



DOI: 10.18413/2658-6533-2020-7-1-0-2

УДК 616-056.7:575

Биоинформатические инструменты и интернет-ресурсы для оценки регуляторного потенциала полиморфных локусов, установленных полногеномными ассоциативными исследованиями мультифакториальных заболеваний (обзор)

А.В. Полоников , Е.Ю. Клёсова , Ю.Э. Азарова 

Федеральное государственное бюджетное образовательное учреждение высшего образования «Курский государственный медицинский университет», ул. Карла Маркса, д. 3, г. Курск, 305041, Российская Федерация
Автор для переписки: Е.Ю. Клёсова (*ecless@yandex.ru*)

Резюме




Актуальность: Полногеномные ассоциативные исследования (genome-wide association studies, GWAS) представляют собой разновидность генетических исследований, целью которых является анализ ассоциаций между геномными вариантами и фенотипическими признаками в популяции. За последние 12 лет было установлено более 60 тысяч ассоциаций между тремя миллионами однонуклеотидных полиморфных вариантов (SNPs) и 829 заболеваниями. Тем не менее, несмотря на достигнутые успехи, большую проблему представляет вопрос патогенетической интерпретации полученных данных, поскольку абсолютное большинство локусов находятся в межгенных областях и некодирующих последовательностях генома. **Цель исследования:** Изучить возможности существующих биоинформатических инструментов, позволяющих оценить возможные фенотипические эффекты SNPs на определенные молекулярные функции и биологические процессы, а также имеющие патогенетическое значение для развития мультифакториальных заболеваний. **Материалы и методы:** Проведен анализ российской и зарубежной научной литературы по биоинформатическим методам анализа и интернет-ресурсам, необходимым для оценки регуляторного потенциала полиморфных локусов, установленных в полногеномных ассоциативных исследованиях мультифакториальных заболеваний. **Результаты:** В обзоре представлены основные итоги изучения спектра применения баз данных и интернет ресурсов для оценки влияния вариантов ДНК на экспрессию генов в различных тканях, метилирование ДНК, характеристики метаболомного профиля, рассмотрены алгоритмические подходы, систематизированы качественные и количественные online-инструменты, а также вычислительные методы. **Заключение:** Полногеномные ассоциативные исследова-

ния открыли новую эру в истории генетических исследований мультифакториальных заболеваний. Биоинформатический анализ *in silico* позволяет дать всестороннюю оценку эффектам SNPs и их роли в развитии того или иного фенотипического признака болезни.

Ключевые слова: ДНК-полиморфизмы; полногеномные ассоциативные исследования; мультифакториальные заболевания; биоинформатические инструменты

Для цитирования: Полоников АВ, Клёсова ЕЮ, Азарова ЮЭ. Биоинформатические инструменты и интернет-ресурсы для оценки регуляторного потенциала полиморфных локусов, установленных полногеномными ассоциативными исследованиями мультифакториальных заболеваний (обзор). Научные результаты биомедицинских исследований. 2021;7(1):15-31. DOI: 10.18413/2658-6533-2020-7-1-0-2

Bioinformatic tools and internet resources for functional annotation of polymorphic loci detected by genome wide association studies of multifactorial diseases (review)

Alexey V. Polonikov , Elena Yu. Klysova , Iuliia E. Azarova 

Kursk State Medical University,
3 Karl Marx St., Kursk, 305041, Russia

Corresponding author: Elena Yu. Klysova (*ecless@yandex.ru*)

Abstract

Background: Genome-wide association studies (GWAS) are a type of genetic research whose purpose is to analyze the associations between genomic variants and phenotypic traits in a population. Over the past 12 years, more than 60000 associations have been established between three million single nucleotide variants (SNPs) and 829 diseases, however, despite the progress achieved, the pathogenetic interpretation of the data is a huge problem, since the vast majority of the loci are located in intergenic regions and non-coding sequences of the genome, or in genes that are not related to metabolic pathways involved in the development of a particular pathology. In this regard, the integrated usage of bioinformatic tools gives an opportunity to evaluate the possible effects of SNPs on certain molecular functions and biological processes related to disease pathogenesis. **The aim of the study:** To examine the capabilities of existing bioinformatics tools to evaluate possible phenotypic effects of SNPs on certain molecular functions and biological processes, as well as having pathogenetic significance for the development of multifactorial diseases. **Materials and methods:** The authors carried out an analysis of the Russian and foreign scientific literature on bioinformatic methods of analysis and Internet resources necessary for the assessment of the regulatory potential of polymorphic loci established in genome-wide associative studies of multifactorial diseases. **Results:** The review presents the main results of studying the spectrum of application of databases and Inter-

net resources for assessing the effect of DNA variants on gene expression in various tissues, DNA methylation, and characteristics of the metabolomic profile. **Conclusion:** Genome-wide associative research has opened a new era in the history of genetic research on multifactorial diseases. In silico bioinformatics analysis provides a comprehensive assessment of the effects of SNPs and their role in the development of a phenotypic trait of disease.

Keywords: DNA polymorphisms; genome-wide association studies; multifactorial diseases; bioinformatics tools

For citation: Polonikov AV, Klyosova EYu, Azarova IE. Bioinformatic tools and internet resources for functional annotation of polymorphic loci detected by genome wide association studies of multifactorial diseases (review). Research Results in Biomedicine. 2021;7(1):15-31. Russian. DOI: 10.18413/2658-6533-2020-7-1-0-2

Введение. Полногеномные ассоциативные исследования (genome-wide association studies, GWAS) представляют собой разновидность генетических исследований, целью которых является анализ ассоциаций между геномными вариантами и фенотипическими признаками в популяции [1, 2]. На сегодняшний день GWAS стали флагманом медицинской геномики открытию новых генов, которые вносят вклад в развитие полигенных мультифакториальных заболеваний. Главная цель этих исследований – формирование более глубокого понимания фундаментальных биологических основ болезни, что *априори* должно способствовать разработке и внедрению более эффективных способов профилактики и лечения заболеваний. Фактически при GWAS сравнивают геномы группы больных людей с геномами контрольной группы, включающей в себя аналогичных по возрасту, полу и другим признакам здоровых людей [1]. GWAS выявляют ассоциации конкретных локусов генома с признаками или заболеваниями с использованием набора однонуклеотидных полиморфизмов или SNP, максимально покрывающих геном и маркирующих блоки взаимосвязанных частых SNPs (*tagSNP*). Современные чипы для GWAS содержат 300000–5000000 *tagSNPs* с максимально возможным покрытием генома. Обнаружение ассоциации *tagSNP* с фенотипом означает, что один или несколько маркированных им коррелирующихся SNP должны контролировать биологические

функции, которые лежат в основе выявленной ассоциации [3]. Однако трансляция накопленных в результате GWAS генофенотипических ассоциаций с позиции патогенеза существенно осложняется тем, что факт статистической ассоциации полиморфного варианта в геноме с каким-либо признаком не всегда является прямым свидетельством влияния гена и отражает механизмы формирования патологического процесса.

