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

3C-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, 2021

# Содержание

1. [Введение](#_bookmark0) 4

[1.1 Актуальность](#_bookmark1) . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

[1.2 Цель](#_bookmark2) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

[1.3 Задачи](#_bookmark3) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

1. [Обзор литературы](#_bookmark4) 5
   1. [Механизмы развития генетических патологий](#_bookmark5) . . . . . . . . . 7
   2. [Типы генетических аномалий, лежащих в основе генетических патологий](#_bookmark6) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
   3. [Функциональные классы генетических вариантов](#_bookmark7) 11
   4. [Методы детекции генетических вариантов](#_bookmark8) 12
   5. [Виды NGS](#_bookmark9) 18
   6. [Базовая схема обработки результатов высокопроизводительно- го секвенирования для поиска и клинической интерпретации](#_bookmark10)

[однонуклеотидных полиморфизмов](#_bookmark10) 21

* 1. [Аннотация, фильтрация и интерпретация результатов](#_bookmark14) 28
  2. [Exo-C: суть метода](#_bookmark15) 32

1. [Материалы и методы](#_bookmark16) 33
2. [Результаты](#_bookmark17) 40
   1. [Результаты секвенирования Exo-C-библиотек](#_bookmark18) 40
   2. [Автоматизация обработки данных секвенирования](#_bookmark20) 41
   3. [Сравнение данных секвенирования клеточной линии K562](#_bookmark22) 43
3. [Обсуждение результатов](#_bookmark26) 46
   1. [Контрольные образцы](#_bookmark27) 46
   2. [Оценка результатов секвенирования Exo-C-библиотек](#_bookmark28) 47
4. [Предварительные выводы](#_bookmark29) 48
5. [План работы](#_bookmark30) 49

**Список сокращений**

**3C** (*англ.* Chromosome Conformation Capture) –– захват конформации хромо- сом

**BAM** (*англ.* Binary sequence Alignment/Map) –– бинарный файловый формат, предназначенный для хранения информации о картированных прочте- ниях

**BQSR** (*англ.* Base Quality Score Recalibration) –– рекалибровка качества про- чтений

**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 обо- гащение экзома и технологии захвата конформации хромосом

**FISH** (*англ.* Fluorescence In Situ Hybridization) –– флуоресцентная *in situ* ги- бридизация

**GATK** (*англ.* Genome Analysis ToolKit) –– набор инструментов для биоинфор- мационного 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 импринтирован- ные полимеры

**MLPA** (*англ.* Multiplex Ligation-dependent Probe Amplification) –– мультиплекс- ная лигаза-зависимая амплификация зонда

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

**NIPT** (*англ.* Non-Invasive Prenatal Testing) –– неинвазивное пренатальное те- стирование

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

**PEC** (*англ.* Primer Extension Capture) –– захват с помощью расширения прай- мера

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

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

**SMART** –– анализ транскриптома одной клетки

**SNV** (*англ.* Single Nucleotide Variant) –– однонуклеотидный генетический ва- риант

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

**VCF** (*англ.* Variant Call Format) –– формат записи генетических вариантов, най- денных 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

# Введение

## Актуальность

Наследственные заболевания являются одной из основных причин младен- ческой и детской смертности в развитых странах. Взрослые люди с такими патологиями требуют огромных затрат средств на медикаменты, оператив- ные вмешательства, специальный уход и социальные льготы. Таким образом, доступные и точные методы диагностики наследственных заболеваний мо- гут 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. Су- ществует множество методов выявления генетических вариантов, каждый ме- тод имеет свои преимущества и границы применения.

Наиболее перспективными в диагностическом и исследовательском плане в настоящее время являются методы секвенирования –– например, полноге- номное и полноэкзомное секвенирование. В Секторе геномных механизмов онтогенеза ИЦиГ СО РАН был разработан новейший метод секвенирования ––

Exo-C, сочетающий технологии экзомного обогащения с захватом конформа- ции хромосом. Потенциальным преимуществом данного метода может быть возможность поиска как крупных перестроек, так и точечных полиморфиз- мов в экзоме при относительно небольшой глубине секвенирования, от кото- рой напрямую зависит цена секвенирования. Широкий спектр применения метода и доступность в финансовом аспекте делают метод Exo-C привлека- тельным как для медико-биологических научных исследований, так и для внедрения в клиническую практику.

## Цель

Целью нашей работы является сравнение эффективности методов Exo-C, пол- ногеномного секвенирования и экзомного секвенирования для поиска точеч-

ных полиморфизмов в геномах клеток человека.

## Задачи

Основные задачи, которые необходимо решить для достижения поставлен- ной нами цели:

1. Разработать биоинформационный протокол анализа данных секвениро- вания Exo-C-библиотек.
2. Проанализировать доступные данные полногеномного, полноэкзомно- го, Hi-C и Exo-C-секвенирования для иммортализованной клеточной линии человека K562.
3. Сравнить точечные генетические варианты в геноме клеток K562, де- тектируемые при использовании полногеномного и экзомного секвени- рования, с таковыми, найденными методом Exo-C.

# Обзор литературы

Генетические варианты, их взаимодействие друг с другом и со средой опре- деляет течение болезней. Существуют генетические варианты, которые опре- деляют предрасположенность и проявляются только во взаимодействии со средой; примером могут служить варианты, определяющие предрасположен- ность к аддикциям (никотин, героин, алкоголь и пр.) [[1](#_bookmark31)]. Бывают и такие ге- нетические варианты, которые повышают восприимчивость к одному факто- ру среды и повышают устойчивость к другому, либо дают позитивный эф- фект в сочетании и негативный по отдельности. Примером может служить бета-талассемия [[2](#_bookmark32)]. Особняком стоят те варианты, которые вне зависимости от средового компонента и генетического окружения приводят к развитию заболевания (например, нейрофиброматоз I типа, который наследуется по аутосомно-доминантному типу и имеет 100% пенетрантность [[3](#_bookmark33)]).

