Математическая биология и биоинформатика
*2020. Т. 15. № 1. С. 1–20. doi: 10.17537/2020.15.****

====== **Информационные и вычислительные технологии в биологии и медицине** =====

УДК: 616-074+57.087.1

Методы поиска клинически значимых полиморфизмов в геноме человека

Валеев Э.С.*[1,2], **Фишман В.С.****[1,2]

[1]*Институт Цитологии и Генетики СО РАН, Новосибирск, Россия*
[2]*Новосибирский Государственный Университет, Новосибирск, Россия*

Аннотация. Клиническая генетика играет важную роль в диагностике редких наследственных заболеваний. Однако её применение не ограничивается только диагностикой ”— она может быть использована при прогнозировании возникновения и течения заболеваний, помогать в подборе терапии. Были разработаны множество методов поиска генетических вариантов, технические и биоинформационные решения. Этот обзор освещает как хорошо себя зарекомендовавшие методы, так и разработки недавнего времени методы, а также их перспективы. В первую очередь затрагиваются особенности наиболее используемых цитогенетических методик, таких как кариотипирование, FISH, CGH и MLPA, их преимущества и недостатки. Затем анализируются технологии секвенирования, при этом особое внимание уделяется секвенированию нового поколения (NGS). В заключение рассматриваются основные проблемы, связанные с биоинформационными методами обработки NGS-данных, и перспективы для будущих разработок, так как, несмотря на скорость развития технологий NGS и связанных с ними биоинформационных методов, диагностика полигенных и генетически гетерогенных заболеваний всё ещё остаётся нерешённой проблемой.

Ключевые слова: *Наследственные заболевания, клиническая генетика, цитогенетика, NGS*

ВВЕДЕНИЕ

Генетические варианты, их взаимодействие друг с другом и со средой определяет течение болезней. Существуют генетические варианты, которые определяют предрасположенность и проявляются только во взаимодействии со средой; примером могут служить варианты, определяющие предрасположенность к аддикциям (никотин, героин, алкоголь и пр.) [1]. Бывают и такие генетические варианты, которые повышают восприимчивость к одному фактору среды и повышают устойчивость к другому, либо дают позитивный эффект в сочетании и негативный по отдельности. Примером может служить бета-талассемия [2]. Особняком стоят те варианты, которые вне зависимости от средового компонента и генетического окружения приводят к развитию заболевания (например, нейрофиброматоз I типа, который наследуется по аутосомно-доминантному типу и имеет 100 % пенетрантность — 3).

Генетические заболевания остаются одной из основных причин младенческой и детской смертности в развитых странах. Врождённые аномалии являются причиной около 20 % смертности до 1 года, а также порядка 10 % в возрасте 1–4 года и 6 %

*emil@bionet.nsc.ru
**minja@bionet.nsc.ru

в возрасте 5–9 лет. Злокачественные новообразования являются причиной смерти в 8 % случаев в возрасте 1–4 лет, и 15 % случаев в возрасте 5–9 лет. Порядка 3 % от смертности в возрасте 1–9 лет связаны с сердечными патологиями [4]. Взрослые люди с генетическими патологиями требуют огромных затрат средств — на радикальные и паллиативные операции, медикаментозную поддержку (иногда пожизненную), создание условий, учреждений и обучение персонала для обеспечения специализированного ухода.

Таким образом, доступные и точные методы диагностики генетических заболеваний могут помочь в сокращении заболеваемости и смертности, а также повысить экономическое благополучие населения.

ЧАСТЫЕ И РЕДКИЕ (ОРФАННЫЕ) ПАТОЛОГИИ

Генетические патологии делятся на группы по частоте встречаемости в популяции. Выделяют частые и редкие (орфанные) заболевания. Определения орфанных заболеваний могут различаться — например, в США, согласно “Health Promotion and Disease Prevention Amendments of 1984”, редкими считаются патологии, поражающие менее 200 тыс. населения страны (примерно 1 : 1630 при текущей численности населения в 326 млн человек) [5]. Европейское Медицинское Агентство определяет границу как 1 : 2000. Систематический анализ показал, что существует более 290 определений, и среднее значение находится в интервале 40–50 на 100 тыс. населения [6].

Также сложность в определении орфанных заболеваний представляет неравномерность их распространённости в тех или иных регионах. Некоторые заболевания могут быть орфанными в одной популяции и частыми в другой (эффект основателя, а также сверхдоминирование). Частным случаем эффекта основателя является атаксия Каймановых островов, связанная с гипоплазией мозжечка и сопутствующими неврологическими проявлениями (задержка развития, дизартрия, нистагм, интенционное дрожание). Это аутосомно-рецессивное заболевание распространено исключительно в одном регионе — Большой Кайманов остров, гетерозиготные носители составляют около 18 % местного населения [7]. Примером сверхдоминирования может служить бета-талассемия — заболевание, связанное с нарушением структуры гемоглобина. Несмотря на то, что у эритроцитов носителей в значительной степени снижена способность переносить кислород, дефектный гемоглобин представляет сложность для развития малярийного плазмодия и таким образом повышает устойчивость носителя бета-талассемии к малярии [2]. Соответственно, бета-талассемия распространена в эпидемически опасных по малярии регионах — Средиземноморье и Юго-Восточная Азия, наибольшая частота встречаемости наблюдается на Кипре (14 %) и Сардинии (10.3 %) при средней частоте по земному шару в 1.5 %.

Несмотря на то, что каждое из орфанных заболеваний само по себе встречается редко, в сумме они поражают значительный процент населения (предположительно 5–8 % европейской популяции). Общее число орфанных болезней неизвестно по причине недостатков стандартизации, наиболее частая оценка — 5–8 тыс. [8]. Около 80 % редких болезней имеют генетическую природу и начинаются в раннем детстве [8].

Таким образом, ключевым моментом для изучения данных заболеваний является понимание механизмов, лежащих в основе их развития. Количество орфанных заболеваний делает эту задачу крайне непростой. Тем не менее, многие механизмы на сегодняшний момент достаточно хорошо изучены. О них речь пойдёт далее.

