

## МЕТОДЫ ФУНКЦИОНАЛЬНОЙ ХАРАКТЕРИСТИКИ ПОЛИМОРФИЗМОВ НЕКОДИРУЮЩИХ РЕГУЛЯТОРНЫХ ОБЛАСТЕЙ ГЕНОМА ЧЕЛОВЕКА

### Обзор

© 2024 А.Н. Уварова<sup>1\*</sup>, Е.А. Ткаченко<sup>2,3</sup>, Е.М. Стасевич<sup>1,4</sup>, Э.А. Жеремян<sup>1</sup>,  
К.В. Корнеев<sup>1</sup>, Д.В. Купраш<sup>1,3</sup>

<sup>1</sup> Институт молекулярной биологии им. В.А. Энгельгардта РАН,  
Центр высокоточного редактирования и генетических технологий для биомедицины,  
119991 Москва, Россия; электронная почта: uvaowwww@gmail.com

<sup>2</sup> Институт молекулярной биологии им. В.А. Энгельгардта РАН, 119991 Москва, Россия

<sup>3</sup> Московский государственный университет имени М.В. Ломоносова,  
биологический факультет, 119234 Москва, Россия

<sup>4</sup> Московский физико-технический институт, 141700 Долгопрудный, Московская обл., Россия

Поступила в редакцию 20.11.2023

После доработки 27.03.2024

Принята к публикации 11.04.2024

В настоящее время благодаря полногеномным исследованиям охарактеризовано множество ассоциаций между генетическими полиморфизмами и различными заболеваниями. Большинство клинически значимых полиморфизмов локализованы в некодирующих областях генома. Современные биоинформатические ресурсы позволяют предсказывать молекулярные механизмы влияния некодирующих полиморфизмов на экспрессию генов, однако такие гипотезы нуждаются в экспериментальной проверке. В представленном обзоре рассмотрены основные методы определения молекулярных механизмов зависимости патогенеза заболеваний от тех или иных генетических вариантов некодирующих последовательностей. В частности, рассмотрены способы определения факторов транскрипции, эффективность связывания которых зависит от варианта полиморфизма. Биоинформатические ресурсы, позволяющие предсказывать механизм влияния полиморфизмов на патогенез заболеваний, достигли впечатляющего развития, однако необходимость в экспериментальных подходах к данному вопросу остается актуальной.

**КЛЮЧЕВЫЕ СЛОВА:** функциональные полиморфизмы, регуляторные области, транскрипционные факторы, репортерный анализ, CRISPR-Cas.

DOI: 10.31857/S0320972524060023 EDN: XMSUZO

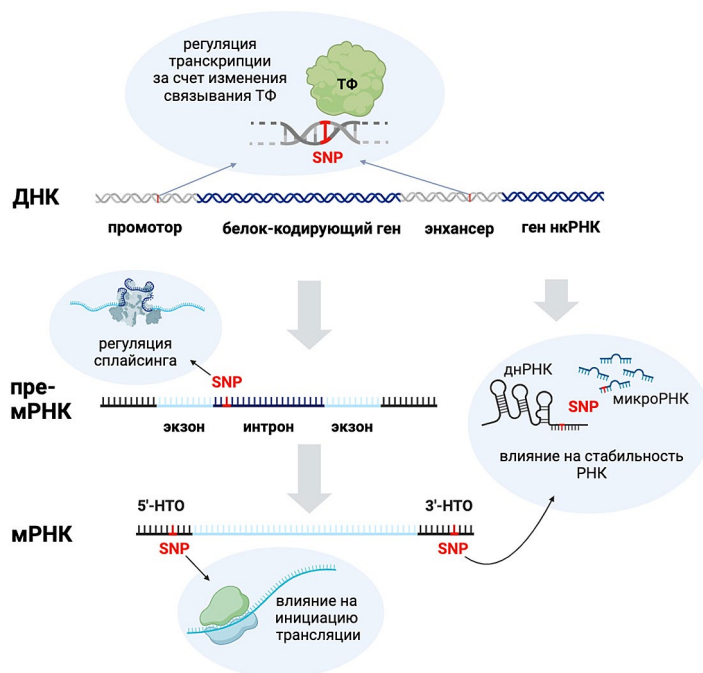
### ВВЕДЕНИЕ

Несмотря на то что геномы людей на 99,9% идентичны, именно оставшиеся 0,1% генетических вариаций лежат в основе фенотипических различий, в том числе восприимчивости к болезням [1]. К таким генетическим вариантам отно-

сят однонуклеотидные полиморфизмы (SNV, Single Nucleotide Variation или SNP, Single Nucleotide Polymorphism), инсерции/делеции (indel), а также структурные вариации длиной более 50 п.н. (SV, Structural Variation) [2]. Наиболее распространенным генетическим вариантом является SNP – вариация (вариант аллеля) последовательности

Принятые сокращения: ДЦР – двухцепочечный разрыв ДНК; НТО – нетранслируемая область; ТФ – транскрипционный фактор; ChIP – иммунопреципитация хроматина; CRISPR – короткие палиндромные повторы, регулярно расположенные группами; EMSA – анализ сдвига электрофоретической подвижности; GWAS – полногеномные исследования; HDR – гомологически направленная репарация; MPRA – массовый параллельный репортерный анализ; QTL – локус количественного признака; eQTL – локус количественного признака, регулирующий экспрессию генов на уровне транскрипции; raQTL – локус количественного признака, определяемый по активности репортерного гена; SNP – однонуклеотидный полиморфизм.

\* Адресат для корреспонденции.



**Рис. 1.** Основные механизмы влияния некодирующих полиморфизмов на регуляцию экспрессии генов (изображение сделано с помощью BioRender.com)

ДНК размером в один нуклеотид у представителей одного вида, встречающаяся в популяции с частотой не менее 1% [3]. SNP встречаются в геноме каждые 200–300 п.н., располагаются как в кодирующих частях генома, так и в регуляторных (промоторы, энхансеры, интроны и нетранслируемые области) [4, 5]. Актуальность изучения SNP заключается в том, что такие генетические варианты часто ассоциированы с различными заболеваниями, что показано с помощью многочисленных полногеномных исследований (Genome-Wide Association Studies, GWAS). Порядка 95% клинически значимых SNP локализованы в некодирующих областях генома [6], и их функциональная значимость может быть связана с изменением регуляторных характеристик областей, окружающих полиморфизм [7]. К таким регуляторным областям эукариотического генома можно отнести промоторы, энхансеры 5'- и 3'-нетранслируемые области (НТО) белок-кодирующих генов, области генов некодирующих РНК (нкРНК) и регуляторные элементы сплайсинга (SRE, splicing regulatory elements) [5, 8]. Промоторы инициируют транскрипцию генов, а энхансерные элементы усиливают эту инициацию [9]. Промоторы являются предпочтительными местами посадки транскрипционных факторов (ТФ) и РНК-полимеразы II на ДНК и включают в себя область первого транскрибируемого нуклеотида транскрипта (сайт начала транскрипции, TSS) [10]. Энхансеры, впервые идентифицированные с помощью репортерного анализа как элементы, способные увеличивать экспрессию репортерного гена [11], представляют собой плат-

формы для связывания ТФ, способные действовать независимо от ориентации, расстояния и расположения по отношению к гену-мишени [12]. 5'- и 3'-НТО играют важную роль в посттранскрипционной регуляции экспрессии генов и являются частью зрелой кодирующей мРНК. Так, 5'-НТО содержат различные регуляторные компоненты, влияющие на инициацию трансляции, а в состав 3'-НТО входят последовательности, связывающие микроРНК и приводящие к деградации транскрипта [5]. Также следует отметить, что некодирующие полиморфизмы в составе НТО могут участвовать и в регуляции транскрипции, поскольку последовательность 5'-НТО обычно пересекается с промоторными регионами генов, а последовательность 3'-НТО может пересекаться с другими регуляторными элементами генов – например, энхансерами [13]. Некодирующие полиморфизмы также располагаются в нкРНК, о влиянии которых на созревание РНК, регуляцию транскрипции, ремоделирование хроматина и посттранскрипционные модификации РНК за последние годы получено много информации [14].