Генетические заболевания остаются одной из основных причин мла- денческой 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 [[4](#_bookmark34)]. Взрослые люди с генетическими патология- ми требуют огромных затрат средств –– на радикальные и паллиативные опе- рации, медикаментозную поддержку (иногда пожизненную), создание усло- вий, учреждений и обучение персонала для обеспечения специализированно- го ухода.

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

**Частые и редкие (орфанные) патологии.** Генетические патологии делят- ся 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 Pre- vention Amendments of 1984”, редкими считаются патологии, поражающие менее 200 тыс. населения страны (примерно 1 : 1630 при текущей численно- сти населения в 326 млн человек) [[5](#_bookmark35)]. Европейское Медицинское Агентство определяет границу как 1 : 2000. Систематический анализ показал, что су- ществует более 290 определений, и среднее значение находится в интервале 40–50 на 100 тыс. населения [[6](#_bookmark36)].

Также сложность в определении орфанных заболеваний представляет неравномерность их распространённости в тех или иных регионах. Некото- рые заболевания могут быть орфанными в одной популяции и частыми в другой (эффект основателя, а также сверхдоминирование). Частным случа- ем эффекта основателя является атаксия Каймановых островов, связанная с гипоплазией мозжечка и сопутствующими неврологическими проявлени- ями (задержка развития, дизартрия, нистагм, интенционное дрожание). Это аутосомно-рецессивное заболевание распространено исключительно в одном регионе –– Большой Кайманов остров, гетерозиготные носители составляют около 18% местного населения [[7](#_bookmark37)]. 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 об- разом повышает устойчивость носителя бета-талассемии к малярии [[2](#_bookmark32)]. Со- ответственно, 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 [[8](#_bookmark38)]. Существуют различные базы данных, собирающие информацию по орфанным заболеваниям, наиболее известными и часто ис- пользуемыми из них являются:

* 1. Global Genes;
  2. Online Mendelian Inheritance in Man (OMIM®) [[9](#_bookmark39)];
  3. Orphanet [[10](#_bookmark40)].

Около 80% редких болезней имеют генетическую природу и начинают- ся в раннем детстве [[8](#_bookmark38)]. Таким образом, ключевым моментом для изучения данных заболеваний является понимание механизмов, лежащих в основе их развития. Количество орфанных заболеваний делает эту задачу крайне непро- стой. Тем не менее, многие механизмы на сегодняшний момент достаточно хорошо изучены. О них речь пойдёт далее.

## Механизмы развития генетических патологий

Механизмы развития генетических патологий делятся на две большие груп- пы. В первую относят изменения белок-кодирующей последовательности ге- на, приводящие к прекращению синтеза белка либо к синтезу изменённого полипептида. Ко второй группе относятся эпигенетические механизмы, не

затрагивающие непосредственно белок-кодирующие последовательности ге- нов.

Изменения белок-кодирующей последовательности гена (экзонов и сплайс- сайтов) могут приводить к замене аминокислот, сдвигам рамки считывания, появлению преждевременных стоп-кодонов и нарушениям сплайсинга. Пре-

кращение синтеза белка снижает дозу гена, а изменённый полипептид спосо- бен как потерять свою функцию, снизив таким образом дозу гена, так и приоб- рести новые свойства (токсичность). Классическим примером приобретения белком токсичности является известное наследственное нейродегенератив- ное заболевание –– аутосомно-доминантный вариант болезни Альцгеймера. Другое нейродегенеративное заболевание –– аутосомно-рецессивная болезнь Паркинсона –– может служить примером потери белком протективной функ- ции [[11](#_bookmark41)].

Также генетические патологии могут развиваться из-за эпигенетиче- ских механизмов, приводящих к изменению экспрессии генов. К таким ме- ханизмам можно отнести метилирование ДНК –– изменение молекулы ДНК без изменения нуклеотидной последовательности, а также ацетилирование гистонов [[12](#_bookmark42)].

В частности, нарушение метилирования ДНК ответственно за развитие синдрома Беквита—Видемана. Экспрессия генов CDKN1C и IGF2 регулиру- ется в зависимости от того, на материнской или отцовской хромосоме они находятся (явление геномного импринтинга). Потеря импринтинга, вызван- ная изменениями регуляторного района, ведёт к изменению экспрессии этих генов и, как следствие, к тяжёлым порокам развития, включающим висцеро- мегалию, виcцеральные грыжи, эмбриональные опухоли, пороки сердца и по- чек [[13](#_bookmark43)]. Изменение ацетилирования гистонов некоторых генов в клетках го- ловного мозга связано с развитием такого заболевания, как шизофрения [[14](#_bookmark44)]. Кроме того, на экспрессию генов в значительной степени влияет трёх-

мерная 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 боль-

шая часть промотор-энхансерных взаимодействий находится внутри тополо- гически ассоциированных доменов (ТАДов) [[15](#_bookmark45)]. В результате разрушения старых или образования новых границ ТАДов формируются структурные ва- рианты, характеризующиеся иными промотор-энхансерными взаимодействи- ями. Подобные изменения лежат в основе таких состояний, как FtM-инверсия пола (ген SOX9) и синдром Кукса (ген KCNJ2) [[16](#_bookmark46)].

