его применение в Национальной генетической
инициативе «100 000+Я»**

Г.М. Выходцев¹, М.Х. Гуржиханова², О.С. Киреева¹, Д.В.
Пустошилов¹

¹ООО «Биотехнологический кампус», Москва, Россия
grirow@yandex.ru

²НМИЦ ДГОИ им. Дмитрия Рогачева, Москва, Россия

Проведён сравнительный анализ инструментов для HLA-типирования по данным NGS, включая широко используемый HLA-HD и новый инструмент T1K [1]. Ранее на данных Национальной генетической инициативы «100 000+Я» HLA-типирование выполнялось с помощью HLA-HD, однако этот инструмент допускал ошибки на образцах пациентов с рецидивом лейкоза после трансплантации гемопоэтических стволовых клеток (ТГСК).

Для валидации результатов Центр им. Рогачёва предоставил образцы таких пациентов, а также экспериментальную панель для HLA-типирования, которая послужила «золотым стандартом».

По результатам бенчмаркинга инструмент T1K продемонстрировал более высокую точность по сравнению с HLA-HD, особенно на сложных клинических образцах.

Выявлено, что в случаях рецидива лейкоза после ТГСК нередко (в ~10–30% случаев) возникает копияно-нейтральная потеря гетерозиготности (CN-LOH) в генах HLA. Данная аномалия позволяет опухолевым клеткам избегать иммунного контроля донорских лимфоцитов и существенно влияет на выбор терапии [2].

Инструмент T1K способен выявлять гомозиготные аллели (возможный индикатор CN-LOH), однако не позволяет оценивать изменения числа копий генов. Поэтому пайплайн был дополнен модулем CNVpytor для анализа копияности, что повысило информативность результатов и позволило обнаруживать события CN-LOH, а также делеции и дупликации в HLA-локусе.

На основе данных Национальной генетической инициативы «100 000+Я» (когорты здоровых доноров и пациентов) была построена база частот аллелей HLA, а также обнаружены отличия в локальных частотах аллелей HLA по сравнению с общемировыми, что подчёркивает значимость популяционно-специфичных баз данных аллельных частот.

1. Song, L., Bai, G., Liu, X. S., Li, B. & Li, H. Efficient and accurate KIR and HLA genotyping with massively parallel sequencing data. *Genome Res.* 33, 923–931 (2023).
2. Arnold, P. Y. Review: HLA loss and detection in the setting of relapse from HLA-mismatched hematopoietic cell transplant. *Hum. Immunol.* 83, 712–720 (2022).

**Молекулярные механизмы адаптации к зимней спячке:
транскриптомный анализ скелетных мышц сони-полчка**

Г.Р. Газизова¹, Р.М. Девятяров^{2,6}, О.С. Козлова¹, А.С.
Комиссаров³, О.В. Тяпкина⁴, Е.И. Шагимарданова^{5,6}, О.А. Гусев^{2,6}

¹*Казанский федеральный университет, Кремлевская, 18, Казань,
Россия, grgazizova@gmail.com*

²*Университет Джунтендо, Токио, Япония*

³*Университет ИТМО, Санкт-Петербург, Россия*

⁴*Казанский институт биохимии и биофизики Казанского научного
центра РАН, Казань, Россия*

⁵*Сколковский институт науки и технологий, Москва, Россия*

⁶*ООО «ЛИФТ Центр», Москва, Россия*

Для переживания неблагоприятных условий окружающей среды, таких как снижение температуры и отсутствие пищи, ряд животных обладает уникальной способностью впадать в торпор – особое состояние организма, при котором происходит замедление метаболизма, снижение температуры тела, частоты сердечных сокращений и дыхательных циклов [1]. В торпор могут впадать некоторые виды птиц и млекопитающих - сумчатых, грызунов и летучих мышей [2]. У ряда видов это происходит в течение одного дня с целью сохранения энергии, в то время как у истинных гибернаторов, или зимнеящих животных, торпор может длиться до нескольких недель при сезонных изменениях окружающей среды [2]. Во время зимней спячки животные подвергаются многократным циклам торпора-пробуждения, а их скелетные мышцы находятся в длительной иммобилизации. При этом у них отсутствуют серьезные повреждения сосудов и органов, а мышцы избегают атрофии [3].

Соня-полчок (*Glis glis*) – мелкий древесный грызун, который при неблагоприятных условиях может находиться в спячке

до 11 месяцев в году [4] – является подходящей моделью для изучения зимней спячки.

В настоящей работе нами проведено исследование транскриптомного профиля в мышцах сонь-полчков в активном и торпидном состоянии с целью определения молекулярных механизмов, участвующих в регуляции процесса зимней спячки и поддержании мышечного гомеостаза во время длительной иммобилизации.

У двух групп животных: активных ($n=3$) и находящихся в фазе торпора ($n=3$) в течение 14 дней в контролируемых климатических условиях, были отобраны мышцы двух типов: «быстрый» длинный разгибатель пальцев (*m. EDL*) и «медленная» камбаловидная мышца (*m. soleus*).

На начальном этапе исследования были проведены секвенирование, сборка и аннотация генома сони-полчка. Для решения этих задач было использовано секвенирование длинных прочтений (HiFi) технологии Pacific Bioscience и метод определения конформации хромосом (Hi-C). Сборка генома была осуществлена с помощью геномного ассемблера Flye [5], аннотация генома – с помощью алгоритма Braker3 [6]. В результате был собран геном на уровне хромосом с размером N50 96,1 Мб и 95,6% полноты по BUSCO.

Для определения изменений профиля экспрессии генов в ответ на торпор был проведен анализ данных RNA-seq. С целью выявления ключевых регуляторных элементов и факторов транскрипции на уровне промоторов, специфичных для мышц, в ответ на зимнюю спячку у сони-полчка был использован CAGE-seq (Cap Analysis of Gene Expression).

В результате исследования было выявлено, что реакция на зимнюю спячку различается у скелетных мышц «быстрого» и «медленного» типов. В ответе на зимнюю спячку в скелетных мышцах участвуют гены, ассоциированные с регуляцией дифференцировки и пролиферации клеток, аутофагии, апоптоза, деградации и стабильности белка, wnt-сигнального пути. Анализ обогащения мотивов показал подавление активности многих

факторов транскрипции, активирующих ключевые гены гомеостаза скелетных мышц.

Дальнейшие исследования в этом направлении приведут к более глубокому пониманию стратегий выживания зимнеящих млекопитающих в экстремальных условиях окружающей среды.

Работа выполнена за счет средств субсидии, выделенной в рамках государственной поддержки Казанского (Приволжского) федерального университета в целях повышения его конкурентоспособности среди ведущих мировых научно-образовательных центров.

1. H.V.Carey et al (2003) Mammalian Hibernation: Cellular and Molecular Responses to Depressed Metabolism and Low Temperature, *Physiol Rev.*, 83(4):1153-81
2. T. Ruf, F. Geiser (2015) Daily torpor and hibernation in birds and mammals, *Biol Rev Camb Philos Soc.*, 90(3):891-926.
3. Y. Xu et al (2013) Molecular signatures of mammalian hibernation: comparisons with alternative phenotypes, *BMC Genomics*, 14:567.
4. F. Hoelzl et al (2015) How to spend the summer? Free-living dormice (*Glis glis*) can hibernate for 11 months in non-reproductive years, *J Comp Physiol B*, 185:931–939
5. M. Kolmogorov et al. (2019) Assembly of long, error-prone reads using repeat graphs, *Nat Biotechnol*, 37:540–546.
6. L. Gabriel et al (2024). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA, *Genome research*, 34(5):769–777

**Применение метода температурного протеомного
профилирования для изучения взаимодействия лекарств с
белковыми комплексами.**

Лейла Алтаевна Гарибова^{1,2}, Иван Игоревич Федоров^{1,2}, Дарья
Дмитриевна Емексева^{1,2}, Михаил Владимирович Горшков¹, Марк
Витальевич Иванов¹

¹ИНЭПХФ им. В.Л. Тальрозе ФИЦ ХФ РАН

²Московский физико-технический институт (национальный исследовательский университет), garibova.02@gmail.com

Одним из методов, позволяющий идентифицировать взаимодействие малых молекул с белками в рамках всего протеома, а следовательно найти новые мишени лекарственных препаратов и их возможные побочные эффекты, является химическая протеомика, в частности температурное профилирование протеома (TPP) [1]. TPP основан на определении точки денатурации белка при изменении температуры: в результате денатурации белки становятся нерастворимыми в соответствующем буферном растворе, что позволяет разделить их на растворимые и нерастворимые фракции. Кроме того, взаимодействие химического вещества с белком меняет его точку плавления. Масс-спектрометрия позволяет построить кривые растворимости для всех наблюдаемых в протеоме белков и выявить белковые мишени лекарств.

Несмотря на свои преимущества, метод TPP имеет ряд недочетов. Согласно стандартной теории, кривые растворимости одного и того же белка в контрольных и обработанных лекарством образцах совпадают при исходной температуре 37°C (рис.1). При этом стандартные подходы к проведению экспериментов не позволяют это наблюдать и подтвердить экспериментально. Изменив схему эксперимента, мы обнаружили, что для десятков белков это не выполняется и уже на начальной температуре отмечаются различия в интенсивности сигнала (рис. 2). Предположительно, лекарственные препараты связываются только с отдельными белками в белковом комплексе, но другие белки меняют свою растворимость при воздействии на комплекс и при переходе белков в свободную форму.

В данной работе предложен альтернативный подход на основе ультрабыстрого полнопротеомного анализа DirectMS1 [2], который позволяет выявлять не только прямые мишени препаратов,

но и белки, ассоциированные с ними в комплексах. В качестве модельной системы использованы клетки линии рака яичников A2780, обработанные противоопухолевыми препаратами (топотекан, 8-азагуанин) и экспериментальными соединениями.

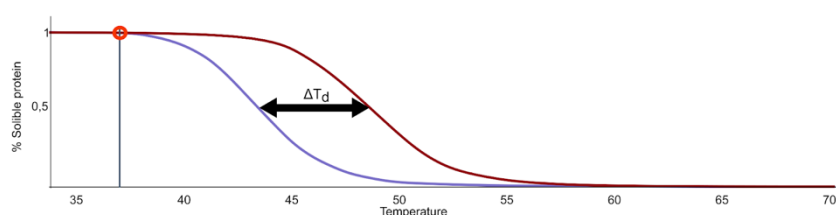


Рис. 1. Теоретические кривые растворимости белка в контрольном и обработанном лекарственным препаратом образцах. ΔT_d - разность температур плавления.

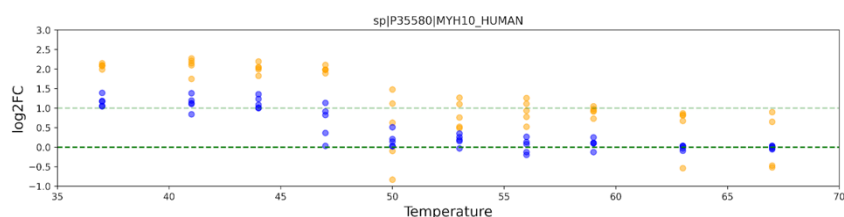


Рис. 2. График кривых растворимости целевого белка sp|P35580|MYH10_HUMAN со статистически значимым отклонением в температуре плавления между контрольными пробами и пробами, обработанными онкопрепаратом Топотекан.

1. И.И.Федоров и др. (2022) Химическая протеомика на основе масс-спектрометрии в задачах поиска лекарственных мишеней, БИОХИМИЯ, 87:1232–1245.
2. M.V.Ivanov et al. (2020) DirectMS1: MS/MS-Free Identification of 1000 Proteins of Cellular Proteomes in 5 Minutes, Analytical Chemistry, 92: 4326–4333.

Поиск вторичных структур ДНК (G-квадруплексов и Z-ДНК) в геноме Осьминога (*Octopus Bimaculoides isolate*)

Д.А.Гарипова, М.С.Попцова

*Международная лаборатория Биоинформатики ИИИиЦН
ФКН НИУ ВШЭ, dagaripova@edu.hse.ru*

Осьминоги представляют собой уникальную модель для изучения молекулярных механизмов эволюционной адаптации благодаря сложно организованному геному и высоко развитому поведению [2]. Одной из ключевых особенностей этих головоногих моллюсков является их способность к активному редактированию РНК с помощью ферментов семейства ADAR. [3] Этот процесс приводит к изменению кодирующих последовательностей мРНК и, как следствие, к образованию альтернативных вариантов белков, что может влиять на пластичность нервной системы и адаптацию к изменяющимся условиям среды. Важную роль в управлении экспрессией генов, репликацией ДНК и поддержанием стабильности генома играют вторичные структуры ДНК, такие как G-квадруплексы и Z-ДНК.

В настоящее время известно, что вторичные структуры ДНК могут влиять на редактирование РНК, поскольку ADAR-ферменты специфически связываются с двухцепочечными участками и содержат Z-РНК- и квадруплекс-связывающий домен. [4] Однако остается неясным, насколько часто в геноме осьминога встречаются G-квадруплексы и Z-ДНК; связаны ли эти структуры с генами, подверженными редактированию РНК; какие биологические функции могут быть ассоциированы с участками ДНК, образующими неканонические структуры.

В данной работе был проведен комплексный анализ вторичных структур ДНК (G-квадруплексов и Z-ДНК) в геноме осьминога *Octopus bimaculoides*. Поиск G-квадруплексов осуществлялся поиском по паттернам, а Z-ДНК программами Z-Hunt и моделью глубокого обучения Z-DNABERT. [1] В результате

анализа полного генома (2.7 Gb) было выявлено 74947 уникальных участков с G-квадруplexами, а при исследовании трех хромосом обнаружено 2159 участков с Z-ДНК.

Анализ распределения вторичных структур ДНК по геномным областям показал, что G-квадруplexы чаще встречаются в генах, контролирующих транскрипцию (T-box факторы) и посттрансляционные модификации (убиквитинирование), что подтверждает их роль в регуляции экспрессии генов. Было обнаружено обогащение G-квадруplexов в промоторных областях генов, кодирующих цинковые пальцы (ZFPs) и протокадгерин, что указывает на их потенциальную роль как универсальных регуляторных элементов, участвующих в развитии нервной системы, клеточной идентичности и ответе на стресс.

Также наблюдалось обогащение Z-ДНК в промоторных областях или вблизи сайтов начала транскрипции, что свидетельствует о ее возможном участии в регуляции экспрессии генов. Функциональный анализ генов, содержащих Z-ДНК, выявил несколько ключевых категорий: (1) адгезивные GPCR (LOC106873634, LOC106882816), участвующие в межклеточной коммуникации и нейробиологических процессах; (2) цинковые пальцы (LOC106873635, LOC106877807), связанные с регуляцией транскрипции; (3) ферменты, обеспечивающие антиоксидантную защиту и метаболизм (глутамат-цистеинлигаза, STEAP4); (4) транспортеры, важные для нейротрансмиссии (натрий-зависимый дофаминовый транспортер); (5) белки репарации ДНК и сплайсинга, поддерживающие стабильность генома (DNX37, ARMT1).

Полученные результаты подчеркивают значимость вторичных структур ДНК в регуляции ключевых биологических процессов у осьминога. G-квадруplexы и Z-ДНК демонстрируют статистически значимую ассоциацию с генами, участвующими в транскрипции, нейробиологии и клеточном метаболизме, что согласуется с их предполагаемой ролью в модуляции экспрессии генов. Особенно интересна их связь с нейробиологическими

функциями, что может отражать адаптационные механизмы, характерные для высокоорганизованной нервной системы головоногих. Дальнейшие исследования в этом направлении могут пролить свет на эволюционные аспекты регуляции генома у беспозвоночных.

1. Dmitry Umerenkov, Alan Herbert et al. (2023) Z-Flipon Variants reveal the many roles of Z-DNA and Z-RNA in health and disease, *Science*, 12:13–16. [Science]
2. Caroline B. Albertin, Oleg Simakov, Therese Mitros et al. (2015) The octopus genome and the evolution of cephalopod neural and morphological novelties, *Nature*, 12:13–16. [Nature]
3. Ting Zhang, Chaoran Yin et al. (2022) ADAR1 masks the cancer immunotherapeutic promise of ZBP1-driven necroptosis, *Nature*, 12:13–16. [Nature]
4. Alan Herbert 1, Oleksandr Cherednichenko et al. (2025) Z α and Z β Localize ADAR1 to Flipons That Modulate Innate Immunity, Alternative Splicing, and Nonsynonymous RNA Editing, *PubMed*, 12:13–16. [PubMed]

Хроматин-РНК взаимодействия: анализ РНК-компоненты интерактома

Л.Д. Гаркуль, А.А. Жарикова, А.А. Миронов

г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73, Факультет биоинженерии и биоинформатики, lidia.garkul@yandex.ru

Известно, что РНК играют важную роль в ядерной организации эукариотических клеток [1]. Относительно недавно начали появляться полногеномные эксперименты, позволяющие получать информацию о полном РНК-хроматиновом интерактоме,

без предварительного знания последовательностей РНК, ДНК или белков, участвующих в контактах [2]. В настоящей работе используются данные полногеномных протоколов, основанных на концепции "все-против-всех", где первым шагом является фиксация клеток формальдегидом, затем пространственно близкие РНК и ДНК части лигируются с использованием нуклеотидного линкера. Во всех таких экспериментах, на выходе получают химерные чтения - РНК и ДНК последовательности, которые в момент фиксации оказались пространственно близки. Большинство работ, посвященных анализу такого типа данных сконцентрированы на анализе ДНК-частей контактов. Однако, теоретически, информация о том, в какое место молекулы приходится РНК-часть контакта, может дать новые знания о природе взаимодействия РНК с ДНК-локусами. Целью данной работы является полный анализ РНК-частей контактов данных экспериментов "все-против-всех".

В ходе исследования использовались данные из различных экспериментальных протоколов, таких как GRID-seq, RADICL, Red-C и RedChIP, проведенные на клеточных линиях человека (*Homo sapiens*) и мыши (*Mus musculus*). Был предложен подход, основанный на фоновой модели, для выявления участков генов, которые могут играть важную роль во взаимодействии РНК и хроматина. С помощью этого подхода для каждого из датасетов были определены потенциально значимые участки РНК для взаимодействия с хроматином. Анализ показал, что выявленные участки, встречающиеся во всех датасетах, в основном относятся к длинным некодирующим РНК. Например, для длинной некодирующей РНК NEAT1 были обнаружены три потенциально значимых участка, которые совпадают с уже описанными "доменами" NEAT1 в экспериментальной литературе.

Дальнейшие планы включают более подробный анализ находок алгоритма и расширение ортологического анализа.

1. J.S.Mattick, P.P.Amaral et al. (2023) Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* 24: 430–447.

2. X. Li, XD. Fu (2019) Chromatin-associated RNAs as facilitators of functional genomic interactions. Nat Rev Genet 20: 503–519.

Транскрипционный фактор Nrf2 и эволюция защитных систем клетки: Вариабельность структуры и числа элементов антиоксидантного ответа (ARE)

К.В. Гордеев¹, А.А. Коковихина², Г.А. Шиловский³

*¹Кубанский государственный медицинский университет
Министерства Здравоохранения Российской Федерации, Россия,
350063, Краснодар, Митрофана Седина, 4,
gordeev.kirill.loremipsum@gmail.com*

*²Ижевский государственный технический университет имени
М.Т. Калашикова, Россия, 426069, Ижевск, Студенческая ул., 7*

*³Московский государственный университет имени М.В.
Ломоносова, Россия, 119992, Москва, Ленинские горы, д. 1*

Motivation and Aim: Транскрипционный фактор Nrf2 участвует в различных клеточных процессах, таких как реакция на окислительный стресс, митохондриальное дыхание, покой стволовых клеток, трансляция мРНК, аутофагия и ответ на несвернутые белки [1]. Lewis и соавт. [2] предполагают, что те виды, которые проявляют высокую конститутивную активность Nrf2, быстрее реагируют и на окислительный стресс, что позволяет достигать более высокой продолжительности здоровой жизни. Nrf2 занимает центральное место в сложной регуляторной сети, существенно варьирует в зависимости от клеточного типа, а также в зависимости от различных стимулов, распознавания элемента антиоксидантного ответа (antioxidant response element, ARE), наличия партнеров по связыванию, конкуренции с другими активаторами и репрессорами [1]. В частности, Nrf2 индуцирует

экспрессию антиоксидантных ферментов II фазы, которые нейтрализуют активные формы кислорода и тем самым обеспечивают защиту клетки от повреждений, вызванных окислительным стрессом. ARE, представляя собой цис-регуляторные элементы, локализованные в промоторных участках генов-мишеней транскрипционного фактора Nrf2, играют ключевую роль в поддержании окислительно-восстановительного гомеостаза клетки. Цис-регуляторным элементом ARE служит участок, содержащий последовательность нуклеотидов 5'-A/GTGAC/TnnnGCA/G-3' («ядро») [3]. Разнообразие в последовательностях ARE не только обеспечивает регуляторную селективность разнообразных ТФ, но также обуславливает функциональную сложность пути Keap1 / Nrf2 / ARE. [4]. Все ARE могут быть разделены на четыре структурно-функциональных класса на основании различий в нуклеотидной последовательности [4]. Изучение различий ARE-содержащих последовательностей помогает понять различия в уровне экспрессии генов-мишеней Nrf2 в разных клеточных типах у разных видов.

Methods and Algorithms: Выравнивание последовательностей проводили с помощью MAFFT. Сравнивали количество и нуклеотидную последовательность четырех типов функциональных последовательностей ARE у 25 видов позвоночных, включая человека (*Homo sapiens*) и ряда модельных видов (мышь *Mus musculus*, крысы *Rattus norvegicus* и данио *Danio rerio*), а также голого землекопа (*Heterocephalus glaber*).

Results: Нами предложен новый подход идентификации и анализа ARE. В результате обширного биоинформатического исследования промоторных последовательностей конкретных генов-мишеней Nrf2 были идентифицированы функциональные последовательности ARE у 25 видов позвоночных, принадлежащих к 18 отрядам. Вариабельность по меньшей мере трех из четырех видов последовательности ARE различается в разных отрядах позвоночных.

Conclusion: Считается, что гены, промоторы которых содержат последовательности ARE первого и второго классов, в ответ на

активацию Nrf2 отвечают большей степенью индукции и находятся под меньшим негативным контролем других транскрипционных факторов семейства bZip. Гены с ARE первого и третьего классов, в силу содержания последовательности TRE, отвечают также на активацию транскрипционного фактора AP-1 [4]. Таким образом, вариабельность нуклеотидной последовательности ARE, а также тот факт, что некоторые гены (например, ген гемоксигеназы человека и грызунов) содержат в промоторах несколько последовательностей ARE, относящихся к разным классам, делает возможной тонкую и гибкую регуляцию экспрессии подконтрольных генов в ответ на стрессовые воздействия. Исследование выполнено за счет Госзадания, биологического факультета МГУ имени М.В. Ломоносова

1. Tonelli C., Chio I.I.C, Tuveson D.A. Transcriptional Regulation by Nrf2. *Antioxid. Redox Signal.* 2018;29(17): 1727-1745. doi 10.1089/ars.2017.7342.
2. Lewis K.N., Wason E., Edrey Y.H., Kristan D.M., Nevo E., Buffenstein R. Regulation of Nrf2 signaling and longevity in naturally long-lived rodents. *Proc. Natl. Acad. Sci. USA.* 2015;112(12):3722-3727. doi 10.1073/pnas.1417566112.
3. Wasserman W.W., Fahl W.E. Functional antioxidant responsive elements. *Proc. Natl. Acad. Sci. USA.* 1997;94(10):5361-5366. doi 10.1073/pnas.94.10.5361.
4. Hayes J.D., McMahon M., Chowdhry S., Dinkova-Kostova A.T. Cancer chemoprevention mechanisms mediated through the Keap1-Nrf2 pathway. *Antioxid. Redox Signal.* 2010;13(11):1713-1748. doi 10.1089/ars.2010.3221.

Исследование редких генетических вариантов на основе результатов секвенирования экзома

А.С. Грачева^{1,2}, М.Б. Хаджиева^{1,3}, Д.А. Кашатникова², И.В. Редкин¹, В.Е. Захарченко¹, Т.Н. Крылова¹, А.Н. Кузовлев¹, Л.Е. Сальникова^{1,2,3}

¹*Федеральное государственное бюджетное научное учреждение "Федеральный научно-клинический центр реаниматологии и реабилитологии" (ФНКЦ РР), Россия, г. Москва*

²*Федеральное государственное бюджетное учреждение науки Институт общей генетики имени Н.И. Вавилова Российской академии наук, Россия, г. Москва
palesa@yandex.ru*

³*Федеральное государственное бюджетное учреждение "Национальный медицинский исследовательский центр детской гематологии, онкологии и иммунологии имени Дмитрия Рогачева" Министерства здравоохранения Российской Федерации, Россия, г. Москва*

В представленной работе авторы провели исследование редких потенциально патогенных (оказывающих высокое деструктивное воздействие на белок, далее «НП» - high-impact) генетических вариантов у пациентов в критическом состоянии или с высоким риском его развития. Анализ редких вариантов проведен по результатам секвенирования экзома исследуемых пациентов в контексте омнигенной гипотезы, согласно которой, ассоциативные сигналы распространяются по большей части генома через сети геной регуляции от генов, находящихся вне основных путей (периферических), к генам, непосредственно связанным с заболеванием (коровым) [1]. При этом любой ген, экспрессирующийся в ткани, поражаемой при заболевании, может вносить вклад в фенотип. Дальнейшие исследования показали, что

подход, лежащий в основе омнигенной гипотезы и разработанный на основе данных общегеномных исследований распространенных генетических вариантов, справедлив и для редких вариантов [2-4]. Целью данной работы являлся анализ кумулятивных эффектов редких вариантов в нескольких когортах тяжелобольных пациентов: с COVID-19, с черепно-мозговой травмой (далее ЧМТ), сопутствующей анемией.

В исследование были включены пациенты с критическим заболеванием или с высоким риском его развития (77 пациентов с острым заболеванием COVID-19; 50 хронических пациентов с умеренной/тяжелой ЧМТ в анамнезе). Редкие НІ-варианты (сдвиг рамки считывания, преждевременное образование стоп-кодонов, мутации акцепторного и донорного сайтов сплайсинга, мутации иницирующих и терминальных кодонов) представлены вариантами с аллельной частотой (AF) $<0,001$ и без данных AF в популяционных базах данных. Подробный анализ методов подробно представлен в предыдущих работах авторов [5-7].

В результате исследования у пациентов с тяжелой формой COVID-19 был выявлен избыток редких НІ-вариантов, при этом самый высокий ассоциативный эффект выявлен в наборе генов, ассоциированных с первичными иммунодефицитами. У пациентов с ЧМТ наблюдалось значительное преобладание редких НІ-вариантов в генах, связанных с заболеваниями нервной системы. В обеих исследуемых когортах пациентов была показана ассоциация генов сигнального пути Notch с развитием анемии.

Таким образом, большое количество нарушений, вызванных суммарным влиянием НІ-вариантов в предполагаемых коровых или близких к коровым генах, может приводить к ослаблению компенсаторных механизмов, эффективных в здоровом состоянии и способствовать развитию тяжелого течения заболевания и долгому восстановлению.

Исследование проведено при поддержке Государственного задания Минобрнауки России № FGWS-2025-0008 «Гипоксия критических состояний - ведущие механизмы развития, персонализированная диагностика и лечение».

1. E.A. Boyle, Y.I. Li, J.K. Pritchard (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic, *Cell*, 169(7): 1177–1186.
2. N.R. Wray et al. (2018) Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model, *Cell*, 173(7):1573–1580.
3. V. Pullabhatla et al. (2018) De novo mutations implicate novel genes in systemic lupus erythematosus, *Hum Mol Genet*, 27(3): 421–429.
4. M.A. Scelsi et al. (2021) Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Alzheimer's Disease Sequencing Project (ADSP). Network propagation of rare variants in Alzheimer's disease reveals tissue-specific hub genes and communities, *PLoS Comput Biol*, 17(1): e1008517.
5. M.B. Khadzhieva et al. (2023) COVID-19 severity: does the genetic landscape of rare variants matter? *Front Genet*, 29: 14:1152768.
6. A.S. Gracheva et al. (2024) Genetics and Traumatic Brain Injury: Findings from an Exome-Based Study of a 50-Patient Case Series, *Curr Issues Mol Biol*, 46(9):10351-10368.
7. D.A. Kashatnikova et al. (2025) Red Blood Cell-Related Phenotype-Genotype Correlations in Chronic and Acute Critical Illnesses (Traumatic Brain Injury Cohort and COVID-19 Cohort), *Int J Mol Sci*, 26(3):1239.

**Оптимизация RFdiffusion для генерации байндеров к Z-ДНК
связывающим белкам путём добавления потенциалов**

Д.А. Громак, М. С. Попцова

109028, г. Москва, Покровский бульвар, д. 11, dgromak@hse.ru

Левосторонне закрученная конформация спирали ДНК, называемая Z-ДНК, играет значительную роль в регуляции транскрипции, иммунном ответе и других клеточных процессах [1]. Белки, способные связываться с этой формой ДНК, являются медиаторами

данных процессов, что делает их привлекательным объектом для изучения. Однако формирование Z-ДНК в естественной среде является динамическим многоэтапным процессом, поэтому изучение активности вышеупомянутых белков вызывает значительные затруднения.

Чтобы решить эту проблему, подходит метод создания байндеров с помощью генеративных нейронных сетей, в частности диффузионных моделей. Данная работа фокусируется на подходе, который основан на подходе, использующем предобученные нейросетевые модели RFdiffusion, ProteinMPNN и Alphafold [2].

На первом этапе генерируются координаты атомов в полипептидном остове будущего байндера. В отличие от других методов [3] [4], предложенных для генерации байндеров, структуры получаемых белков не заданы заранее, а являются абсолютно новыми. Для этого применяется обратный диффузионный процесс: обученная модель RFdiffusion последовательно из гауссовского шума восстанавливает пространственные положения атомов. Второй этап состоит из восстановления аминокислотной последовательности, которая сворачивалась бы в сгенерированный каркас. Для этого используется модель ProteinMPNN, основанная на алгоритме передачи сообщений. На третьем этапе проводится оценка полученных дизайнов. Для этого используется модель Alphafold, предсказывающая по последовательности вторичную структуру получившегося байндера. Для определения качества байндера используются метрики iptm и ipae.

Таким образом, данный подход позволяет установить структуры, которых нет в современных библиотеках белковых мотивов. Однако для конкретных задач процент правдоподобных дизайнов может быть относительно мал. В данной работе были разработаны 2 потенциала, которые направляют процесс генерации в сторону структур, которые похожи на Z-ДНК, а также свободная энергия комплекса этих структур с доменом белком мишени минимальна. Применение этих потенциалов позволило значительно увеличить процент успешных дизайнов. Для лучших из полученных дизайнов с помощью BLAST были найдены белки, домены которых похожи

на сгенерированные структуры и могут потенциально связываться с белком-мишенью.

1. S. Ravichandran, V. K. Subramani и К. К. Kim (2019) “Z-DNA in the genome: from structure to disease”, *Biophysical Reviews*, 11(3):383–387.
2. J.L. Watson et al. (2023) De novo design of protein structure and function with RFdiffusion, *Nature*, 620(7976):1089-1100.
3. L.Cao et al. (2022) Design of protein binding proteins from the target structure alone, *Nature*, 605(7910):551-560.
4. P.Gainza et al. (2023) De novo design of protein interactions with learned surface fingerprints, *Nature*, 617(7959):176-184.

Моделирование регуляторной грамматики с использованием нейросетей для предсказания активной транскрипции с однонуклеотидным разрешением

Н.С. Грызунов¹, Д.Д. Пензар¹, И.В. Кулаковский²

¹119234, г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73,
Факультет биоинженерии и биоинформатики МГУ им. М.В.
Ломоносова
nikgr127@gmail.com

²142290, Московская область, г. Пущино ул. Институтская, 4,
Институт белка РАН

Мотивация и цель: Расшифровка регуляторной грамматики некодирующих участков генома помогает понять молекулярные механизмы регуляции инициации транскрипции и предсказывать влияние генетических вариантов на экспрессию генов. Основным подходом является создание моделей, способных предсказывать участки доступного хроматина и профили инициации транскрипции по геномной последовательности. В последние годы появляются модели, нацеленные на предсказание активной инициации транскрипции с однонуклеотидным разрешением, например по

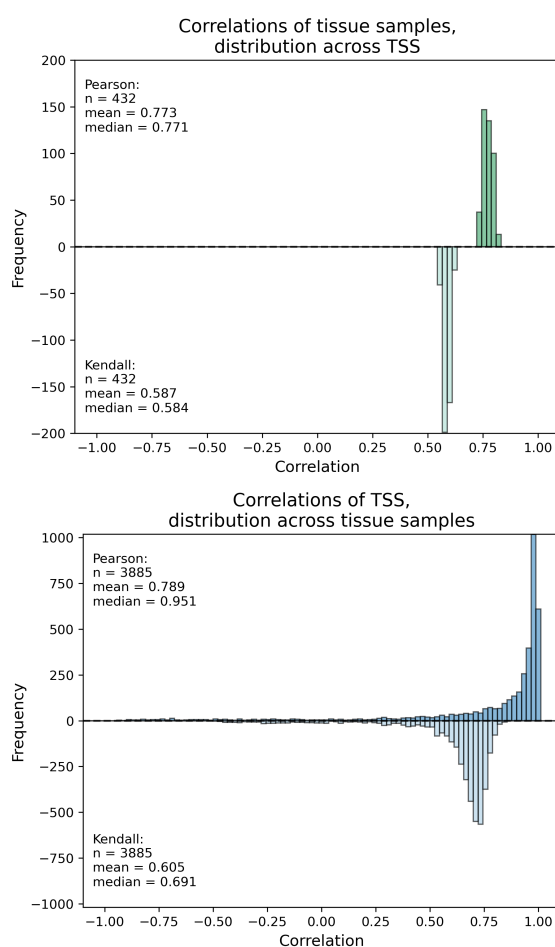
данным экспериментов CAGE-Seq (кэп-анализ экспрессии генов), такие как Puffin [1]. Однако существующие модели имеют ряд недостатков: ограниченная точность, вызванная упрощенной или ошибочно выбранной архитектурой, некорректное использование функции потерь при обучении и, как следствие, низкая точность предсказания эффектов промоторных вариантов [2]. Еще одна фундаментальная проблема — использование в обучении агрегированных данных, объединяющих профили экспрессии из различных тканей, клеточных линий и состояний. Такой подход игнорирует клеточно- и тканеспецифические особенности регуляции.

В связи с этим особенно интересным объектом исследования являются скелетные мышцы, которые обладают выраженным транскрипционным разнообразием, как было показано на образцах крыс и мышей [3], но при этом остаются недоисследованными и слабо представлены в атласах экспрессии генов, а многие особенности регуляции транскрипции, свойственные им, в полной мере не изучены.

Методы: Консорциум FANTOM более 10 лет назад предоставил данные об активности сайтов инициации транскрипции в одонуклеотидном разрешении — профили CAGE-Seq различных человеческих образцов, агрегированных воедино, куда было включено лишь небольшое число образцов скелетных мышц. В то же время, недавно сотрудниками Казанского федерального университета были получены результаты CAGE-Seq для 76 различных скелетных мышц человека. Для построения нейросетей, моделирующих профили транскрипции по данным CAGE, использовались инструменты глубокого обучения, основанные на фреймворках PyTorch и PyTorch Lightning.

Результаты: Мы разработали новую архитектуру нейросети, способную на основе геномной последовательности в одонуклеотидном разрешении предсказывать профиль стартов сайтов транскрипции, специфичный для различных типов мышц. Новая архитектура имеет в 20 раз меньшее число параметров и достигает более высокой точности предсказаний, чем нейросеть Puffin, а также имеет гораздо более высокую скорость обучения. По тотальному покрытию в геномных окнах модель достигает средней корреляции Пирсона 0.79 для отдельных образцов (корреляция по

генам в пределах образца, рис. 1a) и 0.77 в рамках отдельных сайтов инициации транскрипции (корреляция по образцам в пределах гена, рис. 1b). Корреляция однонуклеотидных профилей между предсказанными и измеренными значениями составила 0.52 (рис. 1c). Полученная модель позволяет в дальнейшем провести анализ используемых моделью признаков («объяснимости») и оценить влияние однонуклеотидных промоторных вариантов на предсказанные величины экспрессии генов.



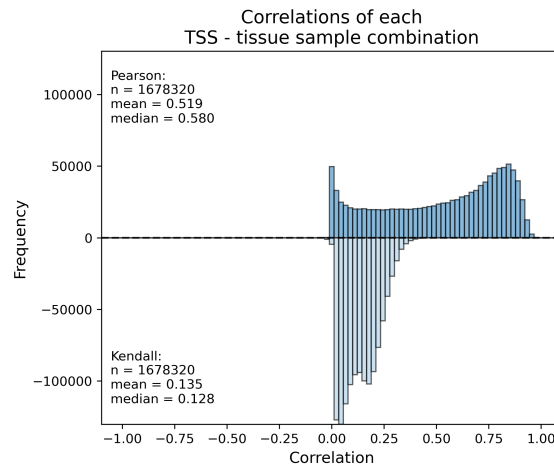


Рисунок 1. Корреляции между истинными и предсказанными значениями: (а) ген-специфичная и (б) тканеспецифичная корреляции суммарных покрытий сайтов старта транскрипции, а также (с) корреляция покрытий в разрешении 1 нт.

1. K.Dudnyk et al (2024) Sequence basis of transcription initiation in the human genome, *Science*, **384** (6694):eadj0116.
2. J.F.Kribelbauer-Swietek et al (2024) EXTRA-seq: A genome-integrated extended massively parallel reporter assay to quantify enhancer-promoter communication, *bioRxiv*.
3. E.E.Terry et al (2018) Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues, *eLife*, **7**:e34613

Модификация линейного индекса инструмента выравнивания minimap2 с учетом известных генетических вариантов

Е. П. Гугучкин¹, А. С. Касьянов², Е. А. Карпулевич¹

¹ИСП РАН, г. Москва, Россия, guguchkin@ispras.ru

²BIOPOLIS-Cibio, г. Вайран, Португалия

Благодаря удешевлению и совершенствованию технологий массового параллельного секвенирования (Next-Generation Sequencing, NGS) секвенирование полного генома человека стало рутинной процедурой в биомедицинских исследованиях. Во многих странах и научных организациях происходит переход от единичных исследований геномов отдельных индивидов к массовому секвенированию групп индивидов и крупных популяционных выборок. Одним из ключевых этапов при анализе выходных данных NGS секвенатора, является выравнивание коротких прочтений (сегментов ДНК, полученных в процессе секвенирования) относительно референсного генома (эталонной последовательности ДНК генома организма соответствующего биологического вида). Несмотря на то, что референсный геном человека существует в нескольких версиях (например, GRCh37, GRCh38, T2T), он все еще представляет собой усредненную последовательность нуклеотидов, без учета частот генетических вариантов конкретной популяции. Большинство алгоритмов выравнивания коротких прочтений на линейный референс для ускорения своей работы используют двухэтапный подход: сначала происходит выравнивание якорей (Рис. 1) и поиск потенциальных позиций расположения короткого прочтения относительно референсного генома, на втором этапе происходит точное выравнивание короткого прочтения, например, с помощью алгоритмов Смита-Вотермана или Нидлмана-Вунша. У генома конкретного человека отличия от референсного генома, такие как точечные различия (Single Nucleotide Variants, SNV), вставки или делеции (Indels) могут встретиться в якорных последовательностях что приведет к снижению качества выравнивания коротких прочтений и качества последующего этапа идентификации генетических вариантов. Наличие генетических вариантов в якорных последовательностях делает задачу корректного выравнивания прочтений нетривиальной, поскольку при наличии в прочтении вариантов, которые не учтены в референсе, часть алгоритмов будет «исправлять» такие отличия или понижать

качество выравнивания короткого прочтения (Mapping quality, MAPQ), а в некоторых случаях даже приводить к неверному выравниванию короткого прочтения относительно референса.

Один из наиболее распространенных подходов к выравниванию коротких прочтений — это метод «seed-and-extend» [1]. Основная идея метода заключается в разбиении как коротких прочтений, так и референсной последовательности на относительно короткие фрагменты (k-меры), с последующим поиском точных совпадений k-меров с помощью специальных структур данных (например, хеш-таблиц, суффиксных массивов и их более компактных или выборочных версий). Затем выполняется построение наиболее вероятных цепочек коллинеарных совпадений (так называемые «цепочки якорей») и уже по ним происходит точное выравнивание с помощью более вычислительно затратных алгоритмов (например, метод Смита–Ватермана). Один из наиболее популярных инструментов выравнивания коротких прочтений — minimap2[2]. Его подход к выравниванию коротких прочтений основан на вычислении «минимайзеров», которые представляют собой особым образом выбранные k-меры, имеющие минимальное лексикографическое значение в скользящем окне заданной длины. Такие минимайзеры формируют «якоря», по которым можно быстро сопоставлять фрагменты прочтения с позициями таких же якорей в

референсе.

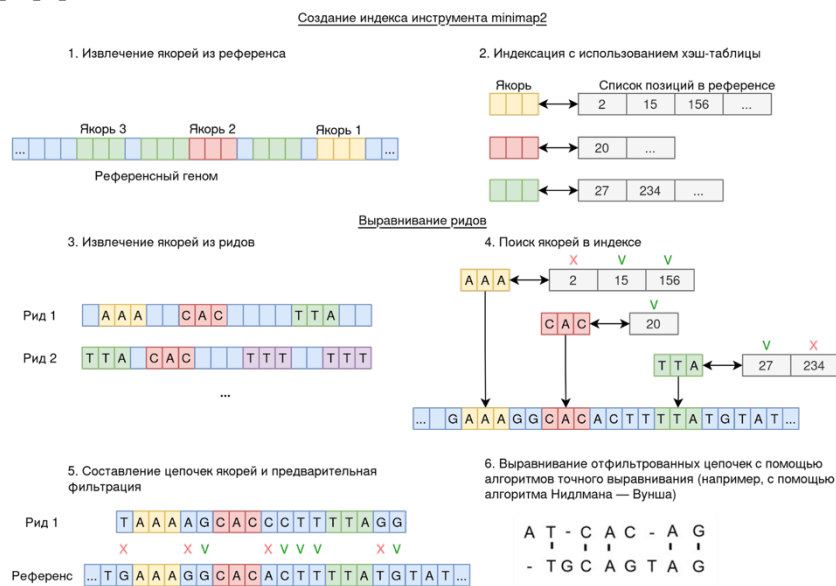


Рис. 1. Схема работы инструмента выравнивания minimap2.

В данном исследовании мы разработали подход[3], позволяющий дополнить индекс популярного инструмента minimap2, ориентированного на линейный референс, дополнительными k-мерами, которые отражают ранее описанные в базах данных генетические варианты. При этом сама эталонная последовательность остается линейной, однако при формировании хэш-таблицы в индекс включаются вспомогательные элементы, содержащие альтернативные нуклеотиды либо короткие вставки/делеции в соответствующих позициях. Благодаря этому, если анализируемое прочтение несет такой вариант, алгоритм minimap2 сопоставляет его с дополнительными k-мерами и корректно определяет положение прочтения без снижения качества выравнивания или искажения результата, которые обычно возникают при отсутствии известных альтернативных аллелей в референсе.

Результаты экспериментов показали, что включение вариантов в индекс minimap2 заметно улучшает качество выявления SNP и Indels по сравнению со стандартным выравниванием. Максимальный прирост был достигнут при использовании данных из Human Pangenome Reference Consortium (HPRC)[4], поскольку эта база содержит варианты, проверенные на уровне сборок, и тем самым обеспечивает присутствие в индексе ряда «нестандартных» аллелей. В итоге общее число ложноположительных и ложноотрицательных вызовов сократилось на несколько тысяч вариантов по всему геному.

В качестве расширения разработанного подхода был реализован метод, позволяющий учитывать заранее заданные длинные инсерции путём модификации референсной последовательности. Инсерции, превышающие базовую длину k-меров, используемых в инструменте minimap2, напрямую добавлялись в референс. В процессе модификации формировался chain-файл для последующего обратного лифтовера, а также список «делений» — для восстановления исходной референсной последовательности. Как и в случае с базовым подходом, в данном методе также используется модифицированный индекс minimap2. Полученные на расширенную референсную последовательность выравнивания впоследствии конвертировались обратно к исходной референсной с помощью инструмента levioSAM2 [5].

В перспективе мы рассматриваем возможность генерации дополнительных контигов, локально содержащих крупные структурные варианты, отсутствующие в линейной референсной последовательности. Эти контиги могут быть интегрированы в основной референс в виде вспомогательных хромосом, на которые будет выполняться выравнивание.

Type	Index	TP	FN	FP	Recall	Precision	F1-Score
SNV	default	334105	2407	421	0.9928	0.99874	0.9957

Type	Index	TP	FN	FP	Recall	Precision	F1-Score
s		3	4	2	5		9
	modified	3341206	23921	4168	0.99289	0.99875	0.99581
	extended	3348388	16739	3650	0.99503	0.99891	0.99697
indels	default	522445	3024	1121	0.99425	0.99786	0.99605
	modified	522453	3016	1116	0.99426	0.99787	0.99606
	extended	522805	2664	1112	0.99493	0.99788	0.99640

Таблица 1. Сравнение полученных VCF-файлов с эталонным VCF образца HG002 при использовании стандартного индекса minimap2 («default»), модифицированного добавлением заданных SNV («modified»), а также индекса, основанного на референсной последовательности, расширенной инсерциями («extended»).

Список литературы

1. Ahmed, N., Bertels, K., & Al-Ars, Z. (2016, December). A comparison of seed-and-extend techniques in modern DNA read alignment algorithms. In 2016 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 1421-1428). IEEE.

2. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), 3094-3100.
3. Guguchkin, E., Kasianov, A., Belenikin, M., Zobkova, G., Kosova, E., Makeev, V., & Karpulevich, E. (2024). Enhancing SNV identification in whole-genome sequencing data through the incorporation of known genetic variants into the minimap2 index. *BMC bioinformatics*, 25(1), 238.
4. Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., ... & Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960), 312-324.
5. Chen, N. C., Paulin, L. F., Sedlazeck, F. J., Koren, S., Phillippy, A. M., & Langmead, B. (2024). Improved sequence mapping using a complete reference genome and lift-over. *Nature methods*, 21(1), 41-49.

Вода и воздух - главные драйверы мутационного спектра митохондриальной ДНК птиц

Ю.С. Гусаров, В.О. Бурская, А.В. Бушуев, А.Г. Михайлова, Б.Э.
Ефименко, К.В. Гунбин, К.Ю. Попадъин

Центр Геномных Исследований БФУ им. И. Канта
yurgusguss@mail.ru

Митохондрия - одна из двух двумембранных органелл животной клетки. Она обладает собственной кольцевой молекулой ДНК (далее - мтДНК) - наследие от ее предка - бактерии. Ключевая функция митохондрий - производство энергии. Осуществляется она благодаря процессу окислительного фосфорилирования. Побочным продуктом данного процесса являются различные химические агенты, которые могут негативно влиять на незащищенные участки мтДНК. Данные участки появляются во время репликации ДНК. Этот процесс происходит неравномерно: во время синтеза дочерней тяжелой цепи с материнской легкой, материнская тяжелая цепь удерживается в одноцепочечном состоянии белками [1]. Именно в

этот момент опасные химические агенты окислительного фосфорилирования могут негативно повлиять на данную цепь, вызывая различные мутации. Самый частый вид мутаций, вызванных данными химическими воздействиями, это замена аденина на гуанин на тяжелой цепи мтДНК (далее A>G).

За последнее время появилось несколько интересных работ нашей лаборатории, указывающих на причины увеличения частот мутаций A>G для некоторых групп животных. Первая - это исследование млекопитающих. Наша команда выяснила, что долгоживущие млекопитающие (например, кит) накапливают за свою жизнь большое количество мутаций данного типа в мтДНК. А короткоживущие млекопитающие, несмотря на более интенсивный метаболизм и, как следствие, частый процесс окислительного фосфорилирования, накапливают меньше данных мутаций [2]. Вторая - исследование рыб. При исследовании тепловодных и холодноводных рыб были открыты некоторые закономерности: обитатели горячих вод накапливают больше мутаций A>G мтДНК, а холодных вод меньше [3]. Также в этой работе проводилось исследование по сравнению количества мутаций A>G у всех групп позвоночных животных. И птицы были лидеры по числу накопленных мутаций [3]. Поэтому это исследование посвящено выяснению первопричин мутагенеза мтДНК птиц.

Работа выполнена на собственном созданном наборе данных, включающем генетическую информацию о мтДНК 766 видов птиц из данных NCBI RefSeq [4], экологическую информацию из базы данных AVONET [5], а также мутационные спектры мтДНК птиц. Самая высокоточные данные данных спектров получены с помощью лабораторного пайплайна NeMu [6].

Исследование выполнялось для нуклеотидов в четырежды-вырожденных нейтральных позициях. Для начала мы рассчитали метрику интенсивности мутагенеза по нашей мутации GhAhSkew (форм. 1).

$$GhAhSkew = \frac{G - A}{G + A}$$

Формула 1. Метрика интенсивности мутагенеза мтДНК по мутации A>G

Данная метрика тем выше, чем больше у нас гуанинов и меньше аденинов в мтДНК. Благодаря ей мы выяснили, что у птиц во всех генах, кроме ND6, который находится на легкой цепи, количество мутаций A>G достигает критической массы. Особенно это хорошо видно при сравнении птиц с млекопитающими. В каждом гене тяжелой цепи метрика GhAhSkew птиц была ~ в 1,5 раза выше по сравнению с млекопитающими.

Мы начали искать причины такого переизбытка количества мутаций у птиц. Для начала мы взяли несколько простых физиологических признаков (масса тела, размер кладки, продолжительность жизни и базальный уровень метаболизма) и начали их сравнивать с нашей метрикой. Статистические анализы не показали каких-либо значимых результатов. Для подтверждения мы также провели PGLS - анализ. Он также не показал статистически значимых результатов, но выдал высокие значения λ Пагеля для нашей независимой переменной - GhAhSkew, что указывает на наследуемость и филогенетическую инерцию нашей метрики.

После неудачных результатов с физиологическими признаками, мы стали искать более сложные экологические признаки, которые могли нам дать ответы. Ключевыми оказались два признака - потеря способности к полету и приобретение способности к нырянию. Группа птиц из инфракласса *Palaeognathae* (страусы, киви и т.д.), потерявших давно способность к полету, а также виды из различных отрядов инфракласса *Neognathae*, независимо потерявших способность к полету (например, совиный попугай), обладали низкими значениями нашей метрики по сравнению с летающими птицами. Единственным исключением был отряд *Sphenisciformes* (пингвинообразные) - у них были повышенные значения GhAhSkew. Но среди всех нелетающих птиц пингвины - единственные ныряльщики. Поэтому мы исследовали этот признак. Все крупные отряды птиц - ныряльщиков (особенно *Anseriformes* (гусеобразные) и *Sphenisciformes*) обладали высокими показателями GhAhSkew по сравнению с птицами, неспособными к нырянию. Для подтверждения своих наблюдений мы также провели PGLS-анализ, показавший статистически значимые результаты.

Для дальнейшего исследования мы восстановили предковые состояния нашей метрики GhAhSkew для трех групп интереса: *Palaeognathae*, *Anseriformes* и *Sphenisciformes*. У предков *Palaeognathae* не было высоких значений GhAhSkew, и они остались на том же уровне у современных птиц. У предков *Anseriformes* и *Sphenisciformes* наблюдается одинаковая картина: как только предок данных групп приобретает способность к нырянию, значение GhAhSkew потомков растет по экспоненте и достигает критических масштабов у современных видов.

Для завершения нашего исследования мы провели анализ мутационных спектров[6]. Мы провели PGLS - анализ наших мутационных спектров для 766 видов птиц с их физиологическими и экологическими признаками. К нашему удивлению, ни один из результатов не был статистически важен. Также мы посмотрели на мутационный спектр предков наших групп интереса: *Palaeognathae*, *Anseriformes* и *Sphenisciformes*. У предков *Palaeognathae* не наблюдалось высоких частот мутаций A>G, а вот у *Anseriformes* и *Sphenisciformes* наоборот - частота мутаций A>G у предков достигла ~0,6, а у современных видов ~0,3. Затем мы проанализировали более точные данные мутационных спектров, полученные с лабораторного пайплайна NeMu [6]. Вследствие ограниченности доступных для исследования видов мы только сравнили частоты мутации A>G для гена CytB с массой и продолжительностью жизни птиц. Линейная модель показала отрицательную корреляцию для частот мутаций A>G с массой и продолжительностью жизни. Все эти результаты подтолкнули нас к некоторым предположениям. Во-первых, долгоживущие птицы с большой массой тела, скорее всего, вырабатывают какие-либо защитные механизмы, способствующие снижению химического урона мтДНК. Это было показано для некоторых представителей *Palaeognathae* [7]. Во-вторых, все современные виды с тяжелым грузом мутаций, вероятно, находятся под действием сильного стабилизирующего отбора, чтобы сохранить оставшиеся аденины в мтДНК. Но эта гипотеза требует дальнейшей проверки.

По итогу, мы выяснили, что именно экологические признаки, как потеря способности к полету и приобретение способности к нырянию, способствуют изменению мутационного спектра мтДНК птиц. Потеря способности к полету предками

Palaeognathae способствовала низкому мутационному грузу современных видов данной группы. Мутационный груз современных птиц - ныряльщиков - это вина их предков, приспособившихся к подводной добыче пищи. На эти результаты указывают как данные нашей мутационной метрики, так и восстановленные мутационные спектры.

Исследование выполнено при финансовой поддержке гранта РФФИ № 21-75-20143

Список литературы:

1. Falkenberg M. (2018) Mitochondrial DNA replication in mammalian cells: overview of the pathway. *Essays Biochem.* 20;62(3):287-296.
2. Mikhailova A.G. et al. (2022) A mitochondria-specific mutational signature of aging: increased rate of A > G substitutions on the heavy strand. *Nucleic Acids Res.* 14;50(18):10264-10277.
3. Alina G. Mikhailova et al. (2020) A mitochondrial mutational signature of temperature in ectothermic and endothermic vertebrates. *BioRxiv* 2020.07.25.221184
4. O'Leary N. A. et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research.* 44;1:733-745.
5. Tobias, J.A, et al. (2022) AVONET: morphological, ecological and geographical data for all birds. *Ecology Letters*, 25, 581–597.
6. Efimenko B et al. (2024) NeMu: a comprehensive pipeline for accurate reconstruction of neutral mutation spectra from evolutionary data. *Nucleic Acids Res.* 5;52:108-115.
7. Calhoun E.A. et al. (2014) Linkages between mitochondrial lipids and life history in temperate and tropical birds. *Physiol Biochem Zool.* 87(2):265-75.

**Динамика молекулярных возрастных изменений у
млекопитающих с различной продолжительностью жизни**

М.С. Давитадзе¹, А.Э. Тышковский², С.Е. Дмитриев³

¹*Московский государственный университет имени
М.В. Ломоносова, факультет биоинженерии и биоинформатики,
Ленинские горы, 1, стр. 73, Москва, 119992,
maria.davitadze@mail.ru*

²*Женская больница Бригема, Гарвардская медицинская школа,
США, Бостон, 02115*

³*НИИ физико-химической биологии имени А.Н. Белозерского,
Ленинские горы, 1, стр 40, Москва, 119234*

Один из основных вопросов геронтологии заключается в том, почему продолжительность жизни даже эволюционно близких видов может варьироваться в крайне широких пределах. Несмотря на значительный прогресс в изучении механизмов старения на клеточном и молекулярном уровнях, причины столь выраженных межвидовых различий в скорости старения и продолжительности жизни до конца не ясны. Хотя известно, что продолжительность жизни коррелирует с такими признаками, как масса тела, скорость развития и возраст достижения половой зрелости, эти факторы лишь частично объясняют наблюдаемые различия. Молекулярные механизмы, лежащие в основе этих различий, до сих пор остаются неясными.

В последние годы значительное внимание уделяется эпигенетическим биомаркерам старения, в особенности на уровне метилирования ДНК. В нескольких работах было показано, что уровень метилирования в определенных участках генома коррелирует с продолжительностью жизни различных видов млекопитающих [1,2]. Однако в этих исследованиях рассматривались в основном статические характеристики –

насколько сильно метилирован тот или иной сайт в среднем у всех представителей того или иного вида. Между тем динамика возрастных изменений метилирования между видами может иметь ключевое значение в контексте старения и регуляции долголетия. До настоящего времени этот аспект практически не изучался.

В настоящей работе мы исследовали связь между максимальной продолжительностью жизни и скоростью возрастных изменений метилирования ДНК ($\Delta mRate$) на межвидовом уровне. Мы проанализировали публично доступные данные 20 862 CpG-сайтов образцов крови 37 видов млекопитающих с различной максимальной продолжительностью жизни [3]. Для каждого сайта мы рассчитали скорость изменения уровня метилирования с возрастом до и после нормализации на продолжительность жизни и оценили, как она связана с продолжительностью жизни вида. Мы также провели аналогичный анализ для пяти видов на уровне экспрессии генов, чтобы сопоставить транскриптомные и эпигенетические паттерны старения, связанные с долголетием.

Мы обнаружили, что и на уровне метилирования ДНК, и на уровне генной экспрессии скорость возрастных молекулярных изменений была отрицательно скоррелирована с максимальной продолжительностью жизни вида. CpG-сайты, скорость изменения метилирования которых с возрастом ($\Delta mRate$) наиболее сильно ассоциирована с максимальной продолжительностью жизни, преимущественно располагаются в 5'-нетранслируемых участках и экзонах, и у долгоживущих видов они дольше сохраняют исходный уровень метилирования. Некоторые из этих CpG-сайтов были объединены в кометилированные кластеры, динамика изменения метилирования которых была ассоциирована с максимальной продолжительностью жизни, массой тела и возрастом половой зрелости, характерных для вида. Функциональное обогащение показало, что CpG-сайты с положительным $\Delta mRate$ связаны с процессами клеточной дифференцировки и развития, тогда как сайты с отрицательным $\Delta mRate$ — с транскрипцией. В то же время на уровне транскриптома мы обнаружили, что гены, скорость

изменения экспрессии которых сильнее всего коррелирует с продолжительностью жизни вида, преимущественно связаны с иммунным ответом.

Таким образом, мультиомный анализ выявил более медленные возрастные глобальные изменения молекулярных паттернов у долгоживущих млекопитающих как на уровне метилирования ДНК, так и на уровне транскриптома. При этом наиболее выраженные сигнатуры замедленного старения на разных уровнях молекулярной организации оказались ассоциированными с различными биологическими процессами, что указывает на многокомпонентную природу регуляции старения.

1. Haghani A et al. (2023) DNA methylation networks underlying mammalian traits, *Science*, 381:eabq5693.
2. Li CZ et al. (2024) Epigenetic predictors of species maximum life span and other life-history traits in mammals, *Sci Adv*, 10:eadm7273.
3. Arneson A et al. (2022) A mammalian methylation array for profiling methylation levels at conserved sequences, *Nat Commun*, 13:783.

Identification of Molecular Biomarkers of Response to Renal Cell Carcinoma Immunotherapy via Gene Co-expression Network Analysis

Artemiy I. Dakhnovets¹, Ilya A. Dyugay², Dmitry M. Chudakov²

¹*Skolkovo Institute of Science and Technology, Moscow, Russia
artemiydah@gmail.com*

²*Genomics of Adaptive Immunity Department, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia*

Differential gene co-expression network analysis has emerged as an efficient approach to identifying discrepancies in relationships between individual genes under different conditions. This approach considers the intrinsic interconnections between genes and the

conditional differences in these relationships, as opposed to a simple analysis of differential gene expression that only examines differences in the expression of individual genes. Gene co-expression analysis, in addition to conventional differential expression techniques, can help identify biomarkers at systemic level. Currently, none of the studied biomarkers for immunotherapy response in metastatic renal cell carcinoma (RCC) are available in clinical practice due to insufficient evidence regarding their predictive value. However, advancements in molecular biology techniques and the emergence of novel genome sequencing technologies provide new opportunities for investigating novel molecular biomarkers that can considerably improve existing prognostic models for response to immunotherapy against RCC. The implementation of gene co-expression analysis can greatly aid in the search for molecular biomarkers and the refinement of predictive and prognostic gene signatures.

In this study we used differential gene co-expression network analysis to identify molecular biomarkers associated with response to RCC treatments, specifically immune checkpoint inhibitor (ICI), atezolizumab, ICI with anti-vascular endothelial growth factor (anti-VEGF), atezolizumab with bevacizumab, and standard tyrosine kinase inhibitors (TKI), sunitinib. The gene expression dataset from the IMmotion150 (NCT01984242) clinical study, which included a diverse cohort of RCC patients, was used to construct a therapy-specific differential gene co-expression network. By exploring this network in responsive versus non-responsive manner, we aimed to identify differentially co-expressed gene modules dysregulated in non-responders. To note, gene expression data was normalized against PTPRC-correlated pan-leukocyte genes in order to reduce variability of immune-related genes for more accurate estimation of their expression. Then, we evaluated the association between the expression patterns of hub genes from differential co-expression modules and clinical outcomes to assess their potential as predictive biomarkers. To refine gene signatures based on findings from differential co-expression networks, we selected co-expression modules whose hub genes demonstrated a significant association with progression-free survival (PFS). This was accomplished using the log-rank test and Cox regression modeling to compare survival distributions among patient groupings based on hub gene signature scores estimated using principal component analysis.

Subsequently, we identified genes from significant hub gene sets that exhibited a notable correlation with PFS. These genes, which showed a strong association with PFS in patients within a specific treatment arm, were compiled into a refined prognostic gene signature for that treatment.

As a result of this study, we identified two refined gene signatures associated with the response to plain atezolizumab and its combination with bevacizumab. The response-to-atezolizumab-associated gene signature included seven hub genes derived from seven distinct co-expression modules obtained from patients treated with either atezolizumab or the combination of atezolizumab and bevacizumab. The response-to-combination-arm gene signature comprises 74 hub genes that are part of various modules that were also derived from both treatment groups. Furthermore, we identified a gene set associated with PFS in two treatment arms: sunitinib and atezolizumab plus bevacizumab. Notably, this gene signature functions in opposite directions: patients treated with sunitinib who have high signature scores tend to benefit from the treatment, whereas patients receiving atezolizumab and bevacizumab exhibit a tendency to benefit when they have low signature scores.

In conclusion, identified gene signatures can potentially be used to guide treatment decisions and improve patient outcomes. However, further research is needed to validate our findings in larger cohorts and to explore the potential clinical applications of gene profiling based on the identified gene sets.

**Сравнение Oxford Nanopore, EM-Seq и HumanMethylationEPIC
BeadChip для обнаружения сайтов 5mC в образцах
злокачественной опухоли и здоровой ткани**

К.А. Дейниченко¹, В.Г. Выноградская¹, П.А. Гребнев¹, В.М.
Микова¹, Д.О. Бобылев¹, А.М. Шаймарданов¹, А.А. Ивашечкин¹,
М.В. Ерохина¹, А.И. Некрасова¹, А.В. Семьянихина^{1,2}, С.И.
Митрофанов¹, К.С. Грамматикати¹, В.С. Юдин¹, С.М. Юдин¹, А.В.
Махотенко¹, А.А. Кескинов¹, С.А. Краевой¹, А.С. Макарова¹, Е.А.
Снигирь¹, Д.В. Светличный¹, В.И. Скворцова³

¹Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства (ФГБУ «ЦСП» ФМБА России). 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1.
KDeynichenko@cspfmba.ru

²Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр онкологии имени Н.Н. Блохина» Министерства Здравоохранения Российской Федерации (ФГБУ «НМИЦ онкологии им. Н.Н. Блохина»). 115478, Россия, г. Москва, Каширское шоссе, д. 23.

³Федеральное медико-биологическое агентство (ФМБА России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.

Полногеномная идентификация метилирования ДНК играет важную роль при профилировании физиологически и клинически значимых эпигенетических изменений. Несмотря на наличие множества экспериментальных методов, остается актуален вопрос исследования их точности, преимуществ и недостатков применения при работе с тканевым биоматериалом. В настоящей работе представлено детальное исследование двух наиболее популярных полногеномных методов профилирования 5-Метилцитозина (5mC) на примере образцов колоректального рака. Первый метод основан на секвенировании длинных прочтений с использованием технологии Oxford Nanopore (ONT), которая выполняет прямое секвенирование ДНК с извлечением информации об эпигенетических модификациях вычислительными методами. Второй метод основан на секвенировании метилирования с использованием ферментирования (Enzymatic methyl sequencing, EM-seq), которое в отличие от бисульфитной конверсии исключает повреждение и фрагментацию ДНК, позволяя более точно детектировать 5mC модификации.

Проведено исследование образцов колоректального рака и нормальной ткани для трех пациентов, с использованием секвенирования методами ONT, EM-seq и микрочиповой платформой Illumina EPIC 850K. Исследованы ограничения этих

подходов, их чувствительность и специфичность, зависимость от глубины покрытия секвенирования и выполнена оценка возможностей детекции метилирования. В целом, как прямое секвенирование ДНК с помощью технологии Oxford Nanopore (ONT), так и EM-seq обеспечивают высокую достоверность идентификации позиций 5mC по всему геному в образцах колоректального рака и здоровой ткани. Полученные результаты демонстрируют, что метод ONT обеспечивает более высокую чувствительность и специфичность по сравнению с EM-seq. Увеличение покрытия EM-Seq до 40X дополнительно улучшает обнаружение метилирования. Данная работа представляет собой исследование полногеномного профилирования 5mC в образцах тканей и обобщает достоинства и недостатки каждого из экспериментальных методов.

Функциональная валидация HLA-ассоциаций с тяжёлыми побочными реакциями на аденовирусные вакцины

А.А. Денисова^{1,5}, А.В. Камелин¹, М.М. Трофимов¹,
Д.Э. Харитонов^{1,3}, С. Мансур^{1,2}, Л. Шахин^{1,2},
А.С. Ракитько^{1,2,3,4}

¹*Genotek Ltd., Moscow, Russia, alexandraa.denisova@gmail.com*

²*Phystech school of Biological and Medical physics, Moscow Institute of Physics and Technology, Moscow, Russia*

³*ITMO University, Saint-Petersburg, Russia*

⁴*National Research University Higher School of Economics, Russia*

⁵*Lomonosov Moscow State University, Moscow, Russia*

Хотя Россия и стала первой страной, одобрившей общедоступную вакцину против COVID-19, прививочная кампания шла достаточно медленно. Среди причин отказа от прививки опрошенные называли боязнь побочных эффектов. Действительно, тяжёлые побочные реакции на вакцинацию представляют собой значимую проблему для общественного здравоохранения. Выявление генетических предикторов таких реакций необходимо как для понимания их механизмов, так и для разработки

персонализированных стратегий вакцинации.

Ранее нами было проведено полногеномное ассоциативное исследование неблагоприятных побочных эффектов вакцинации от COVID-19 среди населения России, в ходе которого были обнаружены ассоциированные варианты, расположенные в HLA-генах.

В настоящей работе мы предприняли более детальный анализ HLA-аллелей с последующей функциональной валидацией ассоциаций с использованием *in silico* моделирования антигенной презентации. Извлечение ДНК и генотипирование проводились на образцах слюны, генотипированных с использованием микрочипов Illumina Infinium Global Screening Array v.1–v.3. Импутация HLA-аллелей выполнялась с помощью алгоритма HIBAG-HLA [1].

Прогноз связывания осуществлялся с помощью NetMHCpan и NetMHCIIpan [2] на основе белковых последовательностей, образующихся после процессинга S-белка, экспрессирующегося после введения первой компоненты вакцины Sputnik V. Информация о последовательности была взята из патента “Фармацевтическое средство и способ его использования для индукции специфического иммунитета против вируса тяжёлого острого респираторного синдрома SARS-CoV-2 (варианты)” (RU 2 731 342 C1).

Анализ проводился на когорте из 31,781 человека (14,899 с тяжёлым течением поствакцинального периода, 16,882 с лёгким; возраст 16–80 лет, ИМТ 18–30). В анализ включались только данные о первой вакцинации первой компонентой.

Для поиска значимых ассоциаций с тяжёлым течением вакцинации для каждого HLA-аллеля строилась модель логистической регрессии с учётом пола, возраста, курения и уровня физической активности, с последующей поправкой на множественное тестирование. В результате были выявлены ассоциированные аллели: HLA-DQB1*02:02 (adjusted p-value = 7.85E-05, OR = 1.16 (1.10–1.23)), HLA-DRB1*07:01 (adjusted p-value = 7.78E-07, OR = 1.16 (1.11–1.23)) и HLA-DQA1*02:01 (adjusted p-value = 2.24E-07, OR = 1.16 (1.10–1.23)).

С помощью NetMHCIIpan мы показали, что аллель HLA-DQB1*02:02 имеет наименьший %Rank среди всех HLA-DQB1 аллелей по связыванию с потенциальными пептидами S-белка

(mean %Rank = 0.0015, std = 0.00349), в то время как HLA-DQA1*02:01 и HLA-DRB1*07:01 находятся с ним в сцеплении.

Полученные данные позволяют объяснить ассоциацию HLA-DQB1*02:02 с более тяжёлыми поствакцинальными реакциями через его высокоаффинное связывание с продуктами распада S-белка.

1. Zheng, X., Shen, J., Cox, C. *et al.* HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J* **14**, 192–200 (2014). <https://doi.org/10.1038/tpj.2013.18>
2. Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., & Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *Journal of immunology* (Baltimore, Md. : 1950), 199(9), 3360–3368. <https://doi.org/10.4049/jimmunol.1700893>

Эволюция геномной организации бактерий рода *Azospirillum*

Н. О. Драненко, М. С. Гельфанд

*Сколковский институт науки и технологий, Большой бул.,
30, стр 1, Москва, 121205, dranenko@phystech.edu*

В течение многих лет существовало единодушное мнение о том, что бактериальный геном состоит из одной кольцевой хромосомы, которая может быть дополнена небольшими вспомогательными кольцевыми плазмидами [1]. С 1979 года, когда были обнаружены линейные плазмиды у *Streptomyces*, эти взгляды претерпели изменения [2]. Десять лет спустя было обнаружено наличие двух хромосом в геноме *Rhodobacter* [3]. Такой тип организации генома называется многочастным геномом. На сегодняшний день около 10% секвенированных видов бактерий отличаются по составу от модельного организма *E. coli*, который содержит одну большую кольцевую хромосому и несколько маленьких кольцевых плазмид [1].

Одним из родов, для представителей которых характерно разделение генома на множество крупных репликонов, является *Azospirillum* [4]. В геномах *Azospirillum* может быть от 6 до 10 репликонов. При таком строении генома для реконструкции эволюции становится важной задача установления соответствия между репликонами, а также анализ движения генов между группами сходных репликонов.

В этой работе проведён соответствующий анализ для 22 полных генов *Azospirillum*. Сравнение репликонов через сходство генов разделения *parAB*, через меру Жаккара для всех генов репликонов, не имеющих паралогов, а также через модифицированную меру Жаккара для всех генов репликонов показали сходные результаты. Все крупнейшие репликоны, содержащие минимум по одной копии генов, кодирующих эссенциальные домены, сходны друг с другом более чем с любыми другими репликонами и образуют одну группу. Также есть одна группа малых репликонов, которая включает в себя по репликону из каждого штамма. Однако остальные репликоны группируются по двум крупным ветвям дерева, но не формируют связей по сходству между этими ветвями. Также на дереве присутствуют два далёких штамма, репликоны в которых слабо сходны по составу генов с репликонами штаммов с двух крупных ветвей.

Также в работе был проведен анализ потоков генов между репликонами и в особенности исследованы гены сегрегации хромосом *parAB* и гены рекомбинации *hcrCD*, предсказаны их сайты связывания.

diCenzo, G.C. & Finan, T.M., 2017. The divided bacterial genome: Structure, function, and evolution. *Microbiology and molecular biology reviews: MMBR*, 81(3).

Martin-Didonet, C.C. et al., 2000. Genome structure of the genus *Azospirillum*. *Journal of bacteriology*, 182(14), pp.4113–4116.

Sakaguchi, K., Hirochika, H. & Gunge, N., 1985. Linear plasmids with terminal inverted repeats obtained from *Streptomyces rochei* and *Kluyveromyces lactis*. *Basic life sciences*, 30, pp.433–451.

Suwanto, A. & Kaplan, S., 1989. Physical and genetic mapping of the *Rhodobacter sphaeroides* 2.4.1 genome: presence of two unique circular chromosomes. *Journal of bacteriology*, 171(11), pp.5850–5859.

Новые функции и эволюция гемоглобинов представителей семейства Chironomidae

Яна Григорьевна Дрозд¹, Андрей А. Пржиборо², Наталья Е. Гоголева³, Елена И. Шагимарданова³, Олег А. Гусев³, Нурислам М. Шайхутдинов⁴, Михаил С. Гельфанд⁴

¹Факультет биоинженерии и биоинформатики МГУ им. Ломоносова, 119991, Москва, drozd.janin@gmail.com

²Зоологический институт РАН, 199034, Санкт-Петербург

³Научный центр регуляторной геномики, Институт фундаментальной медицины и биологии, Казанский федеральный университет, 420012, Казань

⁴Сколковский институт науки и технологий, 121205, Москва

Долгое время считалось, что насекомые в значительной степени не нуждаются в гемоглобинах. Трубчатая, разветвленная сеть трахей у гексапод традиционно считается достаточной для их дыхания. Однако на данный момент гены гемоглобина обнаружены у некоторых семейств насекомых [1]. Также сейчас известно, что, по сравнению с гемоглобинами позвоночных, гемоглобины беспозвоночных показывают гораздо большую изменчивость на аминокислотном и структурном уровнях [2]. На данный момент наиболее интересным объектом исследования в этой области являются представители семейства комаров-звонцов *Chironomidae*, в связи с уникальными свойствами их гемоглобинов и их высокой многокопийностью.

Гемоглобины хирономид обладают очень высокой аффинностью к кислороду, что позволяет эффективно захватывать и транспортировать кислород даже при крайне низких концентрациях. Это свойство, возможно, способствует адаптации к жизни в неблагоприятной среде, где наблюдается гипоксия, так как многие представители семейства являются экстремофильными организмами. Однако, помимо видов, живущих в бедных

кислородом средах, существует большое количество хирономид, приспособившихся к другим условиям, таким как низкие температуры (комары-звонцы рода *Diamesa*), длительное отсутствие воды (*Polypedilum vanderplanki*) или высокие концентрации соли (*Baeotendipes noctivagus*). У перечисленных видов также имеется большое количество генов гемоглобина, что может указывать на то, что помимо высокого сродства к кислороду, гемоглобины хирономид могут иметь и другие особенности, дополнительно способствующие адаптации.

В данной работе были найдены и изучены ортологи гемоглобинов среди 20-ти видов хирономид. Наиболее детально рассматриваются гены гемоглобинов вида *Diamesa permacra* — умеренно криофильного представителя хирономид, личинки которого способны жить при низких температурах. В работе изучаются состав, структуры и функции гемоглобинов в геноме *D. permacra* и их ортологов, а также эволюция гемоглобинов в семействе хирономид в целом. Проводится сравнение количества и состава гемоглобинов между неэкстремофильными и экстремофильными видами, живущими в разных условиях.

Данное исследование помогает пролить свет на функции, уникальные особенности и эволюционное происхождение гемоглобинов вида *D. permacra* и семейства хирономид в целом.

1. T. Burmester, T. Hankeln (2007) The respiratory proteins of insects, *Journal of Insect Physiology*, 4:285–294.

2. H. Herold, S. Davis, D. Grimaldi (2020) Transcriptomes reveal expression of hemoglobins throughout insects and other Hexapoda, *PLoS One*, 2020 6:5-15

Эпигенетическая регуляция и переход от одноклеточного к многоклеточному состоянию в *Dictyostelium discoideum*

Е.С.Егоров¹, И.В.Жегалова²

¹МГУ им. М.В. Ломоносова, Факультет биоинженерии и биоинформатики, eser2281@gmail.com

²Институт биологии гена РАН

Представитель таксона *Amoebozoa Dictyostelium discoideum* характеризуется необычным жизненным циклом. При наличии достаточного количества питательных веществ этот организм представляет собой одноклеточные амёбы (vegetative state). В стрессовых условиях эти клетки формируют многоклеточный организм. Этот переход включает последовательное прохождение стадий «streaming» и «mound», и завершается формированием плодоносящей стадии, продуцирующей споры.

Подобные морфологические перестройки требуют значительных изменений в экспрессии многих генов, что контролируется с помощью регуляторных сетей – совокупности взаимодействующих генов, транскрипционных факторов и других молекулярных компонентов, которые координируют экспрессию генов и, следовательно, обеспечивают синхронное протекание жизненных процессов и морфологических изменений. Такие сети интегрируют генетическую и эпигенетическую информацию, позволяя организму динамично реагировать на изменение условий внешней среды. Однако регуляция всех этих процессов изучена недостаточно, что подчеркивает актуальность исследования регуляторных сетей *Dictyostelium discoideum*.

Целью данного проекта является изучение регуляторных сетей в геноме *Dictyostelium discoideum*. В рамках данного проекта мы исследовали гипотезу возможной недопредсказанности энхансеров с помощью выбранной предсказательной модели, в частности, был проведен анализ дифференциальной экспрессии генов на разных стадиях жизненного цикла *D. discoideum* и сравнение множества высокоэкспрессированных генов на каждой стадии с генами, содержащими в себе предсказанную на этой стадии энхансерную последовательность. Эти множества практически не пересекаются между собой ни на одной из четырёх стадий, что может говорить о сложных паттернах экспрессии генов во время смены жизненных стадий *D. discoideum*.

Было проверено несколько гипотез по поводу возможной регуляции энхансерными последовательностями близлежащих генов, с учётом пространственной структуры генома и без учёта таковой. Полученные результаты предполагают комплексную структуру регуляции экспрессии генов в *Dictyostelium discoideum*.

Помимо этого, был выполнен анализ данных секвенирования транскриптома одиночных клеток *Dictyostelium* на разных стадиях развития: был вычислен индекс возраста транскриптома (transcriptomic age index) для разных стадий, что позволило оценить эволюционную динамику экспрессии генов. С помощью анализа траекторий были выявлены гены, критически влияющие на переход между различными стадиями жизненного цикла. Дополнительно была предпринята попытка кластеризации и аннотации субпопуляций клеток самой гетерогенной, плодовой стадии развития, на основе известных маркерных генов, описанных в литературных источниках.

Полученные результаты демонстрируют сложную и многоуровневую организацию регуляторных сетей в *Dictyostelium discoideum*. Выявленная комплексность регуляции экспрессии генов подчеркивает необходимость дальнейших исследований с использованием интегративных подходов, что может привести к более глубокому пониманию механизмов морфогенеза и эволюционных адаптаций в этом организме.

Анализ событий альтернативного сплайсинга при дегенерации поясничных межпозвонковых дисков человека в когорте RuDDS

Е.Е. Елгаева¹, Д.Е. Маслов², А.А. Бердникова¹, Я.А. Цепилов¹

¹*Институт цитологии и генетики СО РАН, пр. Академика Лаврентьева, 10, Новосибирск, Новосибирская обл., Россия, 630090, elizabeth.elgaeva@gmail.com*

²*Институт искусственного интеллекта, Московский государственный университет, Ломоносовский пр., 27к1, Москва, Россия, 119991*

Введение: Дегенерация поясничных межпозвонковых дисков (ДПМД) – это один из основных факторов риска боли в спине, которая является ведущей причиной нетрудоспособности в мире. Изменения в профиле экспрессии генов при ДПМД были

неоднократно описаны в литературе [1-2], однако до настоящего времени их характеризовали лишь количественно, не уделяя внимания качественному составу изоформ транскриптов. Тем не менее, существуют указания на то, что альтернативный сплайсинг может играть важную роль в дегенерации позвоночника [3-4]. Целью данного исследования стало изучить события альтернативного сплайсинга при ДПМД, проведя анализ дифференциального использования изоформ транскриптов (DTU).

Материал и методы: Мы проанализировали транскриптомы поясничных межпозвонковых дисков 19 условно здоровых людей (степень дегенерации по шкале Pfirrmann < 4) – контрольная группа, и 50 пациентов с тяжелой ДПМД (степень дегенерации по шкале Pfirrmann ≥ 4). Данные транскриптомов получены из когорты RuDDS [5] с использованием цепь-специфичного парного секвенирования РНК на платформе DNBSEQ-G400. Количественная оценка содержания транскриптов проводилась с использованием программы salmon, с последующим анализом DTU с использованием пакета DRIMSeq для моделирования изменений в использовании транскриптов и выявления транскриптов со значительными различиями между условно здоровыми и сильно дегенерированными дисками. С целью контроля общей доли ложноположительных находок (OFDR) была применена двухэтапная процедура, основанная на методе Холм-Шаффер и реализованная в пакете stageR: на первом этапе выбирались гипотезы, значимые на уровне гена, после чего оценивались гипотезы на уровне транскриптов только для значимых генов. Функциональная аннотация результатов проводилась при помощи следующих программ: 1) DIGGER для анализа изменений в сетях белок-белковых взаимодействий в результате альтернативного сплайсинга; 2) IsoformSwitchAnalyzeR для анализа последствий альтернативного сплайсинга для структуры транскрипта и описания паттерна их представленности в группах испытуемых (пороговые значения FDR < 0,05 и dIF ≥ 0,05); 3) NEASE для анализа обогащения молекулярно-биологических путей на уровне экзонов.

Результаты: 109 транскриптов из 89 генов продемонстрировали значительное изменение в их использовании у пациентов с тяжелой ДПМД по сравнению с контрольной группой.

В 20 генах было обнаружено, по крайней мере, одно реципрокное переключение изоформ. DIGGER выявил потерю взаимодействия между BZIP-доменом CREB1 и протеинкиназы доменом AKT1 и нарушение димеризации CREB1 в группе условно-здоровых пациентов. При тяжелой ДПМД выявлено нарушение димеризации CCDC88A по домену HOOK. IsoformSwitchAnalyzeR обнаружил потерю функциональных доменов в HTRA1 и NOP53 у пациентов с тяжелой ДПМД. Среди наиболее интересных генов, активность которых оказалась снижена у пациентов с тяжелой ДПМД, можно выделить *FNDC3B*, который является супрессором дифференцировки и пролиферации остеобластов [6], и регулятор клеточного цикла *CSNK1E*, который является мишенью для экспериментальных лекарств для лечения различных заболеваний. Наиболее вероятно, что эти гены изменяют уровень своей экспрессии по отношению к контролям за счет использования альтернативного сайта старта транскрипции. Анализ обогащения с помощью NEASE для 109 транскриптов показал статистически значимое преобладание терминов, связанных с процессингом, сплайсингом и метаболизмом РНК.

Заключение: Таким образом, альтернативный сплайсинг активирует дифференцировку остеобластов у пациентов с тяжелой ДПМД. По-видимому, это достигается благодаря нормальному функционированию CREB1 и снижению активности *FNDC3B*, а также через нарушение димеризации CCDC88A и выпадение функциональных доменов в HTRA1 и NOP53, которое приводит к активации сигнальных путей VEGFA-VEGFR2 и PI3K/AKT. Авторы выражают благодарность О. Леоновой, И. Исакову и А. Пелеганчуку, А. Иванову, Т. Голубевой и В. Фишману за помощь в сборе и подготовке материалов исследования.

Список литературы:

1. S.M. Riester et al. (2018) RNA sequencing identifies gene regulatory networks controlling extracellular matrix synthesis in intervertebral disk tissues, *J Orthop Res*, **36**:1356-1369.
2. Y.G. Zhang et al. (2010) Gene expression profiles of disc tissues and peripheral blood mononuclear cells from patients with degenerative discs, *J Bone Miner Metab*, **28**:209-19.

3. Y. Yao et al. (2016) Global profiling of the gene expression and alternative splicing events during hypoxia-regulated chondrogenic differentiation in human cartilage endplate-derived stem cells, *Genomics*, **107**:170-7.
4. Y. Yao et al. (2017) General regulatory effects of hypoxia on human cartilage endplate-derived stem cells: A genome-wide analysis of differential gene expression and alternative splicing events, *Mol Med Rep*, **16**:3001-3009.
5. O.N. Leonova et al. (2022) A protocol for recruiting and analyzing the disease-oriented Russian disc degeneration study (RuDDS) biobank for functional omics studies of lumbar disc degeneration, *PLoS One*, **17**:e0267384.
6. K. Kishimoto et al. (2010) Fad104, a positive regulator of adipogenesis, negatively regulates osteoblast differentiation, *Biochem Biophys Res Commun*, **397**:187-91.

Выявление аминокислотных замен, ассоциированных с болезнью Альцгеймера, с использованием масс-спектрометрического *de novo* секвенирования

Д.Д. Емекеева, М.В. Горшков, И.А. Тарасова

*Институт энергетических проблем химической физики им. В.Л. Тальрозе, Федеральный исследовательский центр химической физики им. Н.Н. Семенова РАН, Москва, Россия,
dashabezik65@gmail.com*

Болезнь Альцгеймера (БА) остается одним из наиболее распространенных нейродегенеративных заболеваний, патогенез которого до конца не изучен. Одним из ключевых аспектов исследования БА является изучение изменений в белковых последовательностях, таких как аминокислотные замены (АКЗ), которые могут играть важную роль в развитии заболевания. В данной работе представлен новый подход к выявлению АКЗ, специфичных для БА, с включающий *de novo* секвенирования пептидов, поиск по базе данных и предсказание хроматографического времени удерживания.

Исследование основано на анализе протеомных данных, полученных из образцов ткани мозга пациентов с БА и контрольной группы, собранных посмертно. Данные включали результаты тандемной хроматомасс-спектрометрии (LC-MS/MS) из нескольких крупных консорциумных исследований [1-3]. Для идентификации АКЗ использовался разработанный авторами конвейер, объединяющий *de novo* секвенирование (с помощью инструментов Novor и DenovoGUI) и поиск в базах данных (MSFragger, IdentiPy). Для валидации результатов применялись алгоритмы предсказания времени удерживания (DeepLC), доля ложноположительных результатов оценивалась по метрике False Discovery Rate (FDR < 1%).

Был проведен поиск белков, значимо измененных аминокислотными заменами. Среди них был проведен анализ генных онтологий с целью найти функциональные кластеры. В результате было обнаружено, что значимо модифицированные белки (напр. GFAP, TUBB3, SPTAN1, CAMK2A, CNP) участвуют в ключевых процессах, ассоциированных с БА. Среди них транспорт ионов, АТФ-связывание, связывание рецепторов, сопряженных с G-белками, проекция аксонов и секреция нейротрансмиттеров. Были проанализированы процессы, влекущие аминокислотные замены, и их доли. АКЗ могут возникать из-за однонуклеотидных полиморфизмов (SNP, ~2%), ошибок аминоацилирования тРНК (~10%) и посттрансляционных модификаций (PTM, ~30%). Наиболее частыми заменами, обогащенными при БА были N→D и Q→E, связанные с деамидированием, распространенной модификацией при нейродегенерации, а также A→S, потенциально связанная с SNP, и I→V, которую связывают с аминоацилированием tRNA^{Ile}. А также показано, что при БА увеличивается доля замен H→Y, что может быть связано с пониженной активностью HARS1. При анализе АКЗ, возникающими вследствие SNP, были обнаружены замены HBD (W38S), GFAP (N77D, Q290E, D295N), ACTB (F200Y), GLUD1 (A137P) и NEFL (N98D). Упомянутые выше АКЗ были оценены как патогенные как на уровне гена, так и на уровне белка. Более того они были подтверждены по опубликованным клиническими данным SNP. Разработанный алгоритм позволил выявить специфические для БА аминокислотные замены, что расширяет понимание молекулярных

механизмов заболевания. Полученные данные указывают на потенциальные биомаркеры и мишени для терапии, такие как белки, участвующие в передаче сигналов и поддержании нейрональной целостности. Дальнейшие исследования направлены на изучение функциональных последствий выявленных АКЗ и их роли в прогрессировании БА.

Исследование поддержано Российским научным фондом (грант №23-45-00012).

1. J.M. [Bader](#) et al.(2023) Proteomics Separates Adult-Type Diffuse High-Grade Gliomas in Metabolic Subgroups Independent of 1p/19q Codeletion and across IDH Mutational Status, *Cell Reports Medicine*, **4**, 100877
2. E.Johnson et al. (2022) Large-Scale Deep Multi-Layer Analysis of Alzheimer’s Disease Brain Reveals Strong Proteomic Disease-Related Changes Not Observed at the RNA Level, *Nat Neurosci*, **25**: 213–225.
3. E.[Bereczki](#) et al. (2018) Synaptic Markers of Cognitive Decline in Neurodegenerative Diseases: A Proteomic Approach, *Brain*, **141**, 582–595,

Белок-вариант-фенотипический анализ мультисистемного фенотипа с использованием AlphaFold на примере NBAS-ассоциированных патологий

Л.Р. Жожиков, Н.Р. Максимова

НИЛ «Молекулярная медицина и генетика человека», Медицинский Институт, Северо-Восточный федеральный университет имени М. К. Аммосова, 677000, ул. Кулаковского, 46, г. Якутск, Республика Саха (Якутия), Россия, jlr10@mail.ru

Введение.

Природа мультисистемных патологий является актуальной проблемой медицинской генетики, особенно в контексте корректной классификации, диагностики и терапии возможных осложнений, а также в целом — предсказания развития фенотипа. Одной из таких патологий являются NBAS-ассоциированные

заболевания, внутригенетическая гетерогенность фенотипа которых варьируется от синдрома детской печёночной недостаточности второго типа (ILFS2, OMIM #616483) до синдрома нанизма с атрофией зрительного нерва (SOPH-синдром, OMIM #614800) [1]. Ранее предпринимались попытки деления пациентов на подгруппы по локализации варианта в гене, однако природа этих различий на уровне белка оставалась неясной [2]. С развитием современных технологий искусственного интеллекта и нейросетей появилась возможность предсказания структур белков на основе последовательности с помощью инструмента AlphaFold, особенно в контексте белков структура которых не установлена кристаллографически, к числу которых относится и NBAS. В данной работе мы попытались объяснить мультисистемный фенотип *NBAS*-ассоциированных патологий путём идентификации структуры, субдоменной архитектуры, а также сайтов связывания белка NBAS [3].

Методы.

Для построения модели белка использовалась третья версия AlphaFold. Для оценки качества структур применялись встроенные метрики AlphaFold — pLDDT и PAE. Для идентификации субдоменной архитектуры были объединены PAE-графики пяти моделей. Пороговые значения ожидаемой ошибки позиционирования считались в диапазоне от 2 до 5 Å. Для поиска сайтов связывания применялась метрика «консенсусных сайтов контакта», рассчитанная по параметру contact-probes для всех пяти структур белковых комплексов NBAS с его известными партнёрами. Пороговое значение вероятности контакта составляло 0,5. Поскольку на сегодняшний день остаётся неясным, насколько AlphaFold способен надёжно моделировать мутантные белки, для оценки влияния патогенных вариантов мы ограничились картированием этих вариантов на PAE-график дикого типа NBAS с целью определения поражаемого субдомена. В выборку пациентов включались случаи с гомозиготными миссенс либо компаунд-гетерозиготами вариантами. Для построения графиков и выполнения расчётов применялись скрипты на языках программирования Python3 и R.

Результаты.