Будучи наиболее часто встречающимся классом генетических вариантов, SNP являются основным генетическим маркером при картировании локусов количественных признаков (QTL, Quantitative Trait Loci), которые можно условно разделить на те, которые регулируют экспрессию генов непосредственно на уровне транскрипции и хроматина, что отражается на уровне мРНК (eQTL – expression QTL, регулирующий экспрессию генов на уровне транскрипции), и те,

которые влияют на посттранскрипционные процессы (sQTL – splicing QTL, регулирующие альтернативный сплайсинг пре-мРНК; pQTL – protein QTL, регулирующие экспрессию белка) [15]. Механизмом функционального влияния полиморфизмов на геномном уровне может быть нарушение функционирования регуляторных элементов вследствие изменения последовательности сайтов взаимодействия ТФ с ДНК (как в сторону уменьшения, так и в сторону увеличения эффективности связывания) [16]. На посттранскрипционном уровне некодирующие полиморфизмы могут влиять на активность 5'- и 3'-НТО мРНК, которые играют важную роль в регуляции трансляции и стабильности мРНК, в том числе за счет изменения связывания регуляторных микроРНК [17–19]. Также SNP в последовательности незрелой микроРНК могут влиять на эффективность созревания микроРНК и изменять эффективность связывания с мРНК [20, 21], а аллельные варианты в составе днРНК (длинных некодирующих РНК) – с разной эффективностью модулировать концентрацию комплементарных микроРНК [22]. Значительное количество функциональных генетических вариантов, классифицируемых как sQTL, располагаются в регуляторных элементах сплайсинга, непосредственно изменяя последовательность сайтов сплайсинга или модифицируя сайты связывания РНК-связывающих белков [23]. Основные механизмы влияния некодирующих полиморфизмов на регуляцию экспрессии генов представлены на рис. 1.

В представленном обзоре описаны основные экспериментальные подходы к анализу функциональных некодирующих аллельных вариантов, в том числе рассмотрены способы определения ТФ, эффективность связывания которых зависит от аллельного варианта.

#### **АНАЛИЗ ВЛИЯНИЯ ГЕНЕТИЧЕСКИХ ПОЛИМОРФИЗМОВ НА ЭКСПРЕССИЮ ГЕНОВ С ИСПОЛЬЗОВАНИЕМ ГЕНОВ-РЕПОРТЕРОВ**

Экспериментальные методы изучения влияния полиморфизмов на экспрессию гена можно разделить на две большие группы: исследования с использованием генетических репортерных конструкций и изучение полиморфизмов непосредственно в нативном геномном контексте.

Первая группа методов включает в себя использование репортерных генетических конструкций, в которых влияние генетического варианта на регуляторный элемент определяется по активности репортерного гена (reporter assay QTL, гаQTL). К ним относится и классический метод люциферазного репортерного анализа, когда

аллельные варианты исследуемой регуляторной последовательности (промотора или энхансера) интегрируют в репортерную конструкцию и сравнивают активность гена-репортера в полученных конструкциях после их трансфекции в физиологически релевантную культуру клеток [24]. С помощью метода двойного люциферазного теста описано множество гаQTL в разных типах клеток и регуляторных элементах. Так, ранее были предложены молекулярные механизмы, объясняющие связь между развитием заболеваний и генетическими полиморфизмами, расположенными в регуляторных зонах различных генов: промоторе [25–28], близкорасположенных энхансерных областях [29, 30] и энхансерах, расположенных в отдаленных межгенных локусах [31, 32].

Поиск sQTL в системах с геном-репортером обычно затруднен размером исследуемого гена, превышающим емкость репортерной плазмиды. В такой ситуации используют так называемые репортерные мини-гены. Конструкция мини-гена включает в себя фрагмент исследуемого локуса, содержащий полиморфизм и достаточный для воспроизведения природной картины сплайсинга между репортерами сплайсинга (как правило, между двумя экзонами). Способность исследуемой области влиять на эффективность сплайсинга измеряется по экспрессии целевого транскрипта либо в ядерном экстракте с помощью количественного ПЦР-анализа (кПЦР), либо в живых клетках, если кодируемый мини-геном репортерный белок позволяет это сделать [33]. Так, с помощью репортерного анализа мини-генов были охарактеризованы полиморфизмы, регулирующие сплайсинг гена субъединицы кальциевого канала *SCN1A*, связанного с эпилепсией [34], компонента системы репарации ДНК *RAD51C*, проявляющего себя в качестве онкосупрессора [35] и ряда других.

#### **ВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ РЕПОРТЕРНЫЕ АНАЛИЗЫ**

За последнее десятилетие появилось множество модификаций высокопроизводительных репортерных анализов. Их можно классифицировать по регуляторным областям, которые они позволяют исследовать, а также по техническим особенностям. Так, протокол массового параллельного репортерного анализа (MPRA, Massively Parallel Reporter Assay) включает в себя синтез последовательностей ДНК (потенциальных энхансеров/промоторов, 5'-НТО либо 3'-НТО) с добавлением уникальных баркодов и клонирование этих последовательностей в репортерные плазмиды, которые затем трансфицируются в типы клеток, интересующие исследователей. Анализ активности

регуляторных областей происходит с помощью высокопроизводительного секвенирования и количественного определения баркодов, которые однозначно определяют конкретную регуляторную последовательность и коррелируют с уровнем РНК репортерного гена [36, 37]. Метод MPRA естественным образом подходит не только для изучения функциональности регуляторных элементов, но и для оценки функционального влияния их генетических вариантов [38]. Так, например, с помощью метода MPRA был проведен скрининг полиморфизмов, расположенных в некодирующих областях генома и ассоциированных с шизофренией и с болезнью Альцгеймера. Интересно, что из 148 SNP, показавших аллельные различия в клетках K562 и 53 – в клетках SK-SY5Y, только 9 показали аллельные различия в обеих клеточных линиях, наглядно демонстрируя, что генетические варианты обычно оказывают свое регуляторное действие только в определенных типах клеток [39]. Применение метода MPRA к библиотеке 5'-НТО генов человека позволило выявить 45 ассоциированных с заболеваниями аллельных вариантов, которые значительно влияют на процесс загрузки мРНК на рибосомы, однако интересно, что для большинства из найденных вариантов полученных данных оказалось недостаточно для изменения классификации патогенности в базе данных Clinvar, а наиболее яркий эффект показали 3 полиморфизма, которые создавали новый стартовый кодон, то есть влияли на структуру белка [40]. В другой работе Griesemer в 6 клеточных линиях человека было изучено более 12 000 3'-НТО-вариантов, ассоциированных с заболеваниями человека и/или находящихся под положительным давлением в человеческой популяции [41]. Оказалось, что несколько сотен из них значимо влияли на уровень репортерного транскрипта хотя бы в одной клеточной линии и несколько десятков совпали с ранее охарактеризованными вариантами с каким-либо уровнем клинической значимости. Интересно, что лишь для двух SNP, находящихся в гене вирусной защиты *TRIM14* и в гене *PILRB*, ассоциированном с возрастной макулярной дегенерацией, сочетание новизны и степени влияния на уровень транскрипта оказалось достаточным для проверки гипотезы при помощи Cas9-опосредованной замены аллеля в геномном контексте [41]. Технические ограничения метода MPRA также включают в себя длину тестируемых фрагментов ДНК до 130–230 п.н. и количество тестируемых конструкций – до 100–200 тысяч последовательностей [42].

Существуют высокопроизводительные подходы, в которых используются последовательности, полученные из геномной ДНК. Так, метод SuRE (Survey of Regulatory Elements) был разработан с

использованием данных секвенирования геномов клеточных линий, происходящих от четырех различных этнических групп, и оптимизирован для изучения потенциального влияния однонуклеотидной замены на активность регуляторных элементов [43]. Случайные фрагменты геномной ДНК длиной в несколько сотен п.н. клонируются в репортерную плазмиду без промотора, которая при трансфекции в культивируемые клетки продуцирует транскрипт только в том случае, если вставленный фрагмент несет функциональный сайт начала транскрипции. Так как транскрипты могут производить как активные промоторы, так и энхансеры, этот метод позволяет анализировать активность обоих типов регуляторных элементов. Как и в методе MPRA, транскрипты анализируются с помощью высокопроизводительного секвенирования и количественно оцениваются с использованием баркода, уникального для каждого изучаемого геномного фрагмента. Такой подход позволяет протестировать активность регуляторных элементов, содержащих альтернативные аллели нескольких миллионов различных SNP (то есть большинства из известных).

