

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ
ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

Институт медицины и психологии В. Зельмана НГУ

КУРСОВАЯ РАБОТА

Валеев Эмиль Салаватович
Группа 12452

Тема работы: «Разработка инструментов для поиска клинически значимых
полиморфизмов в геноме человека на основе данных секвенирования ЗС-библиотек»

Научный руководитель:

Фишман Вениамин Семенович,
к.б.н., ведущий научный сотрудник,
заведующий Сектором геномных
механизмов онтогенеза, ИЦиГ СО РАН

ФИО: _____ / _____
«_____» _____ 20____ г.

Оценка: _____

Новосибирск, 2020

Содержание

1	Введение	3
1.1	Актуальность	3
1.2	Цели	3
1.3	Задачи	3
2	Обзор литературы	3
2.1	Механизмы развития генетических патологий	5
2.2	Типы генетических аномалий, лежащих в основе генетических патологий	6
2.3	Функциональные классы генетических вариантов	7
2.4	Методы детектирования генетических вариантов	7
2.5	Виды NGS	11
2.6	Базовая схема обработки результатов секвенирования	12
2.7	Аннотация, фильтрация и интерпретация результатов	14
2.8	Когортный и семейный анализ	16
2.9	Случайные находки	16
2.10	Ехо-С: суть метода	16
3	Материалы и методы	17
4	Результаты	20
5	Обсуждение результатов	20
6	Предварительные выводы	20
A	Данные секвенирования K562	21

Список сокращений

3C (Chromosome Conformation Capture) — определение конформации хромосом

BAM

BQSR (Base Quality Score Recalibration) —

cfDNA (Circulating Free DNA) —

cffDNA (Cell-Free Fetal DNA) —

CGH (Comparative Genomic Hybridization) — сравнительная геномная гибридизация

CNV (Copy Number Variation) — вариация числа копий

Ехо-С

FISH (Fluorescence In Situ Hybridization) — флуоресцентная *in situ* гибридизация

GATK (Genome Analysis ToolKit)

Hi-C (all-vs-all chromosome conformation capture) —

LoF (Loss of Function) —

MAPQ (MAPping Quality) — качество картирования

MLPA (Multiplex Ligation-dependent Probe Amplification) — мультиплексная лигат-зависимая амплификация зонда

NGS (New Generation Sequencing) — секвенирование нового поколения

NIPT (Non-Invasive Prenatal Testing) —

NOR (Nucleolus Organizer Region) — ядрышковый организатор

RG (Read Group)

SNV (Single Nucleotide Variant)

TAD (Topologically Associated Domain) — топологически ассоциированные домены

UTR (UnTranslated Regions) — нетранслируемая область

VCF (Variant Call Format) —

WES (Whole Exome Sequencing) — полноэкзомное секвенирование

WGS (Whole Genome Sequencing) — полногеномное секвенирование

ДНК — дезоксирибонуклеиновая кислота

ПЦР — полимеразная цепная реакция

РНК — рибонуклеиновая кислота

ХМА — хромосомный микроматричный анализ

1. Введение

1.1. Актуальность

1.2. Цели

1.3. Задачи

2. Обзор литературы

Генетические болезни (моногенные заболевания, геномные структурные дефекты, вариации числа копий) — это основная причина смертности детей до 10 лет.

Генетическое детерминирование патологий

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)[38]. 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[39].

Также сложности 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 позитивный отбор). Частным случаем эффекта основателя является атаксия Каймановых островов (ATCAU), 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[53]. Примером позитивного отбора является 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[52]. 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 — 5000–8000[40]. Существуют различные базы данных, собирающие информацию по орфанным заболеваниям, наиболее известными и часто используемыми из них являются:

1. Global Genes
2. Online Mendelian Inheritance in Man (OMIM)[24]
3. Orphadata

Около 80% редких болезней имеют генетическую природу и начинаются в раннем детстве[40]. 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.

2.1. Механизмы развития генетических патологий

Механизмы развития генетических патологий делятся на две большие группы. В первую относят изменения белок-кодирующей последовательности гена, приводящие к прекращению синтеза белка либо к синтезу изменённого полипептида. Ко второй группе относятся эпигенетические механизмы, не затрагивающие непосредственно белок-кодирующие последовательности генов.

Изменения белок-кодирующей последовательности гена (экзонов и сплайсинг-сайтов) могут приводить к замене аминокислот, сдвигам рамки считывания, появлению преждевременных стоп-кодонов и нарушениям сплайсинга. Прекращение синтеза белка снижает дозу гена, а изменённый полипептид способен как потерять свою функцию, снизив таким образом дозу гена, так и приобрести новые свойства (токсичность). Классическим примером приобретения белком токсичности являются некоторые наследственные нейродегенеративные заболевания — в частности, аутосомно-доминантный вариант болезни Альцгеймера. Другое нейродегенеративное заболевание — аутосомно-рецессивная болезнь Паркинсона — может служить примером потери белком протективной функции[57].

Также генетические патологии могут развиваться из-за эпигенетических механизмов, приводящих к изменению экспрессии генов. К таким механизмам можно отнести метилирование ДНК — изменение молекулы ДНК без изменения нуклеотидной последовательности, а также ацетилирование гистонов.

В частности, нарушение метилирования ДНК ответственно за развитие синдрома Беквита—Видемана. Экспрессия генов CDKN1C и IGF2 регулируется в зависимости от того, на материнской или отцовской хромосоме они находятся (явление геномного импринтинга). Потеря импринтинга, вызванная изменениями регуляторного района, ведёт к изменению экспрессии этих генов и, как следствие, к тяжёлым порокам развития, включающим висцеромегалию, висцеральные грыжи, эмбриональные опухоли, пороки сердца и почек[54]. Изменение ацетилирования гистонов некоторых генов в клетках головного мозга связано с развитием такого заболевания, как шизофрения[55].