ЦИТОГЕНЕТИЧЕСКИЕ МЕТОДЫ ДЕТЕКЦИИ ВАРИАНТОВ

Кариотипирование

Данный метод представляет собой микроскопическое исследование клеток, синхронизированных на стадии метафазы митоза. Однако простое микроскопическое исследование хромосом плохо подходит для обнаружения генетических вариантов, поэтому были разработаны различные методы окраски (бэндинга), позволяющие отдифференцировать отдельные хромосомы и хромосомные регионы [9]:

1. Q-окрашивание — позволяет отдифференцировать все хромосомы, применяется для исследования Y-хромосомы (быстрое определение генетического пола, выявление мозаицизма по Y-хромосоме, транслокаций между Y-хромосомой и другими хромосомами). Окрашивание легко снимается, что позволяет использовать этот метод для последовательной окраски и изучения хромосом;

2. G-окрашивание — наиболее часто используемый метод. Позволяет отдифференцировать все хромосомы, гарантирует стойкое окрашивание, легко поддаётся фотографированию.

3. R-окрашивание — визуализирует концы хромосом, а также специфические именно для этого окрашивания бэнды (так называемые R-позитивные бэнды).

4. C-окрашивание — применяется для анализа вариабельной дистальной части Y-хромосомы, а также центромерных регионов прочих хромосом, содержащих конститутивный гетерохроматин. Хорошо подходит для выявления перестроек, затрагивающих гетерохроматиновые регионы. Кроме того, C-окрашиванием хорошо определяются кольцевые и дицентрические хромосомы;

5. NOR-окрашивание — визуализирует ядрышковые организаторы (англ. *Nucleolus Organizer Region, NOR*), богатые рибосомальными генами;

6. DA–DAPI-окрашивание — применяется для идентификации центромерных гетерохроматизированных районов.

Окрашенные хромосомы далее изучаются на предмет формы, количества и наличия перестроек.

Кариотипирование — рутинная методика при диагностике врождённых патологий, аутопсии мертворожденных и злокачественных образований кроветворного ряда. Преимущества кариотипирования в том, что данным методом можно охватить весь геном, визуализации поддаются отдельные клетки и отдельные хромосомы. Ограничения — обязательно требуются живые клетки, также на эффективность влияет размер перестроек (не менее 1–5 Mbp) и процент поражённых клеток в образце (минимум 5–10 %) [10].

В целом классический метод кариотипирования, достаточно дешёвый и простой в исполнении, требует от исследователя значительного опыта при интерпретации. Более поздние методы изучения хромосом, как будет показано далее, развивались не только в направлении увеличения разрешающей способности, но и облегчения интерпретации полученных данных.

Флуоресцентная *in situ* гибридизация (FISH)

Основой флуоресцентной *in situ* гибридизации (англ. *Fluorescence In Situ Hybridization, FISH*) является гибридизация нуклеиновых кислот образца и комплементарных им проб, содержащих флуоресцентную метку. Гибридизация может производиться с ДНК

(метафазные или интерфазные хромосомы) или с РНК. FISH позволяет определить число исследуемых локусов в геноме (при использовании метода 3D-FISH) или последовательность расположения на метафазной хромосоме. Метод является «золотым стандартом» в определении хромосомных патологий — как в клетках с врождёнными перестройками, так и в клетках опухолей.

Данные при помощи метода FISH можно получить, анализируя отсутствие или присутствие сигналов от использованных флюорофоров. Количество различимых цветовых меток равно $(2^x - 1)$, где x — количество флюорофоров. Это позволяет реализовать, например, спектральное кариотипирование (англ. *Spectral Karyotyping, SKY*), при котором каждая хромосома окрашивается в свой собственный цвет и межхромосомные перестройки видны даже начинающему специалисту [11]. Тем не менее, лимитирующими факторами остаются:

- потребность в хорошо обученном персонале. Относительная простота интерпретации результатов сочетается со сложностью протокола приготовления образца, который зависит от характера пробы и образца, и должен быть настроен эмпирически;
- цена реактивов;
- время гибридизации. Кинетика реакций гибридизации в ядре изучена недостаточно, и требуется достаточно долгое время, чтобы получить сигналы, которые можно измерить и сравнить между собой.
- разрешение. Детектировать сигнал от одной молекулы флюорофора очень сложно, такими молекулами должен быть покрыт протяжённый участок ДНК. Поэтому детектировать изменения участков размером менее 100 kbp достаточно затруднительно.

В настоящее время методика FISH значительно усложнилась. Биотехнологические компании предлагают панели олигонуклеотидов, определяющие специфические участки размером от десятков тысяч до миллиона пар оснований, а также олигонуклеотиды с высокой чувствительностью, позволяющие определить сплайс-варианты и даже SNV. Разрабатываются технологии micro-FISH (μFISH), сочетающие FISH с микрофлюидными технологиями (проведение реакций в микроскопических объёмах жидкости). При этом процесс удешевляется, автоматизируется, ускоряется (за счёт уменьшения объёмов, а соответственно, и времени гибридизации) и упрощается для использования в обширных исследованиях и для внедрения в клинику [12].

Сравнительная геномная гибридизация

Как и в случае с методом FISH, основой данного метода (англ. *Comparative Genomic Hybridization, CGH*) является флуоресцентная гибридизация. Однако CGH использует два образца генома — тестовый и контрольный, каждый из которых метится флюорофором, а затем гибридизуется в соотношении $1:1$. Таким образом в тестовом образце можно обнаружить CNV и перестройки.

В отличие от FISH, CGH проверяет весь геном на наличие перестроек и не требует знаний о целевом регионе. К ограничениям анализа относится невозможность выявления полиплоидии, мозаицизма и сбалансированных транслокаций.

В настоящее время CGH используется в виде array-CGH (aCGH), или хромосомного микроматричного анализа (ХМА), при котором CGH комбинируется с микрочиповой гибридизацией [13]. ДНК-микрочипы, или микроматрицы, представляют собой сотни

тысяч или миллионы однонитевых фрагментов ДНК (зондов), которые ковалентно пришиты к основанию (микрочипу). При ХМА на микрочип наносятся контрольные фрагменты генома либо контрольные последовательности генов, которые могут быть связаны с конкретной патологией. Порядок зондов на чипе строго определён, что упрощает локализацию и определение характера перестройки.

С помощью сравнительной гибридизации геномов могут быть обнаружены самые разные структурные вариации — CNV, инверсии, хромосомные транслокации и анэуплоидии. Для этого используются длинные зонды, которые позволяют проводить гибридизацию последовательностей, имеющих некоторые различия. Когда пробы ДНК короткие, эффективность гибридизации очень чувствительна к несовпадениям; такие зонды облегчают сравнение геномов на нуклеотидном уровне (поиск SNV).

Микроматрицы предлагают относительно недорогие и эффективные средства сравнения всех известных типов генетических вариаций. Однако для таких целей, как обнаружение неизвестных или часто повторяющихся последовательностей, эти методы не подходят [14].