Еще один метод, позволяющий идентифицировать *raQTL* – High-resolution Dissection of Regulatory Activity (HiDRA) [42] – также использует фрагментацию геномной ДНК и представляет собой комбинацию подходов ATAC-seq и STARR-seq. ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) позволяет обогатить образцы доступным для транспозазы, т.е. открытым, хроматином, в то время как STARR-seq (Self-Transcribing Active Regulatory Region sequencing) представляет собой репортерный анализ, в ходе которого предполагаемые регуляторные элементы (способные повышать транскрипционную активность репортера) клонируются в 3'-НТО гена-репортера и, таким образом, способствуют собственной транскрипции. Затем активные последовательности ДНК идентифицируют и количественно оценивают с помощью высокопроизводительного секвенирования РНК [44]. Так, с помощью HiDRA в локусе гена *IKZF3* был найден драйверный элемент длиной 76 п.н., включающий в себя rs12946510, который ассоциирован с рассеянным склерозом, что позволяет выделить этот SNP в качестве потенциально функционального [42]. Действительно, при дальнейшем функциональном тестировании было показано, что присутствие рискового аллеля rs12946510 снижало активацию Т-хелперов и экспрессию генов *IKZF3* и *ORMDL3* [45]. Важным этапом обработки результатов каждого из описанных высокопроизводительных методов является использование вероятностных математических моделей, таких, например, как алгоритм SHARPR-RE (Systematic High-resolution Activation and Repression Prediction from Reporter

assays with Random Endpoints) [46], для анализа наложения последовательностей и оценки влияния конкретных нуклеотидов на активность этих последовательностей.

К высокопроизводительным репортерным анализам полиморфных вариантов относится метод MaPSY (Massively Parallel Splicing Assay) [47], который был использован для изучения нарушений сплайсинга при расстройствах аутистического спектра. По результатам скрининга были охарактеризованы генетические варианты в генах *TNRC6C*, *MAPK8IP1* и *USP45* и показано, что белки семейства TNRC6 могут увеличивать риск развития аутизма [48]. Недавно был предложен метод Cre-зависимого MPRA *in vivo* для функционального анализа библиотеки 3'-НТО с генетическими вариантами, ассоциированными с аутизмом. Количественная оценка транскриптов в зависимости от активности регуляторного элемента осуществлялась в определенных типах нейронов за счет трансдукции библиотек в ткани мозга мышей с тканеспецифической экспрессией Cre-рекомбиназы. Этот метод позволяет изучить регуляторный эффект в более релевантном клеточном контексте, поскольку нейроны имеют совершенно другую экспрессию транс-действующих факторов (например, ТФ и микроРНК), чем клеточные линии [49].

Главное ограничение методов, основанных на репортерных тестах, состоит в отсутствии релевантного хроматинового контекста, который сопровождает регуляторный элемент в нативном геноме. Отчасти это ограничение снимается в методе lentiMPRA, когда библиотека с исследуемыми регуляторными элементами создается в лентивирусном векторе, который интегрируется в геном, обеспечивая анализ транскрипции внутри хроматинового контекста [50].

### ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ ГЕНЕТИЧЕСКИХ ПОЛИМОРФИЗМОВ В НАТИВНОМ ГЕНОМНОМ КОНТЕКСТЕ

Говоря о влиянии генетических вариантов на патогенез заболевания, важно учитывать контекст хроматина, который, в свою очередь, различается у разных типов и функциональных состояний клеток. Картирование eQTL само по себе позволяет связать конкретный генотип с изменением уровней мРНК потенциальных генов-мишеней в нативном геномном контексте, в том числе тканеспецифично [51, 52]. Функциональная связь генов с регуляторными локусами, находящимися на расстоянии от них, может быть найдена с помощью определения трехмерной организации хроматина такими методами, как Hi-C (high-throughput chromosome conformation capture),

ChIA-PET (chromatin interaction analysis with paired-end tag sequencing) и их модификациями [53, 54]. Сопоставление трехмерных тканеспецифических геномных карт с болезнью-ассоциированными регуляторными SNP позволяет идентифицировать наиболее вероятные гены-участники патогенеза. Наиболее точным методом проверки построенных таким образом гипотез является редактирование генома и получение клеток с нужными комбинациями вариантов. Точное и эффективное редактирование отдельных нуклеотидов в геноме человека стало трудновыполнимой, но реальной задачей благодаря РНК-программируемым бактериальным нуклеазам, найденным в системе CRISPR (clustered regularly interspaced short palindromic repeats)-Cas [55]. Двухцепочечный разрыв ДНК (ДЦР), индуцированный в целевом сайте нуклеазой Cas9 из *Streptococcus pyogenes* (наиболее популярным на настоящий момент геномным редактором), запускает клеточные механизмы репарации ДНК, в том числе гомологически направленную репарацию HDR [56], которая задействована в методах семейства CRISPR-HDR, когда восстановление целевого участка происходит в присутствии гомологичной последовательности ДНК, содержащей нужный аллельный вариант.

Данный метод, использованный во многих работах по изучению полиморфизмов [57, 58], имеет существенное ограничение по эффективности, так как у млекопитающих репарация ДЦР преимущественно происходит с помощью нехомологического соединения концов (NHEJ) [59]. Исходя из этих особенностей, редактирование CRISPR-HDR требует больших усилий и может приводить к нарушениям экспрессии близлежащих генов [45]. Другим подходом к точному редактированию генома, хорошо проявляющим себя в подходящем нуклеотидном контексте, является редактирование азотистых оснований (BE, base editing) с помощью каталитически неактивной dCas9 (dead Cas9) или Cas9 с низкой активностью (nCas9), слитой с ферментом дезаминазой. В зависимости от специфичности фермента выделяют цитозинные (преобразуют С•G в Т•A) и адениновые (А•T в G•C) редакторы, а также редактор с цитидиндезаминазой и урацил-ДНК-гликозилазой (преобразуют С•G в G•C) [60–63]. Так, с помощью технологии цитозинового редактора был исследован полиморфизм rs12603332, ассоциированный с риском развития астмы, и показано его влияние на экспрессию генов регулятора биосинтеза сфинголипидов *ORMDL3* и модулятора клеточного ответа на стресс *ATF6a* в Т-клеточной линии Jurkat [64]. Благодаря отсутствию этапа образования ДЦР метод BE является более безопасным для клеток, чем CRISPR-HDR, однако он имеет ограничения в плане активности ферментов и нецелевого редак-

тирования близлежащих нуклеотидов [60, 61]. Еще один недавно разработанный перспективный подход к редактированию генома – праймированное редактирование (prime editing). Редактор основан на мутантной нуклеазе Cas9, вносящей одноцепочечные разрывы (nCas9), слитой с обратной транскриптазой (MMLV RT), и использует модифицированную направляющую РНК (pegRNA), которая одновременно определяет целевой сайт для nCas9, выступает в роли праймера для MMLV RT и является РНК-матрицей для синтеза новой последовательности ДНК. Отредактированная цепь ДНК затем включается в геном посредством эндогенных клеточных процессов [65].

Благодаря высокой точности редактирования и более широкой, чем у стандартных редакторов оснований, области применения, праймированное редактирование имеет большой потенциал для работы с однонуклеотидными полиморфизмами. С помощью однонуклеотидных замен осуществляют направленную эволюцию, применяемую в селекции сельскохозяйственных культур [66, 67]. В мышинных моделях была показана низкая внецелевая активность праймированного редактирования при изменении варианта некодирующего полиморфизма в сравнении CRISPR-HDR-методом [68]. Также с помощью Prime-редактирования в миобластах человека была скорректирована мутация в белок-кодирующей части гена кальциевого канала *RYR1*, связанная с двигательными нарушениями [69]. Основным принципиальным ограничением метода праймированного редактирования является большой размер и сложности с доставкой редактора в клетки [70]. При построении систем для высокопроизводительного скрининга проблема доставки может быть решена при помощи лентивирусной трансдукции целевых клеток конструктами, кодирующими редактор и pegRNA. Последующее культивирование клеток в течение нескольких недель позволяет достичь эффективности редактирования, достаточной для исследования функционального эффекта сотен и даже тысяч однонуклеотидных замен в одном эксперименте [71].

#### **ИДЕНТИФИКАЦИЯ ТРАНСКРИПЦИОННЫХ ФАКТОРОВ, ОПОСРЕДУЮЩИХ АЛЛЕЛЬ-СПЕЦИФИЧЕСКУЮ РАЗНИЦУ АКТИВНОСТИ РЕГУЛЯТОРНЫХ ЭЛЕМЕНТОВ**

Определение различных типов QTL (eQTL, гаQTL и т.д.) не дает информации о конкретном молекулярном механизме, на который влияет определенный генетический вариант, поэтому остается актуальной дальнейшая функциональная

аннотация. Как упоминалось ранее, к механизмам влияния полиморфизмов на функции регуляторного элемента относятся изменения свойств промоторных и энхансерных областей, 5'-НТО и 3'-НТО, нкРНК, а также нарушения сплайсинга. Наиболее изученной причиной зависимости свойств регуляторных элементов от располагающихся в них SNP является способность однонуклеотидной замены влиять на сродство к функциональному транскрипционному фактору.