Кроме того, на экспрессию в значительной степени влияет трёхмерная структура хроматина. К примеру, энхансерный район не обязательно находится в непосредственной близости от гена, для его работы необходим физический контакт с промотором гена за счёт выпетливания ДНК. Белковый комплекс, связанный с энхансером, привлекает в эту область РНК-полимеразу и увеличивает вероятность её связывания с промотором. Известно, что большая часть промотор-энхансерных взаимодействий находится внутри топологически ассоциированных доменов (TAD). В результате разрушения старых или образования новых границ TAD формируются структурные варианты, характеризующиеся иными промотор-энхансерными взаимодействиями. Это является причиной таких состояний, как FtM-инверсия пола (ген SOX9) и синдром Кукса (ген KCNJ2)[56].

В основе как генетических, так и эпигенетических механизмов чаще всего лежат варианты тех или иных участков ДНК. Эти варианты существенно различаются по размеру, характеру изменения, а также функциональному значению, которое напрямую зависит от затрагиваемых вариантов районов генома.

2.2. Типы генетических аномалий, лежащих в основе генетических патологий

Самыми крупными являются хромосомные аномалии. Основные их типы включают:

- Анеуплоидии — изменение числа хромосом. Примерами анеуплоидий могут служить синдром Дауна (трисомия 21 хромосомы), Эдвардса (трисомия 18 хромосомы), Патау (трисомия 13 хромосомы), а также вариации числа половых хромосом (синдром Тёрнера, Клайнфельтера и другие). Частичная моносомия — синдром кошачьего крика (связан с утратой плеча 5 хромосомы). Прочие анеуплоидии ведут к несовместимым с жизнью нарушениям эмбрионального развития и, как следствие, спонтанным абортam.
- Инверсии — переворот фрагмента хромосомы. Крупные инверсии могут быть причиной изменения границы TAD, а также загираания кроссинговера и образования гаплогрупп.
- Транслокации — перемещение фрагмента хромосомы с одного места на другое. Выделяют сбалансированные транслокации (перемещение фрагмента на негомологичную хромосому), несбалансированные (перемещение на гомологичную хромосому), реципрокные (взаимный обмен участками между негомологичными хромосомами) и Робертсоновские (слияние акроцентрических хромосом с образованием метацентрической или субметацентрической). Фенотипическое проявление транслокаций может быть различным — если сбалансированные, реципрокные и Робертсоновские транслокации почти не проявляются (сказываясь иногда только на фертильности), то несбалансированные могут привести к задержкам развития, тяжёлым порокам и даже летальному исходу.

Отдельно следует выделить вариации числа копий (CNV). К ним относятся дупликации (мультипликации) и делеции хромосомных сегментов размером от тысячи до нескольких миллионов пар оснований. CNV способны увеличивать или уменьшать дозу гена, в значительной степени влияя на его экспрессию. Различия в количестве копий могут носить как положительный характер, так и отрицательный — в частности, дупликации в гене CCL3L1 способны увеличить устойчивость к ВИЧ[58], а крупные CNV в разных частях генома ассоциированы с расстройствами аутистического спектра[59].

Самыми небольшими — но не менее важными — являются точечные полиморфизмы (SNV) и короткие инсерции и делеции (indels) размером 20–50bp, которые могут приводить как к генетическим, так и к эпигенетическим изменениям. Чаще всего это наследуемые генетические варианты, которые из поколения в поколение проявляются у членов семьи как определённый фенотип. Однако существуют и генетические варианты *de novo*, приводящие к развитию патологий. Согласно оценкам, предоставленным [60], в среднем в каждом поколении у человека возникают 44–82 SNV *de novo*, из них 1–2 приходится на белок-кодирующие регионы. Число небольших инсерций и делеций оценивается в 2.9–9 на геном, крупные перестройки встречаются значительно реже. Также известно, что количество генетических вариантов *de novo* непрерывно растёт в течение жизни человека.

2.3. Функциональные классы генетических вариантов

Как уже было упомянуто выше, значение генетических вариантов напрямую зависит от их положения относительно функциональных частей генома. Варианты могут находиться как внутри генов, так и вне их.

Области гена, в которые может попасть генетический вариант:

- Экзоны, непосредственно отвечающие за последовательность белка. Генетические варианты в экзонах могут быть синонимичными (без замены аминокислоты) и несинонимичными — миссенс (замена на другую аминокислоту), нонсенс (замена на стоп-кодон) либо сдвиг рамки считывания, приводящий к изменению значительной части белковой молекулы. Миссенс-варианты редко приводят к утрате функции белка, но они могут повлиять на экспрессию гена, если замена пришлась на регуляторный мотив[46].
- Интроны, которые содержат регуляторные области и сплайсинг-сайты, необходимые для процессинга транскрипта в готовую мРНК, а также 3'-нетранслируемая область (3' UTR) и 5'-нетранслируемая область (5' UTR), вовлечённые в регуляцию транскрипции, трансляции и деградации транскрипта. Влияние генетических вариантов в этих областях недостаточно изучено, и их связь с конкретной патологией у пациента порой достаточно трудно доказать. Тем не менее, существуют специальные инструменты, позволяющие оценить патогенность таких вариантов. Интронные и UTR генетические варианты обычно рассматриваются в случае, если иного объяснения фенотипу пациента не было найдено.

Внегенные варианты могут приходиться на различные регуляторные последовательности, например, энхансеры, сайленсеры, а также сайты связывания белков, отвечающих за процессы метилирования или трёхмерную организацию хроматина.

Как мы видим, типов генетических вариантов существует огромное множество, они в значительной степени различаются между собой, и их определение может представлять трудность для исследователя. На сегодняшний день разработано множество методик, облегчающих эту задачу. О них речь пойдёт ниже.

2.4. Методы детектирования генетических вариантов

Кариотипирование. Данный метод представляет собой микроскопическое исследование клеток, остановленных на стадии метафазы. Однако простое микроскопическое исследование хромосом достаточно затруднительно, поэтому были разработаны различные методы окраски (бэндинга), позволяющие отдифференцировать отдельные хромосомы и хромосомные регионы[61]:

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 варибельной 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 (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 5 Mbp) и процент поражённых клеток в образце (минимум 5–10%)[49].