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

## Типы генетических аномалий, лежащих в основе гене- тических патологий

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

Несбалансированные перестройки в большинстве своём приводят к ле- тальному исходу (в эмбриональном или детском периодах) и грубым измене- ниям фенотипа. К несбалансированным относятся:

* Анэуплоидии –– изменение числа хромосом. Примерами анэуплоидий могут служить синдром Дауна (трисомия 21 хромосомы), Эдвардса (три- сомия 18 хромосомы), Патау (трисомия 13 хромосомы), а также вари- ации числа половых хромосом (синдромы Тёрнера, Клайнфельтера и другие). Частичная моносомия –– синдром кошачьего крика (связан с утратой плеча 5 хромосомы). Прочие анэуплоидии ведут к несовмести- мым с жизнью нарушениям эмбрионального развития и, как следствие, спонтанным абортам.
* Несбалансированные транслокации –– перемещение фрагмента хромо- сомы с одного места на другое с изменением количества генетической информации. Несбалансированные транслокации могут приводить к значимым изменениям фенотипа (например, инверсия пола [[17](#_bookmark47)]) и слу- жить онкогенами [[18](#_bookmark48)].
* Вариации числа копий (*англ.* Copy Number Variations, CNV) –– дуплика- ции (мультипликации) и делеции хромосомных сегментов размером от тысячи до нескольких миллионов пар оснований. Могут возникнуть из несбалансированных транслокаций, амплификаций и собственно деле- ций. CNV способны увеличивать или уменьшать дозу гена, в значитель- ной степени влияя на его экспрессию. Различия в количестве копий мо- гут носить как положительный характер, так и отрицательный –– в част- ности, дупликации в гене CCL3L1 способны увеличить устойчивость к ВИЧ [[19](#_bookmark49)], а крупные CNV в разных частях генома ассоциированы с расстройствами аутического спектра [[20](#_bookmark50)].

Сбалансированные перестройки чаще всего характеризуются более мяг- кими фенотипическими проявлениями, а иногда и их отсутствием. К сбалан- сированным перестройкам относятся:

* Инверсии –– переворот фрагмента хромосомы. Крупные инверсии мо- гут быть причиной изменения границы ТАД, а также запирания крос- синговера и образования гаплогрупп.
* Сбалансированные транслокации –– перемещение фрагмента хромосо- мы с одного места на другое без изменения количества генетической информации. В свою очередь они делятся на реципрокные (взаимный обмен участками между негомологичными хромосомами) и Робертсо- новские (слияние акроцентрических хромосом с образованием мета- центрической или субметацентрической). Сбалансированные трансло- кации могут как не проявляться в фенотипе (сказываясь только на фер- тильности [[21](#_bookmark51)]), так и приводить к серьёзным последствиям –– напри- мер, синдрому Дауна (робертсоновская транслокация является причи- ной синдрома Дауна в 2–4% случаев [[22](#_bookmark52)]).

Самыми небольшими –– но не менее важными –– являются точечные по- лиморфизмы (*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. Так- же варианты делятся на наследуемые, которые передаются от родителей к детям, и варианты *de novo*. Согласно оценкам, предоставленным [[23](#_bookmark53)], в сред- нем 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.

## Функциональные классы генетических вариантов

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

Области гена, в которые может попасть генетический вариант:

* + - Экзоны, непосредственно отвечающие за последовательность белка. Ге- нетические варианты в экзонах могут быть синонимичными (без заме- ны аминокислоты) и несинонимичными –– миссенс (замена на другую аминокислоту), нонсенс (замена на стоп-кодон) либо сдвиг рамки счи- тывания, приводящий к изменению значительной части белковой мо- лекулы. Миссенс-варианты редко приводят к утрате функции белка, но они могут повлиять на экспрессию гена, если замена пришлась на ре- гуляторный мотив [[24](#_bookmark54)].
    - Интроны, которые содержат регуляторные области и сплайс-сайты, необ- ходимые для процессинга транскрипта в готовую мРНК, а также 3’- нетранслируемая область (*англ.* 3’-untranslated region, 3’UTR) и 5’-нетранс- лируемая область (*англ.* 5’-untranslated region, 5’UTR), вовлечённые в регуляцию транскрипции, трансляции и деградации транскрипта. В част-

ности, в 5’UTR находится так называемая консенсусная последователь- ность Козак, важная для инициации трансляции мРНК [[25](#_bookmark55)]. Также из- вестно, что в 5’UTR могут находиться открытые рамки считывания, ко- торые влияют на поведение рибосомы –– могут вызывать её торможе- ние, диссоциацию, либо перекрывать основной старт-кодон гена [[26](#_bookmark56)]. Генетические варианты могут как разрушать канонические сплайс-сайты, так и способствовать образованию новых внутри интронных участков [[27](#_bookmark57)]. Влияние генетических вариантов в этих областях недостаточно изуче- но, и их связь с конкретной патологией у пациента порой достаточно трудно доказать. Тем не менее, существуют специальные инструмен- ты, 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 свя- зывания белков, отвечающих за процессы метилирования или трёхмерную организацию хроматина.

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

## Методы детекции генетических вариантов

**Кариотипирование.** Данный метод представляет собой микроскопическое исследование клеток, синхронизированных на стадии метафазы митоза. Од- нако простое микроскопическое исследование хромосом плохо подходит для обнаружения генетических вариантов, поэтому были разработаны различные методы окраски (бэндинга), позволяющие отдифференцировать отдельные хромосомы и хромосомные регионы [[28](#_bookmark58)]:

1. Q-окрашивание –– позволяет отдифференцировать все хромосомы, при- меняется для исследования Y-хромосомы (быстрое определение гене-

тического пола, выявление мозаицизма по Y-хромосоме, транслокаций между Y-хромосомой и другими хромосомами). Окрашивание легко снимается, что позволяет использовать этот метод для последователь- ной окраски и изучения хромосом;

1. G-окрашивание –– наиболее часто используемый метод. Позволяет от- дифференцировать все хромосомы, гарантирует стойкое окрашивание, легко поддаётся фотографированию.
2. R-окрашивание –– визуализирует концы хромосом, а также специфиче- ские именно для этого окрашивания бэнды (так называемые R-позитив- ные бэнды).
3. C-окрашивание –– применяется для анализа вариабельной дистальной части Y-хромосомы, а также центромерных регионов прочих хромо- сом, содержащих конститутивный гетерохроматин. Хорошо подходит для выявления перестроек, затрагивающих гетерохроматиновые реги- оны. Кроме того, C-окрашиванием хорошо определяются кольцевые и дицентрические хромосомы;
4. NOR-окрашивание –– визуализирует ядрышковые организаторы (*англ.* Nucleolus Organizer Region, NOR), богатые рибосомальными генами;
5. DA–DAPI-окрашивание –– применяется для идентификации центромер- ных гетерохроматизированных районов.