Для понимания патофизиологической взаимосвязи генетического варианта с фенотипом болезни ключевым моментом является структурно-функциональная характеристика и классификация полиморфных молекулы ДНК. Выделяют 4 основных класса полиморфных вариантов ДНК: 1) однонуклеотидные варианты (SNVs или SNPs): их число колеблется по разным данным от 3750000 до 4500000 [4]; 2) короткие (<50 нуклеотидов) инсерции и делеции (InDels): их насчитывается от 700000 до 1000000 [4]; 3) варианты числа копий (CNVs) – в основном представлены тандемными дупликациями и составляют 5-10% генома [5]; 4) структурные варианты (SVs) – составляют в среднем 13% генома [6]. Каждый вариант ДНК характеризуется собственным фенотипическим эффектом. Одни из таких вариантов хорошо изучены, другие – нет. SNPs, локализованные в области дистального промотора или 5'-нетранслируемого региона (5'-UTR), способны влиять на экспрессию гена, в котором они находятся (цис-эффекты), а также и близлежащих генов (транс-

эффекты) [7]. SNPs, находящиеся в консенсусных последовательностях и регуляторных элементах (энхансерах, сайленсерах) могут влиять на сборку сплайсосомы и процесс сплайсинга [8, 9]. Варианты в 3'-нетранслируемой области (3'-UTR) и полиА-сигнальной последовательности (AATAAA) способны снижать стабильность мРНК и уменьшать ее количество, и, следовательно, и количество синтезируемого белкового продукта. Варианты, расположенные в старт- и стоп-кодонах могут существенно изменять качество трансляции, приводя к образованию более длинного или более короткого полипептида с более низкой термодинамической стабильностью по сравнению с нормальным белком [10]. Описаны необычные эффекты вариантов в межгенных элементах, способных включать или выключать экспрессию удаленных генов путем изменения статуса их метилирования.

В таблице 1 суммированы основные достижения в медицинской генетике, полученные в результате полногеномных ас-

социативных исследований [1]. С помощью GWAS установлено более 10,000 надежных ассоциаций SNPs с фенотипами и болезнями. В частности, была установлена выраженная вариация показателя неравновесия по сцеплению (LD) в геноме. Выявлено, что большая доля генетической изменчивости, детерминирующей полигенные признаки и болезни, связана с влиянием частых SNPs. Был подтвержден феномен плейотропии: многие SNPs одновременно влияют на множество признаков. GWAS позволили подтвердить причинно-следственные связи и доказать ложноположительных взаимосвязи генотипов и фенотипов. С помощью GWAS было показано, что генетическая структура может имитировать географическую структуру и представлены доказательства действия естественного отбора. Сочетание GWAS с омиксными технологиями позволило установить, что 2/3 GWAS-ассоциированных SNPs расположены в гене, который не является ближайшим геном к наиболее ассоциированному SNP.

Таблица 1

Достижения GWAS в медицинской генетике

Table 1

GWAS achievements in medical genetics

Анализ	Цель	Открытия
Полногеномный анализ ассоциации (GWAS)	Обнаружение ассоциаций SNP с фенотипами/болезнями	Установлено более 10,000 надежных ассоциаций SNPs с фенотипами и болезнями
Полногеномная оценка неравновесия по сцеплению (linkage disequilibrium, LD)	Количественная оценка архитектуры генома	Выраженная вариация показателя LD в геноме, в том числе и межпопуляционная
Оценка наследуемости SNP	Генетическая архитектура	Большая доля генетической изменчивости связана с частыми SNPs
Оценка генетической корреляции	Обнаружение и количественная оценка плейотропии	Плейотропия вездесуща (многие SNPs влияют на множество признаков)
Полигенные оценки риска (Polygenic risk scores)	Обнаружение плейотропии; валидация результатов GWAS	Обнаружение ассоциаций с новыми признаками. Подтверждение прогнозирования на независимых выборках.
Менделевская рандомизация (Mendelian randomization)	Тестирование причинно-следственных связей	Подтверждение известных причинно-следственных связей; эмпирическое доказательство ложноположительных взаимосвязей.
Популяционные различия в частотах аллелей	Реконструкция истории народонаселения; обнаружение естественного отбора	Генетическая структура может имитировать географическую структуру; доказательство действия естественного отбора
Сочетание GWAS с omics-технологиями	Точное картирование; обнаружение генов-мишеней; функции генов	Две трети GWAS-ассоциированных SNPs расположены в гене, который не является ближайшим геном к наиболее ассоциированному SNP

На май 2018 года каталог полногеномных ассоциативных исследований (GWAS catalog, <https://www.ebi.ac.uk/gwas/home>) включал более 69 млн. ассоциаций SNP с фенотипами/болезнями, обнаруженными более чем в 5000 работ и опубликованными в 3378 научных статьях. На март 2019 года база GWAS central (<https://www.gwascentral.org>) включала 69 986 326 ассоциаций между 2974967 уникальными SNP и 829 уникальными болезнями/фенотипами.

Несмотря на достигнутые успехи, GWAS столкнулись с очень серьезной проблемой – сложностью патофизиологической интерпретации выявленных гено-фенотипических ассоциаций. Как правило, связь между генетическим вариантом и признаком не дает непосредственной информации о гене-мишени или механизме, посредством которого данный вариант связан с фенотипом [11]. Проблема заключается в том, что наборы tagSNP, анализируемые GWAS, несмотря на максимально возможное покрытие генома, зачастую не являются причинами изучаемого заболевания. Причинный SNP может находиться где угодно в пределах гаплотипического блока, который может охватывать более 100 kb и часто содержат более 1000 отдельных SNPs [11].

Некоторые значимые ассоциации в системе полиморфизм – фенотип – метаболический путь могут быть обнаружены при рассмотрении дистантных по отношению к данному SNP генов. Так, в GWAS Catalog было зарегистрировано 12 однонуклеотидных полиморфизмов, ассоциированных с множественной миеломой [11]. Картирование этих SNP в генах, удаленных менее чем на 10 000 килобаз, не выявило генов, которые могли бы быть отнесены к какому-либо метаболическому пути. Увеличение расстояния до 400 килобаз также не дало положительных результатов. Однако, анализ SNP в связи с генами, удаленными от них более чем на 500 килобаз, привел к формированию кластеров, вовлеченных в 2 пути - «Миелома» и «Метабо-

лизм при раке», оба имеющих близкое отношение к фенотипу. Так же, многие SNPs показали ассоциации с цветом глаз, однако, анализ связей между ними и близкорасположенными генами не дал результатов [12]. Отношение этого фенотипа к пути «меланогенез» открылось только при картировании SNPs в генах, расположенных на расстоянии 100 килобаз [13]. Парадигма «один SNP – много генов» может быть весьма полезной в обнаружении молекулярных основ фенотипического признака. Интересно, что геном существует в трехмерном измерении, и это является главным фактором, объясняющим дистанционный эффект SNPs на удаленные гены [14-16]. Накапливающиеся данные о 3D-упаковке генома позволят в ближайшем будущем определить положение SNPs в их трехмерном пространственном окружении. К сожалению, сделать это сейчас не представляется возможным в виду нехватки знаний о структурной и пространственной организации хроматина. Еще одно объяснение дистанционных эффектов SNP заключается в том, что однонуклеотидных полиморфизмы являются маркерами больших структурных вариантов, затрагивающих крупные сегменты хромосом [17].