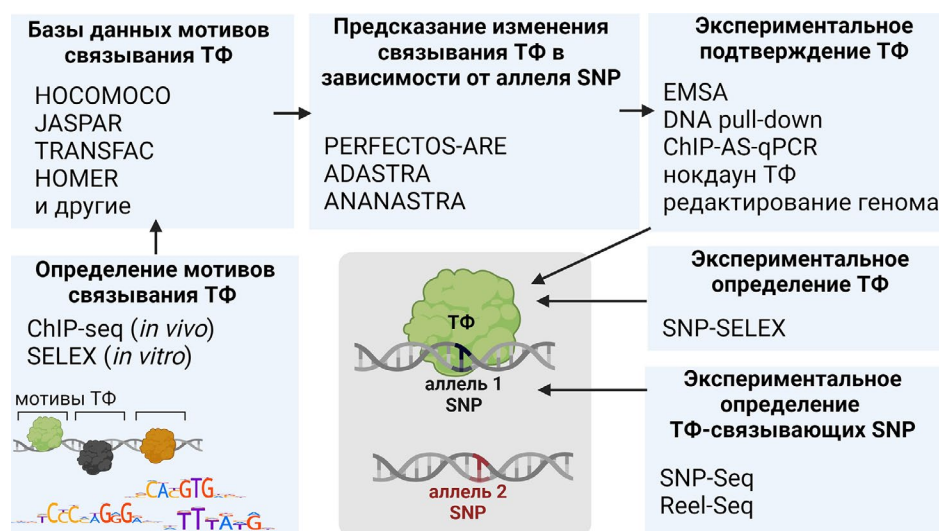
Существуют разнообразные *in silico* подходы для предсказания предпочтительных мотивов связывания ТФ, в большинстве случаев основанные на позиционных весовых матрицах (PWM, Positional Weight Matrix), формируемых с помощью множественного выравнивания ТФ-связывающих последовательностей [72, 73]. В свою очередь, информация о конкретных ТФ-связывающих последовательностях может быть получена с помощью высокопроизводительных методов полногеномного картирования сайтов связывания *in vivo*, например, основанных на иммунопреципитации хроматина ChIP (Chromatin Immunoprecipitation) или на высокопроизводительном HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment) для отбора последовательностей связывания ТФ *in vitro* [74]. Для идентификации последовательностей генома, связанных со специфическими белками в их нативном хроматиновом контексте, используют метод ChIP-seq, сочетающий в себе иммунопреципитацию хроматина с последующим высокопроизводительным секвенированием ДНК [75]. Последовательности, оптимальные для связывания определенного ТФ (возможно, не существующие в природе), ищут при помощи SELEX-семейства методик обогащения библиотек случайно сгенерированных олигонуклеотидов специфическими последовательностями с высоким сродством к данному ТФ [76]. К широко известным базам данных, основанных на методе PWM, относятся TRANSFAC [77], HOCOMOCO [78], JASPAR [79], HOMER [80], iRegulon [81] и др. Биоинформатические приложения позволяют оценивать потенциальное изменение силы связывания ТФ в зависимости от варианта полиморфизма. Эффективность аллель-специфического связывания ТФ может быть оценена непосредственно по данным ChIP-Seq, если глубина секвенирования позволяет обнаружить статистически значимые отклонения в частотах альтернативных аллелей SNP в сайте связывания [82, 83]. Комбинация ChIP с количественным определением аллелей, ChIP-AS-qPCR (ChIP-based allele-specific quantitative PCR), позволяет измерить влияние аллельных вариантов на эффективность связывания ТФ в живой клетке [57]. Был предложен высокопроизводительный вариант анализа связывания ТФ

с полиморфизмами регуляторных областей SNP-SELEX на основе HT-SELEX. Этот метод позволяет анализировать влияние порядка 100 000 аллельных вариантов потенциально регуляторных SNP (GWAS-аннотированных) на связывание нескольких сотен ТФ [84]. К экспериментальным подходам идентификации ТФ можно также отнести классический метод анализа ДНК-белковых взаимодействий, основанный на сдвиге электрофоретической подвижности – EMSA (electrophoretic mobility shift assay). В процессе EMSA происходит специфическое связывание исследуемых белков с меченой олигонуклеотидной пробой с последующим анализом подвижности таких фрагментов при электрофорезе в полиакриламидном геле в нативных условиях, причем количество образующегося комплекса позволяет оценить относительную силу связывания [85]. Специфичность определения белковых компонентов в комплексах достигается путем добавления антитела к конкретному белку в реакции: EMSA-supershift [86]. Также существуют высокопроизводительные методы анализа большого количества SNP на предмет влияния аллельных вариантов на связывание ТФ, основанные на инкубации SNP-содержащих олигонуклеотидов с ядерным экстрактом из интересующего типа клеток, с последующим анализом обогащенных библиотек с помощью секвенирования; к таким методам относятся SNPs-Seq [57] и Reel-Seq [87]. Ни один из этих методов сам по себе не позволяет установить, какой ТФ связывается с конкретным аллельным вариантом, однако такая информация может быть получена с помощью масс-спектрометрии и/или с использованием очищенного ТФ вместо ядерного экстракта [24, 88].

Среди удобных биоинформатических баз данных для точечного анализа интересующего SNP

можно выделить онлайн-ресурс PERFECTOS-ARE <https://opera.autosome.org/perfectosare> [76], в котором собраны предсказанные мотивы сайтов связывания ТФ из ряда баз данных: HOCOMOCO [78], JASPAR [79], HT-SELEX [89] и других. Другой биоинформатический ресурс, ADAstra [82], содержащий обширные данные об аллель-специфическом связывании ТФ с аллельными вариантами в разных типах клеток, базируется на данных HOCOMOCO и SPRy-SARUS [90], а также на аллель-специфичных данных ДНКазного футпринтинга [91]. Ресурс ANANASTRA [92], построенный на основе систематического анализа аллельного дисбаланса в экспериментах ChIP-Seq, позволяет аннотировать большое количество генетических вариантов параллельно.

В качестве примеров использования такой аннотации можно привести функциональную характеристику SNP rs7873784 и rs71327024, расположенных в регуляторных областях генов *TLR4* и *CXCR6* соответственно [13, 31]. По результатам GWAS для обоих SNP была показана ассоциация с заболеваниями: минорный С-аллель rs7873784 ассоциирован с ревматоидным артритом, а минорный Т-аллель rs71327024 ассоциирован с тяжелым течением COVID-19. Репортерные тесты показали, что оба SNP являются *raQTL*, поэтому с помощью биоинформатического анализа были найдены релевантные для соответствующих типов клеток ТФ PU.1 (rs7873784) и c-Myb (rs71327024), аллель-зависимо связывающиеся с сайтами, содержащими SNP. Для подтверждения этой гипотезы был проведен генетический нокдаун ТФ с использованием малых интерферирующих РНК (siRNA) и метод иммунопреципитации «ДНК pull-down» [93]. Последний метод включает в себя инкубацию олигонуклеотидов, содержащих альтернативные



**Рис. 2.** Методы определения функциональных транскрипционных факторов, аллель-специфически связывающихся с областью полиморфизма (изображение сделано с помощью BioRender.com)

варианты SNP, с ядерным экстрактом из релевантных клеток, иммунопреципитацию специфическими антителами к предсказанному ТФ и последующую количественную оценку обогащенных олигонуклеотидов. Описанные методы определения транскрипционных факторов, эффективность связывания которых зависит от аллеля полиморфизма, представлены на рис. 2.

Благодаря постоянно растущему объему данных и современным моделям машинного обучения, биоинформатические расчеты позволяют со все большей точностью аннотировать кандидатные ТФ, аллель-специфически связывающиеся с областью SNP [94–96]. Однако клиническая валидация и тем более использование этих данных в диагностике и, возможно, в терапии заболеваний возможны только после экспериментального подтверждения в разных типах клеток в релевантном функциональном контексте.

## ЗАКЛЮЧЕНИЕ

В настоящее время благодаря метаанализу большого количества экспериментальных данных создаются биоинформатические приложения для поиска наиболее вероятных функциональных генетических вариантов, а также предсказания конкретных механизмов их влияния на патогенез заболеваний. Подавляющая часть генетических вариантов располагается в некодирующих областях генома и влияет на функционирование генов

за счет регуляции их экспрессии. Такая регуляция может широко варьироваться в зависимости от типа клеток и их функционального состояния, что не всегда учитывается в *in silico* подходах, допускающих статистические обобщения. В связи с этим остается актуальной необходимость в разносторонних экспериментальных методах характеристики конкретных генетических вариантов. Наиболее информативным способом изучения влияния генетических вариантов на фенотип является создание точных генетических моделей с помощью технологий редактирования генома. Однако из-за сложности процедуры точного редактирования по-прежнему сохраняет актуальность предварительная характеристика исследуемых аллельных вариантов с использованием репортерных тестов.