Окрашенные хромосомы далее изучаются на предмет формы, количе- ства и наличия перестроек.

Кариотипирование –– рутинная методика при диагностике врождённых патологий, аутопсии мертворожденных и злокачественных образований кро- ветворного ряда. Преимущества кариотипирования в том, что данным мето- дом можно охватить весь геном, визуализации поддаются отдельные клетки и отдельные хромосомы. Ограничения –– обязательно требуются живые клет- ки, также на эффективность влияет размер перестроек (не менее 1–5 млн п.о.) и процент поражённых клеток в образце (минимум 5–10%) [[29](#_bookmark59)].

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 пока- зано далее, развивались не только в направлении увеличения разрешающей способности, но и облегчения интерпретации полученных данных.

**Флуоресцентная *in situ* гибридизация** (*англ.* Fluorescence In Situ Hybridiza- tion, FISH). Основой является гибридизация нуклеиновых кислот образца и комплементарных им проб, содержащих флуоресцентную метку. Гибридиза- ция 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 опреде- лении хромосомных патологий –– как в клетках с врождёнными перестройка- ми, так и в клетках опухолей.

Данные при помощи метода FISH можно получить, анализируя отсут- ствие или присутствие сигналов от использованных флюорофоров. Количе- ство различимых цветовых меток равно (2*x −* 1), где *x* –– количество флюоро- форов. Это позволяет реализовать, например, спектральное кариотипирова- ние (*англ.* Spectral Karyotyping, SKY), при котором каждая хромосома окра- шивается в свой собственный цвет и межхромосомные перестройки видны даже начинающему специалисту [[30](#_bookmark60)]. Тем не менее, лимитирующими факто- рами остаются:

* + - потребность в хорошо обученном персонале. Относительная простота интерпретации результатов сочетается со сложностью протокола при- готовления 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 уча-

сток ДНК. Поэтому детектировать изменения участков размером менее 100 тыс. п.о. достаточно затруднительно.

В настоящее время методика FISH значительно усложнилась. Биотех- нологические компании предлагают панели олигонуклеотидов, определяю- щие специфические участки размером от десятков тысяч до миллиона пар оснований, а также олигонуклеотиды с высокой чувствительностью, позволя- ющие определить сплайс-варианты и даже SNV. Разрабатываются техноло- гии micro-FISH (*µ*FISH), сочетающие FISH с микрофлюидными технология- ми (проведение реакций в микроскопических объёмах жидкости). При этом процесс удешевляется, автоматизируется, ускоряется (за счёт уменьшения объёмов, а соответственно, и времени гибридизации) и упрощается для ис- пользования в обширных исследованиях и для внедрения в клинику [[31](#_bookmark61)].

**Сравнительная геномная гибридизация** (*англ.* Comparative Genomic Hy- bridization, CGH). Как и в случае с методом FISH, основой данного метода является флуоресцентная гибридизация. Однако CGH использует два образ- ца генома –– тестовый и контрольный, каждый из которых метится флюоро- фором, 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 перестро- ек и не требует знаний о целевом регионе. К ограничениям анализа относит- ся 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 комби- нируется с микрочиповой гибридизацией [[32](#_bookmark62)]. ДНК-микрочипы, или микро- матрицы, представляют собой сотни тысяч или миллионы однонитевых фраг- ментов ДНК (зондов), которые ковалентно пришиты к основанию (микрочи- пу). При ХМА на микрочип наносятся контрольные фрагменты генома либо контрольные последовательности генов, которые могут быть связаны с кон- кретной патологией. Порядок зондов на чипе строго определён, что упрощает

локализацию и определение характера перестройки.

С помощью сравнительной гибридизации геномов могут быть обнару- жены самые разные структурные вариации –– CNV, инверсии, хромосомные транслокации и анэуплоидии. Для этого используются длинные зонды, ко- торые позволяют проводить гибридизацию последовательностей, имеющих некоторые различия. Когда пробы ДНК короткие, эффективность гибридиза- ции очень чувствительна к несовпадениям; такие зонды облегчают сравне- ние геномов на нуклеотидном уровне (поиск SNV).

Микроматрицы предлагают относительно недорогие и эффективные сред- ства сравнения всех известных типов генетических вариаций. Однако для та-

ких целей, как обнаружение неизвестных или часто повторяющихся последо- вательностей, эти методы не подходят [[33](#_bookmark63)].

**Мультиплексная лигаза-зависимая амплификация зонда** (*англ.* Multi- plex Ligation-dependent Probe Amplification, MLPA). Основой MLPA является ПЦР-амплификация специальных проб, гибридизующихся с целевыми райо- нами ДНК. Каждая проба представляет собой пару полу-проб; каждая полу- проба имеет комплементарную геному часть и технические последователь- ности –– праймер для ПЦР и вставки, обеспечивающие большой размер про- дукта амплификации. Если полу-пробы гибридизуются с геномом без зазора, они лигируются и впоследствии амплифицируются; лигированные пробы от- личаются 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 определе- ния этих участков также применяют электрофорез, т.к. не подвергшаяся гид- ролизу ДНК по длине значительно превосходит фрагменты гидролизованных рестриктазой метилированных регионов.

Слабым местом MLPA остаётся интерпретация результатов. Опреде- ление гомозиготных CNV не представляет труда –– их распознают по нали- чию/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-ре- акции и различий между образцами [[34](#_bookmark64)].