Моделирование структуры NBAS позволило

идентифицировать ранее предполагаемый домен β -пропеллера, домен Sec39, а также С-конец белка. Идентификация сайтов связывания с белками-партнёрами позволила установить контакты с ключевыми участниками ретроградного мембранного транспорта — USE1 и ZW10 [4]. USE1 — мембранный белок, его сайт связывания локализован на N-конце NBAS, между β -пропеллером и доменом Sec39. Сайт связывания с ZW10, напротив, расположен между доменом Sec39 и С-концом. Примечательно, что данный участок обладает наименьшей структурной упорядоченностью (низкие значения pLDDT), что может свидетельствовать о значительных конформационных изменениях при связывании и предполагаемой активирующей роли данного взаимодействия для функции белкового комплекса. Анализ консенсусных сайтов связывания с другими известными партнёрами (TANGO1, UPF1) не выявил достоверных контактов. Однако в пределах домена Sec39 был обнаружен отдельный консенсусный сайт связывания NBAS с белком Rab18 [5]. Известно, что Rab18 участвует в созревании липидных капель и опосредует контакт между липидной каплей и ЭР через NRZ-комплекс в специализированных типах клеток. Определение субдоменной архитектуры NBAS на основе PAE-графиков позволило выделить шесть субдоменов: β -пропеллер, "шея", Sec39-тело, Sec39-хвост, С-тело и С-хвост. Картирование патогенных вариантов на эти субдомены показало, что варианты, ассоциированные с ILFS2, преимущественно локализуются на участках связывания NBAS с его белками партнерами. Это указывает на то, что нарушение сборки NRZ-комплекса, особенно в условиях клеточного стресса (например, лихорадки), может вызывать его дестабилизацию и функциональную недостаточность — что согласуется с клиническим наблюдением, что триггером ILFS2 является гипертермия. Картирование также показало, что фенотип атрофии зрительного нерва строго ассоциирован с субдоменами С-конца и полностью отсутствует при поражении других участков белка.

Обсуждение.

В данной работе мы проанализировали мультисистемный фенотип *NBAS*-ассоциированных заболеваний, используя структурную информацию с AlphaFold. Однако следует учитывать ряд ограничений. Во-первых, модель NBAS имеет участки с

низкими значениями pLDDT. Во-вторых, возможное влияние посттрансляционных модификаций как самого NBAS, так и его партнёров, также может менять конформацию белка и его взаимодействия. Данные о модификациях NBAS в литературе нами не были найдены. В-третьих, AlphaFold не учитывает влияние внешних условий, таких как pH и температура, на конформацию белка. И наконец, в свете известных ограничений AlphaFold дизайн исследования включал белки дикого типа взаимодействия с которыми были показаны экспериментально. Такой подход позволил сфокусироваться на устойчивых и воспроизводимых результатах, минимизируя риск ошибочных обобщений.

Выводы.

Предложенный в данной работе подход позволяет выделять субдомены белка и картировать на них сайты связывания и патогенные варианты. Анализ белков-партнёров и их функций даёт возможность формировать гипотезы о механизмах действия и исключать предполагаемые прямые контакты. Это позволяет установить уязвимые участки белка и понять молекулярную природу мультисистемных заболеваний, ассоциированных с генами, белки которых обладают несколькими функциями в клетке. Выражаем благодарность Алексею Финкельштейну, Дмитрию Иванкову, а также другим участникам МССМВ23 за ценные замечания и обсуждения, способствовавшие улучшению данной работы на ранних этапах её подготовки.

Список литературы:

1. Л.Р. Жожиков, Ф.Ф. Васильев, Н.Р. Максимова (2023) Литературный обзор *NBAS*-ассоциированных патологий. Все ли найдены?, Генетика, 59(4):371–380.
2. L. Zhodzhirov et al. (2023) Origins of SOPH syndrome: A study of 93 Yakut patients with review of C-terminal phenotype, Clinical Genetics, 103(6):625–635.
3. L. Zhodzhirov, F. Vasilev, N. Maksimova (2025) Protein-Variant-Phenotype Study of NBAS Using AlphaFold in the Aspect of SOPH Syndrome, Proteins, 93(4):871–884.
4. T. Aoki et al. (2009) Identification of the Neuroblastoma-amplified Gene Product as a Component of the Syntaxin 18 Complex Implicated in Golgi-to-Endoplasmic Reticulum

Retrograde Transport, Mol. Biol. Cell, 20(11):2639–2649.

5. D. Xu et al. (2018) Rab18 promotes lipid droplet (LD) growth by tethering the ER to LDs through SNARE and NRZ interactions, J. Cell Biol., 217(3):975–995.

Сравнительный анализ ТАДов и хроматиновых петель в *in vitro* дифференцированных и постмортальных нейронах

Загирова Д.Р.^{1,2}, Морозов К.В.², Кононкова А.Д.², Молодова М.Н.²,
Ваулин Н.С.³, Дудковская А.В.², Плетенев И.А.², Ефимова О.И.²,
Ульянов К.А.², Творогова А.В.⁴, Лагарькова М.А.⁵, Хайтович Ф.Е.²,
Разин С.В.^{4,6}, Ульянов С.В.^{4,6}, Храмеева Е.Е.²

¹*Институт проблем передачи информации им. А.А. Харкевича
РАН, Большой Каретный пер., 19, стр. 1,
Москва, Россия*

²*Сколковский институт науки и технологий, Большой бул., 30,
стр. 1, Москва, Россия
diana.zagirowa@yandex.ru*

³*Медицинский университет Вены, Вена, Австрия*

⁴*ИБГ РАН, ул. Вавилова, 34/5, Москва, Россия*

⁵*ФНКЦ физико-химической медицины им. Ю.М. Лопухина, ул.
Малая Пироговская, 1А, Москва, Россия*

⁶*Биологический факультет МГУ им. М.В.Ломоносова, Ленинские
горы, 1, стр. 12, Москва, Россия*

Нейроны, дифференцированные из индуцированных плюрипотентных стволовых клеток (ИПСК), стали ключевым инструментом для изучения нейродегенеративных заболеваний и нейрональных функций благодаря их доступности по сравнению с постмортальными образцами [1, 2]. Ранее оценка сходства дифференцированных нейронов со зрелыми нейронами

основывалась преимущественно на морфологических данных и анализе транскриптома [3-5], в то время как особенности трёхмерной организации их генома оставались практически неизученными. Наше исследование впервые осуществляет комплексное сравнение пространственной организации генома постмортальных нейронов и нейронов, дифференцированных из ИПСК, устраняя этот пробел.

Нами создана база единообразно обработанных Hi-C данных, включающая 55 человеческих и 35 мышиных образцов (<https://hic.skoltech.ru/datasets/>). База объединяет как оригинальные, так и общедоступные данные, а также включает фетальные образцы и исходные ИПСК для проведения всестороннего сравнения. Выполнен систематический анализ организации хроматина на всех уровнях организации с детальным изучением топологически-ассоциированных доменов (ТАДов) и хроматиновых петель с интеграцией транскриптомных (RNA-seq), эпигеномных (ChIP-seq) и других омиксных данных.

Результаты показывают, что хроматин нейронов, полученных из ИПСК, сохраняет характеристики недифференцированных клеток на всех уровнях организации, в том числе на уровне ТАДов и петель. Анализ главных компонент на основе профиля инсуляции, выраженности границ ТАДов и интенсивности петель показал, что дифференцированные нейроны образуют самостоятельный кластер, близкий к предшественникам, но отдельный от постмортальных зрелых нейронов. Кроме того, нейроны фетального мозга формируют переходную ось от стволовых клеток к зрелым нейронам, отражая этапы нормального онтогенеза. В дифференцированных нейронах наблюдается повышение выраженности границ ТАДов в целом, а также наличие дополнительных более выраженных границ, обогащенных генами, связанными с развитием. На уровне петель в этом типе клеток обнаружено отсутствие части дальних взаимодействий, обогащенных гистоновой меткой H3K27me3. Данные петли предположительно могут служить для репрессии транскрипционных факторов развития во взрослых нейронах. Следовательно, структурный анализ выявил ряд отличий, которые могут объясняться тем, что некоторые репрессивные механизмы, определяющие формирование зрелых ТАДов и хроматиновых

петель *in vivo*, не полностью реализуются в дифференцированных нейронах.

Таким образом, несмотря на морфологическое сходство культуральных нейронов с постмортальными нейронами и экспрессию ими специфических маркеров, хроматин данного типа клеток на уровне ТАДов и хроматиновых петель сохраняет черты клеток-предшественников, не достигая зрелого состояния. Это указывает на необходимость осторожной интерпретации данных, полученных на таких моделях, и оптимизации протоколов дифференцировки. Наше исследование предоставляет важные данные для оценки применимости моделей нейронов, предлагает аналитический подход для анализа хроматина и формирует одну из самых крупных унифицированных баз данных Hi-C, которая может быть полезной широкому научному сообществу.

1. T. A. Kolagar et al. (2020) Human Pluripotent Stem Cells in Neurodegenerative Diseases: Potentials, Advances and Limitations, *Curr. Stem Cell Res. Ther.* **15**:102–110.
2. Di Val Cervo Rivetti et al. (2021) hiPSCs for Predictive Modelling of Neurodegenerative Diseases: Dreaming the Possible, *Nat. Rev. Neurol.* **17**:381–392.
3. A. Jezierski et al. (2022) Electrophysiological- and Neuropharmacological-Based Benchmarking of Human Induced Pluripotent Stem Cell-Derived and Primary Rodent Neurons, *Stem Cell Rev. Rep.* **18**:259–277.
4. A. E. Handel et al. (2022) Assessing Similarity to Primary Tissue and Cortical Layer Identity in Induced Pluripotent Stem Cell-Derived Cortical Neurons through Single-Cell Transcriptomics, *Hum. Mol. Genet.* **25**:989–1000.
6. G. E. Hoffman et al. (2017) Transcriptional Signatures of Schizophrenia in hiPSC-Derived NPCs and Neurons Are Concordant with Post-Mortem Adult Brains, *Nat. Commun.* **8**:2225.

**Генетические факторы варибельности уровня ХС ЛНП
среди пациентов с причинными вариантами в генах LDLR,
APOB и PCSK9**

М. Зайченко^{1,2}, В. Е. Раменский^{2,3,4}, А. В. Киселева² А. В.
Блохина², В. И. Михайлина², А. И. Ершова², О. М. Драпкина², А. Н.
Мешков²

¹*ФГАОУ ВО Московский физико-технический институт
(национальный исследовательский университет), Московская
область, г. Долгопрудный, Институтский пер., 9., 141701
marija.zaichenoka@gmail.com*

²*ФГБУ Национальный медицинский исследовательский центр
терапии и профилактической медицины Минздрава России,
Петроверигский пер., 10, Москва, 101990*

³*Факультет Биоинженерии и биоинформатики, Московский
государственный университет им. М.В. Ломоносова. 19234, г.
Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73*

⁴*Институт перспективных исследований проблем
искусственного интеллекта и интеллектуальных систем,
Московский государственный университет им. М.В. Ломоносова.
119192, Москва, Ломоносовский просп., 27 корпус 1*

Семейная гиперхолестеринемия (СГХС) – заболевание, проявляющееся аномальным повышением уровней холестерина липопротеинов низкой плотности (ХС ЛНП). СГХС является самым распространенным моногенным заболеванием в мире с частотой от 1:250 вплоть до 1:100 в различных популяциях [1]. При этом моногенные причины заболевания удается выявить только примерно у 40% пациентов с клиническим диагнозом СГХС [2]. Около 90% выявляемых у пациентов вариантов находятся в гене LDLR, на остальные 10% приходятся миссенс-замены в генах APOB и PCSK9 [3]. Выделяют гетеро- и гомозиготную формы СГХС, которые, соответственно, вызываются каузальным вариантом в гетеро- и гомозиготном состоянии. При этом уровни ХС

ЛНП при гомозиготной форме заболевания в среднем в два раза выше, чем у гетерозиготной формы заболевания [4]. Как было сказано выше, лишь примерно у половины пациентов с клиническим диагнозом СГХС выявляются моногенные причины заболевания. В случае пациентов без моногенной причины заболевания принято говорить о полигенной гиперхолестеринемии [5]. У таких пациентов наблюдается отягощение частыми вариантами, которые ассоциированы с повышенным уровнем ХС ЛНП, что также известно как высокий полигенный фон [6]. Целью настоящего исследования было изучение генетических факторов вариативности ХС ЛНП среди пациентов НМИЦ ТПМ, у которых были выявлены варианты, вызывающие СГХС, в генах LDLR, APOB и PCSK9.

В работе использовался конвейер обработки данных секвенирования на основе GATK 3.8 и Ensembl VEP. В исследовании анализировали пациентов с патогенными и вероятно патогенными вариантами в трех основных генах СГХС (LDLR, APOB и PCSK9), а также исследовались дополнительные находки в генах липидного обмена: APOE, LDLRAP1, ABCG5, ABCG8, APOA5, APOC2, LMF1, LPL, APOC3, APOA4, APOA1, ANGPTL3, ANGPTL4, GPIHBP1, GPD1, LIPA, MTP, SAR1B, ABCA1, LCAT, CETP, SCARB1, LIPC. Дополнительно были проанализированы данные шкал генетического риска (ШГР) для общего холестерина (ОХС) из работы Willer et al [7].

В ходе анализа данных секвенирования 6113 образцов из биобанка НМИЦ ТПМ было выявлено 432 носителя патогенных и вероятно патогенных вариантов в генах LDLR, APOB и PCSK9 (далее "пациенты"). Из 432 людей 364 являлись носителями вариантов в гене LDLR (83%). При этом у 37 человек из 364 уровень ХС ЛНП был ниже диагностического критерия для СГХС 4,9 ммоль/л без корректирующей терапии, что свидетельствует о неполной пенетрантности вариантов в LDLR – пенетрантность в таком случае составляет порядка 90%. В гене APOB была выявлена одна миссенс-замена p.Arg3527Gln у 68 человек, у двух людей были выявлены варианты в гене PCSK9. Среди 29 носителей вариантов в дополнительных генах, 11 являлось носителями вариантов, понижающих уровень ХС ЛНП, 18 – вариантов, повышающих уровень ХС ЛНП. При анализе данных ШГР для ОХС пациентам

были присвоены перцентили значений ШГР на основе популяционных данных. Среди 432 человек, у 79 наблюдались значения ШГР выше 90-го перцентиля (рисковые), а у 28 -- ниже 10-го перцентиля (протективные). Таким образом, в выборке носителей патогенных вариантов, ассоциированных с СГХС, наблюдается перепредставленность рисковых значений ШГР для ОХС (ОШ=1,75, $p<0,001$) и недопредставленность протективных значений ШГР для ОХС (ОШ=0,63, $p=0,03$) относительно популяции. При этом у обладателей рисковых значений ШГР наблюдались значимо более высокие уровни ХС ЛНП, чем у обладателей протективных значений ШГР ($p<0,001$).

В рамках исследования было показано, что у пациентов с генетическим диагнозом СГХС наблюдается отягощение дополнительными факторами, повышающими уровни ХС ЛНП, и недопредставленность факторов, понижающих уровни ХС ЛНП.

1. M. Kalwick, M. Roth (2025) A Comprehensive Review of the Genetics of Dyslipidemias and Risk of Atherosclerotic Cardiovascular Disease, *Nutrients*, 17(4): 659.

2. P. J. Talmud et al. (2013) Use of low-density lipoprotein cholesterol gene score to distinguish patients with polygenic and monogenic familial hypercholesterolaemia: a case-control study, *The Lancet*, 381.9874: 1293-1301.

3. A. Meshkov et al. (2021) The LDLR, APOB, and PCSK9 variants of index patients with familial hypercholesterolemia in Russia, *Genes*, 12.1: 66.

4. A. Nohara et al. (2021) Homozygous Familial Hypercholesterolemia, *Journal of Atherosclerosis and Thrombosis*, 28(7):665-678.

5. M. Futema et al. (2015) Refinement of variant selection for the LDL cholesterol genetic risk score in the diagnosis of the polygenic form of clinical familial hypercholesterolemia and replication in samples from 6 countries, *Clinical Chemistry*, 61(1):231-8.

6. М. Зайченко и др. (2024) Использование шкал генетического риска для дифференциальной диагностики у лиц с клиническим диагнозом семейная гиперхолестеринемия, *Кардиоваскулярная терапия и профилактика*, 23(12):4251.

7. C. J. Willer et al. (2013) Discovery and refinement of loci associated with lipid levels, *Nature Genetics*, 45(11):1274-1283.

Алгоритм идентификации бактериофагов в метагеномных данных на примере кишечной микробиоты пациентов с меланомой

Н.В. Захаревич, К.М. Климина

*ФГБУ ФНКЦ ФХМ им. Ю.М. Лопухина ФМБА России, Москва,
119435, Малая Пироговская, д. 1а,
zakharevich@yandex.ru*

Идентификация бактериофагов в метагеномных данных представляет собой не простую задачу из-за мозаичности геномов фагов, трудности сборки их полных последовательностей и возможных ошибок в классификации вирусных последовательностей как бактериальных, в связи с их способностью интегрироваться в геном хозяина. Отсутствие универсального маркерного гена, аналога 16S рРНК прокариот, ещё больше усложняет процесс. Современные вычислительные подходы направлены на достижение баланса между чувствительностью, необходимой для успешного преодоления высокого уровня разнообразия и быстрой эволюции фагов, а также обычно низкого и неравномерного покрытия вирусных последовательностей, и скоростью, требуемой для анализа больших массивов метагеномных данных. За последние несколько лет, с развитием методов секвенирования, метагеномики, и внедрением методов машинного и глубокого обучения – был разработан ряд инструментов для идентификации фаговых последовательностей. Все они имеют свои достоинства и недостатки. Методы идентификации вирусных последовательностей можно разделить на две большие группы: (1) методы использующие подходы, основанные на выравнивании (гомологии) с уже известными последовательностями и (2) методы, применяющие машинное/глубокое обучение. Также, в последние годы, активно развиваются гибридные подходы, сочетающие элементы анализа,

основанные на выравнивании и машинного обучения. Выбор инструмента для анализа метагеномных данных будет существенно влиять на предсказанное вирусное сообщество. Вышеизложенное указывает на необходимость, на настоящий момент, использовать многоинструментальный подход для предсказания вирусных геномов в метагеномном контексте, при котором комбинация различных алгоритмов позволит получить более надёжный консенсусный прогноз.

Нами был применён следующий алгоритм при поиске фаговых геномов в каталоге кишечных метагеномов, состоящем из 490 образцов от пациентов с метастатической меланомой с различным ответом на иммунотерапию (из 7-ми различных наборов данных, включая собственный набор данных). Реконструкция вирусных контигов (фаговых геномов) из метагеномных последовательностей проводилась путём сборки сырых (отфильтрованных) метагеномных данных (прочтений) в контиги параллельно с помощью двух инструментов *metaviralSPAdes* и *MEGANIT*. Для формирования набора фаговых геномов проводилось предсказание последовательностей фагов и профагов в собранных метагеномных данных при помощи комбинации восьми алгоритмов: *viralVerify*, *VirSorter2*, *DeepVirFinder*, *Seeker*, *VIBRANT*, *geNomad*, *DBSCAN-SWA* и *Phigaro*, с последующим сравнением полученных результатов и отбором надежных предсказаний. В дальнейший анализ шли только те последовательности, которые были предсказаны двумя и более алгоритмами, а также удовлетворяли следующему критерию по длине > 5000 пар оснований. Для гарантии качества и надёжности идентифицированных фагов, использовался инструмент *CheckV*, и из дальнейшего анализа исключались последовательности с низким уровнем качества (“Low-quality” и “Undetermined quality”). Далее проводилась кластеризация (этап дерепликации) отобранных фаговых геномов (на данном этапе у нас было 17909 последовательностей) с помощью инструмента *dRep* (основные параметры: `--S_algorithm fastANI -nc 0.80 -pa 0.80 -sa 0.95`). На выходе, после дерепликации, был получен избыточный набор фаговых геномов в размере – 11682 последовательности.

Далее, перед нами стояла задача по поиску бактериофагов – потенциальных прогностических биомаркеров эффективности

иммунотерапии. Для её решения были выполнены следующие шаги. Для поиска воспроизводимых фаговых биомаркеров, из-за композиционного характера анализируемых данных, для определения различий между группами пациентов, отвечающих и не отвечающих на иммунотерапию, нами использовался инструмент для дифференциального анализа численности – Songbird, с помощью которого можно описать логарифмическое изменение относительной численности для признаков в наборе. Анализ логарифмических соотношений признаков в контексте их рангов и поданных метаданных осуществлялся при помощи инструмента визуализации Qurro. Таким образом на данном этапе, сначала, с помощью инструмента inStrain и собственного скрипта на языке python нами были созданы матрицы относительных представленностей, для полученных на предыдущем этапе 11682 последовательностей, для каждого анализируемого набора данных отдельно (геномы с шириной покрытия менее 0.5 удалялись). Данные матрицы использовались в качестве входных данных для программы Songbird (с параметрами: `--p-epochs 10000 --p-differential-prior 0.5 --p-min-sample-count 1 --p-min-feature-count 0 --p-summary-interval 1.`), а полученные по результатам работы Songbird значения дифференциалов передавались уже для обработки в программу Qurro. И Songbird и Qurro использовались через QIIME 2 (как плагины). Таким образом на данном этапе Songbird и Qurro использовались нами для поиска дифференциальных рангов анализируемых фаговых геномов, и определения, какие из них преобладают у пациентов, отвечающих на терапию, а какие у пациентов не отвечающих. В качестве отбора наиболее сильно ассоциированных с ответом/не ответом на иммунотерапию фаговых геномов бралась отсечка для дифференциалов > 0.3 (< -0.3) – что соответствует двукратному увеличению представленности. Для каждого набора данных была выполнена оценка статистической значимости относительного содержания отобранных фаговых геномов между группами пациентов, отвечающих и не отвечающих на иммунотерапию.

Описанная выше процедура проводилась независимо для каждого набора данных. В связи с этим, дальнейшим нашим шагом был поиск бактериофагов – кандидатов в прогностические биомаркеры ответа на иммунотерапию – встречающихся в двух и

более наборах. Поиск пересечений биомаркеров между наборами был выполнен с помощью собственного скрипта на языке python со следующими двумя условиями: последовательность фагового генома считалась кандидатом, если присутствовала в двух и более наборах данных; если последовательность фагового генома была ассоциирована с разными исходами в разных наборах данных, она исключалась из списка кандидатов. По результатам проведенного анализа на роль кандидатов в прогностические биомаркеры ответа на терапию было выдвинуто 1185 фаговых последовательностей. При этом прогностических биомаркеров успеха иммунотерапии среди фаговых последовательностей оказалось 1175 – 99,2%, и всего 10 последовательностей – 0,8% от выявленных 1185 оказались биомаркерами (пересекающимися между наборами) для случая отсутствия ответа на иммунотерапию. Таким образом, мы приходим к принципу Анны Карениной, адаптированному к микробиоте кишечника, подчеркивающему ключевой аспект различий между микробиотой отвечающих на иммунотерапию и не отвечающих пациентов. По всей видимости, микробиоты ответчиков демонстрируют более устойчивую конфигурацию с определенным балансом/преобладанием ключевых бактериальных таксонов, а следовательно, и бактериофагов, а также функциональной стабильностью и метаболическим гомеостазом – обеспеченных всё тем же сложившимся сообществом бактерий и фагов. В тоже время, микробиоты не отвечающих на терапию пациентов в некотором роде уникальны – так как дисбаланс в них может возникать из-за множества причин, приводя к разнообразным нарушениям, зачастую индивидуальным и неповторимым. Стоит подчеркнуть, что принцип Анны Карениной в контексте кишечной микробиоты уже не раз отмечался в различных публикациях.

На настоящий момент мы проводим анализ, полученных нами 1185 последовательностей фаговых геномов – кандидатов в прогностические биомаркеры ответа на иммунотерапию, а именно: проводим таксономическую классификацию вирусных контигов; выполняем поиск и предсказание хозяев выявленных бактериофагов; делаем функциональную аннотацию вирусных контигов.

Работа выполнена при поддержке гранта РНФ № 23-75-10125.

Моделирование шума в данных РНК-ДНК взаимодействий методами машинного обучения

Д.С. Звездин^{1,2}, А.А. Жарикова^{1,2}, А.А. Миронов^{1,2}

¹*Московский государственный университет им. М.В. Ломоносова, факультет биоинженерии и биоинформатики, Ленинские горы, 1, стр. 73, Москва, 119992, Россия, dima.zvezdin1@gmail.com*

²*Институт общей генетики им. Н.И. Вавилова Российской академии наук, ул. Губкина, д. 3, Москва, 119991, Россия, mironov@bioinf.fbb.msu.ru*

Существенную часть транскриптов в эукариотических клетках представляют разнообразные классы некодирующих РНК [1]. Функциональная роль большинства из них остается невыясненной, тогда как для некоторых установлена достаточно хорошо. Например, малые ядерные РНК являются ключевыми компонентами сплайсосомы, участвуя в процессе сплайсинга, а малые ядрышковые РНК обеспечивают модификацию оснований рибосомальных РНК в ядрышке. Особый интерес представляет класс длинных некодирующих РНК (днкРНК), выполняющих разнообразные регуляторные и структурные функции [2]. За последние годы появился ряд экспериментальных протоколов, позволяющих фиксировать взаимодействия РНК с ДНК. Среди них есть методы как для определения взаимодействий отдельных заранее известных РНК с хроматином, так и полногеномного определения РНК-ДНК интерактома (Red-C [3], RADICL-seq [4], GRID-seq [5]).

Эти методы открыли новые возможности для изучения процессов, в которые вовлечены различные типы некодирующих РНК в ядрах эукариотических клеток. Однако все они обладают существенным ограничением - высоким уровнем шума, вызванным контактами, не несущими регуляторной роли, а являющимися следствием диффузии молекул РНК в ядре и неспецифических сшивок в процессе пробоподготовки. Для борьбы с шумом применяются различные нормировки: на фоновый уровень

контактируемости с данным регионом ДНК для учета открытости хроматина и технических факторов и на скейлинг для учета расстояния до гена РНК. В нашей лаборатории был разработан инструмент BaRDIC для поиска статистически значимых взаимодействий (пиков) в данных полногеномных экспериментов, учитывающий эти нормировки [6]. Проблемой является разрешение полученных данных, ширина пиков зачастую достигает десятков тысяч пар оснований, что не позволяет точно локализовать локусы функциональных взаимодействий. Перспективной задачей является разработка метода, который позволит повысить разрешение разметки значимых взаимодействий, что требует более качественной оценки уровня шума.

Для нормировки на фоновый уровень контактируемости в данных полногеномных экспериментов как правило используется покрытие генома ДНК-частями контактов мРНК, чьи гены находятся на других хромосомах (транс-контакты). Этот метод основан на предположении о том, что подавляющее большинство контактов мРНК с чужими хромосомами не являются функциональными, а являются следствием диффузии молекул РНК в ядре клетки. Известно, что контакты РНК в целом скоррелированы с открытостью хроматина [4], однако на данный момент не было произведено комплексного исследования, какие факторы влияют на интенсивность фонового контактирования. Помимо открытости хроматина, другие геномные характеристики могут быть скоррелированы с шумом в данных РНК-ДНК. Различные признаки последовательности ДНК и РНК также могут быть связаны с интенсивностью фоновых взаимодействий. Определив набор признаков, коррелированных с шумом, их можно использовать для его предсказания с помощью моделей машинного обучения. Обученную модель можно использовать для импутации и удаления артефактов, что может улучшить оценку уровня шума, в особенности, для поиска пиков на высоком разрешении. Интерпретация предсказаний модели может позволить установить новые взаимосвязи в данных РНК-хроматиновых взаимодействий.

Мы провели комплексное исследование устройства шума в данных РНК-хроматиновых взаимодействий для двух экспериментальных протоколов и двух клеточных линий мыши: RADICL-seq: mESC (эмбриональные стволовые клетки) и mOPC

(предшественники олигодендроцитов), GRID-seq: mESC. Мы получили по большей части ожидаемые результаты касательно корреляции геномных данных с шумом: он коррелирует с данными экспериментов ATAC-seq, DNase-seq, активными гистоновыми метками и связан с различными аспектами структуры хроматина. При этом сильнее всего шум коррелирует с признаками последовательности ДНК: GC-составом и числом сайтов рестриктазы, использованной для фрагментации ДНК в эксперименте.

Неожиданным оказался эффект разного разрешения при сравнении уровня шума между наборами данных: на большем размере бина ожидаемо данные на mESC, полученные разными протоколами, коррелируют сильнее, чем данные одного экспериментального протокола между mESC и mOPC. Однако, на более маленьких бинах ситуация обратная: эффект технических отличий в протоколах больше сказывается на отличиях в экспериментальном шуме. На малых бинах, в отличие от больших, сильно перепредставлены нулевые значения, что косвенно говорит об их технической природе. Эти наблюдения говорят о необходимости применять дополнительную обработку и импутацию фонового трека взаимодействий для более качественной оценки шума.

Для решения этой задачи мы предложили использовать характеристики последовательности ДНК, а также тканеспецифичные данные сторонних экспериментов для построения модели шума. Мы протестировали ряд моделей и показали, что для эффективного предсказания шума модель должна иметь два свойства: быть способной описывать нелинейные взаимосвязи признаков и учитывать локальный контекст региона генома. При этом нелинейная модель не должна быть сложной: для эффективного предсказания достаточно двухслойной полносвязной нейронной сети, содержащей примерно две тысячи параметров.

Для проверки консистентности мы сравнили предсказания моделей, обученных на данных двух разных протоколов. Несмотря на одинаковый набор обучающих признаков (данные на одной клеточной линии), визуально предсказанные треки шума существенно отличаются. Тем не менее, корреляция между предсказанными треками существенно выше, чем между

наблюдаемыми значениями шума, что говорит о том, что модели выучивают консистентные взаимосвязи между признаками генома и шумом, и косвенно подтверждает валидность построенных моделей.

Работа финансировалась РНФ грант №23-14-00136.

1. S. Djebali, C. A. Davis, et al. (2012) Landscape of transcription in human cells. *Nature*. 489:101–108.
2. X. Li, X.-D. Fu (2019) Chromatin-associated RNAs as facilitators of functional genomic interactions. *Nat. Rev. Genet.* 20:503–519.
3. A. A. Gavrilov, A. A. Zharikova, et al. (2020) Studying RNA–DNA interactome by Red-C identifies noncoding RNAs associated with various chromatin types and reveals transcription dynamics. *Nucleic Acids Res.* 48:6699–6714.
4. A. Bonetti, F. Agostini, et al. (2020) RADICL-seq identifies general and cell type-specific principles of genome-wide RNA–chromatin interactions. *Nat. Commun.* 11:1018.
5. X. Li, B. Zhou, et al. (2017) GRID-seq reveals the global RNA–chromatin interactome. *Nat. Biotechnol.* 35:940–950.
6. D. E. Mylarshchikov, A. I. Nikolskaya, et al. (2024) BaRDIC: robust peak calling for RNA–DNA interaction data. *NAR Genomics Bioinforma.* 6.

Сравнительный анализ метагеномных образцов с помощью библиотеки MetaFX

А.Б. Иванов¹, Е.И. Олехнович², В.И. Ульянов¹

¹*Университет ИТМО, Кронверкский пр., д.49, Санкт-Петербург, Россия, abivanov@itmo.ru*

²*ФНКЦ физико-химической медицины им. академика Ю.М.Лопухина ФМБА, Москва, Россия*

Микробные сообщества населяют различные экологические ниши, включая почву, водоемы и кишечник человека. Анализ данных метагеномного секвенирования применяется для

исследования функций микробных сообществ, например, при исследовании здоровья человека, диагностировании заболеваний и выборе стратегий лечения [1, 2].

Разработка методов для сравнительного анализа метагеномных данных является актуальной задачей, поскольку позволяет изучать зависимости между свойствами сообществ и характеристиками образцов в различных группах. Например, может производиться сравнение пациентов с заболеванием и контрольной группы для установления различий в микробиоте кишечника и формулировки биологических гипотез о возникновении заболевания и оптимальных способах лечения.

Существующие методы для сравнения метагеномов можно разделить на две категории. Методы из первой категории основаны на использовании баз данных для таксономической и функциональной аннотации образцов [3, 4] и последующем сравнении. Недостатком таких методов является зависимость от баз, а также потеря информации об особенностях образцов и вариациях в геномах, которые не содержатся в базах данных. Методы из второй категории используют k -меры (строки длины k из прочтений) и рассчитанные с их помощью статистики для сравнения метагеномных образцов [5, 6]. Это позволяет проводить детальное сравнение, однако полученные результаты сложно интерпретировать. Короткие k -меры плохо аннотируются, поэтому дальнейший анализ и формулирование биологических выводов становится невозможным.

Нами предложен метод извлечения признаков и сравнительного анализа метагеномных образцов, которые объединяет сильные стороны существующих методов. Идея метода состоит в двухэтапной обработке данных. На первом этапе из каждой группы образцов извлекаются опорные k -меры, которые позволяют различать образцы разных групп между собой. Поиск таких k -меров производится с учетом частоты встречаемости в образцах разных групп и статистических тестов. На втором этапе полученные группы опорных k -меров объединяются в длинные интерпретируемые признаки с помощью графа де Брейна. Производится локальное извлечение подграфов из графа де Брейна вокруг вершин, соответствующих опорным k -мерам. Это позволяет сократить число признаков и увеличить их длину, что упрощает

дальнейшую аннотацию и интерпретацию.

Предложенные методы были реализованы в виде библиотеки MetaFX для сравнительного анализа и классификации образцов полногеномного метагеномного секвенирования коротких прочтений. Библиотека расположена в свободном доступе для скачивания по адресу: <https://anaconda.org/bioconda/metafx>.

Кроме реализации методов извлечения признаков, библиотека содержит пайплайны для классификации образцов. На основе извлеченных признаков для каждого образца рассчитывается численный вектор покрытия подграфов графа де Брейна k-мерами из образца. Полученная матрица признаков используется для обучения классификационной модели с помощью случайного леса деревьев решений или алгоритмов бустинга. Обученный классификатор в дальнейшем применяется для определения классов для неразмеченных метагеномных образцов, например, для помощи при раннем диагностировании заболеваний у пациентов по данным секвенирования микробиоты кишечника.

Библиотека была применена для анализа четырех открытых наборов данных метагеномов пациентов с воспалительными заболеваниями кишечника (ВЗК) [7-10]. Всего было обработано 590 образцов из трех групп: 256 пациентов с болезнью Крона (БК), 156 пациентов с язвенным колитом (ЯК) и 178 пациентов из контрольной группы.

Была обучена модель случайного леса со 100 решающими деревьями. Проводилась как 5-кратная кросс-валидация на одном наборе данных, так и обучение классификатора на одном из наборов и предсказание классов образцов из других наборов. Сравнение проводилось с использованием результатов таксономической аннотации Kraken2 для получения признаков и обучения на них классификационной модели.

Производилась трехклассовая классификация, результаты оценивались с помощью метрики ассигасу. Для кросс-валидации прирост качества составил от 6 до 36 % по сравнению с таксономическими признаками. Для независимых наборов данных прирост качества составил от 1 до 20 %.

Также был проведен дальнейший анализ извлеченных графовых признаков и установление их таксономической принадлежности с помощью программы Kraken2. Показано, что

микрофлора кишечника людей из контрольной группы обогащена продуцентами short-chain fatty acids (SCFAs) и бутирата, такими как *F. Prausnitzii*, *R. Hominis*, *A. Hadrus* и *R. Lactatiformans*. При болезни Крона выявлен ряд условно-патогенных микроорганизмов, которые могут запускать процессы фиброза и продуцировать микробно конъюгированные желчные кислоты, что негативно сказывается на течении заболевания. Также в группах БК и ЯК повышено содержание человеческой ДНК, что связано с истончением слизистой кишечника, воспалением и гибелью клеток.

Разработанная библиотека MetaFX позволит упростить сравнительный анализ метагеномных данных и может лечь в основу систем поддержки принятия медицинских решений.

Иванов А.Б. и Олехнович Е.И. финансово поддержаны грантом РНФ №23-75-10125.

1. The Human Microbiome Project Consortium. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**:207–214
2. J. Lloyd-Price et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, **569(7758)**:655-662.
3. A. Blanco-Míguez et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nature Biotechnology*, **41(11)**:1633-1644.
4. F. Beghini et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*, **10**:e65088.
5. B.D. Ondov et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome biology*, **17**:1-14.
6. Y. Wang et al. (2020). KmerGO: a tool to identify group-specific sequences with k-mers. *Frontiers in microbiology*, **11**:2067.
7. E.A. Franzosa et al. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nature microbiology*, **4(2)**:293-305.
8. J. Lloyd-Price et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, **569(7758)**:655-662.
9. G. Lo Sasso et al. (2021). Inflammatory bowel disease-associated

- changes in the gut: focus on Kazan patients. *Inflammatory bowel diseases*, **27(3)**:418-433.
- Q. He et al. (2017). Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience*, **6(7)**.

Иммуноинформатический дизайн и молекулярно-динамическая валидация субъединичной вакцины-кандидата против болезни Лайма

А.С. Ивановский, А.А. Тюленев, В.И. Тимофеев, Ю.В. Писаревский, Ю.А. Дьякова

*пл. Академика Курчатова, 1, Москва, 123098,
a.Iwanowskiy@gmail.com*

Болезнь Лайма, вызываемая спирохетами *Borrelia*, остается значимой проблемой здравоохранения, особенно в условиях отсутствия лицензированных профилактических средств [1]. В работе представлен инновационный подход к разработке субъединичной вакцины на основе флагеллина (UniProt: P11089) — ключевого поверхностного белка, отвечающего за движение бактерии. Комбинирование методов иммуноинформатики и молекулярной динамики (МД) позволило оптимизировать антиген для индукции устойчивого иммунного ответа.

На первом этапе иммуноинформатического анализа было проведено предсказание эпитопов с использованием инструментов NetCTL[2] (Т-клеточные эпитопы) и SEMA2 [3] (В-клеточные линейные/конформационные эпитопы) выявлено 12 Т- и 4 В-эпитопа с высокой консервативностью. Пространственная структура флагеллина была предсказана с помощью AlphaFold3 [4], что позволило визуализировать пространственное расположение предсказанных эпитопов (рис. 1).

```

sp|P11089|FLA1_BORBU 1      10      20      30      40      50      60
MIINHTSAINASRNNGINANLSIQENISSGRINRASDDAAGMGVSGKINAQIRGLS
sp|P11089|FLA1_BORBU 70     80     90     100    110    120
QASFTSKAINETCTTEGVINEVEKVIANRHKIAYVGGGSAKQISFQITG
sp|P11089|FLA1_BORBU 130    140    150    160    170    180
LHSLAQQAQYNQHMILSSASQVETAEELMQPAILNTPESLSSQASNTLRVHVGA
sp|P11089|FLA1_BORBU 190    200    210    220    230    240
QDEAIAVNIYAANVANLISGGACITQAIVQEGVQEGGACQPAFATAPSSQGVNSFVNV
sp|P11089|FLA1_BORBU 250    260    270    280    290    300
CTIVDANTSIAKIEAIRMISDQRANLGAFQNRLESIKDSTFAIENIRASTAQIKDAIV
sp|P11089|FLA1_BORBU 310    320    330
GGGVAAITNSSLTQSAAMNTSQANQVFQWVSTLIP

```

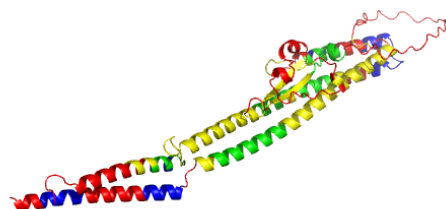


Рис. 1. Пространственная структура рассматриваемого белка флагеллин (Т-клеточные эпитопы выделены зелёным, В-клеточные эпитопы выделены красным, их пересечение синим).

Аллергенные участки идентифицированы с помощью AllerTOP2 [5] и удалены. Повторное моделирование подтвердило сохранение стабильности структуры.

Далее была проведена молекулярное динамическое моделирование длительностью 200-нс в Gromacs-2023 при 310 K, 0.15 M NaCl. Анализ среднеквадратичных отклонений (СКО: 1.5–2.0 Å), флуктуаций (СКФ) и радиуса гирации (Rg) подтвердил структурную устойчивость конструкции (рис. 2).

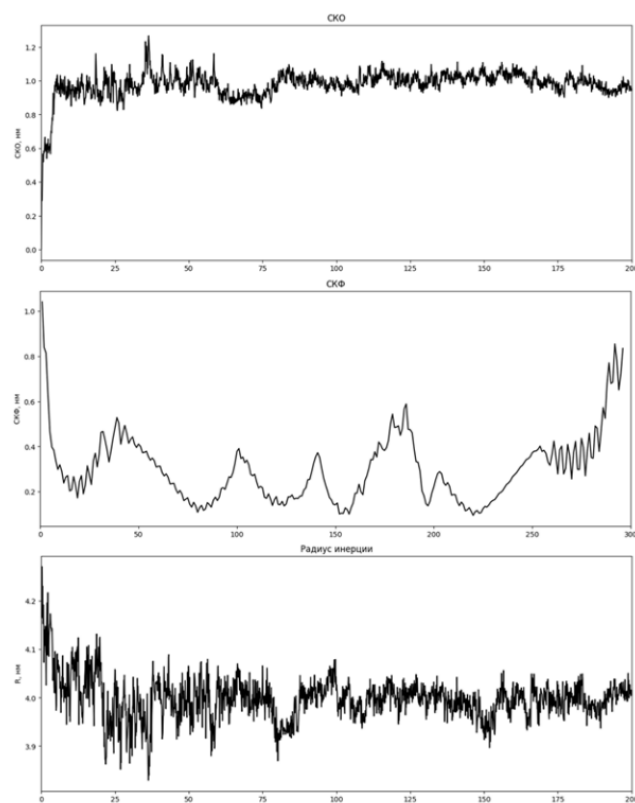


Рис. 2. Среднеквадратичное отклонение (СКО), среднеквадратичная флуктуация (СКФ) и радиус инерции аминокислотных остатков вакцины-кандидата в процессе молекулярной динамики.

Финальным этапом моделирования стала симуляция иммунного ответа при помощи C-IMMSIM [6], выявившая двухэтапную активацию: врожденного иммунитета (пик активности пролиферации Т- и В-клеток на 3–5 сутки) и адаптивный ответ (формирование клеток памяти). Отмечена продукция IgG, IgM, IL-2, IFN- γ без признаков гиперреактивности. Структурная стабильность: МД-анализ показал минимальные отклонения (СКО < 2 Å), что соответствует нативным белкам. Имуногенность: Вакцина-кандидат индуцирует

сбалансированный гуморальный и клеточный ответ, включая активацию CD4⁺/CD8⁺ Т-лимфоцитов и В-клеток памяти. Безопасность: Отсутствие аллергенных участков и цитокинового шторма подтверждает низкий риск побочных реакций.

Предложенная методология, интегрирующая иммуноинформатику и МД, позволила провести рациональный дизайн вероятной стабильной и иммуногенной конструкции. Результаты *in silico*-моделирования обосновывают перспективность дальнейшего рассмотрения как всей конструкции на основе флагеллина *B. Burgdorferi*, так и отдельных предсказанных эпитопов для профилактики болезни Лайма. Дальнейшие этапы включают экспериментальную проверку *in vitro* и *in vivo*. Разработанный подход может быть адаптирован для рационального дизайна вакцин-кандидатов против других патогенов со сложной антигенной архитектурой.

Работа выполнена в рамках государственного задания Национального исследовательского центра «Курчатовский институт». В ходе выполнения данной работы были активно задействованы ресурсы совместного исследовательского центра «Комплекс моделирования и обработки данных исследовательских установок мегакласса» Национального исследовательского центра «Курчатовский институт», <http://ckp.nrcki.ru/>.

1. S. A. Plotkin, E. D. Shapiro (2024). The Current and Future State of Vaccines for Lyme Disease, *Clinical Infectious Diseases*, ciae639.
2. M.V. Larsen, et al (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*. 31;8:424.
3. N.V. Ivanisenko, et al. (2024) SEMA 2.0: web-platform for B-cell conformational epitopes prediction using artificial intelligence, *Nucleic Acids Research*, 52(W1), 533–539.
4. J. Abramson, et al (2024). Accurate structure prediction of biomolecular interactions with AlphaFold. *Nature*. 630. 493–500.
5. Dimitrov I, et al (2014). AllerTOP v.2--a server for *in silico* prediction of allergens. *J Mol Model*. 20(6). 2278.
6. N. Rapin, et al. (2010). *PLoS One*. 5(4). P. 9862–9867.

Защитные системы внутриклеточных паразитов

Р.А. Ириоглов, А.В. Алексеевский

119992, Москва, Ленинские горы, дом 1, стр 40, irio-roman@yandex.ru

Борьба между вирусами-бактериофагами и прокариотами активно изучается в последние годы не только как фундаментальные вопросы о системах иммунитета у бактерий и архей, но со стороны практических возможностей фаговой терапии и влияния фагов на колонии бактерий и биоплёнки. В настоящее время известно более 150 защитных систем прокариот. Арсенал защитных систем бактерий, живущих в определенных экологических нишах может отражать распространенность и роль бактериофагов. Созданы методы для предсказания наличия генов большого числа защитных систем в геномах прокариот с помощью НММ-профилей. Один из таких методов - онлайн-сервис PADLOC [1], был использован нами для анализа распространенности защитных систем прокариот.

Целью нашей работы было сравнить состав защитных систем у преимущественно свободноживущих бактерий и внутриклеточных паразитов.

Мы идентифицировали защитные системы в геномах 22 видов прокариот - 4 облигатных внутриклеточных паразитов, 11 факультативных внутриклеточных паразитов, и для контроля в 7 бактериях и археях, не имеющих в жизненном цикле стадии внутриклеточных паразитов (Таблица 1).

У облигатных внутриклеточных паразитов защитные системы либо не были обнаружены, либо находилось в среднем от 2 до 4 систем на геном. Найденные у них системы SoFic, dGTPase и PD-T4-6 состоят из одного белка, и имеют дополнительные функции в бактериальной клетке, что способствовало их сохранению при редукции генома. Кроме того, данные системы часто представлены у факультативных внутриклеточных паразитов и экстремофилов.

Полученные данные свидетельствуют в пользу выдвинутого нами ранее предположения о зависимости количества и

разнообразия защитных систем у бактерий от среды обитания и наличия фагов ней. В дальнейшем мы планируем изучить системы, найденные у облигатных внутриклеточных паразитов и исследовать состав систем защиты у других категорий экстремофилов.

Список литературы:

1. A.B. Isaev, O.S. Musharova, K.V. Severinov (2021) Microbial Arsenal of Antiviral Defenses – Part I. Biochemistry Moscow 86, 319–337.
2. Payne L.J., Todeschini T.C., Wu, Y., Perry B.J., Ronson C.W., Fineran P.C., Nobrega F.L. and Jackson S.A. (2021) Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. Nucleic Acids Research 49, 10868-10878.

Таблица 1. Защитные системы факультативных и облигатных внутриклеточных паразитов

Вид	Клас с	Число исследованн х штаммов	Системы защиты на геном			
			Ми п	Ма х	Сред .	Представленн ые системы
Rickettsia japonica	obl	30	2	2	2	SoFic (100%), dGTPase (100%)
Chlamydia trachomatis	obl	30	0	0	0	-
Chlamydia psittaci	obl	30	0	0	0	-
Mycobacterium leprae	obl	5	3	4	4	PD-T4-6 (10 шт), dGTPase (5 шт), AbiE (4 шт)
Staphylococcus aureus	fac	30	1	7	5	SoFic (73%), Mokosh тип II(63%)

Brucella suis	fac	30	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella ovis	fac	1	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella abortus	fac	44	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella melitensis	fac	80	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella canis	fac	10	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella microti	fac	1	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella pinnipedialis	fac	3	2	2	2	RM тип I (100%), dGTPase (100%)
Brucella ceti	fac	2	2	2	2	RM тип I (100%), dGTPase (100%)
Shigella flexneri	fac	30	2	6	3	RM тип I (100%),

						VSPR (100%),
Shigella sonnei	fac	30	4	8	5	RM тип II (100%), VSPR (100%) Mokosh тип II (57%)

Обозначения классов: obl - облигатный внутриклеточный паразит, fac - факультативный внутриклеточный паразит

**Роль опухоль-ассоциированных фибробластов и
эндотелиальных клеток в прогрессировании заболевания у
пациентов с глиобластомой.**

А.М. Исмаилов, М.С. Попцова

*Международная лаборатория биоинформатики,
Национальный исследовательский университет «Высшая школа
экономики», Москва, Россия
amismailov@hse.ru*

Глиобластомы являются одними из самых смертоносных онкологических заболеваний в медицине. Уже многие годы не удается добиться значимых результатов лечения пациентов с этим заболеванием. Средняя продолжительность жизни стагнирует около 18 месяцев с момента установления диагноза уже многие годы, несмотря на агрессивное лечение включающее тотальное хирургическое удаление и последующие курсы химио- и лучевой терапии. Основными причинами таких удручающих результатов лечения является крайняя гетерогенность и иммуносупрессивность микроокружения опухоли, гарантирующие иммунологическую невидимость клеток глиобластомы [1]. В данной работе мы провели анализ транскриптомных данных одноклеточного секвенирования на открытых данных. Данные содержали образцы глиобластом одиннадцати пациентов, взятых интраоперационно. В результате

работы мы демонстрируем комплексное и разностороннее проопухолевое влияние опухоль-ассоциированных фибробластов и эндотелиальных клеток на микроокружение глиобластомы, влияние данных клеток на физиологию микроокружения простирается от пролиферативно-протекторного воздействия на сами опухолевые клетки, и в частности на популяцию стволовых клеток глиобластомы, M2 поляризующего эффекта на макрофаги/микроглию и как прямого так и опосредованного ингибирующего влияния на популяцию эффекторных Т лимфоцитов. Мы также демонстрируем и обратные поддерживающие связи клеток микроокружения на популяцию опухоль-ассоциированных фибробластов. В совокупности находки указывают на важную роль рассматриваемых клеток в структуре микроокружения глиобластомы в поддержании как гетерогенности, так и иммуносупрессивности опухоли. Полученные данные свидетельствуют о необходимости включения иммунологического воздействия на популяцию опухоль-ассоциированных фибробластов и эндотелиальных клеток для повышения иммунореактивности опухоли, что потенциально открывает возможность для более эффективного применения как альтернативных видов иммунотерапии, так и конвенциональной химиотерапии.

1. Ismailov A, Spallone A, Belogurov A Jr, Herbert A, Poptsova M. Molecular biology of the deadliest cancer - glioblastoma: what do we know? Front Immunol. 2025 Mar 21;16:1530305. doi: 10.3389/fimmu.2025.1530305. PMID: 40191211; PMCID: PMC11968700.

РНК-опосредованные взаимодействия: Алгоритм поиска eRNAi-мишеней для изучения влияния метагенома.

Иштуганова В.В.¹, Кравченко М.А.¹, Адонин Л.С.², Смутин Д.В.^{2,3}

¹Санкт-Петербургский государственный университет,
Институт биоинформатики e-mail:valeriishtuganova@gmail.com

²*Санкт-Петербургский государственный университет
телекоммуникаций им. проф. М.А. Бонч-Бруевича*

³*Национальный исследовательский университет ИТМО*

РНК-интерференция представляет собой механизм, основанный на подавлении экспрессии генов с помощью малых молекул РНК (siRNA). Это естественный процесс регуляции, играющий ключевую роль в защите клетки от чужеродной генетической информации и контроле собственных генетических программ [1]. Изначально считалось, что РНК-интерференция распространяется только на внутриклеточные процессы, однако, сейчас активно изучаются механизмы передачи малых интерферирующих молекул, которые могут попадать в клетку из окружающей среды и модулируют активность генов (eRNAi) [1]. Этот процесс играет ключевую роль в коммуникации у различных живых существ. В настоящее время РНК-интерференция окружающей среды также рассматривается как перспективный метод борьбы с патогенами [2,3]. Тем не менее, процессы транспорта молекул РНК и их потенциальные мишени остаются недостаточно изученными.

РНК-интерференции окружающей среды открывает новые возможности для изучения взаимодействий между микробиомом и организмом-хозяином. Мы предполагаем, что некодирующие участки мРНК прокариот могут выступать в качестве регуляторных элементов, влияя на экспрессию генов-мишеней в клетках эукариотических организмов. Это приводит к сайленсингу специфических генов хозяина и объясняет наблюдаемые изменения в его транскриптоме. Такой механизм может лежать в основе взаимной регуляции генов в системе «микробиом–хозяин», воздействуя на устойчивость к патогенам и метаболические процессы.

Алгоритм разработки для поиска мишеней eRNAi основывается на анализе мультиомных данных путем интеграции метагеномных и транскриптомных данных. В начале работы мы проводим масштабный анализ литературы для сбора подходящих данных и их предобработки. Особое внимание уделяется изучению кросс-царственных взаимодействий, для чего специально

подбираются филогенетически далёкие объекты из разных систематических групп. Анализ метагеномных данных осуществляется с использованием классификатора Kraken2 с последующим уточнением таксономического профиля с помощью программы Bracken. Аннотация метагенома позволяет выявить геномы микроорганизмов, их таксономическую принадлежность. Канонический корреспондентный анализ дифференциальной экспрессии генов организма-хозяина и метагеномных данных определяет устойчивую взаимосвязь между присутствием конкретных таксонов микроорганизмов и изменениями в экспрессии генов хозяина.

Далее геномы микроорганизмов, полученные по результатам анализа таксономии, мы скачиваем с базы данных NCBI и разделяем на k-меры фиксированной длины, которые затем фильтруем по уникальным паттернам для выделения регионов, потенциально вовлеченных в антисмысловые взаимодействия. Картографирование k-меров на транскрибируемые регионы позволяет выявить перекрывающиеся последовательности-кандидаты, способные функционировать как эффекторы eRNAi. Для верификации полученных результатов проводим анализ дифференциальной экспрессии с применением PerMANCOVA.

На следующем этапе в работе мы планируем применить методы машинного обучения. Такие признаки, как позиционная энтропия, частота k-меров и термодинамическая стабильность РНК-дуплексов, будут использоваться в качестве ключевых признаков для построения классификатора. Это позволит сократить список кандидатов и отобрать наиболее значимые мишени, ассоциированные с механизмом РНК-интерференции окружающей среды.

Изучение механизмов РНК-интерференции открывает новые перспективы для исследования взаимодействий между организмами. Важным направлением является выявление потенциальных мишеней eRNAi, что имеет не только фундаментальное, но и прикладное значение. Интеграция биоинформатического анализа с методами машинного обучения позволит более точно определять эффекторы eRNAi. В будущем методы изучения и выявления мишеней РНК-интерференции

окружающей среды могут быть адаптированы для разработки РНК-препаратов в сельском хозяйстве для обеспечения устойчивости растений к патогенам путем подавления ключевых генов микроорганизмов.

Список использованных источников:

1. Whangbo, J. S., Hunter, C. P. (2008). Environmental RNA interference. Trends in Genetics, 24(6): 297-305.
2. Carbonell, A. (2022). RNAi tools for controlling viroid diseases, Virus Research, 313: 198729.
3. Khan, T. and Zhang, X. (2024). CRISPR-Cas9-based genome editing in *Saccharomyces cerevisiae*: current applications and future perspectives, Trends in Biotechnology.

Метапротеомный подход к анализу бактериальных культур и микробиомов

Е. М. Казакова, М. В. Иванов, И. А. Тарасова

*Институт энергетических проблем химической физики им
В.Л.Тальрозе ФИЦ ХФ РАН, Ленинский проспект 38 к.2, г. Москва,
Россия, 119334,
kazakovaem@gmail.com*

Микроорганизмы являются неотъемлемой частью многих экосистем, они обитают в почве, мировом океане, органах и тканях живых организмов. Исследования показывают взаимосвязь между состоянием микробиома кишечника и развитием многих болезней, таких как расстройства нервно-психического развития, заболевания желудочно-кишечного тракта, нейродегенеративные заболевания [1,2] . В то же время, определение эталонного «здорового» микробиома представляет собой сложную задачу, так как микробиомы разных людей в значительной степени отличаются друг от друга [3]. Наиболее широко используемые на сегодняшний день подходы к анализу микроорганизмов и их сообществ это 16s

рРНК секвенирование, метагеномное секвенирование и MALDI TOF масс-спектрометрия. Использование методик, основанных на анализе совокупности генетического материала сообщества, позволяет установить видовой состав микробиома и предсказать его метаболическую активность. Метод MALDI TOF, применяемый для анализа чистых культур микроорганизмов, позволяет с высокой точностью определить видовую принадлежность исследуемого образца по спектрам рибосомальных белков. Метод метапротеомного анализа позволяет установить как количественный таксономический состав микробиома, так и его метаболическую активность. Однако, инструментальные подходы, используемые в метапротеомных исследованиях, обладают существенным недостатком - длительным (порядка нескольких часов на один образец) временем анализа [4]. Использование таких хроматографических градиентов обосновано сложностью исследуемого образца и делает исследование с применением метода метапротеомики очень дорогостоящими.

В данном исследовании представлен подход к метапротеомному анализу бактериальных культур и микробиомов методом ультрабыстрой хромато-масс-спектрометрии. Масс-спектрометрический метод основывается на использовании коротких хроматографических градиентов с измерением масс-спектров первого уровня [5]. Такой подход позволяет существенно сократить время анализа одного образца, а также снимает проблему возникновения химерных спектров ионов-фрагментов. Разработанный биоинформатический подход позволяет решать такие задачи, как идентификация таксономического состава микробиома без предварительной информации об образце («слепой поиск»), сравнение составов различных образцов и анализ метаболической активности микроорганизмов. Разработанные алгоритмы представлены в CLI-приложении MetaDirectMS1 на языке python.

В ходе работы использовались следующие типы образцов: 19 бактериальных культур, 3 набора модельных микробиомов, составленных из штаммов почвенных бактерий, 3 штамма бактерий-деструкторов, для которых производился анализ метаболической активности, а также пробиотические добавки и образцы микробиомов крыс с моделью болезни Паркинсона.

Применение разработанного алгоритма «слепого поиска» позволило с точностью 95% установить таксономический состав для образцов бактериальных культур и 88% - для образцов модельных микробиомов. Алгоритм количественного сравнения составов различных образцов был протестирован на образцах модельных микробиомов. Рассчитанные с применением алгоритма кратные изменения массовой доли штаммов совпали с истинными со средним коэффициентом корреляции Пирсона 0,88. В ходе исследования было также проведено сравнение результатов анализа изменения метаболической активности, полученных по данным MS1-протеомики, с результатами, полученными с применением классических методов количественного протеомного анализа. Было показано, что разработанная методика позволяет определить за существенно меньшее время анализа тот же набор ключевых ферментов, ответственных за изменение метаболизма микроорганизмов, что и классические методики [6]. Анализ протеома штамма *G. alkanivorans* 135 подтвердил механизм деградации дибензотиофена, предсказанный по данным количественного транскриптомного анализа, и выявил механизм дальнейшего расщепления продуктов десульфуризации [7]. Разработанная в данном исследовании методика анализа микроорганизмов и их сообществ позволяет по результатам одного масс-спектрометрического эксперимента охарактеризовать образцы с точки зрения их состава и метаболической активности. За счет короткого экспериментального времени метод применим для анализов больших популяционных выборок или скрининга состояния пациента во времени.

Авторы благодарят В.Г. Згуду (ЦКП «Протеом человека»), А.В. Третьякова и А.А. Шелепчикова (ФГБУ «ВГНКИ»), В. Горшкова и Ф. Кьелдсена (факультет биохимии и молекулярной биологии, Университет Южной Дании) за реализацию метода ультрабыстрой протеомики DirectMS1 и И.П. Соляникову и Я.А. Делеган (Институт биохимии и физиологии микроорганизмов им. Г.К. Скрыбина РАН) за подготовку образцов почвенных бактерий.

1. J. Manos (2022) The human microbiome in disease and pathology, *APMIS*, **130(12)**:690-705.
2. S.G. Sorboni et al. (2022) A Comprehensive Review on the Role of

- the Gut Microbiome in Human Neurological Disorders, *Clin. Microbiol. Rev. American Society for Microbiology*, **35(1)**.
3. J. Lloyd-Price et al. (2016) The healthy human microbiome, *Genome Med.*, **8(51)**.
 4. Van Den Bossche T. et al. (2021) Critical Assessment of MetaProteome Investigation (CAMPI): a multi-laboratory comparison of established workflows, *Nat. Commun. Nature Publishing Group*, **12(7305)**.
 5. Ivanov M.V. et al. (2021) Boosting MS1-only Proteomics with Machine Learning Allows 2000 Protein Identifications in Single-Shot Human Proteome Analysis Using 5 min HPLC Gradient, *J. Proteome Res. American Chemical Society*, **20(4)**: 1864–1873.
 6. Kazakova E. et al. (2024) Ultrafast metaproteomics for quantitative assessment of strain isolates and microbiomes, *Microchem. J.*, **207**
 7. Frantsuzova E. et al. (2025) Integrated Omics Approaches to Explore a New System of Genetic Control of Dibenzothiophene Desulfurization and Aromatic Ring Cleavage by *Gordonia alkanivorans* Strain 135, *Biology. Multidisciplinary Digital Publishing Institute*, **14(2)**:188

Поиск редких белковых структур и исследование на них качества работы AlphaFold, ESMFold и ProteinMPNN

А.Н. Калинин, Д.Н. Иванков

*Сколковский институт науки и технологий,
Aleksei.Kalinkin@skoltech.ru*

В последние годы инструменты структурной биоинформатики, основанные на методах глубокого обучения, получили широкое распространение. Эти модели основаны на закономерностях, выученных из тренировочных данных. Однако остаётся открытым вопрос, способны ли такие модели выучить нижележащие физические принципы, хотя бы в ограниченной степени. Поэтому необходима проверка таких методов при работе с редкими или отсутствующими в базах данных структурами, работа с которыми является одним из наиболее значимых применений этих

моделей.

Нами был разработан алгоритм поиска редких структур для оценки работы на них современных методов, основанных на глубоком обучении. На первом этапе для каждой структуры из базы данных PDB выполняется поиск структурных аналогов с использованием Foldseek [1]. Затем проводится фильтрация результатов: исключаются совпадения с идентичной аминокислотной последовательностью и $e\text{-value} > 0.001$. Таким образом, сохраняются только те результаты поиска, которые имеют отличные от запроса аминокислотные последовательности и высокое структурное сходство. Число найденных таким образом структур используется как метрика оценки редкости исходной структуры: чем более редкая структура, тем ниже значение метрики. Дополнительно были отсеяны результаты поиска для структур, полученных при помощи метода ЯМР и структур, образованных короткими цепями (менее 20 аминокислотных остатков).

С помощью разработанного алгоритма были сформированы выборки наиболее редких и распространённых структур. На них мы протестировали AlphaFold Server [2], ESMFold [3], как на оригинальных последовательностях, так и на последовательностях, полученных при помощи ProteinMPNN [4] с исходной структуры. Оценка качества работы модели производилась при помощи структурного выравнивания оригинальной и предсказанной структуры при помощи TM-align. Качество работы этих методов на выборке редких структур оказалось нестабильным: наблюдались как хорошие ($TM\text{-score} > 0.9$), так и плохие предсказания ($TM\text{-score} < 0.3$). Указанное выше касается в том числе предсказаний AlphaFold Server, имеющего прямой доступ к PDB и, соответственно, к той структуре, которая должна быть предсказана. Предсказания ESMFold, как правило, были хуже, чем предсказания AlphaFold. То же справедливо и для результатов предсказаний, полученных с использованием ProteinMPNN; однако, для некоторых случаев, предсказания с таких последовательностей были значительно лучше предсказаний с оригинальных последовательностей.

1. M.van Kempen, S.S.Kim, C.Tumescheit et al. (2024) Fast and accurate protein structure search with Foldseek, *Nat Biotechnol* **42**: 243–

246.

2. J.Abramson, J.Adler, J.Dunger et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3, *Nature* **630**: 493–500.
3. Z.Lin et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model, *Science* **379**: 1123-1130.
4. J.Dauparas et al. (2022) Robust deep learning–based protein sequence design using ProteinMPNN, *Science* **378**: 49-56.

Поиск полинуклеотидилированных коротких РНК в транскриптомах *E. coli*

Н.М. Камолдинов¹, Н.П. Кольжецов², Н.Ю. Маркелова², В.В. Панюков³, К.С. Шавкунов², О.Н. Озолин²

¹*факультет биотехнологии, Московский Государственный
Университет имени М.В. Ломоносова, Москва 119991, Ленинские
горы, 1
NMK44@mail.ru*

²*Институт Биофизики Клетки РАН (ФИЦ ПНЦБИ РАН), Пушкино
142290, ул. Институтская, 3*

³*Институт Математических Проблем Биологии РАН, Пушкино
142290, ул. профессора Виткевича, 1*

Особенностью текущего периода в молекулярной биологии является быстро растущий интерес к регуляторному потенциалу фрагментов бактериальных РНК, причём не «малых регуляторных РНК» (sRNAs), которые кодируются специальными генами, имеют относительно большой размер и небольшое число целевых партнёров, а коротких РНК (до 40 н.), вырезаемых из транскриптов функциональных генов и приобретающих самостоятельную, но пока мало предсказуемую функцию в клетках. В ходе процессинга такие олигонуклеотиды часто подвергаются различным модификациям, расширяющим функциональный потенциал

конечных продуктов. Примером такой модификации, обнаруженной нами ранее в идентичных фрагментах тРНК LeuT, V, P, Q (LeuTVPQ) и LeuU, но отсутствующей у гомологичных LeuX, LeuZ и LeuW, является удлинение 5'-концевых 19-меров (GCGAAGGTGGCGGAATTGG) нематричными нуклеотидами [1].

Целью исследования была более полная характеристика фрагментов лейциновых тРНК, а также поиск аналогичных транскриптов в данных RNA-seq кишечной палочки (*E. coli*). Для анализа было использовано 2 набора коротких (16-50 н.) прочтений, полученных из транскриптома *E. coli* K-12 MG1655 и 2 набора прочтений для делеционного мутанта этого штамма по гену структурного белка нуклеоида Dps. Библиотеки прочтений были получены с использованием концевых адаптеров, фиксирующих оба конца РНК и, после удаления адаптеров, отфильтрованы по качеству на портале Galaxy [2]. Они содержали 1169138 ± 348072 и 1168606 ± 96600 последовательностей, соответственно.

Для поиска целевых прочтений был разработан алгоритм RandExt, картирующий на геном 5'-концевые k-меры каждого прочтения заданной длины; оценивающий совпадение с геномом соседних последовательностей разной длины (L) и записывающий удлиненные олигонуклеотиды в выводной файл вместе с координатами 5'-конца k-мера, картированного на геном. Игнорировались гомонуклеотидные треки, появляющиеся в результате полинуклеотидилирования бактериальных РНК, а также прочтения с переменными фланками, содержащими правильные нуклеотиды, числом больше, чем L-4.

Анализ был проведен для k-меров длиной от 12 до 23 нуклеотидов, имеющих нематричные фланки длиной от 4 до 10 н. Явной разницы в количестве прочтений с переменными 3'-концами между транскриптомами дикого и мутантного штаммов обнаружено не было. Проявляя зависимость от длины k-мера и переменного фланка их общее число для каждого штамма без учета копий превысило 2000. Среднее число позиций в геноме, продукты которых оказались модифицированными в сходных по размеру транскриптомах составило 248 ± 14 и 228 ± 12 для исходного и мутантного штаммов, соответственно. На Рисунке 1 в качестве примера показано распределение контекста фланков на фрагментах тРНК-LeuTVPQ с заменой U на A в первой позиции. Несмотря на

малый размер транскриптома, в нём обнаружено 24 из 27 возможных тетрамеров, соответствующих критериям отбора, что свидетельствует о случайности процесса встраивания нематричных нуклеотидов.

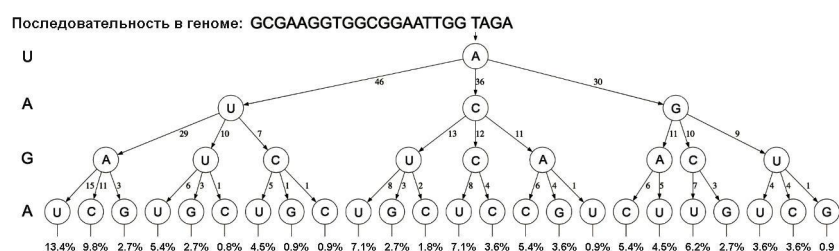


Рисунок 1. Фрагмент схемы распределения последовательностей фланков для 19-меров, принадлежащих tRNA-LeuTVPQ в транскриптоме *E. coli* K-12 MG1655 размером 821065 прочтений.

Распределение модифицированных олигонуклеотидов по геному, наоборот, оказалось явно неслучайным. Хотя основная часть прочтений с негеномными последовательностями на концах была представлена единичными копиями, от 48 до 82 геномных локусов в разных экспериментах продуцировали множественные олигонуклеотиды с варибельными концами. Из них 24 были найдены в транскриптомах всех экспериментов и имели идентичные 5'-концы. В их число входят продукты 8 локусов, расположенных в генах LeuTVPQ, 4 продукта межгенных областей перед генами tРНК IleV(UT) и GltW, 7 транскриптов из промоторных областей генов 16S-рРНК, 3 фрагмента малых регуляторных РНК RdlA(BC), РНК антисмысловая к sRNA RyeA и фрагмент мРНК. Процесс случайного включения нематричных нуклеотидов на 3'-концы РНК, следовательно, не является уникальным свойством LeuTVPQ тРНК, хотя они вносят большой вклад в число таких транскриптов в клетках.

Функциональная роль такой пост-транскрипционной модификации пока мало понятна. Наличие случайных последовательностей на 3'-конце сообщает фрагментам РНК способность формировать «якорные» дуплексы со многими нуклеиновыми кислотами, влияя на их стабильность или

функциональность.. Такие олигонуклеотиды могут быть универсальными «губками» (sponges) [3], функционально извлекающими нетранслируемые РНК из метаболизма клеток, или «РНК-хищниками» (predators) [4], способствующими их разрезанию по типу РНК-интерференции. Они также могут быть универсальными праймерами в реакциях обычной или обратной транскрипции, приводя к образованию нефункциональных химер. Детальное изучение роли модифицированных фрагментов LeuTVPQ тРНК, их биогенеза и поиск структурных факторов, отличающих фрагменты этих тРНК от аналогичных фрагментов LeuWXX тРНК станет предметом следующих исследований. Работа была поддержана грантом РФФИ 24-14-00433.

1. K.S. Shavkunov et al. (2024) Products of abortive transcription can prime synthesis of chimeric oligonucleotides, *Mathematical Biology and Bioinformatics*, 19:453-471.
2. The Galaxy Community (2024) The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update, *Nucleic Acids Research*, 52:W83–W94
3. G. Matera et al. (2022) Global RNA interactome of Salmonella discovers a 5' UTR sponge for the MicF small RNA that connects membrane permeability to transport capacity, *Molecular Cell*, 82:629-644.e4
4. W. Kim and Y. Lee (2020) Mechanism for coordinate regulation of rpoS by sRNA-sRNA interaction in Escherichia coli, *RNA Biology*, 17: 176-187

Микробные сигнатуры кишечника как маркеры эффективности иммунотерапии рака

В.А. Канаева, Е.И. Олехнович

*Малая Пироговская ул., д. 1а, Москва, Россия, 119435,
vera.a.kanaeva@gmail.com*

Иммунотерапия, основанная на ингибиторах иммунных контрольных точек (ИИКТ), в значительной степени изменила

подходы к лечению метастатических злокачественных новообразований, демонстрируя впечатляющие клинические результаты. Несмотря на это, клинические исходы остаются гетерогенными: до 50% пациентов демонстрируют первичную резистентность или ранний рецидив после терапии [1]. Кроме того, применение ИИКТ сопряжено с риском развития тяжелых иммуноопосредованных осложнений [2], что актуализирует поиск достоверных предикторов ответа для персонализации терапии. Многочисленные исследования отмечают связь между составом кишечной микробиоты и ответом на терапию [3,4]. Однако механизмы этого взаимодействия и надёжные предикторы результатов лечения остаются неясными. Первоначальные исследования роли кишечного микробиома в модуляции ответа на терапию ИИКТ преимущественно фокусировались на меланоме [3,5]. Тем не менее, в свете современной гипотезы об универсальных механизмах микробиом-иммунного взаимодействия, возникает необходимость расширения исследований на другие нозологические формы.

В этой работе были объединены данные из 11-ти ранее опубликованных исследований для комплексного метагеномного анализа. Набор данных насчитывает 624 образца кишечных метагеномов, охватывает более пяти типов рака и включает различные географические регионы на трех континентах. После контроля качества и фильтрации человеческой ДНК была выполнена метагеномная сборка и биннинг, с последующей дерепликацией. В результате был получен избыточный каталог из 3,649 геномов, собранных из метагеномов (metagenome-assembled genome - MAG). Для определения метагеномных маркеров был получен профиль представленности MAG'ов и проведено их ранжирование по степени ассоциации с исходом иммунотерапии. Аннотация маркеров включала определение их таксономической принадлежности и происхождения (кишечник, другие участки тела человека, пищевые продукты), и функциональный анализ с использованием баз данных CAZy [6] и KEGG [7]. Статистическая оценка проводилась на основе логарифмических отношений [8] представленности маркерных MAG'ов в группах ответивших (response – R) и не ответивших (non-response – NR) пациентов по принципу GSEA (Gene-set enrichment analysis – анализ обогащения

наборов генов).

В результате анализа выявлены 637 метагеномных маркеров (298 ассоциированы с R и 339 с NR). Статистический анализ тестом Тьюки с поправкой Бонферрони (скорр. $p < 0,001$) подтвердил высокую дифференцирующую способность этих маркеров, что подкрепляется большим размером эффекта (Cohen's $d = 0,91$). Прогностическая модель на основе логистической регрессии с "leave-one group out" кросс-валидацией продемонстрировала хорошую точность (AUC = 0,72) в предсказании исхода лечения. Таксономический анализ показал ассоциированность Bacteroidota с благоприятным исходом лечения, а Proteobacteria с отсутствием ответа. В числе бактериальных родов связанных с R выявляются Blautia, Bacteroides, Fusicatenibacter, Gemmiger, напротив Acetatifactor, Dysosmobacter, Prevotella, Veillonella ассоциированы с NR. Род Faecalibacterium показал двойственное распределение, но со значимым преобладанием в группе R. Также была выявлена значимая ($p < 0,001$) связь между присутствием оральных и пищевых микробов в профиле кишечной микробиоты с отсутствием ответа на иммунотерапию. В функциональном профиле группы R было выявлено значимое обогащение путей метаболизма гистидина, 2-оксокарбоновой кислоты и серы, биосинтеза аминокислот, кофакторов и взаимопревращения пентозы и глюкуроната (KEGG), а также гликозидазной активности (CAZy). В группе NR оказались обогащены такие пути KEGG, как биосинтез липополисахаридов и сборка жгутика.

Одним из важных результатов исследования стало выявление рода Blautia как маркера положительного ответа на терапию. Представители этого таксона обладают высоким гликозидазным потенциалом (CAZy), позволяющей метаболизировать сложные углеводы с образованием короткоцепочечных жирных кислот — ацетата и пропионата [9], которые участвуют в поддержании барьерной функции кишечника. Роль Faecalibacterium оказалась неоднозначной, что согласуется с данным Gunjur et al. [10] и может быть связано с функциональным полиморфизмом различных штаммов. Обнаружение оральных и пищевых контаминантов в кишечнике у NR-пациентов может свидетельствовать о деградации микробного сообщества кишечника [11], что, в свою очередь, может быть связано с

ослаблением иммунной регуляции. Анализ KEGG-путей показал, что микробиота пациентов в группе R характеризуется повышенной метаболической активностью и функциональной пластичностью, что может способствовать более выраженному противоопухолевому иммунному ответу. Напротив, у NR-пациентов наблюдалось доминирование представителей отдела Proteobacteria, обладающих жгутиковой подвижностью. Известно, что такие виды способны вытеснять симбиотические популяции [12], нарушая экосистему кишечника. Кроме того, увеличение биосинтеза липополисахаридов у этих бактерий создаёт провоспалительный фон через активацию TLR4/5 [13,14], что может негативно влиять на эффективность иммунотерапии.

Исследование позволило определить потенциальные бактериальные маркеры ответа на иммунотерапию, включая перспективный род *Blautia*. Его участие в метаболизме углеводов и поддержании иммунного гомеостаза открывает возможности для новых подходов к модуляции микробиоты с целью повышения эффективности лечения. Полученные данные требуют дальнейшей экспериментальной и клинической верификации, однако уже сейчас демонстрируют значимость микробиоты в контексте индивидуализированного противоопухолевого иммунитета.

Исследование поддержано грантом РФФИ (№ 22-75-10029, <https://rscf.ru/project/22-75-10029/>).

1. Larkin J. et al. (2015) "Combined nivolumab and ipilimumab or monotherapy in untreated melanoma.", *New England journal of medicine*, 373.1: 23-34.
2. Robert C. et al. (2019) "Pembrolizumab versus ipilimumab in advanced melanoma (KEYNOTE-006): post-hoc 5-year results from an open-label, multicentre, randomised, controlled, phase 3 study.", *The Lancet Oncology*, 20.9: 1239-1251.
3. A.E. Frankel et al. (2017) "Metagenomic shotgun sequencing and unbiased metabolomic profiling identify specific human gut microbiota and metabolites associated with immune checkpoint therapy efficacy in melanoma patients.", *Neoplasia*, 19.10: 848-855.
4. Z. Peng et al. (2020) "The gut microbiome is associated with clinical response to anti-PD-1/PD-L1 immunotherapy in gastrointestinal cancer.", *Cancer Immunology Research*, 8.10: 1251-1261.

5. E.N. Baruch et al. (2021) "Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients." *Science* 371.6529: 602-609.
6. E. Drula et al. (2022) "The carbohydrate-active enzyme database: functions and literature.", *Nucleic acids research*, 50.D1: D571-D577.
7. M. Kanehisa et al. (2025) "KEGG: biological systems database as a model of the real world.", *Nucleic acids research*, 53.D1: D672-D677.
8. J.T. Morton et al. (2019) "Establishing microbial composition measurement standards with reference frames.", *Nature communications*, 10.1: 2719.
9. S.M. Holmberg et al. (2024) "The gut commensal *Blautia* maintains colonic mucus function under low-fiber consumption through secretion of short-chain fatty acids.", *Nature Communications*, 15.1: 3502.
10. A. Gunjur et al. (2024) "A gut microbial signature for combination immune checkpoint blockade across cancer types.", *Nature Medicine*, 30.3: 797-809.
11. C. Liao et al. (2024) "Oral bacteria relative abundance in faeces increases due to gut microbiota depletion and is linked with patient outcomes.", *Nature microbiology*, 9.6: 1555-1565.
12. B. Routy et al. (2018) "The gut microbiota influences anticancer immunosurveillance and general health.", *Nature Reviews Clinical Oncology*, 15.6: 382-396.
13. L. Mazgaeen, and P. Gurung (2020) "Recent advances in lipopolysaccharide recognition systems.", *International journal of molecular sciences*, 21.2: 379.
14. F. Hayashi et al. (2001) "The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5.", *Nature*, 410.6832: 1099-1103.

Gene expression normalization in problematic data by DEAr

Karpenko D.V.

National Medical Research Center for Hematology, Russian Federation, Moscow, Novy Zykovsky Proezd, 4, d_@list.ru

Gene expression is a key analysis in modern medical and biological studies. High-throughput sequencing provides vital information for modern science. However, the amount of information

provided is coupled with the high cost of a measurement. It is common practice to complement next-generation sequencing with real-time PCR data. Real-time PCR is the method of choice for measuring the expression of one or a few genes of interest. Studying the expression of specific genes is required in clinical diagnostics and experimental biology to test a hypothesis under certain or different conditions.

The essential step in analysis is normalization of acquired data to make gene expression in different samples comparable. Sample processing results in unknown inconsistency in quality and quantity of analyzed RNA. The standard solution is to use internal controls in the output data. Expression of genes, known as housekeeping genes, is considered a stable inner reference. However, expression of housekeeping genes can vary significantly depending on conditions of an experiment or different cell subpopulations in a study. High-throughput sequencing allows a researcher to select stable genes from a large number of all genes in a dataset. Real-time PCR provides a more constrained set of choices, limited by preselected genes. The strategy here is to select less variable genes. However, genes that are used as references can change their expression depending on a sample. Although we can identify instability in a reference gene, there are no standard protocols to correct for this instability. Standard protocols suggest additional measurements and analysis for a new gene to check if it is stable enough. There are also difficulties when there are missing measurements of genes, especially housekeeping genes. Missing values are usually a subject of imputation by data, which is at best a neutral introduction of biased information, and at worst can significantly disturb an analysis. To solve these problems, I introduce DEAr (Differential Expression Analyzer).

DEAr performs normalization by fitting an expression pattern of each sample to a general pattern of a data set. Abnormalities in a pattern are dynamically identified for all genes in all samples. An abnormal value of gene expression in a sample is penalised in terms of its contribution to a general pattern. An original algorithm allows integration of penalties according to inaccuracies in provided data and preliminarily known instabilities in gene expressions. A penalty can make an associated value effectively insignificant. This means that such a value could be a missing value. The architecture of the algorithm allows it to work with data sets with missing values without imputation or

deletion. This highlights the key difference with standard algorithms, which require all expression values in housekeeping genes to be identified to form a summarized normalization level.

At the core of the algorithm are nested recursions, which enable the identification of anomalies and their correction. DEAr is the specific machine intelligence designed to solve specific tasks. DEAr can take and return structured data as an Excel or Python file. DEAr is packaged in an executable file and requires no programming skills or software to run. In summary, DEAr is an essential tool for working with gene expression analysis on limited or problematic data sets. DEAr can perform normalization of data from different sources, including counting for a number of molecules or reads for a gene. Accurate and imprecise measurements can be handled by DEAr, with more attention paid to accurate and reproducible data. Zeros or values associated with high or low outbreaks from a dynamic range of measurements could be processed as part of the DEAr analysis. Missing values could be handled without prior input or deletion, and without them during the DEAr run. Information from external sources could be provided to DEAr for integrated analysis. A detailed description and the program are presented in a preprint [1].

1 D.V. Karpenko (2023) DEAr – Differential Expression Analyzer, *[Preprint]*, doi:10.21203/RS.3.RS-2957165/V3.

Молекулярное моделирование механизма связывания глицина и ионов с транспортером GlyT1

С.А. Карпычева¹, Л.А. Ивонцин², Е.В. Машковцева^{1,2}, Я.Р. Нарциссов^{2,3}

¹ФГАОУ ВО "Российский Национальный Исследовательский Медицинский Университет им. Н.И. Пирогова" МЗ РФ, Москва, Россия, *karpycheva98@mail.ru*

²НИИ цитохимии и молекулярной фармакологии, Москва, Россия

³*Группа биомедицинских исследований, БиДиФарма ГмбХ, Зик, Германия*

Глицин – один из основных тормозных нейромедиаторов центральной нервной системы, играющий ключевую роль в реализации нейротрансмиссии. Его концентрация в синаптической щели регулируется за счет обратного захвата с помощью глициновых транспортеров (GlyT1,2), которые относятся к классу нейротрансмиттерных натриевых симпортеров NSS. Переносчик GlyT1 локализуется в основном на мембранах астроцитов и глутаматергических нейронов, в то время как GlyT2 – в пресинаптических терминалях глицинергических нейронов. Оба транспортера осуществляют перенос глицина внутрь клетки совместно с ионами Na⁺ и Cl⁻ [1]. Изучение механизма работы глициновых переносчиков имеет большое значение для понимания процессов синаптической передачи на молекулярном уровне, а также для разработки терапевтических стратегий при лечении шизофрении, нейропатической боли и других неврологических расстройств [2]. При этом, особый интерес представляет GlyT1, поскольку он выполняет более универсальные функции: участвует не только в завершении тормозной нейропередачи, но и в модуляции возбуждающей активности через влияние на глицин-зависимую коактивацию NMDA-рецепторов. Кроме того, нарушения его работы могут быть вовлечены в патогенез эпилепсии, инсульта и нейродегенеративных заболеваний, что подчеркивает его клиническую значимость.

За последние пять лет с помощью методов рентгеноструктурного анализа и криоэлектронной микроскопии впервые были получены кристаллические структуры транспортера GlyT1 в высоком разрешении [3-5], что позволило изучить механизмы транслокации, распознавания и связывания лигандов во время транспортного цикла переносчика на атомистическом уровне. В данной работе проводилось исследование координации ионных и субстратных взаимодействий в транспортере GlyT1 методами молекулярного моделирования.

Начальные координаты атомов белка были извлечены из структуры GlyT1, разрешенной в открытой во внеклеточную среду конформации (PDB ID: 8WFL). Недостающие остатки 183-201 во

внеклеточной петле между трансмембранными спиралями TMC3 и TMC4 были достроены с использованием программного пакета PyRosetta 4.0. Мембранно-белковый комплекс собран при помощи веб-сервиса CHARMM-GUI. Переносчик был встроен в липидный бислой, состоящий из молекул фосфатидилхолина (POPC), а затем сольватирован 150 мМ раствором NaCl. В качестве естественного субстрата GlyT1 в среду была добавлена молекула глицина в виде цвиттер-иона, что соответствует форме этого соединения при физиологическом значении pH 7,4 в синаптической щели. Классическое молекулярно-динамическое (МД) моделирование выполнялось при помощи программного обеспечения GROMACS 2023.5 с использованием силового поля CHARMM36m. После начальной процедуры минимизации энергии и уравнивания МД моделирование производилось без ограничений в NPT ансамбле в течение 1,5 мкс.

Однако известно, что ассоциация субстратов с сайтами связывания белка представляет собой «редкое событие», характеризующееся несколькими метастабильными состояниями, переход между которыми происходит на временных масштабах, выходящих за рамки возможностей классической молекулярной динамики. В связи с этим также было проведено моделирование с использованием методов метадинамики, которые позволяют ускорить исследование конформационного пространства реакции. Моделирование методом метадинамики выполнялось в соответствии с расширенным протоколом Funnel-Metadynamics (FM) при помощи программных пакетов GROMACS 2023.5, PLUMED 2.9 и VMD. Параметры FM подбирались таким образом, чтобы обеспечить эффективную оценку свободной энергии связывания лигандов с белком.

Изучение последовательности проникновения ионов в переносчик важно для понимания эффективности транспорта и выявления механизмов, лежащих в основе нейрохимических нарушений. В результате анализа МД траекторий был установлен порядок связывания субстратов: сначала происходила ассоциация катиона Na1, затем наблюдалось последовательное связывание аниона Cl и катиона Na2. Завершающим этапом данного процесса стало связывание молекулы глицина. Также, были

идентифицированы основные аминокислотные остатки, расположенные на трансмембранных спиральях TMC1, 3, 6 и 8, которые отвечают за перенос и координацию лигандов в соответствующих сайтах связывания. Обнаруженные аминокислоты согласуются с данными экспериментальных исследований GlyT1 и его гомологов [3-4]. Кроме того, был получен энергетический профиль, отражающий процесс связывания глицина с транспортером, и проведена оценка свободной энергии связывания $\Delta G = -8,4$ ккал/моль, характеризующая аффинность глицина к GlyT1.

Полученные данные имеют важное значение для фундаментального понимания механизмов функционирования переносчика GlyT1 и других представителей NSS, имеющих общий LeuT-фолдинг. Кроме того, результаты исследования в дальнейшем могут быть использованы для поиска и изучения потенциальных фармакологических соединений, способных корректировать работу глицинергической и глутаматергической систем.

1. К.С. Зайцев, Е.В. Машковцева, Я.Р. Нарциссов (2012) Мембранные переносчики аминокислоты глицин в нервной ткани: структура, локализация, основные функции и регуляция, Успехи современной биологии, 132(4):391–400.
2. R.J. Harvey, B.K. Yee (2013) Glycine transporters as novel therapeutic targets in schizophrenia, alcohol dependence and pain, Nature Reviews Drug Discovery, 12(11):866–885.
3. A. Shahsavari et al. (2021) Structural insights into the inhibition of glycine reuptake, Nature, 591:677–681.
4. Y. Wei et al. (2024) Transport mechanism and pharmacology of the human GlyT1, Cell, 187(7):1719–1732.e14.
5. N. Li et al. (2025) Modulation of the human GlyT1 by clinical drugs and cholesterol, Nature Communications, 16:2412.

Сравнение пространственной организации хроматина нейронов и книдоцитов у *Nematostella vectensis*

А. В. Каштанова¹, Я. В. Ким², Arnau Sebé-Pedrós²

¹г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73,
Факультет биоинженерии и биоинформатики,
anastasiia.kashtanova@gmail.com

²88 of Dr. Aiguader St. Barcelona, Spain, The Centre for Genomic
Regulation (CRG)

Пространственная организация хроматина играет важную роль не только в упаковке генетического материала в ядре, но и в регуляции экспрессии генов. Среди ключевых уровней укладки генома выделяют петли хроматина, топологически ассоциированные домены (ТАДы) и хроматиновые фонтаны, А/В-компарменты — участки, обогащённые транскрипционно активным или неактивным хроматином соответственно, — а также хромосомные территории и межхромосомные взаимодействия. Эти архитектурные элементы описаны в основном у модельных организмов и не являются универсальными для всех групп животных. Их наличие, а также механизмы формирования могут существенно различаться, особенно у представителей ранних ветвей эволюционного древа.

Книдарии, как сестринская группа билатерий, играют важную роль в понимании эволюции пространственной организации генома и механизмов регуляции транскрипции. Одной из ключевых особенностей этих животных являются книдоциты — уникальные специализированные клетки, способные к мгновенному выбросу стреляющего органоида, которые используются для охоты и защиты. Считается, что книдоциты эволюционно произошли от нейронов. Это подтверждается рядом данных: книдоциты и нейроны развиваются из общих клеточных предшественников, имеют сходные транскрипционные регуляторы, и во многих случаях разделяют компоненты молекулярной сигнальной и экзоцитарной машинерии. В связи с этим *Nematostella vectensis* — морская анемона и представитель книдарий — широко

используется как модельный объект в эво-дево исследованиях и представляет особый интерес для изучения ранних этапов эволюции нейрональных клеток.

В данной работе мы провели сравнительный анализ пространственной организации хроматина в книдоцитах и нейронах *Nematostella vectensis*, с целью выявления различий в укладке генома, связанных с функциональной специализацией этих клеток.

Для сравнительного анализа мы использовали данные Micro-C, полученные отдельно для нейронов и книдоцитов *Nematostella vectensis*. Сравнение контактных карт показало различия в укладке хроматина между этими двумя типами клеток. Мы провели идентификацию участков инсуляции — границ доменов, ограничивающих взаимодействия хроматина, — и обнаружили как общие участки инсуляции, присутствующие в обоих типах клеток, так и специфичные для нейронов. В то же время участки инсуляции, характерных исключительно для книдоцитов, оказалось значительно меньше.

Анализ геномного контекста выявленных участков инсуляции показал, что они обогащены генами. Более того, во многих из них наблюдалось наличие пар генов с конвергентной ориентацией транскрипции, что позволяет предположить возможный механизм формирования изолирующих структур хроматина на основе сближения терминальных участков транскрибируемых генов.

Дополнительно, в тех участках инсуляции, которые перекрываются с регионами открытого хроматина (по данным ChIP-seq), был проведён поиск обогащения мотивами. Среди них мы обнаружили значимое обогащение мотива YY1 — фактора, известного своей ролью в петлевом взаимодействии и формировании трёхмерной архитектуры генома у других животных.

Convergent 3D genome architecture alterations in SIRT6 deficiency, aging, and neurodegeneration

E. Kashuk¹, D. Smirnov^{1,2,3}, E. Eremenko^{2,3}, A. Golova¹, D. Toiber^{2,3},
E. Khrameeva¹

¹*Center for Molecular and Cellular Biology, Skolkovo Institute of
Science and Technology, Moscow 121205, Russia,
ekaterina.kashuk@skoltech.ru*

²*Department of Life Sciences, Ben-Gurion University of the Negev,
Beer Sheva 8410501, Israel*

³*The Zlotowski Center for Neuroscience, Ben-Gurion University of the
Negev, Beer Sheva 8410501, Israel*

Background

SIRT6 is an NAD⁺-dependent deacetylase and mono-ADP-ribosyltransferase. It selectively deacetylates histone H3, enforcing chromatin compaction and transcriptional repression of genes involved in stress and inflammation. Beyond histones, SIRT6 mono-ADP-ribosylates and activates PARP1, thereby accelerating repair of DNA double-strand breaks – a process especially crucial in nondividing cells such as neurons [1].