Подавляющее большинство SNP (около 90%), ассоциированных с болезнями в результате GWAS, располагаются в некодирующих областях генома [18]. В расшифровке регуляторного потенциала некодирующих SNPs важную роль могут играть *in silico* инструменты, оперирующие доступными базами данных с достаточными характеристиками генной экспрессии, эпигенетических маркеров, 3D-контактов хроматина и других геномных параметров, включая гены-мишени фармакотерапии.

Использование уже разработанных биоинформатических инструментов будет способствовать пониманию функциональной значимости ассоциированных с фенотипом SNPs, что уже успешно продемонстрировано рядом работ [18-23]. Например, Watanabe K. и др. [24], опираясь на данные полногеномного исследования ас-

социаций однонуклеотидных вариантов с индексом массы тела, обрабатывают результаты, используя биоинформатическую онлайн-платформу FUMA (<http://fuma.ctglab.nl>) для функционального аннотирования приоритетных SNPs и интерактивной визуализации взаимодействующих генов. Chen и соавторы [12] также использовали биоинформатические инструменты и ресурсы VEP, Regulome DB, ANNOVAR, HaploReg для предсказания функционального и регуляторного потенциала 9 184 некодирующих вариантов базы NHGRI, и описали регулируемые механизмы 96% изученных вариантов. Более того, 3 случайно выбранных варианта из этого списка были подвержены функциональному тестированию и проявили энхансерную или сайленсерную активность. Однако фактически с начала 2015 года менее 40% GWAS использовали биоинформатические инструменты для приоритезации и предсказания функции некодирующих SNPs [25].

Биоинформатические методы и подходы к оценке регуляторного потенциала полиморфных локусов.

В обзорном исследовании Nishizaki и Boyle [26] представлена детальная характеристика современных онлайн-инструментов для оценки функциональной роли SNPs в геноме.

Функциональный анализ и интерпретация локусов, ассоциированных с болезнью в результате GWAS представляет собой сложную и многоэтапную задачу, начиная с точного картирования причинных SNPs до исследования функции гена-мишени. При этом на ключевых этапах данного процесса используются различные биоинформатические инструменты. На рисунке 1 представлены наиболее популярные подходы к выявлению функциональных полиморфизмов, ассоциированных с развитием патологии [3]. Данная схема анализа выявленных ассоциаций GWAS-локусов направлена на выделение приоритетных SNPs из общей массы ассоциированных с фенотипом генетических

вариантов. Этот подход подразумевает интеграцию генетических данных, биоинформатического анализа и вычислительных процедур. Однако функциональная оценка некодирующих регуляторных вариантов требует поэтапного применения значительного арсенала программных вычислений, в том числе обращение к большим массивам биологических данных, *in silico* инструментам и результатам экспериментальных молекулярных этиологий.

Первый этап – точное картирование, которое направлено на обнаружение т.н. причинных SNP, влияющих на молекулярные и клеточные процессы, связанные с фенотипом болезнью. Достигается посредством, например, более “плотного” генотипирования участка генома или процедурой импутации недостающих данных генотипирования, а также статистическими процедурами (логистическая регрессия, тест отношения правдоподобия, анализ гаплотипов и др.) для выявления SNP, наиболее значимо ассоциированного с фенотипом - с наибольшим размером эффекта (effect size) и наиболее значимым уровнем значимости (P-value) влияния на фенотип [14, 27, 28].

Следующий этап – *in silico* аннотирование SNPs, направлен на выяснение механизма, посредством которого данный генетический вариант может влиять на экспрессию гена или активность его продукта. Сложность данной задачи заключается в интерпретации некодирующих SNP, что требует применения множества вычислительных процедур, включая анализ специальных баз данных, содержащих экспериментально подтвержденную информацию о регуляторном потенциале SNP (например, Regulome, TRANFAC, ChiA) и применение специальных биоинформатических инструментов. Сканируя мотивы в базах данных TRANSFAC, JASPAR и UniPRODE, можно легко оценить аффинность транскрипционных факторов в отношении связывания с заданными участками ДНК [19-21].



Рис.1. Способы функционального анализа и интерпретации локусов GWAS
Fig. 1. Workflow for functional analysis and interpretation of GWAS loci

Следующий этап – оценка функциональности SNP и идентификация гена-мишени. Для некодирующих регуляторных вариантов используются: методы анализа экспрессии генов, включая выявление eQTL (локусов в геноме, ассоциированных с количественными изменениями экспрессии генов) в различных тканях с полногеномным уровнем значимости; анализ влияния 3-мерной структуры хроматина на регион ДНК с SNP, люциферазный тест, *in vitro* тест связывания белка с ДНК [29, 30].

Для некодирующих вариантов РНК, используются инструменты для поиска их генов-мишеней и таким образом оценивают потенциал связывания микроРНК с областью SNP (TargetScan, MiRanda) [31]. Для некодирующих эпигенетических вариантов оценивают уровень метилирования ДНК, проводят иммунопреципитацию

хроматина в сочетании с высокоэффективным секвенированием [32-34].

Формулировка гипотезы о биоинформатически предсказанном эффекте SNP на фенотип затем используется для тестирования этого эффекта в эксперименте. На заключительном этапе исследуют функции гена-мишени с использованием культивированных клеточных линий человеческих тканей, нокаут-моделей животных, технологии геномного редактирования (CRISPR) и других методологий [35].

Таким образом, именно некодирующие SNP являются главным и наиболее сложным объектом для анализа и биологической интерпретации связи с фенотипом болезни. Рассмотрим основные особенности регуляторного потенциала некодирующих SNP. Многие некодирующие SNPs находятся в регуляторных последовательностях генома и способны влиять на экс-

прессию генов на транскрипционном, посттранскрипционном и посттрансляционном уровнях [36,37]. Некодирующие варианты в энхансерах – одни из главных кандидатов для функциональной интерпретации GWAS-локусов. Регуляторные сигналы могут действовать на больших расстояниях по всему геному и вступать в контакт с промоторами-мишенями посредством трехмерной упаковки ДНК. Доступность транскрипционных факторов зависит от структурных изменений хроматина, обусловленных посттрансляционными модификациями гистонов, такими как метилирование и ацетилирование [38]. В отличие от закрытого хроматина, т.н. пермиссивный хроматин достаточно динамичен для факторов транскрипции, инициируя ремоделирование доступности специфической последовательности ДНК и обеспечивая открытую конформацию хроматина [39].