Мультиплексная лигаза-зависимая амплификация зонда

Основой мультиплексной лигаза-зависимой амплификации зонда (англ. *Multiplex Ligation-dependent Probe Amplification, MLPA*) является ПЦР-амплификация специальных проб, гибридизующихся с целевыми районами ДНК. Каждая проба представляет собой пару полу-проб; каждая полу-проба имеет комплементарную геному часть и технические последовательности — праймер для ПЦР и вставки, обеспечивающие большой размер продукта амплификации. Если полу-пробы гибридизуются с геномом без зазора, они лигируются и впоследствии амплифицируются; лигированные пробы отличаются от полу-проб с праймером по длине. Длину готового ПЦР-продукта определяют методом электрофореза.

Данная методика подходит для определения CNV, включающих целые гены, а также аномалий метилирования ДНК. Во втором случае используют метил-чувствительные рестриктазы — ферменты, которые по определённым сайтам гидролизуют исключительно метилированную ДНК. Для определения этих участков также применяют электрофорез, т.к. не подвергшаяся гидролизу ДНК по длине значительно превосходит фрагменты гидролизованных рестриктазой метилированных регионов.

Слабым местом MLPA остаётся интерпретация результатов. Определение гомозиготных CNV не представляет труда — их распознают по наличию/отсутствию пика в сравнении с контрольным образцом. Гетерозиготные CNV видны как пики отличающейся высоты, и их поиск требует серьёзную биоинформационную обработку с учётом особенностей конкретной ПЦР-реакции и различий между образцами [15].

2. МЕТОДЫ СЕКВЕНИРОВАНИЯ

Как мы видим, перечисленные методы имеют один серьёзный недостаток — они могут определить наличие или отсутствие, совпадение или несовпадение, но не способны прочитать априори неизвестную последовательность ДНК. Специально для этого были разработаны методы секвенирования.

Секвенирование по Сэнгеру

Исторический метод, позволяющий с высокой точностью анализировать короткие (до 1 kbp) фрагменты ДНК [16]. Суть его состоит в проведении обычной реакции амплификации ДНК, только в смесь дезоксирибонуклеотидов (dNTP) добавлены

дидезоксирибонуклеотиды (ddNTP), которые при присоединении к ДНК обрывают синтез и имеют флуоресцентную или радиоактивную метку (соотношение примерно 100 : 1 соответственно). Таким образом, в процессе амплификации в пробирках образуется смесь из меченых цепей разной длины. При разделении этой смеси на электрофорезе проявляется характерная «лестница», последовательность флуоресцентных сигналов в которой совпадает с последовательностью исследуемой ДНК.

Основным недостатком секвенирования по Сэнгеру является ограничение длины исследуемого фрагмента ДНК.

В настоящее время метод Сэнгера используется для подтверждения вариантов, найденных с помощью методов секвенирования нового поколения.

Секвенирование нового поколения

Секвенирование нового поколения (англ. *New Generation Sequencing, NGS*) — это комплекс технологий, позволяющих прочитать за сравнительно небольшое время миллионы последовательностей ДНК. Благодаря этому единовременно можно проанализировать несколько генов, либо весь геном.

В методах NGS наблюдается развитие двух основных парадигм, различающихся по длине прочтений. Секвенирование короткими прочтениями характеризуется меньшей ценой и более качественными данными, что позволяет применять данные методы в популяционных исследованиях и клинической практике (поиск патогенных генетических вариантов). Секвенирование длинными прочтениями хорошо подходит для сборки новых геномов и изучения отдельных изоформ генов [17]. Количество различных методов в настоящее время значительно, но самым часто используемым является метод Illumina (короткие прочтения).

Основные проблемы данных NGS:

- Финансовые вложения и время, затраченные на секвенирование и анализ данных. По-прежнему остаются лимитирующим фактором применения NGS в клинической практике;
- Ошибки секвенирования и ПЦР. Их значимость уменьшается с увеличением покрытия, но не исчезает полностью;
- Неоднородность покрытия генома или таргетных регионов прочтениями. Это может быть связано как с недостатками приготовления библиотеки, так и с проблемами картирования.

МЕТОДИКИ NGS

Полногеномное секвенирование

Приготовление библиотек при полногеномном секвенировании (англ. *Whole Genome Sequencing, WGS*) производится из всего клеточного материала, либо только из ядер. ДНК фрагментируется таким образом, что достигается относительно ровное покрытие генома.

WGS при достаточной глубине покрытия вполне пригодно для поиска SNV, небольших делеций и инсерций. Полногеномное секвенирование со слабым покрытием может быть использовано для определения CNV — например, при неинвазивном пренатальном тестировании (англ. *Non-Invasive Prenatal Testing, NIPT*), когда используется свободная ДНК плода (англ. *Cell-Free Fetal DNA, cffDNA*), циркулирующая в крови матери [18].

Таргетные панели

Основой данных методов является обогащение целевых регионов генома. Методов обогащения существует достаточно много, но все они делятся на четыре основные категории [19]:

1. Твердофазная гибридизация. Для этого используют комплементарные целевым регионам короткие ДНК-пробы, зафиксированные на твёрдом основании (микрочипе). После гибридизации нецелевую ДНК вымывают, а целевые фрагменты остаются на чипе.

2. Жидкофазная гибридизация. Эти методы характеризуются тем, что ДНК-пробы находятся в растворе и помечены специальной молекулой (например, биотином). После гибридизации с целевой ДНК пробы вылавливают бусинами, поверхность которых способна связывать молекулы биотина.

3. Полимеразно-опосредованный захват. В этих методах ПЦР производят на стадии обогащения. Например, методы молекулярно импринтированных полимеров (англ. *Molecularly Imprinted Polymers, MIP*) и анализа транскриптома одной клетки (англ. *SMART*) используют длинные пробы, содержащие как праймер, так и регион для остановки элонгации и инициации лигирования. После элонгации и лигирования получаются кольцевые молекулы, содержащие целевой регион; линейные молекулы в последующем удаляют из раствора. Метод захвата с помощью расширения праймера (англ. *Primer Extension Capture, PEC*) использует биотинилированные праймеры, которые гибридизуются с целевыми регионами и элонгируются; далее их вылавливают бусинами, как в методах жидкофазной гибридизации.

4. Захват регионов. Включает в себя сортировку и микродиссекцию хромосом, благодаря чему можно обогатить библиотеку фрагментов последовательностями отдельной хромосомы или даже её части. Это методы, требующие чрезвычайно сложных техник и хорошо обученный персонал, но очень полезные в отдельных ситуациях.