Through these actions, SIRT6 preserves genomic stability, supports telomere maintenance, and suppresses pro-inflammatory gene expression, positioning it as a key modulator of cellular senescence and organismal longevity [2]. Age-related and Alzheimer's disease models consistently show reduced SIRT6 levels, correlating with elevated DNA damage, compromised chromatin organization, metabolic dysregulation, and activation of aging pathways [3]. In mice, loss of SIRT6 yields a progeroid phenotype – spinal deformities, metabolic defects, and cognitive decline – underscoring its vital role in healthspan and lifespan determination [4].

In neuronal tissues, SIRT6 is highly expressed in the cortex and hippocampus, where it facilitates both homologous recombination (HR) and nonhomologous end-joining (NHEJ) repair mechanisms. It also counteracts amyloid- β_{42} -induced genomic lesions via a p53-dependent pathway, suggesting that maintaining SIRT6 activity may mitigate Alzheimer's pathology. Transcriptomic studies in AD mouse models link SIRT6 deficiency to upregulation of inflammatory and stress-response genes, further driving neuronal dysfunction [5].

Data and Methods

To investigate chromatin dynamics in aging and under SIRT6-deficiency, we analyzed Hi-C datasets from cortical neurons of adult wild-type (WT), aged (Old) and SIRT6-knockout (SIRT6-KO) mice. Parallel open-access Hi-C datasets from CK-p25 mice, a validated AD model recapitulating neurodegeneration and tau pathology were included to compare AD-induced chromatin changes [6]. CK-p25 cortical mouse neurons were stratified into three groups: (1) Control (high neuronal nuclei marker NeuN, baseline DNA damage marker γ H2AX), (2) Step 1 (high NeuN, elevated γ H2AX), and (3) Step 2 (low NeuN, high γ H2AX), reflecting progressive neuronal dysfunction. For transcriptomic analysis, RNA sequencing was performed on cortical tissues from WT and shRNA-mediated SIRT6-knockdown (shSIRT6) mice, alongside publicly available datasets from neurodegeneration stages (Control, Step 1, Step 2) and aging (Control, Old) [7, 8].

Results

Hi-C analysis demonstrated progressive disruption of chromatin compartmentalization, showing significantly reduced interactions between transcriptionally active (A) and inactive (B) compartments in both Step 1 and Step 2 neurons, as well as in aged neurons. These findings indicate diminished compartment mixing during both normal aging and neurodegenerative progression. Compartment shifts were dominated by transitions from weak B (wB) to strong B (sB) compartments (18.5% in Step 1, 20.3% in Step 2) and A $\rightarrow$ B shifts (1.6–1.8%), mirroring SIRT6-KO neurons. Functional enrichment of wB $\rightarrow$ sB regions identified immunoglobulin heavy chain variable (Ighv) genes on chromosome 12 (59 genes in aging neurons, 38 in SIRT6-KO neurons,

47 in Step 1, 21 in Step 2), implicating immune pathway dysregulation in aging, SIRT6 deficiency and neurodegeneration. Topologically-associating domains prominence increased in Step 1 and aged neurons compared to controls but decreased in Step 2 and SIRT6-KO neurons, suggesting early chromatin compaction followed by structural decay in advanced neurodegeneration. Chromatin loops were diminished in Step 1/Step 2 and SIRT6-KO neurons, with architectural loops (non-regulatory) preferentially lost and enhancer-promoter loops retained or gained. This selective reorganization indicates a shift from structural chromatin interactions to regulatory element connectivity during progression of neurodegeneration and under SIRT6 deficiency.

Analysis of transcriptomic profiles revealed shared transcriptional signatures between SIRT6-KO, aging, and neurodegeneration. Upregulated genes in shSIRT6/WT and Step1/Step2 neurons were enriched for chromatin modifiers such as histone acetylation and chromatin organization. Key transcription factors associated with altered chromatin interactions and differentially expressed genes (DEGs) in aging, SIRT6 deficiency, and neurodegeneration included REST, which was prominently linked to upregulated interactions in shSIRT6 neurons. TBP, TP53, MYC, and E2F1 – known regulators of cell cycle progression, apoptosis, and stress responses – were associated with increased chromatin interactions and upregulated DEGs in both SIRT6-deficient and neurodegenerative contexts [9]. SUZ12, EZH2, and BMI1, components of Polycomb repressive complex involved in epigenetic silencing and stem cell maintenance, were linked to upregulated interactions and DEGs in SIRT6 deficiency, aging, and neurodegeneration [10].

Conclusion

Together, these findings demonstrate that SIRT6 loss drives chromatin reorganization and transcriptional programs mirroring both natural aging and Alzheimer's-like neurodegeneration, highlighting its central role in maintaining neuronal genome integrity.

1. Akter R. et al. (2021) A Comprehensive Analysis into the Therapeutic Application of Natural Products as SIRT6 Modulators in Alzheimer's Disease, Aging, Cancer, Inflammation, and Diabetes, *Int. J. Mol. Sci.*, 22:4180.
2. Dzidek A. et al. (2023) The Role of Increased Expression of Sirtuin 6 in the Prevention of Premature Aging Pathomechanisms, *Int. J. Mol. Sci.*, 24:9655.
3. Li X. et al. (2021) SIRT6 in Senescence and Aging-Related Cardiovascular Diseases, *Front. Cell Dev. Biol.*, 9:641315.
4. Raj S. et al. (2020) Sirt6 Deacetylase: A Potential Key Regulator in the Prevention of Obesity, Diabetes and Neurodegenerative Disease, *Front. Pharmacol.*, 11:598326.
5. Pradhan R. et al. (2022) Blood Circulatory Level of Seven Sirtuins in Alzheimer's Disease: Potent Biomarker Based on Translational Research, *Mol. Neurobiol.*, 59(3):1440–1451.
6. Dileep V. et al. (2023) Neuronal DNA double-strand breaks lead to genome structural variations and 3D genome disruption in neurodegeneration, *Cell*, 186(20):4404–4421.e20.
7. Welch G.M. et al. (2022) Neurons burdened by DNA double-strand breaks incite microglia activation through antiviral-like signaling in neurodegeneration, *Sci. Adv.*, 8(39):eabo4662.
8. Allen W.E. et al. (2023) Molecular and spatial signatures of mouse brain aging at single-cell resolution, *Cell*, 186(1):194–208.e18.
9. J. Pflaum, S. Schlosser, M. Müller (2014) p53 Family and Cellular Stress Responses in Cancer, *Front. Oncol.*, 4:285.
10. L. Gong et al. (2022) CK2-mediated phosphorylation of SUZ12 promotes PRC2 function by stabilizing enzyme active site, *Nat. Commun.*, 13:6781.

Поиск однонуклеотидных полиморфизмов, приводящих к появлению энхансеров, в клеточной линии K562

Я.С. Киселёва¹, Р.С. Султанов², Г.П. Арапиди²

¹МФТИ (НИУ), 141701, Московская область, г. Долгопрудный,
Институтский переулок, д. 9., kiseleva.ias@phystech.edu

²ФГБУ ФНКЦ ФХМ им. Ю.М. Лопухина ФМБА России, Россия,
Москва, 119435, Малая Пироговская, д. 1а

Энхансеры, одни из ключевых регуляторов экспрессии генов, представляют собой небольшие некодирующие последовательности ДНК и контролируют транскрипцию с промотора гена или группы генов. Транскрипционные факторы (ТФ) связываются с энхансерами через специфические последовательности ДНК, обладающие высокой аффинностью к ТФ [1]. Различные заболевания оказываются связанными с нарушением функции энхансеров из-за хромосомных перестроек, генетических вариаций внутри энхансеров или эпигенетических модификаций [2]. Целью данной работы стал поиск однонуклеотидных полиморфизмов (single nucleotide polymorphism, SNP) клеточной линии K562, приводящих к появлению de novo сайтов связывания транскрипционных факторов (ССТФ) и de novo энхансеров.

В данных полногеномного секвенирования (WGS) клеточной линии K562 (миелогенная лейкемия человека, ENCODE Project) с помощью программы GATK HaplotypeCaller было найдено 3451952 SNP. С использованием базы данных HOCOMOCO, содержащей мотивы связывания ТФ с ДНК, были определены мутации, которые потенциально могут приводить к появлению ССТФ. С помощью модели негативного биномиального распределения мы определили 1154 SNP (суммарно для всех 59 проанализированных ТФ), обладающих значимым аллельным дисбалансом по данным ChIP-Seq на соответствующий ТФ (AI-SNP). Также были проанализированы данные DNase-Seq, ChIP-Seq

на гистоновые модификации H3K27ac/H3K4me1, характерные для активных энхансеров, а также данные Fiber-Seq для изучения влияния AI-SNP на эпигенетический ландшафт линии K562. Для поиска целевых генов предполагаемых *de novo* энхансеров был проведен анализ пространственной структуры хроматина (Hi-C) а также использовались данные известных ассоциаций между SNP и экспрессией генов проекта GTEx (Genotype-Tissue Expression project). Таким образом, мы определили AI-SNP (rs12623363), который приводит к возникновению ССТФ GATA2 и потенциально может влиять на экспрессию белок-кодирующего гена SERTAD2, который является прото-онкогеном, и AI-SNP (rs3748136), связанный с повышением экспрессии длинной некодирующей РНК RP11-10A14.5, для которой показано участие в регуляции метастазирования опухоли [3, 4].

Литература

1. F. Spitz, E. Furlong. Nature Reviews Genetics 2012, doi: 10.1038/nrg3207
2. A. Claringbould, JB. Zaugg. Trends in molecular medicine 2021, doi: 10.1016/j.molmed.2021.07.012
3. Y. Chen, X. Jie, B. Xing et al. Cell Death & Disease 2022, doi: 10.1038/s41419-022-04567-5
4. Z. Lin, F. Feng, J. Liang et al. Discover Oncology 2022, doi: 10.1007/s12672-022-00493-2

Identification of Immunocompromised Bees using 16S rRNA Sequencing and Assessment of Changes in Core Composition due to Disease

Aleksandra Kozlova¹, Anna Izotova², Gennadiy Kachmazov³, Stepan Toshchakov²

¹ *Higher school of economics, Analysis in biology and medicine*

² *Kurchatov Centre for Genome Research, National Research Centre "Kurchatov Institute", 123182 Moscow, Russia*

³ Severo-Osetinskiy Gosudarstvennyy Universitet Imeni K. L.
Khetagurova
Moscow, Russia, amiandwho@gmail.com

The Western honey bee (*Apis mellifera*) is a vital pollinator in global ecosystems and agriculture. Beyond environmental threats such as habitat fragmentation and pesticide exposure, microbial symbionts within the bee gut substantially influence colony health. The gut microbiome of *A. mellifera* is a highly conserved community composed predominantly of seven core genera: *Gilliamella*, *Snodgrassella*, *Bifidobacterium*, *Lactobacillus*, *Frischella*, *Bartonella*, and *Commensalibacter* [1,2]. These microbes are involved in carbohydrate metabolism, immune stimulation, and protection against pathogens [3,4]. However, variations in this core composition – driven by diet, habitat, or disease – can significantly affect bee resilience [5]. While previous studies have characterized the functional roles of these microbes, few have addressed their spatial variability within relation to any factors [5,6]. The study aims to identify pathogenic microflora, identify clusters of diseased bees and characterize changes in core microorganisms in connection with diseases.

The collection of samples for the study was carried out in 2020 in the Republic of North Ossetia-Alania. For the experiment, 40 worker forager bees belonging to the same breed - the North Caucasian gray mountain bee - were selected. The intestines of half of the bees were dissected and sequenced, samples from the other half of the bees were homogenized and sequenced as a whole bee to assess the difference between the microbiome of the whole bee and the intestine. Gut and whole-body DNA was extracted using QIAampPowerFecal kits. The V4 region of the 16S rRNA gene was amplified with a two-step PCR protocol and sequenced on Illumina platforms. Sequencing yielded between 6,600 and 16,000 reads per sample, with rarefaction analysis confirming sufficient depth.

Sequence data were processed using DADA2 [7], and taxonomy was assigned via the Silva138 version 1 database [8]. Next, a primary analysis was performed to assess differences in the microbiome due to different sample processing. Since no differences were found in the microbiome of whole bees and gut samples, an identification of clusters counts was performed with NbClust package [9] and assess the

optimality of attitudes towards these clusters using k-means method from the basic R functions on central relative logarithm (CRL) [10] transformed data (Figure 1A). To determine the microorganisms that influence the allocation of clusters, an indicpecies [11] analysis was conducted. The cluster whose allocation was explained by the prevailing presence of pathogenic organisms of bees was designated as a cluster with reduced immunity. In contrast, the cluster with the absence of pathogenic microorganisms was called healthy. Next, determination of differences in core genres abundance was performed using the t-test on the CLR transformed data on the cluster of healthy bees and for bees with reduced immunity.

The analysis revealed common bacterial genera *Gilliamella*, *Snodgrassella*, *Lactobacillus*, *Bifidobacterium* and *Bartonella* in the microbiome samples. *Frischella* and *Commensalibacter* were present in a smaller number of samples. Microorganisms classified as minor genera were also detected. Unsupervised clustering based on CLR transformed data identified two distinct microbiome groups. Cluster A comprised healthy bees with high core microbial abundance (>90%) and Cluster B included samples enriched in non-core bacteria such as *Klebsiella*, *Morganella*, *Serratia*, and *Hafnia*, suggesting pathological shifts (Figure 1B) [12–15]. High *Frischella* abundance likely reflects immune activation rather than pathogenesis, reinforcing its role as a beneficial commensal during stress [16].

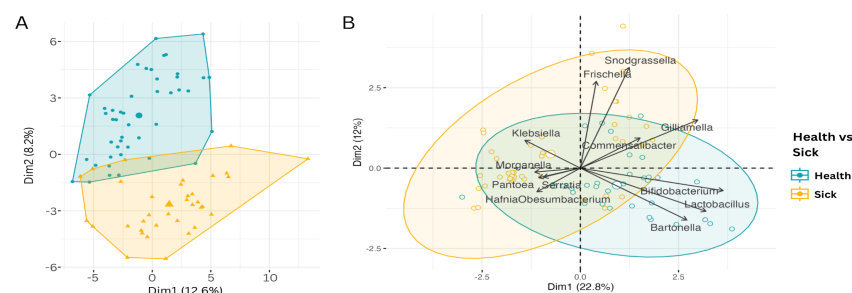


Figure 1. *Apis mellifera* microbiome diversity. A - PCoA display of the assignment of samples to clusters using the K-means method on the Aitchison distance between samples. Each point corresponds to one microbiota sample. B - PCA biplot with displaying samples and features

(bacterial genres). “Sick” - cluster with reduced immunity, “Health” - cluster with absence of pathogenic microorganisms.

In analysis of variation of core microbiome was found the increase in *Gilliamella* and *Snodgrassella* and a marked decrease in *Lactobacillus* in the cluster with reduced immunity. This shift was consistent across different pathogens, suggesting a universal stress response or impaired immune homeostasis.

Our results confirm that although the honeybee microbiome is generally stable, environmental pressures, particularly diet and pathogens, can disrupt this equilibrium. We also show patterns of changes in the abundance of the core microbiome composition and provide ecological evidence for the interaction of immune-stimulating symbionts in the honeybee gut. Our results are consistent with previous studies showing the response of the honeybee microbiome to pathogen challenge and highlight the need for routine microbiome monitoring in beekeeping.

1. Kwong W.K., Moran N.A. Gut microbial communities of social bees // Nat. Rev. Microbiol. Springer Science and Business Media LLC, 2016. Vol. 14, № 6. P. 374–384.
2. Engel P. et al. The Bee Microbiome: Impact on Bee Health and Model for Evolution and Ecology of Host-Microbe Interactions // mBio / ed. Hurst G.B., Collier R.J. 2016. Vol. 7, № 2. P. e02164-15.
3. Motta E.V.S., Moran N.A. The honeybee microbiota and its impact on health and disease // Nat. Rev. Microbiol. 2024. Vol. 22, № 3. P. 122–137.
4. Kešnerová L. et al. Disentangling metabolic functions of bacteria in the honey bee gut // PLoS Biol. 2017. Vol. 15, № 12. P. e2003467.
5. Kešnerová L. et al. Gut microbiota structure differs between honeybees in winter and summer: 3 // ISME J. Nature Publishing Group, 2020. Vol. 14, № 3. P. 801–814.
6. Tran L. et al. Gut microbiome metagenomic sequences of honey bees (*Apis mellifera*) exposed to crops // Microbiol. Resour. Announc. / ed. Stewart F.J. 2025. Vol. 14, № 3. P. e00731-24.
7. Callahan B.J. et al. DADA2: High resolution sample inference from Illumina amplicon data // Nat. Methods. 2016. Vol. 13, №

7. P. 581–583.
8. Sczyrba A. et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software // Nat. Methods. 2017. Vol. 14, № 11. P. 1063–1071.
9. Charrad M. et al. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set // J. Stat. Softw. 2014. Vol. 61. P. 1–36.
10. Gloor G.B. et al. Microbiome Datasets Are Compositional: And This Is Not Optional // Front. Microbiol. 2017. Vol. 8. P. 2224.
11. De Cáceres M. et al. indicpecies: Relationship Between Species and Groups of Sites. 2010. P. 1.8.0.
12. Lakhman A. et al. Bees Klebsiellosis: Key Aspects of Pathogenesis // Adv. Anim. Vet. Sci. 2021. Vol. 9, № 8.
13. Salas B. et al. *Morganella morganii* (Enterobacteriales: Enterobacteriaceae) Is a Lethal Pathogen of Mexican Fruit Fly (Diptera: Tephritidae) Larvae // Fla. Entomol. 2017. Vol. 100, № 4. P. 743–751.
14. Raymann K. et al. Pathogenicity of *Serratia marcescens* Strains in Honey Bees // mBio. American Society for Microbiology, 2018. Vol. 9, № 5. P. 10.1128/mbio.01649-18.
15. Severiche-Bueno D.F. et al. *Hafnia alvei* pneumonia: from bees to human beings // GERMS. 2021. Vol. 11, № 2. P. 306–309.
16. Emery O., Schmidt K., Engel P. Immune system stimulation by the gut symbiont *Frischella perrara* in the honey bee (*Apis mellifera*) // Mol. Ecol. 2017. Vol. 26, № 9. P. 2576–2590.

Анализ и интерпретация геномных данных

Ф.А. Колпаков

*Научно-технический университет «Сириус», Сириус, 354340,
Россия kolpakov.fa@talantiuspeh.ru*

Введение

Стремительное развитие методов секвенирования нового поколения (NGS) позволило сделать геномные исследования более доступными и широко используемыми в клинической практике. Геномные данные становятся ключевым инструментом в диагностике наследственных заболеваний, изучении

онкологических процессов, анализе микробиомов и применении прецизионной медицины. Однако обработка, аннотация и интерпретация геномных данных совместно с клиническими показателями по-прежнему представляют собой серьёзную проблему для врачей и исследователей.

Основные сложности включают:

- Фрагментированность решений. На сегодняшний день для анализа данных используются сотни специализированных инструментов, каждый из которых решает ограниченный круг задач: выравнивание прочтений, вызов вариантов (SNV, CNV), аннотация и интерпретация данных. Эти инструменты часто сложно интегрировать в единую систему, что приводит к необходимости использования множества программ, требующих ручной настройки и значительных временных затрат.
- Недостаток стандартизации. Интерпретация геномных данных часто зависит от используемых методов, версий инструментов и баз данных. Это создаёт проблемы с воспроизводимостью результатов. Стандарты, такие как рекомендации ACMG[1] по интерпретации SNV, существуют, но их внедрение в автоматизированные системы остается фрагментированным.
- Ограниченная доступность платформ. Существующие биоинформатические платформы, такие как Illumina BaseSpace, часто являются коммерческими и требуют значительных финансовых вложений. Это ограничивает доступ к качественным инструментам для научных учреждений с ограниченным бюджетом и небольших клинических лабораторий.
- Ограниченная поддержка совместной работы. Закрытые и архитектуры ряда платформ затрудняют участие сторонних исследователей в их развитии, что снижает скорость обновления методов анализа и их адаптацию к новым данным.
- Закрытые данные. Наличие сложностей с интерпретацией клинически значимых вариантов, в особенности по информации о частотах встречаемости вариантов в конкретной популяции.

Цели и задачи проекта

При поддержке гранта федеральной территории «Сириус» начат проект "Анализ и интерпретация геномных данных".

Научной целью проекта является разработка открытой модульной расширяемой биоинформатической платформы с набором готовых сценариев для анализа и интерпретации геномных данных и последующий анализ ряда геномных данных.

Образовательной целью проекта является создание магистратуры по клинической биоинформатике для подготовки врачей, способных интерпретировать геномные и метагеномные данные на современном уровне, что особенно важно в условиях роста персонализированной медицины.

Основные задачи проекта:

- Разработать открытую модульную расширяемую биоинформатическую платформу для анализа и интерпретации геномных данных.
- Разработать сценарии для анализа и интерпретации геномных данных. Создать открытое хранилище сценариев анализа геномных данных.
- Разработать набор модулей и правил в соответствии с мировыми и российскими рекомендациями для интерпретации SNV.
- Протестировать предложенные подходы на большом объеме геномных данных.
- Разработать программу и курсы магистратуры по клинической биоинформатике для подготовки врачей, способных интерпретировать геномные и метагеномные данные на современном уровне.

Принципы разработки

- Открытый код. Код биоинформатической платформы будет полностью открытым. В качестве стартовой точки будет использоваться код платформы BioUML, который выложен на:
 - GitHub - <https://github.com/Biosoft-ru/BioUML>;
 - GitVerse - <https://gitverse.ru/biouml/BioUML>.
- Хранилище готовых сценариев для анализа и интерпретации

геномных данных, подобно Dockstore[2]. Однако, в связи с заинтересованностью коммерческих компаний в участии в проекте, хранилище будет включать как открытые бесплатные так и коммерческие сценарии.

- Формирование сообщества для разработки и развития биоинформатической платформы и сценариев. Также будет сформирован рабочий комитет из заинтересованных научных организаций, клинических лабораторий, производителей и дистрибьюторов секвенаторов для формулирования требований, апробации и развития биоинформатической платформы и сценариев. Таким образом основные решения будут приниматься экспертами, представляющих разные группы заинтересованных пользователей.
- Итерационный процесс разработки. Будет определен 2-4 недельный цикл разработки с последующим показом и обсуждением рабочему комитету новых версий биоинформатической платформы и сценариев. Это позволит получать быструю обратную связь от пользователей платформы.
- Использование прототипа. В ходе конференции МССМВ-2025 планируется демонстрация первого прототипа биоинформатической платформы и хранилища сценариев анализа данных. Взаимодействие с заинтересованными коллегами будет способствовать формулированию требований дальнейшего развития и совершенствования платформы.

Стартовая страница проекта, где можно найти всю информацию (в разработке): <https://genespace.ru>

Результаты получены при финансовой поддержке проекта, реализуемого в рамках государственной программы федеральной территории «Сириус» «Научно-технологическое развитие федеральной территории «Сириус» (Соглашение №3-03 от 18.02.2025).

4. S. Richards et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and

Genomics and the Association for Molecular Pathology, *Genetics in medicine*, **17(5)**:405-424.

5. D. Yuen et al. (2021) The Dockstore: enhancing a community platform for sharing reproducible and accessible computational protocols, *Nucleic acids research*, **49(W1)**:W624-W632.

Анализ неаннотированных хроматин-ассоциированных некодирующих РНК

М.А. Компаниец¹, Г.К. Рябых^{1,2}, А.А. Миронов^{1,2}

¹ФББ МГУ им. М.В. Ломоносова
mariakompa@mail.ru

²ИОГен РАН

У эукариот значительная часть генома транскрибируется с образованием множества РНК, включая мРНК и различные некодирующие РНК (нкРНК), такие как микроРНК, малые ядерные и ядрышковые РНК, энхансерные РНК, длинные нкРНК и другие. Эти транскрипты функционируют не только в цитоплазме, но и в ядре, где участвуют в регуляции транскрипции, ремоделировании хроматина и поддержании его трехмерной структуры. Среди нкРНК большой интерес представляют хроматин-ассоциированные РНК. Они выполняют свои функции путем взаимодействия с определенными участками хроматина. Эти взаимодействия мы будем называть контактами.

В настоящее время существует множество экспериментальных протоколов, позволяющих определить сайты связывания нкРНК с хроматином, в том числе и методы, предоставляющие возможность получить полный РНК-хроматиновый интерактом (т.е. данные о всех РНК-ДНК взаимодействиях в клетке). С их помощью стало возможно изучение новых, неаннотированных хроматин-ассоциированных РНК без предварительной информации об их последовательности или белковых партнерах, опосредующих взаимодействие с хроматином. Данные таких экспериментов содержат довольно много контактов,

РНК-части которых не принадлежат ни одному известному гену. С целью анализа подобных взаимодействий в нашей группе был разработан унифицированный протокол *de novo* сборки потенциальных хроматин-ассоциированных РНК на основе этих РНК-частей. Собранные таким образом транскрипты получили название *ucaRNAs* (unannotated chromatin-associated RNAs, неаннотированные хроматин-ассоциированные РНК). Координаты *ucaRNA* представлены в базе данных RNA-Chrom [1]. Итак, целью настоящей работы является исследование возможных функций *ucaRNA* и их аннотация путем поиска ортологов в различных организмах, основанного на предположении о синтении, а также разработка процедуры статистической оценки этих РНК и критериев отбора значимых аннотаций.

Был произведен поиск ортологов *ucaRNA*, белок-кодирующих и длинных некодирующих генов человека и мыши в 10 позвоночных организмах при помощи пакета *ortho2align* [2]. Согласно полученным результатам, только 4% генов *ucaRNA* человека имеют хотя бы одного ортолога во всех организмах, 12% обладают ортологами в приматах. Для некоторых *ucaRNA* нашлись аннотированные ортологи (например, 1179_HELAS3, KANSL1L-AS1 и ARNILA – человеческие ортологи *ucaRNA* мыши; Mir410, Ndufa12-ps, gm12472 – мышинные ортологи человеческих *ucaRNA*). Помимо поиска ортологов была также проведена статистическая оценка сборки *ucaRNAs* человека и мыши. Для статистической оценки нами была предложена модель, основанная на предположении о равномерном распределении чтений, присвоенных *ucaRNA*, по всей длине не аннотированных участков генома. Основная цель данной оценки – выявление потенциально значимых хроматин-ассоциированных РНК, взаимодействие которых с хроматином нельзя объяснить случайным распределением чтений. По ее результатам можно утверждать, что качество сборки *ucaRNA* для мыши выше, чем для человека.

Следующим этапом исследования будет комплексный анализ найденных ортологов во всех организмах и более детальная оценка эволюционной консервативности *ucaRNAs*. Планируется также запустить протокол от других организмов, для которых есть данные РНК-хроматинового интерактома, помимо человека и мыши, и продолжить поиск ортологов других типов РНК,

представленных в базе данных RNA-Chrom. Окончательным этапом станет добавление информации про ортологов в базу данных RNA-Chrom.

Список литературы

1. G K Ryabykh, S V Kuznetsov, Y D Korostelev, A I Sigorskikh, A A Zharikova, A A Mironov, RNA-Chrom: a manually curated analytical database of RNA–chromatin interactome, Database, Volume 2023, 2023, baad025
2. Mylarshchikov, D.E., Mironov, A.A. (2022) ortho2align: a sensitive approach for searching for orthologues of novel lncRNAs. BMC Bioinformatics 23, 384

Модульная модель нейрональной активности, учитывающая динамику основных молекулярных мишеней противоэпилептических препаратов

П.Ю. Кондрахин, Ф.А. Колпаков

*Научно-технический университет «Сириус», Сириус, 354340,
Россия kondrakhin.py@gmail.com*

Аннотация

В работе представлена разработанная с использованием модульного подхода математическая модель нейрональной активности, описывающая динамику большинства основных молекулярных мишеней противоэпилептических препаратов и позволяющая проводить моделирование воздействия на них медикаментозных средств. Построение модели проводилось путем объединения и модификации ряда существующих моделей синаптической передачи, учитывающих клеточные процессы, наиболее важные в патологической составляющей эпилепсии. В частности, большое внимание уделено динамике ионных токов (Na^+ , K^+ , Ca^{2+}), рецепторов (АМПА, НМДА, ГАМК_A, ГАМК_B и mGluR) и нейромедиаторов (глутамат и ГАМК). Приведены примеры симуляции тормозного воздействия на синаптическую передачу. Результаты численных расчетов согласуются с

экспериментальными данными, представленными в литературе.

Введение

Несмотря на наличие более 30 противоэпилептических препаратов с разнообразными механизмами действия, примерно треть пациентов с эпилепсией не достигает контроля над приступами при медикаментозной терапии [1].

Цель работы заключается в построении математической модели нейрональной активности, описывающей механизмы функционирования нервных клеток с учетом динамики основных мишеней противоэпилептических препаратов.

Структура модели

Модульная структура модели представлена на Рисунке 1. Основу составила модель [2], которая была расширена на учет регуляции динамики ГАМК и глутамата астроцитарными транспортерами, как это описано в [3]. Также были учтены ГАМК_A-, ГАМК_B-, метаботропные глутаматные рецепторы (mGluR) [3, 4], и тормозное воздействие интернейрона [5].

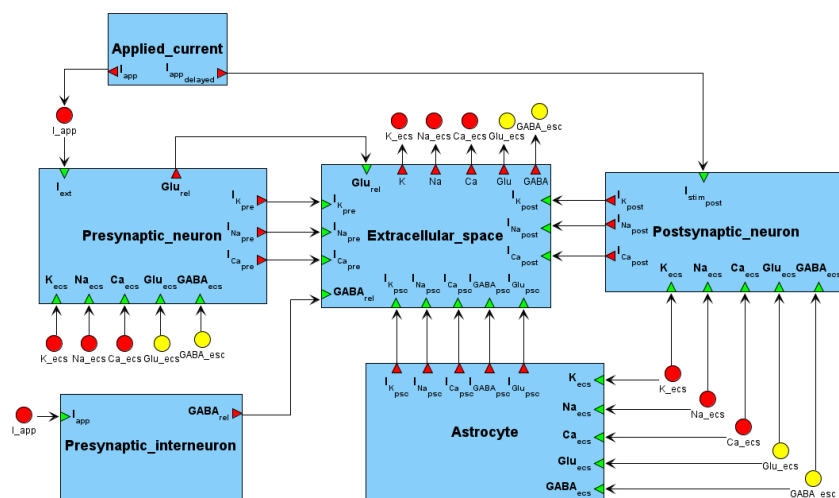


Рисунок 1. Модульная структура модели.

Модель симулирует передачу возбуждения от пресинаптического к постсинаптическому нейрону через

синаптическую щель, где динамика концентраций ионов и нейромедиаторов активно регулируется астроцитом через соответствующие каналы и транспортеры. Дополнительно в системе присутствует интернейрон, который способен оказывать тормозное воздействие на синаптическую передачу путем высвобождения ГАМК.

Модель содержит 247 переменных, 41 дифференциальное уравнение, 174 параметра и 235 операции присваивания.

Симуляция лекарственного воздействия

Одной из наиболее распространенных мишеней ПЭП являются потенциал-зависимые натриевые каналы [6]. Результаты симуляции в условиях ингибирования потенциал-зависимых натриевых каналов приведены на Рисунке 2.

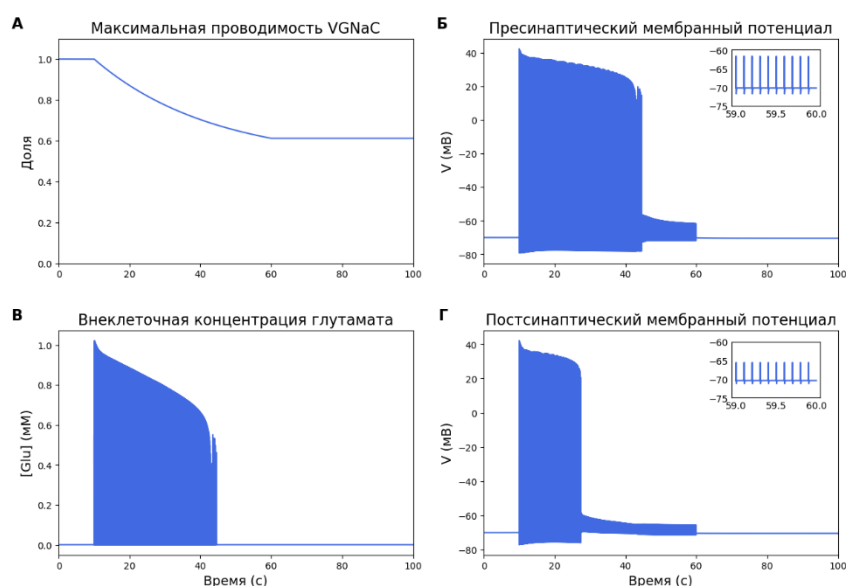


Рисунок 2. Результаты численного расчета модели в условиях ингибирования потенциал-зависимых натриевых каналов.

В случае снижения максимальной проводимости менее чем на 20% натриевые каналы по-прежнему обеспечивают достаточный натриевый ток для генерации потенциала действия. При снижении максимальной проводимости более чем на 20% пресинаптический

нейрон все еще может генерировать потенциалы действия под воздействием внешне приложенного стимула, но концентрации высвобождаемого им глутамата оказывается недостаточно для возбуждения постсинаптического нейрона. Снижение максимальной проводимости более чем на 30% ведет к полному торможению синаптической передачи.

Воспроизводимость результатов

Модульная модель размещена в открытом репозитории вместе с Jupyter-ноутбуком, позволяющим воспроизвести результаты ее расчета, по ссылке: <https://gitlab.sirius-web.org/brain/neuronal-activity-model-with-aed-targets>.

Финансирование

Исследование выполнено за счет гранта Российского научного фонда № 24-14-20031, <https://rscf.ru/project/24-14-20031/>.

Список литературы

1. Z. Chen et al. (2018) Treatment Outcomes in Patients With Newly Diagnosed Epilepsy Treated With Established and New Antiepileptic Drugs, *JAMA Neurol*, **75**:279.
2. M. Toman et al. (2023) The influence of astrocytic leaflet motility on ionic signalling and homeostasis at active synapses, *Sci Rep*, **13**.
3. B. Flanagan et al. (2021) A Computational Study of Astrocytic GABA Release at the Glutamatergic Synapse: EAAT-2 and GAT-3 Coupled Dynamics, *Front. Cell. Neurosci.*, **15**.
4. L. Li et al. (2020) A Computational Model to Investigate GABA-Activated Astrocyte Modulation of Neuronal Excitation, *Computational and Mathematical Methods in Medicine*, **2020**:1–13.
5. M. Borjkhani et al. (2018) Computational modeling of opioid-induced synaptic plasticity in hippocampus, *PLoS ONE*, **13**:e0193410.
6. M.A. Rogawski, W. Löscher (2004) The neurobiology of antiepileptic drugs, *Nat Rev Neurosci*, **5**:553–564.

Reference methylome and episignature-based analysis for Sotos syndrome diagnostics

M.I. Komkova¹, N.V. Zernov¹, I.V. Volodin², D.V. Pustoshilov¹

¹*“Biotech Campus” LLC, Moscow, Russia,
komkova8@mail.ru*

²*“Bochkov Research Centre for Medical Genetics” FSBSI, Moscow,
Russia*

Sotos syndrome (SS) is a genetic overgrowth disorder characterized by distinctive facial appearance, learning disability and overgrowth in childhood. It is primarily caused by mutations in the NSD1 gene, encoding a histone methyltransferase. This disease belongs to chromatinopathies – diseases that develop as a result of pathogenic variants in genes, the products of which are members of epigenetic mechanisms.

Conventional genetic testing has limitations, including the fact that the detectability of pathogenic changes in the NSD1 gene varies from 35% to 90% depending on the method of research and the sample of patients studied [1] and difficulties in interpreting variants of uncertain significance (VUS). DNA methylation profiling may offer a solution to both aforementioned problems by identifying disease-specific epigenetic signatures [2, 3]. In addition, episignature could work not only for the disease announced but also for disease with opposite cause of appearing. For example, gene deletion and gene duplication are opposite effects, so their episignature is anticollinear and difunctional [4]. Another use of episignatures is to identify patients with mosaicism - people who have different genetic sets in different cells of the body [4].

We constructed a pioneering reference methylome using long-read sequencing data from 900 healthy individuals. The developed reference methylome was designed to facilitate reliable detection of DNA methylation signatures associated with SS, enabling differentiation from healthy controls in similar conditions.

To build the reference methylome, we processed and combined the data of 900 methylation levels for each of approximately 30 million CpGs to generate per-base methylation levels across the human genome. Confidence intervals and interquartile ranges for each CpG site were calculated to establish baseline methylation variability in healthy individuals. Additionally, problematic sequencing regions (blacklist regions) were excluded from analysis to avoid false-positive results.

We applied this refined reference methylome to analyze blood samples from individuals clinically suspected to have Sotos syndrome. Patient methylation profiles were compared against the established normal ranges. Significant deviations from normal methylation levels in regions previously reported as characteristic of Sotos syndrome epismature were identified based on both statistical significance (95% confidence interval) and biological relevance (fold-change thresholds), as defined by Equations (1) and (2), respectively.

$$\text{Fold Change} = \log_2 \frac{\text{Mvalue}_{\text{patient}}}{\text{median Mvalue}_{\text{healthy}}} \quad 1$$

$$\text{CI}_{95\text{healthy}} = \text{quantile}_{0.975} ; \text{quantile}_{0.025} \quad 2$$

Figure 1 illustrates the results of this comparative analysis for a representative patient.

As a result, we computed the quality statistics for the reference methylome, such as analyzing the required minimum coverage, total sample size; we identified biological and instrumental complex regions and performed comparative genomic analysis to determine the diagnosis of patients. Our diagnosis coincided with the clinical diagnosis in 86% of cases.

The authors would like to express their gratitude to the staff of the biomedical department and the Laboratory of Biotechnological Campus, LLC for organizing the collection of samples and the experimental part of the work. In addition to that, the authors would like to thank the organizers of the 12th Moscow Conference on Computational Molecular Biology (MCCMB'25).

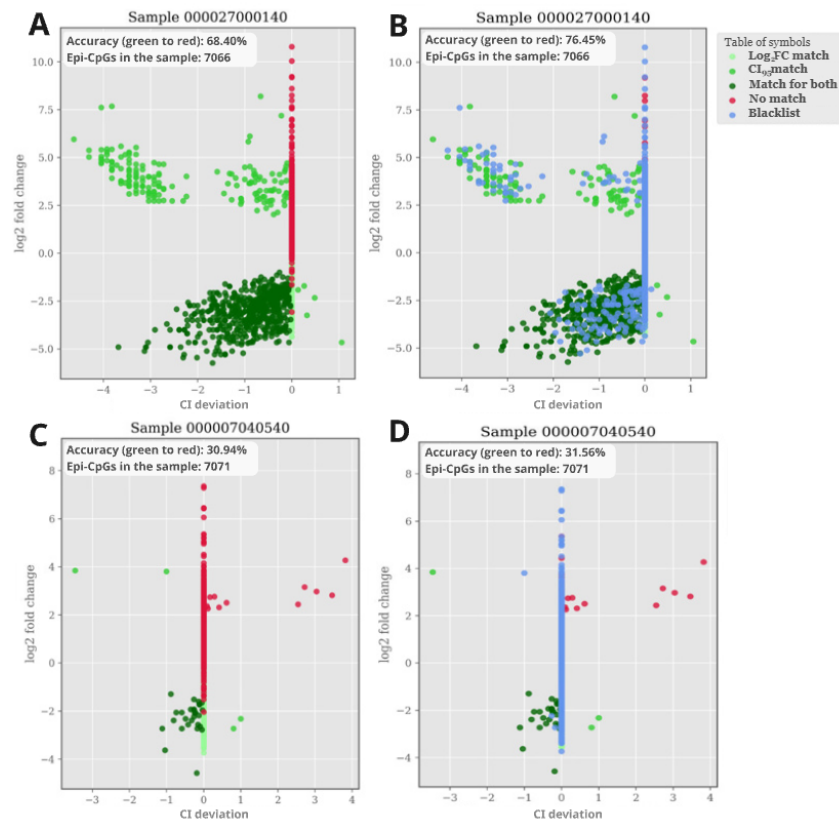


Figure 1. Comparative methylation signature analysis for a patient sample (ID 000027000140) and for a healthy individual (ID 000007040540):

A: Initial analysis for a patient (accuracy 68.40%);
 B: Analysis after filtering blacklist regions for a patient (accuracy improved to 76.45%, an increase of 8%);
 C: Initial analysis for a healthy individual (accuracy 30.94%);
 D: Analysis after filtering blacklist regions for a patient (accuracy improved to 31.56%, an increase of 0.6%)

1. Volodin I.V. et al. (2021) Spectrum of detected genetic variants in NSD1 gene in Patients with Sotos syndrome, Medical

- Genetics, 11:11
2. Choufani S. et al. (2015) NSD1 mutations generate a genome-wide DNA methylation signature, *Nature Communications*, 6:10207.
 3. M.A. Levy et al. (2022) DNA methylation signatures for chromatinopathies: current challenges and future applications, *Clinical Epigenetics*, 14:23.
 4. Treppendahl M.B. et al. (2022) Genome-wide DNA methylation signatures predict outcome of patients with MDS treated with hypomethylating agents, *Leukemia*, 36:1233–1243.

GQ-DNABERT демонстрирует взаимодействия энхансерного и промоторного квадруплекса в открытом хроматине

Коновалов Д.Л.¹, Умеренков Д.Е.², Попцова М.С.¹, Герберт А.³

¹*НИИ ВШЭ*
dkonovalov@hse.ru

²*Sber*

³*InsideOutBio*

Альтернативные конформации ДНК, такие как гуаниновые квадруплексы (GQ), как полагают, играют важную роль в регуляции различных геномных процессов: регуляции транскрипции, организации хроматина, репарации ДНК, поддержание теломер, сплайсинге, трансляции, реакции на стресс и ряде других процессов. Образование GQ исследовалось многими экспериментальными методами, такими как ChIP-seq, CUT&Tag, G4-seq. При этом результаты могут заметно отличаться, в ряде случаев на 2 порядка. Модели основанные на простых закономерностях или на одном эксперименте не могут полностью охватить динамический ландшафт структур. Поэтому существует необходимость в моделях основанных на большой совокупности наборов экспериментальных данных, способной в полной мере описать возможные сайты образования GQ..

В данной работе мы обучили большую языковую модель

DNABERT[1], используя данные базы эндогенных квадруплексов EndoQuad[2] основанной на более 1000 наборов экспериментальных данных. В зависимости от числа экспериментов, подтверждающих образование структуры, квадруплексам в этой базе присвоены уровни поддержки. Наиболее достоверными являются предсказания с высоким уровнем поддержки, поэтому для уменьшения числа ошибок, обучение производилось только структурах с уровнем поддержки 4-6.

Мы изучили распределение предсказанных структур в геноме и обнаружили *de novo* квадруплексы в межгенных и интронных областях. При этом наблюдалось, обогащение участков потенциальными цис-регуляторными элементами (cCRE) и пиками ATAC-seq.

Для оценки потенциальной роли квадруплексов в энхансер-промоторных взаимодействиях мы определили пары энхансер-промотор (EP) с предсказаниями GQ-DNABERT, используя аннотации энхансеров из открытых баз данных. В результате были получены списки пар EP, которые потенциально могут регулироваться с участием пары GQ. Для генов, соответствующих данным промоторам был проведен анализ GO.

Для проверки связи между квадруплексами в промоторах и энхансерах мы также изучили шесть наборов данных секвенирования отдельных клеток 10x Genomics Chromium из клеток и sCell G4 CUT&Tag из опухолевых линий, обнаружив, что пары энхансер-промотор с предсказанным G4 положительно коррелируют с экспрессией генов, независимо от ориентации цепи, при этом некоторые пары потенциально действуют как тканеспецифичные переключатели.

1. Ji, Y., Zhou, Z., Liu, H. and Davuluri, R.V. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37, 2112-2120.
2. Qian, S.H., Shi, M.W., Xiong, Y.L., Zhang, Y., Zhang, Z.H., Song, X.M., Deng, X.Y. and Chen, Z.X. (2024) EndoQuad: a comprehensive genome-wide experimentally validated endogenous G-quadruplex database. *Nucleic Acids Res*, 52, D72-D80.

Характеристика ядрышкового транскриптома клеток линии HeLa

Е. В. Коновалова¹, А.А. Валяева², Е.В. Шеваль²

¹*Факультет биоинженерии и биоинформатики, МГУ имени М.В. Ломоносова, Ленинские горы, д. 1, стр. 73, 119234 Москва, Россия, konzhenya@mail.ru*

²*НИИ ФХБ имени А.Н. Белозерского, Ленинские горы, д. 1, стр. 40, 119234 Москва, Россия*

Ядрышко – немембранный компартмент клеточного ядра, главными функциями которого является синтез и процессинг рибосомных РНК (рРНК) и сборка прерибосом. Ядрышко формируется ядрышковыми организаторами – участками хромосом, содержащими кластеры рРНК генов [1, 2]. Помимо своей основной функции, ядрышко играет важную роль в поддержании архитектуры генома и его стабильности, а также участвует в регуляции многих клеточных процессов, таких как, дифференцировка и ответ на стрессы различной природы [1, 3]. Ядрышковый транскриптом представлен не только пре-рРНК, но и другими разнообразными РНК [4, 5]. Некоторые из них транскрибируются внутри ядрышка на участках генома, разделяющих повторы рРНК генов [6], другие – за пределами ядрышка и впоследствии накапливаются в нем [5]. Такие РНК могут играть роль в поддержании структуры ядрышка или участвовать в регуляции процесса биогенеза пре-рибосом [7]. Поскольку на данный момент закономерности и причины ядрышковой локализации РНК еще не выявлены, изучение его транскриптома представляет большой интерес. Целью данной работы является характеристика транскриптома ядрышка клеток культуры HeLa на основе данных РНК-секвенирования.

Для анализа была изолирована РНК из фракции ядрышек и тотальная РНК клеток линии HeLa. После рРНК деплеции было проведено секвенирование РНК по протоколу Illumina TruSeq с получением одноконцевых цепь-специфических прочтений. Для каждой фракции было отсеквенировано 2 биологические реплики. В качестве референсного генома, на который проводилось

картирование, был выбран T2T-CHM13v2.0, что позволило проанализировать экспрессию с повторяющихся участков генома, в частности, с кластеров рРНК генов.

В данной работе мы оценили экспрессии транскриптов генов согласно аннотации RefSeq, новых транскриптов, предсказанных при помощи StringTie, а также транскриптов различных семейств мобильных элементов. При этом характер картирования на референсный геном позволил предположить совыделение геномной ДНК в ядрышковых образцах, которое, скорее всего, была обусловлено особенностями их пробоподготовки. В связи с этим была проведена коррекция оценки экспрессии анализируемых локусов генома, чтобы компенсировать возможный вклад от геномной ДНК [8].

Анализ дифференциальной локализации РНК позволил выделить группу РНК, накапливающихся в ядрышке, или ядрышко-ассоциированных РНК (яАРНК), среди которых были широко представлены длинные некодирующие РНК. Также в ядрышковом транскриптом были детектированы новые, еще не аннотированные транскрипты и транскрипты мобильных элементов, в особенности сателлитных повторов. При этом кодирующие яАРНК обогащены транскриптами генов, составляющих семейства тканеспецифичных рецепторов и транскрипционных факторов.

Анализ взаимоположения генов яАРНК и ассоциированных с ядрышками доменов генома свидетельствует о том, что подавляющее большинство яАРНК синтезируется вдали от границ ядрышка. Согласно одностороннему тесту Фишера была выявлена ассоциация между некодирующими генами яАРНК и ядрышко-ассоциированными доменами, однако это верно лишь для небольшого числа генов. При этом гены яАРНК ассоциированы с неактивными состояниями хроматина согласно разметке ChromHMM.

Гены яАРНК являются более GC-богатыми в сравнении с остальными генами: это особенно характерно для белок-кодирующих генов и в меньшей степени для некодирующих. При этом некодирующие гены яАРНК обогащены вставками мобильных элементов, а белок-кодирующие, наоборот, обеднены. Таким образом, РНК, составляющие ядрышковый транскриптом, вероятнее всего, локализуются в ядрышке благодаря своим

свойствам. яРНК преимущественно происходят из неактивных регионов генома, которые, в некоторых случаях, составляют непосредственно околядрышковый хроматин. При этом гены яРНК обладают особенностями с точки зрения своей последовательности, которые дифференцируют их от остальных РНК клетки, а также белок-кодирующие и некодирующие яРНК между собой.

Работа выполнена при поддержке Российского научного фонда (проект 24-24-00168).

1. Bersaglieri C, Santoro R. (2019). Genome Organization in and around the Nucleolus. *Cells*, 8:579.
2. Henras AK, Plisson-Chastang C, O'Donohue M, Chakraborty A, Gleizes P. (2015). An overview of pre-ribosomal RNA processing in eukaryotes. *WIREs RNA*, 6:225–42.
3. Lee TA, Han H, Polash A, Cho SK, Lee JW, Ra EA, et al. (2022). The nucleolus is the site for inflammatory RNA decay during infection. *Nat Commun*, 13:5203.
4. Pederson T. (1998). The plurifunctional nucleolus. *Nucleic Acids Res.*, 26:3871–6.
5. Böğürçü-Seidel N, Ritschel N, Acker T, Németh A. (2023). Beyond ribosome biogenesis: noncoding nucleolar RNAs in physiology and tumor biology. *Nucleus*, 14:2274655.
6. Feng S, Manley JL. (2022). Beyond rRNA: nucleolar transcription generates a complex network of RNAs with multiple roles in maintaining cellular homeostasis. *Genes Dev.*, 36:876–86.
7. Sun Y-M, Zhu S-X, Chen X-T, Pan Q, An Y, Chen T-Q, et al. (2024). lncRNAs maintain the functional phase state of nucleolar prion-like protein to facilitate rRNA processing. *Mol Cell*, 84:4878-4895.e10.
8. Li X, Zhang P, Wang H, Yu Y. (2022). Genes expressed at low levels raise false discovery rates in RNA samples contaminated with genomic DNA. *BMC Genomics*, 23:554.

Анализ кросс-контаминации прочтений в данных DNBSEQ секвенирования

Д. Н. Конанов¹, В. Ю. Терещук², И. В. Сонец¹, Е. В. Корнеев¹,
А. В. Лукина-Гронская¹, А. С. Сперанская³, Е. Н. Ильина¹

¹*НИИ системной биологии и медицины, Москва,
konanovdmiriy@gmail.com*

²*Московский физико-технический институт, Долгопрудный*

³*Институт общей генетики им. Н.И.Вавилова, Москва,*

Сегодня технология секвенирования DNBSEQ рассматривается как одна из альтернатив Illumina. MGI DNBSEQ обеспечивает высокопроизводительное секвенирование на базе коротких ридов с разнообразной конфигурацией длин, от SE50 до SE400 и PE300. Вариант PE300 главным образом рекомендован MGI для секвенирования сравнительно длинных ампликонов (16S/18S/ITS). В запуске PE300, с которого началась представленная работа, было совмещено пять разных типов библиотек, включая ампликоны 16S, полные геномы индивидуальных бактерий, WGS-метагеномы и РНКомы. В процессе анализа данных мы обнаружили кросс-контаминацию секвенируемых объектов, которая проявлялась даже на уровне геномных сборок в виде контигов, принадлежащих нецелевым организмам. Было обнаружено, что парные прочтения, порождающие эти контиги, в большинстве некорректно соотнесены, то есть прямое и обратное прочтения из пары несли разные вставки. В данной работе нашей целью стало детальное рассмотрение подобных артефактов, возникающих при анализе данных DNBSEQ.

Необходимо отметить, что, когда длина вставки менее чем 260 нуклеотидов, прочтения PE300 позволяют прочитать последовательность баркода внутри прочтения. Пользуясь этим, мы выяснили, что контаминирующие риды, как правило, несут нецелевой баркод, то есть были неверно демультиплексированы прибором. При этом прямые прочтения были неверно демультиплексированы в 2.54% случаев, а обратные в 0.56%.

Эффект неверного соотнесения пар был воспроизведен также на внешних демо-данных MGI (PE100 и PE150). Было установлено, что короткие вставки значительно более подвержены как некорректному соотнесению пар прочтений, так и некорректному демультиплексированию.

Мы заметили, что описываемые артефактные прочтения зачастую имеют риды-дубликаты с идентификатором рида, отличным на единицу. В данных DNBSEQ номер рида определяется позицией ДНК-наноболлов (DNB) в ячейке, т.е. наблюдаемые дубликаты относятся к оптическим дубликатам и вероятнее всего являются следствием некорректного разрешения сигнала с соседних сайтов посадки DNB. Было установлено, что и некорректное соотнесение, и некорректное демультиплексирование пар прочтений сильно ассоциировано с наличием у них оптических дубликатов. Полная доля оптических дубликатов в запуске PE300 составила 2.516%, в демо-датасете MGI - 1.65%.

Дискордантные пары прочтений характерны и для результатов Illumina, и необходимо отметить, что данные DNBSEQ-секвенирования сопоставимы с ними по частоте подобных событий, что также было показано на внешних данных. Тем не менее, доля оптических дубликатов в данных остается весьма высокой, поэтому дедупликация данных выглядит необходимым шагом при обработке данных. Однако существующие инструменты дедупликации ориентированы на данные секвенирования Illumina и несовместимы со структурой FASTQ-файлов, возвращаемых приборами MGI DNBSEQ. Так, например Picard MarkDuplicates, который является наиболее распространенным методом оценки числа дубликатов в данных Illumina, для данных MGI возвращает 0% оптических дубликатов, при их фактическом количестве около 2%, что подчеркивает необходимость разработки методов дедупликации данных, совместимых с данными DNBSEQ.

Работа выполнена при поддержке гранта РНФ № 24-15-00419.

OpenPtmFinder: аннотация и визуализация пост-трансляционных модификаций в ТМТ-меченных данных масс-спектрометрии методом открытого поиска

А.С. Копейкина, И.А. Тарасова

ИНЭПХФ им. В.Л. Тальрозе ФИЦ ХФ РАН, koreykina.00@bk.ru

Полнопротеомный поиск пост-трансляционных модификаций (ПТМ) до сих пор остаётся сложной и не до конца решённой задачей как с методологической, так и с концептуальной точек зрения. Это обусловлено, во-первых, вариабельностью и лабильностью модификаций, особенно выраженных при использовании жёстких методов выделения белка, и, во-вторых, химическим разнообразием и структурными особенностями многочисленных ПТМ, одновременно находящихся в динамике в клетке (по оценкам на 2021 год, было описано более 400 типов модификаций [1]). Кажется, что охватить всё это разнообразие с помощью одного стандартного протокола анализа представляется невозможным. Однако современные омиксные протеомные технологии, основанные на масс-спектрометрии (МС), наиболее приблизились к решению этой задачи.

Наивысшим разрешением и производительностью при анализе ПТМ в настоящее время обладает подход «снизу вверх» (также известный как «протеомика дробовика»), который предполагает использование протеаз и анализ на уровне пептидов с привлечением базы данных белковых последовательностей. При этом МС-спектры первого и второго уровней сопоставляются с теоретическими спектрами из базы данных в пределах заданной ошибки m/z , описываемой как масс-окно. Для точной идентификации пептида стандартная МС-протеомика использует узкое окно (в среднем ± 0.03 Da). Во многих случаях это окно не охватывает PSM (peptide spectrum match) с массовыми сдвигами, вызванными химическими модификациями при пробоподготовке, процессами ионизации в масс-спектрометре или биологическими вариациями (например, аминокислотными заменами или ПТМ). Подход открытого поиска предполагает значительное расширение масс-окна (до ± 500 Da), что позволяет учитывать пептиды с любыми

массовыми сдвигами. Это, в свою очередь, расширяет пространство поиска и увеличивает сложность интерпретации результатов, что требует разработки специализированных биоинформатических инструментов.

В настоящей работе мы использовали биоинформатический конвейер, включающий поисковую машину MSFragger [2], программу постпоисковой обработки данных открытого поиска AA_stat [3] и новый разработанный нами инструмент OpenPtmFinder для аннотации, визуализации и статистического анализа ПТМ. Для увеличения покрытия протеома и повышения частоты обнаружения различных типов ПТМ были выбраны МС-данные с изобарным мечением (TMT) и глубоким фракционированием из открытого репозитория мультиомиксных данных Synapse [4]. Всего было проанализировано 190 образцов тканей височной доли головного мозга (post mortem) от здоровых доноров и пациентов с диагностированной болезнью Альцгеймера. С помощью программы OpenPtmFinder были интерпретированы массовые сдвиги, что позволило обнаружить 37 955 пептидов с различными модификациями. Наиболее часто встречающимися модификациями оказались: дезамидирование, окисление, фосфорилирование, дегидратация, образование хинона и метилирование. Статистический анализ выявил 125 сайтов с достоверными различиями в представленности модификаций между контрольной и больной группами. Мы предполагаем, что такие изменения ландшафта ПТМ могут отражать патологические процессы, вовлечённые в нейродегенерацию. Кроме того, ряд выявленных модификаций может обладать стабильной представленностью в популяции, что делает их потенциальными биомаркерами заболевания. Например, недавнее исследование показало, что фосфорилирование тау-белка в позициях серина 262 и 356 коррелирует с уровнем нейрофибриллярных клубков и степенью когнитивных нарушений у пациентов с болезнью Альцгеймера, что делает эти модификации отличными диагностическими маркерами [5].

В заключение, по нашим данным, программа OpenPtmFinder не имеет прямых аналогов для интерпретации и статистического анализа ПТМ. Сравнение с литературными источниками подтвердило воспроизводимость результатов и позволило выявить

новых кандидатов, потенциально вовлечённых в патогенез нейродегенерации. Всё это делает OpenPtmFinder эффективным инструментом для рутинных МС-экспериментов.

Работа была выполнена при поддержке Российского научного фонда, грант № 23-45-00012.

1. S. Ramazi, J. Zahiri (2021) Post-translational modifications in proteins: resources, tools and prediction methods, Database, 2021.
2. A. Kong et al. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics, Nat Methods, 14:513–520.
3. L.I. Levitsky et al. (2021) AA_stat: Intelligent profiling of in vivo and in vitro modifications from open search results, Journal of Proteomics, 248.
4. Synapse. (n.d.). <https://www.synapse.org>. (Accessed: 10.05.2025).
5. T. Islam et al. (2025) Phospho-tau serine-262 and serine-356 as biomarkers of pre-tangle soluble tau assemblies in Alzheimer's disease, Nat Med, 31:574–588.

Сравнение инструментов обнаружения структурных вариантов на данных полногеномного секвенирования с короткими чтениями

А.В. Косарева¹, А.Н. Мешков²

¹Москва; Научный исследовательский университет “Высшая Школа Экономики”, факультет биологии и биотехнологии
avkosareva_1@edu.hse.ru

²Москва; «Национальный медицинский исследовательский центр терапии и профилактической медицины» Министерства здравоохранения Российской Федерации

В отличие от небольших мутаций, структурные варианты труднее поддаются детекции на данных секвенирования с

короткими чтениями, поскольку их размеры могут значительно превышать длину фрагментов секвенирования. Отсутствие стандартизированных валидационных наборов данных дополнительно осложняет оценку эффективности методов детекции структурных вариантов. Современные технологии секвенирования с длинными чтениями позволяют более точно идентифицировать структурные варианты благодаря снижению вероятности ошибочного картирования на референсный геном на этапе анализа и точному определению точек разрыва за счет полного охвата участков структурных перестроек.

В данном исследовании для проверки инструментов поиска структурных вариантов на данных с короткими чтениями в качестве истинных валидационных вариантов использовались результаты поиска на данных с длинными чтениями, полученные с помощью инструмента Sniffles2. В качестве сравниваемых инструментов были использованы DELLY, GridSS, Manta, LUMPY и svABA. Исследование выполнено на восьми образцах, инструменты показывают следующие метрики:

1. GridSS: Полнота 0,13, Точность 0,36, F1-мера 0,19
2. DELLY: Полнота 0,12, Точность 0,29, F1-мера 0,17
3. LUMPY: Полнота 0,12, Точность 0,17, F1-мера 0,14
4. Manta: Полнота 0,22, Точность 0,49, F1-мера 0,28
5. svABA: Полнота 0,19, Точность 0,40, F1-мера 0,26

Общее количество детектируемых структурных вариантов при использовании данных с длинными чтениями существенно превосходит количество вариантов, обнаруженных при поиске на данных секвенирования с короткими чтениями для всех протестированных инструментов. Анализ выявил, что инструменты для анализа секвенирования с короткими чтениями чаще обнаруживают структурные варианты длиной более 10 000 п.н., многие из которых могут быть артефактами. Для дальнейшего применения на тестовой выборке образцов предложен инструмент Manta, который показал наилучшие значения метрик качества и продемонстрировал наиболее оптимальное время анализа на образце.

Hmardic: поиск значимых РНК-ДНК взаимодействий

М.Н. Косимов, А.А. Жарикова, А.А. Миронов

*Московский государственный университет им. М.В. Ломоносова,
факультет биоинженерии и биоинформатики, Ленинские горы, 1,
стр. 73, Москва, 119992, Россия
mihamzik00@gmail.com*

За последние годы активно разрабатываются экспериментальные протоколы для анализа РНК-ДНК интерактомов, в результате которых мы получаем информацию о взаимодействии РНК с разными участками хроматина. Эти методы открывают возможность исследовать механизмы регуляции генома с высокой точностью. Все эти методы можно условно разделить на два класса: «один-против-всех», где изучается взаимодействие конкретной РНК с хроматином, и «все-против-всех», которые позволяют исследовать полную картину взаимодействий между всеми РНК и геномной ДНК.

Однако работа с такого рода данными сопряжена с рядом трудностей. В частности, нужно учитывать особенности экспериментальных протоколов, фоновый шум, возникающий из-за неспецифичных контактов, неоднородность хроматина, различия между взаимодействиями с цис-хромосомой (хромосома гена-источника РНК) и транс-хромосомами (все остальные), а также эффект скейлинга — снижение вероятности контакта по мере удаления от гена, в котором экспрессируется РНК. Эти факторы значительно усложняют интерпретацию результатов и требуют применения специальных алгоритмических решений.

Для решения этой задачи была разработана программа Hmardic, предназначенная для поиска специфических контактов — так называемых «пиков» — между РНК и ДНК. Метод основывается на использовании скрытых марковских моделей (Hidden Markov Models, HMM), которые позволяют классифицировать геномные области, присваивая им статус «пик» или «фон». Модель учитывает ключевые источники искажений в данных: гетерогенность хроматина, различия в вероятностях цис- и транс-контактов, а также скейлинг.

Hmardic адаптирован для анализа как «один-против-всех», так и «все-против-всех» данных. Для оценки его эффективности был проведён сравнительный анализ с существующим инструментом BaRDIC [1]. Также Hmardic был применен к данным «все-против-всех», и позволил выявить новые, ранее не описанные паттерны взаимодействий длинных некодирующих РНК с хроматином.

Работа финансировалась грантом РФФИ №23-14-00136.

1. D.E. Mylarshchikov, A.I. Nikolskaya, O.D. Bogomaz, A.A. Zharikova, A.A. Mironov (2024) BaRDIC: robust peak calling for RNA–DNA interaction data, NAR Genomics and Bioinformatics, 6(2):lqae054.

Последствия редактирования гена *galnt17* у домашнего кролика (*Oryctolagus cuniculus*)

Г.Ю. Косовский, О.И. Скобель, Д.Э. Высоцкий, Д.В. Попов

ФГБНУ НИИ пушного звероводства и кролиководства имени В.А. Афанасьева, 140143 Россия, Московская обл., Раменский р-н, пос. Родники, ул. Трудовая, 6, niipzk@mail.ru

Galnt17 принадлежит семейству генов N-ацетилгалактозаминилтрансфераз, кодируемый белок выполняет функцию шаперона других представителей этого семейства в процессе гликозилирования, локализован на фланге комплекса, включающего 25-28 генов, большинство из которых кодирует факторы регуляции транскрипции. Делеция/дупликация комплекса лежит в основе гиперсоциальности у человека (синдром Вильямса-Берена (СВБ)), домашней собаки, ряда видов грызунов. Эволюционно консервативная петля хроматина из трех генов - *Auts2*, *Galnt17* и *Caln1* лежит на фланге этого комплекса [1]. Последствия редактирования гена *galnt17* у кроликов представляют особый интерес, поскольку указанная петля из трех генов, локализованная на хромосоме 6 у домашнего кролика, в отличие от других видов млекопитающих, утрачивает генетическое сцепление с другими генами блока СВБ, находящимися, по-видимому, на 7-й или 11-й хромосомах кролика, судя по результатам гибридизации

их фрагментов с хромосомой 7 человека [2].

По литературным данным, редактирование *Galnt17* у лабораторных линий мышей приводит к высокой летальности в зависимости от генотипической среды: дает потомство только от скрещивания C57BL/6JxC3H/He, но не от «чистых» линий [3].

В наших исследованиях, выполненных совместно с ИБГ РАН, с использованием технологии CRISPR/Cas9 и оптимизации метода эмбриотрансплантации у кролика выполнена вставка 1 нуклеотида во второй экзон гена *galnt17*, приводящая к смещению рамки считывания у домашнего кролика. В конечном итоге получен половозрелый самец с нокаутом *galnt17*. При его спаривании с самками дикого типа получено поколение F1 (23 потомка). У животных проявлялись различные фенотипические особенности, характерные для людей и других млекопитающих с СВБ, к которым относятся такие, как деформация лицевого скелета, в частности, за счет ассиметричной гипоплазии верхнечелюстных пазух, легкий симметричный двусторонний экзофтальм, уменьшение размеров мозжечка и обедненность сосудистой сети паутинной оболочки, выявляемое при патоморфологическом исследовании. У отдельных животных отмечалось изменение цвета окраса с серо-голубого на черный. Кроме того, у всех полученных животных обнаружены различные поведенческие особенности, нетипичные для кроликов – частые вертикализация (вставание на задние конечности), вокализация (попытки издавать звуки), повышенная стресс-реактивность.

Полученные данные позволяют предполагать, что у домашнего кролика, несмотря на разрушение генетического сцепления петли *auts2*, *galnt17* и *caln1* с комплексом генов СВБ, сохраняются элементы топологических взаимодействий между ними, способствующие проявлению нетипичных анатомо-морфологических и поведенческих характеристик у животных с отредактированным геном *galnt17*.

Работа выполнена с использованием УНУ «Коллекция линий генно-редактированных кроликов» ФГБНУ НИИПЗК.

1. O. Engmann et al. (2017) Cocaine-Induced Chromatin Modifications Associate With Increased Expression and Three-Dimensional Looping of *Auts2*. Biol Psychiatry. Dec 1;82(11):794-805.

2. R. Korstanje et al. (1999) Complete homology maps of the rabbit (*Oryctolagus cuniculus*) and human by reciprocal chromosome painting. *Cytogenet. Cell. Genet.*, 86(3-4): 317-322
3. C.Y. Chen et al. (2022) Galnt17 loss-of-function leads to developmental delay and abnormal coordination, activity, and social interactions with cerebellar vermis pathology. *Dev Biol.* Oct;490:155-171.

**Реконструкция контекст-зависимых потоковых моделей для
клеточных линий NCI-60 с использованием транскриптомных
и протеомных данных**

А. А. Котельников, М.А. Куляшов, Ф.А. Колпаков

*Научно-технологический университет «Сириус», 354349,
Российская Федерация, Краснодарский край, федеральная
территория «Сириус», Олимпийский проспект, д.1, artail@mail.ru*

Введение

Панель NCI-60 представляет собой репрезентативную коллекцию из 60 человеческих опухолевых клеточных линий. Она используется для фенотипирования опухолей, тестирования противоопухолевых агентов и изучения молекулярных особенностей опухолевых клеток, включая их метаболическую активность. Метаболическое перепрограммирование признано одним из ключевых признаков опухолевого фенотипа. Эффективным инструментом для изучения клеточного метаболизма является потоковое моделирование. В сочетании с анализом баланса потоков (FBA) это позволяет моделировать метаболизм опухолевых клеток и оценивать их адаптацию к условиям микроокружения. Увеличение прогностической мощности моделей может достигаться за счёт использования омиксных данных. Одним из недавно разработанных алгоритмов контекстной спецификации на основании транскриптомных данных является RIPTiDe [1]. В отличие от подходов, основанных на

ферментативных ограничениях [2], требующих детальных данных о кинетике ферментов, RIPTiDe использует ранжирование транскриптомных данных, что делает метод более применимым при ограниченной доступности кинетических параметров, часто недоступных для раковых клеток. Возможность использования протеомных данных обосновано их прямой связью с активностью ферментов, что критично для метаболического моделирования. В отличие от транскриптома, уровни белков учитывают посттрансляционные модификации и регуляцию, что может повысить точность прогнозов. Основной целью данной работы является сравнение результатов реконструкции контекст-зависимых потоковых моделей с использованием транскриптомных и протеомных данных на примере раковых клеточных линий панели NCI-60.

Материалы и методы

В качестве исходной модели использовалась Recon3D [3] — это одна из ключевых потоковых моделей человека, содержащая 10600 реакций, 5835 метаболитов и 2248 генов. Для адаптации модели к условиям роста раковых клеток была определена "минимальная среда" из 50 метаболитов, общих для всех линий NCI-60 на основании литературных данных [4]. Контекст-зависимые модели с использованием омиксных данных были реконструированы для 12 наиболее изученных линий (на основе публикационной активности) с использованием алгоритма RIPTiDe. Протеомные данные [5] обработаны с восемью типами нормализации (TSS, Median, Quantile и др.) [6]. Для интеграции в модель использовались правила ген-белок (gene protein rules), связывающие гены с ферментами. Ограничения потоков задавались на основе максимальных значений обмена для всех линий, с вариациями:

- Минимальные значения потоков;
- Минимальные значения $\times 1.25$;
- Округление до ближайшего десятичного значения.

Транскриптомные данные нормализовались с помощью RPM (Reads per million mapped reads).

Результаты

Полученные распределения потоков и значения роста сравнены с экспериментальными значениями и с моделями,

ограниченными транскриптомом (Рис. 1). Использование ограниченных с коэффициентом 1.25 моделей показало в среднем лучшие показатели. Несмотря на то, что интеграция протеомных данных успешно прошла не для всех клеточных линий, зачастую давая нулевые скорости роста, отмечается более высокая корреляция полученного распределения потоков с интегрируемыми данными; таким образом можно быть уверенным, что протеомика более биологически достоверные результаты. Определенные типы нормализации протеомных данных (TSS, Median) дают результаты роста в целом точнее остальных.

Рис. 1: Тепловая карта демонстрирует процентные отклонения предсказанных значений роста модели от экспериментальных данных при различных ограничениях максимальной скорости роста. По вертикали отображены клеточные линии, по горизонтали — варианты ограничений с шагом 0.05.

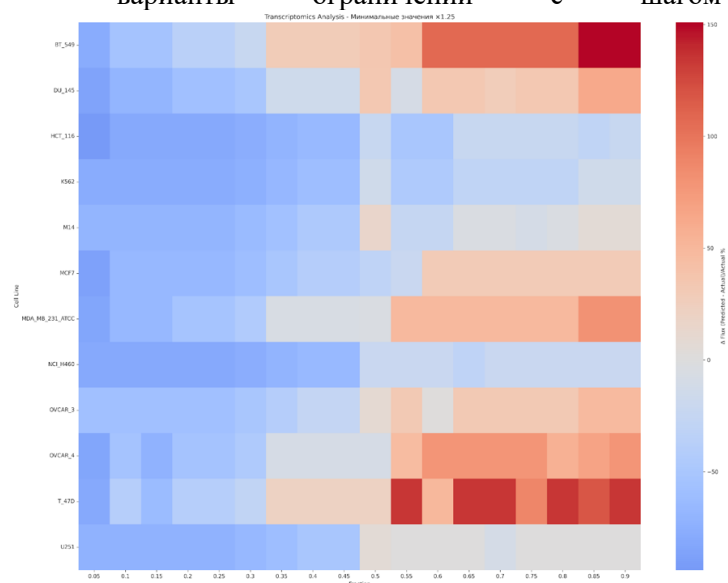
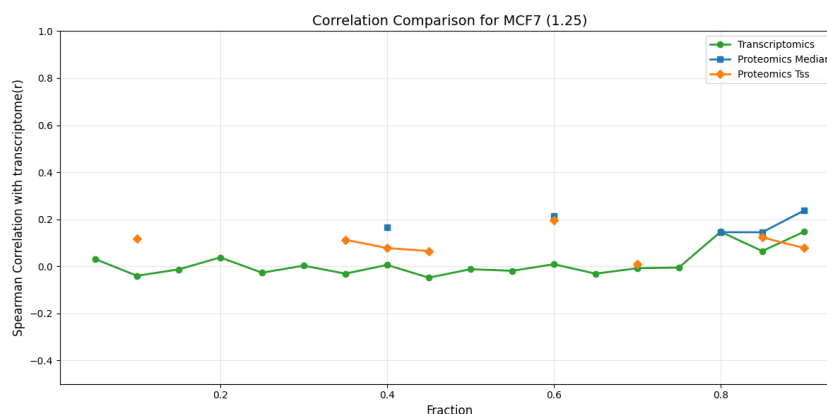


Рис. 2: Корреляция Спирмана между омиксными данными и распределением метаболических потоков для клеточной линии MCF7. Сравнение моделей, ограниченных транскриптомными данными (зеленый) и протеомными данными с нормализацией Median (синий) и Total Sum Square (TSS, оранжевый).



Исследование выполнено за счет гранта Российского научного фонда No 24-14-20031

1. M. L. Jenior et al. (2020) "Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments" *PLoS Comput Biol*, 4:e1007099
2. B. J. Sánchez et al. (2017) "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints" *Molecular Systems Biology*, 8:935
3. E. Brunk et al. (2018) "Recon3d enables a three-dimensional view of gene variation in human metabolism" *Nat Biotechnol*, 272–281
4. J. L. Robinson et al. (2020) "An Atlas of Human Metabolism" *Sci Signal*, 13(624):eaaz1482
4. T. Guo et al. (2019) "Quantitative proteome landscape of the NCI-60 cancer cell lines" *iScience*, 21:664–680
5. T. Välikangas et al. (2016) "A systematic evaluation of normalization methods in quantitative label-free proteomics" *Brief Bioinform*, bbw095

Молекулярные корреляты социальной эволюции: митохондриальный мутационный спектр Blattodea

А.А. Кошель, Г.Г. Осадчий, А.А. Михайлова, А.Г. Михайлова, Б.Э.
Ефименко, К. Ю. Попадьин

*Центр геномных исследований БФУ им. Канта,
anastasiakoshel2001@gmail.com*

Аннотация

Внутригрупповые различия по продолжительности жизни у насекомых с кастовой системой дают уникальную возможность для изучения взаимосвязи между жизненной историей и мутагенезом митохондриальной ДНК (мтДНК). В данной работе мы исследуем, как эусоциальность и связанное с ней долголетие у различных групп термитов отражаются в спектре мутаций, нуклеотидном составе и аминокислотных свойствах митохондриального генома. Используя вычислительные методы реконструкции филогении и анализа нейтральных сайтов, мы показываем, что мутационный сигнатур долголетия — преобладание замен аденина на гуанин (далее A>G) на тяжёлой цепи мтДНК (далее h)— выражен у социальных видов и проявляется на всех уровнях геномной организации.

Введение

Мутационный спектр мтДНК может отражать ключевые биологические характеристики вида, включая его продолжительность жизни и уровень окислительного стресса. Ранее было показано, что у млекопитающих долголетие коррелирует с накоплением Ah>Gh замен на тяжёлой цепи митохондриального генома. Мы применили подобный подход к представителям отряда Blattodea, объединив данные по социальным и асоциальным видам, включая термитов разных уровней социальной организации.

Материалы и методы

Выборка и данные. Анализ проводился на 397 видах Blattodea: 48 тараканов и 331 термит. Были отобраны последовательности гена COX1, а также полные митохондриальные геномы из базы MIDORI Reference 2. Для каждой последовательности определялись 4-кратно

вырожденные синонимичные сайты, принятые за наиболее нейтральные в эволюционном плане.

Филогенетическая реконструкция. С помощью ранее опубликованного пайплайна NeMu были построены филогенетические деревья и проведена реконструкция предковых состояний. Мутации фиксировались по терминальным ветвям. Все замены были нормализованы по количеству исходных нуклеотидов.

Расчёт мутационных и композиционных метрик. Оценивалась частота всех 12 типов однонуклеотидных замен, а также метрики GhAhSkew и ThChSkew, отражающие частоту замен Ah>Gh и Ch>Th соответственно. Анализ проводился как на уровне отдельных видов, так и по группам: тараканы, субсоциальные тараканы *Cryptocercus*, термиты без рабочей касты (низшие) и термиты с выраженной кастовой системой (высшие).

Результаты

1. Мутационный спектр:

У термитов с рабочей кастой (высших) доля Ah>Gh мутаций достоверно выше, чем у всех других групп ($p < 0.05$). Уровень Ah>Gh коррелирует с социальной сложностью и, предположительно, с долголетием репродуктивных особей.

2. Нуклеотидный состав и асимметрия:

У термитов наблюдается снижение содержания А и увеличение G на нейтральных сайтах. GhAhSkew наиболее выражен у высокоорганизованных термитов, в то время как ThChSkew не показывает значимых различий между группами.

3. PCA-анализ:

Первый главный компонент разделяет группы по оси GhAhSkew и содержанию А и G.

Это подтверждает, что основным вкладом в межгрупповую вариацию является именно мутационная направленность Ah>Gh.

4. Кодоновый и аминокислотный состав:

Частота NNG-кодонов и «приобретаемых» аминокислот (Val, Leu, Phe) выше у термитов. У тараканов преобладают «теряемые» аминокислоты (Lys, Thr, Asn), отражая менее

направленный мутационный спектр.

Выводы

Результаты демонстрируют, что социальная структура, а в частности продолжительность жизни, оказывает системное влияние на митохондриальный мутагенез. Повышенная частота Ah>Gh замен у термитов ассоциирована не только с особенностями мутационного спектра, но и с нейтральным нуклеотидным составом и аминокислотной организацией белков.

Список источников

1. Arbeithuber, Barbara, James Hester, Marzia A. Cremona, Nicholas Stoler, Arslan Zaidi, Bonnie Higgins, Kate Anthony, Francesca Chiaromonte, Francisco J. Diaz, and Kateryna D. Makova. (2020) Age-related accumulation of de novo mitochondrial mutations in mammalian oocytes and somatic tissues, PLOS Biology. – P. 3000745
2. Mikhailova, Alina G., Victor Shamansky, Alina A. Mikhailova, Kristina Ushakova, Evgenii Tretyakov, Sergey Oreshkov, Dmitry Knorre, et al. (2020) A mitochondrial mutational signature of temperature in ectothermic and endothermic vertebrates, bioRxiv. – P. 221184
3. Tasaki Eisuke, Takata Mamoru and Matsuura Kenji (2021) Why and how do termite kings and queens live so long?, Phil. Trans. R. Soc. – №. 376. – P. 0740.
4. A. G. Mikhaylova, A. A. Mikhailova, K. Ushakova, E.O. Tretyakov, D. Iliushchenko, V. Shamansky, A. Iurchenko, M. Zazhytska, E. Kozenkova, E. Zdobnov, V. Makeev, V. Yurov, M. Tanaka, I. Gostimskaya, Z. Fleischmann, S. Annis, M. Franco, K. Wasko, W.S Kunz, D.A. Knorre, I. Mazunin, S. Nikolaev, J. Fellay, A. Reymond, K. Khrapko, K. Gunbin, K. Popadin (2022) A mitochondria-specific mutational signature of ageing: increased rate of A>G substitutions on a heavy chain, bioRxiv. – P. 460832
5. Bogdan Efimenko, Konstantin Popadin, Konstantin Gunbin (2024) NeMu: A Comprehensive Pipeline for Accurate Reconstruction of Neutral Mutation Spectra from Evolutionary Data, Nucleic Acids Research. – 52:108–115

Особенности транскрипции генов с однонаправленных и двунаправленных промоторов в генетическом и эпигенетическом контексте генома человека

Г.И. Кравацкая, Н.А. Чуриков, Ю.В. Кравацкий

119991, Москва, ул. Вавилова, 32, Институт Молекулярной Биологии РАН, gk@eimb.ru

Геном человека содержит довольно большое число пар генов, транскрипция которых идет в разных направлениях по разным цепям ДНК и точки стартов транскрипции которых расположены на расстоянии не более 1000 нуклеотидов друг от друга. Участок ДНК между точками старта таких близкорасположенных генов называют двунаправленным промотором. Целью работы было выяснить, чем двунаправленные промоторы отличаются от однонаправленных с точки зрения строения нуклеотидной последовательности, генетического и эпигенетического геномного окружения соответствующих им точек старта транскрипции, а также изучить особенности транскрипции соответствующих им генов. С использованием созданных нашим коллективом оригинальных статистических и аналитических критериев для оценки попарных корреляций и колокализаций [1] проведен полногеномный биоинформатический анализ двунаправленных и однонаправленных промоторов генома человека для линии клеток K562 в окрестностях точек старта транскрипции. Было рассмотрено более 200 различных признаков, связанных с организацией генома в промоторных областях (расположение сайтов связывания различных регуляторных белков, плотность расположения нуклеосом в промоторных областях, чувствительность к ДНКазе I, частоты двухцепочечных разрывов ДНК [2], метилирования ДНК, распределение эпигенетических меток хроматина, а также распределение участков, образующих некоторые неканонические структуры ДНК).

Проведено полногеномное исследование колокализации разных типов промоторов и модификации гистона H2A.Z [1,3] в

линии клеток K562 *H.sapiens*. Обнаружено, что связывание H2A.Z с ДНК в областях двунаправленных промоторов происходит в 1.7 раза чаще, чем в областях однонаправленных промоторов. Также в областях двунаправленных промоторов наблюдается статистически значимая колокализация с перекрытием регионов пиков H2A.Z и TSS; в случае однонаправленных промоторов TSS чаще оказываются за пределами регионов, содержащих пики H2A.Z. Таким образом, нами впервые показано, что в этом смысле модификация гистона H2A.Z может быть маркером двунаправленной транскрипции в клеточной линии K562 *H.sapiens*. Впервые обнаружено, что промоторные области размером 200-600 п.н. вокруг TSS часто являются мишенями эндогенных двухцепочечных разрывов DSB RAFT, особенно это касается областей вокруг TSS двунаправленных промоторов экспрессирующих генов [4].

На начальном этапе работа была поддержана грантом РФФИ 20-04-01134, далее - в рамках Программы фундаментальных научных исследований в Российской Федерации (тема 124031500019-8)

1. Kravatsky Y.V. et. al. (2022) Genome-Wide Study of Colocalization between Genomic Stretches: A Method and Applications to the Regulation of Gene Expression. *Biology* 11(10):1422.
2. Tchurikov N.A. et al. (2016). Hot spots of DNA double-strand breaks in human rDNA units are produced in vivo. *Sci Rep* 6:25866.
3. Li S. et al. (2023) Histone variant H2A.Z modulates nucleosome dynamics to promote DNA accessibility. *Nat Commun* 14(1):769.
4. Tchurikov N.A. et al. (2022) Genes Possessing the Most Frequent DNA DSBs Are Highly Associated with Development and Cancers, and Essentially Overlap with the rDNA-Contacting Genes. *Int J Mol Sci* 23(13):7201.

Метилот *Porphyromonas gingivalis* и особенности его геномной организации

Д.В. Кривонос¹, А.Д. Проскурин¹, А.А. Ахмедзянова¹, Т.А. Савинова¹, К.А. Микозян¹, А.В. Введенский¹, Д.Н. Конанов¹, А.В. Лукина-Гронская¹, В.М. Говорун¹, Е.Н. Ильина¹, З.Э. Ревазова², М.О. Царгасова², Е.И. Выборная², В.А. Бризгалова², И.С. Бобр², О.О. Янушевич²

¹НИИ системной биологии и медицины Роспотребнадзора,
danil01060106@gmail.com

²ФГБОУ ВО «Российский университет медицины»

Porphyromonas gingivalis относится к группе пародонтопатогенов и участвует в развитии пародонтита. Особое место локализации данного патогена в глубине пародонтального кармана дает ему своеобразную защиту, снижая локальную доступность и, как следствие, силу воздействия антибиотиков и антимикробных препаратов, что затрудняет применение классических методов лечения бактериальных заболеваний. Таким образом, возникает потребность в подборе более оптимальных методов борьбы с данным патогеном. Для этого прежде всего необходимо получить полное представление об устройстве данного патогена. Иными словами, получить некоторую системную характеристику, с целью выявления особенностей строения и регуляции генетического материала вида *Porphyromonas gingivalis*.

В ходе работы была сформирована коллекция из 10 клинических изолятов *P. gingivalis* и 20 метагеномных образцов содержимых парадонтальных карманов. Изоляты и метагеномные образцы были секвенированы на платформе Oxford Nanopore. Так, в результате работы были получены сборки 10 изолятов *P. gingivalis* и 20 геномов собранных из метагеномов (MAGs).

Немаловажную роль в регуляции *P. gingivalis* играют его мобильные элементы. Особенно инсерционные последовательности (IS элементы), которые повсеместно распространены в его геноме. В результате данной работы были

выявлены различные особенности в структуре IS элементов. Такие как, «короткие» рамки считывания транспозаз, сильная недопредставленность IS элементов в отдельных штаммах бактерии.

Среди других мобильных элементов, ранее описанных для *P. gingivalis* встречаются также и умеренные бактериофаги, обнаруженные в некоторых опубликованных сборках геномов. Это, в свою очередь, дает надежду на применение фаговой терапии для лечения пародонтита и других воспалительных заболеваний вызываемых *Porphyromonas gingivalis*. В ходе своей работы мы также обнаружили лизогенного бактериофага порядка *Caudoviricetes*, который мог бы в перспективе стать тем самым решением для будущей фаговой терапии.

Полученные результаты приблизили нас к пониманию генома *Porphyromonas gingivalis*, однако на ряд вопросов еще предстоит найти ответы в будущих работах. Для 10 штаммов изолятов, отсеквенированных на платформе Oxford Nanopore были определены 15 сайтов рестрикции модификации, среди которых 12 содержат 6meA и 4 4meC. Было также отмечено, что различные штаммы *P. gingivalis* имеют разный процент фракции метилированных оснований, что, по-видимому, объясняется различиями в работе присутствующих метилаз. Помимо всего прочего было показано, что некоторые метилтрансферазы переносятся IS элементами *P. gingivalis*, влияя тем самым на способность штамма защищаться от специфичных бактериофагов.

Определение абсолютной свободной энергии связывания Дануглипрона с рецептором ГПП-1 методом AWH

А. С. Кривошеев, А. В. Головин

*Научно-технологический университет «Сириус», Олимпийский
просп., д. 1, Сириус, Краснодарский край, Российская Федерация,
354349, krivosheev-sci@yandex.ru*

В настоящее время ожирение и диабет 2 типа стали крайне широко распространёнными заболеваниями, каждое из которых

признано Всемирной организацией здравоохранения эпидемией [1]. Одним из подходов для лечения этих заболеваний является применение агонистов рецептора глюкагоноподобного пептида 1 (ГПП-1) [2, 3]. Секреция пептидного гормона ГПП-1 происходит в дистальной части подвздошной кишки в ответ на поступление таких питательных веществ, как белки и углеводы, после чего он связывается с рецептором ГПП-1 в бета-клетках поджелудочной железы, где его активация приводит к высвобождению инсулина и ингибированию высвобождения глюкагона, а также снижает за счёт задержки опустошения желудка чувство голода, что уменьшает объём принимаемой пищи [4].

В последние годы активно ведётся разработка альтернативы пептидным препаратам-агонистам рецептора ГПП-1, вводимых в основном инъекционным путём, — низкомолекулярных агонистов. К основным преимуществам низкомолекулярных препаратов относят возможность перорального приёма, что более удобно для пациентов, потенциально лучшую тканевую проницаемость, долгий период полувыведения и более низкую цену производства [5]. Применение вычислительных подходов на этапе доклинической разработки лекарственных препаратов стало способом справиться с высокой сложностью этого процесса. Одной из важных характеристик лигандов является свободная энергия связывания их с мишенью, например, рецептором. Молекулярная динамика представляет из себя подход, применяющий классическую механику для изучения динамики молекулярных систем, и позволяет, в частности, определить свободные энергии связывания [6].

Нами были произведены расчёты свободной энергии связывания низкомолекулярного препарата Дануглипрона с рецептором ГПП-1. Молекула представляет особый интерес из-за сходного с естественным агонистом рецептора ГПП-1 механизма действия [7]. Исследование проводилось с применением метода ускоренной гистограммы весов (accelerated weight histogram, AWH) в варианте, позволяющем производить алхимические расчёты [8]. В основе алхимического метода лежит проведение полноатомных молекулярно-динамических симуляций в явном растворителе, что позволяет учесть эффект как растворителя, так и энтропии [9]. Для расчёта абсолютной свободной энергии связывания лиганд

нефизическим образом удаляется из раствора и вставляется в сайт связывания белка [9].

Параметризация и генерация топологий осуществлялась с использованием ACPYPE, частичные заряды были получены с помощью PsiRESP [10], использованное силовое поле — GAFF2. Структура комплекса лиганд-рецептор была взята из базы данных PDB (PDB: 6X1A [7]), из структуры были удалены все цепи белков кроме самого рецептора. Восстановление недостающих аминокислотных остатков в цепи производилось с помощью Modeller [11]. Рецептор был помещён в липидный бислой, состоящий из молекул фосфатидилхолина POPC. Молекулярно-динамические расчёты проводились с помощью программы GROMACS 2024.3 [12] при температуре 310 K.

В результате свободная энергия десольватации Дануглипрона $\Delta G_{elec+vdw}^{solv}$ равна $78,3 \pm 0,5$ ккал/моль (здесь и далее среднеквадратичное отклонение в качестве неопределённости), причём свободная энергия растворения лиганда может быть определена как $\Delta G^{solv} = -\Delta G_{elec+vdw}^{solv}$. Свободная энергия нефизического удаления Дануглипрона из сайта связывания белка $\Delta G_{elec+vdw+restr}^{prot}$ равна $95,8 \pm 4,3$ ккал/моль. Значение ограничения Бореша $\Delta G_{restr_{on}}^{solv}$ [13] было определено аналитически и равно 6,8 ккал/моль. Итоговое значение абсолютной свободной энергии связывания может быть определено по формуле [9]

$$\Delta G_b^o = -\Delta G_{elec+vdw+restr}^{prot} + \Delta G_{elec+vdw}^{solv} + \Delta G_{restr_{on}}^{solv}$$

и равно для Дануглипрона $-10,7 \pm 4,4$ ккал/моль.

Такое высокое значение неопределённости в результатах может затруднять в дальнейшем сравнение различных лигандов между собой на основании значения свободной энергии связывания. Возможные пути решения этой проблемы следующие: модификация применённого протокола, например, добавлением обмена репликами [14], либо использование другого метода для расчётов, например, метадинамики с ограничивающим потенциалом в форме воронки [15]. Применённый в настоящей работе подход, как и любые его модификации, может быть использован в дальнейшем для создания альтернатив существующей запатентованной лекарственной молекуле, которые

будут патенто-чистыми и показывать характеристики связывания с рецептором на уровне Дануглипрона.