На сегодняшний день существует внушительный арсенал *in silico* инструментов и интернет-ресурсов для анализа регуляторного потенциала локусов, ассоциированных с болезнями [39]. Биоинформатические подходы к определению потенциальных регуляторных эффектов не кодирующих SNPs были значительно усилены экспериментальными исследованиями полногеномного формата. Данные проектов ENCODE и проект Национального Института здоровья «Дорожная карта эпигенома» могут быть использованы для оценки регуляторного потенциала не кодирующих вариантов и их проявления различных тканях [40, 41]. Регуляторная функция полиморфизма может быть проявлением эпигенетической модификации генома, включая модификацию гистонов, регуляцию открытости хроматина, связывающую способность транскрипционных факторов. При этом можно оценить потенциальное влияние варианта посредством оценки различных геномных характеристик, таких как количественная оценка экспрессии гена в различных тканях (картирование eQTL), секвенирование хроматина (CHIP-seq технологии), секвенирование гиперчувствительных участков для

ДНК-азы I, анализ взаимодействия хроматина, идентификация ДНК-мотивов, специфически связывающих транскрипционные факторы.

Активно используемые на сегодняшний день онлайн биоинформатические ресурсы направлены на оценку влияния открытого хроматина (ресурсы ENCODE, RegulomeDB, данные проекта «Дорожная карта эпигенома человека»); предсказание связывания участка транскрипционных факторов с ДНК (TRANSFAC, JASPAR); оценку ДНК-белковых взаимодействий (ENCODE, RegulomeDB, HarloReg), оценку метилирования ДНК (ENCODE, MethDB, проект «Дорожная карта эпигенома человека»); анализ экспрессии РНК, модификации гистонов и взаимодействия хроматина. В таблице 2 представлены онлайн-ресурсы для доступа к наиболее популярным биоинформатическим инструментам оценки регуляторного потенциала полиморфизмов. Данные инструменты позволяют аннотировать и предсказывать регуляторные эффекты SNPs с использованием трех основных методологических подходов: функционального аннотирования, оценки консервативности и технологии машинного обучения.

Экспериментальные подтверждения функциональности SNPs реализуется посредством современных биотехнологий. Исследования репортного гена являются дополнением к вышеописанным поисковым системам и предлагают прямое измерение функционального эффекта варианта на уровень экспрессии гена. Для этого регуляторный элемент помещают выше промотора и вводят плазмиду, содержащую интересующий ген [23]. Также, трансгенные линии животных, включая мышей и рыб, представляют собой ценный способ оценки фенотипического эффекта мутации *in vivo* [54]. С открытием редактирования регуляторных коротких палиндромных повторов (технология геномного редактирования CRISPR), не кодирующие варианты и структурные изменения могут быть легче изучены на таких более сложных модельных системах [55].

Таблица 2

Online-ресурсы для доступа к *in silico* инструментам аннотирования полиморфизмов

Table 2

Online resources for access to *in silico* polymorphism annotation tools

Ресурс	Описание, URL	Ссылка
VEP	Включает аннотации из базы Ensembl, предсказывает эффекты SNPs на полногеномном уровне, а также прогнозирует тканеспецифическую активность для 13 клеточных линий человека. http://www.ensembl.org/info/docs/tools/vep/script/index.html	McLaren et al., 2010 [42]
RegulomeDB	Использует бальную систему оценки функциональности полиморфизма с использованием данных более чем 100 типов клеток. http://regulomedb.org	Boyle et al., 2012 [43]
FunciSNP	Использует в качестве вводных пользовательские аннотации для приоритизации SNPs, позволяя пользователям настраивать свои аннотации для запроса интересующего клеточного типа. http://www.bioconductor.org/packages/release/bioc/html/FunciSNP.html	Coetzee et al., 2012 [44]
ANNOVAR	Инструмент командной строки, который использует аннотации в привязке регионам для аннотирования некодирующих вариантов (включая indels) в дополнение к сравнению их с известными базами данных. http://annovar.openbioinformatics.org	Wang et al., 2010 [45]
HaploReg	Поисковый репозиторий для SNPs и indels Проекта 1000 геномов, представляет сводные данные известных аннотаций для SNPs внутри LD блока. http://www.broadinstitute.org/mammals/haploreg/haploreg.php	Ward and Kellis, 2012 [46]
GWAS3D	Оценивает SNPs и indels посредством анализа их трехмерных хромосомных взаимодействий и нарушений связывания транскрипционных факторов. http://jjwanglab.org/gwas3d	Li et al., 2013 [25]
fitCons	Использует метод INSIGHT для предсказания возможности того, что SNP будет влиять на конформацию посредством скрининга сигнатур положительного и отрицательного отбора с на основе трех типов клеток. http://compgen.bscb.cornell.edu/fitCons/	Gulko et al., 2015 [47]
GWAVA	GWAVA основан на алгоритме random forest, использует базы данных HGMD и контрольные варианты проекта 1000 геномов для предсказания функциональности SNP. ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/	Ritchie et al., 2014 [48]
CADD	CADD основан на методе опорных векторов (SVM), используя симулированные варианты как патологические и аллели, сходные у человека и шимпанзе в качестве контрольных. http://cadd.gs.washington.edu	Kircher et al., 2014 [49]
DANN	DANN основан на алгоритме нелинейного обучения нейронных сетей (фиксированные аллели сравнивают со стимулированными вариантами) подобно CADD. https://cbcl.ics.uci.edu/public_data/DANN/	Quang et al., 2015 [51]
FATHMM-MKL	Реализует Kernel-классификатор для оценки сложных нелинейных моделей с использованием патогенных вариантов (HGMD) и обучающих вариантов Проекта 1000 геномов. http://fathmm.biocompute.org.uk	Shihab et al., 2015 [10]
deltaSVM	Использует алгоритм gkm-SVM машинного обучения для оценки эффекта варианта в специфических типах клеток. http://www.beerlab.org/deltasvm/	Lee et al., 2015 [52]
DeepSEA	Использует многослойную иерархическую структурированную модель последовательности глубокого обучения для прогнозирования функциональных SNP с чувствительностью к одному нуклеотиду с использованием данных ENCODE и Roadmap Epigenomics. http://deepsea.princeton.edu/job/analysis/create/	Zhou and Troyanskaya, 2015 [53]

Метилирование ДНК – это фундаментальная эпигенетическая характеристика, контролирующая включение/выключение генной экспрессии. Тем

не менее, взаимосвязь между характером последовательности ДНК и степенью метилирования до конца не ясна [56]. Исследования позволили выявить корреляции

между профилем метилирования ДНК и индивидуальными генотипами для идентификации локусов, способных повлиять на статус метилирования генов. Было открыто множество генных локусов, объясняющих различную степень метилирования т.н. CpG-островков в зависимости от популяции или типа клеточной линии. Однако далеко не все варианты метилирования могут быть интерпретированы только лишь с учетом одних генетических факторов [56, 58]. Поэтому изучение роли SNP в формировании того или иного профиля метилирования становится одним из главных объектов для биоинформатического анализа.