Данный вид тестов позволяет анализировать гены, ответственные за отдельные группы заболеваний — например, существуют таргетные панели для иммунодефицитов, почечных, неврологических болезней, болезней соединительной ткани, сетчатки, а также предрасположенности к отдельным видам онкологических заболеваний. Таргетные панели позволяют анализировать и клетки опухолей — некоторые приспособлены к выявлению общих для многих раковых линий мутаций, другие же разработаны для специфического типа опухолей [20].

Полноэкзомное секвенирование

Техника заключается в секвенировании обогащённого экзома — совокупности белок-кодирующих последовательностей клетки. Для этого используют специальные экзомные таргетные панели. Несмотря на то, что существует множество методов таргетного обогащения, конкретно для полноэкзоного секвенирования (англ. *Whole Exome Sequencing, WES*) могут быть использованы лишь немногие из них, а именно — твердофазная и жидкофазная гибридизация [19].

У человека экзом составляет примерно 1 % от генома, или примерно 30 Mbp (суммарно). При этом более 80 % генетических вариантов, которые представлены в

базе данных известных геномных вариантов CLINVAR [21], и из них более 89 % вариантов, которые отмечены как «патогенные», относятся к белок-кодирующим областям генома; эта цифра приближается к 99 %, если учитывать ближайшие окрестности экзонов [22]. Таким образом, полноэкзомное секвенирование намного лучше подходит для обычной клинической практики, нежели полногеномное. Кроме того, полноэкзомное секвенирование значительно дешевле, что увеличивает его доступность и позволяет, например, произвести тестирование ребёнка и родителей (так называемый трио-тест) и, как следствие, улучшить интерпретацию вариантов [20].

Технологии захвата конформации хромосом

Методики захвата конформации хромосом (англ. *Chromosome Conformation Capture, 3C*) позволяют определить расстояние в 3D-пространстве ядра между двумя точками генома. Принцип состоит в том, что интактное ядро фиксируют формальдегидом, ДНК гидролизуют, лигируют, затем продукты лигазной реакции секвенируют при помощи NGS. Во время лигирования ковалентно связанными могут оказаться только те участки, которые физически находятся близко друг от друга. Картирование химерных прочтений с помощью специальных инструментов позволяет узнать, какие именно участки генома были связаны, а значит, располагались близко друг к другу в пространстве ядра [23]. При обработке большого количества 3C-данных геном разделяют на районы фиксированной длины, называемые бинами. Длина бинов называется разрешением; чем меньше длина, тем более высоким считается разрешение. Прочтение, части которого были картированы на два разных бина, называется контактом между этими районами. Практическое значение имеет информация об относительной частоте контактов между бинами.

В настоящее время существует множество вариантов протокола 3C. Самым известным и широко применяемым является метод Hi-C, сочетающий 3C с методами массового параллельного секвенирования. С его помощью можно подсчитать количество контактов во всём геноме — как внутри-, так и межхромосомные контакты [24].

БИОИНФОРМАЦИОННАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ

Результаты NGS представляют собой гигантские блоки данных, содержащие всевозможные ошибки. Биоинформационная обработка данных секвенирования — это высокотехнологичная отрасль, которая позволяет получить из этих данных практически значимую информацию и минимизировать влияние ошибок на эту информацию.

Ниже представлена базовая схема биоинформационной обработки результатов:

Демультиплексикация

В процессе приготовления NGS-библиотеки к целевым фрагментам ДНК лигируют так называемые адаптерные последовательности, или адаптеры. Очень часто потенциальное количество прочтений, которое способен выдать секвенатор за один запуск, значительно превышает требуемое количество прочтений для отдельной библиотеки, поэтому из соображений экономии и повышения производительности на одном чипе секвенируют сразу несколько библиотек. Для этого в адаптеры вставляют баркоды — последовательности, с помощью которых можно отличить прочтения, относящиеся к разным библиотекам или образцам. Процесс сортировки данных секвенирования по баркодам называется демультиплексикацией. Демультиплексикация происходит на стадии обработки данных в секвенаторе, поэтому особенности алгоритма зависят от конкретной используемой платформы.

Удаление адаптерных последовательностей

Если целевой фрагмент ДНК короче длины прочтения, то фрагменты адаптерной последовательности могут попасть в готовые данные. Это замедляет работу алгоритма картирования, а порой в значительной степени ухудшает его результаты, поэтому встаёт вопрос об удалении адаптерных последовательностей. Также присутствие адаптера в прочтениях может быть признаком контаминации библиотеки, и такие прочтения следует исключить из дальнейшего анализа [25]. Основная сложность данного этапа состоит в наличии ошибок в NGS-данных, то есть алгоритм должен допускать наличие в адаптерных последовательностях некоторого количества несоответствий.

Основным инструментом для удаления адаптеров является cutadapt [25]. Он представляет наиболее простое, гибкое и оптимальное решение. Также стоит отметить GATK MarkIlluminaAdapters[26], использующий в качестве входных данных некартированный BAM-файл (англ. *Unmapped Binary sequence Alignment/Map, uBAM*). Это позволяет сохранить важные метаданные секвенатора. Однако uBAM должен использоваться как выходной формат на уровне секвенатора, что не является общепринятой практикой.

Картирование прочтений

Извлечение информации из необработанных результатов секвенирования затруднительно, так как прочтения содержат много ошибок (как в результате ПЦР-реакции, так и допущенные в процессе секвенирования) и не имеют никакой информации о регионе, из которого они произошли. Поэтому прочтения необходимо картировать на некую референсную геномную последовательность. Алгоритм картирования представляет собой очень сложную систему, которая учитывает последовательность букв в прочтении и качество прочтения. Качество прочтения отражает вероятность того, что буква, прочитанная секвенатором, совпадает с реальным нуклеотидом в данной позиции. Обычно качество прочтения записывается в шкале Phred, к которой приводится формулой

$$Q = -10 \log_{10} P, \tag{1}$$

где P — вероятность того, что нуклеотид прочтен правильно.

Было разработано множество алгоритмов картирования, но в настоящее время «золотым стандартом» являются утилиты, использующие алгоритм Берроуса—Уиллера [27].

Алгоритм картирования выставляет выравниванию коэффициент, называемый качеством выравнивания (англ. *MAPping Quality, MAPQ*). MAPQ отражает вероятность правильности картирования и также записывается в шкале Phred (формула 1). В силу размеров референсной последовательности в ней существует огромное множество повторов и похожих регионов. Современные алгоритмы могут находить несколько потенциальных мест картирования для одного прочтения, и их количество влияет на качество выравнивания.