1. A. Chobot, K. Górowska-Kowolik, M. Sokołowska, P. Jarosz-Chobot (2018) Obesity and diabetes — Not only a simple link between two epidemics, *Diabetes Metabolism Research and Reviews*, 34:e3042.
2. A. DeMarsilis et al. (2022) Pharmacotherapy of type 2 diabetes: An update and future directions, *Metabolism — Clinical and Experimental*, 137:155332.
3. A. Cignarella, L. Busetto, R. Vettor (2021) Pharmacotherapy of obesity: An update, *Pharmacological Research*, 169:105649.
4. M. O. Mahgoub, I. I. Ali, J. O. Adeghate, K. Tekes, H. Kalász, E. A. Adeghate (2023) An Update on the Molecular and Cellular Basis of Pharmacotherapy in Type 2 Diabetes Mellitus, *International Journal of Molecular Sciences*, 24(11):9328.
5. Z. Zheng et al. (2024) Glucagon-like peptide-1 receptor: mechanisms and advances in therapy, *Signal Transduction and Targeted Therapy*, 9(1):234.
6. V. T. Sabe et al. (2021) Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review, *European Journal of Medicinal Chemistry*, 224:113705.
7. X. Zhang et al. (2020) Differential GLP-1R Binding and Activation by Peptide and Non-peptide Agonists, *Molecular Cell*, 80(3):485.
8. M. Lundborg, J. Lidmar, B. Hess (2021) The accelerated weight histogram method for alchemical free energy calculations, *The Journal of Chemical Physics*, 154(20):204103.
9. M. Aldeghi, J. P. Bluck, P. C. Biggin (2018) Absolute Alchemical Free Energy Calculations for Ligand Binding: A Beginner's Guide, In: *Computational Drug Discovery and Design*, M. Gore et al. (Eds.), 199–232 (Springer New York).
10. L. Wang, M. L. O'Mara (2022) PsiRESP: calculating RESP charges with Psi4, *Journal of Open Source Software*, 7(73):4100.
11. B. Webb, A. Sali (2016) Comparative protein structure modeling using MODELLER, *Current Protocols in Bioinformatics*, 54(1):5.6.1–5.6.37.
12. M. J. Abraham et al. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to

supercomputers, *SoftwareX*, 1–2:19–25.

13. S. Boresch, F. Tettinger, M. Leitgeb, M. Karplus (2003) Absolute Binding Free Energies: A Quantitative Approach for Their Calculation, *The Journal of Physical Chemistry B*, 107(35):9535–9551.

14. F. Fratev, S. Sirimulla (2019) An Improved Free Energy Perturbation FEP+ Sampling Protocol for Flexible Ligand-Binding Domains, *Scientific Reports*, 9(1):16829.

15. S. Raniolo, V. Limongelli (2020) Ligand binding free-energy calculations with funnel metadynamics, *Nature Protocols*, 15(9):2837–2866.

Genetic and epigenetic determinants of species longevity

Dmitrii Kriukov

*Artificial Intelligence Research Institute, Moscow, Russia;
Skolkovo Institute of Science and Technology, Moscow, Russia.
kriukov@airi.net*

One of the most widely accepted criteria for a “good” theory of aging is its ability to explain the most conspicuous inter-specific difference—the maximum lifespan of a species, hereafter referred to as “longevity”. Only a limited set of traits can be meaningfully compared across species in the context of aging research, owing to marked differences in baseline metabolite levels, phenotypic characteristics, or even gene-expression profiles. Consequently, it is expedient to begin such an investigation with the DNA sequence—the most comprehensive repository of biological information described to date and, presumably, the primary determinant of species identity.

The use of DNA sequences for longevity explanation is not a new idea: earlier work showed that the density of CpG sites in the promoters of conserved genes across species explains observed longevity rather well [1]. Yet this line of enquiry was not pursued exhaustively; many researchers, recognising a familiar CpG pattern, shifted their focus from genetic to epigenetic determinants [2, 3], namely, DNA methylation. These efforts were likewise fruitful, demonstrating that both the mean level of methylation [2] or the rate of change in methylation marks [3] can account for inter-specific differences in longevity. Mechanistic interpretations of these findings have also been

proposed [4], describing an epigenetic mechanism that maintains genomic stability and apparently underpins variation in longevity.

In the present study I revisit and reproduce these results, integrating genetic and epigenetic datasets from multiple species with curated longevity records. I examine which sequence-level patterns best explain species-specific longevity, relate them to epigenetic modifications, and provide a more detailed characterisation of the underlying mechanisms. Finally, I position these findings within contemporary gerontological frameworks and outline a theory of aging capable of encompassing the full range of observations.

1. Mayne B. et al. A genomic predictor of lifespan in vertebrates //Scientific Reports. – 2019. – T. 9. – №. 1. – C. 1-10.
2. Li C. Z. et al. Epigenetic predictors of species maximum life span and other life-history traits in mammals //Science Advances. – 2024. – T. 10. – №. 23. – C. eadm7273.
3. Crofts S. J. C., Latorre-Crespo E., Chandra T. DNA methylation rates scale with maximum lifespan across mammals //Nature aging. – 2024. – T. 4. – №. 1. – C. 27-32.
4. Bertucci E. M., Parrott B. B. Is CpG density the link between epigenetic aging and lifespan? //Trends in Genetics. – 2020. – T. 36. – №. 10. – C. 725-727.

Lipoxygenases and CFTR inhibitory factors might expand the host ranges of bacteria

G. Kurakin

Independent (unaffiliated) researcher
georgykurakin@gmail.com

We previously hypothesized that bacterial lipoxygenases are linked to cross-kingdom host jumps — which means the ability of a bacterium to affect or colonize two or more species belonging to two or more different kingdoms [1]. In particular, it was noted that bacterial lipoxygenases are associated with affecting plants and humans, or both simultaneously. In my recent preprint, I aimed to elaborate this idea with the means of more detailed phylogenetic analysis, network text analysis

and statistical analysis of active site metrics [2].

In this additional research, I have found that (ω -6)S-lipoxygenases are associated with affecting both plants and vertebrates and proposed a mechanism for this effect. According to the hypothesis, (ω -6)S-lipoxygenases contribute to lipoxin production in humans and to jasmonate production in plants. The both compound classes can suppress the defence responses in their hosts when being overproduced. This could explain why some bacterial lipoxygenases show this dual association with plants and vertebrates.

However, the attempts to publish this preprint formally faced significant doubts and discussions among reviewers. Some of the reviewers' comments could be considered irrelevant, but the major comments regarded methodological issues: manual entity extraction without any automatic tools such as *nltk*, absence of mathematical way to express the correspondence between phylogenetics and network text analysis, and absence of negative or positive controls.

This paper is aimed to partially address the last problem. The selection of a control itself is difficult in this case: it is unclear which dataset could be used that could match the lipoxygenase dataset by all parameters except carrying lipoxygenase. However, analysis of similar cases could provide valuable insights that confirm our hypothesis and serve like a quasi-positive control.

In our recent review, we highlighted a parallel between the action of lipoxygenases and CFTR inhibitory factors (Cif) in the human lung invasion by *Pseudomonas aeruginosa* [3]. Cif is an epoxide hydrolase enzyme with relatively broad substrate specificity. It has been originally noted that it decreases the expression of CFTR on human cells thus phenocopying cystic fibrosis [4], but new results show that it can spoof human oxylipin signalling by cleaving oxylipins formed by the epoxygenase pathway [5]. These parallels urged me to start a new bioinformatic study of bacterial Cif to compare the results with the case of lipoxygenase.

Here, I provide preliminary results of constructing a MAFFT neighbour joining phylogeny of the homologues of *Pseudomonas aeruginosa* Cif found by a BLAST search. *Pseudomonas aeruginosa* Cif shares the common cluster of closely related epoxide hydrolases with other confirmed functional Cif carried by *Acinetobacter nosocomialis*. By assuming that all proteins descending from their common ancestor

are also functional Cif's, I identified a set of putative bacterial Cif's and the respective bacterial species.

The total number of identified species/strains without species name was 101. Of them, 34 species (34%) have been described in literature as pathogens or symbionts of a specific host or multiple hosts. As in the case of lipoxxygenase, they together covered an extremely broad host range. Of these 34 species, 9 (26%) were human/vertebrate pathogens, 8 (23%) were associated with plants, and 6 (18%) had the ability to affect both plants and humans. Smaller proportion (15%) was associated with aquatic organisms (but here we should be aware of a possible data availability bias). For lipoxxygenases (as per my preprint data [2]) the respective proportions were 35%, 30%, and 11% for the sample of 37 lipoxxygenase-carrying pathogens/symbionts, which is a comparable sample size. So, Cif and lipoxxygenase samples have no statistically significant differences by these characteristics and look like two parts of the same statistical universe. There was also no significant difference in the percentage of xenobiotic degraders (5 for the Cif sample and 2 for lipoxxygenase sample).

The similarity looks much deeper as the found bacteria shows the same ecological properties as lipoxxygenase carriers do: tend to be endophytic in plants and opportunistic in humans, show antimicrobial resistance or have emerging status.

If the functions and phylogeny of lipoxxygenase and Cif were unknown, we might just state that they are not involved in host-microbe interactions. But lipoxxygenases and Cif have been experimentally characterized as pathogenicity factors in *Pseudomonas aeruginosa* (Cif — also in *Acinetobacter nosocomialis*) [4, 6, 7]. Moreover, lipoxxygenases and Cif's of pathogenic and symbiotic bacteria cluster together on phylogenetic trees, and these clusters include experimentally characterized *Pseudomonas aeruginosa* proteins. This leads to an opposite assumption: both lipoxxygenases and Cif's might be universal colonization factors.

For the case of lipoxxygenases, I proposed a potential mechanism of universal action as a colonization factor in plants and humans [2]. This is not the case of Cif's. Obviously, they interfere with the epoxxygenase pathway in humans. But we are still at a rudimentary stage of understanding the epoxxygenase pathway in plants: fatty acid epoxxygenase has been discovered in flowering plants only in 2021 [8],

and the details about downstream enzymes are still unknown.

At the current stage, I am working on the scientific and technical challenges for a new network text analysis for Cif, which could be less sensitive to data availability bias.

Potentially, our results on lipoxygenases and Cif could pave the way to an integrated view of bacterial strategies of oxylipin spoofing for cross-kingdom host jumps and expanding host ranges. However, the verification of falsification of our assumptions will take time and new research. Thus, despite the current challenges with formal publishing of the paper on lipoxygenase [2], I will keep it available as a preprint. Meanwhile, I am continuing the research on Cif and other markers which could indirectly confirm the role of oxylipins as universal factors of host-microbe interaction.

Figures and diagrams are available in our ISMB conference poster deposited on *F1000Research* [9].

Abbreviations: CFTR — Cystic fibrosis transmembrane conductance regulator

1. G.F. Kurakin, A.M. Samoukina, N.A. Potapova (2020) Bacterial and Protozoan Lipoxygenases Could be Involved in Cell-to-Cell Signaling and Immune Response Suppression, *Biochemistry (Moscow)*, 85:1048–1063.
2. G.F. Kurakin (2022) Bacterial lipoxygenases are associated with host-microbe interactions and may provide cross-kingdom host jumps, *bioRxiv*, 2022.06.21.497025.
3. A.S. Amoah, N.B. Pestov, T.V. Korneenko, I.A. Prokhorenko, G.F. Kurakin, N.A. Barlev (2024) Lipoxygenases at the Intersection of Infection and Carcinogenesis, *International Journal of Molecular Sciences*, 25(7): 3961.
4. D.P. MacEachran, S. Ye, J. Bomberger, D.A. Hogan, A. Swiatecka-Urban, B.A. Stanton, G.A. O'Toole (2007) The *Pseudomonas aeruginosa* secreted protein PA2934 decreases apical membrane expression of the cystic fibrosis transmembrane conductance regulator, *Infection and Immunity*, 75(8): 3902-3912.
5. K.L. Hvorecny, C.D. Bahl, S. Kitamura, K.S.S. Lee, B.D. Hammock, C. Morisseau, D.R. Madden (2017) Active-Site Flexibility and Substrate Specificity in a Bacterial Virulence Factor: Crystallographic Snapshots of an Epoxide Hydrolase, *Structure*, 25: 697-707.e4.

6. E. Morello, T. Pérez-Berezo, C. Boisseau et al. (2019) *Pseudomonas aeruginosa* lipoxygenase LoxA contributes to lung infection by altering the host immune lipid signaling, *Frontiers in Microbiology*, 10: 1826.
7. C.D. Bahl, K.L. Hvorecny, A.A. Bridges et al. (2014) Signature motifs identify an *Acinetobacter* Cif virulence factor with epoxide hydrolase activity, *Journal of Biological Chemistry*, 289(11): 7460-9.
8. E. Pineau, V. Sauveplane, E. Grienberger, J.E. Bassard, F. Beisson, F. Pinot (2021) CYP77B1 a fatty acid epoxigenase specific to flowering plants, *Plant Science*, 307, 110905
9. G. Kurakin (2024) Oxylin "spoofing" by bacteria could contribute to cross-kingdom host jumps and affecting cystic fibrosis patients [version 1; not peer reviewed]. *F1000Research*, 13:1373 (poster).

Интерпретация предсказаний сверточных нейронных сетей для регуляторных регионов генома

В.А. Лабанов, А.О. Зинкевич, Д.Д. Пензар

*Факультет биоинженерии и биоинформатики, МГУ им. М.В.
Ломоносова, Москва, labanovvlad@yandex.ru*

Понимание механизмов, лежащих в основе регуляции активности генов является одной из важнейших задач современной биоинформатики. В число этих механизмов входит связывание транскрипционных факторов (ТФ), их взаимодействие между собой и позиционно-специфичные эффекты. Каждая позиция регуляторной последовательности в геноме может вносить уникальный вклад в активацию или подавление экспрессии. Глубокие нейронные сети широко используются для предсказания активности регуляторных последовательностей ДНК благодаря их способности выявлять как простые паттерны, наподобие мотивов связывания ТФ, так и более сложные взаимодействия. Однако интерпретация предсказаний нейронных сетей остается важной

проблемой, так как информация о структуре паттернов и их влиянии на предсказание модели необходима для изучения механизмов регуляции генов. Эту проблему позволяют решить современные методы интерпретируемого искусственного интеллекта (ХАИ). В данной работе было проведено сравнение различных методов интерпретации предсказаний нейронных сетей и был создан фреймворк, позволяющий выявлять паттерны регуляторных элементов, влияющие на экспрессию соответствующих генов.

Фреймворк состоит из глубокой сверточной нейронной сети LegNet [1], разработанной нашей лабораторией, и объясняющего алгоритма. Он принимает на вход последовательность промотора или энхансера и предсказывает влияние каждого нуклеотида на экспрессию соответствующего гена. Для выбора наиболее точного объясняющего алгоритма были сравнены различные алгоритмы, написанные самостоятельно или взятые из других фреймворков. При сравнении алгоритмов и валидации фреймворка использовался набор данных, созданный с помощью насыщающего мутагенеза (Kircher et al, 2019) и доступный публично [2]. Результаты сравнения объясняющих алгоритмов представлены на рисунке 1. Также для сравнения использовались данные из базы данных аллель-специфических сайтов связывания транскрипционных факторов человека (ADAstra) [3] и атласа аллель-специфической доступности хроматина (UDACHA) [4].

Итак, в данной работе было проведено сравнение различных методов интерпретации предсказаний нейронных сетей для задач регуляторной геномики. Было показано, что ни один из алгоритмов не смог достичь точности ISM, который считается золотым стандартом интерпретации. При этом хорошие результаты показываются и менее вычислительно затратными, а значит, применимыми к большим объемам данных, методами, например, TISM [5], Occlusion, GradientShap [6] и IG [7].

1. Penzar D, et al. "LegNet: a best-in-class deep learning model for short DNA regulatory regions." *Bioinformatics* 39.8 (2023): btad457.
2. Kircher M, et al. "Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution." *Nature*

communications 10.1 (2019): 3583.

3. Abramov S, et al. "Landscape of allele-specific transcription factor binding in the human genome." *Nature communications* 12.1 (2021): 2751.

4. Buyan A, et al. "Statistical framework for calling allelic imbalance in high-throughput sequencing data." *Nature Communications* 16.1 (2025): 1739.

5. Sasse A, et al. "Quick and effective approximation of in silico saturation mutagenesis experiments with first-order taylor expansion." *Iscience* 27.9 (2024).

6. Kokhlikyan N, et al. "Captum: A unified and generic model interpretability library for pytorch." *arXiv preprint arXiv:2009.07896* (2020).

7. Sundararajan M, et al. "Axiomatic attribution for deep networks." *International conference on machine learning*. PMLR, 2017.

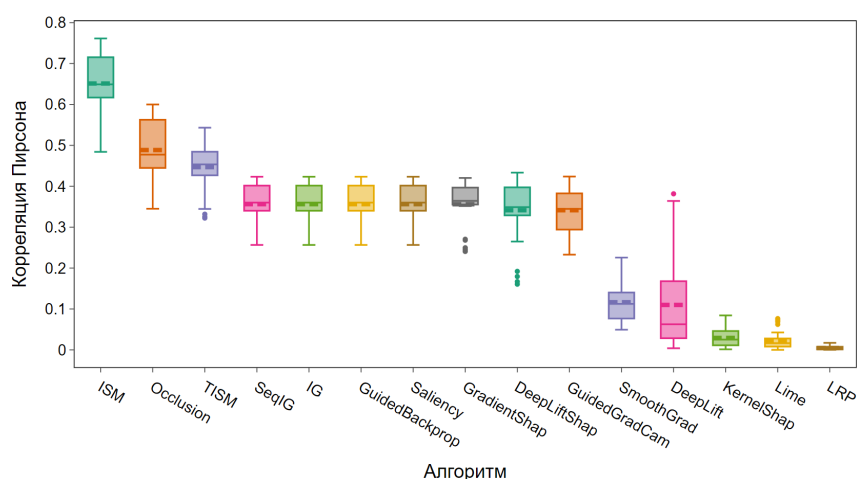


Рис. 1. Сравнение точности предсказания влияния нуклеотидов регуляторных элементов на экспрессию различными объясняющими алгоритмами. На box-plot для каждого объясняющего алгоритма (ось абсцисс) показано распределение корреляций Пирсона атрибуций выбранной группы элементов (ось ординат).

**Функциональные особенности микросателлитов в
сравнительной геномике: анализ факторов, влияющих на
полиморфизм tandemных повторов**

А.В. Леонов, Д.В. Попов, Г.Ю. Косовский

*ФГБНУ НИИ пушного звероводства и кролиководства им. В.А.
Афанасьева, 140143 Россия, Московская обл., Раменский р-н, пос.
Родники, ул. Трудовая, 6, thelastmanintheworld@gmail.com*

Микросателлитные маркеры широко применяются для генотипирования в геномике различных видов благодаря своей высокой полиморфности и информативности. Несмотря на универсальность применения микросателлитов в геномах различных видов — от насекомых до млекопитающих — ключевые факторы, определяющие вариабельность их полиморфизма, остаются недостаточно изученными. Ряд исследований указывает на зависимость полиморфизма микросателлитов от таких факторов, как частота репликационных ошибок (например, slippage-мутаций), функциональный контекст локуса (кодирующие vs. некодирующие регионы) и влияние естественного отбора [1]. Однако сравнительный анализ этих факторов между таксономически удаленными группами остается редким.

Ранее нами было выявлено сходство полиморфизма у разных видов млекопитающих геномных фрагментов, фланкированных инвертированными повторами микросателлитных локусов, при выраженных межвидовых отличиях по изменчивости геномных участков, фланкированных длинными концевыми повторами эндогенных ретровирусов [2], несмотря на тесную связь между формированием микросателлитов и ретротранспозонами [3]. Для выявления особенностей вариабельности данных генетических элементов у представителей трех классов (нескольких видов млекопитающих (Mammalia), пресмыкающихся (Reptilia) — прыткая ящерица (*Lacerta agilis*) и насекомых (Insecta) —

медоносная пчела (*Apis mellifica*), в настоящей работе выполнен сравнительный анализ полиморфизма геномных фрагментов, фланкированных инвертированными повторами последовательностей (СТС)6С, (GAG)6С и (ACC)6G с использованием полимеразной цепной реакции (ПЦР), получения продуктов амплификации (ампликонов) и расчета характеристик генетического разнообразия (полиморфного информационного содержания - PIC).

При сравнительном анализе электрофоретических профилей продуктов ПЦР, полученных с использованием в качестве праймеров микросателлитных повторов (СТС)6С, (GAG)6С и (ACC)6G, была выявлена устойчивая межвидовая закономерность: спектры двух первых праймеров характеризовались сниженным количеством выявляемых ампликонов (в среднем 4–7 на праймер) относительно последовательности (ACC)6G (в среднем 9–11 ампликонов). Наблюдаемая диспропорция в количестве ампликонов, сохраняющаяся у филогенетически удаленных групп, подчеркивает роль структуры повторов в формировании спектров продуктов амплификации.

Анализ PIC позволил получить следующие данные. Внутривидовая изменчивость у трёх пород и синтетического кросса домашнего кролика по праймеру (ACC)6G варьировала от 0,12 до 0,15, у крупного рогатого скота (КРС) от 0,07 до 0,28; по (СТС)6С – у домашнего кролика от 0,17 до 0,35, у КРС – от 0,09 до 0,30; по (GAG)6С у домашнего кролика от 0,11 до 0,22, у КРС – от 0,08 до 0,22.

Межвидовые значения PIC у других видов млекопитающих (куньи – соболь, норка; псовые – собака, волк, шакал; полорогие – лошадь, овца, муфлон) попадали примерно в тот же диапазон: по праймеру (ACC)6G – от 0,15 до 0,39; по (СТС)6С – от 0,09 до 0,38; по (GAG)6С – от 0,16 до 0,38. Сходные значения наблюдались и у ящериц – по (ACC)6G – 0,25, по (GAG)6С – 0,17, и у пчел – по (ACC)6G – 0,35, по (СТС)6С – 0,36, по (GAG)6С – 0,11.

Исследования Yun Xia et al. [4] показали, что триплетный мотив (ACC)_n занимает второе место по распространённости в зонах высокой плотности микросателлитов (HMDPs) генома человека. Среди триплетных повторов (ACC)_n доминирует вместе с мотивом (GGT)_n, причём его повышенная представленность в экзонах предполагает роль в стабильности белков или регуляции транскрипции. Распределение ACC-содержащих HMDPs коррелирует с GC-составом и некоторыми архитектурными характеристиками хроматина, что потенциально может подтверждать их эволюционную консервативность и функциональную значимость.

Согласно данным Toth et al. [5], тринуклеотидные повторы ACC характеризуются низкой частотой встречаемости в межгенных и интронных регионах у большинства таксонов, за исключением представителей отряда Rodentia. Важно отметить, что мотив ACC кодирует треонин (Thr) — незаменимую аминокислоту, играющую критическую роль в метаболических, иммунных и структурных процессах у позвоночных [6;7]. Высокая консервативность ACC-повторов в экзонах может быть связана с их функциональной значимостью, включая поддержание стабильности белковых структур и участие в посттрансляционных модификациях.

Проведённое исследование подтверждает, что вариабельность микросателлитных локусов, в частности, триплетных повторов (ACC)_n, определяется сложным взаимодействием филогенетически обусловленных видоспецифичных мутационных процессов, функциональных ограничений и структурных особенностей генома.

1. E.R.Moxon, C.Wills (1999) DNA microsatellites: agents of evolution? *Scientific American*, 280(1):94-99.
2. Д.В.Попов и др. (2022) Контроль гетерозиготности с применением ДНК-маркеров у сельскохозяйственных видов животных, Кролиководство и звероводство. 3:44-51.
3. В.И.Глазко и др. (2023) ДНК маркеры и «микросателлитный код» (обзор), *Сельскохозяйственная биология*. 58(2):223-248.

4. Y.Xia et al. (2024) Microsatellite density landscapes illustrate short tandem repeats aggregation in the complete reference human genome, BMC genomics. 25(1):960.
5. G.Tóth et al. (2000) Microsatellites in different eukaryotic genomes: survey and analysis, Genome research. 10(7):967-981.
6. Q.Tang et al. (2021) Physiological functions of threonine in animals: beyond nutrition metabolism, Nutrients. 13(8):2592.
7. T.L.McGaha et al. (2012) Amino acid catabolism: a pivotal regulator of innate and adaptive immunity, Immunological reviews. 249(1):135-157.

Relationship between mtDNA nucleotide composition and life cycle and environmental traits in Nematoda

V. Lobanova, B. Efimenko, K. Popadin

*Immanuel Kant Baltic Federal University, Kaliningrad, Russia,
valeriavlobanova@gmail.com
konstantinpopadin@gmail.com*

The mitochondrion is an organelle involved in cellular metabolism and synthesizes energy in the form of ATP molecules. It contains mitochondrial DNA encoding some of the respiratory chain proteins. Intensive cellular respiration is accompanied by the formation of reactive oxygen species (ROS), which are by-products of the oxidative phosphorylation. ROS molecules can interact with DNA and change its sequence as mutagens.

The Nematoda phylum includes organisms with a very wide range of environmental conditions and life cycle variations. The diversity of nematodes is so great that it is estimated at 20 million species, more than other types of Metazoa [1]. Free-living nematodes live in various ecological niches: soil, terrestrial, aquatic. Among them, there are

species capable of feeding on plants, bacteria (including reducents), and even other worms.

We hypothesized that the evolution of organisms under certain environmental conditions may influence metabolism and cellular respiration and, as a consequence, mutagenesis and changes in the nucleotide composition of mitochondrial DNA. We have analyzed 148 complete sequenced mitochondrial genomes of nematodes from the RefSeq database using NeMu pipeline [2] to obtain a mutational spectrum, in which the most frequent types of substitutions were C>T, complementary G>A and A>G.

We applied the principal component analysis to nucleotide composition of complete mtDNA sequences and separately for synonymous and fourfold degenerate positions of protein-coding genes. The analysis of the most neutral (fourfold degenerate) positions showed well the relationships between the PC1 component with the content of thymine and cytosine, and the PC2 component with the content of adenine and guanine, which correspond to the main transitions in the observed mutational spectrum.

Because complete mitochondrial genomes of nematodes are not sequenced abundantly, some groups by feeding type or ecological niche appeared to be represented by only a few species. Therefore, we also performed the same analyses on 1837 COX1 sequences from the MIDORI2 database.

By looking at species extreme in the content of any nucleotide, we hypothesized that the distribution of nematodes along the PC1 axis, respectively the thymine and cytosine content, is related to reproduction rate, fertility and metabolism. To confirm this hypothesis, we used a ranking of free-living nematodes on a “colonizers-persisters” scale, where “colonizers” with rank 1 are similar to r-strategists and have high fertility, small body size and short juvenile period, while “persisters” with rank 5 are similar to K-strategists and have low fertility, large body size and longer juvenile period [3]. The thymine content of rank 1 nematodes is statistically significantly higher than of rank 5 ($p = 0.0007$). We also performed a similar analysis using a Pearson correlation for PC1 and body length, which is one of the parameters that approximates

selection strategy and life cycle, among parasitic and free-living nematodes ($r = -0.38$, $p = 0.0004$). The increased thymine content and its opposite to cytosine content may be a consequence of frequently occurring C>T substitutions, which is one of the metrics for the rate of molecular evolution. This is supported by the lifestyle and morphological characterization of the nematodes.

Also, we hypothesized that the distribution of nematodes along the PC2 axis, respectively adenine and guanine content, might be related to habitat aerobicity. To confirm it, we compare PC2 values among groups pooled by feeding type and habitat (the final hosts for parasitic nematodes). Preliminary results show that free-living nematodes contain more adenine and less guanine, while parasitic nematodes are much more diverse in the distribution of adenine and guanine content. This may be due to different tissue oxygen saturation within host organisms, while for free-living nematodes the habitat oxygen content is relatively more constant.

Acknowledgments: this work is supported by RSF № 21-75-20143.

1. Hodda, M. (2022). Phylum Nematoda: a classification, catalogue and index of valid genera, with a census of valid species. *Zootaxa*, 5114(1), 1-289.
2. Efimenko, B., Popadin, K., & Gunbin, K. (2024). NeMu: a comprehensive pipeline for accurate reconstruction of neutral mutation spectra from evolutionary data. *Nucleic Acids Research*, 52(W1), W108-W115.
3. Bongers, T. (1999). The Maturity Index, the evolution of nematode life history traits, adaptive radiation and cp-scaling. *Plant and soil*, 212, 13-22.

От публикаций к системному пониманию (и перспективным направлениям): агрегация и анализ данных miRNA-дисрегуляции AGE/RAGE-сигналинга при метаболическом синдроме и коморбидных патологиях

Михаил Михайлович Лопатин, Мария Михайловна Бограя, София Сергеевна Воронова, Мария Александровна Вульф, Лариса Сергеевна Литвинова

*Балтийский федеральный университет им. И. Канта,
Россия, г. Калининград,
mikhaillopatin28@gmail.com*

Метаболический синдром (МС) представляет собой комплекс взаимосвязанных факторов риска сердечно-сосудистых заболеваний, сахарного диабета 2-го типа, неалкогольной жировой болезни печени (НАЖБП) и других патологий. Ведущую роль в патогенезе МС играет дисрегуляция ключевых сигнальных путей, среди которых особое место занимает ось AGE/RAGE. Активация данного пути происходит при связывании конечных продуктов гликирования (advanced glycation end products, AGEs) с рецептором RAGE, что запускает внутриклеточные каскады (NF- κ B, MAPK и др.), приводящие к хроническому воспалению, оксидативному стрессу и фиброзу [1]. В настоящее время система AGE/RAGE представляет интерес для рассмотрения как интегральное звено, объединяющее метаболические нарушения с их коморбидными осложнениями. Таргетинг этого сигнального пути признается перспективным направлением в терапии МС и ассоциированных заболеваний [2–4].

Нобелевская премия по физиологии и медицине 2024 года подчеркнула ключевую роль микроРНК в эпигенетической регуляции. Эти короткие некодирующие РНК функционируют, главным образом, как посттранскрипционные репрессоры генов. Некоторые микроРНК продемонстрировали способность модулировать передачу сигнала через AGE-RAGE-путь, влияя на воспалительные процессы, клеточную пролиферацию и

дифференцировку. Например, в экспериментах с сосудистыми эндотелиальными клетками, обработанными избытком глюкозы, миметик miR-126 снижал воспаление и оксидативный стресс за счёт прямого подавления miR-126 своей доказанной мишени HMGB1, что в свою очередь опосредовало ингибирование NADPH-оксидазы [5].

Несмотря на прогресс в изучении микроРНК в контексте МС и ассоциированных заболеваний [6], их роль в регуляции сигнального пути AGE/RAGE остаётся недостаточно исследованной. Существующие данные носят фрагментарный характер и, что критически важно, не учитывают коморбидность патологий. Это существенно ограничивает понимание интегративного влияния микроРНК на континуум нарушений, ассоциированных с МС.

Авторы решили провести агрегацию и анализ существующих экспериментальных данных о дисрегулированных микроРНК и их подтверждённых мишенях в сигнальном пути AGE/RAGE при МС и коморбидных заболеваниях с целью представить целостную картину их взаимодействий.

В исследовании применялся биоинформатический подход, выходящий за рамки узкого его понимания: как инструмента для анализа последовательностей или предсказания мишеней — что типично для многих работ в области микроРНК-исследований. Вместо этого акцент был сделан на интеграцию и синтез уже существующих эмпирических сведений в единую концептуальную модель.

В PubMed и Gene Expression Omnibus были отобраны 10 исследований микроРНК-профилирования для пациентов с МС и его коморбидностями (сахарный диабет 2 типа, атеросклероз, ожирение, НАЖБП, а также пародонтоз и синдром поликистозных яичников). Локализации образцов были релевантны патологиям (например, печень для НАЖБП, жировая ткань для ожирения), профилирование проводилось исключительно методом микрочипов (по причине приемлемого баланса между ошибками 1-го и 2-го рода, а также доступности данных).

Дифференциально экспрессируемые микроРНК (ДЭМ) были идентифицированы на основе данных, представленных в оригинальных исследованиях. Для каждой работы выгружались

списки микроРНК, которые авторы этих работ обозначили как статистически значимые ДЭМ (без единого порога logFC или p-value между списками) — для сохранения авторской интерпретации результатов и репрезентации текущего исследовательского поля. Учитывались как повышено-, так и понижено-экспрессируемые микроРНК. Экспериментально подтвержденные взаимодействия ДЭМ с генами-мишенями были получены из MiRTarBase v.9.0, затем отобраны гены, связанные с сигнальным путем AGE/RAGE (дополненный путь Wikipathways WP2324).

По результатам работы было собрано 304 уникальных ДЭМ, из которых 41 ДЭМ встречалась более чем в одной патологии-локализации, причем 25 ДЭМ имели схожую динамику (только повышены или только понижены) независимо от патологии и локализации, что указывает на их универсальную роль в регуляции AGE/RAGE. Обнаружено, что в каждой паре патология-локализация ось AGE/RAGE подвергается весомой дерегуляции, наибольшие значения наблюдались в атеросклеротических бляшках, где было зафиксировано 284 экспериментально подтвержденных взаимодействия с генами (miRNA-target interactions, MTI), при МС в мононуклеарах периферической крови (203 MTI) и при ожирении в подкожной жировой ткани (147 MTI). Выявлены общие регуляторные узлы для различных коморбидных состояний, например miR-92a-3p, ранее не рассматривавшаяся как связующее звено между патологиями, оказалась пониженной при ожирении, поликистозе и атеросклерозе. Полученные данные указывают на тканеспецифичный характер микроРНК-дерегуляции, сочетающийся с универсальными паттернами, что открывает перспективы для разработки комбинированных диагностических панелей.

Фокусирование на лиганды и рецепторы оси AGE/RAGE показало, что, наибольшей дерегуляции подвергается HMGB1 (37 MTI) и DDOST (10 MTI) — ген рецептора, конкурирующего с RAGE. По остальным сигнальным агентам сведения скудные, что указывает на необходимость заполнения «белых пятен». Наша работа систематизирует имеющиеся данные, выявляет пробелы и задает вектор для будущих исследований. Универсальность подхода позволяет применять его к другим сигнальным путям.

Исследование выполнено при финансовой поддержке ГДЗ
№ FZWM-2024-0012.

- 1 C.-Y. Shen, C.-H. Lu, C.-H. Wu, K.-J. Li, Y.-M. Kuo, S.-C. Hsieh, et al. The Development of Maillard Reaction, and Advanced Glycation End Product (AGE)-Receptor for AGE (RAGE) Signaling Inhibitors as Novel Therapeutic Strategies for Patients with AGE-Related Diseases. (2020) *Molecules*. 25, 5591.
- 2 K. Taguchi, K. Fukami RAGE signaling regulates the progression of diabetic complications. (2023) *Front. Pharmacol.* 14, 1128872.
- 3 R. Pal, S. K. Bhadada AGEs accumulation with vascular complications, glycemic control and metabolic syndrome: A narrative review. (2023) *Bone*. 176, 116884.
- 4 K. Gutowska, K. Czajkowski, A. Kuryłowicz Receptor for the Advanced Glycation End Products (RAGE) Pathway in Adipose Tissue Metabolism. (2023) *IJMS*. 24, 10982.
- 5 S. Tang, F. Wang, M. Shao, Y. Wang, H. Zhu MicroRNA-126 suppresses inflammation in endothelial cells under hyperglycemic condition by targeting HMGB1. (2017) *Vascular Pharmacology*. 88, 48–55.
- 6 F. Ramzan, M. H. Vickers, R. F. Mithen Epigenetics, microRNA and Metabolic Syndrome: A Comprehensive Review. (2021) *IJMS*. 22, 5047.

**Разработка и внедрение методов точной детекции коротких
тандемных повторов (STR) в российской популяции**

Е.И. Лукьянчикова¹, О.С. Киреева², Д.В. Пустошилов²

¹*Московский физико-технический институт (государственный
университет), Москва, Россия
lukianchikova.ei@phystech.edu*

²*ООО «Биотехнологический кампус», Москва, Россия*

Короткие tandemные повторы (STR) — повторяющиеся последовательности ДНК (2–6 п.н.), играющие критическую роль в регуляции генов, стабильности генома и развитии заболеваний, включая нейродегенеративные расстройства (болезнь Хантингтона, спиноцереbellарные атаксии), онкологию и эндокринные патологии. В рамках проекта "100000+Я" поставлена задача создания универсального биоинформатического пайплайна для высокоточной детекции STR в больших популяционных когортах с использованием данных NGS (MGI).

Основные этапы работы включали:

1. сравнение алгоритмов детекции (ExpansionHunter, STRipy, TideHunter), где STRipy выбран за возможность визуализации (тепловые карты, гистограммы) и интеграции с клиническими панелями (CODIS, BRCA);
2. анализ популяционной вариабельности через построение графиков распределения длин STR с привлечением данных gnomAD (мировые и нефинские когорты), что выявило уникальные паттерны в российской популяции, ассоциированном с неаутоиммунным гипотиреозом;
3. оценку генетической дифференциации методом AMOVA, показавшую повышенный вклад межпопуляционной изменчивости для STR ($F_{ST} = 0.08$ в генах HTT и FMR1) по сравнению с SNP, что подтверждает их ценность для этнических исследований.

Разработанный пайплайн интегрирован в платформу Wizard, обеспечив автоматизированную обработку образцов проекта. Среди ключевых находок — редкие аллели в ATXN1 и DMPK, связанные с ранними формами атаксии и миотонической дистрофией, отсутствующие в референсных базах.

Перспективы работы включают:

- углубленный анализ STR-профилей этнических групп России для выявления региональных генетических особенностей;

- внедрение моделей машинного обучения для прогнозирования патогенности STR.

Результаты проекта позволят создать новые диагностические панели, улучшить интерпретацию клинических данных и раскрыть роль STR в эволюции популяций.

Оценка влияния гексуроновых кислот на экспрессию генов ключевых регуляторов формирования биопленок у *E. coli*

А.Т. Магкаев^{1,2}, У.Д. Кузнецова^{1,3}, Т.А. Бессонова^{1,4,5},
М.Н. Тутукина^{1,5,6}

¹*Институт общей генетики им. Н.И. Вавилова РАН,*

²*НИУ ВШЭ, Москва, Россия, atmagkaev@edu.hse.ru*

³*РНИМУ им. Н.И. Пирогова, Москва, Россия*

⁴*Институт биофизики клетки РАН - ФИЦ ПНЦБИ РАН*

⁵*Институт проблем передачи информации РАН*

⁶*Сколковский институт науки и технологий*

Биоплёнки — это бактериальные сообщества, обладающие повышенной устойчивостью к антибиотикам и иммунным механизмам, которые способствуют развитию хронических и трудно поддающихся лечению инфекционных заболеваний. Ключевую роль в регуляции формирования биоплёнок *E. coli* играют белки-регуляторы, а также, по-видимому, сигнальные некодирующие РНК и метаболиты пути Эшвелла, однако их интегрированные механизмы до конца не изучены.

Цель работы - оценить влияние интермедиатов пути Эшвелла - D-галактурановой и D-глюкуроновой кислот — на экспрессию генов ключевых регуляторов формирования биопленок у *E. coli*.

Среди прочих, была поставлена задача анализа экспрессии генов, кодирующих белок-регулятор формирования биопленок и подвижности CsgD и некодирующие РНК SsrS и CsrC в

планктонных формах бактерий *E. coli* выращенных при 30°C и при 37°C и бактериях, образующих биопленку при 37°C, при росте на различных источниках углерода, и поиск корреляции с биопленкообразованием.

Из бактерий, растущих в указанных условиях была выделена тотальная РНК, из которой путем обратной транскрипции с помощью M-MuLV RevertAid (Thermo) была получена кДНК. Реакция количественной ПЦР проводилась на амплификаторе DT-lite (“ДНК-Технология” Россия) с использованием qPCRMix-HS SYBR (“Евроген”, Россия). Контролем служил результат обратной транскрипции пробы с РНК без фермента. Количественный анализ уровня экспрессии производился при помощи метода 2– $\Delta\Delta C_t$ относительно *Escherichia coli*, выращенной при 30°C в присутствии глюкозы в качестве источника углерода. На основе этих данных были построены столбчатые диаграммы при помощи языка программирования Python в среде Google Colab.

Из столбчатой диаграммы на рисунке 1А видно, что экспрессия *csgD* в планктонных формах бактерий при росте на глюкозе и галактуронате практически не зависит от температуры. При 37°C экспрессия активируется при росте на D-галактуронозойной кислоте, тогда как в случае биопленок в присутствии D-галактуронозойной кислоты она понижается. На основе диаграммы на рисунке 1В видно, что экспрессия *ssrS* в планктонных клетках не зависит ни от температуры, ни от источника углерода, тогда как в биопленке она была существенно активирована при росте на D-галактуронозойной кислоте.

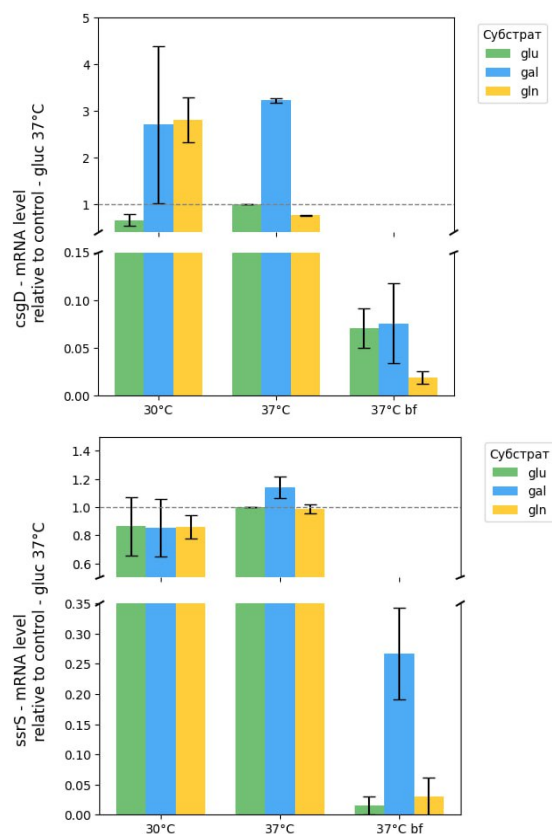


Рис 1. (А – для белка-регулятора CsgD, В – для некодирующей РНК SsrS) Столбчатые диаграммы, демонстрирующие результаты анализа экспрессии исследуемых генов методом 2-ΔΔCt. Цветами указаны субстраты. Условия роста указаны под столбцами.

Выводы:

Гексуроновые кислоты по-разному модулируют экспрессию ключевых регуляторов биопленкообразования *E. coli* K-12.

D-галактуриновая кислота усиливает экспрессию *csgD* и *ssrS*, что может усиливать формирование биоплёнки и согласуется с данными о необходимости галактуриновой кислоты для успешной колонизации организма хозяина клетками *E. coli*. D-глюкуроновая

кислота, напротив, снижает экспрессию регуляторов и, вероятно, ингибирует биопленкообразование - что также согласуется с ранее полученными нами данными.

Планы на будущее:

В дальнейших экспериментах мы планируем проанализировать экспрессию других генов — регуляторов биопленкообразования у *E. coli*, а также клонировать некодирующие РНК SsrS и CsrC в низкокопийные плазмиды. Это позволит проверить, как сверхэкспрессия этих РНК влияет на подвижность, формирование биопленок и экспрессию ключевых регуляторных генов.

Работа поддержана РНФ 24-24-00435.

1. T.A.Bessonova, U.D. Kuznetsova, A.T. Magkaev et al. (2024) Connection between Hexuronate Metabolism and the Ability of Escherichia coli to Adhesion and Biofilm Formation, Microbiology, 93:497–501. <https://doi.org/10.1134/S002626172460472X>
2. M.N. Tutukina, A.I. Dakhnovets, A.D. Kaznadzey, M.S. Gelfand and O.N. Ozoline (2023), Sense and antisense RNA products of the *uxuR* gene can affect motility and chemotaxis acting independent of the UxuR protein, Front. Mol. Biosci. 10:1121376.doi: 10.3389/fmolb.2023.1121376

Полногеномный анализ для оценки безопасности пробиотических штаммов семейства *Bacillaceae*

К.А. Максимова, М.И. Маркелова, М.Н. Синягина, Д.Р.
Камальдинова, Т.В. Григорьева

*Казанский Федеральный Университет, Казань, ул. Кремлёвская,
18, amksusha44@mail.ru*

Пробиотики представляют собой живые микроорганизмы, которые приносят пользу здоровью хозяина при употреблении в

достаточном количестве. В настоящее время наблюдается значительный интерес к использованию пробиотиков в различных отраслях, таких как пищевая и косметологическая промышленности, фармацевтика, сельское хозяйство. Наиболее часто используемыми пробиотиками для здоровья человека являются представители родов *Lactobacillus* и *Bifidobacterium*, однако в последние годы интерес к использованию бактерий семейства *Bacillaceae* значительно возрос благодаря их уникальным свойствам.

Так, способность различных видов *Bacillaceae* к спорообразованию позволяет им проявлять высокую устойчивость к различным условиям окружающей среды, таким как нагревание, изменение pH, высушивание, что делает их стабильными во время обработки и хранения пищевых и фармацевтических препаратов. Также представители *Bacillaceae* отличаются способностью продуцировать антимикробные вещества, подавляющие рост патогенных микроорганизмов, что даёт возможность использовать их для коррекции микробиоты. Кроме того, *Bacillaceae* могут оказывать противовоспалительное и иммуномодулирующее действия и поддерживать целостность и функционирование кишечного и кожного барьера [1].

Однако, несмотря на пользу, некоторые бактерии способны представлять риск для здоровья, поэтому чрезвычайно важно производить оценку безопасности микроорганизма, который потенциально можно использовать в качестве пробиотика. В данной работе с помощью биоинформатического анализа генома было проведено исследование безопасности трёх штаммов потенциальных кожных пробиотиков: *Bacillus subtilis* 3KD, *Bacillus velezensis* 7RV и *Priestia megaterium* 14SM, выделенных с кожи предплечья здоровых добровольцев.

Была выделена ДНК микроорганизмов, подготовлены библиотеки с использованием набора NEBNext Ultra II и произведено секвенирование на приборе GeneMind FASTASeq 300 в режиме 2*150 п.о. Была произведена оценка качества полученных прочтений с помощью FastQC v0.12.1 [2], тримминг адаптеров и

низкокачественных прочтений был произведен с помощью Cutadapt v4.4 [3]. Сборка проводилась с использованием алгоритма SPAdes v3.14.0 [4], полнота и чистота сборки оценивалась при помощи CheckM v1.0.13 [5], затем геномы были проаннотированы инструментом Prokka v1.14.6 [6]. Геном штамма *B. subtilis* 3KD составляет 3997095 п.н. с содержанием ГЦ 43,76%; штамма *B. velezensis* 7RV – 3918980 п.н. с содержанием ГЦ 46,46%; штамма *B. megaterium* 14SM – 5069076 п.н. с содержанием ГЦ 38,01%.

Для идентификации генов вирулентности и токсинов, присутствующих в геноме микроорганизмов, была использована база данных VFDB [7]. В штаммах *B. velezensis* 7RV и *P. megaterium* 14SM не было обнаружено генов вирулентности, соответствующих критериям >80% идентичности и >60% покрытия, однако в штамме *B. subtilis* 3KD был найден ген *bslA* (*yuaB*). Он кодирует белок BslA (Biofilm Surface Layer A), который играет ключевую роль в формировании биоплёнок. Биоплёнки выполняют функцию защиты бактериальных колоний от внешних воздействий, и в случае патогенных микроорганизмов они будут затруднять лечение. Однако при отсутствии других генов вирулентности наличие биоплёнки в пробиотиках позволяет бактериям адаптироваться к неблагоприятным условиям окружающей среды и конкурировать с другими микроорганизмами, поэтому наличие этого гена не было определено как фактор небезопасности штамма.

Затем был произведён поиск генов антибиотикорезистентности. Для этого использовались базы данных CARD [8] и ResFinder [9]. В результате было найдено 12 генов антибиотикорезистентности у *B. subtilis* 3KD (*aadK*, *blt*, *bmr*, *rphB*, *ykkD*, *ykkC*, *vmlR*, *tmrB*, *lmrB*, *mphK*, *mprF*, *tetL*) и 1 ген антибиотикорезистентности у *B. velezensis* 7RV (*clbA*); в геноме бактерии *P. megaterium* 14SM генов антибиотикорезистентности обнаружено не было. Каждый из найденных генов имел идентичность более 98% и покрытие более 97%, кроме гена *rphB* с идентичностью равной 80,44%.

Наличие генов антибиотикорезистентности представляет опасность тем, что они могут передаваться другим бактериям в

организме-хозяине. Поэтому для того, чтобы установить возможность горизонтального переноса генов антибиотикорезистентности, нами был произведён поиск мобильных элементов в геноме бактерий *B. subtilis* 3KD и *B. velezensis* 7RV. Наличие профагов и плазмид было проверено с помощью баз данных PHASTEST [10] и PlasmidFinder [11] соответственно. В результате анализа в геноме бактерий не было обнаружено плазмид; в *B. subtilis* 3KD был найден один регион, соответствующий профагу, полнота которого была отмечена как сомнительная, а в *B. velezensis* 7RV – 2 интактных региона. Затем был произведён поиск инсерционных элементов (IS) при помощи базы данных ISfinder [12], и было обнаружено 5 IS в геноме *B. subtilis* 3KD и 23 IS в геноме *B. velezensis* 7RV ($E < 0,0001$). Хотя в геноме бактерий были найдены IS и профаги, ни один из этих элементов не был обнаружен вблизи генов антибиотикорезистентности, что позволяет нам сделать вывод о том, что вероятность их передачи другим бактериям невелика. Кроме того, мы определили при помощи базы данных ICEfinder [13], что в геноме бактерий отсутствуют интегративные конъюгативные элементы.

Таким образом, во всех штаммах исследованных бактерий не было найдено генов вирулентности, представляющих опасность для организма человека. Несмотря на то, что гены антибиотикорезистентности присутствовали в геноме штаммов *B. subtilis* 3KD и *B. velezensis* 7RV, ни один из них не является транспортируемым. Таким образом, на основе анализа полных геномов исследуемые штаммы могут считаться безопасными для человека.

1. J. Payne, D. Bellmer, R. Jadeja, P. Muriana (2024) The Potential of *Bacillus* Species as Probiotics in the Food Industry: A Review, *Foods* (Basel, Switzerland), 13(15):2444.
2. S. Andrews (2010) FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]

3. M. Martin (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads, *EMBnet.journal*, 17(1):10–12.
4. A. Prjibelski, D. Antipov, D. Meleshko, A. Lapidus, A. Korobeynikov (2020) Using SPAdes De Novo Assembler, *Current protocols in bioinformatics*, 70(1):e102.
5. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes, *Genome research*, 25(7):1043–1055.
6. T. Seemann (2014) Prokka: rapid prokaryotic genome annotation, *Bioinformatics (Oxford, England)*, 30(14):2068–2069.
7. S. Zhou, B. Liu, D. Zheng, L. Chen, J. Yang (2025) VFDB 2025: an integrated resource for exploring anti-virulence compounds, *Nucleic acids research*, 53(D1):D871–D877.
8. B.P. Alcock et al. (2023) CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database, *Nucleic acids research*, 51(D1):D690–D699.
9. V. Bortolaia et al. (2020) ResFinder 4.0 for predictions of phenotypes from genotypes, *The Journal of antimicrobial chemotherapy*, 75(12):3491–3500.
10. D. S. Wishart, S. Han, S. Saha, E. Oler, H. Peters, J. R. Grant, P. Stothard, V. Gautam (2023) PHASTEST: faster than PHASTER, better than PHAST, *Nucleic acids research*, 51(W1):W443–W450.
11. A. Carattoli, E. Zankari, A. García-Fernández, M. Voldby Larsen, O. Lund, L. Villa, F. Møller Aarestrup, H. Hasman (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing, *Antimicrobial agents and chemotherapy*, 58(7):3895–3903.
12. P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, M. Chandler (2006) ISfinder: the reference centre for bacterial insertion sequences, *Nucleic acids research*, 34:D32–D36.
13. M. Liu, X. Li, Y. Xie, D. Bi, J. Sun, J. Li, C. Tai, Z. Deng, H. Y. Ou (2019) ICEberg 2.0: an updated database of bacterial integrative and conjugative elements, *Nucleic acids research*, 47(D1):D660–D665.

Исследование СЭ локусов в областях формирования Z-ДНК

Ю.В. Макус¹, Г.А. Ашниева^{2,3}, А.В. Орлов⁴

¹*Национальный исследовательский университет «Высшая школа экономики», ул. Мясницкая, 20, 101000 Москва, Россия,
ymakus@yandex.ru*

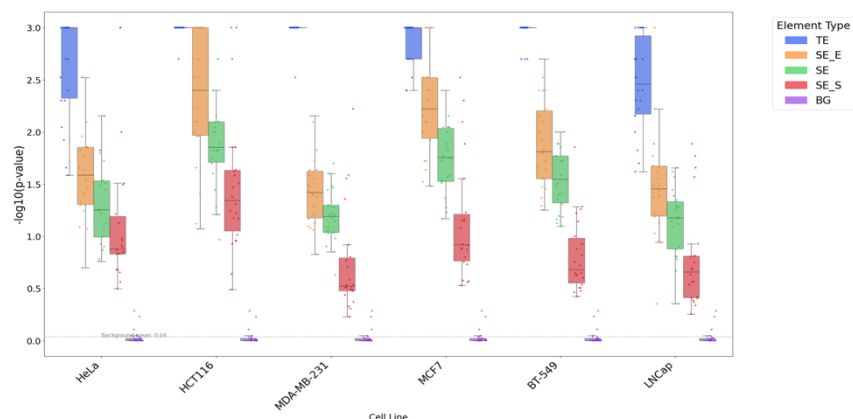
²*Институт проблем передачи информации РАН, Большой
Каретный пер., д. 19, стр. 1, 127051 Москва, Россия*

³*Институт общей генетики РАН, ул. Губкина, 3, 119991 Москва,
Россия*

⁴*Институт общей физики им. М.Н. Прохорова РАН, ул. Вавилова,
д. 38, 119991 Москва, Россия*

Исследование направлено на изучение функционального взаимодействия супер-энхансеров (СЭ) и Z-ДНК в регуляции транскрипции при злокачественной трансформации клеток. Супер-энхансеры, кластеры энхансеров с высокой плотностью транскрипционных факторов [1], и Z-ДНК, левозакрученная конформация ДНК, ассоциированная с активными регуляторными регионами [2], предположительно кооперируют в модуляции экспрессии генов. Для анализа их колокализации использованы предсказанные координаты Z-ДНК (Z-DNABERT) и СЭ шести эпителиальных опухолевых клеточных линий (HeLa, HCT116, MDA-MB-231, MCF7, BT-549 и LNCap). Перестановочный тест подтвердил неслучайный характер пересечений Z-ДНК с энхансерами, что свидетельствует об их функциональной взаимосвязи (рис. 1).

Рисунок 1 – Боксплоты и точечные даиграммы распределения – $\log_{10}(\text{p-value})$, полученные в результате пермутационных тестов.



Выявлены функционально консервативные супер-энхансеры, активные исключительно в опухолях, имеющие пересечения с предсказаниями Z-ДНК. Эти регионы ассоциированы с ключевыми онкогенными процессами: нарушением клеточного цикла, поддержанием плюрипотентности и эпигенетическим ремоделированием. Среди значимых транскрипционных факторов выделены CLOCK(bHLH) и Zscan4c(Zf), регулирующие 9 молекулярных путей, включая метаболическое репрограммирование и сигналинг. Анализ мутаций в локусах пересечения СЭ и Z-ДНК (данные 1000 Genomes) выявил их статистически значимое обогащение в регуляторных регионах ($p < 10^{-73}$, тест Манна-Уитни), что подтверждает влияние эволюционного отбора на функциональность этих элементов.

Результаты демонстрируют двойную роль Z-ДНК в составе СЭ: как платформы для рекрутирования транскрипционных факторов и как структурного элемента, формирующего транскрипционно-активные хроматиновые домены. Обнаружена согласованная активация консервативных СЭ в различных эпителиальных опухолях, что указывает на универсальные механизмы перестройки регуляторных сетей при канцерогенезе.

1. A. Claringbould, JB. Zaugg (2021) Enhancers in disease: molecular basis and emerging treatment strategies, Trends Mol Med, 27(11):1060-1073.
2. G. Wang, KM. Vasquez (2007) Z-DNA, an active element in the genome, Front Biosci, 12:4424-38.

Comparative Genomics of Bats in the Context of Aging Research

Leonid Malaev, Anna Rybina, Evgeniy Efimov, Dmitrii Kriukov,
Ekaterina Khrameeva

Skolkovo Institute of Science and Technology, Moscow, Russia

Representing approximately 20 % of mammalian diversity [1], Chiroptera exhibits remarkable variation in lifespans and this phenomena remains largely unexplored. A comprehensive comparative genomic analysis was conducted on 17 publicly available bat genomes, focusing on genes associated with aging and the extracellular matrix in order to identify biological mechanisms underlying bats' exceptional longevity. A custom bioinformatic pipeline was developed specifically for this analysis, which is suitable for the robust identification of genes experiencing positive selection across closely-related species. The pipeline incorporates modern HYPHY tools such as BUSTED-E [2] in order to avoid the false positive results usually present in such studies. An addition to the HYPHY toolkit was introduced specifically for this study as now root-to-tip values can be calculated easily using HYPHY MG94 model with new 'lineage' option.

A suite of genes under positive selection in long-lived bat species was identified. Enrichment analysis demonstrated that regulators of stem cell exhaustion and secreted extracellular matrix factors constitute the most significantly over-represented categories among these loci.

Most of the extracellular matrix genes remain highly conserved among different bats, yet secreted factors experience positive selection

more often compared to most other gene categories which makes them a promising target for the deeper investigation.

Stem cell exhaustion is an important aging hallmark and the extensive amount of positively selected genes associated with it suggest that long-living bats might have improved tissue maintenance. However, further research is required to fully assess the impact that these genes might have on lifespans of bats.

The identification of positive selection in epigenetic regulator genes such as *Eu-HMTase1* suggests that bat epigenetics requires close inspection, while positive selection of the gene *BNIP3L* (*BCL2/adenovirus E1B 19 kDa protein-interacting protein 3-like*) raises questions regarding whether long-lived bats have enhanced immunity and viral resistance—traits already recognized as hallmarks of bat biology [3]. Taking into account other identified positively selected genes such as, for example, *Wnt-3a* and *L-VEGF*, it seems that long-living bats may also have improved tissue homeostasis. We do not confidently claim that these biological traits are the reasons behind bat longevity but we suggest that they are promising prospects for future research.

These results establish a foundation for the identification of candidate genetic mechanisms underlying exceptional bat longevity. With this study we laid the groundwork for further research as obtained results allow us to focus attention on the key biological processes that show variance in bats for the forthcoming studies instead of desperately trying to analyze everything all at once.

The pipeline developed for this study is universal and it can be applied to any other subset of species. It allows to find signatures of positive selection and check the association of the *dN/dS* values with LQ (longevity quotient) with the least amount of technical issues as currently possible.

Acknowledgements: I want to thank professor Sergei Pond for the help with the pipeline creation.

1. I. Agnarsson et al. (2011) A time-calibrated species-level phylogeny of bats (chiroptera, mammalia), *PLoS Curr.*, 3:RRN1212

2. A. Selberg et al. (2024) Minus the error: Estimating dn/ds and testing for natural selection in the presence of residual alignment errors, bioRxiv
3. A. Banerjee et al. (2020) Novel insights into immune systems of bats. *Frontiers in Immunology*, 11

Графовый анализ макромолекулярных взаимодействий

А.Д. Малышев¹, А.А. Жарикова^{1,2}, А.А. Миронов^{1,2}

¹*Московский государственный университет им. М.В. Ломоносова, факультет биоинженерии и биоинформатики, Ленинские горы, 1, стр. 73, Москва, 119992, Россия, malyshev.a.d@yandex.ru*

²*Институт общей генетики им. Н.И. Вавилова РАН, ул. Губкина, 3, Москва, 119991*

Количество длинных некодирующих РНК (днкРНК) сопоставимо с количеством белок-кодирующих генов, но функции и механизмы действия установлены лишь для небольшого их числа, и даже эта информация позволяет понять, что днкРНК принимают участие практически во всех процессах жизнедеятельности клетки. Одним из подходов к изучению биологии днкРНК является анализ их взаимодействий с другими макромолекулами, такими как ДНК и белки.

Разработано большое число экспериментальных протоколов, направленных на выявление взаимодействий между молекулами различных типов. Анализ полученных таким образом данных может рассматриваться как задача на графе, где вершины – молекулы ДНК, РНК или белков, а рёбра – известные для них парные контакты. Преимущества такого подхода заключаются, во-первых, в возможности выявлять взаимодействия более высокого порядка, то есть находить функциональные комплексы; во-вторых, в учёте не только прямых, но и опосредованных контактов; и, в-третьих, в интеграции всей доступной информации для выявления

эмерджентных свойств системы, причём важно учитывать артефакты, возникающие из-за гетерогенности, неполноты и высокого уровня шума в данных. При этом становится возможным применение классических алгоритмов и машинного обучения на графах для решения задач кластеризации, предсказания свойств вершин и поиска новых связей. В частности, интерес представляют графовые нейронные сети, позволяющие учитывать информацию как о свойствах самой вершины, например, днкРНК, так и о её взаимодействиях различного типа.

В данной работе решена задача интеграции и приведения к единому формату экспериментальных данных, полученных в рамках различных экспериментов. Для линии клеток K562 мы использовали данные RIC-seq и KARR-seq (РНК-РНК взаимодействия), различные протоколы в рамках проекта ENCODE (главным образом eCLIP и RIP-seq, РНК-белковые контакты), а также базы данных белок-белковых взаимодействий: BioGRID, STRING, IntAct. Важно отметить, что для протоколов типа RIC-seq [1] разработан собственный программный конвейер, позволяющий производить более эффективную обработку данных.

Для построенного графа мы провели статистический анализ, позволивший выявить характерные особенности взаимодействий вершин различного типа, в частности различные паттерны взаимодействия с локусами хроматина для РНК и белков. Так, белки преимущественно взаимодействуют с энхансерами и сайтами начала транскрипции, а РНК – с молчащим, транскрибируемым и репрессированным хроматином. Для графа в целом показаны свойства, известные для графов белок-белковых взаимодействий: малый диаметр и распределение степени вершин, напоминающее таковое для безмасштабных сетей [2]. Использование GO-аннотации для белков позволило показать связь кластеров в графе с отдельными биологическими процессами: транскрипции, сплайсинга, регуляции состояния и конформации хроматина. Выявленные недостатки и артефакты, связанные с гетерогенностью, неполнотой и высоким уровнем шума в исходных данных, остаются материалом для дальнейшей работы.

Работа финансировалась РНФ грант №23-14-00136.

1. Cai, Z. et al. (2020) RIC-seq for global in situ profiling of RNA–RNA spatial interactions, *Nature*, 582:432–437.
2. Meng, X., Li, W., Peng, X., Li, Y. & Li, M. (2021) Protein interaction networks: centrality, modularity, dynamics, and applications, *Frontiers of Computer Science*, 25.

Использование карты электроспектральной настройки каротиноида для рационального дизайна каротинопротеинов

А.А. Мамчур, М.М. Сурков, И.А. Ярошевич

*Биологический факультет, МГУ им. М.В. Ломоносова, Москва,
Россия, al.mam4ur@yandex.ru*

Каротиноиды – это класс органических пигментов, выполняющих разнообразные биологические функции, в том числе, использующихся фотосинтезирующими организмами для регуляции фотохимических процессов. Оптические свойства каротиноидов, зачастую определяющие их биологическую роль, в значительной степени зависят от двух основных параметров: химического состава каротиноида и влияния его молекулярного окружения. В частности, близость заряженных групп локальном окружении каротиноида приводит к значительному сдвигу их полосы поглощения, что указывает на широкие возможности данных пигментов к электроспектральной настройке (ЭСН). Данная работа посвящена оценке влияния внешнего электростатического потенциала на спектральные свойства каротиноидов. Для достижения цели был использован подход с построением карты ЭСН, описанный в статье Kabir et al. [1], позволяющий оценить взаимосвязь между положением заряда относительно хромофора и направлением и силой сдвига оптического спектра поглощения.

Вокруг предварительно оптимизированной в вакууме структуры кантаксантина (PBE0/DEF2-TZVPP), создана

поверхность из 313 точек, расположенных на расстоянии радиуса Ван-дер-Ваальса от атомов исследуемой молекулы. В одну из точек на поверхности помещался точечный заряд $+0.1e$ («позитивная» карта), после чего проводился расчёт энергии оптически активного $S_0 \rightarrow S_2$ ($1Ag \rightarrow 1Bu$) перехода в присутствии данного заряда (CAM-B3LYP-D3/cc-pVTZ). Относительное изменение энергии оптического перехода для каждой из 313 уникальных позиций и составляет карту ЭСН. Аналогичная процедура построения карты проводилась для точек с зарядом $-0.1e$ («негативная» карта).

Полученная карта ЭСН показала, что эффект заряженных точек в окружении каротиноида на его спектральные свойства значительно зависит от положения точки относительно хромофора. Помещение положительного заряда в область бета-иононовых колец приводит к сдвигу спектра поглощения в более красную область (батохромный сдвиг), тогда как расположение заряженной точки рядом с π -сопряжённой цепью – к сдвигу в синюю область (гипсохромный сдвиг). «Негативная» карта продемонстрировала обратный эффект, однако не была полным отражением «позитивной».

Карта ЭСН далее будет использована для рационального дизайна каротинопротеинов – белков, нековалентно связанных с каротиноидами. Одним из них является оранжевый каротиноидный белок (orange carotenoid protein, OCP), играющий важную роль в фотосинтетическом аппарате цианобактерий и ответственный за нефотохимическое тушение. OCP является водорастворимым фоторецептором, что делает его интересным для прикладных задач синтетической биологии, в которых он может выступать в качестве оптического фотопереклювателя. Точечный мутагенез OCP позволит создать набор разнообразных переключателей, отличающихся максимумами поглощения.

Взаимодействие белковой части OCP с лигандом-каротиноидом обеспечивается, но не ограничивается наличием водородных связей кето-группы хромофора с аминокислотами W288 и Y201. Более того, анализ кристаллических структур OCP указывает на существование градиента электростатического потенциала, наведенного аминокислотными остатками белка [2]. Положительный полюс этого градиента расположен в С-доме белка (в окрестности упомянутых выше водородных связей) и

отрицательный в противоположном N-домене. Предполагается, что именно наведённый электростатический потенциал и его изменение в процессе фотоактивации может являться ключевым механизмом, лежащим в основе спектральных свойств всего фоторецептора.

В связи с описанными свойствами оранжевого каротиноидного белка логично предположить, что внесение точечных мутаций в области, отмеченные как наиболее чувствительные к изменению потенциала на карте ЭСН, является перспективным подходом к созданию каротинопротеинов с заданными свойствами.

Исследование выполнено за счёт гранта Российского научного фонда №25-74-20001.

1. M.P. Kabir, Y. Orozco-Gonzales, S.Gozem (2019) Electronic spectra of flavin in different redox and protonation states: a computational perspective on the effect of the electrostatic environment, *Phys. Chem. Chem. Phys.*, **21**:16526–16537.
2. Y. Mori (2016) Computational study on the color change of 3'-hydroxyechinenone in the orange carotenoid protein, *Chem. Phys. Lett.* **652**: 184–189.

Population Prevalence and Heritability of Mitochondrial DNA Heteroplasmy

Saleem Mansour^{1,2}, Layal Shaheen^{1,2}, Michil Trofimov¹, Dmitrii Kharitonov^{1,3}, Artem Leushin^{1,4}, Aleksandra Denisova^{1,5}, Alexander Rakitko^{1,2,3,4}

¹*Genotek Ltd., Moscow, Russia, mansours456@gmail.com*

²*Phystech school of Biological and Medical physics, Moscow Institute of Physics and Technology, Moscow, Russia*

³*ITMO University, Saint-Petersburg, Russia*

⁴*National Research University Higher School of Economics, Russian Federation*

⁵*Lomonosov Moscow State University, Moscow, Russia*

Background

Heteroplasmy is defined as the presence of more than one mitochondrial DNA (mtDNA) genotype within an individual, tissue, cell, or mitochondrion. It arises from the coexistence of mutant and wild-type mtDNA molecules, resulting in a mixed mtDNA population within a mitochondrial network. Previous studies have demonstrated that heteroplasmy occurs in somatic cells, tends to accumulate with age, and can also be maternally inherited. These properties underscore the importance of exploring heteroplasmy dynamics further, particularly their potential utility for improved estimation of matrilineal relationships.

Methods

Mitochondrial DNA from 16,451 samples was extracted following Good Laboratory Practices, sequenced, and processed using Next-Generation Sequencing instrumentation according to established best practices and quality control guidelines. Raw sequencing data (BCL files) were converted to FASTQ format, aligned to the Revised Cambridge Reference Sequence (rCRS) of the human mitochondrial genome using the BWA-MEM aligner. Variant calling was performed using GATK's Mutect2 in mitochondrial mode. Called variants were filtered for strand bias, retaining only variants with forward/reverse strand read ratios between 0.2 and 0.8. Sites with sequencing depth below 40× were excluded. Variant calling was additionally performed using Mutserve for validation; variants showing allele fraction differences greater than 10% between methods were excluded. To focus specifically on true heteroplasmic variants, we retained only variants with allele fractions between 10% and 90%. Applied thresholds were selected empirically and may require further optimization.

Results

Accurate detection of heteroplasmy largely depends on sequencing depth; thus, for samples with intermediate sequencing coverage (between 40× and 400×), we applied a threshold of 10% for defining heteroplasmic variants. Our analysis shows that 10807 out of 16,451 (35%) studied individuals carry at least one heteroplasmic position at this threshold. Multiple mtDNA positions showed elevated frequencies of heteroplasmy, suggesting hotspot regions potentially

affected by sequencing artifacts related to homopolymeric stretches or indel calling. These positions, therefore, require cautious interpretation in heteroplasmy analyses. Several identified hotspots align with previously reported problematic regions (“caution zones”), including positions 302–310, 500–510, 567, and 16182–16193.

Previous literature has reported the heritability of heteroplasmy, influenced by mitochondrial bottleneck dynamics. Among 346 mother-child pairs analyzed, we identified 169 heteroplasmic mutations in both mothers and offspring, with 35 mutations shared between mothers and offspring. The correlation in heteroplasmy levels between mother-offspring pairs was statistically significant ($R^2 = 0.68$, $p < 0.0001$).

Conclusions

This study highlights the prevalence, patterns, and inheritance dynamics of mitochondrial heteroplasmy in human samples. While heteroplasmic variants above the 10% threshold are observed in a minority of individuals (35%) Our results reinforce prior evidence that heteroplasmy can be maternally inherited, with significant correlation observed between mother-child pairs.

Структурные основы каталитической активности метилтрансферазы THUMPD2

С. С. Марьясина¹, Е.А. Пахомова², В.И. Польшаков¹, П.В. Сергиев¹

¹*МГУ имени М.В. Ломоносова, Россия, г. Москва, Ленинские горы,
1с73, sofia.mariasina@yandex.ru*

²*Школа на Юго-Востоке имени Маршала Чуйкова, Россия, г.
Москва, Таможенный проезд, 4*

THUMPD2 представляет собой метилтрансферазу, осуществляющую модификацию m2G72 в мяРНК U6 — важном компоненте сплайсосомы. Данная модификация критически влияет на эффективность сплайсинга пре-мРНК у млекопитающих. Как и многие другие метилтрансферазы, THUMPD2 использует S-аденозил-L-метионин (SAM) в качестве донора метильной группы.

Кроме того, для проявления каталитической активности ферменту требуется белок-партнер TRMT112.

В ходе исследования были определены ключевые аминокислотные остатки THUMPD2, участвующие в связывании с SAM и TRMT112. Методами молекулярной динамики (GROMACS) проведены симуляции нативной и мутантных форм белка. Анализ траекторий позволил установить три критических остатка (T311, D329 и S357), обеспечивающих стабилизацию SAM в активном центре. Кроме того, выявлена значимость остатка D386 для образования комплекса с TRMT112.

Данные расчетов сопоставили с экспериментальными данными по активности мутантных вариантов THUMPD2, экспрессированными в клеточной культуре. Сравнительный анализ выявил хорошую корреляцию между предсказанными *in silico* структурными изменениями и экспериментально измеренным уровнем каталитической активности. Полученные данные позволяют установить причинно-следственную связь между специфическими аминокислотными заменами и снижением функциональности фермента на молекулярном уровне."

Работа выполнена при финансовой поддержке РФФИ, проекты 24-14-00081 и 24-74-00052.

Temperature-mediated virulence shifts in *Verticillium alfalfae*: mutant phenotyping and whole-genome analysis

A. Meshcherov¹, L. Sidorov¹, S. Boldyrev^{1,2}, A. Baik¹, L. Gentzbittel³,
C. Ben³

¹*Skolkovo Institute of Science and Technology, Moscow, Russia,
a.meshcherov@skoltech.ru*

²*LLC Oil Gene, Russia*

³*Université de Toulouse, Toulouse INP, CNRS, IRD, CRBE, Toulouse,
France*

Verticillium wilt is a devastating fungal, soil-borne, infectious disease that leads to significant losses across a wide range of crops. One of the most dangerous members of the *Verticillium* genus is *Verticillium*

alfalfae, the major causative agent of Verticillium wilt in alfalfa (*Medicago sativa*) and its wild relative *Medicago truncatula*. In previous studies utilizing an experimental evolutionary approach, it was revealed that *V. alfalfae* exhibits remarkable adaptation to thermal environmental conditions that is associated with increased virulence (1). The general aim of this study is to investigate how such thermal adaptation influences genomic composition and fungal phenotypic traits, such as the secretion of extracellular enzymes associated with fungal virulence and aggressiveness.

Twenty-three strains from a warm-temperature-adapted mutant collection of *V. alfalfae* (and wild-type) (1) that differed most in their ability to infect *M. truncatula* plants with varying resistance levels to wilt were phenotyped in depth for traits associated with fungal fitness, aggressiveness, and virulence, such as colony growth rate, sporulation rate, and extracellular enzymatic activities responsible for degradation of plant cell components (2, 3): cellulase, polygalacturonase, lipase, and protease. We found that, despite sharing a common wild-type ancestor, the analyzed strains showed significant differences in sporulation rate, polygalacturonase, and cellulase activities. To understand the genetic background of the observed phenotypic diversity, we performed whole-genome sequencing of the five most phenotypically contrasting strains (including the wild type) and identified unique SNPs associated with fungal metabolism, which may explain the observed phenotypic diversity.

Identified candidate SNPs likely driving these adaptive shifts offer the first genetic clues into thermal adaptation in *V. alfalfae*. These results lay a crucial foundation for forecasting how climate change may influence the epidemiology of Verticillium wilt. In future work, we will extend our analysis to the full set of 120 temperature-adapted strains, enabling genome-wide association studies (GWAS) to systematically pinpoint genomic regions linked to thermal adaptation and virulence. Such insights are essential for developing climate-resilient strategies to protect crops.

1. Sbeiti A. A. L. et al. (2023) Temperature increase modifies susceptibility to Verticillium wilt in *Medicago* spp and may contribute to the emergence of more aggressive pathogenic strains, *Frontiers in Plant Science*, 14:1109154.

2. Kubicek C. P. et al. (2014) Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi, *Annual Review of Phytopathology*, 52(1):427-451.

3. Gaillardin C. (2010) Lipases as pathogenicity factors of fungi, In: *Handbook of Hydrocarbon and Lipid Microbiology*, T.N. Kenneth et al. (Eds.), 3259–3268 (Springer)

We would like to thank our colleagues from Skoltech Genomic Core Facility to help with library preparation for the whole genome sequencing of fungal strains.

**Расширение базы AAindex на неканонические аминокислоты
является ключевым методом повышения точности
предсказания вторичной и локальной структуры
коллагеноподобных участков белков**

Ю.В. Мильчевский, А. Кравцов, А.Н. Кретьева, Г.И. Кравацкая,
Ю.В. Кравацкий

*119991, Москва, ул. Вавилова, 32, Институт Молекулярной
Биологии РАН, milch@eimb.ru*

Введение. В текущий момент времени может складываться впечатление, что фундаментальная проблема предсказания пространственной структуры белка по последовательности уже решена методами машинного обучения [1, 2]. Тем не менее, эту проблему нельзя считать решённой с научной точки зрения ввиду обоснованных сомнений [3] в том, что AlphaFold и ESMFold предсказывают структуру белка, а не распознают её, опираясь на постоянно пополняющиеся базы данных [4, 5]. Этот вывод подтверждается результатами CASP: так, для негомологических белков AlphaFold 2 точно предсказал лишь 35.7% структур, а 58.7% были предсказаны с удовлетворительной точностью [6]. Таким образом, создание и разработка методов предсказания третичной структуры белка по последовательности на основе фундаментальных научных подходов остаются актуальными.

Постановка задачи. Одним из этапов предсказания

третичной структуры белка может быть предсказание его вторичной структуры по 8 классам DSSP либо локальной структуры по 16 классам де Бреверна [7]. Для решения этих задач необходимы статистически значимые наборы предикторов, состоящие из структурных и физико-химических признаков [8, 9]. Одним из наиболее широко используемых источников таких признаков является база данных AАindex [10], в которой представлены практически все опубликованные численные характеристики физико-химических свойств аминокислот. Однако AАindex охватывает исключительно 20 канонических аминокислот. Неканонические аминокислоты (ncAA) при этом не являются редкостью: PDB [4] содержит информацию о белках, включающих более 1000 различных ncAA. Более того, для некоторых классов белков именно ncAA определяют структурные особенности и функциональные свойства белка. В частности, гидроксипролин (HYP) составляет примерно 14% от общего числа аминокислот в фибриллярном коллагене, особенно в коллагене типа I [11].

Для расширения базы AАindex путём предсказания физико-химических свойств ncAA необходимо учитывать информацию о химических компонентах, общих для канонических и неканонических аминокислот. Для каждой ncAA должен быть сформирован набор признаков, отражающих её химический состав. Эти признаки могут быть сопоставлены с известными физико-химическими свойствами из базы AАindex с использованием пошагового регрессионного анализа. В качестве набора признаков могут быть использованы предикторы, полученные из строковой SMILES-кодировки [12], описывающей молекулярную структуру каждой аминокислоты. После построения расширенной базы физико-химических свойств на основе предсказанных значений для ncAA, такие данные могут быть включены в обучающие модели при анализе белков, содержащих неканонические аминокислоты.

Результаты. Нами разработан метод оценки физико-химических свойств неканонических аминокислот (ncAA) на основе данных базы AАindex, а также проведена оценка точности предсказаний. Расчёт физико-химических свойств осуществлялся для всех ncAA, представленных в PDB, на основе их химической структуры, закодированной в формате SMILES из базы данных PDBechem. Качество предсказаний оценивалось по значению коэффициента

корреляции Пирсона между экспериментальными и предсказанными значениями. Качество обучающей модели оценивалось вычислением среднеквадратичной ошибки (RMSE) предсказанного значения. Физические свойства, связанные с гидрофобностью и широко применяющиеся в задачах моделирования белковых структур (KUHL950101, WOLR790101, EISD840101, KIDA850101, PRAM900101, ENGD860101, BLAS910101, GOLD730101, COWR900101, CIDH920102) предсказаны с точностью свыше 80%. Более 20 свойств из базы AAindex предсказаны для ncAA с точностью, превышающей 95%. Как известно, точность предсказаний белковых структур существенно снижается для белков, содержащих коллагеноподобные участки [13]. Это может быть связано с высоким содержанием неканонической аминокислоты HYP в коллагеноподобных регионах, поскольку он играет ключевую роль в стабилизации структуры коллагена [14, 15]. Одной из причин снижения точности может служить отсутствие различия между пролином (PRO) и гидроксипролином (HYP) в символьном представлении белков: в однобуквенной кодировке обе аминокислоты обозначаются символом P. Между тем, анализ физико-химических свойств демонстрирует существенные различия между PRO и HYP. Так, для свойства EISD840101 (с точностью предсказания 92,4%) наблюдается пятикратное расхождение между измеренным значением для PRO и предсказанным значением для HYP. Ещё более выраженное различие отмечено для свойства ROSM880101 (точность предсказания 97,4%), значение которого меняет знак при переходе от PRO к HYP. Расширенная база данных AAindexNC была использована для предсказания вторичной структуры белков по девяти классам, включая новый класс, соответствующий спиралям типа полипролин-II. Полученная точность предсказания составила $Q_9 = 76.7\%$, что, согласно текущим данным, является наивысшим результатом в данной области на момент публикации. Все результаты свободно доступны на <https://aaindexnc.eimb.ru/> Работа поддержана грантом РФФ 24-24-00493.

1. Jumper J. et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583-589.

2. Lin Z. et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379(6637):1123-1130.
3. Finkelstein A.V. (2022) Does AlphaFold predict the spatial structure of a protein from physics... *bioRxiv*:2022.2011.2021.517308.
4. Berman H.M. et. al. (2012) The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* 20(3):391-396.
5. UniProt Consortium (2025) UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 53(D1):D609-D617.
6. Tunyasuvunakool K. et al. (2021) Highly accurate protein structure prediction for the human proteome. *Nature* 596(7873):590-596.
7. Etchebest C. et al. (2005) A structural alphabet for local protein structures: improved prediction methods. *Proteins* 59(4):810-827.
8. Мильчевский Ю.В. и др. (2023) Метод комплексного формирования... *Молекулярная Биология* 57(1):127-139.
9. Milchevskiy Y.V. et. al. (2023) Effective Local and Secondary Protein Structure Prediction... *Int J Mol Sci* 24(21):15656.
10. Kawashima S. et al. (2008) AIndex: amino acid index database, progress report 2008. *Nucleic Acids Res* 36(Database issue):D202-205.
11. Colgrave M.L. et. al. (2008) Hydroxyproline quantification for the estimation of collagen in tissue... *J Chromatogr A* 1212(1-2):150-153.
12. Weininger D. (1988) SMILES, a chemical language and information system. *J Chem Inf Comput Sci* 28(1):31-36.
13. O'Brien K.T. et. al. (2020) Prediction of polyproline II secondary structure propensity in proteins. *R Soc Open Sci* 7(1):191239.
14. Bella J. et. al. (1994) Crystal and molecular structure of a collagen-like peptide at 1.9 Å resolution. *Science* 266(5182):75-81.
15. Burjanadze T.V. (1979) Hydroxyproline content and location in relation to collagen thermal stability. *Biopolymers* 18(4):931-938.

**Концепция создания генетической лаборатории и
крупномасштабной ИТ-инфраструктуры для получения,
хранения и обработки данных WGS человека**

С.И. Митрофанов¹, Н.А. Сухих¹, Л.А. Голубникова¹,
К.С. Грамматикати¹, М.Н. Пилипенко¹, М.В. Иванов¹,
Д.А. Каштанова¹, Т.А. Шпакова¹, Л.В. Фролова¹, А.П. Сергеев¹,
А.И. Некрасова¹, С.В. Романова¹, А.А. Ивашечкин¹,
О.В. Варламова¹, В.В. Максютин¹, П.Г. Казакова¹, Н.В. Буланова¹,
П.А. Гребнев¹, А.В. Махотенко¹, Е.А. Снигирь¹, О.М. Марченко¹,
Е.Д. Маралова¹, А.С. Шингалиев¹, А.В. Закубанский¹,
В.В. Даниэль¹, В.С. Юдин¹, В.В. Макаров¹, А.А. Кескинов¹,
С.А. Краевой¹, С.М. Юдин¹, В.И. Скворцова²

¹ФГБУ «ЦСП» ФМБА России, Mitrofanov@cspfmbaru

²ФМБА России, Skvortsova@cspfmbaru

Развитие технологий и удешевление процесса секвенирования привели к росту количества запускаемых международных и локальных проектов изучения генома человека. На рубеже веков секвенирование одного человеческого генома стало мировым достижением^{1,2}. Спустя 25 лет ряд стран уже подходит к планке 100 тысяч секвенированных геномов^{3–5} и запускают новые проекты по секвенированию 1 миллиона геномов^{6,7}. Одной из главных сложностей таких проектов на начальных этапах является формирование стандартов и правил для единообразного сбора и обработки образцов, контроля качества, а также подготовки лаборатории и ИТ-инфраструктуры для хранения и биоинформатического анализа множества образцов и больших объемов данных. Для построения всех элементов молекулярно-генетической лаборатории и центра обработки данных, а также создания на их основе единого программно-аппаратного комплекса необходимы профессионалы разных профилей: научные сотрудники, медики, молекулярные биологи, биоинформатики, ИТ-специалисты и многие другие.

Одной из главных сложностей таких проектов на начальных этапах является формирование стандартов и правил для

единообразного сбора и обработки образцов, контроля качества, а также подготовки лаборатории и ИТ-инфраструктуры для хранения и биоинформатического анализа множества образцов и больших объемов данных. Для построения всех отдельных элементов молекулярно-генетической лаборатории и центра обработки данных и создания на их основе единого программно-аппаратного комплекса необходимы профессионалы разных профилей: научные сотрудники, медики, молекулярные биологи, биоинформатики, ИТ-специалисты и пр.

В 2019 году на базе ФГБУ «ЦСП» ФМБА России началось формирование инфраструктуры для проведения полногеномного секвенирования человека с целью создания базы данных популяционных частот генетических вариантов населения Российской Федерации. В процессе работы с нуля построены лабораторный комплекс для пробоподготовки и полногеномного секвенирования образцов и ИТ-инфраструктура для хранения и обработки больших объемов данных, а также сформирован и валидирован биоинформатический пайплайн для единообразного массового потокового анализа данных секвенирования человека. Дополнительно создана электронная система для сбора метаинформации в виде электронных индивидуальных регистрационных карт участников исследования, позволяющая детально проверять и контролировать всю вносимую информацию. Дополнительно создан крупный криобанк для долговременного хранения биообразцов. На протяжении всей цепочки жизненного цикла – от сбора биообразцов крови и метаинформации до принятия решения о включении результатов биоинформатического анализа (малых генетических вариантов) в базу данных – сформированы критерии контроля качества используемого материала и данных. Результатом работы стала коллекция биообразцов более 120 тысяч человек, проживающих на территории 85 регионов Российской Федерации, представленная в виде: (1) аликвот крови в криобанке, (2) деперсонализированной метаинформации во внутренней информационной системе, (3) сырых данных секвенирования и (4) результатов биоинформатического анализа. Полностью деперсонализированная часть информации опубликована в виде базы данных популяционных частот генетических вариантов населения Российской Федерации, содержащей частоты

встречаемости более 550 млн герминальных генетических вариантов: однонуклеотидных вариантов, коротких вставок и делеций (<https://nir.cspfmba.ru/info>).

В данной работе представлены основные стадии жизненного цикла образца. Дополнительно представлена концепция системы хранения больших массивов разнородных данных: метаинформации, результатов лабораторных исследований, данных с секвенаторов и результатов биоинформатического анализа. Особое внимание уделено контролю качества и критериям оценки пригодности используемого материала на всех стадиях жизненного цикла.

Авторы выражают благодарность следующим сотрудникам: Д.А. Шестаков, А.Н. Ломов, Д.В. Николаев, В.С. Носов, В.С. Саласин, Е.Ю. Ерохина, Н.В. Фелиз, Б.А. Бурков, В.А. Пахомов, В.Г. Харламов, К.Д. Конуреева, О.И. Блинова, М.В. Терехов, А.Ю. Яковчик, А.А. Мамчур, А.М. Румянцева, О.А. Антонова, М.В. Ерохина, Д.М. Загребина, Н.А. Кумар, Т.А. Костюкова, Д.С. Мацнев, В.А. Петрова.

1. Venter, J. C. et al. The Sequence of the Human Genome. *Science* (2001) doi:10.1126/science.1058040.
2. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).
3. Leong, I. U. S. et al. Large-Scale Pharmacogenomics Analysis of Patients With Cancer Within the 100,000 Genomes Project Combining Whole-Genome Sequencing and Medical Records to Inform Clinical Practice. *J Clin Oncol* 43, 682–693 (2025).
4. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 590, 290 (2021).
5. Feng, Y.-C. A. et al. Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genomics* 2, (2022).
6. Genomic data in the All of Us Research Program. *Nature* 627, 340–346 (2024).
7. Li, Z. et al. CMDB: the comprehensive population genome variation database of China. *Nucleic Acids Res* 51, D890–D895 (2022).

Мультиомиксный подход к исследованию роли моно- и диметилирования H3K36 и устанавливающих его ферментов в пространственной организации хроматина

В.Д. Митюков, С.В. Ульянов

Институт биологии гена РАН, vd.mityukov@yandex.com

Метилирование гистона H3 по лизину-36 (H3K36me) является важной модификацией, которая участвует в регуляции транскрипции, сплайсинга и репарации различных повреждений ДНК [1]. На данный момент установлено по меньшей мере 8 трансфераз, катализирующих моно-, ди- и триметилирование H3K36. Моно- и диметилирование лизина-36 обеспечивает семейство NSD белков, но роль этих ферментов и влияние меток на пространственную структуру хроматина не ясны. Для ответа на этот вопрос в рамках данной работы было проведено исследование пространственной организации интерфазного хроматина мышинных эмбриональных стволовых клеток дикого типа и на фоне нокаутов H3K36-специфичных моно- и диметилтрансфераз Nsd1, Nsd2, Nsd3. Установленная с помощью техники Hi-C пространственная организация хроматина позволила сделать ряд выводов о влиянии перечисленных метилтрансфераз, в то же время профили метилирования H3K36me2 и H3K27me3 и экспрессия генов полученные с помощью SET-seq [2] расширили возможности интерпретации результатов.

В данной работе показано, что при большом разрешении не обнаружено отличий в структуре хроматина в перечисленных нокаутах. Частота дальних взаимодействий и границы компартментов остаются практически неизменными в линиях с нокаутами по сравнению с контролем. При этом на локальном уровне генома при небольшом разрешении видны отличия, что подтверждается данными SET-seq. Несмотря на то что границы ТADов остаются в среднем неизменными, в некоторых случаях наблюдается ослабление силы взаимодействия внутри. При нокауте NSD2 наблюдается ослабление петель, в нокаутах NSD1 и NSD3 напротив заметно их усиление. Эти изменения происходят не по всему геному, но в отдельных локусах, что свидетельствует об

избирательном влиянии белков Nsd на структуру хроматина в определённых геномных контекстах (см рисунок 1 и 2).
 Данная работа поддержана грантом РФФ 24-44-00023.