Как мы видим, перечисленные методы имеют один серьёзный недоста- ток –– они могут определить наличие или отсутствие, совпадение или несов- падение, но не способны прочитать априори неизвестную последовательность ДНК. Специально для этого были разработаны методы секвенирования.

**Секвенирование по Сэнгеру.** Исторический метод, позволяющий с высо- кой точностью анализировать короткие (до 1 тысячи п.о.) фрагменты ДНК [[35](#_bookmark65)]. Суть его состоит в проведении обычной реакции амплификации ДНК, только в смесь дезоксирибонуклеотидов (dNTP) добавлены дидезоксирибонуклео- тиды (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 об- разуется смесь из меченых цепей разной длины. При разделении этой смеси на электрофорезе проявляется характерная «лестница», последовательность флуоресцентных сигналов в которой совпадает с последовательностью иссле- дуемой ДНК.

Основным недостатком секвенирования по Сэнгеру является ограниче- ние длины исследуемого фрагмента ДНК.

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

**Секвенирование нового поколения** (*англ.* New Generation Sequencing, NGS). Это комплекс технологий, позволяющих прочитать за сравнительно неболь- шое время миллионы последовательностей ДНК. Благодаря этому единовре- менно можно проанализировать несколько генов, либо весь геном.

В методах NGS наблюдается развитие двух основных парадигм, раз- личающихся по длине прочтений. Секвенирование короткими прочтениями характеризуется меньшей ценой и более качественными данными, что позво- ляет применять данные методы в популяционных исследованиях и клиниче-

ской практике (поиск патогенных генетических вариантов). Секвенирование длинными прочтениями хорошо подходит для сборки новых геномов и изу- чения отдельных изоформ генов [[36](#_bookmark66)]. Количество различных методов в на- стоящее время значительно, но самым часто используемым является метод Illumina (короткие прочтения).

Основные проблемы данных NGS:

* + - Финансовые вложения и время, затраченные на секвенирование и ана- лиз данных. По-прежнему остаются лимитирующим фактором приме- нения NGS в клинической практике;
    - Ошибки секвенирования и ПЦР. Их значимость уменьшается с увели- чением покрытия, но не исчезает полностью;
    - Неоднородность покрытия генома или таргетных регионов прочтения- ми. Это может быть связано как с недостатками приготовления библио- теки, так и с проблемами картирования.

## Виды NGS

**Полногеномное секвенирование** (*англ.* Whole Genome Sequencing, WGS). Приготовление библиотек при полногеномном секвенировании производит- ся 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 поис- ка SNV, небольших делеций и инсерций. Полногеномное секвенирование со слабым покрытием может быть использовано для определения CNV –– напри- мер, при неинвазивном пренатальном тестировании (*англ.* Non-Invasive Pre- natal Testing, NIPT), когда используется свободная ДНК плода (*англ.* Cell-Free Fetal DNA, cffDNA), циркулирующая в крови матери [[37](#_bookmark67)].

**Таргетные панели.** Основой данных методов является обогащение целе- вых регионов генома. Методов обогащения существует достаточно много, но все они делятся на 4 основные категории [[38](#_bookmark68)]:

1. Твердофазная гибридизация. Для этого используют комплементарные целевым регионам короткие ДНК-пробы, зафиксированные на твёрдом основании (микрочипе). После гибридизации нецелевую ДНК вымыва- ют, а целевые фрагменты остаются на чипе.
2. Жидкофазная гибридизация. Эти методы характеризуются тем, что ДНК- пробы находятся в растворе и помечены специальной молекулой (на- пример, биотином). После гибридизации с целевой ДНК пробы вылав- ливают бусинами, поверхность которых способна связывать молекулы биотина.
3. Полимеразно-опосредованный захват. В этих методах ПЦР производят на стадии обогащения. Например, методы молекулярно импринтиро- ванных полимеров (*англ.* Molecularly Imprinted Polymers, MIP) и ана- лиза транскриптома одной клетки (SMART) используют длинные про- бы, содержащие как праймер, так и регион для остановки элонгации и инициации лигирования. После элонгации и лигирования получают- ся кольцевые молекулы, содержащие целевой регион; линейные моле- кулы в последующем удаляют из раствора. Метод захвата с помощью расширения праймера (*англ.* Primer Extension Capture, PEC) использу- ет биотинилированные праймеры, которые гибридизуются с целевыми регионами и элонгируются; далее их вылавливают бусинами, как в ме- тодах жидкофазной гибридизации.
4. Захват регионов. Включает в себя сортировку и микродиссекцию хро- мосом, благодаря чему можно обогатить библиотеку фрагментов после- довательностями отдельной хромосомы или даже её части. Это методы, требующие чрезвычайно сложных техник и хорошо обученный персо- нал, но очень полезные в отдельных ситуациях.

Данный вид тестов позволяет анализировать гены, ответственные за от- дельные группы заболеваний –– например, существуют таргетные панели для иммунодефицитов, почечных, неврологических болезней, болезней соедини- тельной ткани, сетчатки, а также предрасположенности к отдельным видам онкологических заболеваний. Таргетные панели позволяют анализировать и

клетки опухолей –– некоторые приспособлены к выявлению общих для мно- гих раковых линий мутаций, другие же разработаны для специфического ти- па опухолей [[39](#_bookmark69)].

**Полноэкзомное секвенирование** (*англ.* Whole Exome Sequencing, WES). Техника заключается в секвенировании обогащённого экзома –– совокупно- сти 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 жидкофаз- ная гибридизация [[38](#_bookmark68)].