В целом классический метод кариотипирования, достаточно 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 является гибридизация нуклеиновых кислот образца и комплементарных им проб, содержащих флуоресцентную метку. Гибридизация может производиться на ДНК (метафазные или интерфазные хромосомы) и на РНК. FISH позволяет определить количественные характеристики нуклеиновых кислот и их пространственное расположение в ядре. 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 различных цветовых меток равно $(2^x - 1)$, где x — количество флюорофоров. Это позволяет реализовать, например, спектральное кариотипирование (SKY), при котором каждая хромосома окрашивается в свой собственный цвет и межхромосомные перестройки видны

даже начинающему специалисту[62]. Тем не менее, лимитирующими факторами остаются:

- потребность в хорошо обученном персонале. Относительная простота интерпретации результатов сочетается со сложностью протокола приготовления, 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.

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 сплайсинг-варианты и даже SNV. Разрабатываются технологии micro-FISH (μ FISH), 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[41].

Сравнительная геномная гибридизация (CGH). Как и в случае с методом FISH, основой данного метода 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, не требует знаний о целевом регионе и может быть использован на интерфазных клетках. Однако, как и у классических методов, разрешение CGH ограничено 5–10Mbp. 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[42]. Гибридизация между комплементарными цепями ДНК позволяет исследовать неизвестную ДНК путем сравнения с ДНК известной последовательности. Для этой цели были созданы ДНК-микрочипы, или микроматрицы. Они 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 анеуп-

лоидии. 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[50].

Мультиплексная лигат-зависимая амплификация зонда (MLPA). Основой 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 целых генов, 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 гидролизированных рестриктазой метилированных регионов.

Слабым местом MLPA остаётся интерпретация результатов. Определение гомозиготные 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[43].

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 короткий (до 1kbp) фрагмент ДНК[29]. Суть его состоит в проведении обычной реакции репликации 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 репликации в пробирках образуется смесь из меченых цепей разной длины. При разделении этой смеси на электрофорезе образуется характерная «лестница», последовательность флуоресцентных сигналов в которой совпадает с последовательностью исследуемой ДНК.

Основным недостатком секвенирования по Сэнгеру является ограничение длины исследуемого фрагмента ДНК. Также в случае, если исследуемая матрица была клонирована, а не амплифицирована методом ПЦР, она может быть контаминирована вектором клонирования, что затрудняет прочтение последовательности.

В настоящее время метод Сэнгера используется для подтверждения вариантов, найденных с методов NGS.

NGS. Секвенирование нового поколения (NGS) — это комплекс технологий, позволяющих прочитать за сравнительно небольшое время миллионы коротких последовательностей ДНК. Благодаря этому одновременно можно проанализировать несколько генов, либо весь геном (в отличие от традиционных методов).

Проблемы данных NGS:

- Ошибки секвенирования.
- Неоднородность покрытия генома.
- Ошибки ПЦР и ПЦР-дубликаты.
- Неточное выравнивание инделов и повторяющихся последовательностей (например, поли-А трактов).

В настоящее время NGS используется и для выявления крупных перестроек (метод Hi-C).

2.5. Виды NGS

Полногеномное секвенирование (WGS). Приготовление библиотек при полногеномном секвенировании производится из всего клеточного материала, либо только из ядер. ДНК фрагментируется с помощью специфических рестриктаз либо точно подобранных доз ДНКаз. Таким образом достигается относительно ровное покрытие генома.

WGS при достаточной глубине покрытия вполне пригоден для поиска SNP и небольших инделов. Полногеномное секвенирование со слабым покрытием может быть использовано для определения CNV — например, NIPT, когда используется свободная ДНК плода (cfDNA), циркулирующая в крови матери[37]. WGS также может быть использовано в диагностике микробиома с целью определения источника хронической инфекции, реконструкции путей передачи инфекции, а также выявления антибиотикорезистентных штаммов[28].

Полноэкзомное секвенирование (WES) Техника заключается в секвенировании экзона — совокупности белок-кодирующих последовательностей клетки. Это достигается с помощью обогащения последовательностей, гибридованных с олигонуклеотидами, фланкирующими экзоны.

В помощь WES нельзя производить поиск вариантов в некодирующих регионах, но, тем не менее, этот метод неплохо себя зарекомендовал, так как значительная часть

патогенных вариантов располагается в экзонах и вблизи них. 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[27].

Таргетные панели. Как и в случае с WES, основой является обогащение целевых регионов. Отличие заключается в том, что целевые регионы подбираются специфически для конкретной цели. 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. Также таргетные панели позволяют анализировать клетки опухолей — некоторые приспособлены к выявлению общих для многих раковых линий мутаций, другие же разработаны для специфического типа опухолей[27].

Технологии захвата конформации хромосом (3C). Данные методики позволяют определить расстояние между двумя точками генома, используя информацию об относительной частоте контактов между ними. Принцип состоит в том, что интактное ядро (бактериальная хромосома) фиксируется, ДНК разрезается, лигируется, затем снова разрезается и секвенируется. Во время лигирования связанными могут оказаться участки, которые физически находятся близко друг от друга. Картирование химерных прочтений с помощью специальных инструментов позволяет узнать, какие именно участки генома были связаны[44].

В настоящее время существует множество вариантов протокола 3C — 4C, 5C, Hi-C, TCC, ChIA-PET и scHi-C. 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[45].

2.6. Базовая схема обработки результатов секвенирования

Демультиплексирование. В процессе секвенирования к целевым фрагментам ДНК могут пришиваться так называемые адаптерные последовательности. Эти адаптеры могут содержать так называемые баркоды — последовательности, с помощью которых можно отличить ДНК различных образцов. Процесс сортировки данных секвенирования по баркодам называется демультиплексированием. Чаще всего демультиплексирование производится самим секвенатором, но иногда его приходится производить вручную.

Удаление адаптерных последовательностей. Если целевая ДНК короче длины прочтения, то фрагменты адаптера на 3' конце могут попасть в готовые данные, и встаёт вопрос об их удалении. Также присутствие адаптера в прочтениях может быть признаком контаминации, и такие прочтения следует исключить из дальнейшего анализа[13].

Картирование. Прочтения необходимо картировать на некую референсную геномную последовательность. Алгоритм картирования представляет собой очень сложную систему, которая учитывает последовательность букв в прочтении и их качество. В настоящее время «золотым стандартом» являются утилиты, использующие алгоритм Берроуса–Уиллера[30].