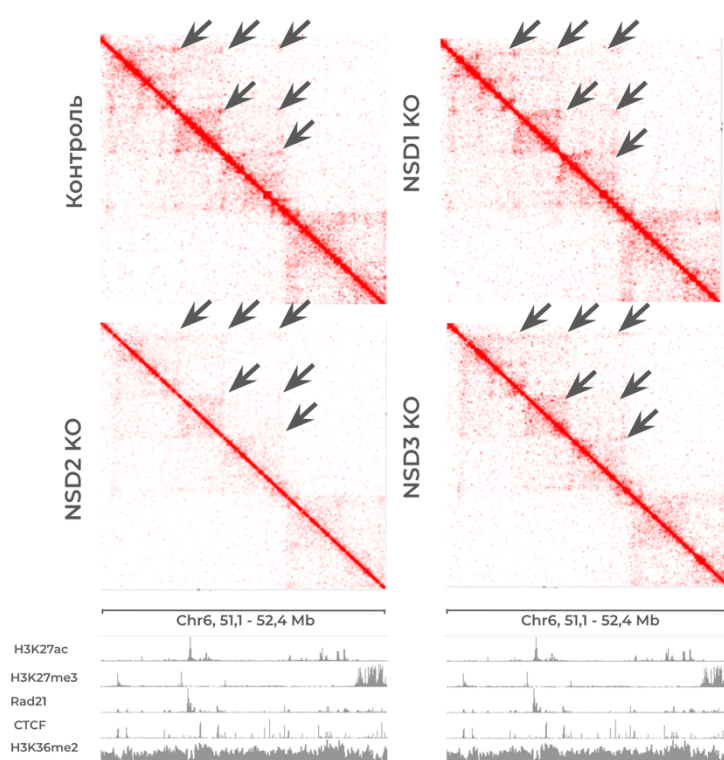


Рисунок 1 – Регион на 6й хромосоме, 51,1 - 52,4 Мб

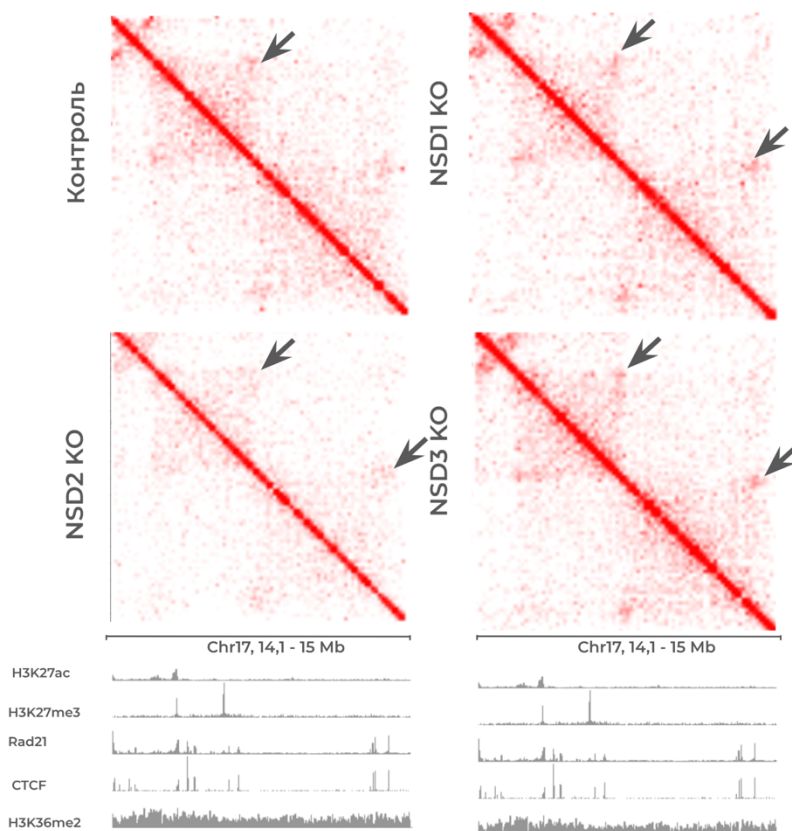


Рисунок 2 – Регион на 17й хромосоме, 14,1 - 15 Mb

1. Lam, U. T. F., Tan, B. K. Y., Poh, J. J. X., & Chen, E. S. (2022). Structural and functional specificity of H3K36 methylation. *Epigenetics & Chromatin*, 15(1), 17.
2. Sun, Z., Tang, Y., Zhang, Y., Fang, Y., Jia, J., Zeng, W., & Fang, D. (2021). Joint single-cell multiomic analysis in Wnt3a induced asymmetric stem cell division. *Nature communications*, 12(1), 5941.

Long-range Polycomb-mediated contacts in human neurons

Maria Molodova¹, Ilya Pletenev¹, Diana Zagirova¹, Nikita Vaulin², Kirill Morozov¹, Ekaterina Kuznechekova³, Anna Kononkova¹, Kirill Ulianov¹, Maria Bazarevich¹, Alexander Cherkasov¹, Olga Efimova¹, Maria Lagarkova⁴, Sergey Razin^{5,6}, Philipp Khaitovich¹, Sergey Ulianov^{5,6}, Ekaterina Khrameeva¹

¹*The Skolkovo Institute of Science and Technology, Moscow, Bolshoy Bulvar, 30 - 1, Maria.Molodova@skoltech.ru*

²*Medical University of Vienna, Spitalgasse 23, 1090 Wien, Vienna, Austria*

³*Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Leninskie Gory, 1, bldg. 73, Moscow*

⁴*Lopukhin Scientific Research Institute of Physical-Chemical Medicine, 1A, Moscow, Malaya Pirogovskaya Str., 1A*

⁵*Institute of Gene Biology, Russian Academy of Sciences, Moscow, Vavilova St., 34/5*

⁶*Lomonosov Moscow State University, Faculty of Biology, Leninskie Gory, 1, bldg. 12, Moscow*

The mammalian genome is hierarchically folded into structures of increasing complexity, including chromatin loops, topologically associating domains (TADs), compartments, and chromosome territories. This multiscale organization is essential for regulation of gene expression and other key processes in the nucleus [1-3]. Our research has previously identified long-range Polycomb (PcG) contacts specific to mature neurons, absent in other brain cells [4]. These interacting regions, enriched in H3K27me3, contain developmental transcription factors with repressed expression, suggesting a role in maintaining neuronal identity. However, the precise function of these contacts in neurons, as well as the mechanism of their formation remains unknown.

Here, we enhanced our dataset by generating high-resolution Hi-

C maps for neurons and non-neurons, and validated PcG contacts using ChIP-seq. Moreover, we analyzed chromatin organization in iPSC-derived neurons, revealing that they lack mature neuron-specific PcG contacts.

First, to confirm that these contacts are mediated by PcG proteins, we analyzed publicly available ChIP-seq data for BMI1 in neurons and oligodendrocytes. Consistent with our hypothesis, BMI1 was highly concentrated at gene promoters in neurons that form long-range contacts. We further performed ChIP-seq for Ring1B, a core Polycomb subunit, despite initial technical challenges due to extended postmortem intervals. Optimizing donor selection allowed us to successfully obtain Ring1B ChIP-seq data of sufficient quality.

Second, we tested the potential of iPSC-derived neurons as an *in vitro* model for studying PcG contacts. Our analysis revealed that iPSC-derived neurons exhibit reduced long-range Polycomb contacts, similar to fetal neurons. They also showed other chromatin features resembling neural progenitor cells, such as strong compartmentalization and TAD boundaries enriched for developmental genes. These findings highlight the limitations of iPSC-derived neurons in accurately modeling the chromatin architecture of mature brain neurons.

Together, our study provides a comprehensive analysis of Polycomb-mediated chromatin contacts in human neurons, elucidating the mechanisms behind their formation, particularly the role of Polycomb subunits. We also demonstrate that iPSC-derived neurons do not fully replicate the chromatin architecture of mature neurons, emphasizing the need for cautious interpretation of these models in neuroepigenetic research. These findings contribute to the broader understanding of 3D chromatin organization in the human brain and may help clarify the mechanisms underlying neurodevelopmental and neurodegenerative disorders.

1. J.R. Dixon, I. Jung, S. Selvaraj et al. (2015) Chromatin architecture reorganization during stem cell differentiation, *Nature*, 518:331–336.
2. B. Bonev, N. Mendelson Cohen, Q. Szabo et al. (2017) Multiscale 3D Genome Rewiring during Mouse Neural Development, *Cell*, 171:557–572.
3. G. Li (2011) Chromatin higher-order structures and gene regulation,

Current Opinion in Genetics & Development, 21:175–186.

4. I.A. Pletenev, M. Bazarevich, D.R. Zagirova et al. (2024) Extensive long-range polycomb interactions and weak compartmentalization are hallmarks of human neuronal 3D genome, Nucleic Acids Research, Nucleic Acids Research, 52(11):6234–6252.

Связь интеграционной активности ретроэлементов человека с факторами геномной нестабильности в солидных опухолях

А.А. Модестов, М.В. Сунцова, Г.С. Захарова, М.А. Золотовская,
Е.В. Поддубская, А.А. Буздин

*ФГАОУ ВО Первый МГМУ им. И.М. Сеченова Минздрава России
(Сеченовский Университет), Москва, Россия,
modestov_a_a@staff.sechenov.ru*

Геномная нестабильность является одной из ключевых характеристик опухоли [1]. Она служит источником как «драйверных», так и «пассажирских» мутаций, направляющих клетку по пути опухолевой трансформации и ускоряющих молекулярную эволюцию опухоли. Активность мобильных генетических элементов может способствовать развитию геномной нестабильности в опухолевых клетках. В данном исследовании мы количественно определили число вставок функционально активных семейств ретроэлементов (РЭ) LINE L1 (L1Hs и L1PA2) и Alu (AluYa5/Yb8) в солидных опухолях. Мы также оценили активность РЭ в более широких выборках из экспериментальной когорты и базы данных TCGA, а затем сопоставили её с факторами геномной нестабильности, такими как: РНК-редактирование, экспрессия гена теломеразы (*TERT*), уровень активации путей репарации ДНК, мутационная нагрузка опухоли (tumor mutation burden, TMB), вариации числа копий (copy number variation, CNV), потеря гетерозиготности (loss of heterozygosity, LOH), количество онкогенных химерных генов и микросателлитная нестабильность (microsatellite instability, MSI).

В рамках исследования использованы доступные

транскриптомные и экзомные профили наиболее распространённых солидных опухолей из экспериментальной когорты (415 опухолевых образцов) и базы данных TCGA (2488 опухолевых образцов): рак шейки матки, колоректальный рак, рак лёгкого, рак поджелудочной железы и рак желудка. Уровни активации (pathway activation level, PAL) 47 наиболее значимых путей репарации ДНК (по базам данных Reactome, NCI, KEGG, Qiagen) рассчитаны с использованием биоинформатической платформы Oncobox [2], [3].

Количественно определен уровень редактирования аденозин-инозин (A-to-I(G)) аденозиндеаминазой, действующей на РНК (adenosine deaminase acting on RNA, ADAR), с использованием набора из 20 генов, редактируемых как минимум в 80% образцов согласно базе данных REDportal [4]. В анализ были включены 24 сайта в экзонах следующих генов: *UQCRHL*, *COPA*, *RPL32*, *RPL15*, *FLNB*, *RPL24*, *IGFBP7*, *RPS23*, *EEF1A1*, *AZIN1*, *RPL7A*, *FTH1*, *COG3*, *RPLP1*, *NEIL1*, *ZNF358*, *GIPC1*, *RPS9*, *BLCAP* и *FLNA*. Частоты аллелей оценивались с использованием инструмента bam-readcount [5] для расчёта уровня редактирования в каждом сайте. В качестве альтернативного подхода рассчитан индекс редактирования Alu (Alu editing index, AEI) инструментом RNAEditingIndexer [6].

Основываясь на транскриптомных данных описанных выше факторов с количеством вставок Alu и L1, мы составили генную подпись, отражающую интеграционную активность РЭ. Для этого были протестированы комбинации генов, экспрессии которых демонстрировали значимую корреляцию с интеграционной активностью РЭ: *L1Ns*, *L1PA2*, *APOBEC3A*, *APOBEC3B* и *TERT*, чтобы оценить их способность различать образцы в соответствии с наблюдаемой активностью вставок РЭ. Наилучшая комбинация включала *L1Ns*, *L1PA2*, *TERT* и *APOBEC3B*. Значения площади под ROC-кривой (AUC), рассчитанные для полученной генной подписи, превышали 0,798 для вставок Alu и 0,758 для вставок L1 во всех протестированных перцентилях опухолевых образцов. Эти результаты указывают на то, что построенная генная подпись демонстрирует высокую надёжность в качестве биомаркера интеграционной активности РЭ в экспериментальных образцах опухолей.

Корреляционный анализ полученной подписи с

описанными выше факторами геномной нестабильности показал, что для объединённой выборки опухолевых образцов наблюдаются общие тенденции для экспериментальных данных и TCGA: преобладают положительные корреляции с путями репарации ДНК, а число отрицательно коррелирующих путей в обеих базах данных относительно невелико. При этом экспрессионная подпись интеграционной активности Alu и L1 продемонстрировала опухолеспецифические ассоциации с исследуемыми факторами. Большинство путей репарации ДНК положительно коррелировали с подписью интеграционной активности Alu и L1 по крайней мере в одном типе опухоли. Среди путей, демонстрирующих отрицательную корреляцию, были обнаружены пути, связанные с арестом клеточного цикла.

При анализе корреляции экспрессии генов РНК-редактирования с экспрессионной подписью выяснилось, что для экспериментальных образцов интеграционная подпись преимущественно отрицательно коррелирует с экспрессией гена *ADARBI*. Положительная корреляция наблюдалась для *APOBEC3B*. Для данных TCGA отрицательная корреляция также в основном наблюдалась для *ADARBI*, тогда как положительная корреляция была характерна для генов *APOBEC3B*, *AICDA*, *APOBEC1*, *APOBEC3C*, *APOBEC3D* и *APOBEC3F*.

В объединённой выборке экспериментальных образцов интеграционная активность положительно коррелировала с TMB. В данных TCGA наблюдалась положительная корреляция с TMB при раке лёгкого, шейки матки и поджелудочной железы. Данные по CNV и LOH были доступны для базы TCGA, и интеграционная активность элементов Alu и L1 положительно коррелировала как с CNV, так и с LOH при колоректальном раке, раке лёгкого, поджелудочной железы и желудка. Количество онкогенных химерных генов, информация о которых также была доступна в TCGA, положительно коррелировало с экспрессионной подписью, что наблюдалось при раке лёгкого, поджелудочной железы и желудка. При анализе MSI для колоректального рака наблюдалась общая тенденция между экспериментальными данными и TCGA: значения экспрессионной подписи были выше в образцах с микросателлитной стабильностью (MSS). Аналогичные значения в образцах с MSS также наблюдались при раке желудка.

Таким образом, в данном исследовании разработана экспрессионная подпись интеграционной активности РЭ, которая эффективно отражает количество вставок Alu и L1, и выявлены статистически значимые корреляции подписи с факторами геномной нестабильности. Данные исследования будут способствовать более подробному изучению молекулярных механизмов канцерогенеза и раскрытию роли РЭ в опухолевом процессе.

Работа выполнена при финансовой поддержке РНФ (грант № 20-75-10071-П).

1. J.-K.Lee et al. (2016) Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies, *Annu Rev Pathol*, **11**:283–312.
2. A.Buzdin et al. (2021) Using proteomic and transcriptomic data to assess activation of intracellular molecular pathways, *Adv Protein Chem Struct Biol*, **127**:1–53.
3. N.Borisov et al. (2020) Quantitation of Molecular Pathway Activation Using RNA Sequencing Data, *Methods Mol Biol*, **2063**:189–206.
4. L.Mansi et al. (2021) REDiportal: millions of novel A-to-I RNA editing events from thousands of RNAseq experiments, *Nucleic Acids Res*, **49**:D1012–9.
5. A.Khanna et al. (2021) Bam-readcount - rapid generation of basepair-resolution sequence metrics, *ArXiv*, arXiv:2107.12817v1.
6. S.H.Roth et al. (2019) Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity, *Nat Methods*, **16**:1131–8.

МОЛЕКУЛЯРНЫЕ МЕХАНИЗМЫ СБОЕВ В ПЕРЕДАЧЕ НЕРВНЫХ ИМПУЛЬСОВ ПО НЕЙРОНАМ В НАЧАЛЬНОЙ СТАДИИ БОЛЕЗНИ АЛЬЦГЕЙМЕРА

П.В. Мокрушников

Новосибирский государственный архитектурно-строительный
университет (Сибстрин), Новосибирск, Россия,
mokrusnikovpavel617@gmail.com

Болезнь Альцгеймера, или сенильная деменция — это

серьёзное нейродегенеративное заболевание [1]. В научном сообществе наиболее популярна гипотеза, согласно которой причиной болезни является возникновение большого количества амилоидных бляшек β -амилоидного пептида ($A\beta$) в головном мозге и следующее за этим появление гиперфосфорилированного таубелка в форме фибриллярных клубков в нейронах [1]. Амилоидные бляшки амилоидного β -амилоидного пептида представляют собой уложенные в стопку антипараллельные β -структуры, направленные перпендикулярно оси фибриллы [1]. Процесс образования этих амилоидных фибрилл начинается с образования протофибриллярных образований, богатых β -структурами [1]. Но до сих пор нет понимания, как эти структуры воздействуют на передачу потенциалов действия по нейронам. Неясность молекулярных механизмов развития болезни тормозит разработку лекарств, которые излечивают эту болезнь.

Цель работы: определить молекулярные механизмы изменения переноса потенциала действия в нейронах в начальной стадии болезни Альцгеймера.

Результаты и обсуждение. При болезни Альцгеймера происходит увеличение доли β -амилоидного пептида $A\beta$ в головном мозге, возникают сбои в передаче нервных импульсов по нейронам головного мозга. Ранее установлено, что при увеличении доли β -структур во вторичной структуре мембранных белков в мембранах происходит образование белок-липидных доменов [2]. В этих белок-липидных доменах липиды начинают сильнее притягиваться к мембранным белкам, в которых увеличилось количество β -структур, поверхностная плотность липидов увеличивается. Такие же процессы происходят в мембранах нейронов и глиальных клеток при попадании в них мономеров и димеров β -амилоидного пептида $A\beta$, имеющих конформацию в виде β -структуры. Образование белок-липидных доменов в мембранах нейронов ведёт к увеличению механических напряжений в мембранах. Как следствие, увеличивается активность Na^+, K^+ -АТФазы в мембранах нейронов [3], что, в свою очередь, ведёт к увеличению концентрации ионов Na^+ снаружи мембраны. В результате увеличивается разность электрохимических потенциалов снаружи и внутри нейрона. На мембране нейрона возникает гиперполяризация (тормозящий постсинаптический потенциал).

Для возникновения в нейроне потенциала действия требуется большее раздражение. Это является первой причиной нарушения деятельности головного мозга при увеличении количества β -амилоидного пептида $A\beta$ в головном мозге.

Второй причиной является изменение скорости передачи нервных импульсов по нейронам при увеличении количества белка $A\beta$ в шванновских клетках, создающих миелиновую оболочку нейронных аксонов. В этом случае шванновские клетки могут разрываться, как разрываются плазматические мембраны из-за образования в них белок-липидных доменов при взаимодействии с кортизолом [4, 5]. Эти разрывы шванновских клеток влияют на скорость переноса потенциала действия вдоль аксонов, которая становится меньше.

Третьей причиной изменения передачи нервных импульсов по нейронным сетям человека являются структурные изменения в пресинаптических окончаниях. Внедрение большого количества мономеров и димеров β -амилоидного пептида $A\beta$ в мембраны пресинаптических окончаний приводит к увеличению механических напряжений в мембране. Ca^{2+} -каналы зависят от механических напряжений в мембране, поэтому $A\beta$ оказывает влияние на перенос Ca^{2+} из межклеточной среды внутрь нейрона, изменяя выброс медиаторов в синаптическую щель [1].

Вывод. Выявлено три молекулярных механизма возникновения сбоев в передаче нервных импульсов по нейронам в начальной стадии болезни Альцгеймера. Сбои возникают из-за структурных изменений в мембранах нейронов и шванновских клеток. Эти структурные изменения возникают из-за внедрения мономеров и димеров $A\beta$ в мембраны. В результате в мембранах образуются белок-липидные домены, которые способствуют возникновению сбоев в передаче потенциалов действия вдоль нейронов.

Список литературы:

1. E. Liu et al. (2024) Updates in Alzheimer's disease: from basic research to diagnosis and therapies, Transl. Neurodegener., 13(1). doi: 10.1186/s40035-024-00432-x.
2. П.В. Мокрушников, В.Я. Рудяк (2025) Роль β - и α -структур мембранных белков в изменении структуры плазмалеммы, Биохимия, 90(1): 87–106.

3. П.В. Мокрушников и др. (2019) Структурные переходы в мембранах эритроцитов (экспериментальные и теоретические модели). Новосибирск, НГАСУ, 286.
4. P.V. Mokrushnikov (2019) Mechanical Stresses in the Lipid Bilayer of Erythrocyte Membranes, In: Lipid Bilayers: Properties, Behavior and Interactions, Mohammad Ashrafuzzaman (Eds.), 43–91 (NY: – Nova Science Publishers).
5. П.В. Мокрушников, В.Я. Рудяк (2023) Механическая модель образования складок на плазматической мембране, Доклады Академии наук высшей школы Российской Федерации, 2(59): 29–40. DOI: 10.17212/1727-2769-2023-2-29-40

Comparative analysis of chromatin compartment transitions during in vitro neuronal differentiation

Morozov Kirill¹, Kononkova Anna¹, Zagirova Diana^{1,2}, Dudkovskaia Anastasia¹, Vaulin Nikita³, Molodova Maria¹, Pletenev Ilya¹, Efimova Olga¹, Ulianov Kirill¹, Tvorogova Anna⁴, Lagarkova Maria⁵, Khaitovich Philipp¹, Razin Sergei^{4,6}, Ulianov Sergei^{4,6}, Khrameeva Ekaterina¹

¹*Skolkovo Institute of Science and Technology, Bolshoy bul., 30, build. 1, Moscow, Russia
kvmoroz67@gmail.com*

²*Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoy Karetny per. 19, build.1, Moscow, Russia*

³*Medical University of Vienna, Vienna, Austria*

⁴*Institute of Gene Biology Russian Academy of Sciences, Vavilova st., 34/5, Moscow, Russia*

⁵*Federal Research & Clinical Center of Physical-Chemical Medicine of Federal Medical Biological Agency, Malaya Pirogovskaya st., 1A, Moscow, Russia*

⁶Lomonosov Moscow State University, Faculty of biology, Leninskie Gory, 1, build. 12, Moscow, Russia

Changes in neuronal functions in neurodegenerative and mental diseases are an important applied problem in medicine and biology. The process of obtaining post-mortem samples is complex and the samples are heterogeneous, which creates a number of challenges for researchers. However, neurons differentiated from induced pluripotent stem cells have emerged as a key tool for addressing this issue, due to their availability, reproducibility under various conditions and great potential for modification [1]. However, the question of their similarity to mature neurons obtained from postmortem samples remains open. Previously, assessments were mainly based on morphological [2] and transcriptomic data [3]. No studies have been conducted on changes in chromatin structure during the neuronal differentiation of iPSC cells as well as their relationship with changes in expression compared to mature neurons. Nevertheless, studies on the three-dimensional organisation of the iPSC-derived cell genome already exist [4].

The aim of our work is to comprehensively compare the spatial organisation of the genome in post-mortem and induced pluripotent stem cell (iPSC)-differentiated neurons, thereby eliminating this gap. We have created a new database of uniformly processed Hi-C data that combines original data with that from open sources. A total of 141 human and 78 mouse samples were collected (<https://hic.skoltech.ru/datasets/>). We carried out a systematic analysis of changes in chromatin structure at various levels of organisation during the differentiation of iPSCs into adult neurons, as well as comparing them with postmortem samples. In particular, we studied changes in the organization of active (A) and inactive (B) chromatin compartments. To evaluate the importance of these changes, we integrated transcriptomic (bulk RNA-seq), epigenomic (ChIP-seq) and other omics data.

The results demonstrate that the chromatin of neurons derived from induced pluripotent stem cells (iPSCs) retains the characteristics of progenitor cells at all levels of organisation, including the compartmental level. Principal component analysis based on the compartmentalization profile revealed that iPSC-derived neurons form an independent cluster

that lies between progenitor cells and mature neurons obtained from post-mortem samples. Meanwhile, the iPSCs themselves form a separate cluster, positioned more distantly from the differentiated neurons than the post-mortem neurons. During neuronal differentiation, expansion of the B compartment in the genome is clearly visible, accompanied by an increase in its strength and a weakening of the A compartment. The expression of genes located in genomic regions that change compartment is highly correlated with changes in chromatin structure. This suggests that a large number of differentially expressed genes between post-mortem and differentiated neurons may be associated with developmental mechanisms that are not fully realised in vitro, but which determine the formation of mature neurons in vivo.

Therefore, the morphological similarity between iPSC-derived and post-mortem neurons and their expression of specific neuronal markers does not indicate that the cell types are completely similar. The chromatin of iPSC-differentiated neurons retains features of progenitor cells without reaching maturity. Our study provides important data for evaluating the suitability of iPSC-derived cell models, particularly neurons, and also establishes one of the largest unified Hi-C databases, which may be useful to other researchers in this field.

1. Di Val Cervo Rivetti et al. (2021) hiPSCs for Predictive Modelling of Neurodegenerative Diseases: Dreaming the Possible, *Nat. Rev. Neurol.* 17:381–392.
2. A. Jezierski et al. (2022) Electrophysiological- and Neuropharmacological-Based Benchmarking of Human Induced Pluripotent Stem Cell-Derived and Primary Rodent Neurons, *Stem Cell Rev. Rep.* 18:259–277.
3. A. E. Handel et al. (2016) Assessing Similarity to Primary Tissue and Cortical Layer Identity in Induced Pluripotent Stem Cell-Derived Cortical Neurons through Single-Cell Transcriptomics, *Hum. Mol. Genet.* 25:989–1000.
4. Prashanth Rajarajan et al (2018). Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science*, 362, 4311.

Геномное профилирование штаммов *Bifidobacterium adolescentis* в контексте ответа на иммунотерапию меланомы

М.Д. Морозов¹, М. С. Дьячкова¹, М.В. Одорская², В.В. Штоль¹,
В.В. Бабенко¹, П.Ю. Зорук¹, В.А. Веселовский¹, К.М. Климина¹

¹ФГБУ ФНКЦ ФХМ им. Ю.М. Лопухина ФМБА России, Москва,
119435, Малая Пироговская, д. 1а, taxim_d_morozov@mail.ru

²Институт общей генетики им. Н.И. Вавилова РАН, Москва,
Россия, 119991, ГСП-1, Москва, ул. Губкина, д. 3,
maya_epifanova@mail.ru

Ингибиторы иммунных контрольных точек (ИИКТ) произвели революцию в лечении рака. Однако их эффективность остается непредсказуемой и часто сопровождается побочными эффектами, что подчеркивает необходимость разработки новых стратегий, улучшающих клинические результаты [1]. Одним из перспективных направлений является модуляция противоопухолевого иммунного ответа через влияние микробиоты кишечника. Недавние исследования выявили связь между присутствием *Bifidobacterium adolescentis* в микробиоте пациентов и положительным ответом на иммунотерапию меланомы [2, 3]. При этом роль данного вида в иммуномодуляции остаётся неоднозначной: различные штаммы способны как усиливать, так и ослаблять противоопухолевый ответ.

Предполагается, что такая изменчивость, обусловлена различиями в молекулярных структурах бактериальных компонентов, таких как пептидогликан и метаболитов, особенно экзополисахаридов (ЭПС), которые модулируют активацию и дифференцировку иммунных клеток [4, 5]. Цель исследования - охарактеризовать штаммы *B. adolescentis*, потенциально усиливающие эффективность ИИКТ, а также выявить генетические и функциональные маркеры, связывающие штаммоспецифические особенности с влиянием на прогрессию опухоли.

Было проведено культивирование 32 штаммов *B. adolescentis* из микробной коллекции ИОГен РАН и выделена

высокомолекулярная ДНК. Приготовлены библиотеки с помощью набора КАРА HyperPlus (100 нг ДНК на образец) и секвенированы на платформах NextSeq 1000 (Illumina) и PromethION (Oxford Nanopore). Данные секвенирования были задепонированы в NCBI GenBank (PRJNA1217612). Гибридные сборки de novo были получены с помощью Unicycler, metaFlye и Pilon. Аннотацию генома проводили с помощью NCBI PGAP и KEGG. Пангеномный анализ проводили с помощью BPGA. Филогенетические связи реконструированы с помощью OrthoFinder и RAxML-NG и визуализированы в iTOL. На основе референсных генных кластеров из предыдущих исследований, с помощью ruGenomeViz и пользовательских скриптов была проведена идентификация потенциальных иммуномодулирующих генов, включая те, которые участвуют в биосинтезе ЭПС.

Мы получили 28 "complete" и 4 "draft" геномных сборок. Для проведения сравнительного анализа, были аннотированы и проанализированы 43 "complete" генома (включая 15 общедоступных геномов из базы данных NCBI GenBank), с целью выявления генетических особенностей, связанных с положительным ответом на иммунотерапию. Анализ выявил выраженную геномную гетерогенность, особенно в генах, связанных с синтезом ЭПС. Была идентифицирована группа из 11 генов: 4 участвуют (rmlA, rmlB, rmlC, и rmlD) в синтезе рамнозы, а 7 — в интеграции рамнозы в гликаны. Предполагается, что присутствие рамнозы в экзополисахаридах может способствовать ингибированию роста опухоли. На основе наличия идентифицированных генов штаммы были классифицированы по их предполагаемому влиянию на рост опухоли. Семь штаммов обладали частичным набором этих генов, в то время как только четыре штамма содержали все идентифицированные гены. Остальные штаммы содержали два или менее генов. Штаммы, несущие полный набор генов, вероятно, могут оказывать противоопухолевое действие и усиливать иммунный ответ.

Таким образом, выявление определенных генетических детерминант позволяет классифицировать штаммы *B. adolescentis* по их предполагаемого противоопухолевого потенциала и формирует основу для разработки системы рационального отбора

штаммов-кандидатов с наибольшей способностью усиливать иммунный ответ хозяина. На данном этапе были выбраны два штамма - 150 и DSM20083 - обладающие противоположным эффектом на рост опухоли. Полученные результаты послужат основой для *in vivo* скрининга всех штаммов и откроют перспективы их трансляционного применения в области иммунотерапии.

Работа выполнена при поддержке гранта РНФ № 25-75-10195.

1. Ling, Sia Pei et al. 2022. "Role of Immunotherapy in the Treatment of Cancer: A Systematic Review." *Cancers* 2022, Vol. 14, Page 5205 14(21): 5205. <https://www.mdpi.com/2072-6694/14/21/5205/htm> (February 7, 2025).
2. Matson, Vyara et al. 2018. "The Commensal Microbiome Is Associated with Anti-PD-1 Efficacy in Metastatic Melanoma Patients." *Science* 359(6371): 104–8. <https://www.science.org/doi/10.1126/science.aao3290> (February 7, 2025).
3. Zakharevich, Natalia V. et al. 2024. "Systemic Metabolic Depletion of Gut Microbiome Undermines Responsiveness to Melanoma Immunotherapy." *Life Science Alliance* 7(5). <https://www.life-science-alliance.org/content/7/5/e202302480> (February 7, 2025).
4. Zhao, Haobin et al. 2019. "Iturin A-like Lipopeptides from *Bacillus Subtilis* Trigger Apoptosis, Paraptosis, and Autophagy in Caco-2 Cells." *Journal of cellular physiology* 234(5): 6414–27. <https://pubmed.ncbi.nlm.nih.gov/30238995/> (February 7, 2025).
5. Yu, Rui, Fanglei Zuo, Huiqin Ma, and Shangwu Chen. 2019. "Exopolysaccharide-Producing *Bifidobacterium Adolescentis* Strains with Similar Adhesion Property Induce Differential Regulation of Inflammatory Immune Response in Treg/Th17 Axis of DSS-Colitis Mice." *Nutrients* 2019, Vol. 11, Page 782 11(4): 782. <https://www.mdpi.com/2072-6643/11/4/782/htm> (February 7, 2025).

Анализ гетероплазмии мтДНК на основе данных NGS

Г.С. Муравьев¹, Н.Д. Кашко², Д.А. Кнорре²

¹*Факультет биоинженерии и биоинформатики, Московский государственный университет имени М.В. Ломоносова, Россия, 119992, Москва, Ленинские горы, д. 1, turgeo13@gmail.com*

²*НИИ физико-химической биологии имени А.Н. Белозерского, Московский государственный университет имени М.В. Ломоносова, Россия, 119992, Москва, Ленинские горы, д. 1*

Состояние, когда в одной эукариотической клетке присутствует несколько вариантов мтДНК с разной нуклеотидной последовательностью, называется гетероплазмией. С митохондриальной гетероплазмией сопряжено большое число генетических заболеваний человека. Поэтому изучение закономерностей наследования мтДНК в череде поколений клеток и многоклеточных организмов представляет как общий, так и прикладной интерес.

Обычно за счёт случайного дрейфа генов через несколько поколений в каждой клетке остается только какой-то один вариант мтДНК (состояние гомоплазмии). Однако для некоторых организмов описаны случаи стабильной гетероплазмии, сохраняющейся в течение многих поколений.

В данных высокопроизводительного секвенирования (NGS) содержится информация о несогласующихся парных чтениях (discordant paired-end reads). Эта ситуация возникает, когда парные чтения картируются далеко друг от друга, или оказываются направлены в одну и ту же сторону. Такие чтения могут указывать на наличие хромосомных перестроек, в том числе делеций.

Но для эффективного анализа данных NGS и поиска структурных перестроек мтДНК требуется специальный биоинформационный конвейер. Тем не менее существует не так много эффективных программных конвейеров. Причём часть пакетов перестали поддерживаться вскоре после публикации, такие как MitoDel [1] или eKLIPse [2]. Другой существующий метод Delly [3] был написан для линейных хромосом и, как отмечают, авторы зависит

от глубины покрытия генома. Как следствие с кольцевой мтДНК дрожжей, обладающей неравномерным покрытием из-за наличия АТ-богатых областей, данный программный конвейер может работать не вполне корректно.

Поэтому в данной работе нами был разработан программный конвейер, реализующий анализ данных NGS. С помощью данного программного конвейера было показано наличие вторичных делеций в митохондриальных геномах нашей коллекции мутантных штаммов, а также в контрольном штамме дикого типа. С помощью данного программного конвейера было показано наличие вторичных делеций в митохондриальных геномах нашей коллекции мутантных rho- штаммов (имеющих первичные делеции мтДНК), а также в rho+ штамме дикого типа (имеющим полноразмерную мтДНК). Стоит отметить, что среди находок оказались варианты, совпадающие с координатами первичных делеций других штаммов. Также, используя результаты секвенирования rho0 штаммов (утративших мтДНК), мы показали отсутствие в наших штаммах митохондриально-ядерных псевдогенов (NUMTs). Так же мы планируем использовать программный конвейер, чтобы в дальнейшем производить подсчёт соотношения разных вариантов мтДНК в штамме. Причём мы ожидаем, что количественная оценка уровня гетероплазии может быть сделана как по результатам анализа NGS. В дальнейшем на основе этих данных мы планируем оценивать порог патогенности разных вариантов мтДНК.

Работа выполнена при финансовой поддержке РФФИ, проекты 22-14-00108.

1. Bosworth, C. M., Grandhi, S., Gould, M. P., & LaFramboise, T. (2017). Detection and quantification of mitochondrial DNA deletions from next-generation sequence data. *BMC bioinformatics*, 18(Suppl 12), 407. <https://doi.org/10.1186/s12859-017-1821-7>
2. Goudenège, D., Bris, C., Hoffmann, V., Desquirit-Dumas, V., Jardel, C., Rucheton, B., Bannwarth, S., Paquis-Flucklinger, V., Lebre, A. S., Colin, E., Amati-Bonneau, P., Bonneau, D., Reynier, P., Lenaers, G., & Procaccio, V. (2019). eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. *Genetics in medicine : official journal of the American*

College of Medical Genetics, 21(6), 1407–1416.
<https://doi.org/10.1038/s41436-018-0350-8>

3. Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* (Oxford, England), 28(18), i333–i339. <https://doi.org/10.1093/bioinformatics/bts378>

Создание инструмента для аннотации регуляторных вариантов в геноме человека

В.В. Мурашка¹, А.О. Зинкевич¹, Д.Д. Пензар²

¹*Факультет биоинженерии и биоинформатики, МГУ имени М.В. Ломоносова, Москва, Россия, bushsebz@gmail.com*

²*Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, Россия*

По текущим оценкам, порядка 80-90% геномных вариантов человека, ассоциированных с клинически значимыми заболеваниями, приходится на регуляторные участки генома.

На данный момент наиболее применяемым инструментом для аннотации VCF файлов является CADD. При составлении аннотации он полагается на разнообразные биологические параметры: метрики консервативности, расстояние до ближайших экзон-интронных границ, уровень экспрессии и др. Для предсказания эффектов регуляторных мутаций в CADD 1.6 были добавлены подходы на основе искусственных нейронных сетей - в его основе лежит конволюционная нейронная сеть [1]. Однако на соревновании DREAM Challenge 2022 CADD показали довольно скромные результаты, заняв только 20 место, не превзойдя по качеству даже базовое решение, предложенное организаторами конкурса [2].

В этом же соревновании была предложена архитектура LegNet, которая стала победителем во всех категориях [3]. Помимо этого, существует несколько различных вариантов

аннотации подобных вариантов, каждый из которых обладает своими плюсами и недостатками. В CADD реализован лишь один, выбранный авторами.

На фоне этого было решено создать новый инструмент для автоматической аннотации VCF файлов на основе LegNet. Были предобучены версии модели для разных тканей и клеточных линий. Полученный инструмент был протестирован на данных насыщающего мутагенеза для оценки, насколько модель чувствительная к однонуклеотидным заменам. Несмотря на то, что коэффициент корреляции предсказания и целевой переменной невысок (0.4), LegNet оказывается способен найти сайты, показавшие себя важными в MPRA эксперименте.

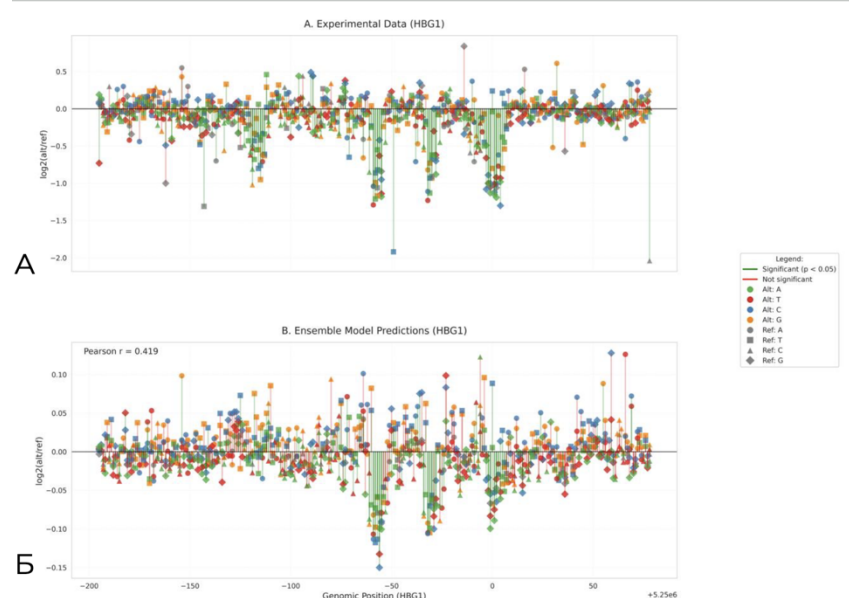


Рисунок 1. Логарифм влияния однонуклеотидных замен (SNV) в промоторной области гена HBG1 А. измеренный экспериментально Б. предсказанный с помощью LegNet

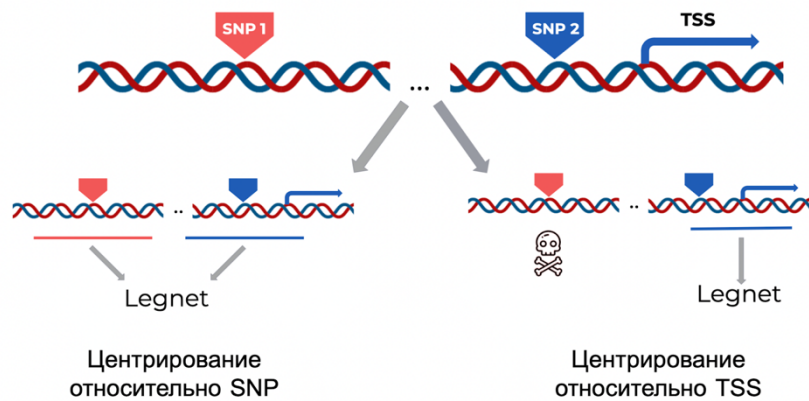


Рисунок 2. Оценка эффекта SNP реализована в двух вариантах: в окне, центрированном относительно мутации, и в окне, картированном относительно TSS.

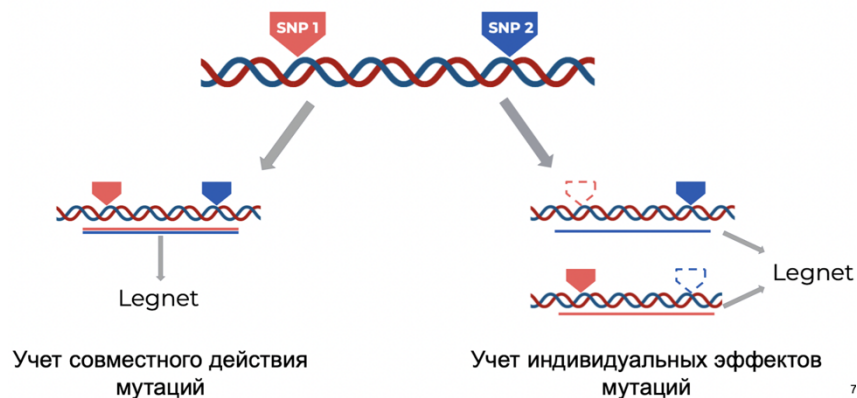


Рисунок 3. Вторая опция заключается в том, следует ли учитывать все мутации в заданном окне или же сосредоточиться на одной конкретной. Это позволит рассматривать как и взаимодействие целого ансамбля мутаций, так и сделать предсказания индивидуальных SNP.

1. P. Rentzsch et al. (2021) CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores.

Genome Med.

2. A.M. Rafi et al. (2024) A community effort to optimize sequence-based deep learning models of gene regulation, Nat Biotechnol.
3. D.D. Penzar et al. (2023) LegNet: a best-in-class deep learning model for short DNA regulatory regions., Nat Biotechnol.

Биомаркеры и генетические предикторы для продления спортивной результативности в силовых видах спорта

А. Д. Мухаметшина, Р. Д. Мухаметшина

*119234, г. Москва, ГСП-1, Ленинские горы МГУ 1, стр. 73,
Факультет биоинженерии и биоинформатики, 1709alb@gmail.com*

Проблема сохранения высокой спортивной подготовки с возрастом актуальна для всех дисциплин, однако особенно остро она стоит в силовых видах спорта, где пик физической формы сопряжён с высокой нагрузкой на опорно-двигательный аппарат. Несмотря на физиологическое старение, некоторые спортсмены сохраняют хороший уровень выступлений до 30–40 лет, что указывает на наличие устойчивых механизмов адаптации.

Целью настоящей работы является систематизация данных о молекулярных биомаркерах и генетических предикторах, связанных с адаптацией к силовой нагрузке, эффективностью восстановления и механизмами замедления старения скелетных мышц. Исследование включает систематический анализ публикаций и биоинформатический анализ геномных данных, что позволяет выделить ключевые мишени для управления тренировочным процессом и спортивным долголетием.

1. Введение

Спортсмены высокого уровня представляют собой наглядную модель для анализа комплексного влияния генетических факторов, тренировочных воздействий и условий окружающей среды на реализацию предельных адаптационных резервов организма. Установлено, что с возрастом происходит снижение анаболической активности, ухудшение регенераторных

способностей и усиление воспалительных процессов, что ограничивает возможность сохранения высокой силы и мощности. Поэтому возрастает значимость поиска надёжных молекулярных маркеров, способных отражать эффективность адаптации, восстановления и потенциальную продолжительность спортивной карьеры.

Молекулярная реакция на физические нагрузки демонстрируют значительную зависимость от типа тренировочного воздействия: силовые упражнения вызывают иные молекулярные адаптации, чем упражнения на выносливость, несмотря на схожие ответы на уровне общих паттернов генной экспрессии. Поэтому для адекватного анализа необходим учет данных, специфичных для силовых тренировок [1].

3. Материалы и методы

Анализ проводился с использованием баз данных PubMed, ScienceDirect и Google Scholar в период с 1 апреля по 22 апреля 2025 года. Были проанализированы публикации, посвящённые изучению биомаркеров и экспрессии генов, связанных с силовыми показателями, адаптацией к нагрузке, восстановлением и спортивным долголетием.

Для оценки молекулярных маркеров была использована база GEO (данные GSE59880), содержащие транскриптомные данные скелетной мускулатуры по возрастным группам молодых и возрастных спортсменов. Для получения списка дифференциально экспрессируемых генов данные обрабатывались с использованием пакета R «limma» и GEOparse в python.

Результаты дифференциальной экспрессии позволили выделить гены, значимо активируемые или подавляемые в ответ на нагрузку, включая *MYH1*, *MYBPC2*, *DDIT4*, *PDK4*, *DDIT4*, *CITED2*. Эти гены вовлечены в ключевые биологические процессы, отражённые в анализе Gene Ontology (GO): регуляция метаболизма, окислительный стресс, пролиферация клеток и ангиогенез. KEGG-анализ продемонстрировал обогащение в путях, связанных с сигналингом инсулина, биогенезом митохондрий, гликолизом.

3. Результаты

3.1 Генетические предикторы

Исходя из анализа литературных источников наиболее валидированными генетическими маркерами, связанных с

физической работоспособностью являются полиморфизмы ACTN3 R577X и ACE I/D, ассоциированные соответственно с взрывной силой и эффективностью восстановления [2]. Вместе с тем, недавние GWAS выделяют ряд новых SNP, например, полиморфизмы в *PPARA*, *VEGFA* и *ADRB2*, вовлечённых в регуляцию митохондриальной функции, воспалительных каскадов и метаболизма белков в скелетных мышцах [3].

3.2 Биомаркеры адаптации и восстановления при силовых тренировках

Молекулярный профиль скелетной мускулатуры у атлетов, занимающихся силовыми видами спорта, характеризуется специфическими адаптационными изменениями. Так выявлены следующие ключевые особенности:

- Повышенная активность обмена разветвлённых аминокислот (валин, лейцин, изолейцин) — отражает усиленный синтез белка и восстановление после нагрузки [1].
- Активация транскрипционных факторов MYF5, MYOD1, MYOG — критично для миогенеза и гипертрофии.
- Повышенное содержание фосфолипидов (PC и PE) — связано с пластичностью мембран и чувствительностью к инсулину.
- Устойчивый ответ на непривычные нагрузки (например, аэробные) с активацией путей HIF1A, CITED2, PDK4 — говорит о сохранённой адаптивности даже в узко специализированных спортсменах [1].

Таким образом, биомаркеры, такие как BCAA, MYF6, PDK4, а также уровни фосфолипидов, могут быть использованы для мониторинга эффективности тренировочного процесса и восстановления.

3.3 Влияние поведенческих и эпигенетических факторов

Образ жизни, питание, качество сна и социальная поддержка усиливают или ослабляют генетически обусловленные адаптационные механизмы [4]. Эпигенетические модификации, включая метилирование ДНК и гистонов, оказывают влияние на тренируемость и регенерацию.

4. Заключение

Индивидуализация тренировочного процесса с учётом молекулярных биомаркеров, генетических полиморфизмов и физиологических показателей открывает путь к научно обоснованному продлению спортивной карьеры. Использование мультиомных подходов позволяет точно воздействовать на ключевые адаптационные механизмы, улучшая восстановление и предотвращая возрастные потери мышечной массы и силы.

Будущие исследования должны быть направлены на трансляцию этих данных в практическую спортивную медицину и разработку персонализированных программ тренировок и реабилитации для спортсменов силовых видов спорта.

Литература

1. Reitzner, S. M., Emanuelsson, E. B., Arif, M., Kaczkowski, B., Kwon, A. T. J., Mardinoglu, A., Arner, E., Chapman, M. A., & Sundberg, C. J. (2023). *Molecular profiling of high-level athlete skeletal muscle after acute exercise – a systems biology approach*. bioRxiv. <https://doi.org/10.1101/2023.04.11.535671>
2. Kumagai, H., Miller, B., Kim, S. J., Torgan, C. E., Murakami, H., Higashida, K., Ishiura, H., & Ishiura, S. (2023). Novel insights into mitochondrial DNA: Mitochondrial microproteins and mtDNA variants modulate athletic performance and age-related diseases. *Genes (Basel)*, 14(2), 286. <https://doi.org/10.3390/genes14020286>
3. Bıçakçı, B., Ciężczyk, P., & Humińska-Lisowska, K. (2024). Genetic determinants of endurance: A narrative review on elite athlete status and performance. *International Journal of Molecular Sciences*, 25(23), 13041. <https://doi.org/10.3390/ijms252313041>
4. Kroshus, E., Wagner, J., Wyrick, D., Athey, A., & Bell, L. (2019). Wake up call for collegiate athlete sleep: Narrative review and consensus recommendations from the NCAA Interassociation Task Force on Sleep and Wellness. *British Journal of Sports Medicine*, 53(12), 731–736. <https://doi.org/10.1136/bjsports-2019-100590>

Ссылка на данные GSE59880
<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59880>

**Поиск моноклональных антител к isoD7-A β с помощью
кластерного анализа репертуаров В-клеток
иммунизированных мышей**

Михаил Юрьевич Мышкин ¹, Елизавета Андреевна Колобова ^{2,3},
Николай Андреевич Николаев ^{1,4}, Владимир Александрович
Митькевич ³, Ирина Леонидовна Григорова ^{2,3}, Екатерина
Марковна Мерзляк ^{1,2}, Ольга Владимировна Британова ^{1,2}

¹ ФГБУН ГНЦ РФ Институт биоорганической химии им.
академиков М.М. Шемякина и Ю.А. Овчинникова РАН, Москва,
улица Миклухо-Маклая, д. 16/10, mikhailmyshkin.ibch@gmail.com

² Институт Трансляционной Медицины, Российский
Национальный Исследовательский Медицинский Университет им.
Н.И. Пирогова, Москва, ул. Островитянова д. 1.

³ ФГБУН «Институт молекулярной биологии им. В.А.
Энгельгардта» РАН, Москва, ул. Вавилова, 32.

⁴ Факультет биоинженерии и биоинформатики МГУ имени М.В.
Ломоносова, Москва, Ленинские горы д. 1, стр. 73.

Болезнь Альцгеймера (БА) является наиболее распространенной формой деменции, на долю которой приходится до 70% всех случаев и сопровождается формированием амилоидных бляшек из бета-амилоидных пептидов (A β). Бета-амилоидные олигомеры оказывают токсическое воздействие на нейроны, нарушая их связи и способствуя когнитивным нарушениям. Одной из наиболее токсических форм бета-амилоидов, провоцирующих олигомеризацию пептидов, является изомеризованный по Asp7 A β (isoD7-A β), который накапливается у людей с возрастом [1].

Недавно одобренные FDA моноклональные антитела, такие как леканемаб и адуканумаб [2], представляют собой современные

подходы к терапии БА. Эти препараты направлены на удаление амилоидных отложений или предотвращение их накопления. Однако их использование связано с рядом побочных эффектов и приводит к незначительному или кратковременному клиническому улучшению [3].

Основной целью настоящего исследования является получение моноклональных антител к isoD7-A β с целью разработки новых методов диагностики и терапии болезни Альцгеймера. Антитела, специфичные к isoD7-A β , должны селективно распознавать модификацию аспартата в 7 положении, что позволит избежать возможных побочных эффектов в результате связывания с нормальными формами амилоида [4].

В данной работе были проанализированы репертуары В-клеточных рецепторов лимфоцитов 5 мышей (линия BALB/c), иммунизированных конъюгатом овальбумина и N-концевого фрагмента isoD7-A β (OVA-isoD7-A β 1-16). С помощью FACS В-клетки мышей были отсортированы на три группы: специфичные либо к isoD7-A β , но не к A β (“специфичные”), либо к ним обоим (“кросс-специфичные”), либо ни к одному из них (“неспецифичные”, “контроль”). кДНК библиотеки репертуаров В-клеточных рецепторов (тяжелые HC и легкие LC цепи) были получены методом 5'RACE и затем отсекарованы на платформе Illumina. Был проведен кластерный анализ последовательностей В-клеточных рецепторов, а для тяжелых цепей еще дополнительно были построены филогенетические деревья. С помощью фильтрации кластеров по присутствию в них “специфических” и “кросс-специфических” клонотипов, анализа филогенетических деревьев, а также на основе сравнения встречаемости этих клонотипов в различных репликах, были отобраны наиболее вероятные пары HC+LC, которые могли быть в составе антител, ответивших на иммунизацию. Затем были получены рекомбинантные антитела в состав которых входили данные цепи и проанализирована их способность связывать isoD7-A β и A β .

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (Контракт в системе электронный бюджет № 075-15-2024-530).

1. T. Shimizu et al. (2000) Isoaspartate formation and neurodegeneration

- in Alzheimer's disease, Arch Biochem Biophys, 381 № 2: 225–234.
2. K.I. Avgerinos et al. (2021) Effects of monoclonal antibodies against amyloid- β on clinical and biomarker outcomes and adverse event risks: A systematic review and meta-analysis of phase III RCTs in Alzheimer's disease, Ageing Res Rev, 68: P. 101339.
2. D. Jeremic et al. (2021) Past, present and future of therapeutic strategies against amyloid- β peptides in Alzheimer's disease: a systematic review, Ageing Res Rev, 72: P. 101496.
4. Gnoth K. et al. (2020) Targeting isoaspartate-modified A β rescues behavioral deficits in transgenic mice with Alzheimer's disease-like pathology, Alzheimers Res Ther, 12 № 1: 149.

**Гетерогенность скелетных мышц человека:
транскрипционные маркеры быстрых и медленных
мышечных волокон**

Ш. Р. Низамов¹, Г. Р. Газизова¹, Е. И. Шагимарданова^{2,3}, О. А.
Гусев^{2,4}, Р. М. Девятяров², Й. Хаяшизаки⁵

¹НИЛ “Регуляторная геномика”, Институт фундаментальной
медицины и биологии Казанского
(Приволжского) Федерального Университета, Казань, Россия
nizamov.1999@gmail.com

²ООО «ЛИФТ Центр», Москва, Россия

³Сколковский институт науки и технологий, Москва, Россия

⁴ФГБУ «Федеральный центр мозга и нейротехнологий» ФМБА,
Москва, Россия

⁵K.K. DNAFORM / K.K. Mirai Genomics, Yokohama, JAPAN

В теле человека насчитывается более 650 скелетных мышц, каждая из которых обладает рядом уникальных характеристик, в том числе различной степенью устойчивости к миопатиям и другим заболеваниям. Так, например, экстраокулярные мышцы не

подвержены дегенеративным изменениям при мышечной дистрофии Дюшенна [1]. Однако причины этих различий до конца не изучены. Открытым остаётся вопрос о вкладе клеточного состава, иннервации, микроокружения, анатомической локализации и внутренних свойств миоцитов и мышечных волокон в формирование такой устойчивости. Для исследования этого вопроса мы применили транскриптомный подход, позволяющий оценить клеточный состав мышечных образцов и выявить гены, специфичные для соответствующих клеточных типов.

Материалы и методы

В работе использовались образцы скелетных мышц от 4 человек посмертно (3 мужчин и 1 женщина) в возрасте от 56 до 67 лет. Всего было отобрано 225 образцов из 75 различных мышц. Образцы тотальной РНК использовали для подготовки библиотек CAGE-seq и дальнейшего секвенирования на платформе Illumina. Анализ данных секвенирования включал: выравнивание с помощью программ BWA [2] и HISAT2 [3] на геномную сборку GRCh38, кластеризацию с экспрессией >10 TPM как минимум в одном образце (расстояние между ближайшими сайтами инициации транскрипции CAGE не более 20 п.о.) [4], аннотацию кластеров в пределах ± 500 п.о. от начала транскрипта [5].

Для оценки доли каждого типа клеток в образцах CAGE-seq (от англ. Cap Analysis Gene Expression sequencing) был использован метод деконволюции с помощью пакета MuSiC (от англ. MUlti-Subject Single Cell deconvolution) [6] для R. Метод подразумевает использование в качестве референса данные single cell экспериментов, на основе которых составляются пропорции клеточных типов и производится предсказание клеточных пропорций на bulk данных. В качестве референса использовались данные single cell RNA-seq, полученные из скелетных мышц человека в рамках проекта Heart Cell Atlas [7]. На основе полученных данных был проведен анализ с использованием ROC-кривых (от англ. Receiver Operating Characteristic curves) для определения оптимального процентного соотношения клеток быстрого и медленного типа.

Результаты

Мы идентифицировали 21 532 активных точек инициации транскрипции. Из них 16 087 CAGE-пигов было аннотировано как

соответствующие промоторам. Анализ деконволюции показал, что для некоторых мышц пропорции клеточных типов существенно различались между образцами одной и той же мышцы. Это отличие было обусловлено тем, что образцы отбирались из разных участков мышцы. Особенно ярко такая вариабельность была видна на примере миоцитов, поскольку соотношение быстрых и медленных волокон в мышцах варьируется в зависимости от их локализации [8]. Анализ ROC-кривых позволил определить оптимальное пороговое значение в таком соотношении для классификации образцов на быстрые и медленные. Был проведен анализ дифференциальной экспрессии генов, в рамках которого сравнивались профили экспрессии быстрых образцов относительно медленных образцов мышц. В результате анализа было выявлено 782 гена, ассоциированных с быстрыми и медленными мышечными волокнами. Из них 392 гена оказались характерны для быстрых волокон, 390 — для медленных. Среди выявленных генов были обнаружены характерные маркеры, такие как MYH2, MYH1, MYH4, MYBPC2, TNNT3 для быстрых [9, 10], и MYH7, TPM3, LDHB, TNNI1, MYL2 – для медленных [9, 10].

Выводы

Анализ транскриптомных данных показал, что на основе профиля экспрессии возможно охарактеризовать образцы как содержащие преимущественно быстрые или медленные мышечные волокна, независимо от анатомического происхождения мышцы. В ходе исследования нами были идентифицированы гены и, что особенно важно, соответствующие им промоторы, специфичные для быстрых и медленных волокон. Полученные результаты могут быть использованы для учёта типа волокон при анализе клинических и физиологических особенностей различных мышц, что позволит минимизировать влияние этого фактора в последующих исследованиях.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации (№075-15-2021-601).

1. Stuelsatz, P., Shearer, A., Li, Y., Muir, L.A., Ieronimakis, N., Shen, Q.W., Kirillova, I., & Yablonka-Reuveni, Z. (2015). Extraocular muscle satellite cells are high performance myo-

- engines retaining efficient regenerative capacity in dystrophin deficiency. *Developmental biology*, 397.1:31-44.
2. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754-1760.
 3. Kim, D., Paggi, J.M., Park, C. et al. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37:907–915
 4. Kouno, T., Moody, J., Kwon, A. T. J., et al. (2019). C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution. *Nature Communications*, 10.1:360.
 5. Deviatiiarov, R.M., Gams, A., Kulakovskiy, I.V. et al. (2023) An atlas of transcribed human cardiac promoters and enhancers reveals an important role of regulatory elements in heart failure. *Nature Cardiovascular Research*, 2:58–75.
 6. Wang, X., Park, J., Susztak, K. et al. (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications* 10:380.
 7. Litviňuková, M., Talavera-López, C., Maatz, H. et al. (2020) Cells of the adult human heart. *Nature* 588:466–472.
 8. Семенова, Е. А., Хабибова, С. А., Борисов, О. В., Генерозов, Э. В., & Ахметов, И. И. (2019). Вариабельность структуры ДНК и состав мышечных волокон человека. *Физиология человека*, 45.2:128-136.
 9. Li, Y., Ma, Q., Shi, X., Yuan, W., Liu, G., & Wang, C. (2022). Comparative Transcriptome Analysis of Slow-Twitch and Fast-Twitch Muscles in Dezhou Donkeys. *Genes*, 13.9:1610.
 10. Alto, S. I., Chang, C. N., Brown, K., Kioussi, C., & Filtz, T. M. (2021). Gene Expression Profiling of Skeletal Muscles. *Genes*, 12.11:1718.

Ecological determinants of the specific and individualized responses of the gut microbiome to dietary fibre in humans

Daria D. Nikolaeva¹, Anissa A. Armet³, Fuyong Li⁴, Marcus Claesson², Mikhail S. Gelfand¹, Jens Walter²

¹*Skoltech, Moscow, Russia*

²*APC Microbiome Ireland, University²College Cork, Cork, Ireland, daranikolaeva@gmail.com*

³*Department of Agricultural, Food & Nutritional Science, University of Alberta, Edmonton, Canada*

⁴*Department of Animal Science, College of Animal Sciences, Zhejiang University, Hangzhou, China*

Dietary fibers (DFs) contribute to human health and induce specific shifts in gut microbiome composition and functions. However, there is a substantial inter-individual variation in responses, and the determinants of this variability, as well as the consequences for health effects, are poorly understood. The objectives of this work were to characterize effects of distinct fibers on microbiome dynamics in a well-powered human trial, determine the ecological principles that drive responses, and elucidate whether individualized effects of fiber on the microbiome influence the physiological effects of fiber and can be predicted using baseline microbiome features.

In a six-week, parallel-arm, randomized controlled trial, adults with excess weight added 25 g (females) or 35 g (males) of one of three DF supplements to their usual diet daily: acacia gum (AG; n = 75), resistant starch type 4 (RS4; n = 75), or microcrystalline cellulose (MCC; non-fermentable; n = 45). A multiomics approach was applied to fecal samples collected at baseline, week 1, and week 6 and changes of microbiome composition and genomic functional capacity, metabolites (SCFAs and bile acids) and mechanistic and clinical host markers were assessed.

MCC had no effect on fecal microbiota diversity and composition. The fermentable fibers decreased α -diversity and led to

changes in the composition of the fecal microbiota in a subset of subjects (PERMANOVA $p < 0.001$). Compositional shifts were highly specific to the chemical structure of the fiber, with AG increasing *Bifidobacterium longum* and *Faecalibacterium prausnitzii* while RS4 increased *Bifidobacterium adolescentis* and *Parabacteroides distasonis* ($p < 0.001$). Responses were largely stable between week 1 and week 6 (45 - 80% of responders), but *F. prausnitzii* showed a transient response to AG with an increase at week 1 followed by a decrease at week 6 (40% of responders). Although statistically significant in the overall population, responses were highly individualized, with individual taxa showing a response in only 30-70% of participants.

AG and RS4 altered fecal microbiome diversity and composition and induced highly specific microbiome shifts with diverse individual and temporal patterns. Ongoing work is aimed at the characterization of the ecological drivers of responses and investigation of the question if individualized responses are predictable via the microbiome features and linked to the physiological effects of the fibers.

Предсказание профилей 3'-концевых поли(А)- последовательностей мРНК в клеточных линиях человека

А. И. Никольская^{1,2}, Н. С. Грызунов¹, А. А. Миронов^{1,2}

¹*Факультет биоинженерии и биоинформатики, Московский
государственный университет имени М.В.Ломоносова, Москва,
Россия, arnikol31@gmail.com*

²*Институт общей генетики им. Н.И. Вавилова РАН, Москва,
Россия*

Мотивация и цель: Пространственно-временные паттерны экспрессии генов играют ключевую роль в развитии и функционировании организмов. Хотя тканеспецифическая регуляция транскрипции изучена широко и уже используется для разработки терапий, посттранскрипционный уровень контроля

экспрессии по-прежнему остается менее изученным. Одной из самых распространенных посттрансляционных модификаций является полиаденилирование — присоединение поли(А)-хвоста к 3'-концу мРНК. После синтеза в ядре поли(А)-последовательность постепенно укорачивается, и характер этих изменений зависит от свойств конкретного транскрипта. Несмотря на то, что поли(А)-хвост играет важную регуляторную роль на всех этапах жизни мРНК, правила, определяющие варибельность его длины, остаются неясными. Таким образом, создание модели для предсказания профилей поли(А)-последовательностей мРНК представляет собой важный шаг к пониманию посттранскрипционной регуляции и ее варибельности между клеточными типами.

Материалы и методы: Длины поли(А)-последовательностей для обучающей выборки были получены из данных одномолекулярного секвенирования РНК клеточных линий K562, MCF7 и A549, представленных в рамках проекта SG-NEx [1]. Первичная обработка включала расшифровку сигнала и оценку длины поли(А)-хвостов с помощью программы Dorado. Картирование прочтений на референсный транскриптом человека осуществлялось с помощью программы minimap2 с параметрами -ax splice -k14 -uf. В дальнейшем были отобраны выравнивания на изоформы транскриптов с MAPQ > 0. Мы рассчитали медианную длину поли(А)-последовательности для каждой изоформы, при условии, что она была представлена не менее чем 10 прочтениями в соответствующей клеточной линии.

Результаты: Для предсказания профилей поли(А)-хвостов мРНК мы разработали гибридную архитектуру, объединяющую сверточные и рекуррентные нейросетевые слои (Рис. 1). Модель опирается на ранее предложенную архитектуру Saluki для предсказания времени полужизни РНК [2] и дополнена EfficientNetV2-подобными блоками из архитектуры LegNet [3]. В качестве обучающих данных были взяты 8 тысяч последовательностей транскриптов, разрешенных до конкретных изоформ с информацией о клеточной линии. Данные разбивались на тренировочные и тестовые наборы без утечки изоформы одного и того же гена в тестовую выборку. Обучение проводилось в двух режимах: (1) на полной последовательности транскрипта и (2)

только на участке 3'UTR длиной 500 нуклеотидов, прилегающем к поли(А)-сайту. Предварительный анализ показал, что использование полной последовательности с добавлением информации о кодонах и сайтах сплайсинга не приводит к улучшению точности предсказания. Текущая версия модели достигает коэффициента корреляции Спирмена $r_s = 0.55$, что приближается оценке согласованности измерений длины поли(А)-хвоста между репликами одной и той же клеточной линии ($r_s = 0.6–0.7$).

Перспективным направлением развития модели является учет вторичной структуры РНК наряду с последовательностью для повышения качества предсказаний. Кроме того, мы планируем перейти от оценки медианной длины к предсказанию формы распределения поли(А)-хвостов. В совокупности это позволит выявить структурные и регуляторные элементы мРНК, лежащие в основе гетерогенности поли(А)-профилей мРНК в клетках человека. Работа поддержана грантом РФФ № 23-14-00136.

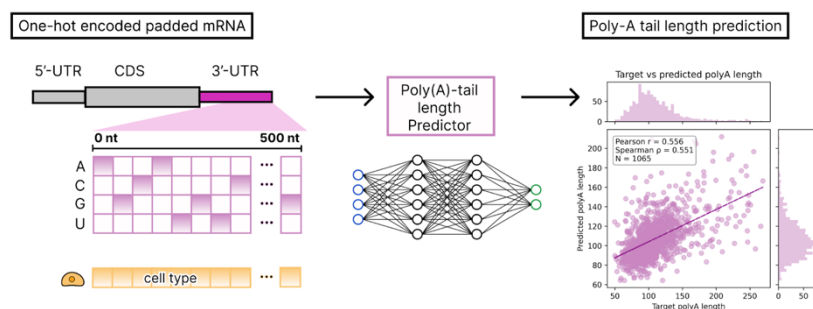


Рис. 1. Схема обучения нейросети для предсказания медианной длины поли(А)-хвостов мРНК по последовательности.

1. Y. Chen et al. (2021) A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines, bioRxiv 2021.04.21.440736
2. V. Agarwal et al. (2022) The genetic and biochemical determinants of mRNA degradation rates in mammals. Genome Biol 23, 245
3. D. Penzar, D. Nogina et al. (2023) LegNet: a best-in-class deep learning model for short DNA regulatory regions, Bioinformatics, btad457

SIRT6 Modulates Splicing Programs Governing Neural Cell Fate Decision.

T. S. Okeowo¹, D. Smirnov¹, Y. Rabuah², D. Toiber², E. Khrameeva¹

¹*Skolkovo Institute of Science and Technology, Moscow, Russia,
Tunde.Okeowo@skoltech.ru*

²*Ben-Gurion University of the Negev, Beersheba, Israel*

SIRT6, an NAD⁺-dependent chromatin remodeler, plays a vital role in brain development, primarily through its function in maintaining genomic stability^{1,2}. Although chromatin-modifying proteins like SIRT6 lack an RNA-binding domain, growing evidence suggests that they can influence RNA splicing³. Here, we investigate the role of SIRT6 in splicing regulation using both mouse and human models.

Using transcriptomic data from embryonic stem cells (ESCs), neural stem cells (NSCs), and cortical tissue, we performed differential splicing analysis between SIRT6 knockout (KO) and wild-type (WT) mice. Loss of SIRT6 led to widespread splicing alterations, including in genes involved in interkinetic nuclear migration and forebrain progenitor proliferation—processes essential for neurodevelopment. Notably, we observed splicing changes in key signaling genes such as FGFR2 and PTK2, which are implicated in the proliferation, differentiation, and survival of neural progenitors.

To investigate the temporal dynamics of SIRT6-dependent splicing regulation, we used a doxycycline-inducible RNA interference system to gradually silence SIRT6 in human SH-SY5Y neuroblastoma cells. A slight reduction in SIRT6 levels affected pathways related to sister chromatid separation. Further reduction disrupted additional processes, including immunity, proteasome-mediated protein degradation, metabolism, and DNA repair. With prolonged strong depletion—approximating a knockout state—we observed broader impacts on FGFR1 signaling, MAPK signaling (including oncogenic cascades), axon guidance, and L1CAM interactions. These progressive, dose- and time-dependent changes in pathway activity recapitulate the

splicing alterations that were previously not detected in the mouse SIRT6-KO brain.

Finally, to identify potential regulators mediating these effects, we analyzed CLIP-seq data and found seven RNA-binding proteins (RBPs) with enriched binding profiles in the SIRT6 context—four spliceosome-associated proteins linked to proliferation and survival and three key RBPs essential for cell fate decision and neuronal differentiation timing. All but EFTUD2 and PTBP1 were upregulated in the SIRT6 knockout brain, suggesting possible coregulation. While EFTUD2 was not differentially expressed, EFTUD2 and PRPF8 interact with SIRT6, and their hyperacetylation under the absence of SIRT6 points to a post-translational regulatory role. The acetylation status of PTBP1 remains to be confirmed. These findings indicate that SIRT6 can influence RNA splicing both by controlling chromatin remodeling and by modulating the expression and acetylation of RBPs, highlighting a previously underappreciated regulatory axis with potential relevance to neurodevelopment and disease.

1. R. Mostoslavsky et al. (2006) “Genomic instability and aging-like phenotype in the absence of mammalian SIRT6”, *Cell*, 315-29.
2. D. Toiber et al. (2013) “SIRT6 recruits SNF2H to DNA break sites, preventing genomic instability through chromatin remodeling”, *Molecular Cell*, 454-68.
3. R. Rahhal, E. Seto (2019) “Emerging roles of histone modifications and HDACs in RNA splicing” *Nucleic acids*, 4911-4926.

Регуляторные варианты при легочной гипертензии: идентификация и механизмы действия

Г.С. Охрименко¹, Н.В. Сафронова¹, В.И. Замятин¹, И.И.
Боровикова², Д.А. Затеищikov³, М.С. Попцова¹

¹ФГАОУ ВО «Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия
gsokhrimenko@hse.ru

²ФГОУ ВПО «Московский государственный университет имени М.В.Ломоносова» Москва, Россия

³ГБУЗ «ГКБ № 29 им. Н.Э. Баумана ДЗМ», Москва, Россия

Легочная гипертензия (ЛГ) — тяжелое заболевание, которое характеризуется повышением давления в легочной артерии. Точные молекулярные механизмы ЛГ остаются не до конца изученными. Особый интерес вызывают варианты, находящиеся в регуляторных элементах, микроРНК и вторичных структурах, влияющих на экспрессию и функциональную активность генов. Для исследования механизмов влияния таких вариантов использовались данные полногеномного секвенирования 348 пациентов с ЛГ. Предобработка данных включала обрезку адаптеров (Trimmomatic), выравнивание прочтений на референсный геном GRCh38.p14 (BWA v0.7.17) и маркирование дубликатов (MarkDuplicatesSpark v4.3.0.0). Идентификация вариантов осуществлялась с помощью DeepVariant (v1.4.0). Анализировались варианты, прошедшие фильтр «pass». Эпигенетические разметки были получены из ChIP-seq анализа эндотелиальных клеток [1] и фибробластов [2] легочной артерии. Для исследования eQTL использовался GTEx Portal (v10). В качестве контроля рассматривались частоты вариантов европейской популяции из dbGAP ALFA. Для статического анализа использовался критерий Фишера с поправкой на множественное тестирование.

В когорте 348 пациентов было выявлено 31 повторяющихся вариантов, локализованных в эпигенетических областях. Среди этих вариантов 14 имели значимое отличие в частоте встречаемости в группе больных по сравнению с контролем. Три варианта rs55896308, rs7681432, rs7669110 повышают экспрессию KDR в фибробластах. KDR кодирует рецептор фактора роста эндотелиальных клеток, гиперэкспрессия которого способствует ангиогенезу и наблюдается у пациентов с ЛГ [3]. Два варианта rs4754985 и rs11226468, понижающие экспрессию PDGFR в лёгочной и сосудистой ткани, могут ослаблять пролиферативный сигнал в гладкомышечных клетках, участвующий в

ремоделировании сосудов при ЛГ. Однако данные о нокаут-моделях *Pdgfd*^{-/-} свидетельствуют, что этот фактор роста приводит лишь к умеренным сосудистым изменениям и, вероятно, не является обязательным для развития гипертензии [4]. Девять энхансерных вариантов rs11813415, rs33979898, rs4934258, rs7096781, rs7096774, rs11599018, rs191221414, rs144979758, rs117347487 показывают разнонаправленное влияние (3 ↑, 6 ↓) на экспрессию *BMPR1A* в различных тканях. Пониженная экспрессия *BMPR1A* ведет к трансформации эндотелия в мезенхиму [5] и характерна для пациентов ЛАГ [6]. В то же время умеренное повышение экспрессии может играть защитную или компенсаторную роль. Кроме того, у одного из пациентов были найдены потенциально регуляторные варианты при отсутствии патогенных или вероятно-патогенных вариантов в генах, ассоциированных с ЛГ. Восемь вариантов rs80021253, rs78398092, rs75063252, rs112333165, rs6843323, rs4336284, rs7656717, rs7682468 расположены в двух энхансерах (Рис 1) и увеличивают экспрессию гена *KDR* (normalized effect size 0.81). Также энхансеры находятся в одном TAD с промотором этого гена (Рис 2).

Таким образом, генетические варианты, локализованные в энхансерных регионах, могут иметь потенциальный регуляторный эффект и способствовать развитию легочной гипертензии.

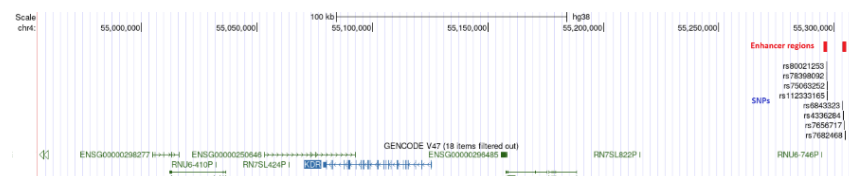


Рисунок 1. Пространственное расположение регуляторных вариантов, локализованных в энхансерных регионах, влияющих на экспрессию гена *KDR*

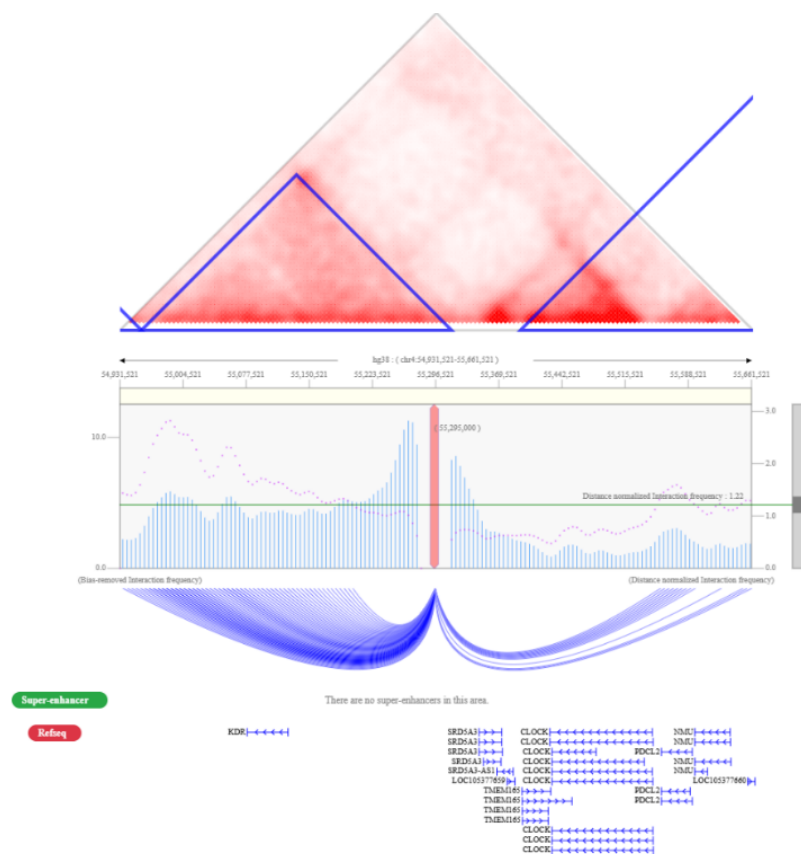


Рисунок 2. Взаимодействия некодирующего варианта rs80021253 с промоторной областью гена KDR

Список литературы

1. Reyes-Palomares, Armando et al. "Remodeling of active endothelial enhancers is associated with aberrant gene-regulatory networks in pulmonary arterial hypertension." Nature communications vol. 11,1 1673. 3 Apr. 2020, doi:10.1038/s41467-020-15463-x
2. Chelladurai, Prakash et al. "Epigenetic reactivation of transcriptional programs orchestrating fetal lung development in human pulmonary hypertension." Science translational medicine vol. 14,648 (2022): eabe5407.

doi:10.1126/scitranslmed.abe5407

3. 3. Tudor, R M et al. "Expression of angiogenesis-related molecules in plexiform lesions in severe pulmonary hypertension: evidence for a process of disordered angiogenesis." The Journal of pathology vol. 195,3 (2001): 367-74. doi:10.1002/path.953
4. 4.Gladh, Hanna et al. "Mice Lacking Platelet-Derived Growth Factor D Display a Mild Vascular Phenotype." PloS one vol. 11,3 e0152276. 31 Mar. 2016, doi:10.1371/journal.pone.0152276
5. 5. Lee, Heon-Woo, et al. "Abstract 17070: Bmpr1a and Its Role in Endomt in Pulmonary Arterial Hypertension." Circulation, vol. 138, suppl. 1, 2018, doi:10.1161/circ.138.suppl_1.17070.
6. 6. El-Bizri, Nesrine et al. "Smooth muscle protein 22alpha-mediated patchy deletion of Bmpr1a impairs cardiac contractility but protects against pulmonary vascular remodeling." Circulation research vol. 102,3 (2008): 380-8. doi:10.1161/CIRCRESAHA.107.161059

**Пайплайн детекции изменения копийности генов и его
применение в Национальной генетической инициативе «100
000+Я»**

Е.И. Павлова¹, М.Х. Гуржиханова², О.С. Киреева¹, Д.В.
Пустошилов¹

¹ООО «Биотехнологический кампус», Москва, Россия
lena.pavlova.9797@mail.ru

²НМИЦ ДГОИ им. Дмитрия Рогачева, Москва, Россия

Изменения копийности генов (CNV, copy number variations) представляют собой значимые структурные вариации в геноме человека, проявляющиеся делецией или дупликацией участков ДНК. Эти изменения играют важную роль в генетической вариабельности популяций и могут влиять на риск развития различных заболеваний, включая онкологические, неврологические

и врождённые расстройства [1], [2]. В связи с этим разработка высокоэффективных методов и подходов к детекции CNV является актуальной задачей современной биоинформатики и медицинской генетики.

Цель данной работы заключается в создании надёжного и точного биоинформатического пайплайна для детекции CNV и последующем его применении для анализа геномных данных, полученных от образцов Национальной генетической инициативы «100 000+Я».

Выбор подходящих инструментов и алгоритмов для анализа CNV проводился на основании обширного обзора современных публикаций и анализа существующих подходов. Основные критерии выбора инструментов включали в себя активность разработки и поддержку проекта, доступность документации и технической поддержки, а также подтверждённые результаты точности и чувствительности на различных типах геномных данных.

В ходе работы был проведён предварительный отбор инструментов на основе таких критериев, как частота обновлений на платформе GitHub, количество звёзд, популярность среди научного сообщества и литературные данные. Среди рассмотренных инструментов были CNVpytor, Spectre, CNVkit, Parascopy, ParseCNV2, ConanVarvar, ifCNV, PennCNV, Sniffles2 и SVIM. На основании указанной оценки были отобраны два наиболее перспективных инструмента — CNVpytor и Spectre, демонстрирующие стабильно высокие показатели и рекомендованные в последних научных публикациях.

Для проверки и сравнения качества детекции CNV, произведенного выбранными инструментами, были использованы данные, полученные с применением технологий секвенирования от Oxford Nanopore Technologies (ONT, инструмент Spectre) и MGI (инструмент CNVpytor). Данные представлены образцами, предоставленными НМИЦ ДГОИ им. Дмитрия Рогачева. Полученные результаты сравнений детекции CNV для различных размеров окон (длины окна: 1k, 10k, 100k, 80k+) позволили оценить применимость и универсальность выбранных методов, что особенно важно в контексте использования разнообразных технологических решений в клинической и исследовательской

практике.

Дополнительно была проведена работа по созданию и валидации специализированного «чёрного списка» (Blacklist), содержащего наиболее часто встречающиеся CNV у здоровых людей. Использование этого списка позволяет исключить из анализа геномные вариации, которые не имеют клинической значимости, и таким образом существенно снизить количество ложноположительных результатов при диагностике заболеваний, связанных с CNV.

Особое внимание в работе было уделено детекции CNV, ассоциированных с различными заболеваниями. В результате анализа выявлены геномные вариации, способные выступать маркерами или потенциальными мишенями для диагностики и персонализированного подхода в медицине. Эти вариации охватывают спектр от онкологических до неврологических расстройств, что подтверждает важность дальнейших исследований и разработки новых диагностических инструментов.

Таким образом, внедренный пайплайн для анализа вариаций числа копий генов представляет собой качественный и оптимальный по требуемым ресурсам подход, способный существенно расширить возможности клинической интерпретации в Национальной генетической инициативе «100 000+Я».

1. Ondrej Pös, et. al. (2021) DNA copy number variation: Main characteristics, evolutionary significance, and pathological aspects, *Biomed J.* T. 44(5). C. 548–559. doi: 10.1016/j.bj.2021.02.003.
2. Lupski J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits, *Trends Genet.* T. 14. C. 417–422. doi: 10.1016/s0168-9525(98)01555-8.

Метод поиска распространяющихся вирусов на основе данных метагеномного секвенирования

В.В.Панова^{1,2}, Н.С.Попов^{1,3}, С.А.Гуров⁴, М.В.Молчанова⁴,
Д.Е.Федоров¹, А.И.Манолов¹

¹*Научно-исследовательский институт системной биологии и
медицины Роспотребнадзора, 117246, Москва, Научный проезд, 18*

²*Московский государственный университет имени М.В.
Ломоносова, факультет биоинженерии и биоинформатики,
119234, Москва, Ленинские горы, д. 1, стр. 73
panova_vv@sysbiomed.ru*

³*Институт медицинской паразитологии, тропических и
трансмиссивных заболеваний им. Е.И. Марциновского, МГМУ им.
И.М. Сеченова, 119435, Москва, Малая Пироговская ул*

⁴*Московский физико-технический институт, 141701, г.
Долгопрудный, Институтский переулок, д.9*

После пандемии COVID-19 возрос интерес к изучению новых инфекций, и метагеномные данные стали важным инструментом для поиска неизвестных патогенов. Существующие подходы к обнаружению инфекций требуют высокого уровня гомологии с уже известными последовательностями. Мы разрабатываем метод идентификации вирусов, предположительно имеющих тенденцию к увеличению распространенности. Он способен анализировать временную динамику распространенности вирусов, не используя референсные базы данных на начальном этапе анализа. Мы хотим продемонстрировать возможность обнаружения последовательностей SARS-CoV-2, не используя последовательность этого вируса в алгоритме поиска. Мы используем три подхода для поиска новых распространяющихся вирусов. Первый из них основан на использовании последовательностей фиксированной длины (k-меров). По изменению представленности и распространенности определенных k-меров, не принадлежащих уже известным

организмам, можно детектировать наличие новых вирусов. Второй подход основан на способности больших языковых моделей создавать эффективные низкоразмерные представления генетических последовательностей. По заполненности ячеек латентного пространства предполагается определить наличие нового вируса. Третий подход предполагает анализ ширины покрытия контигов. Из сырых прочтений собираются контиги. Затем на каждый контиг, собранный из образцов, взятых после появления инфекции, картируется каждое прочтение. У контигов, полученных после появления новой инфекции, ширина покрытия прочтениями, полученными до распространения нового вируса, должна быть больше, чем прочтениями, полученными после распространения нового вируса.

В настоящий момент подходы тестируются на массиве данных тотальной РНК из верхних дыхательных путей до и после пандемии COVID-19 (2020 год) из 367 образцов. Массив данных был собран из NCBI SRA по ключевым словам nasopharyngeal swab, nasopharynx, respiratory OR nasal OR oral NOT gut, Homo Sapiens, Human, RNA-Seq[Strategy], ампликонные данные исключались из рассмотрения вручную. Сумма объемов данных до и после пандемии имеет минимальную разницу между собой.

Эксперимент на смоделированном наборе данных продемонстрировал принципиальную применимость алгоритма, основанного на k-мерах. Получен набор k-меров длиной 15, 17 и 22, встречающихся после 2020 года, но не встречающихся до. Уменьшение вычислительных затрат было произведено следующим образом: каждый k-мер был преобразован в натуральное число в четверичной, а затем в десятичной системе. Множеству k-меров был сопоставлен вектор таких чисел. K-меры длиной 15 и 17 из образцов после 2020 года содержат последовательность вируса SARS-CoV-2 и покрывают почти весь референсный геном SARS-CoV-2 (k-меры длиной 17 покрывают 99.92% генома SARS-CoV-2, длиной 15 – 85.87% процентов генома). Однако k-меры длиной 15 и 17 слишком часто встречаются в уже известных организмах, поэтому в дальнейшем длина k-мера была взята равной 22. Построена база данных kraken2 для k-меров длиной 22 для исключения k-меров, относящихся к геномам организмов, представленных в базах данных, из дальнейшего рассмотрения.

Для тестирования второго подхода была сделана симуляция появления нового вируса во временной серии 10 единообразных образцов, полученных комбинацией в равных пропорциях из организмов, не относящихся к вирусам. Для этого было взято 10 геномов клеточных организмов (про- и эукариот), а также один геном представителя Egovirales, чья последовательность не присутствовала в обучающей выборке модели. Первые 5 образцов – это референсные безвирусные образцы. Ко вторым 5 образцам добавляется по 100, 300, 600, 800, 1000 последовательностей Egovirales. Были извлечены активации feed-forward слоя 6-ого блока трансформера ViraLM для каждой последовательности. Размерность данных была снижена с 789 до 2 с помощью PCA. При этом PCA была обучена на данных без вирусов. Чтобы проверить гипотезу, что в вирусных образцах присутствует повышенное заполнение одной из зон PCA-пространства, были выполнены визуализации сравнения заполненности безвирусных и вирусных образцов (тепловые карты), а также двунаправленный тест Колмогорова-Смирнова. Было показано различие распределений проекции в латентном пространстве (последовательностей геномов) в вирусных и безвирусных образцах.

Из прочтений до и после 2020 года с помощью MEGANIT были собраны контиги – всего 22827613 контигов. К этим контигам планируется применить третий подход. К собранным с помощью MEGANIT контигам из вышеупомянутых реальных данных также была применена модифицированная ViraLM. Визуализируются области со статистически значимыми различиями в представленности между временными интервалами.

Наш подход потенциально применим и показывает результаты на длине $k=15$ и $k=17$. Мы показали возможность отличать вирусные последовательности от невирусных на низкоразмерном представлении последовательностей. Методы нуждаются в дальнейшей доработке.

Исследование проводилось на базе НИИ системной биологии и медицины в рамках государственного задания «Разработка алгоритмов выявления новых уникальных последовательностей ДНК или РНК в метагеномах и их фенотипических характеристик in vitro», номер 12203090069-4.

1. C. Peng et al. (2024) ViraLM: empowering virus discovery through the genome foundation model, *Bioinformatics*, 40(12)

Структурные перестройки генома у пациентов с раком простаты

Е.С. Песковацкова, А.А. Жарикова, А.А. Миронов

*Московский государственный университет имени
М.В.Ломоносова, Факультет биоинженерии и биоинформатики,
Москва, Россия
es-pesk@kodomo.fbb.msu.ru*

Структурные варианты (SV) определяются как генетические варианты размером не менее 50 пар оснований и включают в себя делеции, вставки, дубликации/амплификации, инверсии, транслокации, анеуплоидии, а также более сложные перестройки ДНК. SV могут амплифицировать, изменять структуру онкогенов, или перепрофилировать не кодирующие регуляторные элементы ДНК нарушая экспрессию генов [1]. Многонуклеотидные полиморфизмы переменной длины (MNLP) - небольшие участки от 5 до 18 нуклеотидов, отличающиеся невысоким сходством последовательностей, которые также могут влиять на экспрессию генов аллель специфично [2].

Многие онкологические заболевания сопровождаются структурными перестройками генома. Соматическое образование структурных вариантов происходит на разных стадиях развития опухолей, однако роль герминальных структурных вариантов и других неоднонуклеотидных изменений в геноме недостаточно изучена. Известно, что полиморфизм переменной длины в локусе риска рака простаты 5p15.33 влияет на экспрессию гена *IRX4* [2]. В основном имеющиеся на данный момент исследования полногеномных данных разворачиваются вокруг поиска SNV (single nucleotide variant - однонуклеотидный вариант) и SNP (single nucleotide polymorphism - однонуклеотидный полиморфизм). Геномы больных раком простаты содержат геномные изменения, которые требуют изучения.

Появление технологий секвенирования третьего поколения открыло больше возможностей для анализа SV: длинные прочтения решают проблему некорректного или неоднозначного картирования геномных повторов и, за счет длины, покрывают структурный вариант целиком, что также упрощает обнаружение SV [3]. Многие известные локусы, ассоциированные с заболеваниями, могут быть пересмотрены: ранее найденные ассоциированные с раком SNP могут оказаться лишь в группе сцепления с первопричинной мутацией. С большими возможностями для изучения структурных вариантов, стоит пересмотреть организацию рисков локусов. Был проведен анализ геномных модификаций для больных раком простаты по данным секвенирования Oxford Nanopore. Для этого были проанализированы данные ДНК-секвенирования пациентов с диагностированным раком простаты, разработан биоинформатический протокол для детекции и фазирования SV в данных секвенирования Oxford Nanopore, разработан инструмент для детекции MNLP с помощью подсчета коэффициента вариации по покрытию секвенирования.

Разработанный нами биоинформатический протокол на детектирует и фазирует структурные варианты (SV): их число соответствует значениям, описанным в литературе, а учёт метилирования и гаплотипной фазы расширяет возможности функционального анализа. Обнаруженные SV преимущественно локализируются в мобильных элементах и повторах. Значительная часть SV располагается вблизи SNP, ассоциированных с риском рака предстательной железы (РПЖ), что подчёркивает потенциальную роль крупных перестроек в наследуемой предрасположенности. Дополнительно разработан метод выделения переменных участков генома на основе коэффициента вариации (CoV) покрытия и, позволяющий группировать локальные сигналы изменения покрытия в непрерывные сегменты – предположительно регионы с делециями; такие сегменты часто пересекают повторы и сайты связывания транскрипционных факторов. Часть сегментов, как и SV, перекрывает SNP, ассоциированные с риском РПЖ, что указывает на способность вариаций изменить интерпретацию точечных ассоциаций и углубить понимание генетической архитектуры заболевания.