Секвенирование генома нового поколения (NGS) позволяет измерить в полногеномном масштабе экспрессию генов, привязку транскрипционных факторов, доступность/открытость хроматина, модификации гистонов и метилирование ДНК. Огромные усилия были приложены для характеристики вариаций генома на транскриптомом уровне в различных клеточных линиях и тканях. Проект GENCODE (www.genencodegenes.org) содержит огромный пласт экспериментальных данных о функциональных элементах генома и представляет собой высококачественный каталог транскриптов [57]. Количественные данные о функциональной значимости вариантов могут отличаться в разных базах данных. Так, McCarthy et al [58] показали, что конкордантность Loss of Function (LoF) вариантов между ANNOVAR и VEP составляет 65%, хотя обе поисковые системы используют один и тот же набор транскриптов. Что требует стандартизации представления данных, The Sequence Ontology Project – это первый ресурс, направленный на стандартизацию описательных характеристик генома, опирающийся на базы VEP и ANNOVAR [59].

Таким образом, вся выше представленная информация в полном объеме депонирована в сети баз данных, доступных в Internet, однако, понимание биологического смысла этой информации представляет не меньшую трудность, чем сам про-

цесс их получения. Интерпретация выявленных ассоциаций с позиций формальной логики системы ген-мРНК-белок-метаболит в большинстве случаев крайне проблематична, поскольку абсолютное число локусов, обнаруживших ассоциации с различными фенотипами находятся в некодирующих областях генома, межгенных пространствах или в генах, не имеющих прямого отношения к изучаемому заболеванию. Путь от генотипа к фенотипу в таком случае удастся проложить с помощью биоинформатического инструментария, позволяющего предсказать эффект варианта ДНК на различные аспекты молекулярной жизни в микромире, включая транскрипцию, связывание транскрипционных факторов, созревание пре-мРНК, сплайсинг, трансляцию, эпигенетические модификации (метилирование ДНК, открытость хроматина). Онлайн ресурсы позволяют дать всестороннюю оценку эффектам SNPs и их роли в развитии того или иного фенотипического признака болезни.

В перспективе, увеличение объема выборки до 100000 и более человек позволит в будущих GWAS открыть новые варианты ассоциаций с известными заболеваниями, что поможет конкретизировать диагноз вплоть до его молекулярных основ и выбрать персонализированное лечение болезней. Безусловно, с течением времени, GWAS на основе SNP-панелей будут замещены GWAS на основе полногеномного секвенирования, что, вероятно, прольет свет на неизвестные по сей день аспекты взаимосвязей в системе генотип-фенотип-среда. Если 10-15 лет назад технология генотипирования была лимитирующим фактором в сфере генетических исследований, то сейчас этим фактором является полнота фенотипической характеристики обследуемых лиц. Современный биоинформатический анализ подразумевает стратификацию по фенотипическим признакам с целью выявления причинно-следственных отношений между ними и понимания того, как той или иной фактор среды опосредует воздействие генотипа на формирование

признака. В конечном счете, результаты полногеномных ассоциативных исследований должны быть всесторонне проанализированы и имплементированы в практическое здравоохранение в виде более точных (по чувствительности и специфичности) диагностических предсказательных тестов и алгоритмов персонализированного лечения и профилактики социально значимой мультифакториальной патологии.

Заключение. Полногеномные ассоциативные исследования открыли новую эру в истории генетических исследований мультифакториальных заболеваний. В данном обзоре мы представили способ поэтапного функционального анализа и интерпретации локусов GWAS, описали инструменты и online-ресурсы, позволяющие аннотировать и предсказывать регуляторные эффекты SNPs, основываясь на трех основных методиках. Выяснили, что биоинформатический анализ *in silico* позволяет дать всестороннюю оценку эффектам SNPs и их роли в развитии того или иного фенотипического признака болезни. Полученные результаты были успешно дополнены результатами изучения экспрессии генов в различных тканях, метилировании ДНК и характеристиками метаболомного профиля.

Информация о финансировании

Работа выполнена при финансовой поддержке Российского научного фонда (проект № 20-15-00227).

Financial support

The study was supported by the Russian Science Foundation (№20-15-00227).

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Conflict of interests

The authors have no conflict of interest to declare.

Список литературы

1. Visscher PM, Wray EM, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human*

Genetics. 2017;101(1):5-22. DOI: 10.1016/j.ajhg.2017.06.005

2. Ermann J, Glimcher LH. After GWAS: mice to the rescue? *Current opinion in immunology.* 2012;24(5):564-570. DOI: 10.1016/j.coi.2012.09.005

3. Edwards SL, Beesley J, French JD, et al. Beyond GWASs: illuminating the dark road from association to function. *The American Journal of Human Genetics.* 2013;93(5):779-797. DOI: <https://doi.org/10.1016/j.ajhg.2013.10.012>

4. Wu J, Yu Z, Chen G. PD-1/PD-Ls: A New Target for Regulating Immunopathogenesis in Central Nervous System Disorders. *Current drug delivery.* 2017;14(6):791-796. DOI: <https://doi.org/10.2174/1567201814666161123152311>

5. Zarrei M, MacDonald J, Merico D, et al. A copy number variation map of the human genome. *Nature reviews genetics.* 2015;16(3):172-183. DOI: <https://doi.org/10.1038/nrg3871>

6. Grimm JB, English BP, Chen J, et al. A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nature methods.* 2015;12(3):244-250. DOI: <https://doi.org/10.1038/nmeth.3256>

7. Butkiewicz M, Bush WS. In silico functional annotation of genomic variation. *Current protocols in human genetics.* 2016;88(1):6.15.1-6.15.17. DOI: <https://doi.org/10.1002/0471142905.hg0615s88>

8. Lower KM, Hughes JR, De Gobbi M, et al. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proceedings of the National Academy of Sciences.* 2009;106(51):21771-21776. DOI: <https://doi.org/10.1073/pnas.0909331106>

9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences.* 1977;74(12):5463-5467. DOI: 10.1073/pnas.74.12.5463

10. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics.* 2015;31(10):1536-1543. DOI: <https://doi.org/10.1093/bioinformatics/btv009>

11. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? *Nucleic acids research.* 2016;44(13):6046-6054. DOI: <https://doi.org/10.1093/nar/gkw500>