Основные термины картирования:

Качество выравнивания (MAPQ) — коэффициент, выставляемый конкретному выравниванию алгоритмом.

Первичное выравнивание (primary) — выравнивание наиболее крупного фрагмента прочтения с наиболее высоким MAPQ. Первичное выравнивание только одно.

Вторичное выравнивание (secondary) — выравнивание наиболее крупного фрагмента прочтения с меньшим MAPQ. Вторичных выравниваний может быть несколько.

Добавочное выравнивание (supplementary) — выравнивание менее крупных фрагментов прочтения.

Мягкое клипирование (soft-clip) — отсечение невыровненного конца прочтения с сохранением полной последовательности прочтения. Мягкому клипированию подвергаются прочтения, содержащие адаптеры, а также химерные первичные выравнивания.

Жёсткое клипирование (hard-clip) — отсечение невыровненного конца прочтения без сохранения его последовательности. Жёсткому клипированию подвергаются добавочные выравнивания.

Проблемы картирования:

- Высоковариативные регионы
- Вырожденные (неуникальные) регионы
- Регионы с инделами

Удаление дубликатов. Дубликатные прочтения могут появляться как в результате особенностей приготовления библиотеки (ПЦР-дубликаты), так и из-за ошибок распознавания кластеров амплификации (оптические дубликаты). Согласно принятой практике, дубликаты должны быть удалены или помечены для улучшения поиска вариантов[1].

Тем не менее, 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[31]. Учитывая, что удаление дубликатов может занимать значительную часть потраченного на обработку данных времени и ресурсов компьютера, следует взвесить пользу и затраты данного этапа для конкретной прикладной задачи.

Рекалибровка качества прочтений (BQSR). Приборная оценка качества оснований не соответствует эмпирической, имеют место систематические ошибки прочтения. Вычисление качества каждой отдельной буквы — сложный алгоритм, защищённый авторскими правами производителя секвенатора. Вместе с тем от качества букв напрямую зависит алгоритм поиска вариантов — он использует качество как вес в пользу присутствия или отсутствия варианта в конкретной точке генома.

Решением является рекалибровка качества прочтений, представляющая собой корректировку систематических ошибок, исходя из известных паттернов ковариации. Первоочередное влияние на ошибки оказывают цикл секвенирования и нуклеотидный контекст (динуклеотид). Кроме того, алгоритм рекалибровки учитывает ещё и изменчивость каждого отдельного сайта, используя базы данных известных вариантов.

Разработчики GATK настоятельно рекомендуют использовать BQSR для любых данных секвенирования[1].

Поиск вариантов. Невозможно точно сказать, какая буква находится в каждой позиции генома; анализ производит специальный алгоритм, который оценивает качество оснований (т.е. вероятность, в которой буква присутствует в прочтении) и процент букв в позиции на картированных прочтениях. Отличие генома образца от референсного генома называется вариантом (синонимичные термины «мутация» и «полиморфизм» не рекомендованы к употреблению[32]). Помимо определения буквы, алгоритм может определять зиготность.

Также важным этапом поиска вариантов является так называемое выравнивание по левому краю (left-alignment). Варианты в повторяющихся последовательностях с длиной менее длины одного прочтения невозможно точно локализовать, поэтому они всегда сдвигаются как можно левее. Это чрезвычайно важно при аннотации вариантов, так как все базы данных используют варианты с left-alignment, и неправильная локализация может привести к отсеиванию потенциально патогенного варианта.

2.7. Аннотация, фильтрация и интерпретация результатов

Номер экзона, функциональный класс варианта. (разрыв)

Частота аллеля по основным базам данных. По мере развития методов NGS и увеличения их доступности, начали появляться базы данных, агрегирующие результаты секвенирования различных популяций, а значит — способные определить частоту вариантов в популяции. В настоящее время наиболее крупной является gnomAD[22], поглотившая существовавший ранее ExAC, содержащий исключительно экзомные данные. Она содержит частоты вариантов для всех основных рас, а также некоторых условно-здоровых групп (non-neuro, non-torped и т.п.)

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[21], в которой отражены частоты по популяции Ближнего Востока, ABraOM[26], предоставляющая частоты вариантов среди практи-

чески здорового пожилого населения Бразилии. Также для анализа берутся популяции, в которых велика доля близкородственных связей, например, пакистанская[35].

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 rLI.

Основные проблемы rLI[34]:

- Плохо приспособлен к распознаванию AR вариантов (из-за того, что частота повреждающих вариантов в популяции может быть высокой) и XR вариантов (из-за наличия в популяции здоровых гетерозиготных носителей).
- Плохо приспособлен к распознаванию 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 BRCA1-2.
- Сплайс-варианты рассматриваются как повреждающие, несмотря на то, что вариант в сплайс-сайте 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 rLI применять следует с осторожностью, и без функциональной пробы следует исключить из анализа.
- Также следует относиться с осторожностью к нонсенс-вариантам и сдвигам рамки считывания в последнем экзоне либо в С-терминальной части предпоследнего. Такие транскрипты избегают нонсенс-индуцированного разложения 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 rLI с гаплонедостаточностью конкретного гена в принципе сложно объяснить.

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

Клинические данные из бд и статей. Наиболее достоверным источником данных о патогенности варианта являются исследования, которые изучали конкретный вариант, а также базы данных, агрегирующие информацию из этих статей. Наиболее используемыми в настоящее время являются HGMD[23] и CLINVAR. Тем не менее, CLINVAR считается лишь дополнительным источником, так как часто содержит информацию низкого качества[51].

Анализ и предсказание функционального эффекта *in silico*. *In silico* методы появились в ответ на необходимость как-то классифицировать варианты, по которым недостаточно клинической информации. Существует множество способов проверить патогенность таких вариантов *in vitro*, но проверять таким образом все варианты 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 вариантов, представляют ещё большую проблему.

Поэтому существуют инструменты на основе машинного обучения, предсказывающие консервативность районов и патогенность вариантов на основе имеющихся данных — положения относительно гена и его функциональных элементов, характера замены, а также клинической информации об известных заменах[46].

2.8. Когортный и семейный анализ

Анализ группы, представители которой связаны узами крови (семейный анализ), либо патологией или вариантом фенотипа (когортный анализ).

Семейный анализ нужен для установления путей наследования тех или иных вариантов в родословной. Это позволяет уточнить степень их корреляции с фенотипом. Также с помощью семейного анализа можно находить мутации *de novo*.

Когортный анализ позволяет, например, оценить частоты вариантов в исследуемой и контрольной группе. Помимо этого существует необходимость детекции систематических отклонений покрытия и артефактов выравнивания, связанных с конкретными районами генома и/или особенностями приготовления библиотек. Также анализ нескольких родственных образцов помогает определить зиготность варианта либо импутировать район с недостаточным покрытием.

2.9. Случайные находки

Несмотря на то, что точность определения патогенности вариантов достаточно невысокая, этические правила, регламентирующие работу врача-генетика, рекомендуют сообщать о потенциально патогенных вариантах в некоторых генах, даже если они не связаны с текущим состоянием пациента. К таким генам относятся, например, BRCA1 и BRCA2, связанные с раком молочной железы.

2.10. Ехо-С: суть метода

Основным ограничением NGS-технологий в настоящее время является их цена, напрямую зависящая от глубины секвенирования библиотеки. 3С методы на сегодняшний момент являются наиболее эффективным способом обнаружения хромосомных перестроек, но при небольшой глубине секвенирования обнаружение точечных вариантов затруднено. WGS при соответствующей глубине способно обнаруживать большую часть маленьких инделов и SNP, WES, с другой стороны, позволяет выявить варианты

при небольшой глубине, но только в экземе. Возможности обнаружения хромосомных перестроек для этих двух методов ограничены.

Неплохим компромиссом является метод Ехo-С, сочетающий технологии таргетного обогащения с 3С. С его помощью можно как искать точечные варианты в таргетных регионах, так и перестройки во всём геноме[48].

3. Материалы и методы

Данные секвенирования. Результаты секвенирования клеточной линии K562 были взяты из публичных источников (см. Приложение А).

Контроль качества. Для контроля качества прочтений использовалась утилита FastQC[14], способная оценивать наличие адаптерных последовательностей, распределение прочтений по длине, долю букв по позициям, а также GC-состав. Критерии качества были взяты согласно протоколу разработчика.

Удаление адаптерных последовательностей. Удаление адаптерных последовательностей производилось с помощью утилиты cutadapt[13]. GATK Best Practices рекомендуют использовать в качестве входных данных некартированный ВАМ-файл (uBAM), а для удаления адаптеров использовать собственный инструмент — MarkIlluminaAdapters, так как это позволяет сохранить важные метаданные. Тем не менее, разработчики делают акцент на том, что uBAM должен использоваться как выходной формат на уровне секвенатора, что не является общепринятой практикой[1].

Мы использовали сторонние данные в формате FastQ. Преобразование FastQ файлов в uBAM не позволяет предотвратить потерю метаданных, но значительно увеличивает время обработки данных. Сравнение эффективности cutadapt и MarkIlluminaAdapters не показало каких-либо значимых различий.

Картирование. Картирование производилось с помощью инструментов Bowtie2[15] и BWA[16]. BWA показал лучшие результаты; кроме того, он значительно более эффективно работает с химерными ридями, что немаловажно для используемого нами метода Ехo-С.

Для картирования был взят геном GRCh37/hg19. Из него были удалены так называемые неканоничные хромосомы (некартированные/вариативные референсные последовательности), что позволило улучшить качество выравнивания и значительно упростить работу с готовыми данными.

Кроме того, для правильного функционирования инструментов на дальнейших этапах был разработан скрипт, создающий ReadGroup tag для каждого файла. Конкретных рекомендаций по составлению RG не существует, поэтому мы разработали собственные, основанные на следующих требованиях[1]:

- Поле SM является уникальным для каждого биологического образца и используется при поиске вариантов. Несколько SM в одном файле могут быть использованы при когортном анализе.

- Поле ID является уникальным для каждого RG в BAM-файле. BQSR использует ID как идентификатор самой базовой технической единицы секвенирования.
- Поле PU не является обязательным. Рекомендации GATK советуют помещать в него информацию о чипе секвенирования (баркод), ячейке и баркоде (номере) образца. Во время BQSR, при наличии поле PU является приоритетным по отношению к ID.
- Поле LB является уникальным для каждой библиотеки, приготовленной из биологического образца. Оно отражает различия в количестве ПЦР-дубликатов и поэтому используется инструментом MarkDuplicates.

Объединение BAM-файлов производилось инструментом MergeSamFiles. Сбор статистики по картированию производился с помощью инструмента samtools flagstat.

Удаление ПЦР-дубликатов. Так как мы использовали данные экзомного секвенирования, а количество образцов у нас было относительно небольшим и мы были заинтересованы в максимально качественной подготовке данных, в пайплайн был включён этап удаления ПЦР-дубликатов. Удаление дубликатов производилось инструментом MarkDuplicates от Picard[17], интегрированным в GATK. Оптимальные показатели скорости MarkDuplicates достигаются при запуске Java с параллелизацией сборщиков мусора и количеством сборщиков мусора равным двум[2]. Также, согласно рекомендациям разработчиков, прочтения были предварительно отсортированы по именам, чтобы удалению подверглись не только первичные, но и добавочные выравнивания[1].

Рекалибровка качества прочтений (BQSR). Рекалибровка производилась с помощью инструментов GATK — BaseRecalibrator и ApplyBQSR. Для обучения машинной модели требуются варианты в VCF формате (согласно рекомендациям для Homo sapiens — dbSNP >132).

К сожалению, предоставленная Broad Institute база данных оказалась сильно устаревшей и не вполне подходила для сделанной нами геномной сборки, поэтому было решено подвергнуть обработке dbSNP v150, скачанную с NCBI[47]. База данных потребовала замену и сортировку контигов в соответствии с референсным геномом, а также удаление «пустых» вариантов, содержащих точки в полях REF и ALT. Далее база данных была архивирована с помощью bgzip, а затем проиндексирована IndexFeatureFile от GATK (этот же инструмент одновременно проверяет БД на пригодность для BQSR).

В [2] было показано, что оптимальные показатели скорости BaseRecalibrator достигаются, как и в случае с MarkDuplicates, запуском Java с двумя параллельными сборщиками мусора; кроме того, BaseRecalibrator поддаётся внешнему распараллеливанию путём разделения картированных ридов на хромосомные группы. Хромосомные группы формировались вручную для используемой сборки генома, каждая запускалась с помощью bash-скрипта. Нам удалось усовершенствовать данный этап — запуск BaseRecalibrator производился с помощью библиотек Python3 subprocess, а параллелизация осуществлялась библиотекой multiprocessing, таким образом, можно было делить файл с картированными прочтениями по хромосомам и обрабатывать их отдельно,

так как multiprocessing автоматически распределяет процессы по имеющимся потокам. Также для повышения отказоустойчивости скрипта у BaseRecalibrator и ApplyBQSR была устранена разница в фильтрации прочтений, из-за которой при малых размерах библиотек пайплайн экстренно завершал работу.

Оценка покрытия и обогащения. Покрытие и обогащение в экзOME оценивались с помощью скрипта на основе bedtools[18].

Поиск вариантов. Поиск вариантов производился с помощью GATK HaplotypeCaller. Инструмент запускался с дополнительным параметром `--dont-use-soft-clipped-bases`, который не позволял использовать для поиска вариантов невыравненные химерные части.

Как и в случае с BaseRecalibrator, HaplotypeCaller поддается внешнему распараллеливанию[2]. Мы также осуществили параллелизацию с помощью сочетания subprocess и multiprocessing, достигнув 10-12-кратного ускорения по сравнению с запуском на одном потоке.

Рекалибровка и ранжирование вариантов. В арсенале GATK также присутствуют инструменты для рекалибровки и ранжирования вариантов, с использованием моделей машинного обучения и баз данных с частыми вариантами (CNNScoreVariants и FilterVariantTranches).

Анализ показал, что при наличии этапа рекалибровки вариантов время обработки результатов секвенирования увеличивается почти вдвое. Между тем, рекалибровка и ранжирование с помощью инструментов GATK не исключает необходимость проверки вариантов вручную. Таким образом, от этого этапа решено было отказаться.

Аннотация вариантов. Аннотация вариантов производилась вначале с помощью инструмента Ensembl VEP[12], затем мы мигрировали на ANNOVAR[11].

Используемые базы данных:

1. Human Gene Mutation Database (HGMD®)[23]
2. Online Mendelian Inheritance in Man (OMIM®)[24]
3. GeneCards®: The Human Gene Database — <https://www.genecards.org/>
4. ClinVar — <https://www.ncbi.nlm.nih.gov/clinvar/>
5. dbSNP — <https://www.ncbi.nlm.nih.gov/snp/>
6. Genome Aggregation Database (gnomAD)[22]
7. 1000 Genomes Project — <https://www.internationalgenome.org/>
8. Great Middle East allele frequencies (GME)[21]
9. dbNSFP: Exome Predictions[20]

10. dbSCSNV: Splice site prediction[25]
11. RegSNPintron: intronic SNVs prediction[19]

Фильтрация вариантов. (разрыв) Пограничным значением pLI было взято 0.9, согласно рекомендациям в оригинальной статье[33].

Интерпретация данных и составление отчёта производилось в соответствии с рекомендациями Американского колледжа медицинской генетики и геномики (ACMG) и Ассоциации молекулярной патологии[32].

Из данных Banaszak et al. также были удалены варианты в гене DNMT3A. Исключены все варианты в интервале chr2:25455845-25565459 с фланкированием 1kb.

4. Результаты

5. Обсуждение результатов

6. Предварительные выводы

А. Данные секвенирования K562

Name	Article	Type	Reads, M	Accession Codes
GSM1551618_HIC069	Rao et al.[3]	Hi-C	456.8	SRR1658693
GSM1551619_HIC070	Rao et al.[3]	Hi-C	591.9	SRR1658694
GSM1551620_HIC071	Rao et al.[3]	Hi-C	79.9	SRR1658695 SRR1658696
GSM1551621_HIC072	Rao et al.[3]	Hi-C	79.6	SRR1658697 SRR1658698
GSM1551622_HIC073	Rao et al.[3]	Hi-C	77.4	SRR1658699 SRR1658700
GSM1551623_HIC074	Rao et al.[3]	Hi-C	80.8	SRR1658702 SRR1658701
ENCSR025GPQ	Zhou et al.[4]	WGS	130.0	ENCFF574YLG ENCFF921AXL ENCFF590SSX
ENCSR053AXS	Zhou et al.[4]	WGS	796.2	ENCFF004THU ENCFF066GQD ENCFF313MGL ENCFF506TKC ENCFF080MQF
ENCSR711UNY	Zhou et al.[4]	WGS	449.7	ENCFF471WSA ENCFF826SYZ ENCFF590SSX
SRX3358201	Dixon et al.[5]	WGS	366.3	SRR6251264
GSE148362_G1	Wang et al.[6]	Repli-seq	24.8	SRR11518301
GSE148362_S1	Wang et al.[6]	Repli-seq	30.9	SRR11518302
GSE148362_S2	Wang et al.[6]	Repli-seq	45.4	SRR11518303
GSE148362_S3	Wang et al.[6]	Repli-seq	49.8	SRR11518304
GSE148362_S4	Wang et al.[6]	Repli-seq	44.1	SRR11518305
GSE148362_S5	Wang et al.[6]	Repli-seq	38.4	SRR11518306
GSE148362_S6	Wang et al.[6]	Repli-seq	35.2	SRR11518307
GSE148362_G2	Wang et al.[6]	Repli-seq	33.0	SRR11518308

Name	Article	Type	Reads, M	Accession Codes
INSITU_HS1	Ray et al.[7]	Hi-C	86.3	SRR9019504
INSITU_HS2	Ray et al.[7]	Hi-C	127.1	SRR9019505
INSITU_NHS1	Ray et al.[7]	Hi-C	86.4	SRR9019506
INSITU_NHS2	Ray et al.[7]	Hi-C	128.5	SRR9019507
PD_STABLE_REP1	Moquin et al.[8]	Hi-C	67.2	SRR5470535 SRR5470534
PD_STABLE_REP2	Moquin et al.[8]	Hi-C	52.9	SRR5470536 SRR5470537
PD_TRANSIENT	Moquin et al.[8]	Hi-C	81.3	SRR5470539 SRR5470538
PDDE_TRANSIENT	Moquin et al.[8]	Hi-C	55.2	SRR5470541 SRR5470540
GSM2588815_R1	Belaghzal et al.[9]	Hi-C	72.9	SRR5479813
GSM2536769_WT	Banaszak et al.[10]	WES ¹	39.2	SRR5345331
GSM2536770_WT_TF	Banaszak et al.[10]	WES ¹	49.4	SRR5345332
GSM2536771_MT2	Banaszak et al.[10]	WES ¹	42.0	SRR5345333
GSM2536772_MT3	Banaszak et al.[10]	WES ¹	43.7	SRR5345334
GSM2536773_MT4	Banaszak et al.[10]	WES ¹	39.9	SRR5345335
GSM2536774_MT5	Banaszak et al.[10]	WES ¹	40.8	SRR5345336

Список литературы

- [1] Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013;43(1110):11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43
- [2] Heldenbrand JR, Baheti S, Bockol MA, et al. Recommendations for performance optimizations when using GATK3.8 and GATK4 [published correction appears in BMC Bioinformatics. 2019 Dec 17;20(1):722]. BMC Bioinformatics. 2019;20(1):557. Published 2019 Nov 8. doi:10.1186/s12859-019-3169-7
- [3] Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping [published correction

¹Варианты в гене DNMT3A были исключены из выборки.

- appears in Cell. 2015 Jul 30;162(3):687-8]. Cell. 2014;159(7):1665-1680. doi:10.1016/j.cell.2014.11.021
- [4] Zhou B, Ho SS, Greer SU, et al. Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* 2019;29(3):472-484. doi:10.1101/gr.234948.118
- [5] Dixon JR, Xu J, Dileep V, et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet.* 2018;50(10):1388-1398. doi:10.1038/s41588-018-0195-8
- [6] Yuchuan Wang, Yang Zhang, et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. *bioRxiv* 2020.03.09.982967; doi: <https://doi.org/10.1101/2020.03.09.982967>
- [7] Ray J, Munn PR, Vihervaara A, et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proc Natl Acad Sci U S A.* 2019;116(39):19431-19439. doi:10.1073/pnas.1901244116
- [8] Moquin SA, Thomas S, Whalen S, et al. The Epstein-Barr Virus Episome Maneuvers between Nuclear Chromatin Compartments during Reactivation. *J Virol.* 2018;92(3):e01413-17. Published 2018 Jan 17. doi:10.1128/JVI.01413-17
- [9] Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods.* 2017;123:56-65. doi:10.1016/j.ymeth.2017.04.004
- [10] Banaszak LG, Giudice V, Zhao X, et al. Abnormal RNA splicing and genomic instability after induction of DNMT3A mutations by CRISPR/Cas9 gene editing. *Blood Cells Mol Dis.* 2018;69:10-22. doi:10.1016/j.bcmed.2017.12.002
- [11] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164. doi:10.1093/nar/gkq603
- [12] McLaren, W., Gil, L., Hunt, S.E. et al. The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122 (2016). doi: 10.1186/s13059-016-0974-4
- [13] MARTIN, Marcel. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, [S.l.], v. 17, n. 1, p. pp. 10-12, may 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200.
- [14] Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- [15] Langmead, B., Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357–359 (2012). doi: 10.1038/nmeth.1923

- [16] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760. doi:10.1093/bioinformatics/btp324
- [17] "Picard Toolkit." 2019. Broad Institute, GitHub Repository. <http://broadinstitute.github.io/picard/>; Broad Institute
- [18] Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26, 6, pp. 841–842.
- [19] Lin, H., Hargreaves, K.A., Li, R. et al. RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol* 20, 254 (2019). doi:10.1186/s13059-019-1847-4
- [20] 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. *Hum Mutat*. 2016;37(3):235-241. doi:10.1002/humu.22932
- [21] Scott EM, Halees A, Itan Y, et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet*. 2016;48(9):1071-1076. doi:10.1038/ng.3592
- [22] Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020). doi:10.1038/s41586-020-2308-7
- [23] Stenson PD, Mort M, Ball EV, 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;136(6):665-677. doi:10.1007/s00439-017-1779-6
- [24] Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*. 2015;43(Database issue):D789-D798. doi:10.1093/nar/gku1205
- [25] Jian X, Boerwinkle E, Liu X. In silico tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med*. 2014;16(7):497-503. doi:10.1038/gim.2013.176
- [26] Naslavsky MS, Yamamoto GL, de Almeida TF, Ezquina SAM, Sunaga DY, Pho N, Bozoklian D, Sandberg TOM, Brito LA, Lazar M, Bernardo DV, Amaro E Jr, Duarte YAO, Lebrão ML, Passos-Bueno MR, Zatz M. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat*. 2017 Jul;38(7):751-763. doi: 10.1002/humu.23220.
- [27] Yohe S, Thyagarajan B. Review of Clinical Next-Generation Sequencing. *Arch Pathol Lab Med*. 2017 Nov;141(11):1544-1557. doi: 10.5858/arpa.2016-0501-RA

- [28] Balloux F, Brønstad Brynildsrud O, van Dorp L, et al. From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic. *Trends Microbiol.* 2018;26(12):1035-1048. doi:10.1016/j.tim.2018.08.004
- [29] Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.* 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463
- [30] Burrows M, Wheeler DJ. Technical report 124. Palo Alto, CA: Digital Equipment Corporation; 1994. A block-sorting lossless data compression algorithm.
- [31] Ebbert MT, Wadsworth ME, Staley LA, et al. Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics.* 2016;17 Suppl 7(Suppl 7):239. Published 2016 Jul 25. doi:10.1186/s12859-016-1097-3
- [32] Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17(5):405-424. doi:10.1038/gim.2015.30
- [33] Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291. doi:10.1038/nature19057
- [34] Ziegler A, Colin E, Goudenège D, Bonneau D. A snapshot of some pLI score pitfalls. *Hum Mutat.* 2019 Jul;40(7):839-841. doi: 10.1002/humu.23763
- [35] Saleheen D, Natarajan P, Armean IM, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature.* 2017;544(7649):235-239. doi:10.1038/nature22034
- [36] Wutz G, Várnai C, Nagasaka K, et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* 2017;36(24):3573-3599. doi:10.15252/embj.201798004
- [37] Yu D, Zhang K, Han M, 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;7(6):e674. doi:10.1002/mgg3.674
- [38] Herder M. What Is the Purpose of the Orphan Drug Act?. *PLoS Med.* 2017;14(1):e1002191. Published 2017 Jan 3. doi:10.1371/journal.pmed.1002191
- [39] Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, Hughes DA; International Society for Pharmacoeconomics and Outcomes Research Rare Disease Special Interest Group. Rare Disease Terminology and Definitions-A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. *Value Health.* 2015 Sep;18(6):906-14. doi: 10.1016/j.jval.2015.05.008

- [40] The Lancet Neurology. Rare neurological diseases: a united approach is needed. *Lancet Neurol.* 2011 Feb;10(2):109. doi: 10.1016/S1474-4422(11)70001-1. Erratum in: *Lancet Neurol.* 2011 Mar;10(3):205.
- [41] D. Huber, L. Voith von Voithenberg, G.V. Kaigala. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering*, Volume 1, 2018, Pages 15-24, ISSN 2590-0072. doi: 10.1016/j.mne.2018.10.006.
- [42] Theisen, A. (2008) Microarray-based Comparative Genomic Hybridization (aCGH). *Nature Education* 1(1):45
- [43] 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. *Int J Mol Sci.* 2012;13(3):3245-3276. doi:10.3390/ijms13033245
- [44] Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. *Science.* 2002 Feb 15;295(5558):1306-11. doi: 10.1126/science.1067799. PMID: 11847345.
- [45] Oluwadare, O., Highsmith, M. & Cheng, J. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biol Proced Online* 21, 7 (2019). doi: 10.1186/s12575-019-0094-0
- [46] Alejandro j. Brea-Fernandez, Marta Ferro, Ceres Fernandez-Rozadilla, Ana Blanco, Laura Fachal, Marta Santamarina, Ana Vega, Alejandro Pazos, Angel Carracedo and Clara Ruiz-Ponte, An Update of In Silico Tools for the Prediction of Pathogenesis in Missense Variants, *Current Bioinformatics* (2011) 6: 185. <https://doi.org/10.2174/1574893611106020185>
- [47] <https://www.ncbi.nlm.nih.gov/snp/>
- [48] Mozheiko, E.A., Fishman, V.S. Detection of Point Mutations and Chromosomal Translocations Based on Massive Parallel Sequencing of Enriched 3C Libraries. *Russ J Genet* 55, 1273–1281 (2019). <https://doi.org/10.1134/S1022795419100089>
- [49] B. Sampson, A. McGuire. *Genetics and the Molecular Autopsy*. Editor(s): Linda M. McManus, Richard N. Mitchell. *Pathobiology of Human Disease*, Academic Press, 2014. Pages 3459–3467. ISBN 9780123864574. doi: 10.1016/B978-0-12-386456-7.06707-1.
- [50] Gresham, D., Dunham, M. & Botstein, D. Comparing whole genomes using DNA microarrays. *Nat Rev Genet* 9, 291–302 (2008). <https://doi.org/10.1038/nrg2335>
- [51] Ryzhkova O.P., Kardymon O.L., Prohorchuk E.B., Konovalov F.A., Maslennikov A.B., Stepanov V.A., Afanasyev A.A., Zaklyazminskaya E.V., Kostareva A.A., Pavlov A.E., Golubenko M.V., Polyakov A.V., Kutsev S.I. Guidelines for the interpretation of massive parallel sequencing variants. *Medical Genetics*. 2017;16(7):4-17. (In Russ.)

- [52] Galanello R, Origa R. Beta-thalassemia. *Orphanet J Rare Dis*. 2010;5:11. Published 2010 May 21. doi:10.1186/1750-1172-5-11
- [53] Bomar, J., Benke, P., Slattery, E. 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* 35, 264–269 (2003). <https://doi.org/10.1038/ng1255>
- [54] Jin Z, Liu Y. DNA methylation in human diseases. *Genes Dis*. 2018;5(1):1-8. Published 2018 Jan 31. doi:10.1016/j.gendis.2018.01.002
- [55] Tang, B., Dean, B. & Thomas, E. Disease- and age-related changes in histone acetylation at gene promoters in psychiatric disorders. *Transl Psychiatry* 1, e64 (2011). <https://doi.org/10.1038/tp.2011.61>
- [56] Spielmann, M., Lupiáñez, D.G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* 19, 453–467 (2018). <https://doi.org/10.1038/s41576-018-0007-0>
- [57] Winklhofer KF, Tatzelt J, Haass C. The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. *EMBO J*. 2008;27(2):336-349. doi:10.1038/sj.emboj.7601930
- [58] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, Murthy KK, Rovin BH, Bradley W, Clark RA, Anderson SA, O'connell RJ, Agan BK, Ahuja SS, Bologna R, Sen L, Dolan MJ, Ahuja SK. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*. 2005 Mar 4;307(5714):1434-40. doi: 10.1126/science.1101160.
- [59] Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. Strong association of de novo copy number mutations with autism. *Science*. 2007 Apr 20;316(5823):445-9. doi: 10.1126/science.1138659.
- [60] Acuna-Hidalgo, R., Veltman, J.A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol* 17, 241 (2016). <https://doi.org/10.1186/s13059-016-1110-1>
- [61] Schreck RR, Distèche CM. Chromosome banding techniques. *Curr Protoc Hum Genet*. 2001 May;Chapter 4:Unit4.2. doi: 10.1002/0471142905.hg0402s00. PMID: 18428280.
- [62] Guo B, Han X, Wu Z, Da W, Zhu H. Spectral karyotyping: an unique technique for the detection of complex genomic rearrangements in leukemia. *Transl Pediatr*. 2014;3(2):135-139. doi:10.3978/j.issn.2224-4336.2014.01.02