1. M.R.Cosenza, B.Rodriguez-Martin, J.O.Korbel (2022) Structural variation in cancer: role, prevalence, and mechanisms, Annual Review of Genomics and Human Genetics, 23 (1): 123–152.
2. S.Spisak et al. (2023) A biallelic multiple nucleotide length polymorphism explains functional causality at 5p15.33 prostate cancer risk locus, Nature Communications, 14 (1): 5118.
3. M.Mahmoud et al. (2019) Structural variant calling: the long and the short of it, Genome Biology, 20: 1–14

Алгоритмический метод для вычисления потенциала приспособленности

Д. А. Петров, Д. Н. Иванков

*Сколковский институт науки и технологий,
121205, г. Москва, Большой бул., д. 30, стр. 1
dmitry.petrov@skoltech.ru*

Эпистаз – зависимость эффекта мутации от наличия других мутаций – является главным препятствием для предсказания фенотипа по генотипу. Наличие и количество эпистаза зависит от фенотипа, который измеряется в ходе эксперимента: может так оказаться, что при измерении одной характеристики белка эпистаз есть, а по другой его нет. Семейство функций, являющихся линейной комбинацией сделанных в белке мутаций, называют потенциалом приспособленности. Если измеряемый фенотип является монотонной функцией потенциала приспособленности, то тогда можно говорить, что наблюдаемый эпистаз обусловлен исключительно нелинейностью зависимости фенотипа от потенциала приспособленности. Такой эпистаз называют одномерным; в противном случае эпистаз многомерный.

Исследуя конкретный адаптивный ландшафт, возникает вопрос, является ли эпистаз одномерным или нет? Решение такой задачи требует построения потенциала приспособленности для каждого конкретного набора генотип-фенотип. Такая задача была решена с помощью нейронных сетей для ландшафта GFP [Sarkisyan

et al., 2016]. Однако, нейронные сети не интерпретируемы; более того, в общем случае неизвестно, гарантируют ли они нахождение точного решения.

В настоящей работе мы предлагаем точный алгоритм построения потенциала приспособленности для данного адаптивного ландшафта. Мы использовали подход, основанный на методе линейного программирования. В ходе работы алгоритм аппроксимирует зависимость фенотипа от потенциала приспособленности кусочно-линейной функцией и вычисляет степень нелинейности, подбирая оптимальный (минимизирующий дисперсию) метод регрессии для конкретного ландшафта (линейная, Ridge, LASSO, ElasticNet). В результате работы наш алгоритм отвечает на вопрос, является ли эпистаз в рассматриваемом адаптивном ландшафте одномерным или нет. В случае многомерного эпистаза для очень небольших ландшафтов наш алгоритм выдаёт минимальное количество пар генотип-фенотип, которые нужно убрать для получения одномерного эпистаза.

Исследование выполнено за счёт гранта Российского научного фонда № 25-14-00491, <https://rscf.ru/project/25-14-00491/>.

Список литературы:

1. Z.R. Sailer, M.J. Harms (2017) Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps, *Genetics*, 205:1079-1088.
2. L.A. Esteban, L.R. Lonishin, D.M. Bobrovskiy, G. Leleytner, N.S. Bogatyreva, F.A. Kondrashov, D.N. Ivankov (2020) HypercubeME: two hundred million combinatorially complete datasets from a single experiment, *Bioinformatics*, 36(6):1960-1962.
3. K.S. Sarkisyan, D.A. Bolotin, M.V. Meer, D.R. Usmanova, A.S. Mishin, G.V. Sharonov, D.N. Ivankov, N.G. Bozhanova, M.S. Baranov, O. Soylemez, N.S. Bogatyreva, P.K. Vlasov, E.S. Egorov, M.D. Logacheva, A.S. Kondrashov, D.M. Chudakov, E.V. Putintseva, I.Z. Mamedov, D.S. Tawfik, K.A. Lukyanov, F.A. Kondrashov (2016) Local fitness landscape of the green fluorescent protein, *Nature*, 533(7603):397-401.

Таксономический классификатор для метагеномных сборок для заданных групп организмов

И. С. Петрушин¹, Ю. С. Букин², Д. И. Гутник²

¹*Иркутский государственный университет, ivan.kiel@gmail.com*

²*Лимнологический институт СО РАН, bukinyura@mail.ru*

С ростом доступности технологий секвенирования, становится всё более распространённым анализ микробных сообществ не только по набору маркерных генов (16S, 18S, ITS, rbcL и др.), но и по данным shotgun секвенирования. Получаемые в результате сборки наборы контигов позволяют судить не только о составе сообщества, но и наличии определённых генов метаболизма. Разработанные инструменты биоинформатики используют различные алгоритмы поиска и специализированные базы данных для определения таксономической принадлежности контига: например, в Kaiju [1] поиск контига построено на базе алгоритма BWT (преобразование Барроуза – Уилера), в Kraken [2] – на основе последовательностей k-меров, в SqueezeMeta [3] используется алгоритм DIAMOND. Существенными ограничениями подобных инструментов является редкое обновление баз данных, непрозрачный механизм их формирования, прекращение поддержки продукта (Metavir [4]), а также неявный принцип разрешения конфликтов. Кроме того, статистические подходы на основе k-меров демонстрируют низкую чувствительность.

Предлагаемый нами подход позволит преодолеть описанные недостатки. За основу взята постоянно обновляемая база белковых последовательностей (GTDB [5] для прокариот или NCBI NR [6] для всех организмов), в которой последовательности соотнесены с таксономическим идентификатором. Затем пользователь задаёт требуемый набор таксонов с нужной детализацией (род, семейство, вид и т.п.) и построить базу для поиска для этого набора. Такая специализированная база DIAMOND-поиска позволит избежать ложноположительных результатов поиска, сформировать её в соответствии с методологией исследования и воспроизвести

результат независимо в случае необходимости (например, при рецензировании). Открытый исходный код программы позволяет изменять и дорабатывать её в дальнейшем под нужды пользователей и не потребует поддержки от разработчика.

Публикация осуществляется в рамках проекта «Модернизация курса по применению инструментов биоинформатики для анализа геномных данных», реализуемого победителем грантового конкурса для преподавателей 2024/2025 Стипендиальной программы Владимира Потанина.

1. P. Menzel, K. Ng, A. Krogh (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju, Nat Commun, 7:11257
2. D. Wood, S. Salzberg (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments, Genome Biol, 15:R46
3. J. Tamames, F. Puente-Sánchez (2019) SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline, Front. Microbiol, 9:3349.
4. S. Roux et al. (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis, BMC Bioinformatics, 15:76
5. D. H. Parks et al. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy, Nucleic Acids Research, 50:D785–D794
6. E. W. Sayers et al. (2025) Database resources of the National Center for Biotechnology Information in 2025, Nucleic Acids Research, 53:D20–D29

Возможности протеогеномного анализа бактерии на примере *Mycoplasma gallisepticum* S6

Д. А. Петухова, И. К. Чудинов, Е. В. Корнеев, И. С. Рог, А. В.
Коваленко, О. В. Курилова, Д. С. Матюшкина, В. М. Говорун

*ФБУН НИИ СБМ Роспотребнадзора, Научный проезд 18, Москва,
petuhova_da@sysbiomed.ru*

Введение

Интеграция данных геномики, транскриптомики и протеомики открывает новые возможности для системного анализа микроорганизмов [1]. Одним из перспективных направлений является протеогеномика – подход, позволяющий уточнять аннотацию генома и изучать функционирование биологических систем на нескольких уровнях [2]. Особенно актуален данный подход при работе с организмами, служащими моделью для создания синтетических минимальных клеток, таких как *Mycoplasma gallisepticum* [3].

Настоящая работа посвящена оценке применимости протеогеномного анализа для функционального картирования генов, транскриптов и белков штамма *M. gallisepticum* S6 с использованием онтологии генов (Gene Ontology, GO). Также была поставлена задача оценки согласованности транскриптомных и протеомных уровней экспрессии при нормальных условиях роста.

Материалы и методы

Культивирование *Mycoplasma gallisepticum*. Музейный штамм *Mycoplasma gallisepticum* S6 культивировали при 37 °С в жидкой среде (модифицированная среда Эдвардса), содержащей триптозу (20г/л), Трис (3 г/л), NaCl (5г/л), KCl (1,3 г/л), дрожжевой диализат (5%), термически инактивированную при 56 °С в течение 30 мин лошадиную сыворотку (10%) и глюкозу (1%), pH = 7.4. Замороженную культуру клеток проводили через два пассажа культивирования с 10-кратным разбавлением. В работе использовались клетки в экспоненциальной фазе роста.

Подготовка геномных библиотек и высокопроизводительное секвенирование на платформе MGI и Oxford Nanopore Technologies. Выделение ДНК из моноклональной культуры *Mycoplasma gallisepticum* S6 проводили с использованием буфера СТАВ. Оценку концентрации полученной ДНК проводили с использованием флуориметра Qubit 4.0 и набора Qubit dsDNA HS (Thermo Fisher Scientific, США). Подготовка геномной библиотеки проводилась с использованием набора MGIEasy FS DNA Library Prep Kit (MGI, Китай),

секвенирование выполнялось на платформе DNBSEQ-G400 с использованием набора реагентов DNBSEQ-G400RS High-throughput Sequencing Kit (FCL PE150) и проточной ячейки DNBSEQ-G400RS Sequencing Flow Cell (MGI, Китай).

Для получения длинных прочтений было проведено секвенирование на платформе Oxford Nanopore Technologies. Подготовка библиотеки осуществлялась по протоколу Rapid sequencing gDNA - barcoding V14 (SQK-RBK114.96), высокопроизводительное секвенирование на платформе PromethION (Oxford Nanopore Technologies, Великобритания). Для бейсколлинга использовали программу Dorado v0.9.1 с моделью dna_r10.4.1_e8.2_400bps_sup@v5.0.0. Сборка генома была выполнена с помощью программы Autocycler [4] на основе длинных прочтений, с последующей полировкой с использованием программы Medaka [5], а затем дополнительной полировкой короткими прочтениями с использованием инструментов Polca [6] и PyPolca [7]. Геном был переориентирован с использованием Dnaapler [8], качество сборки оценено с помощью QUAST [9], полнота — по BUSCO [10]. Аннотация проведена с помощью NCBI PGAP [11].

Подготовка мРНК библиотек и секвенирование на платформе MGI. Выделение РНК из 3 биологических повторов моноклональной культуры *Mycoplasma gallisepticum* S6 проводилось с использованием реагента TRIzol LS Reagent (ThermoFisher Scientific, США) и набора реагентов FastPure Cell/Tissue Total RNA Isolation Kit V2 (Vazyme, Китай). Оценку концентрации РНК проводили с использованием флуориметра Qubit 4.0 и набора Qubit RNA HS (ThermoFisher Scientific, США). Качество полученной РНК оценивали с использованием биоанализатора нуклеиновых кислот Tape Station 4150 (Agilent, США) и набора реагентов RNA Screen Tape Agilent. Подготовку РНК библиотек проводили с использованием набора реагентов VAHTS Universal V8 RNA-seq Library Prep Kit for MGI (Vazyme, Китай) по протоколу производителя, деплецию рРНК осуществляли путем гибридизации в буфере HEPES и обработкой дуплекс

специфической нуклеазы (Евроген, Россия). Секвенирование проводили на платформе DNBSEQ-G400 с использованием набора реагентов DNBSEQ-G400RS High-throughput Sequencing Kit (FCL PE150) и проточной ячейки DNBSEQ-G400RS Sequencing Flow Cell (MGI, Китай). Обработка данных осуществлялась с использованием пайплайна ProkSeq [12], включавшего картирование прочтений на собранный геном *M. gallisepticum* S6 с помощью Bowtie2 [13], количественный анализ с FeatureCounts [14], нормализацию по TPM (transcripts per million).

Протеомный анализ. Образцы для восходящего протеомного ВЭЖХ-МС анализа были подготовлены с помощью протеолиза в растворе. Белки подвергались восстановлению с помощью дитиотреитола и алкилированию с помощью йодоацетамида, а затем протеолизу трипсином. Анализ полученного препарата пептидов был проведен с разделением в режиме обращённо-фазовой хроматографии при помощи нанопотоковой системы ВЭЖХ Ultimate 3000 RSLCnano с детекцией тандемным масс-спектрометром высокого разрешения Exploris 480 (Thermo Scientific, США) с ионизацией методом электрораспыления. Регистрация ионов проводилась в режиме данные-зависимого анализа (DDA). Идентификация спектров проводилась с помощью MSFragger [15] против набора последовательностей белков, полученного при аннотации собранного генома *M. gallisepticum* S6.

Функциональная аннотация транскриптов и белков по GO-терминам выполнялась через API QuickGO [16], анализ данных — с помощью скриптов на Python.

Результаты

Полученный геном *M. gallisepticum* S6 представлен одним контигом длиной 987 841 п. н. (GC-состав 31,48%), с полнотой по BUSCO - 98,7%, по результатам аннотации было обнаружено 822 гена.

Картирование транскриптомных прочтений на собранный геном показало высокий процент выравнивания (~ 98%). Из

аннотированных 822 генов демонстрировали экспрессию (TPM > 0) 817 (99,4%).

Протеомный анализ позволил идентифицировать 515, 508 и 519 белков в трёх повторах соответственно, что составляет 68,76%, 67,82% и 69,29% от общего числа аннотированных белков ($n = 749$) *M. gallisepticum* S6 (рис. 1). Такое количество идентифицированных белков связано в первую очередь с пробоподготовкой, где в большей степени выделяются цитоплазматические белки, а не мембранные или секретируемые. Диаграмма также отображает количество идентифицированных пептидов и PSM (peptide spectrum matches), что отражает общее качество масс-спектрометрического анализа и его воспроизводимость между повторами.

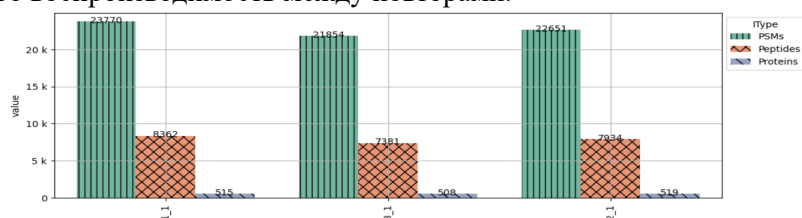


Рисунок 1 – Количество PSM, уникальных пептидов и белков, идентифицированных в трех повторах *M. gallisepticum* S6.

Для оценки соответствия между уровнями транскрипции и трансляции был рассчитан коэффициент корреляции Пирсона между логарифмами средних значений TPM и интенсивности белков. Полученное значение ($r = 0,57$) указывает на умеренную положительную зависимость между экспрессией генов и уровнем соответствующих белков (Рис. 2). Это согласуется с ранее опубликованными данными, в частности с исследованием Бутенко И.О. и соавт. [17], в котором была получена корреляция $r = 0,69$ для 516 белков *M. gallisepticum* S6, выращенной в стандартных условиях.



Рисунок 2 - Диаграмма рассеяния средних значений TPM и средней интенсивности белков (Intensity) для *M. gallisepticum* S6 в логарифмической шкале. Линия тренда — результаты линейной регрессии; $r_p = 0,57$ (коэффициент корреляции Пирсона), $p < 0,001$.

Основные функциональные категории GO-терминов были представлены как среди транскриптов, так и среди белков (Рис. 3, Рис. 4). В частности, наблюдаемые результаты указывают что значительная часть экспрессируемых транскриптов и белков относится к категории “ATP binding”, что отражает важность энергетического обмена в логарифмической фазе роста бактерии и участия АТФ-зависимых процессов в метаболизме клетки. В эту категорию входят ключевые белки, такие как субъединицы АТФ-синтазы, шапероны, транспортеры и ферменты с АТФ-зависимой активностью.

Тем не менее, для комплексного понимания регуляции экспрессии *M. gallisepticum* необходимо проводить протеономный анализ не только в стандартных условиях, но и при различных стрессовых воздействиях. Это позволит выявить чувствительные к

условиям среды пути и расширить функциональную аннотацию генома.

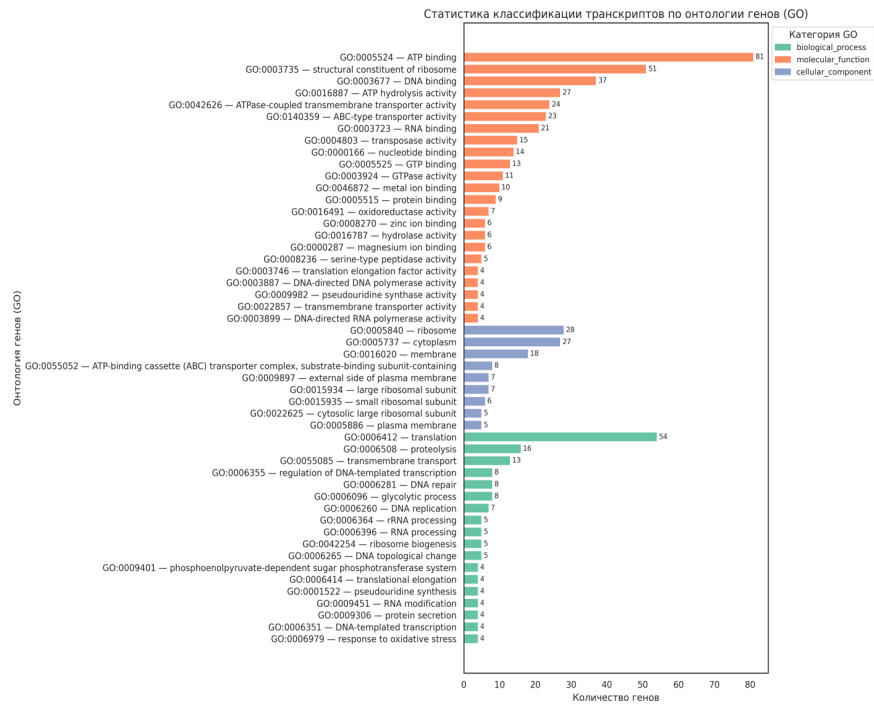


Рисунок 3 - Распределение транскриптов штамма *M. gallisepticum* S6 по категориям онтологии генов (GO).

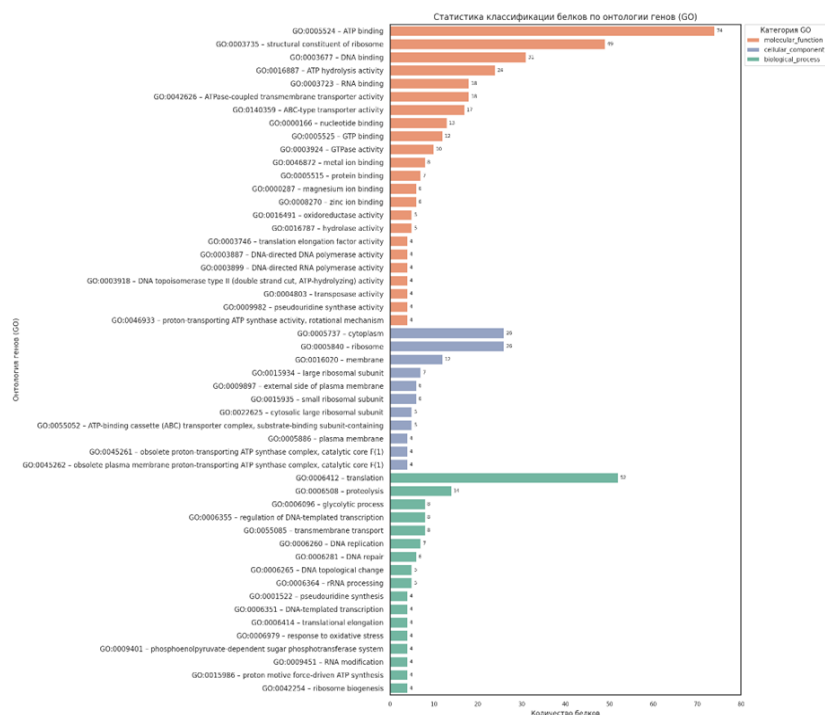


Рисунок 4 – Распределение белков штамма *M. gallisepticum* S6 по категориям онтологии генов (GO).

Работа выполнена на базе НИИ СБМ Роспотребнадзора в рамках госзадания «Резистом».

Использованная литература

1. Aebersold R., Mann M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature*, 537(7620), 347–355.
2. Nesvizhskii A. I. (2014) Proteogenomics: concepts, applications and computational strategies. *Nature Methods*, 11(11), 1114–1125.
3. Hutchison III C. A. et al. (2016) Design and synthesis of a minimal bacterial genome. *Science*, 351(6280), aad6253.

4. Wick R., Stanton T. (2025) rrwick/Autocycler: Autocycler v0.3.0 (v0.3.0). *Zenodo*.
<https://doi.org/10.5281/zenodo.15137029>
5. Medaka: Sequence correction provided by ONT Research. Available from: <https://github.com/nanoporetech/medaka>
6. Zimin A. V., Salzberg S. L. (2020) The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Computational Biology*, 16(6), e1007981.
7. Bouras G. et al. (2024) How low can you go? Short-read polishing of Oxford Nanopore bacterial genome assemblies. *Microbial Genomics*, 10(6), 001254.
8. Bouras G. et al. (2024) Dnaapler: a tool to reorient circular microbial genomes. *Journal of Open Source Software*, 9(93), 596.
9. Gurevich A. et al. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075.
10. Simão F. A. et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212.
11. Tatusova T. et al. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Research*, 44(14), 6614–6624.
12. Mahmud A. K. M. F. et al. (2021) ProkSeq for complete analysis of RNA-Seq data from prokaryotes. *Bioinformatics*, 37(1), 126–128.
13. Langmead B., Salzberg S. L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359.
14. Liao Y., Smyth G. K., Shi W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930.
15. Yu F., Deng Y., Nesvizhskii A. I. (2025) MSFragger-DDA+ enhances peptide identification sensitivity with full isolation window search. *Nature Communications*, 16(1), 3329.
16. QuickGO. Available from: <https://www.ebi.ac.uk/QuickGO/>
17. Butenko I. et al. (2017) Response induced in *Mycoplasma gallisepticum* under heat shock might be relevant to infection process. *Scientific Reports*, 7(1), 11330.

Математическая модель транспорта и накопления доксорибуцина в клетке

С.З. Пирчхадзе, Ф.А. Колпаков, А.В. Звягин

*Научно-технический университет «Сириус», Сириус, 354340,
Россия pirchkhadze@gmail.com*

Аннотация

В данной работе представлен один из модулей многоуровневой модели математической модели фармакокинетики наночастиц при лечении раковых заболеваний, разработанный с использованием платформы BioUML [1]. Модуль отвечает за распределение доксорибуцина в компартментах клетки и при помощи решения системы дифференциальных уравнений рассчитывает концентрацию лекарственного препарата (на примере доксорибуцина) в свободном и связанном состоянии с различными белками, ДНК, в цитоплазме и ядре клетки.

Введение

Раковые заболевания ставят под угрозу здоровье и жизнь миллионов людей по всему миру. В качестве одного из эффективных методов терапии на сегодняшний день используется таргетная доставка действующих веществ при помощи наночастиц – контейнеров обычно из биоразлагаемого материала размерами порядка 1–100 нанометров [2].

Цель работы заключается в построении математической модели, способной учитывать распространение доставляющих препарат наночастиц в организме человека, динамическое развитие раковой опухоли, а также фармакокинетику и фармакодинамику лекарственного препарата.

Структура модели

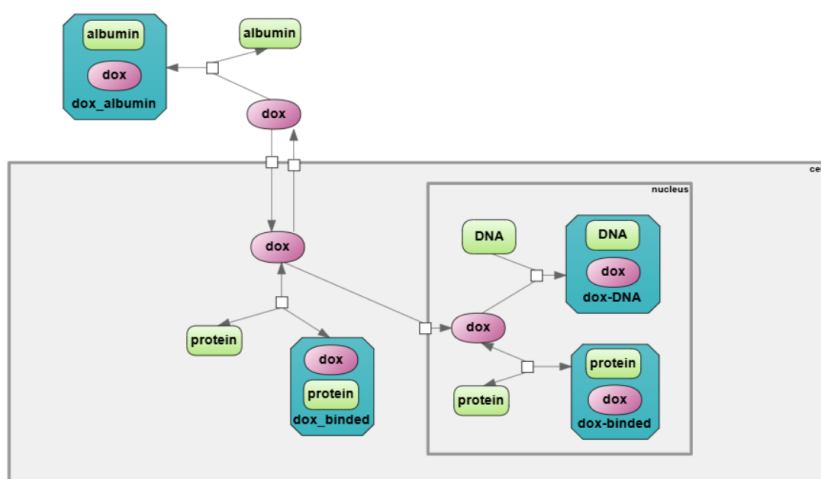


Рисунок 1. Диаграмма модели для учета фармакокинетики доксорубицина в графической нотации SBGN [3].

Структура модели представлена на рисунке 1. За основу взята модель из [4], после чего расширена с учетом дополнительных реакций, поскольку доксорубицин может образовывать связанные комплексы с различными белками как в плазме крови, так и внутри клеток [5]. Около 70% доксорубицина во внеклеточной среде оказывается в связанном состоянии, в котором он гораздо медленнее проникает через клеточную мембрану [6].

Результаты моделирования

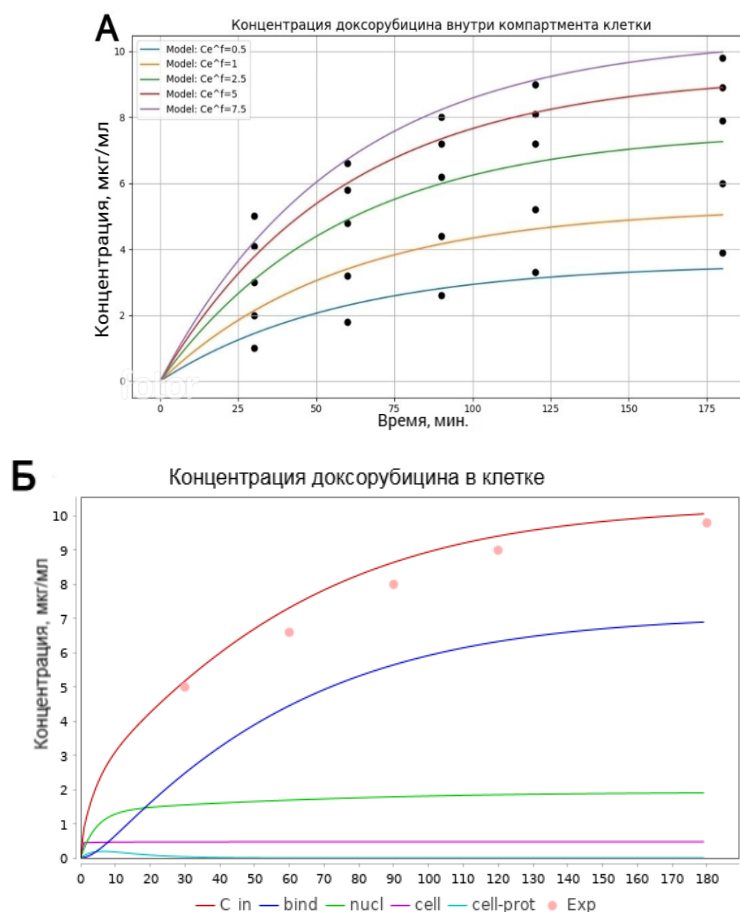


Рисунок 2. Результаты численного расчета модели для оценки концентрации доксорубина в клетке А) При различных начальных дозах препарата; Б) в различных компартментах и состояниях.

На рисунке 2А представлены результаты оптимизации при помощи метода MOPSO [7] с использованием экспериментальных данных из [8]. На рисунке 2Б для одного из экспериментов синим цветом обозначена концентрация доксорубина в связанном с ДНК состоянии. Видно, что его накопление происходит достаточно

быстро и через некоторое время (порядка 180 минут) накопление замедляется, что согласуется с наблюдениями в экспериментах.

Воспроизводимость результатов

Модульная модель доступна в веб-версии платформы BioUML по ссылке: https://biouml2test.biouml.org/bioumlweb/#de=data/Collaboration/Demo/Data/nano/cell_module/

Финансирование

Исследование выполнено за счет гранта Российского научного фонда № 24-14-20031, <https://rscf.ru/project/24-14-20031/>.

Список литературы

1. F. Kolpakov et al. (2022) BioUML—towards a universal research platform, *Nucleic Acids Research*, **50**:W124–W131.
2. M. Yang et al. (2021) The application of nanoparticles in cancer immunotherapy: Targeting tumor microenvironment, *Bioactive Materials*, **6**:1973–1987.
3. N.L. Novère et al. (2009) The Systems Biology Graphical Notation, *Nature Biotechnology*, **27**:735–741.
4. M.T. McKenna et al. (2017) A Predictive Mathematical Modeling Approach for the Study of Doxorubicin Treatment in Triple Negative Breast Cancer, *Scientific Reports*, **7**:5725.
5. O. Tacar et al. (2013) Doxorubicin: an update on anticancer molecular action, toxicity and novel drug delivery systems, *Journal of Pharmacy and Pharmacology*, **65**:157–170.
6. I.V. Zelepukin et al. (2022) Flash drug release from nanoparticles accumulated in the targeted blood vessels facilitates the tumour treatment, *Nature Communications*, **13**(1).
7. Y. Zhang et al. (2015) A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications, *Mathematical Problems in Engineering*, **2015**:1–38.
8. D.J. Kerr et al. (1986) Comparative intracellular uptake of adriamycin and 4'-deoxydoxorubicin by nonsmall cell lung tumor cells in culture and its relationship to cell survival, *Biochemical Pharmacology*, **35**:2817–2823.

**Использование методов докинга и молекулярной динамики
для характеристики влияния замен в генах группы ферментов
ADE**

Пиунов К.А.¹, Гагарин И.Д.¹, Баландин Д.Е.², Мироненко Я.В.²,
Чуров А.В.², Арбатский М.С.², Петрусенко Ю.С.³

¹ ФГАОУ ВО «Уральский федеральный университет им. первого
Президента России Б.Н. Ельцина», Екатеринбург, Россия,
kostya.piunov@gmail.com

² ФГБУ ВО «РНИМУ имени Н.И. Пирогова» Минздрава России
(Пироговский Университет), ОСП «РГНКЦ», Институт изучения
старения, Москва, Россия

³ ООО «Биотек кампус», Москва, Россия

В группе ADE (A β -degrading enzyme) 13 ферментов. Ферменты группы ADE участвуют в деградации амилоида, белка, являющегося одной из причин развития нейродегенеративных заболеваний, в частности, болезни Альцгеймера.

С целью изучения влияния мутаций в генах ферментов группы ADE было проведено полногеномное секвенирование (WGS) образцов крови 585 участников двух исследований - 200 участников из когорты долгожителей (возраст 90+) и 385 участников из когорты Russ-Age (группы с пятилетним шагом от 18 до 90+).

VCF-файлы были проаннотированы по базе ClinVar и dbSNP. Для каждого фермента из группы были получены списки замен, после чего они были нанесены на пространственные структуры этих ферментов из PDB-банка. Затем для каждого фермента была собрана информация о строении активного центра, каталитических остатках, кофакторах и лигандах, необходимых для функционирования фермента.

Для оценки влияния мутаций на функцию фермента был разработан протокол исследования, согласно которому:

1. Проводится подготовка структуры соответствующего фермента из PDB-банка с учетом естественной среды его существования.
2. В аминокислотную последовательность при помощи молекулярной графической системы PyMol вносится замена, соответствующая исследуемой мутации. Методом молекулярной динамики (МД) выполняется релаксация полученной структуры.
3. Проводится сравнение эффективности взаимодействия дикой и мутантной структуры фермента с амилоидом A β 40 при помощи метода молекулярного докинга: по значению scoring-функции, реализованной в гибридном протоколе докинга HDOCK, оценивается энергия связывания амилоидного пептида с активным центром.
4. Проводится молекулярно-динамическое моделирование взаимодействия фермента с амилоидом в водном растворе. Оно выполняется параллельно для дикой и мутантной форм для того, чтобы далее можно было оценить влияние внесенных в структуру замен на функциональные свойства ферментов. Результаты моделирования визуализируются, и оценивается эффективность взаимодействия разных форм фермента с амилоидом (например, по функции расстояния между пептидом и активным центром от времени).

Первым ферментом для исследования из группы ADE был выбран неприлизин – цинк-зависимая мембраносвязанная эндопептидаза. Для проведения моделирования была использована структура с PDB-кодом 1DMT, отражающая строение внеклеточного домена неприлизина (54-749 аминокислотные остатки). Она была предварительно очищена от стабилизирующих лигандов и ингибиторов (оставлен только ион цинка). Была также проведена проверка модели фермента на устойчивость методом молекулярной динамики в течение 100 пикосекунд в водном растворе при 310 К. Очищенная структура показала свою стабильность. Также были подготовлены структуры неприлизина с

найденными в результате WGS мутациями: D377V, Y346C, G224A, I648T, R735L.

С помощью пакета HDOCKlite v1.0 был проведен докинг амилоидного пептида A β 40 (структура 2LFM), состоящего из 40 аминокислотных остатков, к неприлизину. В качестве предпочтительного места стыковки указаны каталитические остатки активного центра фермента: Arg717, Asp650, Glu646, His583, His587, Glu584 и Asn542. Но добиться связывания A β 40 с ними так и не удалось. По всей видимости, для структуры 2LFM не хватает места во внутренней полости неприлизина, содержащей активный сайт, и амилоидный пептид перед каталитическим актом должен изменить свою конформацию. Также известно, что субстраты попадают в полость с активным сайтом через небольшое отверстие в структуре фермента.

В качестве «начальной точки» для проведения молекулярно-динамического моделирования (в программном пакете GROMACS версии 2022.5) взят комплекс с наименьшей энергией A β 40 (2LFM) и неприлизина (очищенный 1DMT), полученный с помощью докинга. В качестве модели растворителя выбрана трехточечная модель воды SPC/E. Электронеутральность системы достигалась путем добавления ионов натрия Na⁺. Моделирование проводилось с использованием силового поля AMBER99SB-ILDN и периодических граничных условий для кубического бокса. Была проведена минимизация энергии для предотвращения стерических столкновений, а также уравнивание системы с приведением к температуре 310 К и давлением 1 бар. Время основного расчета было выбрано 20 нс, с временным шагом 2 фс (10⁶ шагов). По итогам моделирования была визуализирована молекулярно-динамическая траектория системы. Сближения и связывания A β 40 с активным центром неприлизина не произошло. Была также предпринята попытка запустить МД-моделирование при аналогичных параметрах с комплексом, где амилоидный пептид расположен вблизи отверстия в структуре неприлизина (которое, предположительно, должно пропускать молекулы субстрата во внутреннюю полость фермента). В результате моделирования

происходит изменение конформации A β 40: спирализованный участок в центре раскручивается, но тенденции к проникновению амилоидного пептида внутрь неприлизина не наблюдается.

Дальнейшая работа предполагает продолжение исследования неприлизина, а также других ферментов группы ADE с помощью методов молекулярного докинга и динамики. В случаях, когда размер амилоидного пептида мешает связыванию с активным центром исследуемого фермента, рассматривается возможность докинга отдельных отрезков цепи A β 40, состоящих из нескольких аминокислотных остатков.

Использованная литература:

1. C. Oefner et al. (2000) Structure of human neutral endopeptidase (Neprilysin) complexed with phosphoramidon, *J Mol Biol*, 296(2):341–349.
2. T. Sexton et al. (2012) Active site mutations change the cleavage specificity of neprilysin, *PLoS One*, 7(2):e32343.
3. S. Moss et al. (2020) Crystal structure of peptide-bound neprilysin reveals key binding interactions, *FEBS Lett*, 594(2):327–336.
4. S. Moss et al. (2018) High resolution crystal structure of substrate-free human neprilysin, *J Struct Biol*, 204(1):19–25.
5. J.A. Carson & A.J. Turner (2002) Beta-amyloid catabolism: roles for neprilysin (NEP) and other metallopeptidases? *J Neurochem*, 81(1):1–8.
6. N. Schiering et al. (2016) Structure of neprilysin in complex with the active metabolite of sacubitril, *Sci Rep*, 6:27909.
7. Y. Yan et al. (2020) The HDOCK server for integrated protein-protein docking, *Nature Protocols*.
8. M.J. Abraham et al. (2015) GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers, *SoftwareX*, 1:19–25.

Анализ транскриптома Т-клеточных популяций периферической крови пациентов с прогерией

Д.А. Плетенева¹, А.Л. Кунгурцева², П.В. Шелякин³, Д.Б.
Староверов⁴, И.А. Шагина⁴, Е.М. Мерзляк⁴, М.Ю. Мышкин⁴, А.В.
Витебская², О.В. Британова⁴

¹*Институт трансляционной медицины Российского национального
исследовательского медицинского университета имени Н.И.
Пирогова, 117513, Россия, г. Москва, ул. Островитянова, д. 1, стр.
1., andrusyak.darja@yandex.ru*

²*Первый Московский государственный медицинский университет
имени И. М. Сеченова, 119991, Россия, г. Москва, ул. Трубецкая, д.
8, стр. 2.*

³*Центр стволовых клеток Абу-Даби, ул. Махдар Кутуф, Роудхат,
Абу -Даби, 4600, ОАЭ.*

⁴*Институт биоорганической химии имени академиков М.М.
Шемякина и Ю.А. Овчинникова РАН, 117997, Россия, г. Москва,
ГСП-7, ул. Миклухо-Маклая, д. 16/10*

Прогерия (синдром Хатчинсона – Гилфорда, HGPS) – редкое генетическое заболевание, характеризующееся преждевременным старением организма. Оно обусловлено точечной мутацией в гене LMNA, которая приводит к синтезу укороченной формы ламина А – прогерина¹. Накопление прогерина оказывает токсическое воздействие на клетки, нарушая организацию хроматина, и, как следствие, регуляцию экспрессии генов. Предыдущие исследования иммунной системы при HGPS преимущественно проводились на мышинных моделях с нокаутом Lmna². Однако такие модели не отражают в полной мере особенности иммунных клеток при HGPS, поскольку у мышей не происходит накопления прогерина, характерного для этого заболевания.

Целью данного исследования является изучение особенностей Т-клеточного иммунитета пациентов с HGPS. Мы применили анализ данных проточной цитометрии, репертуаров Т-клеточных рецепторов (TCR) и секвенирования РНК единичных клеток для оценки клеточного состава и профилей экспрессии генов периферических клеток крови пациентов с HGPS и здоровых доноров соответствующего возраста.

Согласно данным проточной цитометрии и транскриптома единичных клеток, у пациентов с HGPS не было выявлено снижения доли наивных Т-клеток, что могло бы указывать на иммунологическое старение. Анализ репертуаров TCR косвенно подтвердил это наблюдение, поскольку разнообразие рецепторов у пациентов было сопоставимо с таковым у здоровых сверстников. Снижение разнообразия репертуара TCR является общепризнанным признаком возрастных изменений в иммунной системе³.

Тем не менее, у пациентов с HGPS отмечено накопление GZMK⁺ CD8⁺ Т-клеток – популяции, ранее ассоциированной с иммунным старением⁴. Кроме того, были обнаружены изменения в экспрессии генов, вовлеченных в сигнальные пути WNT, NF-κB и mTOR. Эти изменения могут играть значительную роль в процессах активации, пролиферации и дифференцировки Т-клеток.

Таким образом, данное исследование проясняет ключевые особенности адаптивной иммунной системы у пациентов с HGPS и подчеркивает потенциальные связи между прогерией и физиологическим старением.

1. Eriksson, Maria, et al. (2003) Recurrent de novo point mutations in lamin A cause Hutchinson–Gilford progeria syndrome, *Nature*, 423: 293–298.
2. Toribio-Fernández, Raquel, et al. (2019) Lamin A/C deficiency in CD4⁺ T-cells enhances regulatory T-cells and prevents inflammatory bowel disease, *The Journal of Pathology*, 249: 509–522.

3. Yager, Eric J., et al. (2008) Age-associated decline in T cell repertoire diversity leads to holes in the repertoire and impaired immunity to influenza virus, *The Journal of experimental medicine*, 205:711–723.
4. Mogilenko, Denis A., et al. (2021) Comprehensive profiling of an aging immune system reveals clonal GZMK⁺ CD8⁺ T cells as conserved hallmark of inflammaging, *Immunity*, 54: 99–115.

Алгоритм поиска схожих последовательностей переменных частей Т-клеточных рецепторов с использованием префиксного дерева и редакционного расстояния Левенштейна

Подсытник Михаил Александрович¹, Шугай Михаил Александрович², Власова Елизавета Кирилловна^{1,2}

¹*Факультет информационных технологий и программирования,
Университет ИТМО, Санкт-Петербург, Россия;
michaelpodsytnik@outlook.com*

²*Институт трансляционной медицины, Российский национальный
исследовательский медицинский университет имени Н. И.
Пирогова, Москва, Россия*

Т-клеточные рецепторы (TCR) — гетеродимерные белки на поверхности Т-лимфоцитов, распознающие антигенные пептиды, представленные комплексами МНС, и тем самым запускающие адаптивный иммунный ответ. В каждой цепи рецептора (α и β либо γ и δ) расположены по три комплементарно-определяющие петли (CDR), из которых наиболее переменной является CDR3, формирующаяся при V(D)J-рекомбинации. CDR3-петли обеих цепей образуют центральный контакт с пептидом и в наибольшей степени определяют антигенную специфичность TCR. Сравнение аминокислотных последовательностей CDR3 позволяет группировать рецепторы, которые, как показывают экспериментальные и вычислительные данные, часто распознают

один и тот же эпитоп, поскольку сходство последовательностей коррелирует с близостью пространственных конформаций и, следовательно, со схожей специфичностью.

В данной работе мы представляем алгоритм быстрого поиска похожих CDR3-последовательностей с использованием префиксного дерева (trie), объединенного с алгоритмом вычисления расстояния Левенштейна, адаптированный для анализа крупных репертуаров TCR.

Принцип работы алгоритма можно описать следующим образом. Из заданного репертуара формируется префиксное дерево, в котором каждая вершина кодирует аминокислоту, а полный путь от корня до листа — полную CDR3. Благодаря совместному хранению общих префиксов структура занимает ~40 % меньше памяти, чем плоский список строк того же объёма.

Алгоритм стартует с формирования единственного вектора расстояний Левенштейна между строкой-запросом длиной m аминокислот и пустой последовательностью. Дальнейший обход дерева выполняется рекурсивно: переходя к каждой следующей аминокислоте, мы пересчитываем этот же вектор, добавляя стоимость вставки, удаления или замены. Благодаря тому, что хранится лишь одна строка динамического программирования, объём оперативной памяти остаётся пропорциональным длине запроса $O(m)$ и не зависит от глубины обхода.

Если минимальное расстояние, рассчитанное для текущего узла, превышает заранее заданный порог ошибок, ветвь немедленно отсекается. Напротив, если стоимость остаётся допустимой, все CDR3, заканчивающиеся в этом узле, заносятся в результат.

Для работы с алгоритмом реализован интерфейс, который поддерживает скоринговый режим: вместо простой проверки по расстоянию Левенштейна используется матрица BLOSUM62 (или любая другая пользовательская), и в результат включаются все CDR3, чей суммарный штраф не превышает заданного порога. Реализация отделена от логики обхода, поэтому замена матрицы не требует перестройки дерева.

Поскольку дерево неизменяемо после построения, параллельная обработка пакета запросов распределяется по потокам без синхронизационных накладных расходов, обеспечивая пятикратное ускорение на восьми потоках. На репертуаре объемом 100 тысяч сиквенсов, взятых из базы данных VDJdb, медианное время одного поиска составляет ~2,5 мс на расстоянии Левенштейна равном одному и ~12 мс на расстоянии равном двум, что на два порядка превосходит наивный алгоритм, реализующий попарный перебор. При расчёте всех попарных расстояний между CDR3-последовательностями из VDJdb профильный инструмент TCRdist завершил задачу за 623 секунды, потребовав до 2 Гб оперативной памяти, тогда как наш алгоритм выполнил тот же объём работы за 347 секунд при потреблении не более 100 Мб ОЗУ. Нужно отметить, что полные попарные сравнения являются основной функцией TCRdist, а для нашего алгоритма это лишь одна из возможностей. При этом аналогичных инструментов, реализующих такой функционал, насколько известно авторам, на сегодняшний день не существует.

Высокая вычислительная эффективность алгоритма позволяет подвергать полные TCR-репертуары сканированию на присутствие CDR3-последовательностей, чья антигенная специфичность уже установлена экспериментально (с помощью базы VDJdb и аналогичных ресурсов). Сопоставление с таким референс-набором создаёт широкие практические возможности: выявлять публичные последовательности и тем самым указывать на факт перенесённой инфекции ещё до того, как в крови появляются детектируемые антитела; количественно отслеживать изменение частоты антиген-специфичных клонов при сравнении репертуаров, полученных до и после вакцинации, что упрощает оценку иммуногенности вакцин; отбирать перспективные TCR для клеточной терапии, выделяя клоны, ближайшие по последовательности к рецепторам с уже подтверждённой специфичностью; а также отслеживать динамику тех же клонов в сериях последовательных образцов пациента.

Таким образом, предложенный метод обеспечивает вычислительную основу для масштабируемых, воспроизводимых и клинически ориентированных исследований TCR-репертуаров. В дальнейшем планируется интеграция алгоритма в библиотеку *miru* и разработка специализированных матриц замен, учитывающих частоты аминокислот в позиции CDR3 и вероятность инсерций/делеций. Помимо этого, планируется использование алгоритма для создания полноценной базы данных Т- и В-клеточных репертуаров с возможностью интерактивного поиска схожих последовательностей.

1. Mikhail Shugay et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D419– D427.
2. Dash P. et al., “Quantifiable predictive features define epitope-specific T cell receptor repertoires”, *Nature*, 2017
3. Goncharov, M. et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. *Nat. Methods* 19, 1017–1019 (2022)
4. Anna Postovskaya et al. tcrBLOSUM: an amino acid substitution matrix for sensitive alignment of distant epitope-specific TCRs

Predicting Xenobiotic Metabolism Using a Graph Neural Network and Rule-based approach

Nikita L. Polomoshnov^{1,2}, Anastasia V. Rudik²

¹*Lomonosov Moscow State University, Address: GSP-1, Leninskie Gory, Moscow, 119991, Russian Federation;*

²*Institute of Biomedical Chemistry, Address: 10 bld. 8, Pogodinskaya str., 119121 Moscow, Russia, plomoshnovnikita00@gmail.com*

Xenobiotics (substances that are not present in the human body at concentrations similar to naturally occurring substances) are metabolised by the body to eliminate them [1]. These biotransformations may yield

metabolites that are toxic to humans and may be accompanied by undesirable effects such as carcinogenicity. Acetaminophen, for instance, is metabolised in the liver to form the cytotoxic metabolite N-acetyl-p-benzoquinone imine (NAPQI), which has been demonstrated to cause severe liver damage in overdose [2].

Conventional approaches to metabolism prediction, including in vitro tests, are often costly and time-consuming [3]. Current computational methods, encompassing both rule-based and machine learning approaches, are inherently constrained. A significant number of these methods are confined to the conditional division of metabolism into two phases, the identification of metabolic sites, or the generation of a substantial number of false positives [4][5]. Training machine learning models is also hindered by the paucity of data on the metabolism of drug compounds, which severely limits the size of the model [6].

The fundamental structure of the proposed model encompasses a neural network generator, which is predicated on a set of biotransformation reactions that have been empirically selected, in addition to a filter system that is tasked with the selection of the most probable substrate-metabolite pairs. In this paper, we propose a methodology that integrates a rule-based approach with an attention mechanism and subsequent output filtering using different types of pairwise graph classifiers based on the GATv2 mechanism[10][11]. This approach combines the classical approach via Morgan's fingerprints and a relatively new graph-based approach. The implementation of filtering leads to a reduction in first-order errors, thereby marginally improving metabolite predictions. The training data has been sourced from DrugBank[17], ChEMBL, MetXBioDB[7] and GLORYx[15] databases. Model training to address the lack of metabolism data used pre-training on the enhanced with FAISS[13] algorithm USPTO organic reaction dataset[8], rule augmentation using automatic reaction markup[16], and a positive-unlabelled approach[18] to account for the low representation of real-world examples of metabolites for most drugs and augmentation of the USPTO dataset using automatic retrosynthetic analysis[14]. The hyperparameters of the models were determined using the Optuna[9] framework.

The model showed high accuracy values including F1 score of 0.65, indicating the ability of the model to predict metabolite structure relatively accurately. The new approach can be used for urgent detection of toxic metabolites at the drug development stage, thus avoiding potential health risks.

References

1. Prescott, L.F. (1996) Paracetamol (Acetaminophen): A Critical Bibliographic Review (CRC Press).
2. Andrade, R.J., Chalasani, N., Björnsson, E.S. et al. (2019) Drug-induced liver injury, *Nat Rev Dis Primers*, 5:58.
3. Tyzack, J.D., Kirchmair, J. (2019) Predicting a drug's site of metabolism: a case study, *Chem. Biol. Drug Des.*, 93(4):377–386.
4. Ridder, L., Wagener, M. (2008) SyGMA: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites, *ChemMedChem*, 3:821–832.
5. Rudik, A.V., Bezhtentsev, V.M., Dmitriev, A.V., Druzhilovskiy, D.S., Lagunin, A.A., Filimonov, D.A., Poroikov, V.V. (2017) Meta-Tox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites, *Journal of Chemical Information and Modeling*, 57(4):638–642.
6. Litsa, E.E., Das, P., Kavraki, L.E. (2020) Prediction of drug metabolites using neural machine translation, *Chem. Sci.*, 11:12777–12788.
7. Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., Wishart, D.S. (2019) BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification, *Journal of Cheminformatics*, 11:2.
8. Lowe, D.M. (2012) Extraction of chemical structures and reactions from the literature, Ph.D. thesis (University of Cambridge).

9. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M. (2019) Optuna: A Next-generation Hyperparameter Optimization Framework, arXiv:1907.10902.
10. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y. (2017) Graph Attention Networks, arXiv:1710.10903.
11. Brody, S., Alon, U., Yahav, E. (2021) How Attentive are Graph Attention Networks?, arXiv:2105.14491.
12. Morgan, H.L. (1965) The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service, *Journal of Chemical Information and Modeling*, 5(2):107–113.
13. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., Jégou, H. (2024) The Faiss library, arXiv:2401.08281.
14. Genheden, S., Thakkar, A., Chadimová, V. et al. (2020) AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning, *J Cheminform*, 12:70.
15. de Bruyn Kops, C., Šicho, M., Mazzolari, A., Kirchmair, J. (2021) GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics, *Chemical Research in Toxicology*, 34(2):286–299.
16. Schwaller, P., Hoover, B., Reymond, J.L., Strobelt, H., Laino, T. (2021) Extraction of organic chemistry grammar from unsupervised learning of chemical reactions, *Sci Adv.*, 7(15):eabe4166.
17. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M. (2018) DrugBank 5.0: a major update to the DrugBank database for 2018, *Nucleic Acids Res.*, 46(D1):D1074–D1082.

Влияние эрадикации *Helicobacter pylori* на таксономический состав микробиоты и содержание короткоцепочечных жирных кислот кишечника

Пономарева Е.Р., Маркелова М.И., Синягина М.Н., Куприянова О.В., Абдулхаков С.Р., Григорьева Т.В.

Казанский (Приволжский) федеральный университет, 420008, Казань, Россия elizavettunina@mail.ru

Введение

Инфекционное поражение стенок желудка бактерией *Helicobacter pylori* связывают с развитием гастрита и повышением риска аденокарциномы. Эрадикационная терапия *H. pylori* включает в себя длительный прием двух антибиотиков и является причиной значительных изменений в микробиоте кишечника, что может быть причиной таких нежелательных проявлений, как диарея, запор, вздутие кишечника и др. В связи с этим необходима модификация существующих методов эрадикации *H. pylori*.

Перспективным методом купирования побочных эффектов антибиотикотерапии является применение метабиотиков — метаболитов пробиотических микроорганизмов, и пребиотиков — например, инулина, который стимулирует рост комменсальных микроорганизмов. Наиболее активно изучаемым метаболитом кишечной микробиоты является короткоцепочечная жирная кислота (КЖК) — масляная кислота, продукт метаболизма кишечной микробиоты, а также основной энергетический субстрат для эпителиальных клеток стенки кишечника. Было доказано, что прием масляной кислоты и инулина во время эрадикации *H. pylori* уменьшает побочные эффекты, связанные с приёмом антибиотиков, и способствует восстановлению микробиоты кишечника [1]. В связи с этим актуальным является определение представителей микробиоты, ассоциированных с содержанием в кишечнике короткоцепочечных жирных кислот.

Материалы и методы

В настоящем исследовании приняли участие 15 пациентов с подтвержденной инфекцией *H. pylori*. Эрадикационная терапия включала в себя: в течение 14 дней эзомепразол 20 мг, кларитромицин 500 мг, амоксициллин 1000 мг, висмута трикалия дицитрат 240 мг два раза в день, масляная кислота (250 мг) + инулин (250 мг) (Закофальк) дважды в день. Образцы кала пациентов были собраны до начала лечения (точка 1), сразу после окончания лечения (точка 2) и спустя месяц после окончания лечения (точка 3). Из полученных образцов была выделена ДНК с помощью набора FastDNA Spin Kit. Библиотеки для секвенирования были подготовлены с помощью набора NEBNext Ultra II. Секвенирование было проведено на платформах FASTASeq 300 и HiSeq 1000. Прочтения были картированы на сборку генома человека UCSC hg19 с помощью программы bowtie2 [2]. Риды, которые не были картированы, проанализированы с помощью программы MetaPhlAn4 для оценки таксономического состава [3]. Количественное определение КЖК (уксусной, пропионовой, масляной, изомасляной, валериановой, изовалериановой, капроновой и изокапроновой кислот) в кале осуществляли с помощью газовой хроматографии с пламенно-ионизационным детектированием на приборе CLARUS 680 (PerkinElmer, США). Статистически значимые отличия в содержании КЖК и в таксономическом составе были выявлены с помощью теста Краскела-Уоллеса с поправкой на множественное сравнение методом Бенджамини-Хохберга. Для определения корреляции между содержанием КЖК и таксономическим составом был использован метод Спирмена. За критический уровень значимости было принято значение $P < 0.05$.

Результаты

Сразу после окончания терапии по сравнению с исходной точкой наблюдается достоверное снижение относительной представленности видов *Bifidobacterium adolescentis*, *Bifidobacterium pseudocatenulatum*, *Collensiella aerofaciens*, *Candidatus Cibibacter qucibialis*, *Coprococcus eutactus*, *Fusicatenibacter saccharivorans*, *Roseburia faecis*, *Ruminococcus*

bicirculans и достоверное увеличение представителей видов *Enterococcus faecalis*, *Enterococcus faecium* (Рисунок 1). Спустя месяц после терапии представленность всех видов бактерий, кроме *Bifidobacterium adolescentis* и *Bacteroides uniformis*, имеет тенденцию к восстановлению, причем доля *Enterococcus faecalis*, *Enterococcus faecium*, *Roseburia faecis* и *Ruminococcus bicirculans* достигла исходных значений.

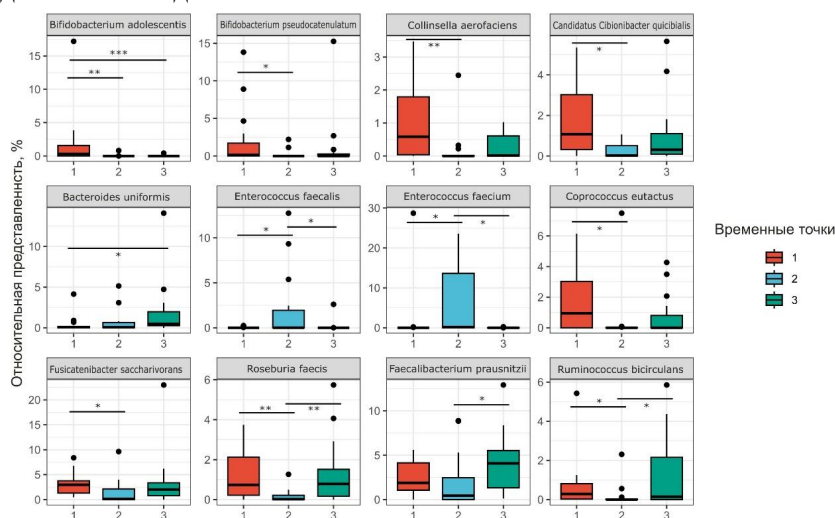


Рис. 1 – Изменения относительной представленности видов во временных точках

Интересно, что несмотря на прием масляной кислоты, ее уровень был достоверно ниже сразу после эрадикационной терапии, и имел тенденцию к восстановлению до исходных значений через месяц после окончания терапии. Помимо этого, было отмечено достоверное снижение уровня валериановой кислоты с тенденцией к возвращению к исходной концентрации (Рисунок 2).

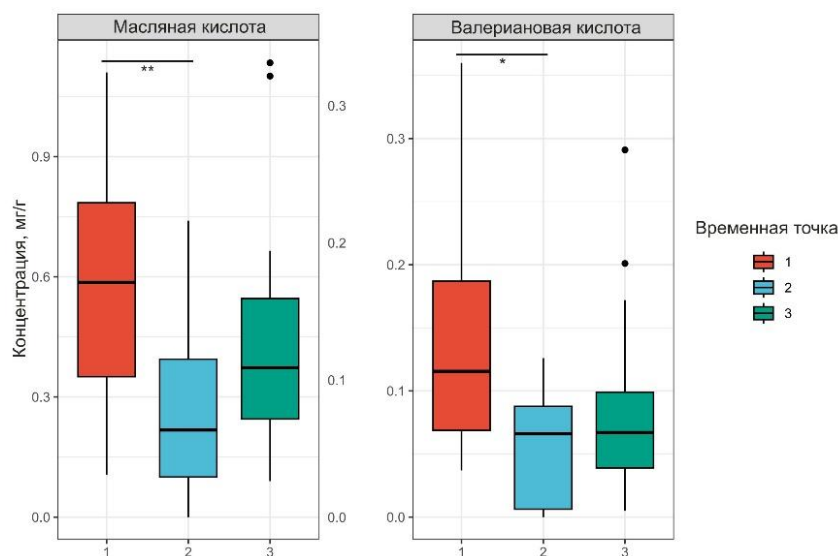


Рис. 2 – Изменение концентрации КЖК в кале во временных точках

Был выявлен ряд представителей микробиоты кишечника, чья относительная представленность достоверно коррелировала с уровнем КЖК (Рисунок 3). Так, с содержанием масляной кислоты положительно коррелировала доля видов *Faecalibacterium prausnitzii*, *Fusicatenibacter saccharivorans*, *Blautia* sp.MCC283, *Dorea longicatena*, *Bifidobacterium pseudocatenulatum*, а отрицательно — *Enterococcus faecium*. Кроме того, были выявлены достоверные положительные корреляции *Faecalibacterium prausnitzii* и *Blautia* sp.MCC283 с концентрацией валериановой кислоты, *Dorea longicatena* с концентрацией изомасляной кислоты, *Bifidobacterium pseudocatenulatum* с концентрацией уксусной кислоты, *Coprococcus eutactus* с концентрацией капролиновой кислоты, и *Escherichia coli* с концентрацией изокапролиновой кислоты, и отрицательные — *Bacteroides uniformis* и *Enterococcus gallinarum* с концентрацией валериановой кислоты, *Enterococcus gallinarum* с концентрациями изовалериановой и изомасляной кислот, *Enterococcus faecium* с

концентрациями уксусной и каприловой кислот и *Phocaeicola dorei* и *Bacteroides uniformis* с концентрацией изокаприловой кислоты.

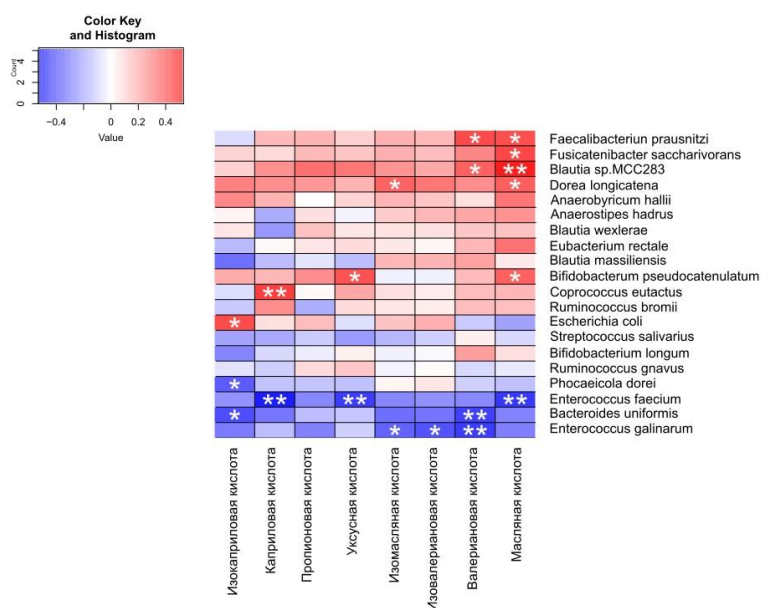


Рис. 3 – Корреляция бактериальных видов с концентрацией КЖК в кале у пациентов

Выводы

Несмотря на использование во время эрадикационной терапии препарата масляной кислоты и инулина наблюдается снижение представленности комменсальных бактерий, а также снижение концентрации масляной кислоты в кале. Виды *Bifidobacterium pseudocatenulatum*, *Coprococcus eutactus*, *Fusicatenibacter saccharivorans* и *Faecalibacterium prausnitzii*, чья представленность была снижена после эрадикации *H. pylori*, положительно коррелировали с содержанием масляной кислоты. Однако через месяц после эрадикации отмечается тенденция к восстановлению исходного состояния микробиоты. Таким образом,

мы выявили микробный паттерн, ассоциированный со снижением содержания масляной кислоты в кишечнике после эрадикационной терапии *H. pylori*.

Работа выполнена за счет средств субсидии, выделенной КФУ на выполнение государственного задания в сфере научной деятельности (№ FZSM-2023-0013).

Источники

1. S. Abdulkhakov et al. (2024) Butyric Acid Supplementation Reduces Changes in the Taxonomic and Functional Composition of Gut Microbiota Caused by *H. pylori*, *Microorganisms* 12(2): 319
2. B. Langmead, S. Salzberg (2012) Fast gapped-read alignment with Bowtie 2, *Nat Methods* 9: 357–359
3. Blanco-Miguez A. et al. (2022) Extending and improving metagenomic taxonomic profiling with uncharacterized species with MetaPhlAn 4, *Biorxiv* 2022-08

Copy number variation (CNV) detection using DNA language model GENA-LM within the framework of preimplantation genetic testing

A.A. Popov^{1,2}, Y.K. Stepanchuk^{3,4}, A.M. Ivanov^{3,4}, M.M. Gridina^{3,4},
V.S. Fishman^{2,3,4}

Corresponding author's email: Andrey.Popov@skoltech.ru

¹*Skolkovo Institute of Science and Technology, Moscow, Russia*

²*Artificial Intelligence Research Institute (AIRI), Moscow, Russia*

³*Institute of Cytology and Genetics, Novosibirsk, Russia*

⁴*Novosibirsk State University (NSU), Russia*

Genetic abnormalities such as copy number variations (CNVs) contribute to developmental disorders and non-viability of embryos^{1,2}, underscoring the critical need for their detection in IVF protocols. Many current sequencing-based CNV analysis methods often rely on control

samples, a dependency that introduces reproducibility challenges due to technical variability in sample preparation or sample quality³. These limitations highlight the urgency for universal, control-independent methods. Our project addresses this gap by leveraging the GENA-LM4 pre-trained DNA language model to predict the representation (coverage) of genomic regions in sequencing data, eliminating control-sample reliance and improving diagnostic accuracy.

DNA coverage in a given region is influenced by two factors: (1) DNA composition biases (GC content, chromatin structure, etc.) and (2) sample preparation specific biases (e.g. type of enzymes, amplification conditions, quality of isolated DNA)^{5–7}. To address the former, our model employs a GENA-LM transformer pre-trained on DNA sequences to capture contextual dependencies. For the latter, sample-specific heads, fully-connected networks, are added to adapt the model to individual sample characteristics. The loss function, PoissonNLLLoss, aligns with the discrete nature of sequencing counts, informed by Lander and Waterman’s DNA representation distribution theory (1988). We trained the model using data from 81 embryo samples with clinically confirmed CNVs.

In this study, we used an unusual approach to model training that deviates from conventional methodologies by omitting the traditional partitioning of genomic data into distinct training and validation sets. Instead, the genome was segmented into contiguous, non-overlapping regions (3–4 kilobases in length), with each bin assigned a real coverage value derived from whole-genome sequencing data. The model was subsequently trained to predict sequence coverage for every genomic bin across samples, utilizing the same dataset for both training and validation phases. This approach enhances the model’s capacity to predict sample-specific coverage in datasets exhibiting high coverage variability not attributable to copy number variation (CNV), such as variability arising from technical biases, sequencing biases or stochastic noise inherent to low-coverage samples. Since most samples are dominated by CNV-free regions, the model systematically underestimated coverage in CNV-containing bins, predicting values expected under normal conditions. This discrepancy enabled effective normalization: the observed-to-predicted coverage ratio amplified the contrast between ‘healthy’ and ‘pathological’ regions, detecting abnormalities without control samples.

By inherently accounting for technical and biological variability, our approach isolates CNV-driven coverage distortions.

Preliminary results demonstrate that our GENA-LM-based tool outperforms established CNV detection methods, including the open-source GATK8 and the commercial Genomenal platform9, reaffirming the utility of language models in deciphering complex genomic patterns4. Advantages of the approach include elimination of control samples, reduced cost and variability of results, and versatility: the program can further be applied, not only in preimplantation genetic testing, but also in other experiments where variability of coverage is observed in different parts of the genome, such as in oncology (analysis of tumour biopsies).

This work was supported by the BlueSkyResearch Grant BSR-2024-42.

1. Lander, J. & Ware, S. M. Copy Number Variation in Congenital Heart Defects. *Curr Genet Med Rep* 2, 168–178 (2014).
2. Zhang, R. et al. Prevalence of chromosomal abnormalities identified by copy number variation sequencing in high-risk pregnancies, spontaneous abortions, and suspected genetic disorders. *Journal of International Medical Research* (2019) doi:10.1177/0300060518818020.
3. Hong, C. S., Singh, L. N., Mullikin, J. C. & Biesecker, L. G. Assessing the reproducibility of exome copy number variations predictions. *Genome Med* 8, 82 (2016).
4. Fishman, V. et al. GENA-LM: a family of open-source foundational DNA language models for long sequences. *Nucleic Acids Research* 53, gkae1310 (2025).
5. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40, e72 (2012).
6. Baumgarten, S. & Bryant, J. Chromatin structure can introduce systematic biases in genome-wide analyses of *Plasmodium falciparum*. *Open Res Eur* 2, 75 (2022).
7. Galià-Camps, C., Pegueroles, C., Turon, X., Carreras, C. & Pascual, M. Genome composition and GC content influence loci distribution in reduced representation genomic studies. *BMC Genomics* 25, 410 (2024).

8. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297–1303 (2010).
9. NGS Wizard - DNA analysis and automation tool | Genomenal - bioinformatics applications. <https://genomenal.com/>.

Phylogenomic and Functional Reassessment of *Pseudogymnoascus destructans* Across Mitochondrial and Nuclear Scales

Ilia V. Popov, Igor V. Popov, Alexey M. Ermakov

*Faculty of Bioengineering and Veterinary Medicine, Don State
Technical University, Rostov-on-Don, Gagarina square, 1,
ivpopov@donstu.ru*

Pseudogymnoascus destructans is a psychrophilic fungus and the causative agent of White-Nose Syndrome, an epizootic disease affecting bat populations in the Northern Hemisphere. Although high-quality nuclear and mitochondrial genome assemblies are available, key aspects of its evolutionary origin and functional genome organization remain insufficiently characterized. In particular, mitochondrial divergence timing and large-scale nuclear protein function remain largely unresolved.

We conducted a two-tiered analysis targeting both mitochondrial and nuclear levels. Phylogenetic reconstruction and divergence dating were performed using 13 mitochondrial protein-coding genes conserved across Leotiomycetes. Topologies were inferred under composition-aware models, and divergence times estimated using a strict clock and mtREV substitution matrix. Substitution rate priors were derived from Sordariomycetes mitochondrial datasets [1]. This approach is justified by the close phylogenetic relationship between Sordariomycetes and Leotiomycetes, and by the fact that both groups exhibit equivalent fossil-calibrated root ages in previous phylogenies [2]. The analysis delineated a monophyletic clade including *P. destructans*, *P. pannorum*, and several cold-adapted Antarctic fungi, with divergence events placed in a paleoclimatically dynamic period.

To address the annotation deficit of the *P. destructans* proteome, we reanalyzed all 20,334 protein sequences available in RefSeq. Of

these, 3,909 were previously annotated, while 16,425 were labeled as “uncharacterized protein”. Using eggNOG-mapper with orthology-based inference, we assigned Clusters of Orthologous Groups (COG) categories to 12,206 proteins in total. Among them, 2,137 COG assignments came from the subset of previously annotated proteins, while the remaining 10,069 were assigned to previously uncharacterized sequences. This represents a 470% increase in COG-based functional classification coverage compared to the initial annotations. The updated functional profile revealed a major enrichment in proteins associated with carbohydrate metabolism, intracellular transport, and stress response. These functional expansions suggest enhanced capabilities for host interaction and environmental persistence, underscoring the ecological adaptability of *P. destructans*.

By linking mitochondrial genome dynamics with nuclear functional expansion, this study uncovers key genomic traits that may underlie the ecological success and pathogenicity of *P. destructans*.

The work was carried out within the framework of the RSF grant № 25-24-00351

1. G.Aguileta et al. (2014) High Variability of Mitochondrial Gene Order among Fungi, *Genome Biology and Evolution*, 6:451-465
2. C.Beimforde et al. (2014) Estimating the Phanerozoic history of the Ascomycota lineages: Combining fossil and molecular data, *Molecular Phylogenetics and Evolution*, 78:386-398

**Изучение эволюции и классификация систем
рестрикции-модификации, содержащих эндонуклеазу
рестрикции семейства RE_AlwI_N6_N4_Mtase**

А.И. Попова, А.В. Алексеевский, И.С. Русинов, С.А. Спирин

*НИИ ФХБ имени А.Н.Белозерского МГУ, 119992, Москва,
Ленинские горы, дом 1, стр 40, diatrimma@mail.ru*

Системы рестрикции-модификации (Р-М) —
прокариотические системы, включающие в себя как минимум два

фермента: сайт-специфическую эндонуклеазу рестрикции (ЭР) и сайт-специфическую ДНК-метилтрансферазу (МТазу) узнающих одну и ту же короткую последовательность ДНК. Они защищают клетку от чужеродной ДНК, внося вклад в выживаемость и разнообразие бактериальных сообществ [1].

Бактериофаги используют множество механизмов для ухода от систем рестрикции-модификации. Однако системы Р-М быстро эволюционируют, что приводит к постоянному противостоянию между бактериями и фагами. Изучение эволюции Р-М систем становится важно, например, для фаговой терапии.

Мы изучали эволюцию систем Р-М одного из гомологичных семейств, а именно, включающих ЭР семейства RE_AlwI и МТазы семейства N6_N4_Mtase

Материал был взят из базы данных REBASE версии 408 (04.08.2024). Для отбора систем использовались HMM-профили из банка Pfam. Отбирались системы, в ЭР которых профилями детектировались эволюционные домены семейства RE_AlwI, и которые включали в себя одну или две МТазы. При этом находками считались все белки, где вес соответствующего профиля был выше рекомендуемого в Pfam порога. Среди систем были отобраны те, в МТазах которых были найдены эволюционные домены семейства N6_N4_Mtase.

Было найдено 43 системы, в том числе четыре системы, у которых аннотированы не все ферменты. Гены недостающих белков были получены при аннотации MetaGeneAnnotator с последующим поиском гомологичных белков при помощи blastx.

Все белковые последовательности были разбиты на кластеры программой CD-HIT по 100% идентичности последовательностей. Из систем, идентичных по составу, было оставлено по одному представителю, всего 27 различных систем. Было построено дерево ЭР (рис. 1).

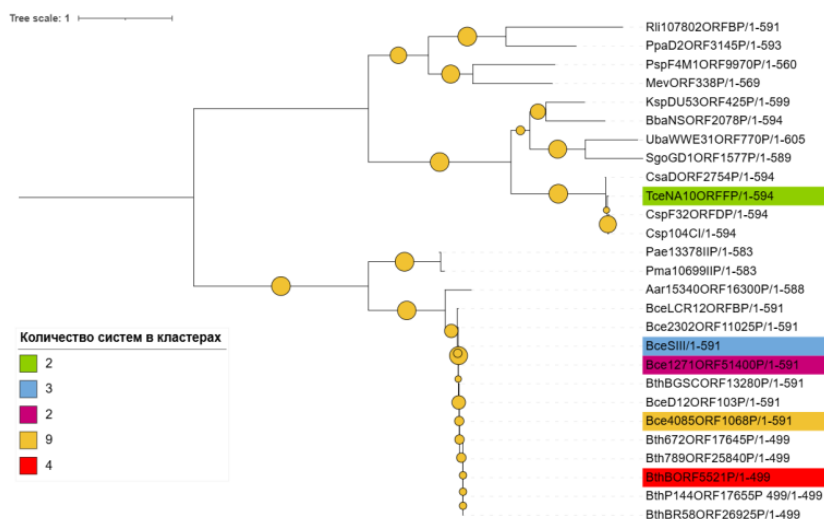


Рис 1. Филогенетическое дерево ЭР класса RE_AlwI + N6_N4_Mtase. Цветами отмечены представители группы идентичных систем. Числа обозначают количество систем в каждой группе

Чтобы выяснить филогенетические связи ЭР данного класса с другими ЭР того же семейства, ЭР из систем классов RE_AlwI + MethyltransfD12 и RE_AlwI + DNA_Methylase (здесь MethyltrasfD12 и DNA_Methylase — семейства МТаз) были разбиты программой CD-HIT на кластеры с порогом 40% идентичности. Затем уникальные ЭР из систем класса RE_AlwI + N6_N4_Mtase были добавлены к референсам этих кластеров и построено общее дерево полностью (рис. 2).

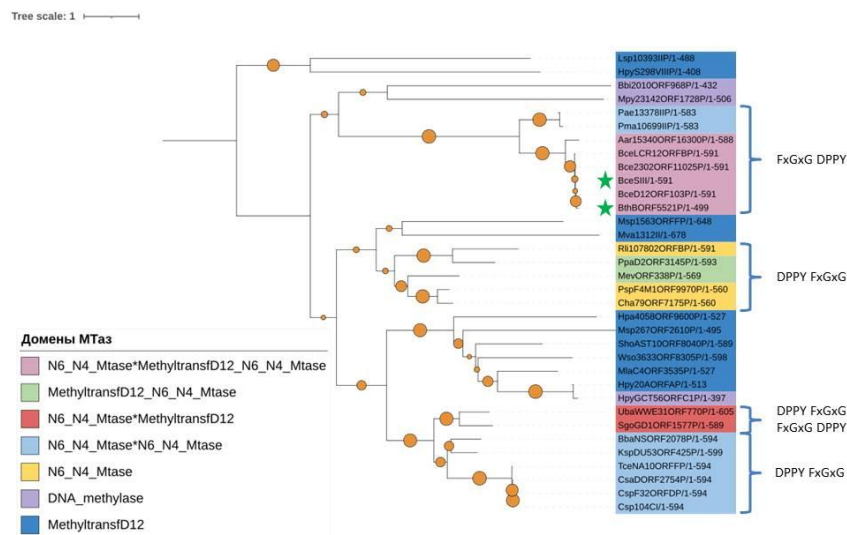


Рис 2. Филогенетическое дерево, реконструированное на основе последовательностей ЭР. Окрашивание названий ЭР обозначает принадлежность данного фермента к системе с определенным набором доменов МТаз. Для ЭР, принадлежащих системам с МТазой семейства N6_N4_Mtase (в части из них присутствуют также МТазы с доменами семейства MethyltrasfD12), указан порядок следования мотивов в МТазе. Размер желтых кругов на ветвях соответствует бутстреп-поддержке. Зелёными звёздами отмечены представители двух групп идентичных ЭР.

Класс систем RE_AlwI + N6_N4_Mtase интересен крайней разнородностью по распространённости систем. Он содержит всего 43 систем, найденных в прокариотических организмах относящихся к разным, в том числе неблизкородственным родам. Нами были отобраны 38 систем с двумя МТазой семейства N6_N4_Mtase и 5 — с одной МТазой того же семейства. В МТазе этих систем консервативные мотивы FxGxG и DPPY могут идти в разном порядке. В МТазе, у которых FxGxG расположен в N-концевой части, основным (т.е., дающим больший вес) профилем Pfam может

быть MethyltransfD12, но может быть и N6_N4_Mtase. В последнем случае профиль N6_N4_Mtase "рвётся": его начальная часть выравнивается в белке после конечной. Тем самым МТазы этих систем фактически относятся к двум разным семействам, различающимся циклической перестановкой субдоменов. ЭР из систем с МТазами, относящихся к этим двум семействам, разделены на филогенетическом дереве. У МТаз, у которых домен FxGxG расположен в N-концевой части, часто в последовательности перекрываются находки профилей N6_N4_Mtase и MethyltransfD12. ЭР из систем с такими МТазами образуют отдельную кладу на дереве всех ЭР семейства RE_AlwI (рис. 2). Был выявлен интересный случай: в ходе эволюции произошла потеря второго домена у МТаз из систем Rli107802ORFB, PspF4M1ORF9970P, Cha79ORF7175P. В первом случае это может быть связано с уменьшением размера генома *Rhodoluna limnophila* (1,4 млн п.н. – размер генома) – хозяина этой системы.

Резюмируя, можно сказать, что класс RE_AlwI + N6_N4_Mtase фактически представляет собой два независимых по происхождению класса. Оба класса содержат мало систем, которые, тем не менее, встречаются в бактериях далёких друг от друга таксономических групп, что может свидетельствовать об их склонности к горизонтальным переносам, а также о том, что они, как правило, быстро утрачиваются бактериями-хозяевами.

1. Williams, R. J. (2003) Restriction endonucleases: classification, properties, and applications, Mol. Biotechnol., 23:225-244.
2. Pitt, A. (2019) *Rhodoluna limnophila* sp. nov., a bacterium with 1.4 Mbp genome size isolated from freshwater habitats located in Salzburg, Austria, Int J Syst Evol Microbiol 69:3946-3954.

Влияние растительной диеты на состав деструкторов сульфохиновозы в кишечном микробиоме

В.О. Пыркин¹, А.А. Ануров², А.А. Рыбина³

¹*Московский Государственный Университет имени М.В.
Ломоносова, Москва, 119991, Россия, vladislw@yandex.ru*

²*Институт молекулярной биологии им. В.А. Энгельгардта РАН,
Москва, 119991, Россия*

³*Сколковский институт науки и технологий, Москва, 121205,
Россия*

Сульфохиновоза (SQ, 6-deoxy-6-sulfo-d-glucopyranose) – сульфосахар, продукт окисления растительного сульфолипида, встречающийся повсеместно от морских до наземных экосистем [1]. Каждый экотоп характеризуется определённым составом деструкторов SQ, в основном это аэробные гетеротрофные бактерии и участники биогеохимического цикла серы [2]. Однако, катаболизм SQ в кишечнике имеет ряд важных особенностей. Данный процесс происходит в анаэробных условиях и некоторые кишечные бактерии способны разлагать SQ до сероводорода (H₂S) в результате межвидовой передачи промежуточных продуктов деградации SQ [3]. В свою очередь H₂S может выступать в качестве потенциального генотоксина, вызывать колоректальный рак и воспаление кишечника [4]. Однако, отмечается, что оба процесса (разложение SQ и продукция H₂S) являются ключевыми атрибутами кишечника человека [3]. Исходя из выше сказанного, исследования метаболических путей деструкции SQ в кишечнике представляет интерес в вопросах диагностики и прогнозирования вышеперечисленных заболеваний. А растительное происхождение SQ, в свою очередь, привлекает интерес анализа различных типов диеты.

Целью данного исследования послужил поиск локусов деградации SQ в метагеномах кишечных микробиомов, а также анализ таксономического состава SQ деструкторов между микробиомами вегетарианцев и всеядных.

В качестве объекта исследования нами были выбраны два набора данных UHGP (Unified Human Gastrointestinal Genome) [5] и AGP (American Gut Project) [6]. UHGP представлял собой метагеномные сборки белковых последовательностей, а AGP набор данных ампликонного и полногеномного (WGS) секвенирования с указанием типа диеты. Для поиска ферментов, участвующих в деградации сульфохинозы (SQ), мы собрали набор белковых последовательностей, включающий как универсальные ферменты, характерные для различных путей SQ-деградации (например, сульфохинозовидаза (YihQ или SqgA), изомеразы (YihS или qvD) и мутаротазы YihR), так и ферменты, специфичные для отдельных путей. В частности, были учтены компоненты сульфотранскетолазного (sulfo-TK: транскетолаза SqwGH, изомеразы SqwI), сульфо-Эмбдена–Мейергофа–Парнаса (sulfo-EMP: киназа (YihV или SqjK), альдолаза (YihT или SqjA)) и сульфотрансальдолазного (sulfo-TAL: трансальдолаза SqvA) путей. На основе ко-локализации соответствующих генов, мы определили SQ-деградирующие локусы, классифицировали их по фенотипу (пути деградации SQ) на основе генов, кодирующих ключевые ферменты и определили таксономический состав деструкторов SQ. Анализ таксономии показал, что большинство SQ-деградирующих локусов в микробиоме кишечника ассоциировано с представителями трёх филумов: Pseudomonadota (sulfo-EMP), Bacteroidota (sulfo-TK) и Firmicutes (sulfo-EMP, -TK, -TAL). Стоит отметить, что Firmicutes обладали наиболее разнообразным набором путей деградации SQ. Локусы также были обнаружены у Actinomycetota (sulfo-TAL), Verrucomicrobiota (sulfo-EMP) и других филумов, однако их доля в кишечном микробном сообществе оказалась значительно ниже.

После кластеризации локусов на них было проведено картирование прочтений из AGP. В результате было рассчитано относительное содержание (в RPKM — количество прочтений на килобазу на картированные риды) локусов SQ метаболизма в микробиоте кишечника человека. Для оценки вклада различных факторов — таких как диета (Diet), возраст (Age), пол (Sex), индекс массы тела (BMI), а также присутствие бактериальных видов (Species) с определенными фенотипами по деградации SQ — были построены две линейные модели: обычная модель наименьших квадратов (OLS) и модель с коррекцией для множественных

переменных (алгоритм MaAsLin2). В обеих моделях использовалась следующая формула: $RPKM \sim Diet * Species + Age + BMI + Sex$. На основании полученных результатов были обнаружены зависимости между типом диеты и составом деструкторов SQ в кишечном микробиоме. Например, бактерии *Mediterraneibacter faecis* и *Phocaeicola dorei* были ассоциированы со всеядной диетой, тогда как *Phocaeicola sartorii* и *Bacteroides acidifaciens* — с растительной.

Таким образом, в ходе работы был разработан алгоритм действий для изучения процессов утилизации SQ, уточнено таксономическое положение основных деструкторов SQ в кишечнике человека и выявлена связь между содержанием отдельных бактериальных видов с типом диеты. Полученные данные могут способствовать разработке новых подходов к диагностике и профилактике заболеваний кишечника, связанных с метаболизмом сульфохинозов.

1. Roy et al. (2003) Glycolytic Breakdown of Sulfoquinovose in Bacteria: a Missing Link in the Sulfur Cycle. *Appl Environ Microbiol* 69
2. E.D. Goddard-Borger, S.J. Williams (2017) Sulfoquinovose in the biosphere: occurrence, metabolism and functions. *Biochem J.*;474(5):827-849.
3. B.T. Hanson et al. (2021) Sulfoquinovose is a select nutrient of prominent bacteria and a source of hydrogen sulfide in the human gut. *ISME J.*;15(9):2779-2791.
4. A. Burrichter et al. (2018) Anaerobic Degradation of the Plant Sugar Sulfoquinovose Concomitant With H₂S Production: *Escherichia coli* K-12 and *Desulfovibrio* sp. Strain DF1 as Co-culture Model. *Front Microbiol.* Nov 27;9:2792.
5. A. Almeida et al. (2021) A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* 39, 105–114 (2021).
6. D. McDonald (2018) American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems*3:10.1128/msystems.00031-18.

Z-DNA and G-quadruplex structures in polycomb-repressive complex anchors in neurons

Bulat Rakhimov¹, Maria Poptsova¹, Vladimir Kurattsev^{1,2},
Ekaterina Khrameeva³

¹*International Laboratory of Bioinformatics, Faculty of Computer
Science, Higher School of Economics, Moscow, Russia
rah.bulat7@yandex.ru, maria.poptsova@gmail.com*

²*Lomonosov Moscow State University, Faculty of Physics, Department
of Biophysics, Moscow, Russia*

³*Skolkovo Institute of Science and Technology
ekhrameeva@gmail.com*

The spatial organization of the genome in human neurons shows significant differences compared to other cell types. One of the key structural features is the formation of long-range strong interactions, known as “neuronal dots”, which are specific to neurons and clearly visualized on Hi-C contact maps away from the main diagonal [1]. These neuronal-specific contacts overlap with peaks of the H3K27me3 ChIP-seq mark, associated with gene repression by Polycomb group proteins. This study focuses on the localization of secondary DNA structures in anchor regions of Polycomb Repressive Complexes (PRC2).

We investigated Polycomb loop anchors to assess enrichment or depletion of promoter regions (2000 bp upstream and 1000 bp downstream from the transcription start site (TSS)) with Z-DNA and G-quadruplex structures. For the overlapping regions, we analyzed separately the genes encoding transcription factors (TFs) and non-TF genes. A permutation test was used to statistically assess enrichment or depletion, allowing comparisons of enrichment in TF and non-TF genes at anchor regions against a background set including all genes. Control analysis was performed on regions that did not overlap anchor points. Gene Ontology (GO) analysis was conducted on the intersecting gene sets. The annotation of anchor regions was obtained from neurons isolated via FACS from the brains of healthy individuals. The data were kindly provided by the laboratory of E. E. Khrameeva (Skoltech).

Genome-wide maps of potential Z-DNA and G-quadruplexes were generated in the laboratory of M. S. Poptsova (HSE University) using deep learning methods [2].

The permutation test (1000 iterations) revealed statistically significant enrichment of both secondary DNA structures at PRC2 anchor regions in neurons, for both TF and non-TF genes. The distribution profiles of Z-DNA and G-quadruplexes were similar, indicating their co-localization in promoter regions of the analyzed genes. GO analysis revealed that TF genes are associated with developmental processes, morphogenesis, organ formation, CNS neuron differentiation, and regulation of gene expression via RNA Polymerase II. Associations with HMG-box containing proteins involved in chromatin remodeling were also identified. Non-TF genes were linked to RNA-induced silencing complexes, translational repression, intercellular communication protein complexes, and inhibitory neuronal signaling pathways. GO terms also included regulation of endothelial cell proliferation and activation of the VEGF signaling pathway during angiogenesis.

Genes at Polycomb-repressive complex anchors containing secondary DNA structures were analyzed for tissue specificity using the TissueEnrich tool, including data from GTEx and the Human Protein Atlas. The identified genes were expressed in neural tissue, adrenal gland, kidney, artery, and other tissues (Fig. 1). Nine tissue-specific genes expressed in the brain were identified (Fig. 2).

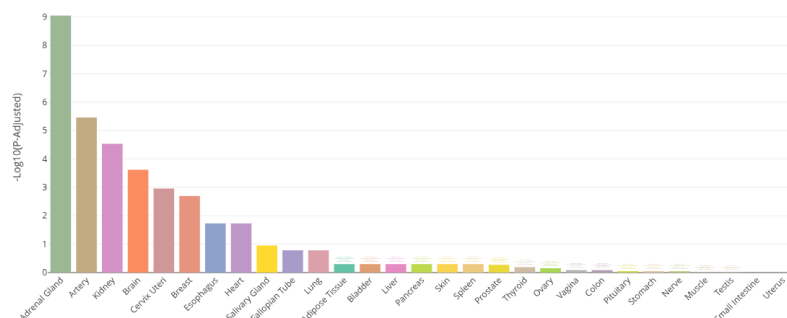


Fig. 1. Tissue enrichment analyses for genes enriched with Z-DNA structures in PRC II.

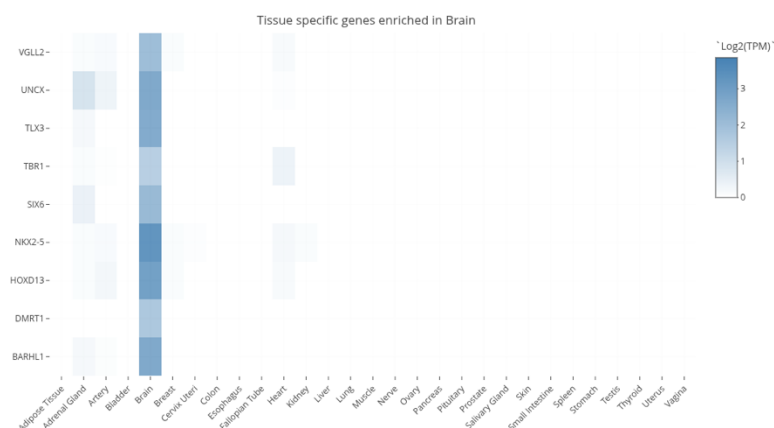


Fig. 2. Brain-specific genes enriched with Z-DNA structures.

We plan to use these data for chromatin structure element prediction via a classification task, where the positive class corresponds to PRC2 anchor regions. Input features will include DNA sequences

encoded using one-hot encoding. A second group of features represents omics data, such as histone marks, DNase I hypersensitive sites (or other chromatin accessibility assays), transcription factors, and RNA Polymerase binding sites. Currently, we are generating vector representations of DNA secondary structure regions based on omics features using UMAP clustering and identifying the most associated histone marks, transcription factors, and other omics signatures. These features are linked to biological processes including embryonic development, morphogenesis, RNA Polymerase II and III dependent transcription, chromatin-binding complexes, transcriptional repression, and chromatin remodeling. The most informative omics features will be used for training classification models.

1. I. Pletenev et al. (2024) Extensive long-range polycomb interactions and weak compartmentalization are hallmarks of human neuronal 3D genome, *Nucleic Acids Research*, 52:6234-6252.
2. N. Beknazarov et al. (2020) Deep learning approach for predicting functional Z-DNA regions using omics data, *Sci Rep* 10:19134.

Альтернативные функции компонентов комплекса дозовой компенсации у дрозофилы

А. Ж. Ревель-Муроз, Н. С. Клименко

*ИБГ РАН, 119334, г. Москва, Вавилова 34/5
revelanastasia@gmail.com*

Комплекс дозовой компенсации (КДК), также называемый MSL комплексом, играет ключевую роль в балансировке уровней продуктов генов между полами у плодовой мухи (*Drosophila melanogaster*). Это происходит за счет двукратного увеличения экспрессии генов на единственной X хромосоме самцов. Главные составляющие данного комплекса это белки: MSL1, формирующий каркас комплекса, MS2, отвечающий за специфичность связывания и стехиометрическую регуляцию, MSL3, считывающий метку активного хроматина H3K36me3, MOF, оставляющий метки активного хроматина H4K16ac и MLE, взаимодействующий с

длинными некодирующими РНК roX1 и roX2, позволяющими удерживать комплекс в рамках X хромосомы [1].

В последнее время представляется все больше свидетельств наличия альтернативных функций компонентов комплекса MSL вне X хромосомы. Так, было показано, что MSL2 у самцов способен связываться с промоторными областями на аутосомах, регулируя экспрессию генов морфогенеза, необходимых для правильного развития крыльев и глаз [2]. А MSL3 у самок плодовой мухи участвует в регуляции дифференциации половых стволовых клеток, способствуя экспрессии и трансляции ключевых факторов, необходимых для трансформации в ооциты [3]. В данной работе мы исследовали места связывания компонентов КДК у самцов и самок на аутосомах с целью определить их функциональные роли, как в составе комплекса, так и белков в отдельности.