У человека экзом составляет примерно 1% от генома, или примерно 30 млн п.о. (суммарно). При этом более 80% генетических вариантов, которые представлены в базе данных известных геномных вариантов CLINVAR [[40](#_bookmark70)], и из них более 89% вариантов, которые отмечены как «патогенные», отно- сятся к белок-кодирующим областям генома; эта цифра приближается к 99%, если учитывать ближайшие окрестности экзонов [[41](#_bookmark71)]. Таким образом, пол- ноэкзомное секвенирование намного лучше подходит для обычной клиниче- ской практики, нежели полногеномное. Кроме того, полноэкзомное секвени- рование значительно дешевле, что увеличивает его доступность и позволяет, например, произвести тестирование ребёнка и родителей (так называемый трио-тест) и, как следствие, улучшить интерпретацию вариантов [[39](#_bookmark69)].

**Технологии захвата конформации хромосом** (*англ.* Chromosome Confor- mation Capture, 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, лиги- руют, затем продукты лигазной реакции секвенируют при помощи NGS. Во время лигирования ковалентно связанными могут оказаться только те участ- ки, которые физически находятся близко друг от друга. Картирование химер- ных прочтений с помощью специальных инструментов позволяет узнать, ка- кие именно участки генома были связаны, а значит, распологались близко друг к другу в пространстве ядра [[42](#_bookmark72)]. 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 картирова- ны на два разных бина, называется контактом между этими районами. Прак- тическое значение имеет информация об относительной частоте контактов между бинами.

В настоящее время существует множество вариантов протокола 3C. Са- мым 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 [[43](#_bookmark73)].

Результаты NGS представляют собой гигантские блоки данных, содер- жащие всевозможные ошибки. Обработка данных секвенирования –– это вы- сокотехнологичная отрасль, которая позволяет получить из этих данных прак- тически значимую информацию и минимизировать влияние ошибок на эту информацию.

## Базовая схема обработки результатов высокопроизводи- тельного секвенирования для поиска и клинической ин- терпретации однонуклеотидных полиморфизмов

**Демультиплексикация.** В процессе приготовления NGS-библиотеки к це- левым фрагментам ДНК лигируют так называемые адаптерные последова- тельности, или адаптеры. Очень часто потенциальное количество прочтений, которое способен выдать секвенатор за один запуск, значительно превыша- ет требуемое количество прочтений для отдельной библиотеки, поэтому из соображений экономии и повышения производительности на одном чипе се- квенируют сразу несколько библиотек. Для этого в адаптеры вставляют бар- коды –– последовательности, с помощью которых можно отличить прочтения, относящиеся к разным библиотекам или образцам. Процесс сортировки дан- ных секвенирования по баркодам называется демультиплексикацией.



Рис. 1: Результаты секвенирования библиотек, содержащих короткие после- довательности, могут быть контаминированы адаптерами. На рисунке показа- ны результаты секвенирования последовательностей, длина которых превы- шает количество циклов секвенирования (сверху), либо значительно меньше количества циклов (снизу).

**Удаление адаптерных последовательностей.** Если целевой фрагмент ДНК короче длины прочтения, то фрагменты адаптерной последовательности мо- гут попасть в готовые данные (Рис. [1](#_bookmark11)). Это замедляет работу алгоритма карти- рования, а порой в значительной степени ухудшает его результаты, поэтому встаёт вопрос об удалении адаптерных последовательностей. Также присут- ствие адаптера в прочтениях может быть признаком контаминации библио- теки, и такие прочтения следует исключить из дальнейшего анализа [[44](#_bookmark74)].

**Картирование прочтений.** Как уже упоминалось выше, результаты NGS –– это прочтения, содержащие небольшие (в пределах 200 п.о.) фрагменты ге- нома. Извлечение информации из необработанных результатов секвенирова- ния 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 после- довательность. Алгоритм картирования представляет собой очень сложную систему, которая учитывает последовательность букв в прочтении и качество прочтения. Качество прочтения отражает вероятность того, что буква, прочи- танная секвенатором, совпадает с реальным нуклеотидом в данной позиции. Обычно качество прочтения записывается в шкале Phred, к которой приво- дится формулой

*Q* = *−*10 log10 *P,* (1)

где *P* –– вероятность того, что нуклеотид прочтен правильно. Было разрабо- тано множество алгоритмов картирования, но в настоящее время «золотым

стандартом» являются утилиты, использующие алгоритм Берроуса–Уиллера [[45](#_bookmark75)].

Обычно алгоритм картирования выставляет выравниванию коэффици- ент, называемый качеством выравнивания (*англ.* MAPping Quality, MAPQ). MAPQ отражает вероятность правильности картирования и также записыва- ется в шкале Phred (Формула [1](#_bookmark12)). В силу размеров референсной последователь- ности в ней существует огромное множество повторов и похожих регионов. Современные алгоритмы могут находить несколько потенциальных мест кар- тирования для одного прочтения, и их количество влияет на качество вырав- нивания.

Также алгоритмы способны разделять прочтение на участки, которые могут быть картированы в разные места генома. По этому признаку прочте- ния 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 направ- лением. Эти участки могут перекрываться, и количество перекрытий также влияет на MAPQ.

Исходя из особенностей алгоритмов картирования, выравнивания де- лятся на следующие классы:

* Первичное выравнивание (*англ.* primary) –– выравнивание наиболее круп-

ного (и содержащего наименьшее количество перекрытий, в случае хи- мерного прочтения) фрагмента прочтения с наиболее высоким MAPQ. Первичное выравнивание только одно. Первичное выравнивание хи- мерного прочтения называется репрезентативным;

* Вторичное выравнивание (*англ.* secondary) –– выравнивание наиболее крупного фрагмента прочтения с меньшим MAPQ. Вторичных вырав- ниваний может быть несколько (в зависимости от выставленного ниж- него порога MAPQ);
* Добавочное выравнивание (*англ.* supplementary) –– выравнивание менее крупных (либо содержащих большее количество перекрытий) фрагмен- тов прочтения. Добавочные выравнивания характерны только для хи- мерных прочтений.