12. Chen G, Yu D, Chen J, et al. Re-annotation of presumed noncoding disease/trait-associated genetic variants by integrative analyses. *Scientific reports*. 2015;5:9453. DOI: <https://doi.org/10.1038/srep09453>
13. Law MH, Bishop DT, Lee JE, et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nature Genetics*. 2015;47(9):987-995. DOI: <https://doi.org/10.1038/ng.3373>
14. Sidore C, Busonero F, Maschio A, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature genetics*. 2015;47(11):1272-1281. DOI: <https://doi.org/10.1038/ng.3368>
15. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. DOI: [10.1038/35057062](https://doi.org/10.1038/35057062)
16. Hagege H, Klous P, Braem, C, et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nature protocols*. 2007;2(7):1722-1733. DOI: <https://doi.org/10.1038/nprot.2007.243>
17. Bulik-Sullivan B, Loh P, Finucane H, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*. 2015;47(3):291-295. DOI: <https://doi.org/10.1038/ng.3211>
18. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Human molecular genetics*. 2015;24(R1):R102-R110. DOI: <https://doi.org/10.1093/hmg/ddv259>
19. Matys V, Fricke E, Geffers R, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*. 2003;31(1):374-378. DOI: <https://doi.org/10.1093/nar/gkg108>
20. Sandelin A, Alkema W, Engstrom P, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*. 2004;32(1):D91-D94. DOI: <http://dx.doi.org/10.1093/nar/gkh012>
21. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research*. 2008;37(1):D77-D82. DOI: <https://doi.org/10.1093/nar/gkn660>
22. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*. 2009;37(2):W202-W208. DOI: <https://doi.org/10.1093/nar/gkp335>
23. Bryzgalov LO, Antontseva EV, Matveeva MYu, et al. Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. *PLoS one*. 2013;8(10):e78833. DOI: <https://doi.org/10.1371/journal.pone.0078833>
24. Watanabe K, Taskesen E, van Bochoven A, et al. FUMA: Functional mapping and annotation of genetic associations. *European Neuropsychopharmacology*. 2019;29(3):S789-S790. DOI: <https://doi.org/10.1016/j.euroneuro.2017.08.018>
25. Li MJ, Yan B, Sham PCh, et al. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Briefings in bioinformatics*. 2014;16(3):393-412. DOI: <https://doi.org/10.1093/bib/bbu018>
26. Nishizaki SS, Boyle AP. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends in Genetics*. 2017;33(1):34-45. DOI: <https://doi.org/10.1016/j.tig.2016.10.008>
27. Ameer A, Rada-Iglesias A, Komorowski J, et al. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. *Nucleic acids research*. 2009;37(12):e85-e85. DOI: <https://doi.org/10.1093/nar/gkp381>
28. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*. 2014;15(2):256-278. DOI: <https://doi.org/10.1093/bib/bbs086>
29. Duggal G, Wang H., Kingsford C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic acids research*. 2013;42(1):87-96. DOI: <https://doi.org/10.1093/nar/gkt857>
30. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*. 2007;35(1):D61-D65. DOI: <https://doi.org/10.1093/nar/gkl842>
31. Nishizaki SS, Boyle A. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends in Genetics*. 2017;33(1):34-45. DOI: <https://doi.org/10.1016/j.tig.2016.10.008>
32. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS genetics*. 2010;6(5):e1000952. DOI: <https://doi.org/10.1371/journal.pgen.1000952>
33. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, et al. Passive and active DNA

methylation and the interplay with genetic variation in gene regulation. *Nelife*. 2013;2:e00523. DOI: 10.7554/eLife.00523

34.Kato N, Loh M, Takeuchi F, et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nature genetics*. 2015;47(11):1282-1293. DOI: <https://doi.org/10.1038/ng.3405>

35.Fisher S, Grice E, Vinton R, et al. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nature protocols*. 2006;1(3):1297-1305. DOI: <https://doi.org/10.1038/nprot.2006.230>

36.Kheradpour P, Ernst J, Melnikov A, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research*. 2013;23(5):800-811. DOI: 10.1101/gr.144899.112

37.Arnold CD, Gerlach D, Stelzer Ch, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. 2013;339(6123):1074-1077. DOI: 10.1126/science.1232542

38.Strahl BD, Allis CD. The language of covalent histone modifications. *Nature*. 2000;403(6765):41. DOI: <https://doi.org/10.1038/47412>

39.Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*. 2019;20:207-220. DOI: <https://doi.org/10.1038/s41576-018-0089-8>

40.Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*. 2012;22(99):1760-1774. DOI: 10.1101/gr.135350.111

41.Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. *Journal of biomedical informatics*. 2011;44(1):87-93. DOI: <https://doi.org/10.1016/j.jbi.2010.03.002>

42.McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26(16):2069-2070. DOI: <https://doi.org/10.1093/bioinformatics/btq330>

43.Boyle A, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*. 2012;22(9):1790-1797. DOI: 10.1101/gr.137323.112

44.Coetzee SG, Rhie SK, Berman BP, et al. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic acids research*. 2012;40(18):e139-e139. DOI: <https://doi.org/10.1093/nar/gks542>

45.Zhang Z, Wang Y, Wang L, et al. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One*. 2010;5(12):e14316. DOI: <https://doi.org/10.1371/journal.pone.0014316>

46.Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*. 2012;30(11):1095-1106. DOI: <https://doi.org/10.1038/nbt.2422>

47.Gulko B, Hubisz M, Gronau I, et al. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature genetics*. 2015;47(3):276-283. DOI: <https://doi.org/10.1038/ng.3196>

48.Ritchie GRS, Dunham I, Zeggini E, et al. Functional annotation of noncoding sequence variants. *Nature methods*. 2014;11(3):294-296. DOI: <https://doi.org/10.1038/nmeth.2832>

49.Kircher M, Witten D, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014;46(3):310-315. DOI: <https://doi.org/10.1038/ng.2892>

50.Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. 2014;31(5):761-763. DOI: <https://doi.org/10.1093/bioinformatics/btu703>

51.Rizvi NA, Hellmann MD, Snyder A, et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 2015;348(6230):124-128. DOI: 10.1126/science.aaa1348

52.Zhou J, Troyanskaya OG, Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*. 2015;12(10):931-934. DOI: <https://doi.org/10.1038/nmeth.3547>

53.Maher B. ENCODE: The human encyclopedia. *Nature News*. 2012;489(7414):46-48. DOI: <https://doi.org/10.1038/489046a>

54.Li MJ, Sham PC, Wang J. FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*. 2010;26(22):2897-2899. DOI: <https://doi.org/10.1093/bioinformatics/btq540>

55. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*. 2012;13(7):484-492. DOI: <https://doi.org/10.1038/nrg3230>