В работе были использованы результаты ChIP-Seq экспериментов, проведенных в лаборатории “Молекулярной организации генома” ИБГ РАН, для белковых компонентов КДК, а также ряда кофакторов. Дополнительно были использованы результаты ChIP-seq и RNA-seq экспериментов, находящиеся в открытом доступе. Сырые данные обработаны с привлечением унифицированного программного конвейера.

В качестве основы для анализа был получен список промоторов аутосомных генов, связывающих один из MSL белков КДК: MSL1,2,3 у самцов и MSL1,3 у самок. Учитывая нормализованный сигнал перечисленных выше белков, данные регионы были кластеризованы. Для промоторов внутри каждого кластера были исследованы уровни нормализованных сигналов других компонентов комплекса MSL, прочих возможных кофакторов, а также гистоновых меток. Был проведен анализ обогащения групп генов (GO), принадлежащих исследуемым кластерам. Также сравнивались уровни экспрессии кластерных генов в разных тканях с использованием базы flyBase.

Во всех 702 исследованных регионах у самок наблюдался высокий сигнал MSL3. Из них 147 промоторов были обогащены сигналами MSL1, MSL2 и MSL3 у самцов, а ещё 105 имели высокий сигнал исключительно для белка MSL2 у самцов. Это разбиение соответствует полученной кластеризации пиков. Как и ожидалось, для всех исследуемых промоторов в телах соответствующих генов

наблюдался высокий уровень метки H3K36me3, распознаваемой белком MSL3. При этом уровни других меток активного хроматина (H3K4me3, H3K27ac, H4K16ac) были высокими для регионов, связывающих компоненты MSL у самцов. Помимо этого, в «мужских» аутосомных промоторах MSL были обогащены белки с ДНК-связывающим C2H2-доменом (CG1529, CG17829), транскрипционные факторы GAF и M1BP, а также компонент комплекса КДК MOF, обеспечивающий ацетилирование H4K16.

Что касается функций генов в кластерах, для промоторов, связывающих все три белка MSL у самцов, были обогащены термины, связанные с дифференциацией клеток, при этом для «женских» MSL-промоторных пиков был выявлен обогащённый путь дифференциации яичников. Анализ экспрессии генов в кластерах показал высокие уровни экспрессии в тканях зрительного диска и ЦНС на стадии личинки, в яичниках у самок и в головном отделе у самцов.

Полученные результаты указывают на возможность участия компонентов MSL-комплекса в морфогенезе самцов и самок за счёт регуляции аутосомных генов как в составе альтернативных комплексов, так и в составе субкомплекса MSL у самцов. При этом механизмы рекрутинга и взаимодействия с другими ко-факторами требуют дальнейшего изучения.

1. Shevelyov YY, Ulianov SV, Gelfand MS, Belyakin SN, Razin SV (2022) Dosage Compensation in : Its Canonical and Non-Canonical Mechanisms. *Int J Mol Sci.*, 23. doi:10.3390/ijms231810976
2. Valsecchi CIK, Basilicata MF, Semplicio G, Georgiev P, Gutierrez NM, Akhtar A. (2018) Facultative dosage compensation of developmental genes on autosomes in *Drosophila* and mouse embryonic stem cells. *Nature Communications*, 9: 1–14.
3. McCarthy A, Sarkar K, Martin ET, Upadhyay M, Jang S, Williams ND, et al. (2022) Msl3 promotes germline stem cell differentiation in female *Drosophila*. *Development*, 149. doi:10.1242/dev.199625

Transcriptomic predictors of therapeutic response to anti-TNF α therapy in inflammatory bowel diseases

G. Reshetnikova¹, E. Scheglova², A. Osipov³, A. Lee⁴, M. Saliutina⁵

¹*Bioinformatics Institute, Kantemirovskaya st. 2A, 197342, St. Petersburg, Russia,
galina.reshetnikova@helsinki.fi*

²*Moscow Institute of Physics and Technology, Institutsky lane 9, 141700, Dolgoprudny, Moscow region, Russia*

³*ITMO University, Kronverkskiy ave, 49A, 197101, St. Petersburg, Russia*

⁴*Koltzov Institute of Developmental Biology of the Russian Academy of Sciences, Vavilova st. 26/1, 119334, Moscow, Russia*

⁵*Charité - University Medicine Berlin, Hindenburgdamm 30, 12200, Berlin, Germany*

Inflammatory bowel diseases (IBD), including ulcerative colitis (UC) and Crohn's disease (CD), are chronic inflammatory disorders that significantly reduce the quality of life of millions of people worldwide. Despite sharing key inflammatory mechanisms, UC and CD differ significantly in their clinical and pathological features. During UC, only the mucosal layers of the colon and rectum are affected. In contrast, during CD, any part of the gastrointestinal tract with transmural inflammation can be affected, leading to complications such as fistulas, strictures, and abscesses.

One of the key immunotherapeutic strategies for IBD treatment is the use of tumor necrosis factor- α (TNF- α) inhibitors, such as infliximab, a chimeric monoclonal antibody that blocks the activity of TNF- α . Despite the overall efficacy of TNF- α blockers, many patients do not experience a clinical response to this therapy. This study aims to identify transcriptomic markers associated with treatment response by comparing gene expression profiles in responders and non-responders before anti-TNF- α therapy. These findings may support the development

of personalized treatment strategies and improve overall therapeutic outcomes.

Transcriptomic microarray datasets GSE12251 (UC; 23 patients: 12 non-responders, 11 responders) and GSE16879 (CD; 37 patients: 17 non-responders, 20 responders) from the Gene Expression Omnibus (GEO) database were used as training cohorts. Differentially expressed genes (DEGs) between non-responders and responders were identified using the limma package in R. Weighted co-expression network analysis (WGCNA) was used to cluster genes into modules based on similar expression patterns and to evaluate their correlation with clinical phenotypes. Gene modules showing a statistically positive correlation with the "non-responder" phenotype were selected for further analysis. From these modules, we selected genes that overlapped with previously identified DEGs and were involved in immune-related processes, based on functional enrichment analysis. STRING and CytoHubba platforms (based on node centrality) were used to construct protein-protein interaction networks and identify hub genes. Validation of the identified hub genes was performed using independent cohorts: GSE23597 microarray dataset (UC; 32 patients: 25 non-responders, 7 responders) and a single-cell RNA-seq dataset GSE282122 (CD; 16 patients: 10 non-responders, 6 responders). For the single-cell cohort, a pseudo-bulk approach analysis was used, aggregating transcriptomic data from 3000 cells per patient. ROC analysis and logistic regression were performed to evaluate the individual and collective prognostic potential of the selected hub genes.

In patients with UC, WGCNA identified gene modules significantly correlated with the absence of response to therapy ($r \geq 0.45$). From these modules, 10 hub genes (IL1B, IL6, CXCL8, TLR2, TLR4, TLR8, ITGAM, S100A12, FCGR3B, and PTGS2) were identified, five of which: IL1B, IL6, CXCL8, TLR2, and TLR4 were also shared with CD. ROC analysis in the UC validation cohort demonstrated high AUC values for IL1B and IL6 (0.771 and 0.817, respectively), suggesting their prognostic potential and ability to distinguish non-responders and responders. A logistic regression model based on genes common to UC and CD achieved an AUC of 0.829.

In patients with CD, WGCNA also identified a module enriched in immune-related genes. From these modules, 10 hub genes (IL1B, IL6, CXCL8, TLR2, TLR4, TLR8, MMP2, TGFB1, PECAM1, and TIMP1)

were identified. A logistic regression model built on these genes demonstrated an AUC of 0,82 in the pseudo-bulk single-cell dataset. Moreover, ROC analysis on the single-cell dataset revealed that IL1B showed an AUC of 0.68. Notably, there was a statistically significant decrease of IL1Blo monocytes among non-responders in comparison to responders. However, the expected increase of monocytes with high IL1B expression was not observed in non-responders.

Our results showed that IL1B, IL6, CXCL8, TLR2, and TLR4 can potentially act as predictors of response to infliximab therapy. IL6 is known to be involved in Th17 differentiation and maintenance of inflammation via STAT3 activation [2]. IL1B and CXCL8 are involved in activation of the innate immune response and neutrophil infiltration in IBD [3, 4]. Increased expression of TLR2 and TLR4 in UC may reflect pathological activation of innate immunity under impaired conditions [5].

UC-specific hub genes such as ITGAM and S100A12 are associated with leukocyte migration and inflammation of neutrophils, respectively. In contrast, CD-specific genes included TGFB1 and TIMP1, whose products are involved in tissue remodeling and fibrosis, as well as PECAM1 and MMP2, reflecting active leukocyte migration and destruction of the extracellular matrix.

Based on the obtained results, IL1B appears to be a shared marker of lack of response to anti-TNF- α therapy in both forms of IBD. These findings are consistent with existing ideas about the functional activity of this cytokine. It has also been shown that the level of IL1B in macrophages of patients with CD negatively correlates with the clinical response to adalimumab [6]. In our work, the abundance of ILBlo monocytes was typical for responders. However, the heterogeneous cell composition of monocytes with diverse patterns of IL1B expression among responders and non-responders requires further functional studies.

1. Rutgeerts, P., Van Assche, G., & Vermeire, S. (2006). Infliximab therapy for inflammatory bowel disease—seven years on, *Alimentary pharmacology & therapeutics*, 23(4): 451-463.
2. Shahini, A. and Shahini, A. (2023). Role of interleukin-6-mediated inflammation in the pathogenesis of inflammatory bowel disease: focus on the available therapeutic approaches and gut microbiome, *J. Cell*

Commun. Signal, 17: 55-74 173.

3. Chen, W., Xu, Z., Jiang, J., Chen, L., Chen, Y., Yu, T., ... & Shi, R. (2023). CXCL8 as a Potential Biomarker for Crohn's Disease is Associated with Immune Infiltration and RNA Methylation, *Biochemical Genetics*, 61(6): 2599-2617.

4. Ngoh, E. N., Weisser, S. B., Lo, Y., Kozicky, L. K., Jen, R., Brugger, H. K., ... & Sly, L. M. (2016). Activity of SHIP, which prevents expression of interleukin 1 β , is reduced in patients with Crohn's disease, *Gastroenterology*, 150(2): 465-476.

5. Tan, Y., Zou, K. F., Qian, W., Chen, S., & Hou, X. H. (2014). Expression and implication of toll-like receptors TLR2, TLR4, and TLR9 in colonic mucosa of patients with ulcerative colitis, *Journal of Huazhong University of Science and Technology [Medical Sciences]*, 34: 785-790.

6. Gole, B., Pernat, C., Jezernik, G., & Potočnik, U. (2023). The expression IL1B correlates negatively with the clinical response to adalimumab in Crohn's disease patients: An ex vivo approach using peripheral blood mononuclear cells, *Life Sciences*, 326: 121822.

Влияние мутации N294S в гене нейраминидазы вируса гриппа на взаимодействие с осельтамивиром

Е.А. Рухлова, Г.С. Онхонова, М.Н. Косенко, И.М. Суслопаров

Государственный научный центр вирусологии и биотехнологии «Вектор» Роспотребнадзора, Кольцово, Новосибирская область, 630559, Россия, e-mail: ruhlovalena@mail.ru

Нейраминидаза вируса гриппа — один из ключевых поверхностных гликопротеинов, ответственный за гидролиз α -2,3- и α -2,6-сиаловых кислот. Ферментативная активность нейраминидазы обеспечивает высвобождение новых вирионов из зараженной клетки, препятствует агрегации и удержанию вируса муцинами слизи верхних дыхательных путей, богатыми сиаловыми кислотами.

Ингибиторы нейраминидазы представляют собой класс противовирусных препаратов, которые, связываясь с активным

центром, блокируют действие этого фермента, тем самым предотвращая распространение вируса в организме. Особый интерес представляют штаммы вируса гриппа, резистентные к ингибиторам нейраминидазы, ввиду снижения эффективности лечения и потенциально повышенных эпидемиологических рисков.

Целью исследования является моделирование взаимодействия нейраминидазы вируса гриппа птиц A(H5N8) с осельтамивиром методами сравнительного моделирования и молекулярного докинга.

В качестве объектов исследования были использованы штаммы A/chicken/Tatarstan/88/2017 (дикий тип) и A/chicken/Tatarstan/112/2017 (мутация N294S), принадлежащие к кладе 2.3.4.4b.

Трехмерная структура нейраминидазы получена методом сравнительного моделирования с использованием MODELLER. В качестве шаблона выбрана структура из базы данных pdb (2HT7). Подготовка молекул белка и лиганда проводилась в AutoDock Tools. Структура комплекса нейраминидазы с занамивиром построена методом молекулярного докинга с использованием AutoDock Vina.

В процессе работы была отобрана модель, расположение лиганда в которой соответствует экспериментальным данным. Для комплекса с осельтамивиром ΔG полученных моделей составила -9,78 ккал/моль для дикого типа и -7,59 ккал/моль для мутантного варианта. Данные значения ΔG указывают на термодинамическую стабильность комплексов, причем меньшее значение для дикого типа предполагает более прочное связывание осельтамивира по сравнению с мутантным вариантом.

Таким образом, замена аспарагина на серин в позиции 294 нейраминидазы, представляет собой значимое изменение аминокислотной последовательности, которое влияет на кинетику связывания с осельтамивиром. Мутация N294S в гене нейраминидазы вируса гриппа может приводить к изменениям в активном центре фермента, снижающим аффинность комплекса с осельтамивиром, и оказывать влияние на эффективность противовирусной терапии.

Dirichlet flow matching for generating RNA UTRs with desired properties

Arseniy K. Rybakov¹, Elizaveta O. Aristova¹, Dmitry D. Penzar²,
Arsenii O. Zinkevich^{1,3}

¹*Faculty of bioengineering and bioinformatics, Lomonosov MSU,
Russia, Moscow, Leninskie gory 1, irisunikova@gmail.com*

²*Institute of protein research, Russian Academy of Sciences, Russia,
Puschino*

³*Vavilov Institute for general genetics, Russian Academy of Sciences,
Russia, Moscow*

Untranslated regions (UTRs), located at the 5' and 3' termini of messenger RNA (mRNA) molecules, are critical cis-regulatory elements that profoundly influence gene expression. While not directly encoding protein, UTRs harbor a wealth of information that dictates mRNA stability, localization, and translational efficiency. The 5' UTR regulates ribosome recruitment and translation initiation, often influenced by secondary structures. In contrast, the 3' UTR is a hub for interactions with RNA-binding proteins (RBPs) and microRNAs (miRNAs), which can either enhance or repress mRNA translation and mediate mRNA degradation. Therefore, constructing UTRs with specific properties is a promising approach to genetic engineering. For instance, this approach could be used to design RNAs with tissue-specific expression, which is an essential property for RNA-based vaccines.

Recent advances in generative modeling provide insight into using deep learning models to generate UTRs. Diffusion- and flow-matching-based models [1, 2] are well-known, state-of-the-art approaches in generative modeling. Recent work [3] shows that diffusion can be potentially applied to UTR generation.

The major disadvantage of applying diffusion to sequence generation, however, is that it doesn't account for the discrete nature of sequences. The recently proposed Dirichlet Flow Matching [4] approach provides a solution to this problem by interpreting the latent representations of the sequence as parameters for a Dirichlet distribution

over the probabilities of each nucleotide in the sequence. To obtain sequences with desired properties, however, it is necessary to employ some form of conditional generation, which can be done by injecting a property embedding directly into the input sequence. Recently, another approach [guidance] has been developed: at each step of a relatively long generation cycle, the object is evaluated with an external pre-trained classifier or regressor model. Then, the gradient of the loss for the given class or value can be subtracted from the object being generated to guide it toward a given property. By using the loss function introduced in Dirichlet flow matching, we trained an U-Net-like [5] architecture for generating RNA UTRs. We demonstrate that this can be done efficiently in both unconditional (i.e. no property constraints provided) and conditional (i.e. generation of sequences with a given expression rate) settings. Our aim is to validate and select the best from different model architectures, and adapt them to use guiding models.

1. J. Ho, A. Jain, P. Abbeel (2020) Denoising diffusion probabilistic models, *Advances in neural information processing systems*, 33: 6840-6851.
2. E. Bengio et al. (2021) Flow network based generative models for non-iterative diverse candidate generation, *Advances in Neural Information Processing Systems*, 34: 27381-27394.
3. D. Penzar et al. (2023) LegNet: a best-in-class deep learning model for short DNA regulatory regions, *Bioinformatics*, 39.8: btad457.
4. H. Stark et al. (2024) Dirichlet flow matching with applications to DNA sequence design, *International Conference on Learning Representations*, <https://openreview.net/forum?id=ehYe5bz8H3>.
5. O. Ronneberger, P. Fischer, T. Brox. (2015) U-net: Convolutional networks for biomedical image segmentation, *Medical image computing and computer-assisted intervention*, 8: 234-241.

Evolution and diversification of the sulfo-Embden–Meyerhof–Parnas pathway in bacteria

Anna Rybina, Mikhail Gelfand

*Skolkovo Institute of Science and Technology, Bolshoy Boulevard, 30,
bld. 1, Moscow 121205, Russia, rybinaann@gmail.com*

Sulfoquinovose (SQ), a sulfonated derivative of glucose synthesized by photosynthetic organisms, plays a key role in the global sulfur cycle, with an estimated annual production exceeding 30 billion kilograms [1,2]. Bacterial degradation of SQ proceeds via six distinct metabolic pathways: the glycolytic-like sulfo-Embden–Meyerhof–Parnas (sulfo-EMP) and sulfo-Entner–Doudoroff (sulfo-ED) pathways; the pentose phosphate-like sulfo-transaldolase (sulfo-TAL) and sulfo-transketolase (sulfo-TK) pathways; and the non-glycolytic sulfo-ASMO and sulfo-ASDO pathways, which involve monooxygenase or dioxygenase activity [3]. Despite their ecological relevance, the evolutionary history of these pathways remains poorly understood.

This study focuses on the phylogenetic origins of the sulfo-EMP pathway, which exists in two non-homologous forms: sulfo-EMP1, represented by the *yih* locus in *Escherichia coli* K-12, and sulfo-EMP2, found in *Alkalicoccus urumqiensis* BZ-SZ-XJ18.

Phylogenetic analysis suggests that sulfo-EMP2 enzymes—SqvB (mutarotase), SqiK (kinase), and SqiA (aldolase)—originated within Firmicutes. These genes appear to have evolved in genomic contexts associated with sulfo-TAL/TK-like pathways before diverging into dedicated sulfo-EMP2 clusters.

In contrast, the sulfo-EMP1 locus likely originated as a minimal *yih*TVW cluster (encoding aldolase, kinase, and a transcriptional regulator) and later expanded to include *yih*S (isomerase) and *yih*Q (SQase). Among Enterobacterales, this locus varies in size: a short version (e.g., *yih*TUVW in *E. coli* Nissle 1917) retains only the core set, while a long version (e.g., *ompL*–*yih*OPQRSTUVWXYZ in *E. coli* K-12) includes additional genes. Transcriptomic analysis revealed that only the long *yih* locus is induced in response to SQ, suggesting functional specialization.

In addition to functional divergence, structural modeling (including molecular docking and dynamics simulations) reveals variation in the CsqR (YihW) transcription factor, which exists in two distinct structural forms—a rare case of regulator diversity in bacterial systems.

Horizontal gene transfer has played a key role in shaping the sulfo-EMP pathway, with Actinomycetes acquiring genes from both Proteobacteria and Firmicutes. These findings shed light on the evolutionary dynamics of SQ catabolism and lay the groundwork for future research into the ecological and biotechnological roles of SQ-degrading bacteria.

Funding: RSF 24-14-00276

1. A.A. Benson, H. Daniel, R. Wiser (1959) A sulfolipid in plants, Proc. Natl. Acad. Sci. U. S. A., 45:1582–1587.
2. J.L. Harwood, R.G. Nicholls (1979) The plant sulpholipid—a major component of the sulphur cycle, Biochem. Soc. Trans., 7:440–447.
3. Y. Wei, Y. Tong, Y. Zhang (2022) New mechanisms for bacterial degradation of sulfoquinovose, Biosci. Rep., 42:BSR20220314.

Классификация эндонуклеаз рестрикции типа III по сходству последовательностей каталитического домена

И.С. Русинов¹, С.А. Спирин^{1,2,3}, А.С. Карягина^{1,4,5},
А.В. Алексеевский^{1,3}

¹НИИ ФХБ имени А.Н. Белозерского МГУ, Москва, Ленинские горы, дом 1, строение 40, isrusinov@gmail.com

²НИУ «Высшая школа экономики», Москва, Покровский бульвар, дом 11

³НИЦ «Курчатовский институт» — НИИСИ, Москва, Нахимовский проспект, дом 36, корпус 1

⁴НИЦ эпидемиологии и микробиологии имени Н.Ф. Гамалеи Минздрава России, Москва, улица Гамалеи, дом 18

⁵ВНИИ сельскохозяйственной биотехнологии, улица Тимирязевская, дом 42

Эндонуклеазы рестрикции (ЭР) – компоненты защитных систем рестрикции-модификации, гидролизующие чужеродную ДНК – широко распространены среди прокариот и хорошо известны благодаря применению в геномной инженерии. Традиционное разделение ЭР на четыре типа основано на составе комплексов и механизме функционирования соответствующих систем рестрикции-модификации, а не на родстве последовательностей белков [1]. Также для ЭР известны четыре неродственные укладки каталитических эндонуклеазных доменов (КЭД) [2,3], соответствие между типом ЭР и укладкой КЭД отсутствует. Целью данной работы является классификация ЭР по сходству последовательностей КЭД – единственного домена, обязательно присутствующего в любой ЭР. На текущем этапе рассматриваются только ЭР типа III, которые были выбраны для отработки методики из-за достаточной обособленности от ЭР других типов и низкого охвата находками профилей известных КЭД из базы Pfam.

Из базы данных REBASE версии 309 были загружены последовательности компонентов систем рестрикции-модификации. Уникальные последовательности были кластеризованы с порогом 50% идентичности. Кластеры, соответствующие ЭР типа III, были отобраны по аннотациям из REBASE, случаи противоречия аннотаций были проверены вручную. Из 684 полученных кластеров были выбраны 243 кластера, репрезентативный белок в которых не короче 500 а.о. и размер кластера не меньше 5. Для каждого кластера было получено множественное выравнивание представителей.

Для поиска КЭД в белках были использованы предсказанные структуры из AlphaFoldDB: ручная разметка КЭД в некоторых структурах была полуавтоматически перенесена на остальные с помощью пространственного совмещения (см. рис. 1). Участки, соответствующие КЭД, были вырезаны из каждого выравнивания и далее повторно кластеризованы и выравнены. С помощью итеративной процедуры поиска и объединения близких кластеров были получены итоговые семейства КЭД для ЭР типа III. По этим семействам были построены НММ-профили.

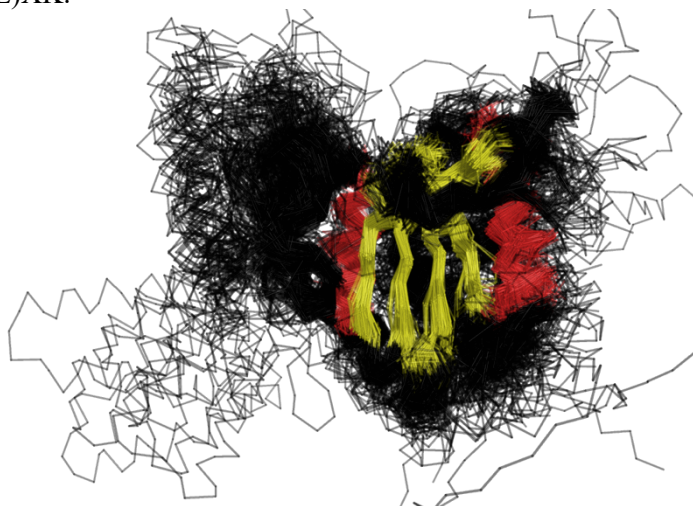
В результате получено семь семейств КЭД. Находки профилей, соответствующих семействам, содержатся в

подавляющем большинстве ЭР типа III из базы данных REBASE. В ЭР других типов обнаружено лишь несколько находок с низким весом. Планируется использовать аналогичную процедуру для классификации ЭР других типов.

Исследование выполнено в рамках государственного задания МГУ имени М.В. Ломоносова.

1. R.J. Roberts et al. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes, *Nucleic Acids Research*, 31(7):1805–1812.
2. J. Orłowski and J.M. Bujnicki (2008) Structural and evolutionary classification of Type II restriction enzymes based on theoretical and experimental analyses, *Nucleic Acids Research*, 36(11):3552–3569.
3. K. Miyazono et al. (2014) A sequence-specific DNA glycosylase mediates restriction-modification in *Pyrococcus abyssi*, *Nature communications*, 5:3178.

Рис. 1. Совмещение предсказанных структур участков КЭД из представителей 243 кластеров. Желтым и красным выделены консервативные β -тяжи и α -спирали, характерные для укладки PD-(D/E)ХК.



**NF-RNACHROM лучше и универсальнее аналогичных
пайплайнов при обработке данных РНК-хроматинового
интерактома животных, растений и прокариот**

Г.К. Рябых¹, И.С. Ильницкий^{1,2}, И.К. Марков², А.И. Сигорских²,
А.И. Никольская², А.А. Миронов^{1,2}

¹*Институт общей генетики имени Н.И. Вавилова, Российская
академия наук, Москва, Россия, ryabukhgrigory@gmail.com*

²*Факультет биоинженерии и биоинформатики, Московский
государственный университет имени М.В. Ломоносова, Ленинские
горы, Москва, Россия*

В настоящее время известно, что в геномах эукариот и прокариот транскрибируются помимо белок-кодирующих РНК множество некодирующих РНК (нкРНК), относящихся к различным биотипам [1, 2]. Для всех этих царств показано, что отдельные нкРНК играют важную роль в регуляции разнообразных биологических процессов, как на уровне отдельных клеток, например, транскрипция и сплайсинг, так и на уровне целых тканей, например, цветение и прорастание семян [3-4].

Сегодня известно уже более ста тысяч человеческих нкРНК, и это число увеличивается [5]. Некодирующие РНК ассоциированы с развитием человеческих патологий, таких как онкологические [6] и врожденные генетические заболевания [7]. В то же время было заключено, что для пролиферации человеческой клетки необходимы около 500 нкРНК [5]. В связи с этим их дальнейшее исследование представляет научный интерес.

В последнее десятилетие увеличивается количество методик поиска РНК-хроматиновых взаимодействий (RAP, CHART-seq, ChIRP-seq, ChOP-seq, CHIRT-seq, MARGI, GRID-seq, ChAR-seq, iMARGI, RADICL-seq, Red-C и ChRD-PET). Разнообразие подходов к дизайну эксперимента, структуре химерного прочтения и вычислительной обработке сырых данных приводит к сложностям в интерпретации и сравнении результатов. Для того чтобы ввести единый стандарт обработки была создана база данных RNA-Chrom [8], все записи которой проходят обработку универсальным

вычислительным протоколом NF-RNACHROM, разработанным нашей группой.

В настоящее время существует множество вычислительных протоколов обработки РНК-хроматиновых данных, например, iMARGI-docker, Flypipe, chartools, GridTools, RADICL-seq, RedClib, ENCODE, ChIRP-seq. Однако, они все ориентированы лишь на данные, полученные соответствующим экспериментальным протоколом, что приводит к проблеме несовместимости результатов.

Мы провели реструктуризацию и расширили протокол NF-RNACHROM, добавив новые модули, в том числе для статистической обработки получаемых РНК-ДНК контактов и использовали его для обработки новых публичных интерактомных данных для разнообразных видов эукариот и прокариот: человек (*Homo sapiens*), мышь (*Mus musculus*), плодовая мушка (*Drosophila melanogaster*), кишечная палочка (*Escherichia coli*), сенная палочка (*Bacillus subtilis*), прокариот *Thermofillum adornatum*, североамериканский красногорлый анолис (*Anolis carolinensis*), резуховидка Таля (*Arabidopsis thaliana*), рис посевной (*Oryza sativa*) и домашняя свинья (*Sus scrofa*). Мы показали, что NF-RNACHROM справился со всеми типами данных различных экспериментов лучше чем оригинальные пайплайны, а обработанные данные хорошо воспроизводят большинство наблюдений в статьях.

Финансирование: работа поддержана грантом РФФ №23-14-00136

1. S.Djebali et al. (2012) Landscape of transcription in human cells, *Nature*, 489:101–108.
2. R.Majumder et al. (2022) Prokaryotic ncRNAs: Master regulators of gene expression, *Current Research in Pharmacology and Drug Discover*, 100136:1–10.
3. G.K.Ryabykh et al. (2012) RNA-Chromatin Interactome: What? Where? When?, *Mol. Biol. (Mosk)*, 56:275-295.
4. P.Jampala et al. (2021) Functions of long non-coding RNA in *Arabidopsis thaliana*, *Plant Signal. Behav.*, 16:1–12.
5. J.S.Mattick et al. (2023) Long non-coding RNAs: definitions, functions, challenges and recommendations, *Nat. Rev. Mol. Cell Biol.*, 24:430–447.

6. M.Huarte (2015) The emerging role of lncRNAs in cancer, Nat. Med., 21:1253–1261.
7. L.Statello et al. (2021) Gene regulation by long non-coding RNAs and its biological functions, Nat. Rev. Mol. Cell Biol., 22:96–118.
8. G.K.Ryabykh et al. (2023) RNA-Chrom: a manually-curated analytical database of RNA-chromatin interactome, Database, 2023:1–10.

**Оценка содержания бактерий и грибов в образцах
песчаных и суглинистых почв после ветровалов в
широколиственном лесу**

Д.С. Савельев^{1,2,3}, М.Н. Тутукина^{2,4}, Т.А. Бессонова^{2,5},
П.В. Шелякин², М.В. Бобровский^{1,6}, Л.Г. Ханина¹

¹*Институт математических проблем биологии РАН –
филиал Института прикладной математики им. М.В. Келдыша
РАН*

²*Институт общей генетики им. Н.И. Вавилова РАН,*

³*Пироговский университет, dsavel27@yandex.ru*

⁴*Сколковский институт науки и технологий,*

⁵*Институт биофизики клетки РАН, обособленное
подразделение ФИЦ ПНЦБИ РАН*

⁶*Институт физико-химических и биологических проблем
почвоведения РАН, обособленное подразделение ФИЦ ПНЦБИ РАН*

Почвенные микроорганизмы участвуют в формировании реакции природных экосистем на потепление, увеличение содержания CO₂, пожары, засухи, вспышки численности насекомых и др. [1]. Важным элементом лесных экосистем является валеж, который необходим для сохранения биоразнообразия, выполнения лесами регулирующей и поддерживающей экосистемных функций. Связь между валежом и почвенными микроорганизмами практически не исследована, хотя валеж может определять пространственно-временную динамику почвенного микробиома, учет которой необходим для оценки долгосрочных последствий изменений климата и стратегий природопользования [2]. Задачей

настоящего исследования было сравнение содержания бактериальной и грибной ДНК в почвенных образцах, отобранных в различных биотопах, под и рядом с валежом деревьев различных видов, путем проведения количественной ПЦР.

Образцы отбирали летом 2024 г. в четырех биотопах, расположенных в заповеднике «Калужские засеки»: на двух участках массового ветровала 2006 года (на песках и на суглинках – дерново-подзолах и дерново-подзолистых почвах) и в двух типах почв (в дерново-подзолистых и темногумусовых) на участке старовозрастного леса на суглинках. Отбор почв проводили под и рядом с валежом из верхнего 5 см слоя горизонта А без подстилки не менее, чем в трехкратной повторности для биотопа и с учетом вида лежащего ствола (всего 63 образца). Для оценки бактериальной и грибной нагрузки из всех образцов была выделена ДНК для последующей количественной ПЦР с праймерами на переменный участок V4 рибосомной рРНК и ITS1-3, соответственно.

Образцы были сгруппированы по четырем биотопам и по месту взятия образца - под и рядом с валежом. Для каждого образца были сделаны две технические реплики. Статистический анализ проводили в среде программирования R. Для оценки различий содержания ДНК между группами образцов проводили серию однофакторных дисперсионных анализов (ANOVA) и рассчитывали χ^2 -критерий [3] с привлечением функции catdes FactoMineR [4].

Содержание ДНК бактерий было значительно ниже в образцах, взятых под стволами, чем в образцах, взятых на участках, свободных от лежащих стволов. Максимальное содержание ДНК бактерий наблюдалось в темногумусовых почвах старовозрастного леса на участках без валежа: оно было незначительным, но значительно выше содержания, наблюдаемого на подзолах под стволами через 18 лет после массового ветровала ($2^{-\Delta Ct} = 0.5$). Эти результаты соответствовали микробной активности суглинистых почв, оцененной ранее методом субстрат- индуцированного дыхания [5, 6]: содержание углерода микробной биомассы и скорость образования CO₂ существенно уменьшались под стволами и на ветровале, и в старовозрастном лесу. Напротив, содержание ДНК грибов было максимальным под валежом (особенно под стволами ели на песках $Ct\_ITS \text{ qPCR} = 19.2 \pm 0.57$) и

минимальным на участке без стволов на темногомусовых почвах (22.1 ± 0.21). Относительно большее содержание микробной ДНК в богатых по С, N и другим элементам темногомусовых почвах и относительно меньшее в них содержание ДНК почвенных грибов при противоположных результатах в более бедных элементами песчаных почвах свидетельствует, вероятно, о разном характере изменений функциональной активности микробиома в почвах разных типов. Учет валежа, лежащего на почве, позволяет более корректно оценивать пространственно-временную динамику почвенных микроорганизмов и точнее прогнозировать отклик лесных сообществ на внешние и внутренние экосистемные воздействия.

Работа поддержана грантом РНФ 24-24-00340.

1. Baldrian, P., López-Mondéjar, R., Kohout, P. (2003) Forest microbiome and global change, *Nature reviews microbiology*. 21: 487–501.
2. Patoine G. et al. (2022) Drivers and trends of global soil microbial carbon over two decades, *Nature Communications*, 13: 4195
3. Husson F., Le S., Pagès J. (2017) *Exploratory multivariate analysis by example using R*. 2nd edition. Chapman & Hall/CRC.
4. Le S., Josse J., Husson F. (2008) FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25 (1): 1–18.
5. Khanina L.G., Ivashchenko K.V., Smirnov V.E., Bobrovskii M.V. (2024) Carbon and nitrogen stocks and microbial activity of the humus horizon of loamy soils after mass windthrow in the broadleaved forest of the Kaluzhskie Zaseki Nature Reserve, *Eurasian Soil Science*, 57(11): 1806–1818.
6. Смирнов В.Э., Ханина Л.Г., Иващенко К.В., Бобровский М.В. (2024) Оценка влияния крупных лежащих стволов дуба на свойства почв на основе статистических моделей со смешанными эффектами, *Математическая биология и биоинформатика, Доклады Международной конференции*. 10: e52. doi: 10.17537/icmbb24.54.

Импутация single-cell ATAC-seq данных с использованием тканеспецифичных геномных аннотаций

Д.Р. Салимов¹, С.А. Панишева², Т.С. Соколова¹, М.А. Куляшов¹,
С.К. Колмыков¹

¹*Научно-технологический университет «Сириус», 354340,
Российская Федерация, Краснодарский край, федеральная
территория «Сириус», Олимпийский пр., д. 1,
salimovdan1988@gmail.com*

²*Новосибирский государственный университет, 630090,
Российская Федерация, Новосибирская область, г. Новосибирск,
ул. Пирогова, д. 1*

Одним из ключевых методов для изучения регуляторной архитектуры хроматина является ATAC-seq, позволяющий определять открытые участки хроматина. В bulk ATAC-seq экспериментах доступность определяется усреднённо по всему образцу/ткани. Single-cell ATAC-seq (scATAC-seq) позволяет изучать ландшафт доступности хроматина на уровне отдельных клеток, раскрывая клеточную гетерогенность и сложную структуру регуляторных программ.

Основной формой представления результатов scATAC-seq экспериментов является матрица пиков: строки соответствуют клеткам, столбцы — участкам открытого хроматина. Однако из-за малого числа фрагментов (и, соответственно, прочтений) на клетку, такие матрицы крайне разрежены — лишь 3–7 % ненулевых значений. Более того, на уровне одной клетки детектируется всего от 1 до 10 % тех пиков, которые обычно наблюдаются в bulk ATAC-seq данных (Chen et al., 2019). Это затрудняет кластеризацию, визуализацию, аннотацию, построение профилей покрытия, а также поиск клеточноспецифичных регуляторных элементов.

В данной работе рассматривается новая методика импутации scATAC-seq данных на основе современных архитектур глубокого обучения. В отличие от большинства существующих подходов, метод позволяет учитывать тканеспецифичную информацию и потенциальные эффекты однонуклеотидных

геномных вариаций (SNP). Мы адаптируем энкодерную часть архитектуры для построения контекстно-зависимых эмбедингов пиков — таким образом учитываются связи между регуляторными элементами, удалёнными друг от друга по геному. Модель обучается на реконструкции исходных сигналов с использованием модифицированной функции потерь. Дополнительно интегрируются внешние геномные аннотации: тканеспецифичные профили открытого хроматина из базы данных GTRD (Kolmykov et al., 2021) и данные о SNP, ассоциированных с изменением доступности хроматина, из базы данных UDACHA (Buyan et al., 2025).

В качестве базового решения был выбран метод SCALE (Single-Cell ATAC-seq analysis via Latent feature Extraction), продемонстрировавший в рамках benchmarking-системы SAPIEnS наилучший рост метрик кластеризации (ARI, AMI) при предварительной фильтрации методом Boruta для наборов пиков до 50 000 (Xiong et al., 2019).

Сравнение разработанного подхода с другими методами импутации scATAC-seq данных scOpen и cisTopic (Li et al., 2019; Bravo González-Blas et al., 2019) выполнялась по протоколу SAPIEnS (Akhtyamov et al., 2024) и включала три этапа:

- Кластеризация: ARI, AMI;
- Визуализация: Silhouette-score для PCA, UMAP и tSNE;
- Цифровой footprinting: MRIS/TFRIS для ключевых факторов транскрипции (GATA1, RUNX1, SPI1).

Первые результаты показывают, что интеграция тканеспецифичных аннотаций и контекстуальных эмбедингов позволяет добиться прироста ARI/AMI по сравнению с существующими подходами, улучшить визуализируемую разделимость кластеров и повысить чувствительность footprinting-анализа в ключевых регуляторных регионах.

В дальнейших планах — автоматизация отбора гиперпараметров модели и расширение метода на более крупные наборы данных (> 50 000 пиков). Предлагаемая архитектура закладывает основу для более точного восстановления структуры открытого хроматина и биологической интерпретации scATAC-seq данных.

Исследование выполнено за счет гранта Российского научного

фонда № 24-24-20108, <https://rscf.ru/project/24-24-20108/>.

Список источников.

1. H.Chen et al. (2019) Assessment of computational methods for the analysis of single-cell ATAC-seq data, *Genome Biology*, 20:241.
2. S.Kolmykov et al. (2021) GTRD: an integrated view of transcription regulation, *Nucleic Acids Research*, 49:104–111.
3. A.Buyan et al. (2025) Statistical framework for calling allelic imbalance in high-throughput sequencing data, *Nature Communication*, 16.
4. L.Xiong et al. (2019) SCALE method for single-cell ATAC-seq analysis via latent feature extraction, *Nature Communication*, 10.
5. Z.Li et al. (2021) Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen, *Nature Communication*, 12.
6. C.Bravo González-Blas et al. (2019) cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data, *Nature Methods*, 16:397–400.
7. P.Akhtyamov et al. (2024) scATAC-seq preprocessing and imputation evaluation system for visualization, clustering and digital footprinting, *Briefings in Bioinformatics*, 25:1–9.

Сравнение методов отбора признаков в геномных данных для обучения модели классификации гетерозисных групп кукурузы

Н.А. Сапожников, Е.И. Григорьева, Д.О. Медведев

*ООО «Пластилин», Россия, 121205, г. Москва, Большой бульвар, д. 42, стр. 1
sapozhnikov@plastilin.team*

Современная селекция кукурузы неразрывно связана с эффективным использованием эффекта гетерозиса. Появление модели для геномной селекции, которая может быстро и достаточно точно классифицировать гетерозисную группу растительной линии облегчает селекционеру задачу определения оптимальной стратегии скрещиваний и потенциал будущих гибридов. Эффективное использование такой технологии остается вызовом, в

связи с высокой размерностью данных, где количество признаков (SNP-маркеров, исчисляемых тысячами или миллионами) значительно превосходит количество образцов. Эта особенность геномных данных приводит к "проклятию размерности", увеличивая риск переобучения моделей, повышая вычислительные затраты и затрудняя интерпретацию результатов. Отбор наиболее информативных признаков становится критически важным этапом не только для повышения точности и устойчивости моделей классификации, но и для выявления биологически значимых маркеров, связанных с гетерозисными группами.

Работа проводилась с использованием публично доступных данных [1]. Полученный VCF файл содержал в себе 604 образца и 186200 SNPs после фильтрации по MAF и импутации. Численность образцов в семи гетерозисных группах имели примерно одинаковую представленность. Метрика f1-score базовой классификации, усредненной по кросс-валидации была равна 0,84. Однако этой модели были выявлены признаки переобучения и запоминания данных.

Для отбора признаков сравнивались подходы разных типов: фильтрационные методы, оценивающие признаки независимо от конечного классификатора; классический GWAS с использованием линейных моделей для поиска ассоциаций; оберточный метод, интегрированный со случайным лесом; а также метод на основе глубокого обучения (автоэнкодер), формирующий новое признаковое пространство с учетом информации о классах. Результаты показали, что классификация набора данных после отбора признаков дает более качественные и стабильные предсказания, чем классификация исходного набора данных "как есть". Предварительные данные показали улучшение метрик классификации после отбора признаков, а также более выраженное формирование кластеров, соответствующих классам, в двумерных представлениях признакового пространства.

Список литературы

1. Li, C., et al. Genomic insights into historical improvement of heterotic groups during modern hybrid maize breeding. *Nat. Plants* 8, 750–763 (2022). <https://doi.org/10.1038/s41477-022-01190-2>

Геномика и филогеномика группы *Bacillus cereus*

С.К. Селяев

*Московский физико-технический институт, Долгопрудный,
Россия,
seliaevsava@gmail.com*

Введение. Группа *Bacillus cereus* представляет собой комплекс бактериальных видов, имеющих важное значение в медицине, сельском хозяйстве и биотехнологии. Эти микроорганизмы демонстрируют удивительное разнообразие - от опасных патогенов, вызывающих пищевые отравления и системные инфекции, до штаммов, используемых в качестве биопестицидов и пробиотиков. Однако точная классификация представителей этой группы долгое время оставалась сложной задачей из-за их высокой генетической схожести.

Основная проблема заключается в том, что традиционные методы идентификации, такие как анализ гена 16S рРНК, часто не позволяют надежно различать виды внутри группы. Это создает серьезные трудности в клинической диагностике, контроле патогенных штаммов и регулировании использования полезных микроорганизмов. Кроме того, существующая таксономическая классификация не всегда отражает реальные эволюционные взаимоотношения между различными представителями группы.

Целью нашего исследования было комплексное изучение группы *B. cereus* с использованием современных методов геномики и филогеномики. В частности, мы ставили перед собой три основные задачи: (1) уточнение таксономической структуры группы на основе полногеномного анализа, (2) реконструкция эволюционных взаимоотношений между различными представителями, и (3) оценка их патогенного и пробиотического потенциала с помощью биоинформатических подходов.

Методы. В нашем исследовании мы применили комплекс современных биоинформатических методов. Для филогенетического анализа использовались два подхода: UBCG2 (Universal Bacterial Core Genes), который анализирует 92

консервативных бактериальных гена, и PanACoTA для изучения динамики пангенома. Эти методы позволяют строить более надежные филогенетические деревья по сравнению с традиционным анализом отдельных генов.

Сравнительно-геномный анализ включал несколько подходов: расчет показателей геномного сходства с помощью OrthoANI (с порогом видового различия 95-96%), TETRA-анализ нуклеотидного состава, и определение цифрового уровня ДНК-ДНК гибридизации (dDDH). Для изучения крупных геномных перестроек и синтении мы использовали три специализированных инструмента: SibeliaZ для выявления гомологичных блоков, PaReBrick для анализа реарранжировок, и BADLON для обнаружения горизонтального переноса генов.

Особое внимание было уделено анализу патогенного потенциала. Мы использовали комбинацию инструментов: PathogenFinder для предсказания общей патогенности, и базу данных VFDB для идентификации конкретных факторов вирулентности. Пробиотический потенциал оценивался с помощью metaProbiotics и ProBioMinServer, которые анализируют наличие генов, ассоциированных с полезными свойствами.

Дополнительно мы применили методы машинного обучения, в частности алгоритм kmer2vec, который позволяет анализировать K-мерные профили целых геномов и выявлять тонкие различия между штаммами.

Результаты. Наше исследование выявило несколько важных закономерностей. Во-первых, мы обнаружили, что традиционный маркер 16S рРНК действительно недостаточен для различения видов в группе *B. cereus*. Причиной этого, по нашим данным, может быть не только высокая консервативность этого гена, но и его многокопийность у некоторых представителей группы (*B. cereus* - 14 копий).

В качестве альтернативы мы предлагаем использовать комбинацию двух генов: ДНК-зависимой РНК-полимеразы (*groB*) и фактора терминации транскрипции (*rho*). Эти гены показали значительно более высокое разрешение при дифференциации видов, что было подтверждено как филогенетическим анализом (значения бутстрэп-поддержки >90%), так и расчетами геномного сходства (OrthoANI).

Особый интерес представляют результаты таксономического анализа. Наши данные убедительно демонстрируют, что *B. cereus* и *B. thuringiensis* следует рассматривать как единый вид. Геномное сходство между ними составляет 98.35% по OrthoANI, а различия в патогенных свойствах объясняются наличием у *B. thuringiensis* специфических плазмид, несущих гены кристаллических токсинов. Аналогично, мы показали, что *B. pacificus* и *B. paranthracis* (96.21% по OrthoANI), а также *B. mycoides* и *B. weihenstephanensis* (97.76% по OrthoANI) представляют собой единые таксоны.

Заключение. Проведенное исследование позволило существенно уточнить таксономическую структуру группы *B. cereus* и выявить новые закономерности в ее эволюции. Полученные результаты имеют важное практическое значение для микробиологической диагностики, эпидемиологического надзора и регулирования использования бактериальных штаммов в биотехнологии.

Integrative transcriptomic and proteomic analysis reveals tolerance mechanisms against primary anthracnose infection in white lupin

Leonid Sidorov^{1†}, Stepan Boldyrev¹, Alina Baik¹, Ekaterina Bondarenko², Polina Volkova³, Laurent Gentzbittel^{4*}, Cécile Ben^{4*}

¹*Skolkovo Institute of Science and Technology, Project Center for Agro Technologies, Moscow, Russia; Department of Molecular and Cellular biology,* ²*Russian Institute of Radiology and Agroecology, 249032 Obninsk, Russia;* ³*Independent Researcher, Geel, Belgium;* ⁴*Université de Toulouse, Toulouse INP, CNRS, IRD, CRBE, Toulouse, France*

[†]*leonid.sidorov@skoltech.ru*

^{*}*previous affiliation: 1*

White lupin (*Lupinus albus* L.) is a promising source of protein, but its production is threatened by anthracnose, caused by *Colletotrichum lupini*. Using transcriptomics and proteomics, the

molecular responses of three cultivars – susceptible (Andromeda), intermediate (SN), and tolerant (Michruinskij) (1) – during primary infection by two aggressive *C. lupini* strains (3 and 139) from a single Russian field were studied. Completely Randomized Design (CRD) was used with three biological replicates, yielding 27 samples (including mock controls). Proteins and total mRNA were extracted from seedling radicles seven days post-inoculation of seeds. Proteomes were analyzed via LC-MS/MS (Label Free Quantification, LFQ), and transcriptomes were sequenced on Illumina (read count analysis).

The biocomputational analysis included the identification of differentially expressed genes (DEGs) using DESeq2 and EdgeR, and differentially abundant proteins (DAPs) with the prolfqua R package (after imputing missing values). A Weighted Gene Co-expression Network (WGCNA) and Gene Regulatory Network (GRN) for a significant module were built (BioNERO R package) based on the transcriptomic data. Protein-protein interaction (PPI) networks for DAPs were generated using STRING database. Finally, the multi-omics analysis was performed using two approaches: (a) nine-quadrant analysis, where the mRNA and protein log₂ fold changes were compared; (b) sparse partial least square modeling proposed by mixOmics R package.

Key pathways and genes associated with tolerance/susceptibility were identified. Transcriptome-proteome integration revealed strain- and cultivar-specific responses: an HSP20-like chaperone protein, associated with plant immunity (2), showed high protein abundance following inoculation with strain 139, despite being transcriptionally downregulated; in contrast, a peroxiredoxin involved in stress-induced germination suppression (3) exhibited elevated mRNA levels but low protein abundance exclusively in the tolerant Michurinskij cultivar. PPI networks implicated glutamate/proline metabolism in tolerance, potentially restricting pathogen nitrogen access. GRN analysis highlighted bHLH and C2C2-CO-like transcription factors as regulators of Michruinskij's tolerance, suggesting candidate targets for future knockout-based functional validation. These results reveal key resistance mechanisms against anthracnose in white lupin and molecular targets for breeding resistant varieties.

1. Samad S. et al. (2023) CHARACTERIZATION OF WHITE LUPIN (*LUPINUS ALBUS*) CULTIVARS' RESPONSE TO PRIMARY

INFECTION BY DIVERSE COLLETOTRICHUM LUPINI STRAINS FROM A SINGLE RUSSIAN FIELD, International Journal of Agriculture and Environmental Research, 9.6: 1057-1079

2. Park C. J., Seo Y. S. (2015) Heat shock proteins: a review of the molecular chaperones for plant immunity, The plant pathology journal, 31.4: 323.

3. Haslekas C. et al. (2003) Seed 1-Cysteine Peroxiredoxin Antioxidants Are Not Involved in Dormancy, But Contribute to Inhibition of Germination during Stress, Plant physiology, 133(3): 1148-1157.

We would like to thank our colleagues from Skoltech Agro, especially Ivan Koteletsev for the creation of the collection of Colletotrichum lupini strains, Samia Samad for the characterization of their aggressiveness on white lupin and Elena Martynova for her help in organizing the proteomic analysis done at Skoltech's Advanced Mass Spectrometry Core Facility. We also thank our collaborators from the Russian Institute of Radiology and Agroecology of National Research Centre «Kurchatov Institute» for the transcriptomic data and All-Russian Scientific Research Institute of Lupin who kindly provided seeds of the white lupin cultivars.

This study was funded by Grant from the Ministry of Science and Higher Education of the Russian Federation No. 075-15-2021-1068.

Распределение мишеней связывания CTCF в гене galnt17 у домашней собаки

О.И. Скобель, А.В. Леонов, Т.Т. Глазко, Г.Ю. Косовский

*Научно-исследовательский институт пушного
звероводства и кролиководства имени В.А. Афанасьева,
140143, Московская обл., Раменский р-он, пгт. Родники,
skobelolga@gmail.com*

Принято считать, что самый древний одомашнированный человеком вид млекопитающих – собака. Накоплены данные, свидетельствующие о том, что домашняя собака отличается от близкородственных диких видов повышенными характеристиками

социальной активности и способности к обучению, что особенно важно при подборе собак-помощников. Обнаружено, что мутации в гене *galnt17* (в частности, вставки транспозонов, хромосома 6, (Cfa6.6 и Cfa6.7)) могут способствовать увеличению эффективности отбора и подбора таких помощников [1].

Galnt17 - N-ацетилгалактозаминилтрансфераза, высоко экспрессируется в коре головного мозга, участвует в функции лизосом, клеточной адгезии и формировании внеклеточного матрикса [2]. У мышей нарушение регуляции *Galnt17* связано с фенотипом, типичным для предрасположенности к аутизму у человека (AUTS2) [3]. У ряда видов млекопитающих, в том числе и собаки, ген *galnt17* входит в эволюционно консервативную петлю хроматина вместе с генами *auts2* и *caln1* (белок, обладающий высоким сходством с кальций связывающими белками семейства кальмодулинов, участвует в физиологии нейронов и потенциально важен для запоминания и обучения) [4]. У млекопитающих в формировании хроматиновых петель принимают участие когезин и фактор регуляции транскрипции (архитектурный белок) CTCF [5].

В геномах собак разных пород ROS_Cfam_1.0 (лабрадор-ретривер), Dog10K_Boxer_Tasha (боксер), UU_Cfam_GSD_1.0 (немецкая овчарка), UNSW_CanFamBas_1.2 (басенджи) и геноме собаки Динго ASM325472v2, представленных в GenBank, с целью оценки присутствия мотивов связывания CTCF в гене *galnt17*, входящего в эволюционно консервативную хроматиновую петлю вместе с генами *auts2* и *caln1*, полиморфизм которого связан с социальной активностью у собак, нами выполнен анализ нуклеотидных последовательностей фрагментов Cfa6.6 и Cfa6.7 с помощью программы PWMScan [6], а также присутствия транспозонов с использованием программы Gini Repbase [7]. Поиск мотива CTCF MA0139.1 (база данных профилей связывания факторов транскрипции JASPAR CORE 2020 [8]) выполнен в обеих цепях исследуемых нуклеотидных последовательностей с допустимым пресечением результатов поиска при $p\text{-value} \leq 0,05$.

Проведенный анализ показал присутствие 53 мотивов связывания CTCF в нуклеотидной последовательности фрагмента Cfa6.6 у всех исследованных геномов, 21 из которых локализованы внутри участка гомологии к неавтономному ретротранспозону VulLag-1.8 из семейства SINE. Кроме того, 4 мотива связывания

CTCF частично перекрывают указанный участок гомологии. Стоит отметить, что последовательность одного мотива, обнаруженного у собаки Динго с $P\text{-value}=2.75\text{e-}02$ на участке 308-327 п.н., отличалась по предпоследнему нуклеотиду от мотива, обнаруженного у собак исследуемых пород с $P\text{-value}=4.29\text{e-}02$ на аналогичном участке (рис).

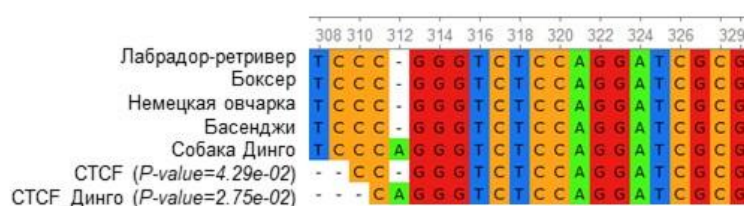


Рис. Участки генома, на котором локализованы сайты посадки, отличающиеся у собаки Динго и исследуемых пород собак

Выявлено 77 мотивов связывания CTCF в нуклеотидной последовательности фрагмента Cfa6.7 у лабрадор-ретривера и басенджи (охотничья подружейная собака и охотничья древняя африканская нелающая собака, соответственно), 20 из которых локализованы на участке гомологии к неавтономному ретротранспозону SINEC2A1_CF, а 34 и 5 - на соответствующих участках гомологии к эндогенному ретровирусу ERV3. При этом у боксера, немецкой овчарки и собаки Динго (спортивная бойцовская порода, исходно пастушья собака и предположительно одичавшая более 5 тыс. лет назад собака Динго, соответственно) выявлено 72 мотива связывания CTCF в нуклеотидной последовательности фрагмента Cfa6.7, 20 из которых также локализованы на участке гомологии к неавтономному ретротранспозону SINEC2A1_CF, а 32 и 5 - на соответствующих участках гомологии к эндогенному ретровирусу ERV3. Остальные мотивы связывания CTCF частично перекрывает указанный участок гомологии.

Таким образом, обнаружено, что в интронах гена galnt17 присутствует более 120 мотивов связывания с архитектурным белком CTCF, часть из которых перекрывается с указанными выше транспозонами, что потенциально может влиять на экспрессию генов как galnt17, так и auts2 и caln1, функционально тесно связанных с высшей нервной деятельностью. Выявлены отличия в

распределении мотивов связывания с CTCF у рассмотренных охотничьих (лабрадор-ретривер, басенджи) собак, пород с охранной функцией (бульдоги, немецкая овчарка) и диких собак Динго, отличающихся по характеристикам поведения.

1. D. Tandon et al. (2019) Homozygosity for Mobile Element Insertions Associated with WBSCR17 Could Predict Success in Assistance Dog Training Programs. *Genes (Basel)*, 10: 439.
2. G.Merla, C.Ucla, M.Guipponi, A.Reymond (2002) Identification of additional transcripts in the Williams-Beuren syndrome critical region. *Hum Genet*, 110: 429–438.
3. P.A.Weisner et al. (2019) A Mouse Mutation That Dysregulates Neighboring Galnt17 and Aut2 Genes Is Associated with Phenotypes Related to the Human AUTS2 Syndrome. *G3 Genes|Genomes|Genetics*, 9: 3891–3906.
4. O.Engmann et al. (2017) Cocaine-Induced Chromatin Modifications Associate With Increased Expression and Three-Dimensional Looping of Aut2. *Biological Psychiatry*, 82: 794–805.
5. J.Glaser, S.Mundlos (2022) 3D or Not 3D: Shaping the Genome during Development. *Cold Spring Harb Perspect Biol*, 14: a040188.
6. G.Ambrosini, R.Groux, P.Bucher (2018) PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. *Bioinformatics*, 34: 2483–2484.
7. O.Kohany, A.J.Gentles, L.Hankus, J.Jurka (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 7: 474.
8. O.Fornes et al. (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, 48: D87–D92.

Использование персонализированных геномов для дообучения LegNet

Д.А. Смирнова¹, Д.Д. Пензар²

¹*Факультет биоинженерии и биоинформатики Московского государственного университета имени М. В. Ломоносова.,
didi2003n@gmail.com*

²*Институт общей генетики им. Н.И. Вавилова Российской академии наук, Москва, Россия*

Задачей работы было использовать персонализированные данные для дообучения моделей по прогнозированию экспрессии генов и затем применить их в фреймворке для другого проекта по vcf аннотатору. Это поможет улучшить качество предсказаний и позволит учитывать особенности типа клеток. Сейчас модели-лидеры демонстрируют нулевую корреляцию между индивидуумами для каждого гена и часто предсказывают неправильные направления влияния даже для генов с сильным генетическим влиянием на экспрессию. Из успешных моделей в этой области Enformer демонстрируют высокую точность, он обладает архитектурой трансформера [2]. Однако такая модель долго учится и делает предсказания, для пользовательского интерфейса она неприменима.

Для дообучения была выбрана модель LegNet, имеющая EfficientNet-подобную архитектуру и победившая в конкурсе DREAM 2022 по прогнозированию экспрессии генов, она была обучена на данных GPRA по дрожжам [1]. Для задачи мы модифицировали архитектуру LegNet, добавив в архитектуру модели дополнительные 2 слоя pooling и 2 блока effective receptive field - это позволило нам расширить рецептивное поле модели с 230 до 1152 нуклеотида и работать с более длинной последовательностью. Мы использовали разные клеточные линии и типы экспериментов для обучения наших моделей, для каждой задачи обучали ансамбль из 10 моделей.

Мы работали с двумя источниками данных - с PRO-CAP от 58 линий LCL для сравнения с моделью Clipnet, также обученной

на этих датасетах, и с данными таргетов модели Enformer [2]. Данные PRO-CAP от 58 линий LCL были обработаны от формата bedgraph до датасета с последовательностями генов и средними значениями экспрессии, вся выборка разделена по хромосомам на тестовую и обучающую. Модель LegNet была дообучена на этих данных, затем сравнение моделей проводилось на MPRA для того, чтобы понять насколько предсказания модели переносятся на другой тип эксперимента. С задачей предсказания по MPRA Clipnet справилась лучше, плохое качество нашей модели может быть объяснимо небольшим размером датасета [3]. На больших обучающих данных LegNet справляется успешнее. Несмотря на это на тестовой выборке данных PRO-CAP, сделанных бутстрепом, корреляция Пирсона была больше - 0.563.

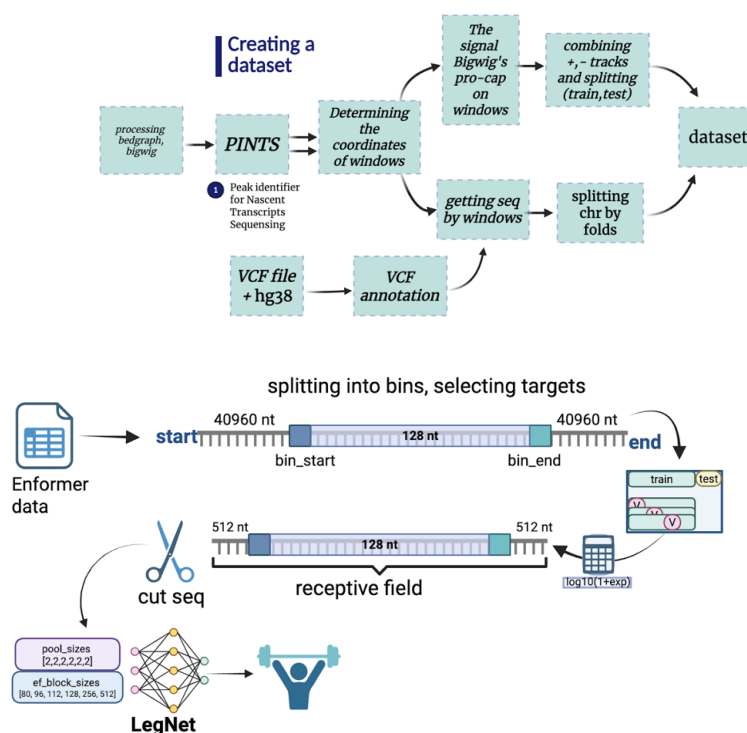


Рис.1 Схемы обработки данных

Также были обучены модели по данным CAGE на клеточных линиях K562, Hek, HepG2. LegNet показала хорошее качество - от 0.624 до 0.755 значения корреляции Пирсона по бутстрефу тестовых данных.

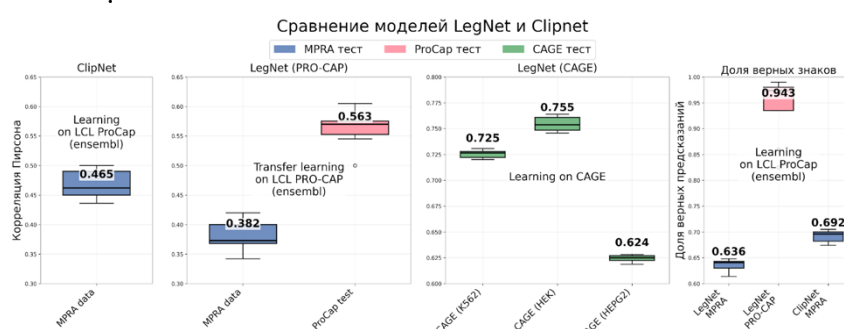


Рис.2 Результат сравнения моделей LegNet и Clipnet

Дальше мы обучали ансамбль моделей на данных Enformer, предварительно подготовив датасеты и отобрав значения для клеточных линий K562, Hek, HepG2 по экспериментам ATAC-seq и DNase-seq. В результате сравнения на тестовых данных после бутстрепа модели показали себя намного лучше модели-трансформера LLM Gene-lm. Были достигнуты корреляции Пирсона 0.7, 0.784 и 0.671 против 0.524, 0.55 и 0.466 от Gene-lm соответственно.

Сравнение моделей LegNet и Gene-lm по клеточным линиям

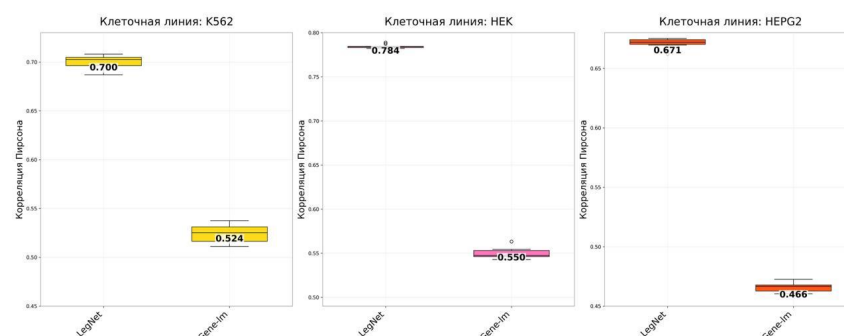


Рис.3 Результат сравнения моделей LegNet и gene-lm

Также мы протестировали обученные модели на известных

базах данных ADAstra и UDACHA для того, чтобы понять насколько хорошо модель может предсказать влияние SNP в регуляторных регионах на связывание транскрипционных факторов и на аллель-специфичную доступность хроматина. Модель угадывает большую долю эффектов по знаку их влияния, однако корреляция низкая и модель не научилась уверенно различать разные эффекты в контексте этих задач.

Список литературы:

1. D.D. Penzar et al. (2023) LegNet: a best-in-class deep learning model for short DNA regulatory regions., Nat Biotechnol.
2. Avsec, Ž., Agarwal, V., Visentin, D. et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods 18, 1196–1203 (2021). <https://doi.org/10.1038/s41592-021-01252-x>
3. He AY, Danko CG. Dissection of core promoter syntax through single nucleotide resolution modeling of transcription initiation. bioRxiv [Preprint]. 2024 Sep 17:2024.03.13.583868. doi: 10.1101/2024.03.13.583868. PMID: 38559255; PMCID: PMC10979970.

Разнообразие Quorum Quenching ферментов у прокариот

Соколов М.Н., Зайцева Ю.В.

*Ярославский государственный университет им. П.Г.
Демидова, Ярославль
melsudbi@yandex.ru*

Введение

Quorum Quenching (QQ) является механизмом конкурентной борьбы у бактерий и включает различные подходы, направленные на снижение эффективности Quorum Sensing систем (QS). Так как QS регулирует экспрессию генов, связанных с проявлением факторов патогенности и образованием биопленок, механизм QQ, как стратегия подавления QS, является

перспективным направлением исследований.

Ферментативная деградация сигнальных молекул QS является одним из подходов, лежащих в основе механизма QQ. Эволюционно QQ складывался не только как стратегия межвидовой конкуренции, он был также связан с системами саморегуляции бактериальных клеток и механизмами устойчивости к антибактериальным веществам, что является причиной разнообразия гидролитических QQ ферментов.

Целью данной работы является анализ разнообразия и распространения генов Quorum Quenching ферментов в геномах прокариот, представленных в базах данных.

Материалы и методы

В ходе работы из базы данных NCBI () при помощи алгоритма BLAST были получены 429 последовательностей белков, не менее, чем на 70 % гомологичных известным из литературы QQ ферментам. Они относились к группам фосфотриэстераз (PTE), металло- β -лактамаз (MBL), α/β -гидролаз и Ntn-ацилаз и принадлежали к грам-положительным и грам-отрицательным бактериям, относящимся к 104 родам и четырем филумам бактерий (Actinomycetota, Bacillota, Bacteroidota, Pseudomonadota), а также 3 родам и одному филуму архей (Thermoproteota). Множественное выравнивание последовательностей было выполнено с использованием программы MAFFT в конфигурации linsi [1]. Филогенетический анализ был произведен в программе IQTree методом максимального правдоподобия (ML) с 1000 реплик бутстрэп-поддержки [2].

Результаты и обсуждения

Наиболее распространенной группой QQ ферментов являются MBL-подобные лактоназы, они были обнаружены у бактерий, относящихся к четырем филумам - Actinomycetota, Bacillota, Bacteroidota и Pseudomonadota. Среди них выделились четыре группы: в первую вошли AiiA, AhlD и AhlS; во вторую AiiB, AidC, RmmL и QlcA; в третью и четвертую обособились белки, высокогомологичные AhlK и AttM соответственно. Последние наиболее гомологичны MBL, не связанным с ингибированием QS.

Фосфотриэстеразоподобные лактоназы, в нашем исследовании были представлены ферментами бактерий QsdA и MCP (Actinomycetota), и тремя ферментами архей - SisLac, SsoPox

и VmoLac (Thermoproteota). Белки этой группы являются наименее распространенным типом ферментов. Среди них явно выделяются две группы, относящиеся к бактериям и археям, однако все эти ферменты имеют общие консервативные участки.

Среди Ntn-ацилаз выделяются две большие группы. Одна представлена ферментами HscB, QuiP, KcPGA и PflA, гомологичными G-пенициллин-амидазам. Вторая группа образована белками AaC, PvdQ и AhlM, гомологичными акулеацин-A-ацилазам. Все эти ферменты распространены, главным образом, среди бактерий из филума Pseudomonadota.

Среди α/β -гидролаз три фермента, AidH, PenV и JydB, высоко гомологичны по отношению друг к другу, и не разделяются достоверно. Первые два белка обнаруживаются в филуме Pseudomonadota, а JydB описан у бактерий, относящихся к Actinomycetota. Два других фермента, AiiM и AidA, обособившиеся в отдельные ветви, описаны у Actinomycetota и Pseudomonadota соответственно.

Показано, что Ntn-ацилазы и α/β -гидролазы проявляют большую степень гомологии по отношению друг к другу, чем к двум другим группам ферментов, выделившихся в отдельные кластеры.

Выводы

Таким образом, различные QQ ферменты распространены у бактерий и архей в различной степени. Наиболее распространенными являются металло- β -лактамазы. Фосфотриэстеразоподобные лактоназы, напротив, среди бактерий встречаются лишь у Actinomycetota, зато обнаруживаются у некоторых архей. Наибольшее разнообразие QQ ферментов выявлено у бактерий из филумов Pseudomonadota и Actinomycetota.

Исследование выполнено за счет гранта Российского научного фонда № 25-16-20136.

Литература

1. Katoh K., Standley D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability //Molecular biology and evolution. – 2013. – Т. 30. – №. 4. – С. 772-780.
2. Hoang D. T. et al. Improving the ultrafast bootstrap

approximation //Molecular biology and evolution. – 2018. – Т. 35. – №. 2. – С. 518-522.

GTRD в эпоху секвенирования одиночных клеток

Т.С. Соколова¹, М.А. Куляшов¹, С.А. Панишева²,
Д.Р. Салимов¹, С.К. Колмыков¹

¹*Научно-технологический университет «Сириус», 354340,
Российская Федерация, Краснодарский край, федеральная
территория «Сириус», Олимпийский пр., д.1,
sokolova.ts@talantiuspeh.ru*

²*Новосибирский государственный университет, 630090,
Новосибирск, ул. Пирогова, 1*

Регуляция транскрипции — многоуровневый процесс, включающий в себя связывание транскрипционных факторов с ДНК, изменение организации хроматина, модификации гистонов и метилирование ДНК. Высокопроизводительное секвенирование (NGS) стало основным инструментом изучения этих процессов. В распоряжении научного сообщества находятся десятки тысяч различных типов экспериментов: ChIP-seq, DNase-seq, ATAC-seq, MNase-seq и других типов. Эти данные позволяют строить интегративные карты регуляции транскрипции и анализировать согласованность и воспроизводимость результатов между различными экспериментами и методами.

Однако следует учитывать, что большинство таких данных получено из биологических образцов, представляющие собой гетерогенные по составу популяции клеток. Это накладывает ограничения на интерпретацию результатов: например, при сравнении профилей открытого хроматина в разных условиях мы фактически анализируем усреднённую реакцию разнородного набора клеток, игнорируя индивидуальные различия между ними. Появление технологий секвенирования одиночных клеток, прежде всего single-cell RNA-seq (scRNA-seq) и single-cell ATAC-seq (scATAC-seq), дало возможность анализировать динамику

регуляторных событий и сопоставлять эпигенетическое состояние с транскрипционной активностью на уровне одиночных клеток и клеточных субпопуляций.

В рамках данной работы в БД GTRD были интегрированы 4413 scATAC-seq эксперимента из БД GEO NCBI. Особый интерес представляет интеграция single-cell данных с другими типами NGS данных. На рисунке 1 приведено сопоставление различных типов NGS экспериментов из БД GTRD с scATAC-seq данными по клеточному типу/ткани.

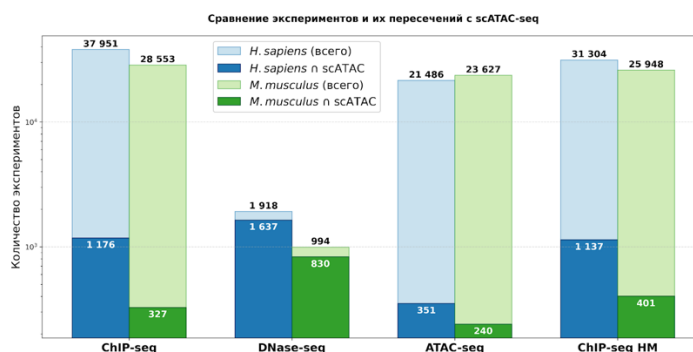


Рисунок 1. Количества NGS экспериментов в GTRD, совпадающих с scATAC-seq данными, по клеточным типам/тканям.

Полуавтоматическая аннотация экспериментов из БД GEO NCBI осуществляется программой Geominer, адаптированной для работы с single-cell данными и поддерживающей основные биологические онтологии клеточных типов и тканей: BRENDA, Uberon, Cell Ontology, Cellosaurus, EFO. Для обработки этих данных разработан специализированный пайплайн (Рисунок 2), включающий этапы контроля качества (FastQC), фильтрации (fastp), выравнивания (Bowtie2) и последующего анализа в фреймворке ArchR. Ключевые этапы анализа включают нормализацию методом TF-IDF, снижение размерности с помощью LSI и визуализацию кластеров клеток через UMAP.

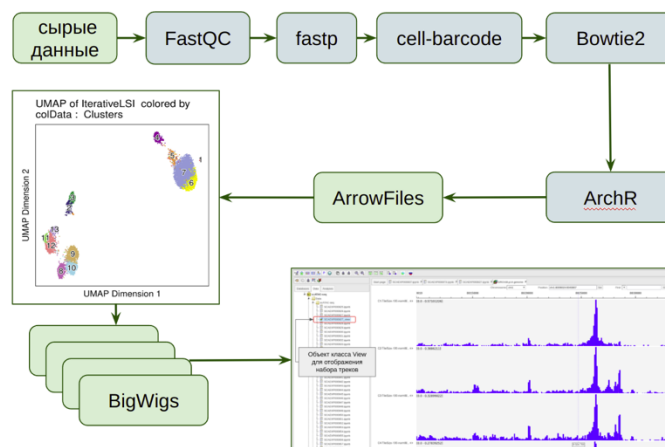


Рисунок 2. Обобщенный конвейер анализа данных single-cell ATAC-seq на платформе BioUML: от обработки сырых данных, оценки качества FastQC, фильтрации данных с помощью fastp, идентификации cell-barcode, картирования с помощью Bowtie2 к формированию объектов в ArchR, масштабированию, кластеризации и визуализации (UMAP), а также экспорту результатов в формате BigWig и их дальнейшей визуализации в геномном браузере.

Встроенные инструменты визуализации на платформе BioUML, включая интеграцию с Jupyter-notebook, позволяют проводить комплексный анализ, совмещая данные scATAC-seq с другими типами экспериментов (ChIP-seq, DNase-seq). Пользователи могут выполнять кластеризацию клеток стандартными методами (UMAP, t-SNE), анализировать специфические кластеры и сравнивать их регуляторные профили. Это открывает новые перспективы для изучения механизмов транскрипционной регуляции и преодоления ограничений, связанных с усреднением сигналов в гетерогенных популяциях клеток.

Таким образом GTRD предоставляет исследователям уникальные возможности для сравнительного анализа bulk- и single-cell данных, позволяя выявлять отдельные клеточные субпопуляции и изучать регуляцию транскрипции на клеточном

уровне.

Исследование выполнено за счет гранта Российского научного фонда № 24-24-20108, <https://rscf.ru/project/24-24-20108/>.

In-depth analysis of WGS metagenomic data unveils signatures explaining features in periodontitis progression

I.V. Sonets¹, I.S. Galeeva¹, D.V. Krivonos¹, A.V. Pavlenko¹,
A.V. Vvedenskiy¹, A.A. Ahkmetzyanova¹, K.A. Mikaelyan¹,
E.N. Ilina¹, O.O. Yanushevich², Z.E. Revazova², E.I.
Vibornaya², G.S. Runova², V.V. Aliamovskii², I.S. Bobr², M.O.
Tsargasova², E.I. Kalinnikova², V.M. Govorun¹

¹*Research Institute for Systems Biology and Medicine (RISBM),
18, Nauchniy Proezd, 117246 Moscow, Russia,
vgovorun@sysbiomed.ru*

²*FSBEI HE "ROSUNIMED" of MOH of Russia, Delegatskaya
Street 20/1, 127473 Moscow, Russia. info@rosunimed.ru*

Periodontitis is a chronic inflammatory disease characterized by the progressive destruction of the tooth-supporting apparatus, including the gingiva, periodontal ligament, and alveolar bone. Clinically, it progresses from mild stages (gingival inflammation, pocket formation) to severe forms involving irreversible bone loss, tooth mobility, and edentulism [1]. Globally, severe periodontitis affects 11% of adults, making it the sixth most prevalent human disease [2]. Beyond oral health, periodontitis is linked to systemic conditions such as diabetes mellitus and cardiovascular disease [3]. These associations underscore its role as a modifier of systemic inflammation, driven by dysbiotic microbial communities that translocate into circulation via ulcerated periodontal pockets [4].

The oral microbiome comprises over 1000 bacterial species, alongside archaea, fungi, and viruses, forming complex biofilms that interact dynamically with the host immune system [5]. Environmental stressors (e.g., poor oral hygiene, smoking) trigger dysbiosis, favoring pathobionts such as *Porphyromonas gingivalis*, *Tannerella forsythia*, and

Treponema denticola—collectively termed the "red complex" [6]. These taxa produce virulence factors (e.g., gingipains, leukotoxins) that degrade host tissues, activate destructive immune responses, and disrupt microbial homeostasis [7]. Despite their established role, the 16S rRNA metagenomic sequencing studies have failed to consistently correlate red complex abundance with disease severity, suggesting unresolved heterogeneity in microbial virulence or host susceptibility [8]. On the other hand, whole-genome shotgun (WGS) metagenomics addresses these limitations by sequencing whole-length microbial DNA in a sample, bypassing PCR amplification biases and providing comprehensive genomic insights [9].

Our objective was to identify microorganisms that are associated with the severity of the disease by performing taxonomy classification, analyzing relationships between the oral microbiome composition of periodontitis-infected patients and covariates, alpha- and beta-diversity estimation and differential abundance testing. Furthermore, we examined the variations in the structure of the microbial interaction network between these two groups.

In our study, we investigated the oral microbiome composition in 28 outpatients with periodontitis associated with *P. gingivalis* by WGS metagenomics. Samples were obtained from 13 patients with mild cases of periodontitis and 15 patients with severe cases of periodontitis upon their admission to the dentist. Patients fulfilled all the following inclusion criteria: 1) men and women aged 18 years and older with a confirmed diagnosis of "Chronic periodontitis (K05.3) newly diagnosed or previously treated in the acute stage" according to PCR (oropharyngeal swab); 2) mild and severe dynamics of the course of periodontitis according to CT scan; 3) signed voluntary informed consent for participation in the study. The metadata were collected based on questionnaires given to the patients. Sequencing was performed on a DNBSEQ-G400 (MGI) instrument using the DNBSEQ-G400RS High-throughput Sequencing Kit (FCL PE200). Sequencing was performed in paired-end mode 2 X 200 bp and dual barcode reading 2 X 10. The statistical analysis of the microbiome data was conducted using R version v.4.4.3. The diversity composition of the bacterial microbiome was assessed using the Shannon index of alpha-diversity. Beta-diversity was calculated with the principal coordinate analysis (PCoA) Unifrac weighted metric. The PERMANOVA was performed to identify

associations between taxa and host parameters with adonis function, running PERMANOVA with 1000 permutation tests on weighted UniFrac distance. For differential abundance analysis at the WGS OTU level, the DESeq2[10] with default Wald test with Benjamin–Hochberg p-value correction was used to assess the statistical significance of log2FC with significance cutoffs $p\text{-value} < 0.1$ and $\text{abs}(\log_2\text{FC}) \geq 1.5$ were applied. The co-abundance networks of WGS data were generated using the SPIEC-EASI algorithm with the Meinshausen–Bühlmann method for detecting correlations. The following parameters were utilized: 150 subsamples, 25 lambda iterations, and a $\text{min}(\text{lambda})=0.001$. From the resulting network, clusters of co-abundant OTUs were identified using the Louvain method. Eigenvector centrality and degree centrality were used for defining hubs/keystone taxa (nodes with a centrality value above the empirical 95% quantile).

The resulting dataset consists of 518 species that belong to 94 families according to the MetaPhlAn[11] taxonomic classification. Among the microbial composition of samples collected from 28 patients with varying degrees of periodontitis severity, the top ten dominant families were identified as Prevotellaceae, Treponemataceae, Fusobacteriaceae, Porphyromonadaceae, Neisseriaceae, Actinomycetaceae, Flavobacteriaceae, Selenomonadaceae, Micrococcaceae and Tannerellaceae. After filtering MAGs, which were de novo assembled with MEGAHIT [12], binned with MetaBAT2 [13], QC-ed and dereplicated with CheckM [14] and dRep [15] respectively, by completeness $\geq 80\%$ and contamination $\leq 10\%$, we obtained 193 MAGs. 424 metabolic pathways is predicted across all samples with HUMANN[16].

Although we found no significant differences between patients with mild and severe periodontitis in their alpha-diversity, we observed statistical differences between male and female patients - men tend to have more diverse microbiome composition. Across three BMI groups (normal weight, obesity and overweight) patients with obesity had the most diverse microbiome. Treatment of periodontitis of patients prior to their admission in this study shows decreasing alpha-diversity of their microbiome. We found no beta-diversity differences between patients of mild and severe groups, however, statistically significant differences were observed between sexes, as well between patients who underwent antibiotics treatment in the last six months prior to their admission to the

dentist, showing profound effect on microbiome composition after such long periods of time. For testing the association between the microbiome and such covariates as age group, severity of periodontitis, IBD, BMI group, hypertension, smoking and coronary artery disease (CAD), we applied a PERMANOVA to MAGs dataset. These covariates explained 46% of the microbiome taxonomic composition with the most extensive and statistically significant contribution coming from the comparison age group (16%) and stage of severity of periodontitis (9%). During the DESeq2 [10] differential abundance analysis, we observed significant differences in the abundance of four OTUs, all of them exhibited associations with the severe degree group in the context of periodontitis. We conducted the same analysis, using prior antibiotics treatment as a grouping factor. This time we discovered significant differences in the abundance of five OTUs, two of them are associated with prior antibiotics treatment and three showing their association with absence of such treatment. Our findings were additionally verified with estimating prevalence of these OTUs and calculating their log10 abundance. Analysis of networks made with NetCoMi [17] was focused on top-100 highly represented OTUs in the dataset, aiming to gain insights into the microbial dynamics and community structure associated with different disease severity levels. Mild group had larger N of components (48 vs 42 in the severe group), but lower clustering coefficient (0.062 vs 0.067 in the severe group) and lower edge density (0.01184 vs 0.01068 in the severe group). The network analysis identified 6 hubs/keystone taxa for the mild group, including GGB1025_SGB1320 (identified as a candidate species *Candidatus Bacteroides pericalifornicus*), SGB1343_Bacteroidota_unclassified, SGB2052_Phocaeicola_abscessus, SGB19844_TM7_phylum_oral_taxon_349, SGB5825_Diaster_invisus, SGB9434_Eikenella_corrodens, and 6 hubs/keystone taxa for the severe group: SGB6629_Filifactor_alocis, SGB2057_Porphyromonas_gingivalis, SGB1580_Prevotella_koreensis, SGB69060_GGB49229 (identified as a member of *Cryptobacteroides* genus with GTDB), SGB5901_Mitsuokella_sp_oral_taxon_131.

This study was focused on the analysis of the oral microbiome samples from patients with periodontitis whose diagnosis was confirmed with CT scan, allowing to clearly distinguish two groups of patients. The decision to concentrate on disease severity is driven by the fact that the

comparison of healthy individuals and periodontitis-infected individuals has been extensively studied in the literature. Although we didn't find significant differences in alpha- and beta-diversity between 2 groups on raw WGS data level, we observed some differences between two groups on MAGs level with PERMANOVA and microbial network analysis. DESeq2 analysis also showed differences in abundances for small number of microorganisms. According to various reports, discovered microorganisms, e.g. *Filifactor alocis*, may be used as a diagnostic indicator of periodontal disease [18]. Impact of other microorganisms on stage of severity of periodontitis is yet to be discovered, e.g. novel species *Prevotella koreensis* first described in [19]. We anticipate that the findings from our study will enhance the understanding of the mechanisms underlying the impact of clinical manifestations of periodontitis on the human oral cavity microbiota.

1. Caton J, Armitage G, Berglundh T, et al. A new classification scheme for periodontal and peri-implant diseases and conditions – Introduction and key changes from the 1999 classification. *J Clin Periodontol.* 2018; 45(Suppl 20): S1–S8. <https://doi.org/10.1111/jcpe.12935>
2. Kassebaum NJ, Bernabé E, Dahiya M, et al. Global burden of severe periodontitis in 1990-2010: a systematic review and meta-regression. *J Dent Res.* 2014;93(11):1045-1053. doi:10.1177/0022034514552491
3. Liccardo D, Cannavo A, Spagnuolo G, et al. Periodontal Disease: A Risk Factor for Diabetes and Cardiovascular Disease. *Int J Mol Sci.* 2019;20(6):1414.doi:10.3390/ijms20061414
4. Hajishengallis, G. Periodontitis: from microbial immune subversion to systemic inflammation. *Nat Rev Immunol* 15, 30–44 (2015). <https://doi.org/10.1038/nri3785>
5. Li, X., Liu, Y., Yang, X., et al. (2022). The Oral Microbiota: Community Composition, Influencing Factors, Pathogenesis, and Interventions. *Frontiers in Microbiology*, 13. <https://doi.org/10.3389/fmicb.2022.895537>
6. Adel-Khattab, D., Groeger, S., Domann, et al. (2021). *Porphyromonas gingivalis* induced up-regulation of PD-L1 in colon carcinoma cells. *Molecular Oral Microbiology*, 36(3), 172–181. <https://doi.org/10.1111/omi.12332>
7. Hajishengallis, G. (2014). Periodontitis: from microbial immune

- subversion to systemic inflammation. *Nature Reviews Immunology*, 15(1), 30–44. <https://doi.org/10.1038/nri3785>
8. Jiang, Y., Song, B., Brandt, B., et al. (2021). Comparison of Red-Complex Bacteria Between Saliva and Subgingival Plaque of Periodontitis Patients: A Systematic Review and Meta-Analysis. *Frontiers in Cellular and Infection Microbiology*, 11. <https://doi.org/10.3389/fcimb.2021.727732>
 9. Ranjan, R., Rani, A., Metwally, A., et al. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, 469(4), 967–977. <https://doi.org/10.1016/j.bbrc.2015.12.083>
 10. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014). <https://doi.org/10.1186/s13059-014-0550-8>
 11. Blanco-Míguez, A., Beghini, F., Cumbo, F. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 41, 1633–1644 (2023). <https://doi.org/10.1038/s41587-023-01688-w>
 12. Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, Tak-Wah Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, *Bioinformatics*, Volume 31, Issue 10, May 2015, Pages 1674–1676, <https://doi.org/10.1093/bioinformatics/btv033>
 13. Kang, D. D., Li, F., Kirton, E., et al. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7, e7359. <https://doi.org/10.7717/peerj.7359>
 14. Parks, D. H., Imelfort, M., Skennerton, C. T., et al. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7), 1043–1055. <https://doi.org/10.1101/gr.186072.114>
 15. Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through

- de-replication. *The ISME Journal*, 11(12), 2864–2868.
<https://doi.org/10.1038/ismej.2017.126>
16. Beghini, F., McIver, L. J., Blanco-Míguez, A., et al. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, 10. <https://doi.org/10.7554/elife.65088>
 17. Peschel, S., Müller, C. L., von Mutius, E., et al. (2020). NetCoMi: network construction and comparison for microbiome data in R. *Briefings in Bioinformatics*, 22(4). <https://doi.org/10.1093/bib/bbaa290>
 18. Aja, E., Mangar, M., Fletcher, H. M., & Mishra, A. (2021). Filifactor alocis: Recent Insights and Advances. *Journal of Dental Research*, 100(8), 790–797. <https://doi.org/10.1177/00220345211000656>
 19. Park, S.-N., Lim, Y. K., Shin, J. H., et al. (2019). *Prevotella koreensis* sp. nov., Isolated from Human Subgingival Dental Plaque of Periodontitis Lesion. *Current Microbiology*, 76(9), 1055–1060. <https://doi.org/10.1007/s00284-019-01720-w>

Heuristic algorithms suitable for implementation into in vitro evolution experiment based on microfluidic engineering

A.V. Spirov¹, E. M. Myasnikova²

¹*The Institute of Scientific Information for Social Sciences
 Russian Academy of Sciences, Moscow, Russia
 alexander.spirov55@gmail.com*

²*I. M. Sechenov Institute of Evolutionary Physiology and
 Biochemistry Russian Academy of Sciences, St. Petersburg, Russia*

In vitro evolution, also referred to as directed evolution, is a cutting-edge bioengineering method that replicates the process of natural selection in order to guide proteins or nucleic acids toward the acquisition of specific and desirable functions [1-3]. Among the most forward-looking directions in the advancement of technologies for in vitro macromolecular evolution is the integration of microdroplet

platforms grounded in the core principles of microfluidics [4-5]. These devices allow for the compartmentalized execution of experimental procedures at a microscale, enhancing the precision and control of evolutionary workflows [6].

A key strategy for improving the efficiency and reducing the cost of in vitro evolution lies in the exploration and application of novel heuristic evolutionary algorithms. These algorithms are tested and optimized through computational modeling of realistic molecular genetic procedures—commonly referred to as in silico experimentation [7-8]. The motivation for our current work stems from a recent central observation: many of these heuristic algorithms, well-established within the field of evolutionary computing, can be meaningfully mapped onto new design principles for the functional modules of microfluidic systems.

In this study, we have developed and computationally tested several innovative approaches to in vitro evolution, with the aim of translating them into viable experimental implementations using microfluidic platforms. The numerical results obtained from our in silico experiments demonstrate that certain algorithms, previously validated in evolutionary computation contexts, show markedly higher performance compared to those algorithms that represent conventional workflows in today's microdroplet-based in vitro evolution [7-8].

Our findings suggest a promising direction for improving experimental evolution techniques by incorporating these advanced algorithms into the design and operation of future microfluidic systems. The prospects for adapting these computationally promising strategies to real-world experimental configurations are also considered and discussed in the context of ongoing technological development in the field.

The work is supported by RSCF grant 22-18-00383 (extension).

1. M. Darmostuk et al. (2015) Current approaches in SELEX: An update to aptamer selection technology, *Biotechnol Adv.*, 33(6 Pt 2):1141-61.

2. M.A. Vorobyeva et al. (2018) Key Aspects of Nucleic Acid Library Design for in Vitro Selection. *Int J Mol Sci.* 19(2):470.

3. N. Komarova, A. Kuznetsov (2019) Inside the Black Box: What Makes SELEX Better?, *Molecules.* 24(19):3598.

4. A.D. Griffiths, D.S. Tawfik (2006) Miniaturising the laboratory in emulsion droplets, *Trends Biotechnol.*, 24(9):395-402.
5. B.M. Paegel, G.F. Joyce (2010) Microfluidic compartmentalized directed evolution, *Chem Biol.*, 17(7):717-24.
6. J. J. Trausch, R.T. Batey (2015) Design of modular “plug-and-play” expression platforms derived from natural riboswitches, *Methods Enzymol*, 550:41-71.
7. A.V. Spirov, E.M. Myasnikova (2022) Heuristic algorithms in evolutionary computation and modular organization of biological macromolecules: Applications to in vitro evolution, *PLoS ONE* 17(1): e0260497.
8. A.V. Spirov, E.M. Myasnikova (2023) Problem of Domain/Building Block Preservation in the Evolution of Biological Macromolecules and Evolutionary Computation, *IEEE/ACM Trans Comput Biol Bioinform*, 20(2):1345-1362.