Также алгоритмы способны разделять прочтение на участки, которые могут быть картированы в разные места генома. По этому признаку прочтения делятся на линейные и химерные. В линейных прочтениях не может быть изменения направления картирования, т.е. картированная часть может иметь только прямое направление, либо только обратное направление относительно генома. Химерные прочтения имеют картированные части с разным направлением. Эти участки могут перекрываться, и количество перекрытий также влияет на MAPQ.

Картированный участок может содержать в себе несовпадения с референсной последовательностью, инсерции и делеции. Это могут быть как ошибки, так и генетические варианты, поэтому данная информация безусловно важна при анализе данных. Кроме того, алгоритмы способны картировать прочтения фрагментарно; это важно, например, при анализе транскриптома (RNA-seq). Также в частично картированных прочтениях могут присутствовать некартируемые участки с 3’- или 5’-конца. В отличие от делеций внутри картированных участков, некартированные концы обычно подвергаются так называемому клипированию и в дальнейшем не учитываются при анализе.

Основные проблемы картирования:

- Высоковариативные регионы. Алгоритм картирования разработан для поиска наиболее полных соответствий, и при большом количестве несовпадений прочтение просто не сможет быть картировано на нужный регион генома;

- Вырожденные (неуникальные) регионы. Соответствие между регионами может привести к неправильному распределению прочтений между ними, а значит — и неправильному картированию генетических вариаций. Кроме того, генетические варианты в регионах с короткими повторами в принципе невозможно картировать точно, поэтому обычной практикой является левое смещение (англ. *left-align*).

- Регионы с инсерциями и делециями. Помимо того, что сами по себе эти варианты сильно ухудшают картирование, содержащие их прочтения могут быть картированы неправильно (из-за того, что алгоритмы картирования используют случайно выбранные позиции в геноме для начала поиска соответствий). Из-за этого могут возникать ложные SNP, а пропорции аллелей могут быть посчитаны неправильно.

Для картирования представлено множество решений, как узкоспециализированных, так и универсальных. Из универсальных можно отметить Bowtie2 [28] и BWA [29]. BWA значительно более эффективно работает с химерными ридами, что немаловажно, например, для метода Exo-C. Узкоспециализированные инструменты включают HISAT2[30], работающий со сплайсированными прочтениями и предназначенный для биоинформационной обработки данных RNA-seq.

Удаление дубликатов

Так как молекулы ДНК очень малы, вероятность их разрушения или возникновения в них ошибок велика, а полученные от них сигналы находятся за пределами чувствительности многих современных приборов. Решением этих проблем является амплификация молекул ДНК. Амплификация может быть как на стадии приготовления библиотеки (ПЦР), так и на стадии секвенирования. При секвенировании амплификация и последующее объединение ампликонов в кластер производятся для усиления сигнала и нивелирования ошибок, происходящих на каждом цикле секвенирования с отдельными молекулами. Соответственно, в процессе секвенирования возникают дубликатные прочтения, которые могут быть как ПЦР-дубликатами библиотеки, так и возникать из-за ошибок распознавания кластеров амплификации (оптические дубликаты).

Согласно принятой практике, для улучшения поиска генетических вариантов дубликаты должны быть удалены или помечены [26]. Однако было показано, что для WGS-данных удаление дубликатов имеет минимальный эффект на улучшение поиска полиморфизмов — приблизительно 92 % из более чем 17 млн вариантов были найдены вне

зависимости от наличия этапа удаления дубликатов и использованных инструментов для поиска дубликатов [31]. Особенно это характерно для WGS-данных.

Инструментарий для удаления дубликатов невелик. В первую очередь это MarkDuplicates от Picard [32], который предоставляет мощное и гибкое решение для биоинформатиков. Он способен удалять не только основные дубликаты, но и добавочные выравнивания, если они были предварительно отсортированы по именам [26]. Это важно при обработке Hi-C данных. Также MarkDuplicates учитывает группы прочтений, соответствующие отдельным библиотекам, что позволяет выявлять особенности их приготовления. Более простое (и намного более быстрое) решение для удаления дубликатов предлагает SAMTools [33].

Рекалибровка качества прочтений

В приборной оценке качества прочтений всегда имеют место систематические ошибки. Это связано как с особенностями физико-химических реакций в секвенаторе, так и с техническими недостатками оборудования. Вычисление качества прочтения — сложный алгоритм, защищённый авторскими правами производителя секвенатора. Вместе с тем от качества прочтений напрямую зависит алгоритм поиска вариантов — он использует данный коэффициент как вес в пользу присутствия или отсутствия генетического варианта в конкретной точке генома.

Решением является рекалибровка качества прочтений (англ. *Base Quality Score Recalibration, BQSR*), представляющая собой корректировку систематических ошибок, исходя из известных паттернов зависимости случайных величин. Следует заметить, что рекалибровка не помогает определить, какой нуклеотид в реальности находится в данной позиции — она лишь указывает алгоритму поиска генетических вариантов, выше или ниже вероятность правильного прочтения нуклеотида секвенатором.

Первоочередное влияние на ошибки оказывают:

1. Собственно прибор (секвенатор) и номер запуска. Большая часть секвенаторов выставляет прочтению более высокое качество прочтения по сравнению с ожидаемым, гораздо реже встречаются модели, занижающие качество прочтения [26]. Каждый отдельный запуск может различаться по параметрам чипа и химических реагентов;

2. Цикл секвенирования. Качество прочтения уменьшается с каждым циклом за счёт накопления ошибок в кластере амплификации;

3. Нуклеотидный контекст. Систематические ошибки, связанные с физико-химическими процессами, влияют на качество прочтения нуклеотида в зависимости от предшествующего ему динуклеотида.

Кроме того, алгоритм рекалибровки учитывает изменчивость каждого отдельного сайта, используя базы данных известных генетических вариантов. Высокая изменчивость повышает вероятность правильного прочтения нуклеотида, не совпадающего с референсным в данной позиции генома.

Инструментарий для BQSR предоставлен Институтом Броуд (англ. *Broad Institute of MIT and Harvard*). Согласно оригинальной статье, BQSR рекомендована к использованию для любых данных секвенирования [26].