56. Schübeler D. Epigenetic islands in a genetic ocean. *Science*. 2012;338(6108):756-757. DOI: [10.1126/science.1227243](https://doi.org/10.1126/science.1227243)

57. Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. *Nucleic acids research*. 2013;41(D1):D48-D55 DOI: [10.1093/nar/gks1236](https://doi.org/10.1093/nar/gks1236)

58. McCarthy DJ, Humburg P, Kanapin A, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome medicine*. 2014;6(3):26. DOI: <https://doi.org/10.1186/gm543>

59. Morin RD, Mendez-Lago M, Mungall A, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*. 2011;476(7360):298-303. DOI: <https://doi.org/10.1038/nature10351>

References

1. Visscher PM, Wray EM, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*. 2017;101(1):5-22. DOI: [10.1016/j.ajhg.2017.06.005](https://doi.org/10.1016/j.ajhg.2017.06.005)

2. Ermann J, Glimcher LH. After GWAS: mice to the rescue? *Current opinion in immunology*. 2012;24(5):564-570. DOI: [10.1016/j.coi.2012.09.005](https://doi.org/10.1016/j.coi.2012.09.005)

3. Edwards SL, Beesley J, French JD, et al. Beyond GWASs: illuminating the dark road from association to function. *The American Journal of Human Genetics*. 2013;93(5):779-797. DOI: <https://doi.org/10.1016/j.ajhg.2013.10.012>

4. Wu J, Yu Z, Chen G. PD-1/PD-Ls: A New Target for Regulating Immunopathogenesis in Central Nervous System Disorders. *Current drug delivery*. 2017;14(6):791-796. DOI: <https://doi.org/10.2174/1567201814666161123152311>

5. Zarrei M, MacDonald J, Merico D, et al. A copy number variation map of the human genome. *Nature reviews genetics*. 2015;16(3):172-183. DOI: <https://doi.org/10.1038/nrg3871>

6. Grimm JB, English BP, Chen J, et al. A general method to improve fluorophores for live-cell and single-molecule microscopy. *Nature methods*. 2015;12(3):244-250. DOI: <https://doi.org/10.1038/nmeth.3256>

7. Butkiewicz M, Bush WS. In silico functional annotation of genomic variation. *Current protocols in human genetics*. 2016;88(1):6.15.1-

6.15.17. DOI: <https://doi.org/10.1002/0471142905.hg0615s88>

8. Lower KM, Hughes JR, De Gobbi M, et al. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proceedings of the National Academy of Sciences*. 2009;106(51):21771-21776. DOI: <https://doi.org/10.1073/pnas.0909331106>

9. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*. 1977;74(12):5463-5467. DOI: [10.1073/pnas.74.12.5463](https://doi.org/10.1073/pnas.74.12.5463)

10. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015;31(10):1536-1543. DOI: <https://doi.org/10.1093/bioinformatics/btv009>

11. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? *Nucleic acids research*. 2016;44(13):6046-6054. DOI: <https://doi.org/10.1093/nar/gkw500>

12. Chen G, Yu D, Chen J, et al. Re-annotation of presumed noncoding disease/trait-associated genetic variants by integrative analyses. *Scientific reports*. 2015;5:9453. DOI: <https://doi.org/10.1038/srep09453>

13. Law MH, Bishop DT, Lee JE, et al. Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nature Genetics*. 2015;47(9):987-995. DOI: <https://doi.org/10.1038/ng.3373>

14. Sidore C, Busonero F, Maschio A, et al. Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nature genetics*. 2015;47(11):1272-1281. DOI: <https://doi.org/10.1038/ng.3368>

15. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. DOI: [10.1038/35057062](https://doi.org/10.1038/35057062)

16. Hagège H, Klous P, Braem C, et al. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nature protocols*. 2007;2(7):1722-1733. DOI: <https://doi.org/10.1038/nprot.2007.243>

17. Bulik-Sullivan B, Loh P, Finucane H, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*. 2015;47(3):291-295. DOI: <https://doi.org/10.1038/ng.3211>

18. Zhang F, Lupski JR. Non-coding genetic variants in human disease. *Human molecular genetics*. 2015;24(R1):R102-R110. DOI: <https://doi.org/10.1093/hmg/ddv259>
19. Matys V, Fricke E, Geffers R, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research*. 2003;31(1):374-378. DOI: <https://doi.org/10.1093/nar/gkg108>
20. Sandelin A, Alkema W, Engstrom P, et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*. 2004;32(1):D91-D94. DOI: <http://dx.doi.org/10.1093/nar/gkh012>
21. Newburger DE, Bulyk ML. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic acids research*. 2008;37(1):D77-D82. DOI: <https://doi.org/10.1093/nar/gkn660>
22. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*. 2009;37(2):W202-W208. DOI: <https://doi.org/10.1093/nar/gkp335>
23. Bryzgalov LO, Antontseva EV, Matveeva MYu, et al. Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. *PLoS one*. 2013;8(10):e78833. DOI: <https://doi.org/10.1371/journal.pone.0078833>
24. Watanabe K, Taskesen E, van Bochoven A, et al. FUMA: Functional mapping and annotation of genetic associations. *European Neuropsychopharmacology*. 2019;29(3):S789-S790. DOI: <https://doi.org/10.1016/j.euroneuro.2017.08.018>
25. Li MJ, Yan B, Sham PCh, et al. Exploring the function of genetic variants in the non-coding genomic regions: approaches for identifying human regulatory variants affecting gene expression. *Briefings in bioinformatics*. 2014;16(3):393-412. DOI: <https://doi.org/10.1093/bib/bbu018>
26. Nishizaki SS, Boyle AP. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends in Genetics*. 2017;33(1):34-45. DOI: <https://doi.org/10.1016/j.tig.2016.10.008>
27. Ameer A, Rada-Iglesias A, Komorowski J, et al. Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. *Nucleic acids research*. 2009;37(12):e85-e85. DOI: <https://doi.org/10.1093/nar/gkp381>
28. Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*. 2014;15(2):256-278. DOI: <https://doi.org/10.1093/bib/bbs086>
29. Duggal G, Wang H., Kingsford C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic acids research*. 2013;42(1):87-96. DOI: <https://doi.org/10.1093/nar/gkt857>
30. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*. 2007;35(1):D61-D65. DOI: <https://doi.org/10.1093/nar/gkl842>
31. Nishizaki SS, Boyle A. Mining the unknown: assigning function to noncoding single nucleotide polymorphisms. *Trends in Genetics*. 2017;33(1):34-45. DOI: <https://doi.org/10.1016/j.tig.2016.10.008>
32. Gibbs JR, van der Brug MP, Hernandez DG, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS genetics*. 2010;6(5):e1000952. DOI: <https://doi.org/10.1371/journal.pgen.1000952>
33. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Nelife*. 2013;2:e00523. DOI: 10.7554/eLife.00523
34. Kato N, Loh M, Takeuchi F, et al. Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nature genetics*. 2015;47(11):1282-1293. DOI: <https://doi.org/10.1038/ng.3405>
35. Fisher S, Grice E, Vinton R, et al. Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nature protocols*. 2006;1(3):1297-1305. DOI: <https://doi.org/10.1038/nprot.2006.230>
36. Kheradpour P, Ernst J, Melnikov A, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome research*. 2013;23(5):800-811. DOI: 10.1101/gr.144899.112
37. Arnold CD, Gerlach D, Stelzer Ch, et al. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*. 2013;339(6123):1074-1077. DOI: 10.1126/science.1232542
38. Strahl BD, Allis CD. The language of covalent histone modifications. *Nature*.