Картированный участок может содержать в себе несовпадения с ре- ференсной последовательностью, инсерции и делеции. Это могут быть как ошибки, так и генетические варианты, поэтому данная информация безуслов- но 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 учитыва- ются при анализе. Клипирование бывает двух типов:

* Мягкое клипирование (*англ.* soft-clip) –– отсечение невыравненного кон- ца прочтения с сохранением полной последовательности прочтения. В отсечённых методом мягкого клипирования регионах могут быть адап- терные последовательности, а также часть химерного прочтения (в ре- презентативном выравнивании).
* Жёсткое клипирование (*англ.* hard-clip) –– отсечение невыравненного кон- ца прочтения без сохранения его последовательности. В регионах, под- вергшихся жёсткому клипированию, обычно находятся репрезентатив- ные участки химерных прочтений (в добавочных выравниваниях).

Основные проблемы картирования:

* Высоковариативные регионы. Алгоритм картирования разработан для поиска наиболее полных соответствий, и при большом количестве несов- падений прочтение просто не сможет быть картировано на нужный ре- гион генома;
* Вырожденные (неуникальные) регионы. Соответствие между региона- ми может привести к неправильному распределению прочтений между ними, а значит –– и неправильному картированию генетических вариа- ций. Кроме того, генетические варианты в регионах с короткими по- вторами 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 прочте- ния 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 по- казан на Рис. [2](#_bookmark13).





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

**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 при- боров. Решением этих проблем является амплификация молекул ДНК. Ам- плификация может быть как на стадии приготовления библиотеки (ПЦР), так и на стадии секвенирования. При секвенировании амплификация и последую- щее объединение ампликонов в кластер производятся для усиления сигнала и нивелирования ошибок, происходящих на каждом цикле секвенирования с отдельными молекулами. Соответственно, в процессе секвенирования воз- никают дубликатные прочтения, которые могут быть как ПЦР-дубликатами библиотеки, так и возникать из-за ошибок распознавания кластеров ампли- фикации (оптические дубликаты). Согласно принятой практике, дубликаты должны быть удалены или помечены для улучшения поиска генетических ва- риантов [[46](#_bookmark76)].

Однако было показано, что для WGS-данных удаление дубликатов име- ет минимальный эффект на улучшение поиска полиморфизмов –– приблизи- тельно 92% из более чем 17 млн вариантов были найдены вне зависимости от наличия этапа удаления дубликатов и использованных инструментов для поиска дубликатов [[47](#_bookmark77)]. Учитывая, что удаление дубликатов может занимать значительную часть потраченного на обработку данных времени, следует взвесить пользу и затраты данного этапа для конкретной прикладной задачи.

**Рекалибровка качества прочтений** (*англ.* Base Quality Score Recalibration, BQSR). В приборной оценке качества прочтений всегда имеют место систе- матические ошибки. Это связано как с особенностями физико-химических реакций в секвенаторе, так и с техническими недостатками оборудования. Вычисление качества прочтения –– сложный алгоритм, защищённый автор- скими правами производителя секвенатора. Вместе с тем от качества прочте- ний напрямую зависит алгоритм поиска вариантов –– он использует данный коэффициент как вес в пользу присутствия или отсутствия генетического ва- рианта в конкретной точке генома.

Решением является рекалибровка качества прочтений, представляющая собой корректировку систематических ошибок, исходя из известных паттер-

нов зависимости случайных величин. Следует заметить, что рекалибровка не помогает определить, какой нуклеотид в реальности находится в данной по- зиции –– она лишь указывает алгоритму поиска генетических вариантов, вы- ше или ниже вероятность правильного прочтения нуклеотида секвенатором.

Первоочередное влияние на ошибки оказывают:

1. Собственно прибор (секвенатор) и номер запуска. Большая часть секве- наторов выставляет прочтению более высокое качество прочтения по сравнению с ожидаемым, гораздо реже встречаются модели, занижаю- щие качество прочтения [[46](#_bookmark76)]. Каждый отдельный запуск может разли- чаться по параметрам чипа и химических реагентов;
2. Цикл секвенирования. Качество прочтения уменьшается с каждым цик- лом 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 вариан- тов. Высокая изменчивость 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 [[46](#_bookmark76)].

**Поиск генетических вариантов.** Невозможно точно сказать, какой нуклео- тид 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 ге- нома образца от референсного генома называется генетическим вариантом. Алгоритм выставляет каждому генетическому варианту коэффициент каче- ства варианта (VCF Qual), записываемый в шкале Phred (Формула [1](#_bookmark12)). Помимо определения генетического варианта, алгоритм может определять его зигот- ность.

Также важным этапом поиска вариантов является уже упомянутое вы- ше левое выравнивание. Варианты в повторяющихся последовательностях с длиной менее длины одного прочтения невозможно точно локализовать, по- этому они всегда сдвигаются как можно левее относительно последователь- ности генома. Это чрезвычайно важно при аннотации генетических вариан- тов, так как все БД используют данные с левым выравниванием, и неправиль- ная 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) [[48](#_bookmark78)], 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.

## Аннотация, фильтрация и интерпретация результатов

Первое, что следует сделать –– это определить, насколько генетический вари- ант 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 ге- нетического 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 [[49](#_bookmark79)], по- этому, во избежание ошибок, рекомендуется использовать их все. Это

особенно важно при дифферециации генетических вариантов с высо- кой вероятностью повреждающего эффекта (сдвиги рамок считывания,

нонсенс-кодоны). Кроме того, различаются алгоритмы определения функ- ционального класса в различных утилитах аннотации, что также созда-

ёт определённые трудности [[50](#_bookmark80)].

1. Клиническая значимость гена. Количество генетических вариантов для поиска можно сузить, зная, какие именно гены могут быть связаны с на- блюдаемым у пациента фенотипом. Для поиска генов по клинической значимости существуют такие базы данных, как OMIM [[9](#_bookmark39)] и OrphaData [[10](#_bookmark40)].
2. 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 та- ких показателей является pLI.