Поиск точковых генетических вариантов

Невозможно точно сказать, какой нуклеотид находится в каждой позиции генома. Анализ производит специальный алгоритм, который оценивает качество прочтения,

качество выравнивания и процент букв в данной позиции на картированных прочтениях. Отличие генома образца от референсного генома называется генетическим вариантом. Алгоритм выставляет каждому генетическому варианту коэффициент качества варианта (англ. *VCF QUAL*), записываемый в шкале Phred (формула 1). Помимо определения генетического варианта, алгоритм может определять его зиготность.

Также важным этапом поиска вариантов является уже упомянутое выше левое выравнивание. Варианты в повторяющихся последовательностях с длиной менее длины одного прочтения невозможно точно локализовать, поэтому они всегда сдвигаются как можно левее относительно последовательности генома. Это чрезвычайно важно при аннотации генетических вариантов, так как все БД используют данные с левым выравниванием, и неправильная локализация может привести к отсеиванию потенциально патогенного варианта.

Как и в случае с картированием, разнообразие инструментов для поиска вариантов значительно, но в данном случае преобладают универсальные инструменты, использующие разные подходы. Одним из самых широко используемых является GATK HaplotypeCaller. Также стоит отметить freebayes, который учитывает цис-транс-ориентацию вариантов, насколько позволяет длина прочтения[34]. Простое решение предлагает SAMTools [33].

Аннотация, фильтрация и интерпретация результатов

После того, как генетические варианты найдены, можно приступать к поиску тех, которые связаны с конкретной патологией у пациента. Однако только в кодирующих областях генома количество генетических вариантов достигает 100 тыс. (из них около 86 % SNV, 7 % инсерций и 7 % делеций) [35], из них с патологиями связаны единицы. Даже после жёсткой фильтрации приходится работать минимум с сотней подходящих генетических вариантов. Это делает серьёзной проблемой поиск нужного варианта и интерпретацию полученных результатов.

Первое, что следует сделать — это определить, насколько генетический вариант значим для нашего исследования, то есть аннотировать его. Существуют две основных парадигмы аннотации генетического варианта — это аннотация по региону и аннотация по координате.

Основные методы аннотации по региону:

1. Функциональный класс. Для определения функционального класса генетического варианта существуют три основных базы данных: knownGene, refGene и ensGene. Они содержат информацию о генах, их частях и транскриптах — координаты, направление, а также номера экзонов и интронов. Координаты в этих базах данных могут различаться [36], поэтому, во избежание ошибок, рекомендуется использовать их все. Это особенно важно при дифферециации генетических вариантов с высокой вероятностью повреждающего эффекта (сдвиги рамок считывания, нонсенс-кодоны). Кроме того, различаются алгоритмы определения функционального класса в различных утилитах аннотации, что также создаёт определённые трудности [37].

2. Клиническая значимость гена. Количество генетических вариантов для поиска можно сузить, зная, какие именно гены могут быть связаны с наблюдаемым у пациента фенотипом. Для поиска генов по клинической значимости существуют такие базы данных, как OMIM [38] и OrphaData [39].

3. Потеря функции (англ. *Loss of Function, LoF*). Различные показатели, отражающие

устойчивость функции гена, основанные на данных о стоп-кодонах, сдвигах рамки считывания и сплайс-вариантах. Одним из таких показателей является pLI.

Основные проблемы pLI [40]:

- Плохо приспособлен к распознаванию аутосомно-рецессивных вариантов (из-за того, что частота повреждающих вариантов в популяции может быть высокой) и X-сцепленных рецессивных вариантов (из-за наличия в популяции здоровых гетерозиготных носителей).
- Плохо приспособлен к распознаванию генетических вариантов в генах, ответственных за патологии, не влияющие на взросление и воспроизводство. Их частота в популяции также может быть высокой. К таким относятся варианты в генах *BRCA1* и *BRCA2*, ответственных за рак молочной железы.
- Сплайс-варианты априори рассматриваются как повреждающие, несмотря на то, что вариант в сайте сплайсинга может не иметь эффекта на сплайсинг, либо приводить к появлению изоформы белка без потери функции.
- Высокая частота распространения заболевания в контрольной группе. Пример — шизофрения.
- К миссенс-вариантам pLI применять следует с осторожностью, и без клинических данных следует исключить из анализа.
- Также следует отнестись с осторожностью к нонсенс-вариантам и сдвигам рамки считывания в последнем экзоне либо в С-терминальной части предпоследнего. Такие транскрипты избегают нонсенс-индуцированной деградации РНК и могут в результате как не привести к каким-либо функциональным изменениям, так и привести к образованию мутантного белка, обладающего меньшей активностью по сравнению с исходным, либо токсичного для клетки.
- В некоторых случаях соотношение pLI с гаплонедостаточностью конкретного гена в принципе сложно объяснить.

Таким образом, высокое значение pLI можно считать хорошим показателем LoF, низкое — с осторожностью.

Аннотация по координате обычно предназначена для миссенс-, интронных и сплайс-вариантов, связь которых с патологическим состоянием значительно сложнее выявить и доказать.

1. Частота аллеля в популяции. Многие тяжёлые генетические патологии испытывают на себе давление отбора, а значит, вызывающие их генетические варианты не могут иметь высокую частоту в популяции. Фильтрация по частоте является одним из базовых способов фильтрации генетических вариантов. Следует заметить, однако, что низкая частота генетического варианта далеко не всегда связана с его патогенностью, поэтому рассматривать низкую частоту как доказательство патогенности некорректно.

   По мере развития методов NGS и увеличения их доступности, начали появляться базы данных, агрегирующие результаты секвенирования различных популяций,

а значит — способные определить частоту генетических вариантов в популяции. В настоящее время наиболее крупной является gnomAD [41], поглотившая существовавший ранее ExAC, содержавший исключительно экзомные данные. Она содержит частоты генетических вариантов для всех основных рас, а также некоторых условно-здоровых групп.

Несмотря на то, что были созданы базы данных для всех рас, очень часто этого недостаточно и необходимо учитывать частоты в популяциях отдельных народов и стран. Такими базами данных являются GME [42], в которой отражены частоты по популяции Ближнего Востока, ABraOM [43], предоставляющая частоты генетических вариантов среди практически здорового пожилого населения Бразилии. Также для анализа берутся популяции, в которых велика доля близкородственных связей, например, пакистанская [44].

2. Клинические данные из БД и статей. Наиболее достоверным источником данных о патогенности генетического варианта являются семейные и популяционные исследования конкретной патологии, а также базы данных, агрегирующие информацию из подобных статей. Наиболее используемыми в настоящее время являются HGMD [45] и CLINVAR [21]. Тем не менее, CLINVAR считается лишь дополнительным источником, так как часто содержит информацию низкого качества [46].