- 2000;403(6765):41. DOI: <https://doi.org/10.1038/47412>
39. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*. 2019;20:207-220. DOI: <https://doi.org/10.1038/s41576-018-0089-8>
40. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*. 2012;22(9):1760-1774. DOI: [10.1101/gr.135350.111](https://doi.org/10.1101/gr.135350.111)
41. Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. *Journal of biomedical informatics*. 2011;44(1):87-93. DOI: <https://doi.org/10.1016/j.jbi.2010.03.002>
42. McLaren W, Pritchard B, Rios D, et al. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26(16):2069-2070. DOI: <https://doi.org/10.1093/bioinformatics/btq330>
43. Boyle A, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*. 2012;22(9):1790-1797. DOI: [10.1101/gr.137323.112](https://doi.org/10.1101/gr.137323.112)
44. Coetzee SG, Rhie SK, Berman BP, et al. FungiSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic acids research*. 2012;40(18):e139-e139. DOI: <https://doi.org/10.1093/nar/gks542>
45. Zhang Z, Wang Y, Wang L, et al. The combined effects of amino acid substitutions and indels on the evolution of structure within protein families. *PLoS One*. 2010;5(12):e14316. DOI: <https://doi.org/10.1371/journal.pone.0014316>
46. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*. 2012;30(11):1095-1106. DOI: <https://doi.org/10.1038/nbt.2422>
47. Gulko B, Hubisz M, Gronau I, et al. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature genetics*. 2015;47(3):276-283. DOI: <https://doi.org/10.1038/ng.3196>
48. Ritchie GRS, Dunham I, Zeggini E, et al. Functional annotation of noncoding sequence variants. *Nature methods*. 2014;11(3):294-296. DOI: <https://doi.org/10.1038/nmeth.2832>
49. Kircher M, Witten D, Jain P, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014;46(3):310-315. DOI: <https://doi.org/10.1038/ng.2892>
50. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. 2014;31(5):761-763. DOI: <https://doi.org/10.1093/bioinformatics/btu703>
51. Rizvi NA, Hellmann MD, Snyder A, et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science*. 2015;348(6230):124-128. DOI: [10.1126/science.aaa1348](https://doi.org/10.1126/science.aaa1348)
52. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*. 2015;12(10):931-934. DOI: <https://doi.org/10.1038/nmeth.3547>
53. Maher B. ENCODE: The human encyclopedia. *Nature News*. 2012;489(7414):46-48. DOI: <https://doi.org/10.1038/489046a>
54. Li MJ, Sham PC, Wang J. FastPval: a fast and memory efficient program to calculate very low P-values from empirical distribution. *Bioinformatics*. 2010;26(22):2897-2899. DOI: <https://doi.org/10.1093/bioinformatics/btq540>
55. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*. 2012;13(7):484-492. DOI: <https://doi.org/10.1038/nrg3230>
56. Schübeler D. Epigenetic islands in a genetic ocean. *Science*. 2012;338(6108):756-757. DOI: [10.1126/science.1227243](https://doi.org/10.1126/science.1227243)
57. Flicek P, Ahmed I, Amode MR, et al. Ensembl 2013. *Nucleic acids research*. 2013;41(D1):D48-D55 DOI: [10.1093/nar/gks1236](https://doi.org/10.1093/nar/gks1236)
58. McCarthy DJ, Humburg P, Kanapin A, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome medicine*. 2014;6(3):26. DOI: <https://doi.org/10.1186/gm543>
59. Morin RD, Mendez-Lago M, Mungall A, et al. Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature*. 2011;476(7360):298-303. DOI: <https://doi.org/10.1038/nature10351>

Статья поступила в редакцию 17 июня 2020 г.
Поступила после доработки 5 сентября 2020 г.
Принята к печати 25 октября 2020 г.

Received 17 June 2020

Revised 5 September 2020

Accepted 25 October 2020

Информация об авторах

Алексей Валерьевич Полоников, доктор медицинских наук, профессор, профессор кафедры биологии, медицинской генетики и экологии, заведующий лабораторией статистической генетики и биоинформатики НИИ генетической и молекулярной эпидемиологии, директор НИИ генетической и молекулярной эпидемиологии, ФГБОУ ВО «Курский государственный медицинский университет», г. Курск, Российская Федерация, E-mail: polonikov@rambler.ru, ORCID: 0000-0001-6280-247X.

Елена Юрьевна Клёсова, инженер-биотехнолог, НИИ генетической и молекулярной эпидемиологии, ФГБОУ ВО «Курский государственный медицинский университет», г. Курск, Российская Федерация, E-mail: ecless@yandex.ru, ORCID: 0000-0002-1543-9230.

Юлия Эдуардовна Азарова, кандидат медицинских наук, доцент кафедры биологической химии, заведующая лабораторией биохимической генетики и метаболомики, НИИ генетической и молекулярной эпидемиологии, ФГБОУ ВО «Курский государственный медицинский университет»,

г. Курск, Российская Федерация, E-mail: azzzzar@yandex.ru, ORCID: 0000-0001-8098-8052.

Information about the authors

Alexey V. Polonikov, Doct. Sci. (Medicine), Professor, Professor at the Department of Biology, Medical Genetics and Ecology, Head of the Laboratory of Statistical Genetics and Bioinformatics, Research Institute for Genetic and Molecular Epidemiology, Director of Research Institute for Genetic and Molecular Epidemiology, Kursk State Medical University, Kursk, Russia, E-mail: polonikov@rambler.ru, ORCID: 0000-0001-6280-247X.

Elena Yu. Klyosova, Engineer-biotechnologist of Research Institute for Genetic and Molecular Epidemiology, Kursk State Medical University, Kursk, Russia, E-mail: ecless@yandex.ru, ORCID: 0000-0002-1543-9230.

Iuliia E. Azarova, Cand. Sci. (Medicine), Associate Professor at the Department of Biological Chemistry, Head of the Laboratory of Biochemical Genetics and Metabolomics of Research Institute for Genetic and Molecular Epidemiology, Kursk State Medical University, Kursk, Russia, E-mail: azzzzar@yandex.ru, ORCID: 0000-0001-8098-8052.