**Вклад авторов.** А.Н. Уварова – концепция и руководство работой; А.Н. Уварова, Е.А. Ткаченко, Е.М. Стасевич, Э.А. Жеремян – написание текста; К.В. Корнеев, Д.В. Купраш – редактирование текста статьи.

**Финансирование.** Работа выполнена при финансовой поддержке Российского научного фонда (грант № 22-24-00987).

**Конфликт интересов.** Авторы заявляют об отсутствии конфликта интересов.

**Соблюдение этических норм.** Настоящая статья не содержит описания каких-либо исследований с участием людей или животных в качестве объектов.

## СПИСОК ЛИТЕРАТУРЫ

1. Ahmed, Z., Zeeshan, S., Mendhe, D., and Dong, X. (2020) Human gene and disease associations for clinical-genomics and precision medicine research, *Clin. Transl. Med.*, **10**, 297-318, <https://doi.org/10.1002/ctm2.28>.
2. Lappalainen, T., Scott, A. J., Brandt, M., and Hall, I. M. (2019) Genomic analysis in the age of human genome sequencing, *Cell*, **177**, 70-84, <https://doi.org/10.1016/j.cell.2019.02.032>.
3. Wright, A. F. (2005) Genetic variation: polymorphisms and mutations, in *eLS*, <https://doi.org/10.1038/npg.els.0005005>.
4. Salisbury, B. A., Pungliya, M., Choi, J. Y., Jiang, R., Sun, X. J., and Stephens, J. C. (2003) SNP and haplotype variation in the human genome, *Mutat. Res.*, **526**, 53-61, [https://doi.org/10.1016/S0027-5107\(03\)00014-9](https://doi.org/10.1016/S0027-5107(03)00014-9).
5. Fabo, T., and Khavari, P. (2023) Functional characterization of human genomic variation linked to polygenic diseases, *Trends Genet.*, **39**, 462-490, <https://doi.org/10.1016/j.tig.2023.02.014>.
6. Orozco, G., Schoenfelder, S., Walker, N., Eyre, S., and Fraser, P. (2022) 3D genome organization links non-coding disease-associated variants to genes, *Front. Cell Dev. Biol.*, **10**, 995388, <https://doi.org/10.3389/fcell.2022.995388>.
7. Johnston, A. D., Simões-Pires, C. A., Thompson, T. V., Suzuki, M., and Greally, J. M. (2019) Functional genetic variants can mediate their regulatory effects through alteration of transcription factor binding, *Nat. Commun.*, **10**, 3472, <https://doi.org/10.1038/s41467-019-11412-5>.
8. Grodecká, L., Buratti, E., and Freiburger, T. (2017) Mutations of pre-mRNA splicing regulatory elements: Are predictions moving forward to clinical diagnostics? *Int. J. Mol. Sci.*, **18**, 1668, <https://doi.org/10.3390/ijms18081668>.
9. Andersson, R., and Sandelin, A. (2020) Determinants of enhancer and promoter activities of regulatory elements, *Nat. Rev. Genet.*, **21**, 71-87, <https://doi.org/10.1038/S41576-019-0173-8>.
10. Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A. M., Taylor, M. S., Engström, P. G., Frith, M. C., Forrest, A. R. R., Alkema, W. B., Tan, S. L., Plessy, C., Kodzius, R., Ravasi, T.,

- Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C. A., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. A., and Hayashizaki, Y. (2006) Genome-wide analysis of mammalian promoter architecture and evolution, *Nat. Genet.*, **38**, 626-635, <https://doi.org/10.1038/NG1789>.
11. Banerji, J., Rusconi, S., and Schaffner, W. (1981) Expression of a  $\beta$ -globin gene is enhanced by remote SV40 DNA sequences, *Cell*, **27**, 299-308, [https://doi.org/10.1016/0092-8674\(81\)90413-X](https://doi.org/10.1016/0092-8674(81)90413-X).
  12. Krivega, I., and Dean, A. (2012) Enhancer and promoter interactions-long distance calls, *Curr. Opin. Genet. Dev.*, **22**, 79-85, <https://doi.org/10.1016/j.gde.2011.11.001>.
  13. Korneev, K. V., Sviriaeva, E. N., Mitkin, N. A., Gorbacheva, A. M., Uvarova, A. N., Ustiugova, A. S., Polanovsky, O. L., Kulakovskiy, I. V., Afanasyeva, M. A., Schwartz, A. M., and Kuprash, D. V. (2020) Minor C allele of the SNP rs7873784 associated with rheumatoid arthritis and type-2 diabetes mellitus binds PU.1 and enhances TLR4 expression., *Biochim. Biophys. Acta Mol. Basis Dis.*, **1866**, 165626, <https://doi.org/10.1016/j.bbadis.2019.165626>.
  14. Panni, S., Lovering, R.C., Porras, P., and Orchard, S. (2020) Non-coding RNA regulatory networks, *Biochim. Biophys. Acta Gene Regul. Mech.*, **1863**, 194417, <https://doi.org/10.1016/j.bbagr.2019.194417>.
  15. Lappalainen, T., and MacArthur, D. G. (2021) From variant to function in human disease genetics, *Science*, **373**, 1464-1468, <https://doi.org/10.1126/science.abi8207>.
  16. Tseng, C. C., Wong, M. C., Liao, W. T., Chen, C. J., Lee, S. C., Yen, J. H., and Chang, S. J. (2021) Genetic variants in transcription factor binding sites in humans: triggered by natural selection and triggers of diseases, *Int. J. Mol. Sci.*, **22**, 4187, <https://doi.org/10.3390/ijms22084187>.
  17. Pan, X., Zhao, J., Zhou, Z., Chen, J., Yang, Z., Wu, Y., Bai, M., Jiao, Y., Yang, Y., Hu, X., Cheng, T., Lu, Q., Wang, B., Li, C. L., Lu, Y. J., Diao, L., Zhong, Y. Q., Pan, J., Zhu, J., Xiao, H. S., Qiu, Z. L., Li, J., Wang, Z., Hui, J., Bao, L., and Zhang, X. (2021) 5'-UTR SNP of *FGF13* causes translational defect and intellectual disability, *eLife*, **10**, e63021, <https://doi.org/10.7554/eLife.63021>.
  18. Cui, Y., Peng, F., Wang, D., Li, Y., Li, J. S., Li, L., and Li, W. (2022) 3'aQTL-atlas: An atlas of 3'UTR alternative polyadenylation quantitative trait loci across human normal tissues, *Nucleic Acids Res.*, **50**, D39-D45, <https://doi.org/10.1093/nar/gkab740>.
  19. Chhichholiya, Y., Suryan, A.K., Suman, P., Munshi, A., and Singh, S. (2021) SNPs in miRNAs and target sequences: role in cancer and diabetes, *Front. Genet.*, **12**, 793523, <https://doi.org/10.3389/fgene.2021.793523>.
  20. Hrdlickova, B., de Almeida, R.C., Borek, Z., and Withoff, S. (2014) Genetic variation in the non-coding genome: Involvement of micro-RNAs and long non-coding RNAs in disease, *Biochim. Biophys. Acta Mol. Basis Dis.*, **1842**, 1910-1922, <https://doi.org/10.1016/j.bbadis.2014.03.011>.
  21. Rykova, E., Ershov, N., Damarov, I., and Merkulova, T. (2022) SNPs in 3'UTR miRNA target sequences associated with individual drug susceptibility, *Int. J. Mol. Sci.*, **23**, 13725, <https://doi.org/10.3390/ijms232213725>.
  22. Feng, T., Feng, N., Zhu, T., Li, Q., Zhang, Q., Wang, Y., Gao, M., Zhou, B., Yu, H., Zheng, M., and Qian, B. (2020) A SNP-mediated lncRNA (LOC146880) and microRNA (miR-539-5p) interaction and its potential impact on the NSCLC risk, *J. Exp. Clin. Cancer Res.*, **39**, 1-12, <https://doi.org/10.1186/s13046-020-01652-5>.
  23. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F., and Guigó, R. (2021) Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome, *Nat. Commun.*, **12**, 727, <https://doi.org/10.1038/s41467-020-20578-2>.
  24. Degtyareva, A. O., Antontseva, E. V., and Merkulova, T. I. (2021) Regulatory snps: Altered transcription factor binding sites implicated in complex traits and diseases, *Int. J. Mol. Sci.*, **22**, 6454, <https://doi.org/10.3390/ijms22126454>.
  25. Gorbacheva, A. M., Korneev, K. V., Kuprash, D. V., and Mitkin, N. A. (2018) The risk G allele of the single-nucleotide polymorphism rs928413 creates a CREB1-binding site that activates IL33 promoter in lung epithelial cells, *Int. J. Mol. Sci.*, **19**, 2911, <https://doi.org/10.3390/ijms19102911>.
  26. Putlyaeva, L. V., Demin, D. E., Korneev, K. V., Kasyanov, A. S., Tatosyan, K. A., Kulakovskiy, I. V., Kuprash, D. V., and Schwartz, A. M. (2018) Potential markers of autoimmune diseases, alleles rs115662534(T) and rs548231435(C), disrupt the binding of transcription factors STAT1 and EBF1 to the regulatory elements of human CD40 gene, *Biochemistry (Moscow)*, **83**, 1534-1542, <https://doi.org/10.1134/S0006297918120118>.
  27. Zhou, J., To, K. K. W., Dong, H., Cheng, Z. S., Lau, C. C. Y., Poon, V. K. M., Fan, Y. H., Song, Y. Q., Tse, H., Chan, K. H., Zheng, B. J., Zhao, G. P., and Yuen, K. Y. (2012) A functional variation in CD55 increases the severity of 2009 pandemic H1N1 influenza A virus infection, *J. Infect. Dis.*, **206**, 495-503, <https://doi.org/10.1093/infdis/jis378>.
  28. Matveeva, M. Y., Kashina, E. V., Reshetnikov, V. V., Bryzgalov, L. O., Antontseva, E. V., Bondar, N. P., and Merkulova, T. I. (2016) Regulatory single nucleotide polymorphisms (rSNPs) at the promoters 1A and 1B of the human APC gene, *BMC Genet.*, **17**, 127-135, <https://doi.org/10.1186/s12863-016-0460-8>.