3. Анализ и предсказание функционального эффекта *in silico*. *In silico* методы появились в ответ на необходимость как-то классифицировать генетические варианты, по которым недостаточно клинической информации. Существует множество способов проверить патогенность таких вариантов *in vitro*, но проверять таким образом все нецелесообразно, а иногда и невозможно. Даже в хорошо изученных генах варианты с неопределённой клинической значимостью могут занимать большую долю — например, в *BRCA1* и *BRCA2* это 33 % и 50 % соответственно. Менее изученные гены, а также пациенты, принадлежащие к популяциям с плохо изученным составом генетических вариантов, представляют ещё большую проблему.

   Поэтому были разработаны инструменты на основе машинного обучения, предсказывающие консервативность районов и патогенность генетических вариантов на основе имеющихся данных — положения относительно гена и его функциональных элементов, характера замены, а также клинической информации об известных заменах [47]. Предсказательная способность отдельных инструментов оставляет желать лучшего, поэтому чаще всего в клинической практике используются агрегаторы, собирающие предсказания с большого числа известных *in silico* инструментов. Таким агрегатором является dbNSFP [48], предоставляющий предсказания для экзомных регионов. Также стоит отметить dbscSNV [49], которая предсказывает эффекты для сплайс-сайтов, и RegSNPIntron [50] — инструмент для предсказания эффектов в интронах.

Значимость вклада каждого отдельного фактора достаточно сложно оценить. Эту проблему решают калькуляторы патогенности, которые по специальным критериям присваивают генетическому варианту ранг, отражающий вероятность повреждающего действия [46].

Аннотация производится специальными инструментами. Они не отличаются по принципу работы, но предоставляют разный набор доступных баз данных. Наиболее широко используемым является Ensembl VEP [51], имеющий веб-интерфейс. Более

простое локальное решение предоставляет ANNOVAR [52]. Недостатком ANNOVAR является неспособность работать с БД в VCF-формате, который в настоящее время являются «золотым стандартом» для хранения данных о генетических вариантах.

ВЫВОДЫ

Несмотря на все достигнутые к настоящему времени успехи, технологиям поиска генетических вариантов предстоит долгий путь. Требуются дальнейшие улучшения как технических методов, так и стратегии обработки данных, чтобы уменьшить количество ошибок и повысить качество обнаружения вариантов.

Для улучшения понимания заболеваний, особенно полигенных, ученым и клиницистам придется объединить информацию из нескольких источников — цитогенетические, геномные, транскриптомные, протеомные и эпигенетические данные. Классические вычислительные методы не в состоянии обрабатывать и извлекать всю возможную информацию из создаваемых наборов данных, и для решения многих задач гораздо лучше подходит искусственный интеллект, который требует гораздо меньше знаний о механизме конкретного процесса. Кроме того, каждый источник данных имеет большое количество особенностей, и работа со всеми слоями информации и их интерпретация невозможна без коллаборации специалистов из разных направлений биологии и медицины.

КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие конфликта интересов.

Список литературы

1. Hiroi, N. & Agatsuma, S. Genetic susceptibility to substance dependence. *Mol Psychiatry*. 2004. No. 10. P. 336–344. doi: 10.1038/sj.mp.4001622.

2. Galanello, R. & Origa, R. Beta-thalassemia. *Orphanet J Rare Dis*. 2010. No. 5. doi: 10.1186/1750-1172-5-11.

3. Jett, K. & Friedman, J. M. Clinical and genetic aspects of neurofibromatosis 1. *Genet Med*. 2009. No. 12. P. 1–11. doi: 10.1097/gim.0b013e3181bf15e3.

4. *When Children Die: Improving Palliative and End-of-Life Care for Children and Their Families*. Ed.: Field, M. J. & Behrman, R. E. Washington (DC): National Academies Press (US), 2003. URL: https://pubmed.ncbi.nlm.nih.gov/25057608. NBK220818[bookaccession].

5. Herder, M. What is the purpose of the orphan drug act? *PLoS Med*. 2017. No. 14. P. e1002191. doi: 10.1371/journal.pmed.1002191.

6. Richter, T. *et al.* Rare disease terminology and definitions—a systematic global review: Report of the ISPOR Rare Disease Special Interest Group. *Value in Health*. 2015. No. 18. P. 906–914. doi: 10.1016/j.jval.2015.05.008.

7. Bomar, J. M. *et al.* Mutations in a novel gene encoding a CRAL-TRIO domain cause human Cayman ataxia and ataxia/dystonia in the jittery mouse. *Nat Genet*. 2003. No. 35. P. 264–269. doi: 10.1038/ng1255.

8. The Lancet Neurology. Rare neurological diseases: a united approach is needed. *The Lancet Neurology*. 2011. No. 10. P. 109. doi: 10.1016/s1474-4422(11)70001-1.

9. Schreck, R. R. & Distèche, C. M. Chromosome banding techniques. *Current protocols in human genetics*. 2001. No. Chapter 4. P. Unit4.2–Unit4.2. doi: 10.1002/0471142905.hg0402s00. 18428280[pmid].

10. Sampson, B. & McGuire, A. Genetics and the molecular autopsy. In *Pathobiology of Human Disease*, 3459–3467 Elsevier, 2014. URL: https://doi.org/10.1016/b978-0-12-386456-7.06707-1.

11. Guo, B., Han, X., Wu, Z., Da, W. & Zhu, H. Spectral karyotyping: an unique technique for the detection of complex genomic rearrangements in leukemia. *Translational pediatrics*. 2014. No. 3. P. 135–139. doi: 10.3978/j.issn.2224-4336.2014.01.02. 26835331[pmid].

12. Huber, D., von Voithenberg, L. V. & Kaigala, G. Fluorescence *in situ* hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*. 2018. No. 1. P. 15–24. doi: 10.1016/j.mne.2018.10.006.

13. Theisen, A. Microarray-based Comparative Genomic Hybridization (aCGH). *Nature Education*. 2008. No. 1. P. 45. URL: https://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-4

14. Gresham, D., Dunham, M. J. & Botstein, D. Comparing whole genomes using DNA microarrays. *Nat Rev Genet*. 2008. No. 9. P. 291–302. doi: 10.1038/nrg2335.

15. Stuppia, L., Antonucci, I., Palka, G. & Gatta, V. Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases. *IJMS*. 2012. No. 13. P. 3245–3276. doi: 10.3390/ijms13033245.

16. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*. 1977. No. 74. P. 5463–5467. doi: 10.1073/pnas.74.12.5463.

17. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016. No. 17. P. 333–351. doi: 10.1038/nrg.2016.49.

18. Yu, D. *et al.* Noninvasive prenatal testing for fetal subchromosomal copy number variations and chromosomal aneuploidy by low-pass whole-genome sequencing. *Mol Genet Genomic Med*. 2019. No. 7. doi: 10.1002/mgg3.674.

19. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics*. 2010. No. 19. P. R145–R151. doi: 10.1093/hmg/ddq333.

20. Yohe, S. & Thyagarajan, B. Review of clinical next-generation sequencing. *Archives of Pathology & Laboratory Medicine*. 2017. No. 141. P. 1544–1557. doi: 10.5858/arpa.2016-0501-ra.

21. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*. 2017. No. 46. P. D1062–D1067. doi: 10.1093/nar/gkx1153.

22. Barbitoff, Y. A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci Rep*. 2020. No. 10. doi: 10.1038/s41598-020-59026-y.

23. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009. No. 326. P. 289–293. doi: 10.1126/science.1181369.

24. Oluwadare, O., Highsmith, M. & Cheng, J. An overview of methods for reconstructing 3D chromosome and genome structures from Hi-C data. *Biol Proced Online*. 2019. No. 21. doi: 10.1186/s12575-019-0094-0.

25. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.*. 2011. No. 17. P. 10. doi: 10.14806/ej.17.1.200.

26. Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*. 2013. No. 43. doi: 10.1002/0471250953.bi1110s43.

27. Burrows, M. & Wheeler, D. A block-sorting lossless data compression algorithm. Tech. Rep., Palo Alto, CA: Digital Equipment Corporation. 1994.

28. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie2. *Nat Methods*. 2012. No. 9. P. 357–359. doi: 10.1038/nmeth.1923.

29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009. No. 25. P. 1754–1760. doi: 10.1093/bioinformatics/btp324.

30. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015. No. 12. P. 357–360. doi: 10.1038/nmeth.3317.

31. Ebbert, M. T. W. *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*. 2016. No. 17. doi: 10.1186/s12859-016-1097-3.

32. Broad Institute. *Picard Toolkit*. 2019. GitHub repository: `http://broadinstitute.github.io/picard/`. Accessed 2021/01/10.

33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009. No. 25. P. 2078–2079. doi: 10.1093/bioinformatics/btp352.

34. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *preprint arXiv:1207.3907 [q-bio.GN]*. 2012. URL: https://doi.org/10.1038/nmeth.3317.

35. Supernat, A., Vidarsson, O. V., Steen, V. M. & Stokowy, T. Comparison of three variant callers for human whole genome sequencing. *Sci Rep*. 2018. No. 8. doi: 10.1038/s41598-018-36177-7.

36. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*. 2014. No. 6. P. 26. doi: 10.1186/gm543.

37. Jesaitis, A. The state of variant annotation: A comparison of AnnoVar, snpEff and VEP. Tech. Rep., The Golden Helix Blog (GHB), `https://blog.goldenhelix.com/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/`. 2014.

38. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research*. 2014. No. 43. P. D789–D798. doi: 10.1093/nar/gku1205.

39. French National Institute of Health and Medical Research (INSERM). *Orphanet: an online database of rare diseases and orphan drugs.*. 1997. Available online at: `http://www.orpha.net`. Accessed 2020/12/08.

40. Ziegler, A., Colin, E., Goudenège, D. & Bonneau, D. A snapshot of some pLI score pitfalls. *Human Mutation*. 2019. doi: 10.1002/humu.23763.

41. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020. No. 581. P. 434–443. doi: 10.1038/s41586-020-2308-7.

42. Scott, E. M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet*. 2016. No. 48. P. 1071–1076. doi: 10.1038/ng.3592.

43. Naslavsky, M. S. *et al.* Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Human Mutation*. 2017. No. 38. P. 751–763. doi: 10.1002/humu.23220.

44. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*. 2017. No. 544. P. 235–239. doi: 10.1038/nature22034.

45. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet*. 2017. No. 136. P. 665–677. doi: 10.1007/s00439-017-1779-6.

46. Ryzhkova, O. *et al.* Guidelines for the interpretation of massive parallel sequencing variants. *Medical Genetics*. 2017. No. 16. P. 4–17. URL: https://www.medgen-journal.ru/jour/article/view/308/224.

47. Brea-Fernandez, A. *et al.* An update of *in silico* tools for the prediction of pathogenesis in missense variants. *CBIO*. 2011. No. 6. P. 185–198. doi: 10.2174/1574893611106020185.

48. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation*. 2016. No. 37. P. 235–241. doi: 10.1002/humu.22932.

49. Jian, X., Boerwinkle, E. & Liu, X. *In silico* tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med*. 2013. No. 16. P. 497–503. doi: 10.1038/gim.2013.176.

50. Lin, H. *et al.* RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol*. 2019. No. 20. doi: 10.1186/s13059-019-1847-4.

51. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol*. 2016. No. 17. doi: 10.1186/s13059-016-0974-4.

52. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010. No. 38. P. e164–e164. doi: 10.1093/nar/gkq603.

Рукопись поступила в редакцию 01.02.2021.

Переработанный вариант поступил 07.02.2021.

Дата опубликования 15.02.2021.

Mathematical Biology and Bioinformatics
*doi: 10.17537/2020.15.****

========== **SECTION** ==========

UDC: 616-074+57.087.1

Methods of clinically significant variants discovery in human genome

Valeev E.*[1,2], **Fishman V.****[1,2]

[1]*Sector of Genomic Mechanisms of Ontogenesis SB RAS, Novosibirsk, Russia*
[2]*Novosibirsk State University, Novosibirsk, Russia*

Abstract. Clinical genetics plays significant role in rare hereditary diseases diagnostics. Its application is not limited to diagnostics only; it can be used to predict and treat diseases, to develop personalized therapeutic methods. There are plenty methods of variant discovery, technical and bioinformatics solutions. This review highlights both well-established and recently developed methods, as well as their possible future. First, we say about cytogenetic methods, such as karyotyping, FISH, CGH and MLPA, their advantages, and disadvantages. Further we analyze sequencing technologies, especially next generation sequencing (NGS). In conclusion, we say about NGS data processing methods and try to prospect their future, because, nevertheless NGS technologies and associated bioinformatics methods have been developed fast, diagnostics of polygenic and genetically heterogeneous diseases remain challenging.

Key words: *Hereditary diseases, clinical genetics, cytogenetics, NGS*