**Region-specific annotation of scRNA-seq of
hypothalamic neurons based on open-source spatial
transcriptomics data**

Z.R. Starinnov¹, M.S. Sabirov¹, E.A. Chikina¹, R.A. Romanov²

¹*Koltzov Institute of Developmental Biology of the Russian
Academy of Sciences, 26 Vavilov Street, Moscow,
z.starinnov@idbras.ru*

²*Department of Molecular Neurosciences, Center for Brain
Research, Medical University of Vienna, Vienna, Austria,
Roman.Romanov@meduniwien.ac.at*

Modern single-cell RNA-sequencing technology is an extremely useful and important method for analyzing various heterogeneous tissues. Also it is very helpful, some of the useful information regarding the localization of each cell was not available for a time. This drawback was fixed by the methods of spatial transcriptomics, which give spatial dimension to the cells in addition to transcriptomal information. Those methods are rapidly developing and currently allow high-quality

subcellular resolution for whole transcriptome. There are many variations of such methods ranging from FISH-based (seq-FISH, MERFISH) to in situ capture methods (Visium, Slide-seq, Stereo-seq).

The main downside for those methods is high cost and complexity of protocols. Not many scientific groups can afford or perform the required procedures to get spatially resolved transcriptomic datasets. And yet there are already huge amount of scRNA-seq datasets, for which spatial resolution would give some interesting insights.

As such, we have decided to use open-access spatial transcriptomics datasets of mice brains to create a pipeline for projection of scRNA-seq data onto spatial datasets. As a base dataset for the model we used MERFISH datasets from Allen Brain Atlas Data Catalog (<https://knowledge.brain-map.org/>) [1], as at a time it had a good balance in high resolution and RNA-capture among available datasets. We used this dataset and open-source python library scVI[2] to create a model, that uses variational autoencoder trained on clustered MERFISH dataset to project the input scRNA-seq data onto the spatially characterized clusters of base dataset. As a result, cells from scRNA-seq data get assigned a spatial characteristic based on the label of spatial cluster from MERFISH dataset.

To test the validity of such pipeline we used hypothalamic neurons, as they are very heterogeneous, diverse and anatomically separated into multiple nuclei[3]. We used nuclei-specific scRNA-seq hypothalamic datasets to test accuracy of spatial prediction, which showed 75-95% depending on the input dataset and nuclei in question. We also used our original hypothalamic scRNA-seq data of P28 rats to find spatial annotations of significant cell clusters, results of which were biologically consistent.

As it stands, we have shown that it is possible to create a stable pipeline for spatial annotation using only open-source libraries and datasets. As more high-quality data becomes available, we will expand the model even further.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2021-1344).

1. M .Zhang et al (2023) Molecularly defined and spatially resolved cell atlas of the whole mouse brain, *Nature*, 624:343-354.
2. A.Gayoso et al (2022) A Python library for probabilistic

analysis of single-cell omics data, Nature Biotechnology, 40:163-166.

3. R.Romanov et al (2020) Molecular design of hypothalamus development, Nature, 582:246-252.

**Анализ сайтов связывания в калий-хлорном
котранспортере KCC2 с использованием молекулярно-
динамического моделирования**

Е.Ю. Степанова¹, Л.А. Ивонцин², Е.В. Машковцева^{1,2},
Я.Р. Нарциссов^{2,3}

¹ФГАОУ ВО "Российский Национальный
Исследовательский Медицинский Университет им. Н.И.
Пирогова" МЗ РФ, Москва, Россия, stepanovaekaterina20@mail.ru

²НИИ цитохимии и молекулярной фармакологии, Москва,
Россия

³Группа биомедицинских исследований, БиДиФарма ГмбХ,
Зик, Германия

Важную роль в поддержании ионного гомеостаза, объема клеток и нейрональной возбудимости играют белки семейства катион-хлорных котранспортеров (CCCs), которые опосредуют электронейтральный вторичный активный симпорт катионов натрия (Na⁺) и/или калия (K⁺) вместе с анионами хлора (Cl⁻) через плазматическую мембрану различных типов клеток. CCCs кодируются семейством генов SLC12 и подразделяются на два широко распространенных подсемейства: Na⁺-зависимые котранспортеры NKCC1,2 и NCC, а также Na⁺-независимые котранспортеры KCC1–4 [1].

Особый интерес представляет калий-хлорный котранспортер KCC2 (SLC12A5), который является одним из основных хлорных переносчиков в центральной нервной системе. Его транспортная активность поддерживает низкую концентрацию ионов хлора в нейронах, что обеспечивает развитие гиперполяризации постсинаптической мембраны при воздействии

тормозных нейромедиаторов гамма-аминомасляной кислоты и глицина на соответствующие рецепторы. Нарушение функции КСС2 ассоциировано с рядом неврологических заболеваний, включая эпилепсию (идиопатическую генерализованную и младенческую с мигрирующими фокальными припадками), шизофрению и аутизм.

Недавние исследования с помощью криоэлектронной микроскопии (Крио-ЭМ) выявили несоответствие между установленной ранее ионной стехиометрией и структурными данными. Согласно классическим представлениям, стехиометрия транспорта в КСС2 составляет $1K^+:1Cl^-$, что согласуется с электронейтральностью. Однако в ряде крио-ЭМ структур калий-хлорных котранспортеров (КСС1–4) были определены три небелковые электронные плотности в области трансмембранного домена переносчика, которые идентифицировали как два аниона Cl^- и один катион K^+ , что противоречит принятой модели. В связи с этим, в некоторых работах предполагается существование как основного хлорного транспортного сайта, так и дополнительного аллостерического, однако их точные функциональные роли и вклад в общий механизм транспорта остаются до конца неизученными [1–3].

В данной работе для анализа ионной селективности сайтов связывания в переносчике КСС2, а также исследования координационных взаимодействий и структурных изменений, возникающих в результате сайт-специфического мутагенеза аминокислотных остатков, формирующих ион-связывающий участок, было проведено молекулярно-динамическое (МД) моделирование с помощью программ NAMD, GROMACS при периодически граничных условиях с использованием силового поля CHARMM36.

С помощью веб-сервиса CHARMM-GUI был собран мембранно-белковый комплекс трансмембранного и внеклеточного доменов одного из мономеров димера человеческого калий-хлорного котранспортера КСС2 (PDB ID: 6M23), встроенного в гомогенный билипидный слой из фосфатидилхолина (PC) и водную среду. В качестве молекул воды использовалась модель TIP3P. Полученная система была нейтрализована ионами калия и хлора в физиологической концентрации 150 ммоль/л. При

конструировании модельной системы были учтены и добавлены в структуру белка две высококонсервативные дисульфидные связи между C287-C302 и C322-C331, стабилизирующие внеклеточный домен и определяющие транспортную активность переносчика KCC2 [4]. Транспортируемые ионы Cl⁻ и K⁺ были размещены в трансмембранном домене исследуемого котранспортера в соответствии с их положениями в ближайшем гомологе KCC4 (PDB ID: 7D99). Координаты аллостерического аниона Cl⁻ были взяты из используемой для модели структуры KCC2. Кроме того, был рассмотрен мембранно-белковый комплекс полного димера человеческого калий-хлорного котранспортера KCC2 (PDB ID: 6M23). Для исследования влияния аутоингибирующего N-конца на стабильность ионов в белке с помощью программы MODELLER была достроена петля длиной 12 аминокислот (N84-P95), неразрешенная с помощью криоэлектронной микроскопии.

На основе МД траекторий были определены области локализации ионов внутри белковой структуры трансмембранного домена KCC2, а также спирали и аминокислотные остатки, формирующие сайты связывания. Критерием участия аминокислоты в координации иона являлась удаленность от него на расстояние, не превышающее 0,4 нм. Ионы считались стабильно ассоциированными в соответствующих участках при их смещении, не превышающем 0,3 нм относительно геометрического центра сайта связывания на протяжении моделирования. Потенциальный хлорный транспортный сайт (SC11) расположен выше калиевого сайта (SK) на расстоянии менее 0,4 нм и определяется амидными группами основной цепи аминокислот Gly113, Val114, Ile115. SK формируется карбонильными группами Asn110, Ile111, Thr412, Pro409 и гидроксильными группами Tyr195 и Thr412. Близкое пространственное расположение SC11 и SK предполагает взаимную координацию ионов Cl⁻ и K⁺ в процессе транспортного цикла. Хлорный аллостерический сайт (SC12) предположительно расположен ниже SK на расстоянии около 0,8 нм и определяется амидными группами основной цепи аминокислот Gly413, Ile414, Met415 и гидроксильной группой Tyr569. Результаты исследования влияния аутоингибирующего N-конца указывают на стабильную локализацию ионов в пределах соответствующих сайтов связывания каждого из мономеров KCC2. Проведенный анализ

также выявил локальные конформационные изменения в структуре сайтов связывания, сопряженные с заменой отдельных аминокислот, что отражает функциональную чувствительность этих участков к точечным мутациям.

Полученные результаты проясняют детали работы калий-хлорного котранспортера KCC2 на атомистическом уровне. Новые сведения в дальнейшем могут быть использованы для разработки сайт-специфических модуляторов, в частности, противоэпилептических препаратов.

1. Y. Xie et al. (2020) Structures and an activation mechanism of human potassium-chloride cotransporters, *Science advances*, 6: eabc5883.

2. S. Liu et al. (2019) Cryo-EM structures of the human cation-chloride cotransporter KCC1, *Science*, 366: 505–508.

3. X. Chi et al. (2021) Cryo-EM structures of the full-length human KCC2 and KCC3 cation-chloride cotransporters, *Cell Research*, 31: 482–484.

4. A. Hartmann et al. (2010) Differences in the large extracellular loop between the K⁺-Cl⁻ cotransporters KCC2 and KCC4, *Journal of Biological Chemistry*, 285: 23994–24002.

Microbial community of the deep underground lake of Baksan Neutrino Observatory

Kirill Tarasov¹, Mikhail Zarubin¹, Alena Yakhnenko¹,
Albert Gangapshev², Elena Kravchenko¹

¹*Joint Institute for Nuclear Research, Dubna, Moscow region,
ktarasov@jinr.ru*

²*Baksan Neutrino Observatory, Institute for Nuclear Research
RAS, Neutrino village, Kabardino-Balkaria*

It is estimated that approximately 12-20% of all biomass on planet is concentrated in the Earth's crust below the surface and since it

is dominated by uncultivated organisms (>99%), it is entitled “dark matter of the biological universe” [1,2]. 16S rRNA gene community profiling is a typical technique for studying communities of such uncultivated “dark matter”. While this technique provides one with the information about typical members, representing the community, it also allows to establish these members’ abundances, giving the valuable information about community metabolic interactions. This approach is widely applied for studying deep communities, though there are only a few sites, where such communities could be sampled. Most of them represent mines, nuclear waste storages, and deep scientific plants. Baksan Neutrino Observatory (BNO), Kabardino-Balkaria, Russia, is one of such sites, where microbiology-related research has never been conducted before. It is a horizontal tunnel 4-km long with the deepest point of 2 km under the mountain Andyrchi surface. There is an outlet of the hot mineralized spring at the far unused end of the tunnel, and the lake was formed 30 m away from the spring and is fed by it.

In current work we performed 16S rRNA gene profiling of microbial community of the lake sediments. The lake community is mainly comprised of Nitrospirata (56%), Proteobacteria (35%) and Planctomycetota (4%) phyla with also minor quantities of Actinobacteriota, Acidobacteriota, Gemmatimonadota. Most part of the community is represented with autotrophic members (at least 64%) with nitrite-oxidizing bacteria (NOB) accounting for 56%. Other members, participating in reactions of nitrogen cycle, are represented with ammonia-oxidizing bacteria (AOB) (1.8%). Methanotrophs of I and II type are also present and together comprise ~7.5%. Nearly 30% of the community members are heterotrophic. No members participating in sulfur cycle are present.

Previously we presented metagenome analysis of the biofilm sample from the puddle at the bottom of the hot mineralized spring outlet [3]. In current work we also present comparison of the lake sediment and the biofilm microbial communities. Both communities mostly shared the same lineages of bacteria, though some lineages were absent in the lake community. AOB/NOB abundances ratio in the lake community was much more less (0.03) than in the biofilm community (8.8) due to enormous amount of Nitrospirata members. Some members of this group are known to perform complete ammonia oxidation, and though the genes for ammonia oxidation were not detected in the Nitrospirata

members of the biofilm community, they were present in the overall biofilm community, and this speaks in favor of the hypothesis that in the lake community Nitrospirota members could be able to perform comammox process. The percentage of heterotrophic members of the lake community is less (30%) than that of the biofilm community (71%), with total absence of predatory bacteria, which are likely unable to prey without high density of cells provided with the biofilm.

Together with our metagenomic work which is currently on review, our work concerning the cultivated member of the community [4], and current work we believe that we provide more broad microbiological overview of the BNO deep site in particular, as well as of the granitic rock habitat in general.

The research was supported by the Russian Science Foundation grant № 24-24-00003.

References:

1. Magnabosco C, Lin L-H, Dong H, Bomberg M, Ghiorse W, Stan-Lotter H, et al. (2018) The biomass and biodiversity of the continental subsurface, *Nat Geosci.*, 11: 707–717.
2. Soares A, Edwards A, An D, Bagnoud A, Bradley J, Barnhart E, et al. (2023) A global perspective on bacterial diversity in the terrestrial deep subsurface, *Microbiology*, 169.
3. Tarasov K, Kravchenko E, Zarubin M, Yakhnenko A. (2024) Deep underground metagenome-assembled genomes from hydrothermal spring, *Microbiol Resour Announc*, 13: e0057424.
4. Tarasov K, Yakhnenko A, Zarubin M, Gangapshev A, Potekhina NV, Avtukh AN, et al. (2022) *Cytobacillus pseudoceanisediminis* sp. nov., A Novel Facultative Methylophilic Bacterium with High Heavy Metal Resistance Isolated from the Deep Underground Saline Spring, *Curr Microbiol.*, 80: 31.

Влияние гаплотипа HLA на уровень антител к β -клеткам поджелудочной железы при сахарном диабете 1 типа

Е.А. Тельная¹, Е.С. Петрайкина¹, Д.И. Тычинин¹, В.П. Богданов², З.Г. Антышева², Д.И. Табаков², Н.В. Лаврушкина¹, Е.А. Снигирь¹, А.В. Махотенко¹, О.М. Марченко¹, А.А. Ивашечкин¹, С.И. Митрофанов¹, П.Ю. Волчков², В.С. Юдин¹, А.А. Кескинов¹, С.М. Юдин¹, Д.В. Светличный¹, М.В. Воронцова², В.И. Скворцова³

¹Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства (ФГБУ «ЦСП» ФМБА России). 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1., etelnaya@cspfmba.ru

²Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр оригинальных и перспективных биомедицинских и фармацевтических технологий». 125315, Россия, г. Москва, ул. Балтийская, д. 8.

³Федеральное медико-биологическое агентство (ФМБА России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.

Сахарный диабет первого типа (СД1) аутоиммунное заболевание, поражающее β -клетки поджелудочной железы. Важными клиническим маркерами аутоиммунного процесса являются аутоантитела к декарбоксилазе глутаминовой кислоты (GAD), тирозин-фосфатазе (IA2), а также трансмембранным белкам-транспортерам ионов цинка типа 8 (ZnT8) [1-2]. Риск развития СД1 имеет высокую генетическую составляющую, а также определяется действием средовых факторов. В частности, определенные гаплотипы человеческого лейкоцитарного антигена (human leukocyte antigen, HLA) класса II, такие как DR4-DQ8, DR3-DQ2, сильно ассоциированы с развитием СД1 [3]. В нашем исследовании выполнен анализ HLA гаплотипов, полученных из данных полногеномного секвенирования 524 пациентов с СД1 со стажем не более двух лет в контексте ассоциации с уровнем аутоантител, характерных для СД1.

В результате исследования установлена связь серотипа DR4 с повышенным уровнем IA2. Для серотипа DQ8 подобной ассоциации выявлено не было. Проведенный статистический анализ выявил, что в ассоциированном с СД1 гаплотипе DR4-DQ8 именно DR4 связан с увеличением уровня IA2. Исследованная выборка пациентов позволила выявить, что у пациентов с серотипом DR15 уровень IA2 понижен, относительно пациентов с другими гаплотипами ($p\_value = 0.005$). Гаплотип DR15-DQ6 считается гаплотипом с низким уровнем риска СД1 [4]. Полученные данные допускают, что серотип DR15 ассоциирован с отсутствием антител к IA-2 даже при гетерозиготности аллели. В результате вычислительного анализа выявлены доминантно-рецессивные взаимоотношения между HLA аллелями в контексте формирования количественного уровня аутоантител при СД1. В частности, для серотипа DR15, уровень антител к IA-2 выше порогового значения лишь в сочетании с ассоциированным с СД1 DR4, что дает основания полагать что в этом сочетании связь DR4 с антителами к IA-2 доминантна над связью DR15 с антителами к IA-2.

Таким образом, результаты работы демонстрируют возможность изолированного действия серотипов DR и DQ внутри одной аллели на формирование профиля островко-специфичных антител. Также в работе показано, что действие одних аллелей может влиять на действие других, в связи с чем профиль антител в случаях HLA-гетерозигот определяется результатом взаимодействия аллелей.

1. K. Decochez et. al. (2000) High Frequency of Persisting or Increasing Islet-Specific Autoantibody Levels After Diagnosis of Type 1 Diabetes Presenting Before 40 Years of Age, *Diabetes care*, 23:838–844.
2. J. Wenzlau, et. al. (2015) Changes in Zinc Transporter 8 Autoantibodies Following Type 1 Diabetes Onset: The Type 1 Diabetes Genetics Consortium Autoantibody Workshop, *Diabetes care*, 38:14-20.
3. J. X. She (1996) Susceptibility to type I diabetes: HLA-DQ and DR revisited, *Viewpoint immunology today*, 17:323-329.
4. N. Kumar et.al. (2013) Genomic evaluation of HLA-DR3+ haplotypes associated with type 1 diabetes, *Annals of the New York Academy of Sciences*, 1283:91–96

Enhancing the Reverse Transcriptase Function in Taq Polymerase via AI-Driven Multiparametric Rational Design

Yulia E. Tomilova¹, Nikolay E. Russkikh², Igor M. Yi²,
Elizaveta V. Shaburova³, Viktor N. Tomilov⁴, Galina B. Pyrinova¹,
Svetlana O. Brezhneva¹, Olga S. Tikhonyuk¹, Nadezhda S. Gololobova¹,
Dmitriy V. Popichenko¹, Maxim O. Arkhipov¹,
Leonid O. Bryzgalov¹, Evgeniy V. Brenner¹, Anastasia A. Artyukh¹,
Dmitry N. Shtokalo^{2,3,5}, Denis V. Antonets^{3*}, Mikhail K. Ivanov^{1,6}

¹*AO VectorBest, Novosibirsk;*

²*AcademGene LLC, Novosibirsk;*

³*MSU Institute for Artificial Intelligence, Lomonosov Moscow State University, Moscow;*

⁴*SibEnzyme Ltd., Novosibirsk; 5 Institute of Informatics Systems SB RAS, Novosibirsk; 6 Institute of Molecular and Cellular Biology SB RAS, Novosibirsk, Russia.*

**Correspondence: antonec@yandex.ru*

Introduction. Modification of natural enzymes to introduce new properties and enhance existing ones is a central challenge in bioengineering. *Thermus aquaticus* DNA polymerase I (Taq pol) is a widely used thermostable enzyme in molecular biology, but its intrinsic reverse transcriptase (RTase) activity is very low. High-temperature RTases are attractive for single-enzyme RTqPCR workflows. Artificial intelligence (AI) methods based on protein language models (PLMs) offer a powerful route to navigate the immense mutational landscape of enzymes when multitrait optimization is required. For example, it was demonstrated that even such a simple PLM as UniRep simulated a fitness landscape accurately enough for engineering applications with using only 24 functionally characterized proteins bearing amino acid substitutions [1]. Here, we report a multiparametric, AI-guided strategy for enhancing the RTase activity of Taq pol without sacrificing other critical properties such as fidelity, thermal stability, 5'→3' exonuclease activity, and the ability to incorporate dUTP and tolerance to LNA-containing substrates [2].

Methods. We first selected 18 single mutants targeting six important sites (E507, S515, D578, H639, F667, M747) implicated in

primer, template or dNTP binding [3]. Each variant, together with the wild type (WT), was characterized in a tenparameter experimental panel comprising: (i) Rtas activity on four RNA templates; (ii) DNAdependent polymerase rate and dissociation constant; (iii) PCR efficiency and probe cleavage performance; (iv) error rate; (v) dTTP→dUTP substitution tolerance; (vi) extension speed on three LNAmodified hairpins; (vii) antibody and aptamer hotstart compatibility. Sequence embeddings were obtained with a ProtT5XL encoder finetuned on 91K Taq pol homologues. Ridge regression models were trained on the 19sequence dataset to predict each parameter; uncertaintyaware Gaussian process (GP) models replaced Ridge after a second data collection round. The GP ensemble was used for in silico screening of 18.3 M variants within a Hamming radius ≤ 3 from the WT.

Two successive cycles prioritized 29 additional single, double and triple mutants for synthesis and testing, expanding the empirical dataset to 47 variants. Top candidates were evaluated in singletube, 60 °C TaqMan RTqPCR targeting synthetic HPIV2 and hMpV RNAs.

Results. Out of 47 tested mutants, 18 showed markedly improved Rtas activity, with some achieving up to two orders of magnitude increase compared to wild-type Taq pol. Enhanced RTase activity was associated with lower Kd values, modestly reduced fidelity (median 1 error/5 200 nt vs 1/6 000 nt for WT), and greater tolerance to LNA modifications. Twelve variants met multitrait acceptance thresholds ($k_{\text{eff}} \geq 0.2 \times \text{WT}$; PCR efficiency ≥ 1.85 ; dT/dU ratio ≤ 1.5 ; error rate $\leq 1/3\,000$) and all supported hotstart control. In singleenzyme RTqPCR they achieved Cq values comparable to a commercial p66 RT/Taq mix, with clean fluorescence rising phases and probe cleavage amplitudes. Notably, double mutant E507KD732N and triple mutant E507QD578SI614M showed $\Delta Cq \leq 0.5$ relative to the twoenzyme control, while maintaining probe digestion and inhibitor tolerance.

Conclusions. We demonstrate that PLM-based embeddings, combined with Gaussian process regression, enable efficient multiparametric optimization of enzyme function with a small experimental dataset. Our AI-guided approach successfully identified Taq pol mutants combining high RTase activity and desirable polymerase traits, providing valuable candidates for singleenzyme RTqPCR diagnostic and research workflows. This methodology is

transferable to other multifunctional enzyme where exhaustive combinatorial mutational scanning is impractical. This research was partially supported by the MSU Institute for Artificial Intelligence. Wetlab reagents were provided by AO VectorBest. We thank the Federal Research Center for Information and Computational Technologies SB RAS for highperformance computing resources.

1. Biswas S. et al. (2021) LowN protein engineering with dataefficient deep learning, Nat. Methods, 18:389–396.
2. Tomilova Y.E. et al. (2024) Enhancing the reversetranscriptase function in Taq polymerase via Aldriven multiparametric rational design, Front. Bioeng. Biotechnol., 12:1495267.
3. Raghunathan G., Marx A. (2019) Engineering DNA polymerases for improved reverse transcription, Trends Biotechnol., 37:1027–1041.

CD4+ Т-клетки в сахарном диабете 1 типа: роль в патогенезе и иммунной регуляции

Д.И. Тычинин¹, Е.С. Петряйкина¹, О.А. Антонова¹, В.П. Богданов^{2,3}, З.Г. Антышева^{2,3}, Е.С. Авсиевич^{2,3}, В.С. Щекина^{2,3}, Ю.А. Крупинова^{2,3}, О.В. Глушкова¹, И.С. Владимиров¹, В.Е. Мухин¹, О.М. Марченко¹, А.А. Ивашечкин¹, О.Д. Фатеев¹, С.И. Митрофанов¹, К.С. Грамматикати¹, А.И. Некрасова¹, В.В. Горев⁴, С.И. Валиева⁴, И.Г. Рыбкина⁴, П.Ю. Волчков^{2,3}, В.С. Юдин¹, А.А. Кескинов¹, С.М. Юдин¹, Д.В. Светличный¹, М.В. Воронцова^{2,3}, В.И. Скворцова⁵

¹Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства. 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1, dtychinin@cspfmba.ru

²Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский

университет)», Центр живых систем. 141700, Россия, Московская обл., г. Долгопрудный, пер. Институтский, д. 9, стр. 7.

³Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр оригинальных и перспективных биомедицинских и фармацевтических технологий». 125315, Россия, г. Москва, ул. Балтийская, д. 8.

⁴ГБУЗ «Морозовская детская городская клиническая больница ДЗМ». 119049, Россия, г. Москва, пер. 4-й Добрынинский, д. 1/9.

⁵Федеральное медико-биологическое агентство (ФМБА России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30

Исследование иммунных изменений и определение клеточных и молекулярных механизмов в основе патогенеза сахарного диабета 1 типа (СД1) является важным для развития методов иммунной терапии и профилактики заболевания. В работе проведен анализ транскриптома единичных клеток мононуклеаров периферической крови (10x scRNAseq PBMC) с применением детального типирования CD4⁺ Т-эффекторных и регуляторных клеток. Исследовались группы детей с диагнозом СД1 (22 ребенка с впервые выявленным, 20 детей со стажем СД1 более года), их прямых родственников (родители - 4 участника, сиблинги - 4 участника) и 12 участников группы здорового контроля, подобранные по возрасту.

По результатам работы было обнаружено, что активное аутоиммунное разрушение β -клеток при СД1 сопровождается повышением доли Th22 среди эффекторных хелперов, ассоциированным с увеличением уровня экспрессии TNF и IL-6, снижением доли и функциональной активности MAIT-клеток, а также повышенным провоспалительным межклеточным сигналингом посредством ADAM10 и ADAM17. Завершение аутоиммунного процесса характеризуется увеличением доли Th17, повышением экспрессии TGF- β 1 и IL-12.

Было выявлено, что рост доли наивных Т-регуляторных клеток (Т-рег) может являться маркером ранних этапов развития СД1. Анализ экспрессии генов выявил значимую корреляцию

между относительной долей наивных Т-рег в структуре CD4-клеток и уровнем экспрессии провоспалительного маркера ADAM10.

Особое внимание привлёк транскрипционный фактор BACH2, который является специфическим маркером наивных Т-рег. Моделирование нокаута BACH2 продемонстрировало наиболее существенные изменения в профиле экспрессий наивных Т-рег пациентов с СД1 в сторону созревания Т-регуляторных клеток. Таким образом, эффекты нокаута BACH2 позволяют имитировать процесс созревания Т-рег, что открывает новые перспективы для понимания патогенеза заболевания и развития стратегии иммунотерапий СД1.

Иммунный профиль отличается у лиц с СД1, находящихся на разных этапах патогенеза заболевания. Уровень наивных Т-регуляторных клеток, баланс CD4+ эффекторных клеток могут считаться маркерами определенных этапов аутоиммунного процесса при СД1. Нарастание доли Th22 характеризует активное аутоиммунное разрушение, а доли Th17, в свою очередь, завершение аутоиммунного процесса.

РАЗРАБОТКА СУБЪЕДИНИЧНОЙ ВАКЦИНЫ-КАНДИДАТА ДЛЯ ПРОФИЛАКТИКИ ЛИХОРАДКИ ДЕНГЕ С ПОМОЩЬЮ МЕТОДОВ ИММУНОИНФОРМАТИКИ

А.А. Тюленев, А.А. Чернявский, А.С. Ивановский,
В.И. Тимофеев, М. А. Марченкова, Ю.В. Писаревский

*пл. Академика Курчатова, 1, Москва, 123098,
tiulenev.aa@phystech.edu*

В настоящее время одной из наиболее крупных проблем, стоящих перед здравоохранением, является проблема распространения эпидемий. Появление высоковирулентных вирусов, таких как SARS-CoV-2, показало невозможность полной изоляции очага эпидемии вследствие мировой тенденции к глобализации. В существующих условиях быстрого распространения эпидемий способы быстрого и достаточно точного получения вакцины-кандидата видятся единственным способом

эффективно решать проблему эпидемий.

Проводимое исследование направленно на ускорение исследования вакцин-кандидатов на этапе лабораторных исследований посредством использования методов биоинформатики. В частности, был исследован вирус лихорадки денге высококонтагиозный и смертоносный вирус[1], которым ежегодно заболевают порядка 50 миллионов человек, и который является причиной порядка 10 тысяч смертей в год[2].

В настоящей работе представлен анализ результатов, исследованных ранее субдоменов надмембранной части белка лихорадки денге[3]. Предсказанная структура белка была исследована на наличие надмембранных доменов с наибольшей концентрацией эпитопов для дальнейшего использования в виде вакцин-кандидатов. Были предсказаны 3 надмембранных домена, богатых эпитопами, после чего были исследованы каждый домен по отдельности.

После проведения предварительного анализа доменов трансмембранного белка лихорадки денге, исследование было продолжено с использованием передовых методов биоинформатики и компьютерного моделирования. Данный этап работы был направлен на более детальное изучение потенциальных вакцин-кандидатов и их иммунологических свойств.

На основе ранее предсказанных субдоменов были созданы три прототипа вакцины. Для моделирования их пространственных структур применили инновационную глубокую нейронную сеть AlphaFold2, использующую архитектуру трансформер. Эта передовая технология позволила получить высокоточные трехмерные модели предполагаемых вакцин-кандидатов.

Следующим шагом стало повторное предсказание конформационных (пространственных) эпитопов. Это было необходимо для уточнения данных с учетом смоделированных трехмерных структур. Параллельно с этим провели тщательное исследование иммунологических свойств, ранее предсказанных Т-линейных эпитопов. В частности, оценивались такие важные характеристики, как иммуногенность, аллергенность, токсичность и консервативность всех выявленных антигенных детерминант по отношению к известным вариантам вируса лихорадки Денге.

Особое внимание было уделено эпитомам, у которых

целевые метрики предсказания аллергенности или токсичности превысили допустимые значения. Для улучшения их характеристик были произведены точечные аминокислотные замены. После внесения этих изменений было проведено повторное предсказание структуры и эпитопов, чтобы удостовериться в положительном эффекте внесенных модификаций.

Следующий этап исследования был посвящен оценке структурной стабильности каждой из предполагаемых вакцин-кандидатов в условиях, приближенных к физиологическим. Для этого были проведены расчеты методом молекулярной динамики с использованием программного пакета Gromacs. Моделирование проводилось в водно-солевом растворе, имитирующем физиологические условия. Использовалось силовое поле AMBER99SB-ILDN в сочетании с моделью воды TIP3P, а также добавлением 0,15 М KCl для создания ионной среды, близкой к условиям в организме.

Заключительным этапом исследования стала симуляция клеточного иммунного ответа на введение вакцин-кандидатов. Для этого использовали специализированный программный пакет C-ImmSim, который позволяет моделировать иммунный ответ организма на введение антигена. Симуляция проводилась для двух временных точек: на 1 и 25 дни после введения вакцины. Это позволило оценить как первичный, так и вторичный иммунный ответ.

Результаты симуляции оказались весьма обнадеживающими. Они продемонстрировали способность разработанных вакцин-кандидатов вызывать не только яркий первичный иммунный ответ, но и формировать стойкий вторичный иммунный ответ. Это указывает на потенциальную эффективность разработанных вакцин в создании длительного иммунитета против вируса лихорадки Денге.

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования, а также в рамках выполнения государственного задания НИЦ “Курчатовский институт”.

1. Gould E, Solomon (2008) T. Pathogenic flaviviruses. *Lancet* . 371(9611):500-509.

2. Ramos-Castañeda J, Barreto dos Santos F, Martínez-Vega R, Galvão de Araujo JM, Joint G, Sarti E (2017). Dengue in Latin America: Systematic Review of Molecular Epidemiological Trends. Beasley DWC, ed. PLoS Negl Trop Dis . 11(1):e0005224.
3. Costa RL, Voloch CM, Schrago CG. (2012) Comparative evolutionary epidemiology of dengue virus serotypes. Infect Genet Evol . 12(2):309-314.

Unproductive splicing in 3'-untranslated regions and its regulation

Y. Uglova, M. Vlasenok, D. D. Pervouchine

*Skolkovo institute of science and technology, Bolshoy
Boulevard, 30, bld. 1, Moscow 121205, Russia,
Yulia.Uglova@skoltech.ru*

The 3'-untranslated regions (3'-UTRs) of messenger RNA (mRNA), located downstream of the stop codon, are known to be tightly involved in regulation of mRNA-based processes such as localization, stability and translation. Mostly the regulation is mediated by RNA-binding proteins (RBPs), which interact with the 3'-UTRs to influence gene expression.

Splicing in 3' untranslated gene regions (3'-UTRs) that occurs more than 50–55 nucleotides downstream of the stop codon triggers nonsense-mediated decay (NMD). The evidence shows that NMD is involved not only in breaking down defective mRNA molecules caused by genetic mutations or errors in RNA processing, but also in controlling the mRNA abundance of a wide range of normal genes. NMD targets include tissue-specific transcripts [1], long 3'-untranslated regions, alternatively spliced and polyadenylated genes, mRNAs with upstream open reading frames. The mechanism involving alternative splicing and NMD to downregulate mRNA levels, is called regulated unproductive splicing and translation (RUST) [2]. It is widespread and highly conserved across different species, including humans.

Tissue-specific splicing is also a common feature of human

genes. It is essential in normal tissue development and function, and dysregulation of tissue-specific splicing has been implicated in various diseases, including cancer [3].

Recently it was shown that 3'-UTR splicing is upregulated in different types of cancer [4]. Also there are examples of regulation of splicing leading to oncogenesis [5]. These findings show that unproductive splicing may have big effects on gene regulation and contribute to disease development.

Our study systematically identifies unproductive splicing events in 3'-UTRs of human protein-coding genes and investigates their regulatory mechanisms. By combining high-throughput transcriptome data, we created a catalog of NMD-associated splicing events and identified their potential regulators.

To identify unproductive splicing events, we analyzed three types of transcriptomic data, including three NMD inactivation experiments with cycloheximide, RBP knockdown data (ENCODE), and tissue-specific gene expression data (GTEx). Differential gene expression and splicing analyses were conducted using custom bioinformatics pipelines and the pyIPSA program. We identified splicing events emerging under NMD inactivation, quantified their abundance, and calculated Percent Spliced In (PSI) values. We examined how 254 common RBP knockdowns affected these events and then analyzed the correlation between the identified events and regulator gene expression across human tissues.

Our analysis led to a catalog of approximately 1900 unproductive splicing events in 3'-UTRs. Regulatory potential was inferred for 199 events within 127 genes, including examples such as the PRPF4B protein-kinase 3'-UTR introns likely regulated by HNRNPU and ILF2, and an intron in the CRKL gene, which may be oncogenic, potentially regulated by GRSF1.

The catalog of NMD-associated splicing events and their predicted regulators can help to better understand how RUST works. To confirm the regulatory role of these interactions, further experimental validation is needed.

1. Mironov, A., Petrova, M., Sergey Margasyuk, Vlasenok, M., Mironov, A., Dmitry Skvortsov, & Pervouchine, D. D. (2023) Tissue-specific regulation of gene expression via unproductive splicing, *Nucleic*

Acids Research, 51(7):3055–3066.

2. Lewis, B. P., Green, R. E., & Brenner, S. E. (2002) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans, *Proceedings of the National Academy of Sciences*, 100(1):189–192.

3. Chou, J., Provot, S., & Werb, Z. (2010) GATA3 in Development and Cancer Differentiation: Cells GATA Have It!, *Journal of Cellular Physiology*, 222(1):42–49.

4. Jia Jia Chan, Zhang, B., Xiao Hong Chew, Adil Salhi, Zhi Hao Kwok, Chun You Lim, Ng Desi, Subramaniam, N., Siemens, A., Tyas Kinanti, Ong, S., Avencia Sánchez-Mejías, Phuong Thao Ly, An, Ö., Sundar, R., Fan, X., Wang, S., Bei En Siew, Kuok Chung Lee, & Choon Seng Chong. (2022) Pan-cancer pervasive upregulation of 3' UTR splicing drives tumourigenesis, *Nature Cell Biology*, 24(6):928–939.

5. Anczuków, O., Rosenberg, A. Z., Akerman, M., Das, S., Zhan, L., Karni, R., Muthuswamy, S. K., & Krainer, A. R. (2012) The splicing factor SRSF1 regulates apoptosis and proliferation to promote mammary epithelial cell transformation, *Nature Structural & Molecular Biology*, 19(2):220–228.

Integrative selection of highly specific CRISPR/Cas9 targets for gene-editing tasks

A. Uzdenov, A. Gribkova, A. Shaytan

*Lomonosov Moscow State University, Moscow, Russia,
azamuzdenov@gmail.com*

An integrated selection of highly specific CRISPR/Cas9 targets is essential for both basic research and applied gene-editing tasks.

This study presents a strategy that combines transcriptomics, epigenetic marks and chemical DNA modifications to choose target genes. It also provides an automatic evaluation of the selected targets with the most reliable methods. We demonstrate its experimental testing in several cell models.

First, for the K562, NHEK and HEK293 lines we collected

ChromHMM, RNA-seq and bisulfite-sequencing data, then classified genes as either active or inactive. Such classification is important when deciding whether to aim for repression or activation.

Comparing the gene classes across the different datasets confirmed the correctness of this stratification (Pearson correlation $\rho = 0.6-0.8$) and showed high agreement between sources.

We implemented a Python algorithm that extracts all 20-nucleotide sites with the NGG motif from promoter sequences and instantly calculates an overall on-target score (Doench–Fusi model, 2016), off-target specificity (Hsu, 2013), and the probability of forming DNA-DNA-RNA triplex structures.

To check how transferable the approach is, we ran the calculations for the K-562 cell line and for a bacterial plasmid expressed in HEK293. From the chosen targets, ten sequences with the best on/off scores were tested in vitro and showed the expected high efficiency.

The result of the work is an open software package that automates sgRNA selection while taking epigenomics and DNA structure into account. Using it increases the accuracy of target-gene selection and lowers the predicted number of off-site cuts. The presented strategy is easy to adapt to any cell system and can serve as a basis for rational CRISPR design.

This research was supported by the Russian Ministry of Science and Higher Education grant No. 075-15-2024-539

1. J. G. Doench, N. Fusi (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9, *Nature Biotechnology*.
2. P. D. Hsu et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases, *Nature Biotechnology*.
3. A. J. Heidersbach et al. (2023) A versatile, high-efficiency platform for CRISPR-based gene activation, *Nature Communications*.
4. R. Singh et al. (2015) Cas9-chromatin binding information enables more accurate CRISPR off-target prediction, *Nucleic Acids Research*.
5. J. Ernst, M. Kellis (2012) ChromHMM: automating chromatin-state discovery and characterization, *Nature Methods*
7. T. Warwick et al. (2022) A universal model of RNA·DNA:DNA triplex formation accurately predicts genome-wide RNA–DNA interactions, *Briefings in Bioinformatics*

8. M. S. Leisegang et al. (2024) RNA-DNA triplexes: molecular mechanisms and functional relevance, *Trends in Biochemical Sciences*
9. Z. Wang, M. Gerstein, M. Snyder (2009) RNA-Seq: a revolutionary tool for transcriptomics, *Nature Reviews Genetics*

Исследование транскрипционного профиля мобильных генетических элементов клеток периферической крови в состоянии противовирусного ответа

О.Д. Фатеев¹, А.В. Болбат¹, А.В. Абдуллатыпов¹, В.Э. Акимов¹,
Д.И. Тычинин¹, В.С. Юдин¹, С.М. Юдин¹, А.А. Кескинов¹,
Д.В. Светличный¹, В.И. Скворцова²

¹Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства (ФГБУ «ЦСП» ФМБА России). 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1.,
OFateev@cspfmba.ru

²Федеральное медико-биологическое агентство (ФМБА России). 123182, Россия, г. Москва, Волоколамское шоссе, д. 30.

Изучение активности мобильных генетических элементов (МГЭ) является ключом к пониманию молекулярных механизмов ответа клеток на стрессовые воздействия, включая вирусные инфекции. МГЭ играют важную роль в регуляции экспрессии генов и формировании структуры хроматина, что делает их исследование важным для понимания патогенеза инфекционных заболеваний.

В данной работе проанализированы данные scRNA-seq моноклеарных клеток периферической крови пациентов с различными вирусными инфекциями (COVID-19, грипп, ВИЧ) и здоровых доноров. Показано, что экспрессия LINE и SINE зависит от типа клеток, нозологии и тяжести заболевания. В клетках T CD8⁺ Naïve группы ВИЧ зафиксировано повышение экспрессии ряда SINE при стабильной экспрессии LINE, что связано с рекрутированием клеточной РНК-полимеразы II через

взаимодействие с вирусным белком Tat. У пациентов с тяжелым течением COVID-19 (штамм Дельта) снижение числа активных LINE коррелирует с увеличением экспрессии нескольких SINE, что указывает на множественную регуляцию SINE одной LINE-полимеразой. В клетках Mon CD16+ выявлено снижение экспрессии одного LINE и трех SINE, что подтверждает участие данного LINE в регуляции связанных SINE.

Результаты исследования демонстрируют значительные изменения активности LINE и SINE в иммунных клетках при различных инфекциях. Эти данные подчеркивают роль МГЭ в адаптации клеток к патогенным воздействиям и открывают новые перспективы для изучения эволюционных механизмов регуляции генома в условиях стресса.

Геномное разнообразие прямых повторов в митохондриальной ДНК летучих мышей: сравнительный анализ и эволюционные перспективы

Федяева И.Р.

*Балтийский федеральный университет имени И.Канта (236041,
Россия, Калининград, ул. Александра Невского, 14),
irina.fedyeva.03@mail.ru*

Летучие мыши представляют собой разнообразную группу млекопитающих, играющих ключевую роль в экосистемах. Изучение их геномов необходимо для понимания механизмов адаптации и эволюции. В данной работе исследовано разнообразие и эволюция прямых повторов (DR) – повторяющихся участков ДНК – в геномах различных видов летучих мышей.

Целью исследования являлось выявление закономерностей в организации DR и их связи с филогенетическими отношениями, продолжительностью жизни и метаболизмом. Для достижения этой цели были проанализированы последовательности митохондриальной ДНК 77 видов летучих мышей, полученные из GenBank. Поиск и анализ DR проводился с использованием биоинформатических инструментов, включая Biopython.

Результаты показали значительные различия в длине и

распределении DR между разными видами летучих мышей. Длина DR варьировала от 10 п.н. до 428 п.н.. Филогенетический анализ выявил, что геномные особенности, в том числе организация DR, связаны с филогенетическими кладами, кроме того наблюдается корреляция количества прямых повторов и типа питания. Наименьшее количество прямых повторов наблюдается в митохондриальной ДНК фрукто- и нектароядных летучих мышей, при этом их продолжительность жизни является одной из самых высоких среди собранной когорты видов, что предполагает отсутствие у них отрицательной корреляции количества прямых повторов и общей их мутагенности и продолжительности жизни, что противоречит данным, полученным Хайдаковым и др. в своей статье [1]. Поскольку, предварительные данные указывают на возможную связь между длиной DR и продолжительностью жизни летучих мышей, что требует дальнейшего изучения.

Полученные данные позволяют лучше понять геномную архитектуру летучих мышей и роль DR в их эволюции. Дальнейшие исследования будут направлены на изучение механизмов влияния DR на экспрессию генов и стабильность генома.

1. Khaidakov M, Siegel ER, Shmookler Reis RJ. Direct repeats in mitochondrial DNA and mammalian lifespan. *Mech Ageing Dev.* 2006 Oct;127(10):808-12. doi: 10.1016/j.mad.2006.07.008. Epub 2006 Sep 7. PMID: 16956646.
2. Lagunas-Rangel FA. Why do bats live so long?-Possible molecular mechanisms. *Biogerontology.* 2020 Feb;21(1):1-11. doi: 10.1007/s10522-019-09840-3. Epub 2019 Oct 10. PMID: 31602545.
3. Wilkinson GS, Chapman AM. Length and sequence variation in evening bat D-loop mtDNA. *Genetics.* 1991 Jul;128(3):607-17. doi: 10.1093/genetics/128.3.607. PMID: 1874418; PMCID: PMC1204534.
4. Wilkinson GS, Mayer F, Kerth G, Petri B. Evolution of repeated sequence arrays in the D-loop region of bat mitochondrial DNA. *Genetics.* 1997 Jul;146(3):1035-48. doi: 10.1093/genetics/146.3.1035. PMID: 9215906; PMCID: PMC1208033.

Новый алгоритм поиска несовершенных палиндромов в ДНК

Г. А. Хазиев, О. А. Зверков, Л. И. Рубанов, А. В. Селиверстов,
В. А. Любецкий

*Институт проблем передачи информации имени А.А. Харкевича
Российской академии наук, 127051, Москва, Большой Каретный
переулок, д.19 стр.1, georgy.haziev@yandex.ru*

Нуклеотидная последовательность называется палиндромом, если она совпадает с комплементарной последовательностью, читаемой в обратном направлении (reverse complement). Последовательность, отличающаяся от (совершенного) палиндрома, называется несовершенным палиндромом. Несовершенные палиндромы, близкие к совершенным, могут быть сайтами кооперативного связывания с ДНК транскрипционных факторов, образующих гомодимеры (Datta and Rister, 2022), а также образовывать шпильки на РНК.

Редакционное расстояние или расстояние Левенштейна между двумя последовательностями вычисляется за квадратичное время (Янковский, 2024).

Нами предложен новый алгоритм, который получает на вход нуклеотидную последовательность x и выдаёт список длин префиксов w , для которых редакционное расстояние $\text{dist}(x, wc(w))$ между исходной последовательностью x и оптимальным палиндромом с таким префиксом w достигает минимума. Каждый префикс w определяет разбиение $x=wz$. Также алгоритм выдаёт минимальное значение редакционного расстояния. При этом надо выбрать палиндром из ансамбля палиндромов, определяемых входной последовательностью. Префиксы палиндромов из этого ансамбля совпадают с префиксами в x . При этом последовательность x может не содержать искомый палиндром целиком. В таком случае она содержит лишь несовершенный палиндром. Но в x содержится половина от искомого палиндрома. Оптимальные разбиения и редакционное расстояние вычисляются за квадратичное время. Вычислительная сложность алгоритма ниже, чем в случае полного перебора вариантов границы инвертированного участка. Для сравнения последовательностей

разной длины удобно использовать функцию $\text{imp}(x)$, значение которой равно отношению длины оптимального палиндрома к длине исходной последовательности. Значение этой функции говорит, насколько последовательность отличается от совершенного палиндрома, и равна нулю для совершенных палиндромов.

Алгоритм реализован на языке Python. Программа доступна по адресу <http://lab6.iitp.ru/~pali>, где также приведен пример использования. Использование библиотеки Numba позволяет значительно уменьшить время работы программы.

Эмпирическая оценка медианы значений $\text{imp}(x)$ для случайных нуклеотидных последовательностей из нескольких сотен нуклеотидов близка к значению 0.26. Но эта медиана немонотонно зависит от длины и заметно выше для коротких последовательностей.

Также разработан второй алгоритм для выделения из последовательности x более короткого несовершенного палиндрома y с меньшим значением $\text{imp}(y) < \text{imp}(x)$, получаемого удалением из исходной последовательности x префикса или суффикса, если это возможно. Эта операция называется триммингом. Основной идеей наших алгоритмов является проверка, верно ли, что одна из оптимальных длин префикса w последовательности x сильно отличается от половины длины x . Если так, то программа усекает x на разницу между длиной w и половиной длины x .

Кроме того, был разработан третий – альтернативный – алгоритм тримминга. Он выполняет заданное число последовательных усечений исходной последовательности x . Первое усечение производится на число нуклеотидов p , заданное как доля от длины x . После каждого усечения выполняется сравнение $\text{imp}(x)$ и $\text{imp}(x')$, где через x' обозначен результат усечения последовательности x на p' . Если выполнено неравенство $\text{imp}(x) > \text{imp}(x')$, то вычисляется следующее усечение уже для x' на p' . Иначе, на следующем шаге вычисляется усечение последовательности x на целую часть от числа $p'/2$ нуклеотидов. Эти алгоритмы также реализованы на языке Python. Программа доступна по адресу <http://lab6.iitp.ru/~trimmers>, где приведены краткое описание и пример использования.

Наш метод применён к последовательностям высококонсервативных элементов (ВКЭ) в некодирующих областях геномов из четырёх млекопитающих (*Homo sapiens*, *Mus musculus*, *Sus scrofa*, *Macaque fascicularis*). Эти ВКЭ были найдены методом, который служит модификацией ранее разработанного авторами алгоритма (Rubanov *et al.*, 2016). Поиск ВКЭ выполнен в синтеничных блоках, расположенных у человека на хромосомах 1–3, в которых предварительно были замаскированы точкой все кодирующие последовательности (CDS) и все участки низкой сложности. К началу и к концу каждого блока добавлялся прилегающий участок генома длиной в 1 Мб.

В результате нами выделено много несовершенных палиндромов с малым значением $\text{imp}(x)$. В частности, многие из этих палиндромов расположены в 5'-некодирующих областях генов.

Опишем более подробно ВКЭ, найденные на начальном локусе первой хромосомы человека и в синтеничных локусах у других рассмотренных видов. Для тех ВКЭ из этих локусов, в составе которых после тримминга найден участок с $\text{imp}(y)$ не выше 0.2, медиана разностей $\text{imp}(x) - \text{imp}(y)$ составила 0.049. Лишь для 20 последовательностей алгоритм не нашёл лучшей последовательности меньшей длины: $\text{imp}(x) = \text{imp}(y)$. Для одного из ВКЭ в результате тримминга был выделен участок с минимальным значением $\text{imp}(y) = 0.128$, хотя у исходной последовательности $\text{imp}(x) = 0.209$. Наибольшее уменьшение величины $\text{imp}(x)$ составляет 0.232, что лишь немного меньше медианы. В этом случае у исходной последовательности $\text{imp}(x) = 0.406$, а после тримминга $\text{imp}(y) = 0.174$.

Отметим, что сами последовательности ВКЭ обычно мало похожи на палиндромы, поскольку обычно обладают высоким значением $\text{imp}(x)$, близким к медиане. Однако предлагаемый алгоритм выполняет усечение префиксов и суффиксов, позволяя выделить несовершенные палиндромы с низким значением $\text{imp}(x)$, но имеющие сравнительно большую длину.

Выполнен поиск тех несовершенных палиндромов в составе ВКЭ, которые попадают в известные промоторные области. Работа выполнена в рамках государственного задания ИППИ РАН, утвержденного Минобрнауки России.

1. В.О. Янковский (2024) Группы изометрий формальных языков относительно обобщенных метрик Левенштейна, *Математические заметки*, **116**(2): 306–315.
2. R.R. Datta, J. Rister (2022) The power of the (imperfect) palindrome: sequence-specific roles of palindromic motifs in gene regulation, *Bioessays*, **44**(4): e2100191.
3. L.I. Rubanov, A.V. Seliverstov, O.A. Zverkov, V.A. Lyubetsky (2016) A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata, *BMC Bioinformatics*, **17**: 385.

Гибридное определение родительского происхождения гаплотипов

Д. А. Хлебников^{1,2}, А. А. Миронов^{1,2}, Alexander Gimelbrant³

¹*Факультет биоинженерии и биоинформатики, Московский Государственный Университет им. М.В. Ломоносова, Москва, Россия*

²*Институт Общей Генетики им. Н.И. Вавилова Российской Академии Наук, Москва, Россия*
dkhlebn@gmail.com

³*Altius Institute for Biomedical Sciences, Seattle, WA*

Гаплотипирование неродственных индивидов базируется на моделировании частот гаплотипов в популяции. Одной из ключевых задач, возникающих в этом контексте, является реконструкция материнских и отцовских гаплотипов в геноме потомка. При наличии данных о тройке (ребёнок и оба родителя), а в некоторых случаях — о двойке (ребёнок и один из родителей), становится возможным определение происхождения гаплотипов каждой хромосомы ребёнка от матери или отца. Однако этот подход сопряжён с существенными финансовыми затратами на секвенирование всех участников трио.

Несмотря на постоянное снижение стоимости секвенирования, сохраняется актуальность разработки более

экономически эффективных алгоритмических методов для решения этой задачи.

В настоящем исследовании мы применили существующий [1] подход определения родительского происхождения гаплотипов для данных метилирования двух доноров, полученных методом Fiber-Seq, для анализа импринтированных дифференциально метилированных регионов хромосом Т-клеток.

Мы изучили возможность кластеризации гаплотипов в две группы ассоциации – отцовского и материнского происхождения – без использования эпигенетических данных. Для этого гаплотипы были вложены в трансформированные с помощью PLS-регрессии пространства данных 1000 Genomes на основании наличия в них полиморфизмов. Затем была применена процедура двустадийной классификации для каждого из гаплотипов: определялась его интерконтинентальная группа, а затем определялось расстояние до интраконтинентальных популяционных групп в пределах одной интерконтинентальной. Такой подход применялся ранее для определения популяционного происхождения индивидов [2-3], но для определения родительского происхождения гаплотипов применяется впервые. Кроме того, нами разработан алгоритм оценки статистической значимости распределения гаплотипов по материнскому или отцовскому геному, что позволяет повысить точность и надёжность результатов.

1. Akbari et al. (2022) Parent-of-origin detection and chromosome-scale haplotyping using long-read DNA methylation sequencing and Strand-seq, *Cell Genom.*, **3**(1):100233.
2. Sun et al. (2019) Application of partial least squares in exploring the genome selection signatures between populations, *Heredity*, **122**:288–293.
3. Alladio et al. (2022) Multivariate statistical approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field, *Sci Rep* **12**:8974.

Методология обучения нейронных сетей для задачи восстановления эффективного размера популяции

Е.В. Хомутов

НИУ Высшая Школа Экономики, г. Москва, ул. Мясницкая, д. 20
ehomutov@hse.ru

Геномные данные представляют собой ценный ресурс для изучения истории популяций. Генеалогии геномных образцов являются выборкой из популяционной истории. В свою очередь для диплоидных организмов (например, человека) в следствие рекомбинаций генеалогии на разных участках генома различаются. Таким образом, из распределения генеалогии можно достаточно точно оценивать, в частности, изменения эффективного размера популяции.

Так, для одного диплоидного образца генеалогия определяется лишь временем до ближайшего общего предка (LCA). Это наблюдение используется в одном из основных методов в арсенале популяционной генетики - методе PSMC [1], основанном на скрытой марковской модели и предсказывающем изменения эффективного размера популяции. В нашей работе мы предлагаем новый метод, использующий глубокое обучение для определения локального времени LCA для полногеномных последовательностей, что является важным методологическим вкладом.

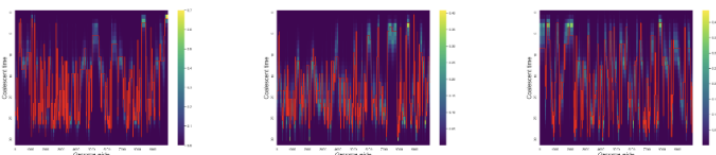
Для проведения экспериментов мы сделали модуль `deergene`, предоставляющий собой программную обертку над генератором геномных данных `msprime` [2]. С их помощью мы генерируем синтетическую обучающую выборку, так как мы не можем использовать реальный геном, ввиду отсутствия знания момента коалесценции.

В терминах машинного обучения, перед нами стоит задача регрессии с учителем - у нас есть обучающая выборка из наблюдаемых геномов $\{G\}$, на ней мы знаем время коалесценции $T_{\text{coal}}^{(j)}(i)$ в геноме $G^{(j)}$ в локусе i . С другой стороны, нас интересует не столько точное время коалесценции, как его порядок - поэтому, для упрощения задачи, мы можем разбить прошлое на временные интервалы, пронумеровать их и решать

задачу классификации. Таким образом, мы приходим к задаче классификации с учителем: по известным $\{G\}$ и $T_{\{coal\}}^{\{(j)\}}(i)$ мы хотим восстановить функционал $F(g) \rightarrow T_{\{coal\}}(i)$.

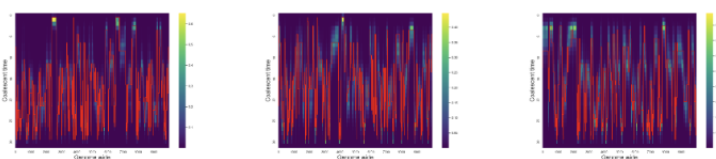
В рамках исследования было рассмотрено три популярные архитектуры нейронных сетей - LSTM, Reformer и GRU. В качестве loss-функций рассматривались KL-divergence и cross-binary entropy. Первоначально для обучения нейронной сети мы брали геномы, сгенерированные при помощи компьютерного моделирования в одной фиксированной демографии, но при таком подходе оказалось, что модель не получает обобщающую способность и плохо ведёт себя на демографиях, которые не использовались при обучении. При этом при обучении на случайных неповторяющихся демографиях тоже не получается хорошей обобщающей способности. Было предложено решение в виде использования фиксированного набора предвыбранных демографий для обучения нейронной сети - при таком подходе модель может предсказывать довольно хорошо и на новых для неё демографиях.

Далее на графиках для простоты интерпретации представлены результаты работы GRU модели с KL-divergence как loss функции для предсказания локального и глобального распределения времени до последнего общего предка (глобальное время суммированием вероятностей вдоль генома и нормировкой к 1).



(a) GRU обученный на константной демографии
(b) GRU обученный на демографии с бутылочным горлышком
(c) GRU обученный на фиксированном наборе демографий

Рис. 1: Локальные предсказания модели GRU на геноме из константной демографии. Красная линия - реальный интервал коалесценции, фон - тепловая карта предсказаний нейронной сети



(a) GRU обученный на константной демографии
(b) GRU обученный на демографии с бутылочным горлышком
(c) GRU обученный на фиксированном наборе демографий

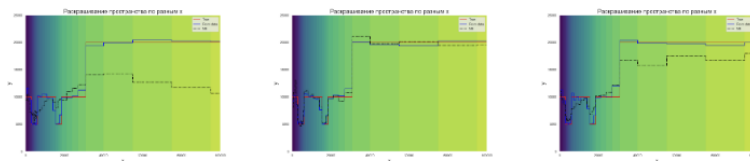
Рис. 2: Локальные предсказания модели GRU на геноме из демографии с бутылочным горлышком. Красная линия - реальный интервал коалесценции, фон - тепловая карта предсказаний нейронной сети

На графиках 1 и 2 видно, что главная гипотеза подтверждается и нейронные сети могут быть использованы для предсказания локального времени до последнего общего предка в каждом локусе диплоидного генома. При этом, при суммировании вдоль генома мы получаем следующее графики:



(a) GRU обученный на константной демографии
(b) GRU обученный на демографии с бутылочным горлышком
(c) GRU обученный на фиксированном наборе демографий

Рис. 3: Глобальные предсказания модели GRU на геноме из константной демографии. Красная линия - реальная демография, синяя - восстановленная из генома, черная прерывистая - предсказание нейронной сети. Фон разделен на цвета, каждый из которых показывает интервалы разбиения (классы в теминах машинного обучения)



(a) GRU обученный на константной демографии
(b) GRU обученный на демографии с бутылочным горлышком
(c) GRU обученный на фиксированном наборе демографий

Рис. 4: Глобальные предсказания модели GRU на геноме из демографии с бутылочным горлышком. Красная линия - реальная демография, синяя - восстановленная из генома, черная прерывистая - предсказание нейронной сети. Фон разделен на цвета, каждый из которых показывает интервалы разбиения (классы в теминах машинного обучения)

На графиках 3 и 4 видно, что после интегрирования модели, видевшие при обучении только одну демографию выдают ошибку, которая связана скорее всего с переобучением. При этом модель,

которая во время обучения видела больше демографий, в среднем справляется хорошо.

Тезисно, во время исследовательской работы получилось прийти к следующим выводам:

- Доказана возможность предсказания локального времени коалесценции с помощью нейронных сетей, а так же дальнейшее предсказание эффективного размера популяции по этим данным;
- GRU самая точная, но самая медленная архитектура;
- Reformer очень быстрый, но менее точный;
- Важнее на чем обучается модель, чем архитектура;

Выражаю искреннюю признательность и благодарность моему научному руководителю Щуру Владимиру Львовичу за помощь на всех этапах проведения исследования.

1. H. Li ,R. Durbin (2011) Inference of human population history from individual whole genome sequences, \textit{Journal, Nature} 475, 2011. P. 493-496.
2. F. Baumdicker et al. (2022) Efficient ancestry and mutation simulation, \textit{Journal, Genetics} Vol. 220, Issue 3 (Eds.), 13–16 (Publisher).

DrugForm: платформа драг-дизайна

И.С. Хохлов¹, А.С. Тащилова¹, Н.А. Бугаев-Макаровский¹, О.В. Глушкова¹, В.С. Юдин¹, А.А. Кескинов¹, С.М. Юдин¹,
Д.В. Светличный¹, В.И. Скворцова²

¹Федеральное государственное бюджетное учреждение «Центр стратегического планирования и управления медико-биологическими рисками здоровью» Федерального медико-биологического агентства (ФГБУ «ЦСП» ФМБА России); 119121, Россия, г. Москва, ул. Погодинская, д. 10, стр. 1, ihohlov@cspfmba.ru

²Федеральное медико-биологическое агентство (ФМБА России);
123182, Россия, г. Москва, Волоколамское шоссе, д. 30.

Драг-дизайн, или компьютерный поиск новых лекарств, требует использования множества различных вычислительных методов и моделей машинного обучения. Обычно исследователи занимаются разработкой отдельных моделей или алгоритмов, выполняющих узкую задачу, например, моделей аффинности, токсичности, или генеративных моделей. В данной работе разработана серия отдельных QSAR (Quantitative Structure–Activity Relationship) моделей, а также программная инфраструктура для их объединения с целью решения задачи генерации структур малых молекул с заданными свойствами.

Разработанная нами платформа для дизайна малых молекул, DrugForm, позволяет рассчитывать ряд физико-химических свойств малой молекулы - лиганда, а также генерировать новые молекулы с необходимыми целевыми значениями ключевых параметров, в том числе аффинности к белку-мишени.

На данный момент платформа DrugForm включает десятки моделей и методов, из которых наиболее значимыми являются модели: аффинности, токсичности, склонности к офф-таргет (нецелевым) эффектам, структурной близости, растворимости, а также методы нахождения подструктур, лекарствовоподобности, синтетической доступности. Модель аффинности позволяет оценивать значения констант аффинности лигандов к произвольным белкам. На ее основе построена модель офф-таргет эффектов, которая возвращает оценки количества нецелевых связываний лиганда с другими белками. Разработанная модель токсичности опирается на концепцию введенной обобщенной токсичности, полученную на основе различных способов введения вещества, - интегративной оценке, полученной при анализе фармакологических данных из различных модельных организмов, и представляет меру общей угрозы жизни и здоровью человека.

Для задач генерации молекулярных структур с заданными свойствами разработана генеративная модель, построенная по принципу обучения с подкреплением. Программный комплекс платформы DrugForm позволяет пользователю выполнить генерацию, опираясь на комбинацию требуемых свойств и

характеристик молекул. Генеративная модель автоматически обращается к необходимым QSAR-моделям и пытается подобрать такие молекулы, которые бы максимально удовлетворяли всем требованиям, поставленным пользователем.

С технической стороны платформа представляет собой распределенную микросервисную архитектуру, позволяющую запускать и одновременно использовать множество моделей, а также стандартным способом добавлять в систему новые модели и методы. Пользовательский веб-интерфейс платформы позволяет запускать генерации и одиночные расчеты, а также визуально оценивать результаты.

С помощью данной платформы был решен ряд практических задач по генерации молекул к мелатониновым, орексиновым и глутаматным рецепторам, и представлены новые перспективные соединения, имеющие высокий терапевтический потенциал.

Evolution of mutually exclusive exons and RNA secondary structures in paralogs

E. Chernyavskaya¹, S.A. Spirin², D. D. Pervouchine¹

¹*Skolkovo institute of science and technology, Bolshoy Boulevard, 30, bld. 1, Moscow 121205, Russia, ekaterina.chernyavskaya@skoltech.ru*

²*Belozersky Institute, Lomonosov Moscow State University, Leninskie Gory, 1, b.40, Moscow, 119234, Russia sergeispirin@gmail.com*

Alternative splicing (AS) plays a crucial role in expanding protein diversity. The most common AS event types include exon skipping, intron retention, and alternative selection of 3' or 5' splice sites. Mutually exclusive exons (MXEs) represent a less frequent but functionally distinct category. MXE splicing is a strictly coordinated mechanism, allowing the inclusion of only one exon from a set of alternatives into the final transcript while others are excluded. A key regulatory feature of MXE splicing often involves RNA secondary structures. A well-characterized example is the Dscam gene in

Drosophila species, the selection of exons in their exon 6 cluster is governed by a set of conserved regulatory sequences forming RNA secondary structure where one docking site could base pair only to one selector sequence at a time. These structures allow only one of MXEs to be included in the transcript, highlighting the critical role of RNA folding in splicing regulation. It has been shown that pre-mRNAs of MXE containing genes tend to form competing secondary structures, which are likely to emerge through the duplication of both exons and their flanking intronic regions with stem-loop structures. [1].

Given that many genes with MXEs belong to paralogous families, it is of particular interest to investigate whether RNA structures controlling their splicing have been conserved or diverged in evolution. Understanding these regulatory patterns in paralogs could advance our knowledge of post-transcriptional regulation, shed light on the evolutionary plasticity of splicing control in gene families with critical biological roles, and potentially provide new therapeutic strategies targeting aberrant splicing in diseases.

In this study, we investigated several paralogous gene families containing mutually exclusive exons, including tropomyosin, fibroblast growth factor receptor (FGFR), actinin, and sodium voltage-dependent channel alpha-subunit (SCN α) families. To trace the evolutionary origins of both gene and exon duplications and their splicing regulatory mechanisms, we compiled comprehensive datasets of homologous genes from over 15 diverse organisms, ranging from vertebrates to mollusks, using publicly available genomic databases.

Comparative genomic approach enabled us to identify previously unannotated exon duplications and conserved complementary intronic sequences potentially involved in MXE splicing regulation across species. For the FGFR gene family, our analysis revealed that the basal chordate *Ciona intestinalis* (possessing only a single FGFR gene) represents the most ancient extant organism containing duplicated MXEs. Through two rounds of whole-genome duplication events, this single gene gave rise to four paralogs that currently exist in humans. Notably, we confirmed a pair of complementary intronic sequences in *Ciona* showing significant evolutionary conservation, likely responsible for MXE splicing regulation. This finding gains particular significance as homologous RNA secondary structures have been experimentally confirmed to regulate alternative splicing in human FGFR2 [2]. These

results demonstrate that for the FGFR family, the RNA secondary structure controlling MXE splicing emerged almost concurrently with the initial exon duplication event and has been conserved throughout vertebrate evolution.

In contrast, we observed a different evolutionary pattern in the voltage-gated sodium channel (SCN α) gene family. The human genome contains ten SCN α paralogs originating through both whole-genome and individual gene duplication events. Our analysis revealed that MXE clusters occurred independently in specific paralogs after their divergence from a common ancestor. Interestingly, we identified a conserved RNA secondary structure in tandem-duplicated genes (SCN1A-SCN3A and SCN9A) that can be traced back to the elephant shark (*Callorhinchus milii*), which possesses only a single ancestral homolog of these genes. This finding suggests that the regulatory mechanism for MXE splicing in these genes was established prior to their tandem duplication events.

These comparative genomic studies reveal distinct evolutionary trajectories of RNA secondary structures regulating MXE splicing in different gene families. The observed patterns – including both deeply conserved mechanisms (as in FGFR) and lineage-specific acquisitions (as in SCN α paralogs) – confirm either extreme divergence from ancestral structures, differential loss in paralogs, or independent convergent evolution, likely driven by strong selective pressure to maintain precise splicing regulation in these functionally critical genes [3]. These evolutionary insights highlight the importance of experimentally validating predicted regulatory structures. Such studies would not only advance our understanding of splicing regulation but could also suggest therapeutic strategies targeting these RNA structures to correct aberrant splicing in diseases associated with these gene families.

1. T. Ivanov and D. Pervouchine (2018) An Evolutionary Mechanism for the Generation of Competing RNA Structures Associated with Mutually Exclusive Exons, *Genes*, vol. 9, no. 7, p. 356.
2. S. J. Muh, R. H. Hovhannisyan, and R. P. Carstens (2002) A Non-sequence-specific Double-stranded RNA Structural Element Regulates Splicing of Two Mutually Exclusive Exons

- of Fibroblast Growth Factor Receptor 2 (FGFR2), *Journal of Biological Chemistry*, vol. 277, no. 51, pp. 50143–50154.
3. M. Petrova, S. Margasyuk, M. Vorobeva, D. Skvortsov, Olga A Dontsova, and D. D. Pervouchine (2018) BRD2 and BRD3 genes independently evolved RNA structures to control unproductive splicing, *NAR Genomics and Bioinformatics*, vol. 6, no. 1.

Type I diabetes-induced hypothalamic changes at a single-cell resolution

Evgeniia Chikina¹, Victoria Melnikova¹, Roman A. Romanov²

¹*Koltzov Institute of Developmental Biology of Russian Academy of Sciences, 26 Vavilov Street, Moscow, 119334, Russia, chikina.evgeniia@gmail.com*

²*Center for Brain Research, Medical University of Vienna, 4 Spitalgasse, Vienna, 1090, Austria, Roman.Romanov@meduniwien.ac.at*

Type 1 diabetes (T1D) is a chronic autoimmune disease characterized by the destruction of pancreatic β -cells, leading to a deficiency in insulin production and chronic hyperglycemia. With over 9 million people affected worldwide as of 2017 [1], the disease typically begins in childhood or adolescence and requires lifelong management. Its early onset, long-term burden, and impact on multiple organ systems make T1D a major public health concern.

Emerging technologies such as single-cell RNA sequencing (scRNA-seq) have revolutionized our ability to dissect complex tissues at the cellular and molecular levels. This is particularly relevant for studying the central nervous system, where cellular heterogeneity and intercellular communication play crucial roles in metabolic regulation. The hypothalamus, a key integrator of energy balance and glucose homeostasis, is especially affected by altered glucose levels in diabetes.

In our study, we performed single-cell and single-nuclei RNA sequencing of hypothalamic cells from mice with streptozotocin-induced

type 1 diabetes and compared them to healthy controls. Two generated datasets allowed us to identify changes in both neuronal and glial cell populations. All samples were preprocessed separately with manually adjusted threshold for cells filtering including removal of low quality cells and potential doublets. Cell type annotation was achieved through both known marker genes from the literature [2] and differentially expressed genes identified in our dataset.

Analyzing changes specific to the cell type we first of all focused on neuronal subset. We were able to identify significant expressional shifts in GABA- and glutamatergic neuron especially connected with their metabolic status. Increasing evidence suggests that not only neurons, but also their surrounding glial environment, are responsive to metabolic cues. For example, insulin receptors on astrocytes have been shown to modulate Pomc-neuron activity in response to elevated glucose levels [3]. Our analysis revealed a significant increase in T-cell abundance in diabetic samples. Further characterization, including expression profiling and comparison with published datasets, indicated that these are activated CD8⁺ T cells with strong effector functions. This infiltration may be linked to alterations in the brain vasculature, as suggested by changes observed in vascular-associated cell types.

Beyond compositional differences, we also identified changes in intercellular communication using the CellChat tool [4]. These included alterations in neuron–astrocyte signaling. For instance, an upregulation of ErbB signaling may reflect astrocyte activation and proliferation, while a reduction in Psap–Gpr3711 interaction points to diminished neuro- and glioprotective signaling.

In summary, our study highlights significant cell type-specific and intercellular changes in the hypothalamus in response to type 1 diabetes. Focusing on neuronal alterations we also demonstrated significant alterations in glial cells, important components of the neuronal microenvironment, suggesting that metabolic dysfunction in diabetes extends beyond neurons to the supportive and regulatory cell types that maintain brain homeostasis.

The research work is supported by the grant № 075-15-2021-1075.

1. WHO: <https://www.who.int/ru/news-room/fact-sheets/detail/diabetes>

2. Zeisel A, et al. (2018) Molecular Architecture of the Mouse Nervous System, *Cell*, 174(4):999-1014.
3. García-Cáceres C, et al. (2016) Astrocytic Insulin Signaling Couples Brain Glucose Uptake with Nutrient Availability, *Cell*, 166(4): 867-880.
4. Jin S, et al. (2021) Inference and analysis of cell-cell communication using CellChat, *Nat Commun.*, 12 (1088).

Влияние замен цистеина на аланин в большом внеклеточном домене Na-зависимого фосфатного транспортера NaPi2b на его пространственную структуру

А.Р. Шайдуллина, Н.И. Акберова

Кафедра биохимии, биотехнологии и фармакологии, КФУ, Казань, Российская Федерация, seuta21@gmail.com

Натрий-зависимый фосфатный транспортер NaPi2b (SLC34A2) сверхэкспрессируется в ряде злокачественных новообразований и рассматривается как потенциальная мишень для таргетной терапии. Его внеклеточный домен (ВКД) содержит иммуногенный эпитоп, распознаваемый моноклональными антителами. В ВКД NaPi2b в положениях 303, 322, 328 и 350 присутствуют остатки цистеина, которые могут участвовать в образовании дисульфидных связей, положение и количество которых в большой мере должны определять структуру и возможные конформации ВКД и влиять в т.ч. на доступность эпитопа для терапевтических антител. В настоящем исследовании проанализировано влияние замен остатков цистеина на аланин в ВКД NaPi2b с использованием методов равновесной молекулярной динамики.

Пространственная структура NaPi2b не определена экспериментально, в работе использовали предсказанную AlphaFold модель структуры NaPi2b (AF-O95436-F1). Замены остатков цистеина на аланин проводили в Swiss-PdbViewer, были подготовлены 4 структуры с одиночными заменами – С303А,

C322A, C328A, C350A, и 3 структуры с двумя заменами - C303A_C322A, C303A_C328A, C322A_C328A. С помощью инструментов CHARMM-GUI все структуры были помещены в липидный бислой POPC и в кубическую водную ячейку $120 \text{ \AA} \times 112 \text{ \AA} \times 172 \text{ \AA}$ с добавлением ионов Na и Cl.

Подготовительные этапы минимизации, нагрева и эквilibрации систем и симуляцию молекулярной динамики длиной 30 нс для каждой системы выполнялись в среде NAMD с силовым полем CHARMM36. С помощью VMD и bio3d R проводили анализ среднеквадратичного отклонения атомов (RMSD), анализ среднеквадратичной флуктуации атомов (RMSF), анализ площади поверхности, доступной для растворителя (SASA), радиус гирации (Rg), энергии электростатических и ван-дер-ваальсовых взаимодействий.

Сравнительный анализ полученных МД-траекторий показал, замены остатков цистеина на аланин изменяют структурные характеристики ВКД NaPi2b. Максимальные отклонения RMSD ВКД и увеличение SASA зафиксированы для C350A и C328A (RMSD: $9.74 \pm 1.74 \text{ \AA}$ для C350A, $7.51 \pm 0.68 \text{ \AA}$ для C328A; SASA: $1100 \text{ \AA}^2$ для обеих систем), (Rg: $33\text{--}34.5 \text{ \AA}$ в системах C350A и C303_322A против $36 \text{ \AA}$ у WT). Наиболее стабильным эпитоп остаётся при заменах C303A и C322A. Во всех мутантных формах наблюдается снижение компактности ВКД ($33\text{--}35.5 \text{ \AA}$ против $36\text{--}36.5 \text{ \AA}$ у WT), а также увеличение площади поверхности, доступной для растворителя (медианное SASA: $1000 \text{ \AA}^2$ для C303A_322A, $1100 \text{ \AA}^2$ для C328A и C350A, $950 \text{ \AA}^2$ у WT). При мутации C328A отмечено усиление электростатических взаимодействий в ВКД (-1000 кДж/моль), у C350A — повышение энергии ван-дер-ваальсовых взаимодействий в районе эпитопа (различия статистически значимы, $p = 0.0002$, диапазон -300 до -400 кДж/моль).

Полученные данные указывают на ключевую роль дисульфидных связей в стабилизации внеклеточного домена NaPi2b и поддержании его иммуногенных свойств. Проведенный анализ может быть использован для оценки конформационной доступности эпитопа и в перспективе — при разработке моноклональных антител.

**Сравнение вероятности делеций мтДНК в регионах,
фланкированных различными комбинациями прямых и
инвертированных повторов**

В.А. Шаманский, К.Ю. Попадьин

*Центр геномных исследований БФУ им. И. Канта,
v.a.shamanskiy@gmail.com*

Накопление крупномасштабных соматических митохондриальных делеций связано как со здоровым старением, так и с возрастными митохондриальными энцефаломиопатиями. Делеции часто фланкируются прямыми нуклеотидными повторами, и было выдвинуто несколько гипотез для объяснения образования делеций с участием прямых повторов. Мы предполагаем, что помимо прямых повторов инвертированные повторы могут играть важную роль в структурировании мтДНК.

Используя коллекцию делеций мтДНК человека и глобальные и локальные свойства мтДНК (распределение прямых и инвертированных повторов), мы пересматриваем риски соматической делеции мтДНК.

Сравнение вероятности делеций мтДНК в регионах, фланкированных различными комбинациями прямых и инвертированных повторов, раскрывает, что в дополнение к прямым повторам, которые, как известно, влияют на делеции, существует сильное влияние вторичной структуры одноцепочечной мтДНК во время репликации. Вторичная структура во время репликации формируется инвертированными повторами, которые могут образовывать стебли и сокращать эффективное расстояние между двумя прямыми повторами, тем самым увеличивая вероятность делеций.

Наши результаты подтверждают недавнее открытие, что делеции мтДНК могут возникать из-за проскальзывания репликации. Понимание механизмов формирования делеций позволяет предсказать генетические риски делеций мтДНК для различных гаплогрупп человека - первыми примерами являются

гаплогруппы D4a и D5a, характеризующиеся повышенной продолжительностью жизни.

Evolutionary Analysis of Bacterial Homologues of the Sirtuin Protein Family

Igor A. Shokodko¹, Mikhail S. Gelfand²

¹*HSE, Pokrovskiy Boulevard, 11b4, Moscow, nero.fact@gmail.com*

²*Skolkovo Institute of Science and Technology, Moscow*

Acetylation is one of the best characterized histone modifications. Histone acetyltransferase and histone deacetylase enzymes control the levels of histone acetylation, making genes, respectively, more or less accessible for the transcription machinery and hence, influencing their expression [1]. The sirtuin family of proteins (sirtuins) is a family of NAD⁺-dependent protein- deacetylases. Diverse members of this family regulate key pathways in all domains of life: in bacteria, archaea, eukaryotes, and even viruses [2]. The NAD⁺-dependent nature of sirtuins makes these proteins a link between nutrition and lifespan of an organism. In this regard, a huge number of studies have been devoted to mammalian homologues (SIRT1-7), describing their structure and functions in a wide variety of processes. At present, information on the functions of bacterial homologues of the sirtuin protein family is very limited to the results of in vivo and in vitro experiments. This is the reason there is data on the functions of only cultured model organisms (*Salmonella enterica*, *Bacillus subtilis*, *Escherichia coli*, etc.). In addition to regulating expression, sirtuins in these bacterial species are actively involved in bacterial virulence and various types of metabolism: degradation of aromatic compounds, metabolism of propionates and fatty acids, chemotaxis, the tricarboxylic acid (TCA) cycle, and nitrogen metabolism [3]. Methods of comparative genomics and phylogenetic analysis can significantly expand our understanding of the functions and evolution of sirtuins. The aim of the present research is to describe the phylogenetics and determine the functions of bacterial homologues of the sirtuin protein family.

Within the complete bacterial proteomes, 8592 bacterial sirtuins were found. Due to the low similarity of the sequences of the proteins found within the bacterial phylum, it was not efficient to construct multiple alignments to construct phylogenetic trees - most sequences in the alignment had near 80% gaps. Because of this, a method was proposed to separate proteins into clusters by pairwise alignment using the MMseqs2 and subsequent clustering of the resulting graph vertices. The UMAP method together with the Leiden algorithm found 13 protein clusters, while the Markov clustering algorithm (MCL) found 16 clusters. The similarity of the two clustering methods was analyzed using the number of matching proteins and Jaccard index values between clusters. The clusters obtained by different methods correspond to each other, with the only difference being that MCL breaks large Leiden clusters into several smaller clusters.

The entire taxonomy of bacteria to which the proteins belong was reconstructed. A total of 2266 unique genres belonging to 51 phyla and 104 classes of bacteria were found. The 4 largest phyla by number of complete genomes (Pseudomonadota, Bacillota, Actinomycetota, Bacteroidota - according to the increase in the number of full genomes in NCBI) significantly prevail in clusters over all other phyla, which only indicates an overrepresentation of certain genomes. When considering the diversity of phyla within clusters, it can be stated that most of the clusters are phylum-specific - one phylum is strongly predominant over others, which reveals an evolutionary trend toward phylum-specificity (several clusters – cluster 4, for example – are exceptions to this).

At this stage, phylogenetic trees were constructed followed by functional enrichment and genomic context analysis (STRING) for clusters 1-4. The cluster 1 tree is clearly divided into two main clades, which correspond to pyridine nucleotides metabolic process. In the other clades, there is frequent colocalization with chaperones - DnaJ and DnaK. The cluster 3 has separate clades where the genomic context is presented mainly by transcription regulators - MarR family transcriptional regulator, iron dependent repressor (DtxR family), phosphate regulon transcriptional regulatory protein PhoB (SphR). Also, ATP-dependent DNA helicases (RecQ family) and stress response systems of OmpR protein family (osmoregulation) may be found in the genomic context. Cluster 4 is the most diverse in taxonomy of all the clusters. The diversity of potential functions (not including the functions

described above) also proved to be high: NAD synthetases and ATP-NAD/AcoX kinases, that involved in the regulation of the intracellular balance of NAD and NADP and is a key enzyme in the biosynthesis of NADP [4]; histon-like protein family (in Aquifex genome); integration host factor (IHF) proteins, which regulate DNA replication origins (OriC) in bacteria [5]; cytochrome P450; peptidyl-lysine desuccinylation. This essential role in the cell cycle, common with the other clusters functional annotations, presence of the ancient hyperthermophiles (Aquifex), and high ecological diversity, makes cluster 4 a candidate for evolutionary ancient bacterial sirtuin cluster. Sirtuins' role in pyridine nucleotides and nicotinamide metabolism and transcription regulation are both expected and already proven in vitro functions, while others, described above functions, are fundamentally new ones.

References:

1. G. Cavalli, E. Heard (2019) Advances in epigenetics link genetics to the environment and disease, *Nature* 571 (7766): 489–499.
2. E. S. Miller et al. (2003) Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage, *Journal of bacteriology* 185.17: 5220-5233.
3. J. Gallego-Jara et al. (2021) Bacterial sirtuins overview: an open niche to explore, *Frontiers in Microbiology* 12: 744416.
4. Z.J. Xiao et al. (2012) Functional study of AcoX, an unknown protein involved in acetoin catabolism, *Advanced Materials Research* 393: 776-779.
5. K. Kasho, K. Tsutomu (2013) DnaA binding locus data promotes DnaA-ATP hydrolysis to enable cell cycle-coordinated replication initiation, *Proceedings of the National Academy of Sciences* 110.3: 936-941.

Polygenic Risk Scores and HLA Genotyping Improve Early Detection of Type 1 and Type 2 Diabetes in Eastern Europeans

Layal Shaheen^{1,2}, Michil Trofimov¹, Kharitonov Dmitrii^{1,3}, Mansour Saleem^{1,2}, Leushin Artem^{1,4}, Denisova Aleksandra^{1,5}, Alexander Rakitko¹

¹*Genotek Ltd., Moscow, Russia, ms.layal.shaheen@gmail.com*

²*Phystech school of Biological and Medical physics, Moscow Institute of Physics and Technology, Moscow, Russia*

³*ITMO University, Saint-Petersburg, Russia*

⁴*National Research University Higher School of Economics, Russian Federation*

⁵*Lomonosov Moscow State University, Moscow, Russia*

Diabetes looms as a global crisis, impacting over 500 million adults worldwide. While Type 1 diabetes (T1D) stealthily targets children and adolescents, Type 2 diabetes (T2D) escalates with age, identified by insulin resistance. Early detection is critical, as delayed diagnosis exacerbates complications such as cardiovascular disease and kidney failure.

In a landmark study focusing on Eastern Europeans—an understudied population—we leveraged polygenic risk scores (PRS) and HLA genotyping to unmask genetic risks. Over 10,000 individuals were analyzed across two cohorts: T2D (1,030 cases, 7,758 controls; ages 40–80) and T1D (421 cases, 4,040 controls; ages 8–40). To transform genetic data into actionable risk profiles, cutting-edge tools such as LDpred2 and PRSice were deployed.

For type 1 diabetes (T1D), genome-wide association studies (GWAS) have identified significant associations with several genes: INS, HLA-DQA1/DQB1, TAP2 and BCL2L15. Variants in the INS gene alter thymic insulin expression, disrupting immune tolerance [2]. The HLA-DQA1/DQB1 region plays a key role in immune regulation, with specific alleles such as DQA1*03:03 predisposing individuals to autoimmune destruction of pancreatic β -cells [1]. TAP2 influences antigen presentation, thereby modulating autoimmune risk [3]. Additionally, BCL2L15, a weakly proapoptotic member of the BCL2 family, regulates apoptosis in pancreatic β -cells [4]. These genes

converge on two fronts: immune misfires and β -cell survival—a roadmap for therapies to intercept T1D. A standout discovery [6] — the HLA-DQA1*03:0X ~ DQB1*03:02 haplotype — emerged as a potent risk predictor (OR = 3.89, $p < 2e-16$). This genetic fingerprint could revolutionize newborn screening, enabling vigilant monitoring before symptoms arise.

For T2D, the GWAS did not identify genome-wide significant associations, likely due to the smaller cohort size (1,030 cases) and the polygenic nature of T2D, where individual variants exert weak effects [5]. Larger samples are needed to detect these subtle signals.

For PRS, LDpred2 outperformed PRSice with an AUC of 0.76 for T1D, flagging individuals at $5.1\times$ higher risk (top 10% vs. median) and AUC = 0.62 for T2D.

Our findings demonstrate the utility of PRS and HLA screening in diabetes risk prediction, with LDpred2 outperforming PRSice, suggesting this dual approach could enable personalized prevention strategies, particularly in underrepresented populations like Eastern Europeans. Potential implementations include personalized prevention, where high-risk individuals identified via PRS and HLA markers could receive tailored interventions such as immune therapies or lifestyle coaching, as well as promoting equity in genomics by bridging a critical gap for Eastern Europeans and empowering underserved populations with precision medicine.

1. Noble JA, Valdes AM, Varney MD, et al. HLA class I and genetic susceptibility to type 1 diabetes: results from the Type 1 Diabetes Genetics Consortium. *Diabetes*. 2010;59(11):2972-2979. doi:10.2337/db10-0699
2. Vafiadis P, Bennett ST, Todd JA, et al. Insulin expression in human thymus is modulated by INS VNTR alleles at the IDDM2 locus. *Nature Genetics*. 1997;15(3):289-292. doi:10.1038/ng0397-289
3. Qu HQ, Lu Y, Marchand L, et al. Genetic control of alternative splicing in the TAP2 gene: possible implication in the genetics of type 1 diabetes. *Diabetes*. 2007;56(1):270-275. doi:10.2337/db06-0865
4. Gurzov EN, Eizirik DL. Bcl-2 proteins in diabetes: mitochondrial pathways of β -cell death and dysfunction. *Trends*

- in Cell Biology. 2011;21(7):424-431. doi:10.1016/j.tcb.2011.03.001
5. Mahajan A, Taliun D, Thurner M, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nature Genetics*. 2018;50(11):1505-1513. doi:10.1038/s41588-018-0241-6
 6. Enczmann J, Balz V, Hoffmann M, Kummer S, Reinauer C, Döing C, Förtsch K, Welters A, Mayatepek E, Meissner T, Jacobsen M, Seyfarth J. Next Generation Sequencing Identifies the HLA-DQA1*03:03 Allele in the Type 1 Diabetes Risk-Associated HLA-DQ8 Serotype. *Genes (Basel)*. 2021 Nov 25;12(12):1879. doi: 10.3390/genes12121879. PMID: 34946827; PMCID: PMC8701008.

The Role of Locus Spatial Structure in the Regulation of Keratin Gene Transcription

Shtompel A.S.^{1,2}, Kalabusheva E.P.³, Luzhin A.V.¹, Molodova M.N.^{1,4}, Zapevalina M.K.⁴, Selivanovsky A.V.^{1,2}, Khrameeva E.E.⁴, Ulianov S.V.^{1,2}, Razin S.V.^{1,2}

¹*Institute of Gene Biology RAS, Moscow, Russia*

²*Lomonosov Moscow State University, Moscow, Russia*
shtompel.an@gmail.com

³*Koltzov Institute of Developmental Biology, RAS, Moscow, Russia*

⁴*Skolkovo Institute of Science and Technology, Moscow, Russia*

Three-dimensional chromatin organization plays a key role in transcription regulation by enabling physical interactions between regulatory elements. One of the best-characterized examples of such regulation is the β -globin gene cluster, where interactions between the locus control region (LCR) and gene promoters ensure stage-specific expression during erythroid differentiation. Similar regulatory principles may underlie the organization of other paralogous gene clusters, such as the human keratin gene cluster (12q13.13). Although several putative

LCR-like elements have been previously identified within the keratin gene locus, their functional role and contribution to gene regulation remain unclear.

The aim of this study was to investigate the contribution of regulatory elements to the spatial organization of the keratin gene locus during different stages of epidermal differentiation. To analyze the chromatin structure of the human keratin gene locus (12q13.13), we used immortalized human keratinocytes (HaCaT), primary human skin keratinocytes, and a model of epidermal differentiation based on induced pluripotent stem cells (iPSCs). Chromatin structure was analyzed using C-TALE technology and fluorescent in situ hybridization (FISH) combined with confocal microscopy. Chromatin immunoprecipitation followed by sequencing (ChIP-seq) was used to assess the binding profiles of transcription factors and histone modifications. In addition, cells were treated with a panel of inhibitors targeting transcription and phase separation. Using the C-TALE, we characterized the 3D structure of the keratin gene locus in the iPSC-derived epidermal differentiation model. The formation of chromatin contacts occurs in a stepwise manner: initially, a contact is established between LCR1 and LCR2, followed by interactions between the LCRs and the KRT5 gene promoter, which correlates with transcription activation. 3D-FISH analysis demonstrated compaction of the locus in differentiating cells and an increased frequency of colocalization between the KRT5 promoter and LCRs. These findings suggest the formation of an active hub that facilitates spatial proximity of regulatory elements and topological insulation of the locus from its genomic surroundings.

To elucidate the mechanisms underlying the establishment of this spatial structure, cells were treated with inhibitors of transcription (flavopiridol, triptolide) and phase separation (JQ1, 1,6-hexanediol). These treatments revealed differential sensitivity of regulatory contacts within the locus. The contact between LCR1 and the KRT5 promoter have the most sensitive, weakening upon treatment with triptolide and phase separation inhibitors, while other contacts remained relatively stable. These observations suggest that distinct mechanisms contribute differentially to the formation and maintenance of specific spatial interactions within the locus.

Taken together, these results highlight the essential role of LCR

elements in regulating transcription and organizing the spatial configuration of the keratin gene locus. The 3D organization of the locus represents a complex network of regulatory interactions that ensures precise gene activation during epidermal differentiation.