Основные проблемы pLI [[51](#_bookmark81)]:

* + Плохо приспособлен к распознаванию аутосомно-рецессивных ва- риантов (из-за того, что частота повреждающих вариантов в попу- ляции может быть высокой) и X-сцепленных рецессивных вари- антов (из-за наличия в популяции здоровых гетерозиготных носи- телей).
  + Плохо приспособлен к распознаванию генетических вариантов в генах, ответственных за патологии, не влияющие на взросление и воспроизводство. Их частота в популяции также может быть вы- сокой. К таким относятся варианты в генах BRCA1 и BRCA2, от- ветственных 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 груп- пе. 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 C-тер- минальной 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 мень- шей 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 показа- телем LoF, низкое –– с осторожностью.

Аннотация по координате обычно предназначена для миссенс-, интрон- ных и сплайс-вариантов, связь которых с патологическим состоянием значи- тельно сложнее выявить и доказать.

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

По мере развития методов NGS и увеличения их доступности, нача- ли появляться базы данных, агрегирующие результаты секвенирования различных популяций, а значит –– способные определить частоту гене- тических вариантов в популяции. В настоящее время наиболее круп- ной является gnomAD [[52](#_bookmark82)], поглотившая существовавший ранее ExAC, содержавший исключительно экзомные данные. Она содержит часто- ты генетических вариантов для всех основных рас, а также некоторых условно-здоровых групп.

Несмотря на то, что были созданы базы данных для всех рас, очень ча- сто этого недостаточно и необходимо учитывать частоты в популяциях отдельных народов и стран. Такими базами данных являются GME [[53](#_bookmark83)],

в которой отражены частоты по популяции Ближнего Востока, ABraOM [[54](#_bookmark84)], предоставляющая частоты генетических вариантов среди практически

здорового пожилого населения Бразилии. Также для анализа берутся популяции, в которых велика доля близкородственных связей, напри- мер, пакистанская [[55](#_bookmark85)].

1. Клинические данные из БД и статей. Наиболее достоверным источни- ком данных о патогенности генетического варианта являются семейные и популяционные исследования конкретной патологии, а также базы данных, агрегирующие информацию из подобных статей. Наиболее ис- пользуемыми в настоящее время являются HGMD [[56](#_bookmark86)] и CLINVAR [[40](#_bookmark70)]. Тем не менее, CLINVAR считается лишь дополнительным источником, так как часто содержит информацию низкого качества [[57](#_bookmark87)].
2. Анализ и предсказание функционального эффекта *in silico*. *In silico* ме- тоды появились в ответ на необходимость как-то классифицировать ге- нетические варианты, по которым недостаточно клинической информа- ции. 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 популяци- ям с плохо изученным составом генетических вариантов, представляют ещё большую проблему.

Поэтому были разработаны инструменты на основе машинного обуче- ния, предсказывающие консервативность районов и патогенность гене- тических вариантов на основе имеющихся данных –– положения отно- сительно гена и его функциональных элементов, характера замены, а также клинической информации об известных заменах [[24](#_bookmark54)]. Предска- зательная способность отдельных инструментов оставляет желать луч-

шего, поэтому чаще всего в клинической практике используются агре- гаторы, собирающие предсказания с большого числа известных *in silico* инструментов.

Значимость вклада каждого отдельного фактора достаточно сложно оце- нить. Эту проблему решают калькуляторы патогенности, которые по специ- альным критериям присваивают генетическому варианту ранг, отражающий вероятность повреждающего действия [[57](#_bookmark87)].

**Когортный и семейный анализ.** В случае, если исследователь имеет до- ступ к группе, представители которой связаны узами крови с пациентом, есть возможность провести семейный анализ. Семейный анализ нужен для уста- новления путей наследования тех или иных генетических вариантов в родо- словной. Это позволяет уточнить их связь с фенотипом. Также анализ несколь- ких родственных образцов помогает определить зиготность варианта, обна- ружить генетические варианты *de novo*, либо импутировать район с недоста- точным покрытием.

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

## Exo-C: суть метода

Как уже упоминалось выше, одним из основных ограничений NGS-техно- логий в настоящее время является их цена, напрямую зависящая от глуби- ны секвенирования библиотеки. Есть ограничения и по возможностям по- иска тех или иных генетических вариантов. 3C-методы на сегодняшний мо- мент являются наиболее перспективным способом обнаружения хромосом- ных перестроек [[58](#_bookmark88)], но при небольшой глубине секвенирования в них обна- ружение точечных полиморфизмов затруднительно [[59](#_bookmark89)]. WGS способно об-

наруживать большую часть SNV, небольших инсерций и делеций, но требу- ет большую глубину секвенирования [[59](#_bookmark89)]; WES, с другой стороны, позволяет выявить генетические варианты при небольшой глубине секвенирования, но только в экзоме. Возможности обнаружения хромосомных перестроек для по- следних двух методов ограничены.

Компромиссом между ценой и возможностями поиска генетических ва- риантов может служить новейший метод Exo-C, сочетающий технологии тар- гетного обогащения с 3C. Суть его заключается в приготовлении Hi-C-биб- лиотеки и последующем обогащении только тех последовательностей, кото- рые связаны с экзомом. Таким образом, с его помощью можно как искать точечные варианты в обогащённых регионах (за счёт большой глубины по- крытия в них), так и хромосомные перестройки во всём геноме (за счёт Hi-C, дающей относительно небольшое, но доступное для анализа покрытие всего генома) [[60](#_bookmark90)].

Тем не менее, как выяснилось, уже существующие биоинформацион- ные методы следует модифицировать для корректной обработки данных Exo-

C. Это связано в первую очередь с особенностями протокола Hi-C, к приме- ру, наличием технических последовательностей (бридж-адаптеров), которые приводят к появлению ложных SNV в экзомных регионах. Таргетное обога- щение, 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 секвениро- вания.

# