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 3С-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 **5**
1.1 Актуальность . . . 5
1.2 Цель . . . 5
1.3 Задачи . . . 5

2 Обзор литературы **6**
2.1 Механизмы развития генетических патологий . . . 7
2.2 Типы генетических аномалий, лежащих в основе генетических патологий 8
2.3 Функциональные классы генетических вариантов . . . 10
2.4 Методы детекции генетических вариантов . . . 11
2.5 Виды NGS . . . 15
2.6 Базовая схема обработки результатов высокопроизводительного секвенирования для поиска и клинической интерпретации однонуклеотидных полиморфизмов . . . 17
2.7 Аннотация, фильтрация и интерпретация результатов . . . 21
2.8 Exo-C: суть метода . . . 24

3 Материалы и методы **25**

4 Результаты **29**
4.1 Результаты секвенирования Exo-C-библиотек . . . 30
4.2 Автоматизация обработки данных секвенирования . . . 30
4.3 Сравнение данных секвенирования клеточной линии K562 . . . 31

5 Обсуждение результатов **33**
5.1 Контрольные образцы . . . 33
5.2 Оценка результатов секвенирования Exo-C-библиотек . . . 33

6 Предварительные выводы **34**

7 План работы **35**

A Данные секвенирования клеточной линии K562 **36**

Список сокращений

3C (*англ.* Chromosome Conformation Capture) — захват конформации хромосом

BAM (*англ.* Binary sequence Alignment/Map) — бинарный файловый формат, предназначенный для хранения информации о картированных прочтениях

BQSR (*англ.* Base Quality Score Recalibration) — рекалибровка качества прочтений

cffDNA (*англ.* Cell-Free Fetal DNA) — свободная ДНК плода

CGH (*англ.* Comparative Genomic Hybridization) — сравнительная геномная гибридизация

CNV (*англ.* Copy Number Variation) — вариация числа копий

Exo-C — метод приготовления NGS-библиотек, сочетающий таргетное обогащение экзома и технологии захвата конформации хромосом

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

GATK (*англ.* Genome Analysis ToolKit) — набор инструментов для биоинформационного анализа, созданный Broad Institute

Hi-C — метод захвата конформации хромосом «все против всех»

LoF (*англ.* Loss of Function) — потеря функции гена

MAPQ (*англ.* MAPping Quality) — качество картирования

MIP (*англ.* Molecularly Imprinted Polymers) — молекулярно импринтированные полимеры

MLPA (*англ.* Multiplex Ligation-dependent Probe Amplification) — мультиплексная лигаза-зависимая амплификация зонда

NGS (*англ.* New Generation Sequencing) — секвенирование нового поколения

NIPT (*англ.* Non-Invasive Prenatal Testing) — неинвазивное пренатальное тестирование

NOR (*англ.* Nucleolus Organizer Region) — ядрышковый организатор

PEC (*англ.* Primer Extension Capture) — захват с помощью расширения праймера

RG (*англ.* Read Group) — группа прочтения

SKY (*англ.* Spectral Karyotyping) — спектральное кариотипирование

SMART — анализ транскриптома одной клетки

SNV (*англ.* Single Nucleotide Variant) — однонуклеотидный генетический вариант

UTR (*англ.* UnTranslated Regions) — нетранслируемая область

VCF (*англ.* Variant Call Format) — формат записи генетических вариантов, найденных в результатах секвенирования

WES (*англ.* Whole Exome Sequencing) — полноэкзомное секвенирование

WGS (*англ.* Whole Genome Sequencing) — полногеномное секвенирование

БД — база данных

ВИЧ — вирус иммунодефицита человека

ДНК — дезоксирибонуклеиновая кислота

мРНК — матричная РНК

п.о. — пары оснований

ПЦР — полимеразная цепная реакция

РНК — рибонуклеиновая кислота

ТАД — топологически ассоциированные домены

ХМА — хромосомный микроматричный анализ

1. Введение

1.1. Актуальность

Наследственные заболевания являются одной из основных причин младенческой и детской смертности в развитых странах. Взрослые люди с такими патологиями требуют огромных затрат средств на медикаменты, оперативные вмешательства, специальный уход и социальные льготы. Таким образом, доступные и точные методы диагностики наследственных заболеваний могут помочь в сокращении заболеваемости и смертности, а также повысить экономическое благополучие населения.

Несмотря на то, что в развитии наследственных заболеваний играют роль множество механизмов, в основе их всегда лежат изменения тех или иных участков ДНК. Эти генетические варианты существенно различаются по размеру, характеру изменения, а также функциональному значению. Существует множество методов выявления генетических вариантов, каждый метод имеет свои преимущества и границы применения.

Наиболее перспективными в диагностическом и исследовательском плане в настоящее время являются методы секвенирования — например, полногеномное и полноэкзомное секвенирование. В Секторе геномных механизмов онтогенеза ИЦиГ СО РАН был разработан новейший метод секвенирования — Exo-C, сочетающий технологии экзомного обогащения с захватом конформации хромосом. Потенциальным преимуществом данного метода может быть возможность поиска как крупных перестроек, так и точечных полиморфизмов в экзоме при относительно небольшой глубине секвенирования, от которой напрямую зависит цена секвенирования. Широкий спектр применения метода и доступность в финансовом аспекте делают метод Exo-C привлекательным как для медико-биологических научных исследований, так и для внедрения в клиническую практику.

1.2. Цель

Целью нашей работы является сравнение эффективности методов Exo-C, полногеномного секвенирования и экзомного секвенирования для поиска точечных полиморфизмов в геномах клеток человека.

1.3. Задачи

Основные задачи, которые необходимо решить для достижения поставленной нами цели:

1. Разработать биоинформационный протокол анализа данных секвенирования Exo-C-библиотек.

2. Проанализировать доступные данные полногеномного, полноэкзомного, Hi-C и Exo-C-секвенирования для иммортализованной клеточной линии человека K562.

3. Сравнить точечные генетические варианты в геноме клеток K562, детектируемые при использовании полногеномного и экзомного секвенирования, с таковыми, найденными методом Exo-C.

2. Обзор литературы

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.)[**?**]. 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[**?**]. 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[**?**]).

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% 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[**?**]. 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)[**?**]. 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[**?**].

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

Кайманових 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[**?**]. 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[**?**]. 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[**?**]. 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®)[**?**];
3. Orphanet[**?**].

Около 80% редких болезней имеют генетическую природу и начинаются в раннем детстве[**?**]. 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 является известное наследственное

нейродегенеративное заболевание — аутосомно-доминантный вариант болезни Альцгеймера. Другое нейродегенеративное заболевание — аутосомно-рецессивная болезнь Паркинсона — может служить примером потери белком протективной функции[**?**].

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[**?**]. Изменение ацетилирования гистонов некоторых генов в клетках головного мозга связано с развитием такого заболевания, как шизофрения[**?**].

Кроме того, на экспрессию генов 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 (ТАДов)[**?**]. В результате разрушения старых или образования новых границ ТАДов формируются структурные варианты, характеризующиеся иными промотор-энхансерными взаимодействиями. Подобные изменения лежат в основе таких состояний, как FtM-инверсия пола (ген SOX9) и синдром Кукса (ген KCNJ2)[**?**].

Несмотря на то, что в развитии наследственных заболеваний эпигенетика безусловно играет важную роль, в основе их всегда лежат изменения тех или иных участков ДНК. Эти генетические варианты существенно различаются по размеру, характеру изменения, а также функциональному значению, которое напрямую зависит от затрагиваемых вариантом районов генома.

2.2. Типы генетических аномалий, лежащих в основе генетических патологий

Генетические аномалии различаются по размеру. Размер непосредственно влияет на способность исследователя обнаружить эту аномалию. Самыми крупными являются хромосомные перестройки. Они делятся на две основных группы — сбалансированные (без изменения количества генетической информации) и несбалансированные (с изменением количества генетической информации).

Несбалансированные перестройки в большинстве своём приводят к летальному исходу (в эмбриональном или детском периодах) и грубым изменениям фенотипа. К несбалансированным относятся:

- Анэуплоидии — изменение числа хромосом. Примерами анэуплоидий могут слу-

жить синдром Дауна (трисомия 21 хромосомы), Эдвардса (трисомия 18 хромосомы), Патау (трисомия 13 хромосомы), а также вариации числа половых хромосом (синдромы Тёрнера, Клайнфельтера и другие). Частичная моносомия — синдром кошачьего крика (связан с утратой плеча 5 хромосомы). Прочие анэуплоидии ведут к несовместимым с жизнью нарушениям эмбрионального развития и, как следствие, спонтанным абортам.

- Несбалансированные транслокации — перемещение фрагмента хромосомы с одного места на другое с изменением количества генетической информации. Несбалансированные транслокации могут приводить к значимым изменениям фенотипа (например, инверсия пола[**?**]) и служить онкогенами[**?**].

- Вариации числа копий (*англ.* Copy Number Variations, CNV) — дупликации (мультипликации) и делеции хромосомных сегментов размером от тысячи до нескольких миллионов пар оснований. Могут возникнуть из несбалансированных транслокаций, амплификаций и собственно делеций. 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[**?**], а крупные CNV в разных частях генома ассоциированы с расстройствами аутического спектра[**?**].

Сбалансированные перестройки чаще всего характеризуются более мягкими фенотипическими проявлениями, а иногда и их отсутствием. К сбалансированным перестройкам относятся:

- Инверсии — переворот фрагмента хромосомы. Крупные инверсии могут быть причиной изменения границы ТАД, а также запирания кроссинговера и образования гаплогрупп.

- Сбалансированные транслокации — перемещение фрагмента хромосомы с одного места на другое без изменения количества генетической информации. В свою очередь они делятся на реципрокные (взаимный обмен участками между негомологичными хромосомами) и Робертсоновские (слияние акроцентрических хромосом с образованием метацентрической или субметацентрической). Сбалансированные транслокации могут как не проявляться в фенотипе (сказываясь только на фертильности[**?**]), так и приводить к серьёзным последствиям — например, синдрому Дауна (робертсоновская транслокация является причиной синдрома Дауна в 2–4% случаев[**?**]).

Самыми небольшими — но не менее важными — являются точечные полиморфизмы (*англ.* Single Nucleotide Variants, SNV) и короткие инсерции и делеции (indels) размером 20–50bp. Чаще всего эти генетические варианты нейтральные и не имеют фенотипических проявлений, но некоторые могут приводить как к генетическим, так и к эпигенетическим изменениям. Также варианты делятся на наследуемые, которые передаются от родителей к детям, и варианты *de novo*. Согласно оценкам, предоставленным

[**?**], в среднем в каждом поколении у человека возникают 44–82 SNV *de novo*, из них 1–2 приходятся 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[**?**].

- Интроны, которые содержат регуляторные области и сплайс-сайты, необходимые для процессинга транскрипта в готовую мРНК, а также 3’-нетранслируемая область (*англ.* 3’-untranslated region, 3’UTR) и 5’-нетранслируемая область (*англ.* 5’-untranslated region, 5’UTR), вовлечённые в регуляцию транскрипции, трансляции и деградации транскрипта. В частности, в 5’UTR находится так называемая консенсусная последовательность Козак, важная для инициации трансляции мРНК[**?**]. Также известно, что в 5’UTR могут находиться открытые рамки считывания, которые влияют на поведение рибосомы — могут вызывать её торможение, диссоциацию, либо перекрывать основной старт-кодон гена[**?**]. Генетические варианты могут как разрушать канонические сплайс-сайты, так и способствовать образованию новых внутри интронных участков[**?**]. 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 детекции генетических 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[**?**]:

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 миллионов п.о.) и процент поражённых клеток в образце (минимум 5–10%)[**?**].

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 (*англ.* Fluorescence In Situ Hybridization, FISH). Основой 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[**?**]. 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 тыс. п.о. 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[**?**].

Сравнительная геномная гибридизация (*англ.* Comparative Genomic Hybridization, 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 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[**?**]. 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[**?**].

Мультиплексная лигаза-зависимая амплификация зонда (*англ.* Multiplex Ligation-dependent Probe Amplification, 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, 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[**?**].

Как мы видим, перечисленные методы имеют один серьёзный недостаток — они мо-

гут 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 тысячи п.о.) фрагменты ДНК[**?**]. 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 (*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[**?**]. 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.

2.5. Виды NGS

Полногеномное секвенирование (*англ.* Whole Genome Sequencing, WGS). Приготовление библиотек при полногеномном секвенировании 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[**?**].

Таргетные панели. Основой данных методов является обогащение целевых регионов генома. Методов обогащения существует достаточно много, но все они делятся на 4 основные категории[**?**]:

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 (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[**?**].

Полноэкзомное секвенирование (*англ.* Whole Exome Sequencing, WES). 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 WES могут быть использованы лишь немногие из них, а именно — твердофазная и жидкофазная гибридизация[**?**].

У человека экзом составляет примерно 1% от генома, или примерно 30 миллионов п.о. (суммарно). При этом более 80% генетических вариантов, которые представлены в базе данных известных геномных вариантов CLINVAR[**?**], 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[**?**]. 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[**?**].

Технологии захвата конформации хромосом (3C). Данные методики 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[**?**]. 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[**?**].

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

2.6. Базовая схема обработки результатов высокопроизводительного секвенирования для поиска и клинической интерпретации однонуклеотидных полиморфизмов

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.

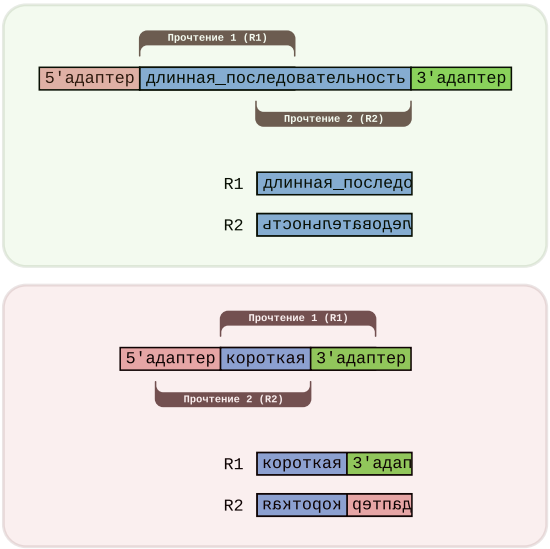

Рис. 1: Фрагменты адаптерных последовательностей в данных секвенирования

Удаление адаптерных последовательностей. Если целевой фрагмент ДНК короче длины прочтения, то фрагменты адаптерной последовательности могут попасть в готовые данные (рис. 1). 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[**?**].

Картирование прочтений. Как уже упоминалось выше, результаты NGS — это прочтения, содержащие небольшие (в пределах 200 п.о.) фрагменты генома. Извлечение информации из необработанных результатов секвенирования затруднительно, так как эти фрагменты 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 за-

писывается 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[**?**].

Обычно 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.

Исходя из особенностей алгоритмов картирования, выравнивания делятся на следующие классы:

- Первичное выравнивание (*англ.* primary) — выравнивание наиболее крупного (и содержащего наименьшее количество перекрытий, в случае химерного прочтения) фрагмента прочтения с наиболее высоким MAPQ. Первичное выравнивание только одно. Первичное выравнивание химерного прочтения называется репрезентативным;

- Вторичное выравнивание (*англ.* secondary) — выравнивание наиболее крупного фрагмента прочтения с меньшим MAPQ. Вторичных выравниваний может быть несколько (в зависимости от выставленного нижнего порога MAPQ);

- Добавочное выравнивание (*англ.* supplementary) — выравнивание менее крупных (либо содержащих большее количество перекрытий) фрагментов прочтения. Добавочные выравнивания характерны только для химерных прочтений.

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. 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. Клипирование бывает двух типов:

- Мягкое клипирование (*англ.* soft-clip) — отсечение невыравненного конца прочтения с сохранением полной последовательности прочтения. В отсечённых методом мягкого клипирования регионах могут быть адаптерные последовательности, а также часть химерного прочтения (в репрезентативном выравнивании).

- Жёсткое клипирование (*англ.* hard-clip) — отсечение невыравненного конца прочтения без сохранения его последовательности. В регионах, подвергшихся жёсткому клипированию, обычно находятся репрезентативные участки химерных прочтений (в добавочных выравниваниях).

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. Пример показан на Рис. 2.

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, дубликаты должны быть удалены или помечены для улучшения поиска генетических вариантов[**?**].

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[**?**]. Учитывая, что удаление дубликатов может занимать значительную часть потраченного на обработку данных времени, следует взвесить пользу и затраты данного этапа для конкретной прикладной задачи.

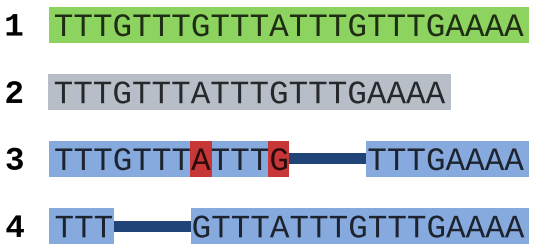

Рис. 2: Неоптимальное картирование прочтения, содержащего делецию. (1) — референсная последовательность, (2) — последовательность прочтения, (3) — картирование, произведённое алгоритмом, включающее две SNV и одну делецию, (4) — оптимальное местоположение делеции

Рекалибровка качества прочтений (*англ.* Base Quality Score Recalibration, BQSR). 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, 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[**?**]. 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 изменчи-

вость повышает вероятность правильного прочтения нуклеотида, не совпадающего с референсным в данной позиции генома.

Broad Institute of MIT and Harvard рекомендует BQSR к использованию для любых данных секвенирования[**?**].

Поиск 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 (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.

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)[**?**], 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.

2.7. Аннотация, фильтрация и интерпретация 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[**?**], поэтому, во избежание ошибок, рекомендуется использовать их все. Это особенно важно при дифферециации генетических вариантов с высокой вероятностью повреждающего эффекта (сдвиги рамок считыва-

ния, нонсенс-кодоны). Кроме того, различаются алгоритмы определения функционального класса в различных утилитах аннотации, что также создаёт определённые трудности[**?**].

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[**?**] и OrphaData[**?**].

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[**?**]:

   - 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[**?**], 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[**?**], в которой отражены частоты по популяции Ближнего Востока, ABraOM[**?**], предоставляющая частоты генетических вариантов среди практически здорового пожилого населения Бразилии. Также для анализа берутся популяции, в которых велика доля близкородственных связей, например, пакистанская[**?**].

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[**?**] и CLINVAR[**?**]. Тем не менее, CLINVAR считается лишь дополнительным источником, так как часто содержит информацию низкого качества[**?**].

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[**?**]. Предсказательная способность отдельных инструментов оставляет желать лучшего, поэтому чаще всего в клинической практике используются агрегаторы, собирающие предсказания с большого числа известных *in silico* инструментов.

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[**?**].

Когортный и семейный анализ. В случае, если исследователь имеет доступ к группе, представители которой связаны узами крови с пациентом, есть возможность провести семейный анализ. Семейный анализ нужен для установления путей наследования тех или иных генетических вариантов в родословной. Это позволяет уточнить их связь с фенотипом. Также анализ нескольких родственных образцов помогает определить зиготность варианта, обнаружить генетические варианты *de novo*, либо импутировать район с недостаточным покрытием.

Если же в распоряжении исследователя находится группа, связанная одной патологией или вариантом фенотипа, можно провести когортный анализ. Когортный анализ позволяет, например, оценить частоты генетических вариантов в исследуемой и контрольной группе. Кроме того, когортный анализ образцов в конкретной лаборатории помогает детектировать систематические отклонения покрытия и артефакты выравнивания, связанные с конкретными районами генома и/или особенностями приготовления библиотек.

2.8. Exo-C: суть метода

Как уже упоминалось выше, одним из основных ограничений NGS-технологий в настоящее время является их цена, напрямую зависящая от глубины секвенирования библиотеки. Есть ограничения и по возможностям поиска тех или иных генетических вариантов. 3C-методы на сегодняшний момент являются наиболее перспективным способом обнаружения хромосомных перестроек[**?**], но при небольшой глубине секвенирования в них обнаружение точечных полиморфизмов затруднительно[**?**]. WGS способно обнаруживать большую часть SNV, небольших инсерций и делеций, но требует большую глубину секвенирования[**?**]; WES, с другой стороны, позволяет выявить генетические варианты при небольшой глубине секвенирования, но только в экзоме. Возможности обнаружения хромосомных перестроек для последних двух методов ограничены.

Компромиссом между ценой и возможностями поиска генетических вариантов может служить новейший метод Exo-C, сочетающий технологии таргетного обогащения с 3C. Суть его заключается в приготовлении Hi-C-библиотеки и последующем обогащении только тех последовательностей, которые связаны с экзомом. Таким образом, с его помощью можно как искать точечные варианты в обогащённых регионах (за счёт большой глубины покрытия в них), так и хромосомные перестройки во всём геноме (за счёт Hi-C, дающей относительно небольшое, но доступное для анализа покрытие всего генома)[**?**].

Тем не менее, как выяснилось, уже существующие биоинформационные методы следует модифицировать для корректной обработки данных Exo-C. Это связано в первую очередь с особенностями протокола Hi-C, к примеру, наличием технических последовательностей (бридж-адаптеров), которые приводят к появлению ложных SNV в экзомных регионах. Таргетное обогащение, со своей стороны, вносит определённые помехи в Hi-C-данные, так как изменяется представленность регионов генома в библиотеке, а значит, и пропорции контактов между регионами.

Данная работа посвящена разработке биоинформационных методов для поиска точковых генетических вариантов в Exo-C-данных и последующего сравнения Exo-C с методами полногеномного и полноэкзомного секвенирования.

3. Материалы и методы

Данные секвенирования. Поиск данных секвенирования производился в базах данных NCBI (GEO DataSets, SRA, PubMed) и ENCODE с использованием ключевых слов “K562”, “K562+WGS”, “K562+WES”, “K562+Hi-C”.

Контроль качества NGS-данных. Для контроля качества прочтений мы использовали утилиту FastQC[**?**], способную оценивать наличие адаптерных последовательностей, распределение прочтений по длине, GC-состав прочтений, а также производить анализ зависимости нуклеотидного состава от позиции в прочтении. Критерии качества были использованы согласно протоколу разработчика[**?**].

Удаление адаптерных последовательностей. Удаление адаптерных последовательностей производилось с помощью утилиты cutadapt[**?**]. В [**?**] рекомендуется использовать в качестве входных данных некартированный BAM-файл (*англ.* Unmapped Binary sequence Alignment/Map, uBAM), а для удаления адаптеров использовать их собственный инструмент — MarkIlluminaAdapters, так как это позволяет сохранить важные метаданные. Тем не менее, был сделан акцент на том, что uBAM должен использоваться как выходной формат на уровне секвенатора, что не является общепринятой практикой.

Мы использовали данные секвенирования в формате FastQ. Пребразование FastQ-файлов в uBAM не предотвращает потерю метаданных, но значительно увеличивает время обработки данных. Сравнение эффективности cutadapt и MarkIlluminaAdapters в процессе удаления адаптеров не показало каких-либо значимых различий.

Картирование. Картирование производилось с помощью инструментов Bowtie2[**?**] и BWA[**?**]. BWA показал лучшие результаты; кроме того, он значительно более эффективно работает с химерными ридами, что немаловажно для используемого нами метода Exo-C.

Для картирования был взят геном GRCh37/hg19, предоставленный NCBI. 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 метку группы прочтений (*англ.* Read Group tag, RG) для каждого файла. Конкретных рекомендаций по составлению RG не существует, поэтому мы разработали собственные, основанные на следующих требованиях[**?**]:

- 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 чипа), ячейке и баркоде (номере) образца. Во время BQSR 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 потому используется инструментом MarkDuplicates.

Объединение BAM-файлов производилось инструментом MergeSamFiles. Сбор статистики по картированию мы осуществляли с помощью инструмента samtools flagstat[**?**].

Удаление ПЦР-дубликатов. Для улучшения данных экзомного секвенирования в пайплайн был включён этап удаления ПЦР-дубликатов. Обычно этот процесс занимает много времени, но количество образцов у нас было относительно небольшим, и мы были заинтересованы в максимально качественной подготовке данных.

Удаление дубликатов производилось инструментом MarkDuplicates от Picard[**?**], интегрированным в GATK. Оптимальные показатели скорости MarkDuplicates достигаются при запуске Java с параллелизацией сборщиков мусора и количеством сборщиков мусора равным двум[**?**]. Также, согласно рекомендациям разработчиков, прочтения были предварительно отсортированы по именам, чтобы удалению подверглись не только первичные, но и добавочные выравнивания[**?**].

Рекалибровка качества прочтений (BQSR). Рекалибровка производилась с помощью инструментов GATK — BaseRecalibrator и ApplyBQSR. Для обучения машинной модели требуются генетические варианты в VCF формате (согласно рекомендациям для *Homo sapiens* — dbSNP v132+).

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,предоставленную NCBI[**?**]. 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 [**?**] 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 прочтений 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 библиотеки 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 экзоме оценивались с помощью скрипта на основе bedtools[**?**].

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

Как и в случае с BaseRecalibrator, HaplotypeCaller поддаётся внешнему распараллеливанию[**?**]. 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 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 исключают необходимость фильтрации генетических вариантов. Таким образом, от этого этапа решено было отказаться.

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

Используемые базы данных:

1. Human Gene Mutation Database (HGMD®)[**?**]
2. Online Mendelian Inheritance in Man (OMIM®)[**?**]

3. GeneCards®: The Human Gene Database[**?**]
4. CLINVAR[**?**]
5. dbSNP[**?**]
6. Genome Aggregation Database (gnomAD)[**?**]
7. 1000 Genomes Project[**?**]
8. Great Middle East allele frequencies (GME)[**?**]
9. dbNSFP: Exome Predictions[**?**]
10. dbscSNV: Splice site prediction[**?**]
11. RegSNPIntron: intronic SNVs prediction[**?**]

Фильтрация генетических вариантов. Аннотации были агрегированы для удобства использования. Так, агрегации подверглись:

- Имена генов по разным БД — для облегчения поиска;
- Описания функциональных классов из разных БД — для устранения несоответствий между ними;
- Ранги инструментов, предсказывающих патогенность генетического варианта. Трёхранговые системы (патогенный, вероятно патогенный и безвредный) были сведены к двухранговой (патогенный и безвредный). Отдельно были агрегированы предсказательные инструменты для экзонов, инструменты для интронов и сплайс-вариантов также учитывались отдельно;
- Ранги инструментов, предсказывающих консервативность нуклеотида. Эмпирическим путём было подобрано пороговое значение 0.7 — нуклеотид считался консервативным, если его предсказанная консервативность была выше, чем у 70% всех нуклеотидов. Это максимальное пороговое значение, которое обеспечивает распределение балла агрегатора от минимального до максимального (от 0 до 7 баз данных, считающих данный нуклеотид консервативным);
- Популяционные частоты — из всех имеющихся в базах данных по конкретному генетическому варианту была выбрана максимальная частота.

Фильтрация происходила в две стадии:

1. Фильтрация отдельных генетических вариантов на основе имеющихся аннотаций. Самая жёсткая фильтрация, которой подвергались все варианты:
    - По глубине покрытия. Генетический вариант считался существующим, если он присутствовал в двух перекрывающихся парных прочтениях, либо в чётырёх независимых прочтениях;

- Частота генетического варианта в популяции не более 3%[**?**].

Прочие фильтры были мягкими — генетический вариант отсеивался только в случае несоответствия всем указанным критериям:

- Присутствие описания связанной с геном патологии в базе данных OMIM;
- Присутствие генетического варианта в базе данных HGMD;
- Балл агрегатора патогенности экзомных вариантов не менее 3[**?**];
- Ранг «патогенный» у агрегаторов интронных или сплайс-вариантов;
- Ранги «патогенный» и «возможно патогенный» по базе данных CLINVAR;
- По функциональному классу: сдвиги рамки считывания, потери стоп- и старт-кодонов, нонсенс- и сплайс-варианты.

2. Фильтрация значимых вариантов на основе аннотаций гена. Все эти фильтры были мягкими — ген мог соответствовать одному любому из перечисленных критериев:

   - Значение pLI более 0.9, согласно рекомендациям в оригинальной статье[**?**];
   - Наследование в гене значится как «доминантное» по базе данных OMIM, либо информации о доминантности нет;
   - Любой значимый вариант в гене является гомозиготным;
   - В гене более одного значимого варианта (вероятность цис-транс-положения).

Интерпретация. Интерпретация данных и составление отчёта производилось в соответствии с рекомендациями Американского колледжа медицинской генетики и геномики (*англ.* American College of Medical Genetics, Bethesda, MD, USA) и Ассоциации молекулярной патологии[**?**].

4. Результаты

На сегодняшний день были выполнены следующие этапы работы:

1. Создание контрольной выборки генетических вариантов, с помощью которой будет проведена оценка пригодности Exo-C-библиотек к поиску генетических вариантов;
2. Проверка качества данных, полученных в результате массового параллельного секвенирования Exo-C-библиотек;
3. Разработка, отладка и тестирование автоматизированного инструмента для обработки данных секвенирования Exo-C-библиотек.

4.1. Результаты секвенирования Exo-C-библиотек

Несмотря на то, что составляющие протокола Exo-C — таргетное обогащение и Hi-C — в настоящее время достаточно отработаны, сочетание этих методик имеет свои подводные камни. Было разработано две вариации протокола Exo-C, обе этих вариации были использованы для приготовления библиотек клеточной линии K562. Результаты секвенирования этих библиотек проверялись биоинформационными методами.

Базовыми параметрами качества библиотек были приняты:

- Доля дубликатов, отражающая качество стадии ПЦР;
- Доля участков, в которых покрытие прочтениями отсутствует, а также тех, в которых оно превышает минимальный порог для анализа (10 прочтений);
- Отношение среднего покрытия вне и внутри экзома, которое можно считать показателем качества таргетного обогащения.

Данные по качеству Exo-C-библиотек представлены в Табл. 1.

Таблица 1: Данные по обогащению Exo-C-библиотек

Название	Глубина, прочтений	Доля дубликатов, %	Доля экзома с глубиной покрытия более 10, %	Среднее покрытие в экзоме	Среднее покрытие вне экзома	Обогащение экзома, раз	Доля непокрытых регионов в экзоме, %	Доля непокрытых регионов вне экзома, %
ExoC-19	136 609 179	18,86	91,68	60,51	5,56	10,89	1,75	28,12
ExoC-20	109 486 529	15,00	72,58	14,88	7,74	1,92	1,66	11,62

4.2. Автоматизация обработки данных секвенирования

При обработке данных секвенирования приходится сталкиваться с проблемами различного характера. Одними из ключевых являются проблемы использования ресурсов компьютера. Результаты секвенирования даже в сжатом виде занимают десятки и сотни гигабайт дискового пространства, и многие инструменты создают файлы с промежуточными результатами, которые занимают дисковое пространство, не неся никакой практической пользы для исследования. Кроме того, из-за вычислительной сложности обработка таких больших блоков данных может занимать дни, недели и даже месяцы работы вычислительного кластера.

Вторая, не менее важная группа проблем, связана с используемыми для обработки инструментами. Как было показано выше, стадий у обработки значительное количество, и не все стадии нужны при обработке конкретного блока данных секвенирования. Ручная настройка и контроль процесса отнимают значительное количество времени исследователя; таким образом, встаёт вопрос стандартизации и автоматизации процесса обработки данных секвенирования.

Существующие инструменты для обработки данных секвенирования были разработаны независимыми группами людей. Эти инструменты различаются по многим аспектам. Так как разработка каждого отдельного инструмента является сложным и трудоёмким процессом, целесообразно использовать их as is, а несоответствия устранять с помощью специально разработанной надстройки. Таким образом, для нами был создан пайплайн, интегрирующий все стадии обработки данных секвенирования.

Решённые задачи:

- Отказоустойчивость: максимально устранены несоответствия форматов входных и выходных данных; процесс разделён на стадии, и в случае экстренного прерывания вычислений (программного или аппаратного) предусмотрен автоматический откат.
- Оптимизация, параллелизация и масштабируемость: все процессы, которые способны использовать стандартные потоки ввода/вывода, объединены вместе, поддающиеся внешнему распараллеливанию были распараллелены, также были подобраны оптимальные параметры запуска приложений, использующих машину Java. Пайплайн может быть использован как на кластерах с большим количеством ядер и оперативной памяти, так и на относительно небольших мощностях офисных компьютеров;
- Значительно упрощены процессы развёртки и использования пайплайна: автоматизировано индексирование референсной последовательности, настройки вынесены в специальный конфигурационный файл, есть возможность обработки пула данных, используя один короткий сценарий;

Код пайплайна доступен на GitHub[**?**].

4.3. Сравнение данных секвенирования клеточной линии K562

Следующим важным этапом работы была проверка эффективности поиска генетических вариантов в Exo-C-библиотеках. Было решено использовать для этого распространённую иммортализованную клеточную линию K562, полученную от пациентки с хроническим миелолейкозом[**?**]. Данная клеточная линия была многократно секвенирована различными лабораториями с использованием различных методик приготовления библиотек. Таким образом, несмотря на то, что в этой клеточной линии наблюдается некоторая гетерогенность между лабораториями из-за большого количества пассажей, несмотря на наличие систематических ошибок при использовании разных методов секвенирования и приготовления библиотек, по K562 существует достаточное количество данных, чтобы использовать эту клеточную линию как стандарт для поиска генетических вариантов.

Результаты секвенирования клеточной линии K562 были взяты из публичных источников[**? ? ? ? ? ? ? ?**]. Использованные в этих статьях методики включают WGS, WES, Hi-C и Repli-seq. Из данных полноэкзомного секвенирования в дальнейшем были исключены все генетические варианты в интервале `chr2:25455845-25565459` с фланкированием 1 тыс. п.о. (ген DNMT3A), так как в одной из работ использовали генетически модифицированную линию с вариантами в данном гене[**?**]. В качестве тестовых Exo-C-образцов мы использовали данные, полученные на основе клеточной линии K562, имеющейся в Институте Цитологии и Генетики СО РАН. Технические данные контроля качества по тестовым и контрольным образцам представлены в Табл. 4 и 5 Приложения.

В общей сложности, объединив варианты из всех контрольных образцов, мы получили 5 496 486 различных генетических вариантов. Также в библиотеках было найдено некоторое количество уникальных генетических вариантов, встречающихся в одной

библиотеке и не встречающихся в остальных (Табл. 2). Наибольший процент уникальных вариантов найден в данных Banaszak et al.

Таблица 2: Уникальные генетические варианты в данных секвенирования контрольных образцов клеточной линии K562

Название	**Протокол**	**Глубина секвенирования, прочтений**	**Общее число вариантов**	**Уникальные варианты**	**Доля уникальных вариантов, %**
Banaszak et al.[**?**]	WES	254 983 225	408 008	41 830	10,25
Belaghzal et al.[**?**]	Hi-C	72 914 268	1 399 457	27 365	1,95
Dixon et al.[**?**]	WGS	366 291 496	4 649 012	327 184	7,03
Moquin et al.[**?**]	Hi-C	256 500 659	2 365 361	67 678	2,86
Rao et al.[**?**]	Hi-C	1 366 228 845	4 218 233	320 508	7,59
Ray et al.[**?**]	Hi-C	428 306 794	1 789 324	89 624	5,00
Wang et al.[**?**]	Repli-seq	301 663 640	2 207 451	37 578	1,70
Zhou et al.[**?**]	WGS	2 621 311 293	4 412 455	166 451	3,77

75 328 генетических вариантов были найдены в данных из всех восьми статей — их было решено использовать как «золотой стандарт». Сразу можно внимание на то, что это составляет лишь 1,37% геномных SNV клеток K562. Такая ситуация может возникнуть в следующих случаях:

1. В одной или нескольких работах обнаружено очень много уникальных вариантов, которые дают существенный вклад в общее число вариантов, но не пересекаются с результатами других исследований;

2. В одной или нескольких работах не найдено подавляющее большинство вариантов, найденных во всех остальных работах;

3. Распределение уникальных вариантов и число общих вариантов между парами работ распределены относительно равномерно, и низкое число общих для всех восьми работ вариантов не может объясняться особенностями какого-то одного или нескольких исследований.

Чтобы проверить, не связана ли низкая доля общих генетических вариантов с особенностями какого-то одного из использованных наборов данных, мы протестировали все комбинации из семи и шести работ. Результаты представлены на Рис. 3.

При исключении из выборки данных Banaszak et al. и Belaghzal et al. общими являются 1 091 331 (19,85%) вариантов. Их решено было использовать как добавочный («серебряный») стандарт.

Также было решено проверить эффективность использованного нами базового фильтра — удаление всех генетических вариантов, в которых глубина альтернативного аллеля составляет менее 4. Поиск вариантов «серебряного» и «золотого» стандартов в наших библиотеках был произведён до и после фильтрации. Результаты показаны в Табл. 3.

Таблица 3: Параметры Exo-C-библиотек. (F–) — до фильтрации по глубине альтернативного аллеля, (F+) — после фильтрации, (Δ) — изменение параметра после фильтрации в процентах

Параметр	ExoC-19			ExoC-20			В обеих			Ни в одной		
	F–	F+	Δ, %	F–	F+	Δ, %	F–	F+	Δ, %	F–	F+	Δ, %
Общее число вариантов в библиотеке	3 173 343	1 396 525	–55,99	3 750 319	2 577 934	–31,26	—	—	—	—	—	—
Вариантов «золотого стандарта»	62 335	52 732	–15,41	72 705	67 270	–7,48	60 728	48 840	–19,58	1 016	4 166	+310,04
Вариантов «серебряного стандарта»	616 375	391 273	–36,52	982 858	821 991	–16,37	580 351	340 833	–41,27	72 449	218 900	+202,14
Доля вариантов «золотого стандарта», %	82,75	70,00	—	96,52	89,30	—	80,62	64,84	—	1,35	5,53	—
Доля вариантов «серебряного стандарта», %	56,48	35,85	—	90,06	75,32	—	53,18	31,23	—	6,64	20,06	—
Доля «золотого стандарта» от всех вариантов библиотеки, %	1,96	3,78	+92,86	1,94	2,61	+34,54	—	—	—	—	—	—
Доля «серебряного стандарта» от всех вариантов библиотеки, %	19,42	28,02	+44,28	26,21	31,89	+21,67	—	—	—	—	—	—

5. Обсуждение результатов

5.1. Контрольные образцы

«Золотой стандарт» с учётом подбора библиотек скорее всего является набором генетических вариантов, относящихся к экзомным регионам. Их было обнаружено 75 тыс., что соответствует оценкам среднего количества генетических вариантов в кодирующих регионах у человека — 100 тыс.[**?**]. Общее число несоответствий с референсным геномом у среднего человека составляет от 4,1 до 5 млн[**?**], что с учётом гетерогенности клеточной линии K562 перекликается с общим количеством найденных нами генетических вариантов (5,5 млн).

Как видно из представленных выше данных, образец Banaszak et al. содержит наибольшее число уникальных вариантов (10,25%). Это может быть связано с тем, что это данные полноэкзомного секвенирования, с высоким покрытием в экзонах, где и были найдены уникальные варианты. В качестве дополнительной гипотезы можно предположить, что в этой работе использовались линии клеток, в значительной степени отличающиеся от классической линии K562.

Прослеживается ожидаемая положительная связь между глубиной секвенирования Hi-C-библиотек и количеством уникальных вариантов в них. В двух WGS-библиотеках подобной связи не наблюдается. Вероятнее всего, это также связано с отличиями использованных линий K562.

5.2. Оценка результатов секвенирования Exo-C-библиотек

В Exo-C-библиотеках глубина секвенирования составляет 136,6 млн прочтений ($2,05 \cdot 10^{10}$ п.о.) и 109,4 млн прочтений ($1,64 \cdot 10^{10}$ п.о.), а среднее покрытие в экзоме — 60,51 и 14,88 прочтений для ExoC-19 и ExoC-20 соответственно. Глубину покрытия более 10 прочтений имеют 91,68% и 72,58% экзома для ExoC-19 и ExoC-20 соответственно. Согласно [**?**], для репрезентативных результатов экзомного секвенирования необходима глубина секвенирования не менее чем в 10^{10} п.о., а для Hi-C — не менее чем

100 млн прочтений. Минимальным порогом глубины для возможности поиска генетических вариантов считается 10 прочтений, практически все гомозиготные SNV могут быть найдены при глубине в 15 прочтений, а гетерозиготные требуют глубину прочтений не менее 33. Приемлемая доля экзома с репрезентативным покрытием (более 10 прочтений) составляет 90%. Таким образом, можно утверждать, что ExoC-19 отвечает требованиям для поиска SNV, а ExoC-20, во-первых, пригодна к поиску только гомозиготных генетических вариантов, а во-вторых, имеет недостаточно хорошее покрытие в экзоме.

«Золотой стандарт» покрыт нашими библиотеками на 82,75% и 96,52%, «серебряный стандарт» — на 56,48% и 90,06% (библиотеки ExoC-19 и ExoC-20 соответственно). Различия объясняются протоколами приготовления: у библиотеки ExoC-20 выше глубина покрытия в экзоме, в 6 раз выше обогащение в экзомных районах (критерий Манна–Уитни $p = 0.0003$). Кроме того, в библиотеке ExoC-19 были использованы адаптерные последовательности, дающие большое количество шума.

Одним из базовых методов фильтрации генетических вариантов является фильтрация по глубине альтернативного аллеля. Сразу можно обратить внимание на следующее:

- В библиотеке ExoC-19 потеряна большая доля вариантов, чем в ExoC-20 — как относительно общего числа, так и относительно вариантов «золотого» и «серебряного» стандартов.
- Доли генетических вариантов «серебряного» и «золотого» стандартов в библиотеках повысились после фильтрации. Сильнее доли увеличились в библиотеке ExoC-19 по сравнению с ExoC-20.

Всё это можно объяснить наличием в библиотеке ExoC-19 большого количества регионов с низким покрытием, генетические варианты в которых были отсеяны фильтрацией по глубине. То есть, фильтрация по глубине является эффективным способом улучшения данных низкого качества.

6. Предварительные выводы

Таким образом, из приведённых нами данных можно сделать следующие выводы:

1. Пайплайн, созданный нами с учётом актуальных рекомендаций для биоинформационной обработки, позволяет эффективно обрабатывать данные NGS, улучшать их качество, а также находить в этих данных SNV.
2. Использование этого конвейера биоинформационных инструментов позволяет обнаружить около 5,5 млн генетических вариантов в контрольных данных клеточной линии K562 (что сопоставимо со средним количеством точечных полиморфизмов в геноме человека), из которых наличие 75 тыс. подтвердилось всеми восемью библиотеками, а 1 млн — шестью библиотеками с наибольшим числом совпадений, не включающими экзомные данные.

3. Сравнение генетических вариантов, полученных из контрольных образцов и Exo-C-библиотек, позволяет утверждать, что метод Exo-C способен детектировать около 90% SNV, подтверждённых всеми библиотеками (экзомные регионы), и 75% SNV, подтверждённых шестью библиотеками (весь геном).

7. План работы

В следующем семестре мы планируем:

1. Произвести анализ генетических вариантов в контрольных и наших образцах по следующим параметрам:
   (a) тип;
   (b) количество альтернативных аллелей;
   (c) распределение в геноме (в том числе с учётом проблемных регионов);
   (d) глубина покрытия;
   (e) зиготность.
2. Произвести анализ данных Exo-C на предмет систематических ошибок поиска генетических вариантов.
3. Произвести анализ результатов секвенирования Exo-C-библиотек у реальных пациентов.

A. Данные секвенирования клеточной линии K562

Таблица 4: Библиотеки данных секвенирования клеточной линии K562

Библиотека	Статья	Репозиторий	Коды доступа	Тип данных	Тип прочтений	Глубина, прочтений	Общее число прочтений	Доля картированных, % от общего числа	Доля добавочных, % от общего числа	Картированные PE прочтения	Картированные синглетоны	Дубликаты PE прочтений	Дубликаты синглетонов	Доля дубликатов, %	Оценка размера библиотеки
Контрольные данные															
GSM1551618_HIC069	Rao et al.	GEO	SRR1658693	Hi-C	PE	456 757 799	1 001 169 248	96,57	8,755	424 945 100	29 290 805	17 848 021	13 182 626	5,56	4 916 114 832
GSM1551619_HIC070	Rao et al.	GEO	SRR1658694	Hi-C	PE	591 854 553	1 314 487 595	98,7	9,949	575 565 379	15 452 072	98 778 796	8 811 532	17,69	1 478 944 337
GSM1551620_HIC071	Rao et al.	GEO	SRR1658695 SRR1658696	Hi-C	PE	79 905 895	173 931 529	98,81	8,118	77 880 938	1 975 600	486 893	269 138	0,79	6 202 732 721
GSM1551621_HIC072	Rao et al.	GEO	SRR1658697 SRR1658698	Hi-C	PE	79 578 049	159 160 116	98,38	0,003	77 155 821	2 265 995	366 805	285 395	0,65	8 088 955 029
GSM1551622_HIC073	Rao et al.	GEO	SRR1658699 SRR1658700	Hi-C	PE	77 353 816	154 710 364	98,33	0,002	74 866 287	2 383 970	240 304	293 115	0,51	11 637 260 975
GSM1551623_HIC074	Rao et al.	GEO	SRR1658702 SRR1658701	Hi-C	PE	80 778 733	175 291 763	98,65	7,835	78 467 294	2 254 814	644 986	321 965	1,01	4 746 870 162
ENCSR025GPQ	Zhou et al.	ENCODE	ENCFF574YLG ENCFF921AXL ENCFF590SSX	WGS	SE	258 022 356	260 044 021	85,39	0,777	—	220 029 156	—	50 689 083	23,04	—
ENCSR053AXS	Zhou et al.	ENCODE	ENCFF004THU ENCFF066GQD ENCFF313MGL ENCFF506TKC ENCFF080MQF	WGS	SE	1 472 492 722	1 592 540 515	91,19	7,538	—	1 332 175 586	—	496 237 198	37,25	—
ENCSR711UNY	Zhou et al.	ENCODE	ENCFF471WSA ENCFF826SYZ ENCFF590SSX	WGS	SE	890 796 215	899 473 769	99,72	0,965	—	888 239 055	—	203 498 352	22,91	—
SRX3358201	Dixon et al.	GEO	SRR6251264	WGS	PE	366 291 496	737 534 099	99,72	0,671	364 794 328	923 254	73 018 048	406 066	20,05	785 091 005
GSE148362_G1	Wang et al.	GEO	SRR11518301	Repli-seq	SE	24 804 095	24 804 396	96,39	0,001	—	23 909 072	—	921 353	3,85	—
GSE148362_G2	Wang et al.	GEO	SRR11518308	Repli-seq	SE	33 032 314	33 033 010	97,61	0,002	—	32 241 907	—	3 881 991	12,04	—
GSE148362_S1	Wang et al.	GEO	SRR11518302	Repli-seq	SE	30 884 788	30 885 298	98,7	0,002	—	30 481 936	—	2 156 480	7,07	—
GSE148362_S2	Wang et al.	GEO	SRR11518303	Repli-seq	SE	45 359 273	45 360 305	98,39	0,002	—	44 630 884	—	1 939 846	4,35	—
GSE148362_S3	Wang et al.	GEO	SRR11518304	Repli-seq	SE	49 807 076	49 807 988	98,79	0,002	—	49 205 535	—	2 889 464	5,87	—
GSE148362_S4	Wang et al.	GEO	SRR11518305	Repli-seq	SE	44 149 029	44 149 770	98,46	0,002	—	43 469 002	—	2 678 091	6,16	—
GSE148362_S5	Wang et al.	GEO	SRR11518306	Repli-seq	SE	38 424 060	38 424 835	97,96	0,002	—	37 640 056	—	3 600 260	9,57	—
GSE148362_S6	Wang et al.	GEO	SRR11518307	Repli-seq	SE	35 203 005	35 203 676	97,51	0,002	—	34 324 742	—	4 177 438	12,17	—
INSITU_HS1	Ray et al.	GEO	SRR9019504	Hi-C	PE	86 294 895	172 589 790	93,3	0	75 521 119	9 982 274	1 841 061	1 615 286	3,29	1 523 677 153
INSITU_HS2	Ray et al.	GEO	SRR9019505	Hi-C	PE	127 093 919	254 187 838	93,36	0	111 730 240	13 858 195	1 923 146	3 048 273	2,91	3 208 280 267
INSITU_NHS1	Ray et al.	GEO	SRR9019506	Hi-C	PE	86 445 594	172 891 188	93,43	0	75 893 138	9 737 847	1 903 981	1 649 376	3,38	1 487 154 386
INSITU_NHS2	Ray et al.	GEO	SRR9019507	Hi-C	PE	128 472 386	256 944 772	93,27	0	112 615 319	14 417 076	1 961 996	3 196 535	2,97	3 194 317 878
PDDE_TRANSIENT	Moquin et al.	GEO	SRR5470541 SRR5470540	Hi-C	PE	55 158 049	110 319 638	95,6	0,003	51 158 920	3 140 556	3 917 308	721 938	8,11	316 780 447
PD_STABLE_REP1	Moquin et al.	GEO	SRR5470535 SRR5470534	Hi-C	PE	67 172 619	134 347 099	97,58	0,001	64 767 511	1 565 427	5 573 966	376 260	8,79	354 373 851
PD_STABLE_REP2	Moquin et al.	GEO	SRR5470536 SRR5470537	Hi-C	PE	52 872 167	105 745 908	98,23	0,001	51 442 087	993 483	2 058 449	217 598	4,17	625 522 723
PD_TRANSIENT	Moquin et al.	GEO	SRR5470539 SRR5470538	Hi-C	PE	81 297 824	162 600 928	95,28	0,003	75 141 163	4 639 787	7 298 377	1 339 404	10,29	361 336 652
GSM2588815_R1	Belaghzal et al.	GEO	SRR5479813	Hi-C	PE	72 914 268	172 533 452	99,39	15,478	72 067 575	648 294	9 694 590	210 273	13,54	243 264 112
GSM2536769_WT	Banaszak et al.	GEO	SRR5345331	WES[1]	PE	39 211 303	78 464 649	99,46	0,054	38 914 993	171 253	7 821 960	91 145	20,17	83 342 746
GSM2536770_WT_TF	Banaszak et al.	GEO	SRR5345332	WES[1]	PE	49 394 206	98 820 633	99,54	0,033	49 068 605	193 565	10 478 814	114 795	21,43	97 869 629
GSM2536771_MT2	Banaszak et al.	GEO	SRR5345333	WES[1]	PE	42 020 936	84 093 776	99,63	0,062	41 772 436	189 177	8 755 216	104 927	21,04	85 177 326
GSM2536772_MT3	Banaszak et al.	GEO	SRR5345334	WES[1]	PE	43 669 613	87 375 385	99,6	0,041	43 414 109	164 448	9 489 133	93 601	21,92	84 242 110
GSM2536773_MT4	Banaszak et al.	GEO	SRR5345335	WES[1]	PE	39 879 263	79 788 847	99,53	0,038	39 609 943	166 651	8 590 165	90 809	21,76	77 577 055
GSM2536774_MT5	Banaszak et al.	GEO	SRR5345336	WES[1]	PE	40 807 904	81 649 292	99,59	0,041	40 559 969	163 957	8 801 283	91 545	21,77	79 383 290
Тестовые данные															
FG_ExoCBel-001	ExoC-19	—	—	Exo-C	PE	136 609 179	359 215 777	99,31	23,940	135 150 334	443 409	25 453 568	159 152	18,86	319 784 450
FG_Quarantine-A	ExoC-20	—	—	Exo-C	PE	53 598 130	140 214 460	99,79	23,150	53 598 130	259 561	7 809 282	68 779	14,60	193 853 459
FG_Quarantine-B	ExoC-20	—	—	Exo-C	PE	55 279 173	144 641 130	99,76	23,108	55 279 173	310 369	8 808 307	90 489	15,97	177 375 163

[a]Варианты в гене DNMT3A были исключены из выборки.

Таблица 5: Образцы данных секвенирования клеточной линии K562

Образец	Тип данных	Тип прочтений	Глубина, прочтений	Общее число прочтений	Доля картированных, % от общего числа	Доля добавочных, % от общего числа	FR PE прочтения, % от картированных	Картированные PE прочтения	Картированные синглетоны	Картированные на разные хромосомы пары, % от картированных	Картированные на разные хромосомы пары (QMAP 4+), % от картированных на разные хромосомы
Контрольные данные											
Rao et al.	Hi-C	PE	1 366 228 845	2 978 750 615	97,95	8,268	27,04	2 617 761 638	53 623 256	21,03	84,23
Zhou et al.	WGS	SE	2 621 311 293	2 752 058 305	93,43	4,751	—	—	—	—	—
Dixon et al.	WGS	PE	366 291 496	737 534 099	99,72	0,671	97,16	729 588 656	923 254	1,25	51,22
Wang et al.	Repli-seq	SE	301 663 640	301 669 278	98,09	0,002	—	—	—	—	—
Ray et al.	Hi-C	PE	428 306 794	856 613 588	93,33	0	35,92	751 519 632	47 995 392	22,77	76,00
Moquin et al.	Hi-C	PE	256 500 659	513 013 573	96,56	0,002	46,64	485 019 362	10 339 253	17,76	75,56
Belaghzal et al.	Hi-C	PE	72 914 268	172 533 452	99,39	15,478	24,77	144 135 150	648 294	34,02	88,02
Banaszak et al.	WES	PE	254 983 225	510 192 582	99,56	0,044	99,41	506 680 110	1 049 051	0,11	81,38
Тестовые данные											
ExoC-19	Exo-C	PE	136 609 179	359 215 777	99,31	23,94	89,22	270 300 668	443 409	5,01	66,93
ExoC-20	Exo-C	PE	109 486 529	284 855 590	99,77	23,13	70,02	217 754 606	569 930	5,87	78,00

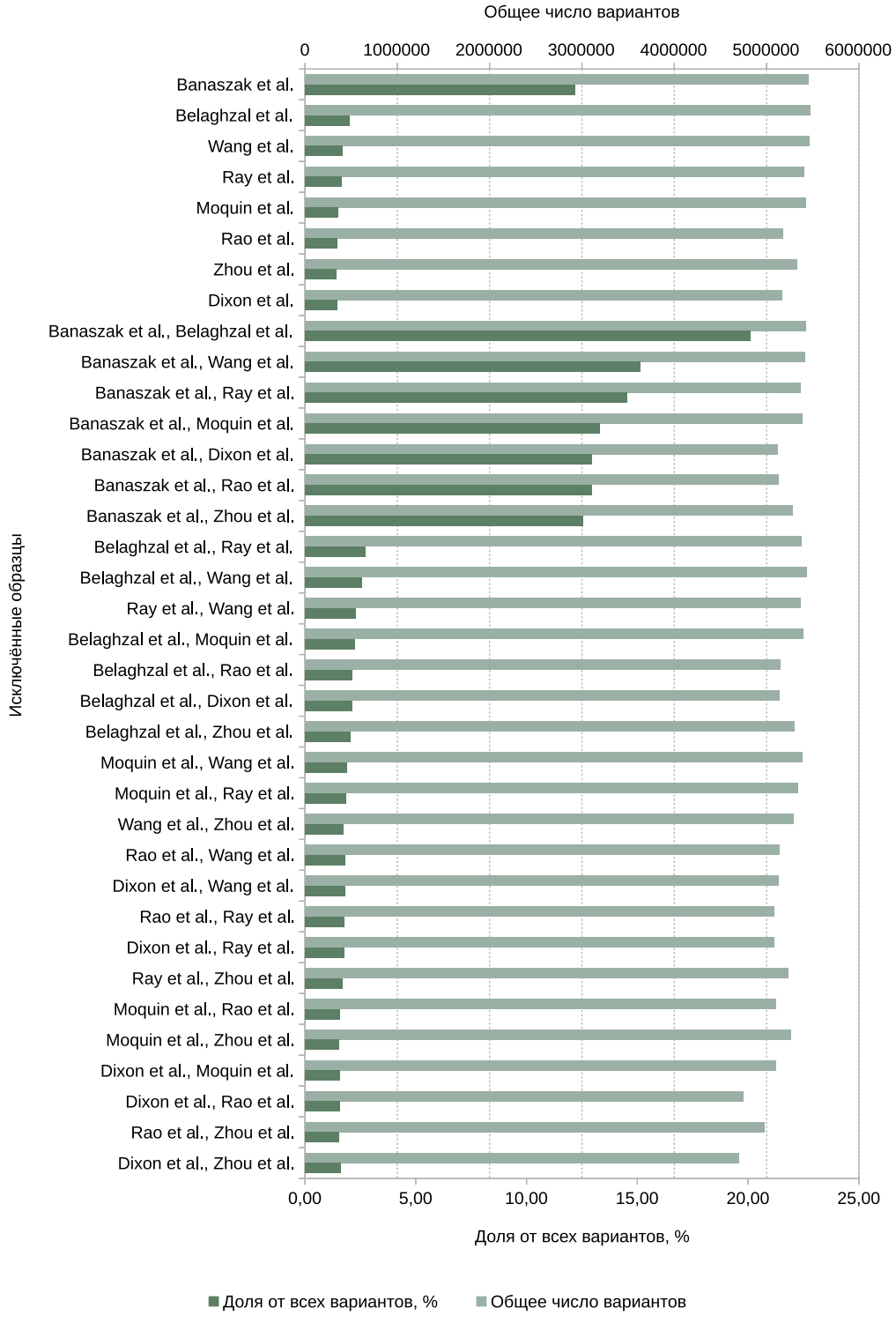

Рис. 3: Исключение образцов из выборки (7 и 6 образцов)