29. Mitkin, N. A., Muratova, A. M., Korneev, K. V., Pavshintsev, V. V., Rumyantsev, K. A., Vagida, M. S., Uvarova, A. N., Afanasyeva, M. A., Schwartz, A. M., and Kuprash, D. V. (2018) Protective C allele of the single-nucleotide polymorphism rs1335532 is associated with strong binding of Ascl2 transcription factor and elevated CD58 expression in B-cells, *Biochim. Biophys. Acta Mol. Basis Dis.*, **1864**, 3211-3220, <https://doi.org/10.1016/j.bbadis.2018.07.008>.
30. Uvarova, A. N., Ustiugova, A. S., Mitkin, N. A., Schwartz, A. M., Korneev, K. V., and Kuprash, D. V. (2022) The minor T allele of the single nucleotide polymorphism rs13360222 decreases the activity of the HAVCR2 gene enhancer in a cell model of human macrophages, *Mol. Biol.*, **56**, 90-96, <https://doi.org/10.1134/S0026893322010095>.
31. Uvarova, A. N., Stasevich, E. M., Ustiugova, A. S., Mitkin, N. A., Zheremyan, E. A., Sheetikov, S. A., Zornikova, K. V., Bogolyubova, A. V., Rubtsov, M. A., Kulakovskiy, I. V., Kuprash, D. V., Korneev, K. V., and Schwartz, A. M. (2023) rs71327024 Associated with COVID-19 hospitalization reduces CXCR6 promoter activity in human CD4<sup>+</sup> T cells via disruption of c-Myb binding, *Int. J. Mol. Sci.*, **24**, 13790, <https://doi.org/10.3390/IJMS241813790>.
32. Ustiugova, A. S., Korneev, K. V., Kuprash, D. V., and Afanasyeva, M. A. (2019) Functional SNPs in the human autoimmunity-associated locus 17q12-21, *Genes*, **10**, 77, <https://doi.org/10.3390/GENES10020077>.
33. Cooper, T. A. (2005) Use of minigene systems to dissect alternative splicing elements, *Methods*, **37**, 331-340, <https://doi.org/10.1016/J.YMETH.2005.07.015>.
34. Sparber, P., Sharova, M., Davydenko, K., Pyankov, D., Filatova, A., and Skoblov, M. (2023) Deciphering the impact of coding and non-coding SCN1A gene variants on RNA splicing, *Brain*, **147**, 1278-1293, <https://doi.org/10.1093/BRAIN/AWAD383>.
35. Sanoguera-Miralles, L., Bueno-Martínez, E., Valenzuela-Palomo, A., Esteban-Sánchez, A., Llinares-Burguet, I., Pérez-Segura, P., García-Álvarez, A., de la Hoya, M., and Velasco-Sampedro, E. A. (2022) Minigene splicing assays identify 20 spliceogenic variants of the breast/ovarian cancer susceptibility gene RAD51C, *Cancers*, **14**, 2960, <https://doi.org/10.3390/CANCERS14122960/S1>.
36. Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., and Gray, J. M. (2016) High-throughput functional comparison of promoter and enhancer activities, *Genome Res.*, **26**, 1023-1033, <https://doi.org/10.1101/GR.204834.116>.
37. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C. G., Kinney, J. B., Kellis, M., Lander, E. S., and Mikkelsen, T. S. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay, *Nat. Biotechnol.*, **30**, 271-277, <https://doi.org/10.1038/nbt.2137>.
38. Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. (2016) Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay, *Cell*, **172**, 1519-1529, <https://doi.org/10.1016/j.cell.2018.02.021>.
39. Myint, L., Wang, R., Boukas, L., Hansen, K. D., Goff, L. A., and Avramopoulos, D. (2020) A screen of 1,049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential, *Am. J. Med. Genet. Part B Neuropsychiatr. Genet.*, **183**, 61-73, <https://doi.org/10.1002/AJMG.B.32761>.
40. Sample, P. J., Wang, B., Reid, D. W., Presnyak, V., McFadyen, I. J., Morris, D. R., and Seelig, G. (2019) Human 5' UTR design and variant effect prediction from a massively parallel translation assay, *Nat. Biotechnol.*, **37**, 803-809, <https://doi.org/10.1038/s41587-019-0164-5>.
41. Griesemer, D., Xue, J. R., Reilly, S. K., Ulirsch, J. C., Kukreja, K., Davis, J. R., Kanai, M., Yang, D. K., Butts, J. C., Guney, M. H., Luban, J., Montgomery, S. B., Finucane, H. K., Novina, C. D., Tewhey, R., and Sabeti, P. C. (2021) Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution, *Cell*, **184**, 5247-5260, <https://doi.org/10.1016/j.cell.2021.08.025>.
42. Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M., and Kellis, M. (2018) High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human, *Nat. Commun.*, **9**, 5380, <https://doi.org/10.1038/s41467-018-07746-1>.
43. Van Arensbergen, J., Pagie, L., FitzPatrick, V. D., de Haas, M., Baltissen, M. P., Comoglio, F., van der Weide, R. H., Teunissen, H., Vösa, U., Franke, L., de Wit, E., Vermeulen, M., Bussemaker, H. J., and van Steensel, B. (2019) High-throughput identification of human SNPs affecting regulatory element activity, *Nat. Genet.*, **51**, 1160-1169, <https://doi.org/10.1038/s41588-019-0455-2>.
44. Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., and Stark, A. (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq, *Science*, **339**, 1074-1077, [https://doi.org/10.1126/SCIENCE.1232542/SUPPL\\_FILE/ARNOLD.SM.PDF](https://doi.org/10.1126/SCIENCE.1232542/SUPPL_FILE/ARNOLD.SM.PDF).
45. Ustiugova, A. S., Dvorianinova, E. M., Melnikova, N. V., Dmitriev, A. A., Kuprash, D. V., and Afanasyeva, M. A. (2023) CRISPR/Cas9 genome editing demonstrates functionality of the autoimmunity-associated SNP rs12946510, *Biochim. Biophys. Acta Mol. Basis Dis.*, **1869**, 166599, <https://doi.org/10.1016/j.bbadis.2022.166599>.

46. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., and Kellis, M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions, *Nat. Biotechnol.*, **34**, 1180-1190, <https://doi.org/10.1038/nbt.3678>.
47. Soemedi, R., Cygan, K. J., Rhine, C. L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W. G. (2017) Pathogenic variants that alter protein code often disrupt splicing, *Nat. Genet.*, **49**, 848-855, <https://doi.org/10.1038/ng.3837>.
48. Rhine, C. L., Neil, C., Wang, J., Maguire, S., Buerer, L., Salomon, M., Meremikwu, I. C., Kim, J., Strande, N. T., and Fairbrother, W. G. (2022) Massively parallel reporter assays discover *de novo* exonic splicing mutants in paralogs of Autism genes, *PLoS Genet.*, **18**, e1009884, <https://doi.org/10.1371/journal.pgen.1009884>.
49. Lagunas, T., Plassmeyer, S. P., Fischer, A. D., Friedman, R. Z., Rieger, M. A., Selmanovic, D., Sarafinowska, S., Sol, Y. K., Kasper, M. J., Fass, S. B., Aguilar Lucero, A. F., An, J. Y., Sanders, S. J., Cohen, B. A., and Dougherty, J. D. (2023) A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of non-coding elements *in vivo*, *Commun. Biol.*, **6**, 1-14, <https://doi.org/10.1038/s42003-023-05483-w>.
50. Gordon, M. G., Inoue, F., Martin, B., Schubach, M., Agarwal, V., Whalen, S., Feng, S., Zhao, J., Ashuach, T., Ziffra, R., Kreimer, A., Georgakopoulos-Soares, I., Yosef, N., Ye, C. J., Pollard, K. S., Shendure, J., Kircher, M., and Ahituv, N. (2020) lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements, *Nat. Protoc.*, **15**, 2387-2412, <https://doi.org/10.1038/s41596-020-0333-5>.
51. GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues, *Science*, **369**, 1318-1330, <https://doi.org/10.1126/science.aaz1776>.
52. Bryois, J., Calini, D., Macnair, W., Foo, L., Urich, E., Ortmann, W., Iglesias, V. A., Selvaraj, S., Nutma, E., Marzin, M., Amor, S., Williams, A., Castelo-Branco, G., Menon, V., De Jager, P., and Malhotra, D. (2022) Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders, *Nat. Neurosci.*, **25**, 1104-1112, <https://doi.org/10.1038/s41593-022-01128-z>.
53. Capurso, D., Tang, Z., and Ruan, Y. (2020) Methods for comparative ChIA-PET and Hi-C data analysis, *Methods*, **170**, 69-74, <https://doi.org/10.1016/j.ymeth.2019.09.019>.
54. Huang, L., Yang, Y., Li, G., Jiang, M., Wen, J., Abnoui, A., Rosen, J. D., Hu, M., and Li, Y. (2022) A systematic evaluation of Hi-C data enhancement methods for enhancing PLAC-seq and HiChIP data, *Brief. Bioinform.*, **23**, 1-14, <https://doi.org/10.1093/BIB/BBAC145>.
55. Khalil, A.M. (2020) The genome editing revolution: review, *J. Genet. Eng. Biotechnol.*, **18**, 68, <https://doi.org/10.1186/S43141-020-00078-Y>.
56. Moon, S. B., Kim, D. Y., Ko, J. H., and Kim, Y. S. (2019) Recent advances in the CRISPR genome editing tool set, *Exp. Mol. Med.*, **51**, 1-11, <https://doi.org/10.1038/s12276-019-0339-7>.
57. Zhang, P., Xia, J. H., Zhu, J., Gao, P., Tian, Y. J., Du, M., Guo, Y. C., Suleman, S., Zhang, Q., Kohli, M., Tillmans, L. S., Thibodeau, S. N., French, A. J., Cerhan, J. R., Wang, L. D., Wei, G. H., and Wang, L. (2018) High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing, *Nat. Commun.*, **9**, 1-12, <https://doi.org/10.1038/s41467-018-04451-x>.
58. Rodríguez-Rodríguez, D. R., Ramírez-Solís, R., Garza-Elizondo, M. A., Garza-Rodríguez, M. D. L., and Barrera-Saldaña, H. A. (2019) Genome editing: a perspective on the application of CRISPR/Cas9 to study human diseases (Review), *Int. J. Mol. Med.*, **43**, 1559-1574, <https://doi.org/10.3892/ijmm.2019.4112>.
59. Yang, H., Ren, S., Yu, S., Pan, H., Li, T., Ge, S., Zhang, J., and Xia, N. (2020) Methods favoring homology-directed repair choice in response to CRISPR/Cas9 Induced-double strand breaks, *Int. J. Mol. Sci.*, **21**, 6461, <https://doi.org/10.3390/IJMS21186461>.
60. Rees, H. A., and Liu, D. R. (2018) Base editing: precision chemistry on the genome and transcriptome of living cells, *Nat. Rev. Genet.*, **19**, 770-788, <https://doi.org/10.1038/s41576-018-0059-1>.
61. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., and Liu, D. R. (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage, *Nature*, **533**, 420-424, <https://doi.org/10.1038/nature17946>.
62. Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I., and Liu, D. R. (2017) Programmable base editing of A·T to G·C in genomic DNA without DNA cleavage, *Nature*, **551**, 464-471, <https://doi.org/10.1038/nature24644>.
63. Zhao, D., Li, J., Li, S., Xin, X., Hu, M., Price, M. A., Rosser, S. J., Bi, C., and Zhang, X. (2021) Glycosylase base editors enable C-to-A and C-to-G base changes, *Nat. Biotechnol.*, **39**, 35-40, <https://doi.org/10.1038/s41587-020-0592-2>.
64. Weng, N., Miller, M., Pham, A. K., Komor, A. C., and Broide, D. H. (2022) Single-base editing of rs12603332 on chromosome 17q21 with a cytosine base editor regulates ORMDL3 and ATF6a expression, *Allergy*, **77**, 1139-1149, <https://doi.org/10.1111/ALL.15092>.
65. Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., and Liu, D. R. (2019) Search-and-replace genome editing without double-strand breaks or donor DNA, *Nature*, **576**, 149-157, <https://doi.org/10.1038/s41586-019-1711-4>.

66. Jiang, Y., Chai, Y., Qiao, D., Wang, J., Xin, C., Sun, W., Cao, Z., Zhang, Y., Zhou, Y., Wang, X. C., and Chen, Q. J. (2022) Optimized prime editing efficiently generates glyphosate-resistant rice plants carrying homozygous TAP-IVS mutation in EPSPS, *Mol. Plant*, **15**, 1646-1649, <https://doi.org/10.1016/j.molp.2022.09.006>.
67. Hassan, M. M., Yuan, G., Chen, J. G., Tuskan, G. A., and Yang, X. (2020) Prime editing technology and its prospects for future applications in plant biology research, *BioDes. Res.*, **2020**, 9350905, <https://doi.org/10.34133/2020/9350905>.
68. Gao, P., Lyu, Q., Ghanam, A. R., Lazzarotto, C. R., Newby, G. A., Zhang, W., Choi, M., Slivano, O. J., Holden, K., Walker, J. A., Kadina, A. P., Munroe, R. J., Abratte, C. M., Schimenti, J. C., Liu, D. R., Tsai, S. Q., Long, X., and Miano, J. M. (2021) Prime editing in mice reveals the essentiality of a single base in driving tissue-specific gene expression, *Genome Biol.*, **22**, 1-21, <https://doi.org/10.1186/s13059-021-02304-3>.
69. Godbout, K., Rousseau, J., and Tremblay, J. P. (2023) Successful correction by prime editing of a mutation in the *RYR1* gene responsible for a myopathy, *Cells*, **13**, 31, <https://doi.org/10.3390/CELLS13010031>.
70. Petrova, I. O., and Smirnikhina, S. A. (2023) The development, optimization and future of prime editing, *Int. J. Mol. Sci.*, **24**, 17045, <https://doi.org/10.3390/IJMS242317045>.
71. Ren, X., Yang, H., Nierenberg, J. L., Sun, Y., Chen, J., Beaman, C., Pham, T., Nobuhara, M., Takagi, M. A., Narayan, V., Li, Y., Ziv, E., and Shen, Y. (2023) High-throughput PRIME-editing screens identify functional DNA variants in the human genome, *Mol. Cell*, **83**, 4633-4645.e9, <https://doi.org/10.1016/J.MOLCEL.2023.11.021>.
72. Ambrosini, G., Vorontsov, I., Penzar, D., Groux, R., Fornes, O., Nikolaeva, D. D., Ballester, B., Grau, J., Grosse, I., Makeev, V., Kulakovskiy, I., and Bucher, P. (2020) Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study, *Genome Biol.*, **21**, 1-18, <https://doi.org/10.1186/s13059-020-01996-3>.
73. Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018) The human transcription factors, *Cell*, **172**, 650-665, <https://doi.org/10.1016/J.CELL.2018.01.029>.
74. Tognon, M., Giugno, R., and Pinello, L. (2023) A survey on algorithms to characterize transcription factor binding sites, *Brief Bioinform.*, **24**, bbad156, <https://doi.org/10.1093/bib/bbad156>.
75. Mundade, R., Ozer, H. G., Wei, H., Prabhu, L., and Lu, T. (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond, *Cell Cycle*, **13**, 2847-2852, <https://doi.org/10.4161/15384101.2014.949201>.
76. Vorontsov, I. E., Kulakovskiy, I. V., Khimulya, G., Nikolaeva, D. D., and Makeev, V. J. (2015) PERFECTOS-APE: Predicting regulatory functional effect of SNPs by approximate P-value estimation, *Bioinforma. 2015 – 6th Int. Conf. Bioinforma. Model. Methods Algorithms, Proceedings; Part 8th Int. Jt. Conf. Biomed. Eng. Syst. Technol., BIOSTEC 2015*, **2**, 102-108, <https://doi.org/10.5220/0005189301020108>.
77. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S., and Urbach, S. (2001) The TRANSFAC system on gene expression regulation, *Nucleic Acids Res.*, **29**, 281-283, <https://doi.org/10.1093/nar/29.1.281>.
78. Vorontsov, I. E., Eliseeva, I. A., Zinkevich, A., Nikonov, M., Abramov, S., Boytsov, A., Kamenets, V., Kasianova, A., Kolmykov, S., Yevshin, I. S., Favorov, A., Medvedeva, Y. A., Jolma, A., Kolpakov, F., Makeev, V. J., and Kulakovskiy, I. V. (2024) HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors, *Nucleic Acids Res.*, **52**, D154-D163, <https://doi.org/10.1093/NAR/GKAD1077>.
79. Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Perez, N. M., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., Vandepoele, K., Wasserman, W. W., Parcy, F., and Mathelier, A. (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles, *Nucleic Acids Res.*, **50**, D165-D173, <https://doi.org/10.1093/NAR/GKAB1113>.
80. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities, *Mol. Cell*, **38**, 576-589, <https://doi.org/10.1016/j.molcel.2010.05.004>.
81. Janky, R., Verfaillie, A., Imrichová, H., van de Sande, B., Standaert, L., Christiaens, V., Hulselmans, G., Hertens, K., Naval Sanchez, M., Potier, D., Svetlichnyy, D., Kalender Atak, Z., Fiers, M., Marine, J. C., and Aerts, S. (2014) iRegulon: from a gene list to a gene regulatory network using large motif and track collections, *PLoS Comput. Biol.*, **10**, e1003731, <https://doi.org/10.1371/journal.pcbi.1003731>.
82. Abramov, S., Boytsov, A., Bykova, D., Penzar, D. D., Yevshin, I., Kolmykov, S. K., Fridman, M. V., Favorov, A. V., Vorontsov, I. E., Baulin, E., Kolpakov, F., Makeev, V. J., and Kulakovskiy, I. V. (2021) Landscape of allele-specific transcription factor binding in the human genome, *Nat. Commun.*, **12**, 2751, <https://doi.org/10.1038/s41467-021-23007-0>.
83. Li, Y., Zhang, X. O., Liu, Y., and Lu, A. (2023) Allele-specific binding (ASB) analyzer for annotation of allele-specific binding SNPs, *BMC Bioinform.*, **24**, 464, <https://doi.org/10.1186/S12859-023-05604-6>.

84. Yan, J., Qiu, Y., Ribeiro dos Santos, A. M., Yin, Y., Li, Y. E., Vinckier, N., Nariai, N., Benaglio, P., Raman, A., Li, X., Fan, S., Chiou, J., Chen, F., Frazer, K. A., Gaulton, K. J., Sander, M., Taipale, J., and Ren, B. (2021) Systematic analysis of binding of transcription factors to noncoding variants, *Nature*, **591**, 147-151, <https://doi.org/10.1038/s41586-021-03211-0>.
85. Hellman, L. M., and Fried, M. G. (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions, *Nat. Protoc.*, **2**, 1849-1861, <https://doi.org/10.1038/nprot.2007.249>.
86. Parés-Matos, E. I. (2013) Electrophoretic mobility-shift and super-shift assays for studies and characterization of protein-DNA complexes, *Methods Mol. Biol.*, **977**, 159-167, [https://doi.org/10.1007/978-1-62703-284-1\\_12](https://doi.org/10.1007/978-1-62703-284-1_12).
87. Zhao, Y., Wu, D., Jiang, D., Zhang, X., Wu, T., Cui, J., Qian, M., Zhao, J., Oesterreich, S., Sun, W., Finkel, T., and Li, G. (2020) A sequential methodology for the rapid identification and characterization of breast cancer-associated functional SNPs, *Nat. Commun.*, **11**, 1-11, <https://doi.org/10.1038/s41467-020-17159-8>.
88. Butter, F., Davison, L., Viturawong, T., Scheibe, M., Vermeulen, M., Todd, J. A., and Mann, M. (2012) Proteome-wide analysis of disease-associated SNPs that show allele-specific transcription factor binding, *PLoS Genet.*, **8**, e1002982, <https://doi.org/10.1371/journal.pgen.1002982>.
89. Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E., and Taipale, J. (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities, *Genome Res.*, **20**, 861-873, <https://doi.org/10.1101/gr100552.109>.
90. Mille, M., Ripoll, J., Cazaux, B., and Rivals, E. (2023) dipwmsearch: a Python package for searching di-PWM motifs, *Bioinformatics*, **39**, btad141, <https://doi.org/10.1093/BIOINFORMATICS/BTAD141>.
91. Maurano, M. T., Haugen, E., Sandstrom, R., Vierstra, J., Shafer, A., Kaul, R., and Stamatoyannopoulos, J. A. (2015) Large-scale identification of sequence variants influencing human transcription factor occupancy *in vivo*, *Nat. Genet.*, **47**, 1393-1401, <https://doi.org/10.1038/ng.3432>.
92. Boytsov, A., Abramov, S., Aiusheeva, A. Z., Kasianova, A. M., Baulin, E., Kuznetsov, I. A., Aulchenko, Y. S., Kolmykov, S., Yevshin, I., Kolpakov, F., Vorontsov, I. E., Makeev, V. J., and Kulakovskiy, I. V (2022) ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs, *Nucleic Acids Res.*, **50**, W51-W56, <https://doi.org/10.1093/nar/gkac262>.
93. Mitkin, N. A., Korneev, K. V., Gorbacheva, A. M., and Kuprash, D. V. (2019) Relative efficiency of transcription factor binding to allelic variants of regulatory regions of human genes in immunoprecipitation and real-time PCR, *Mol. Biol.*, **53**, 346-353, <https://doi.org/10.1134/S0026893319030117>.
94. Yevshin, I., Sharipov, R., Valeev, T., Kel, A., and Kolpakov, F. (2017) GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments, *Nucleic Acids Res.*, **45**, D61-D67, <https://doi.org/10.1093/nar/gkw951>.
95. Zhang, Y., Mo, Q., Xue, L., and Luo, J. (2021) Evaluation of deep learning approaches for modeling transcription factor sequence specificity, *Genomics*, **113**, 3774-3781, <https://doi.org/10.1016/J.YGENO.2021.09.009>.
96. Chen, C., Hou, J., Shi, X., Yang, H., Birchler, J. A., and Cheng, J. (2021) DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks, *BMC Bioinformatics*, **22**, 1-18, <https://doi.org/10.1186/S12859-020-03952-1>.

## METHODS FOR FUNCTIONAL CHARACTERIZATION OF GENETIC POLYMORPHISMS OF NON-CODING REGULATORY REGIONS OF THE HUMAN GENOME

### Review

**A. N. Uvarova<sup>1\*</sup>, E. A. Tkachenko<sup>2,3</sup>, E. M. Stasevich<sup>1,4</sup>, E. A. Zheremyan<sup>1</sup>,  
K. V. Korneev<sup>1</sup>, and D. V. Kuprash<sup>1,3</sup>**

<sup>1</sup> Center for Precision Genome Editing and Genetic Technologies for Biomedicine,  
Engelhardt Institute of Molecular Biology, Russian Academy of Sciences,  
119991 Moscow, Russia; e-mail: uvarowww@gmail.com

<sup>2</sup> Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 119991 Moscow, Russia

<sup>3</sup> Faculty of Biology, Lomonosov Moscow State University, 119234 Moscow, Russia

<sup>4</sup> Moscow Institute of Physics and Technology, 141700 Dolgoprudny, Moscow Region, Russia

Currently, numerous associations between genetic polymorphisms and various diseases have been characterized through Genome-Wide Association Studies. The majority of clinically significant polymorphisms are localized in non-coding regions of the genome. While modern bioinformatic resources make it possible to predict molecular mechanisms that explain the influence of non-coding polymorphisms on gene expression, such hypotheses require experimental verification. This review discusses the methods for elucidating the molecular mechanisms underlying the dependence of disease pathogenesis on specific genetic variants within non-coding sequences. A particular focus is on the methods to identify the transcription factors with binding efficiency contingent upon polymorphic variations. Despite remarkable progress in the bioinformatic resources enabling the prediction of the impact of polymorphisms on disease pathogenesis, the imperative for experimental approaches to this inquiry still persists.

*Keywords:* functional polymorphisms, regulatory genomic regions, transcription factors, reporter analysis, CRISPR-Cas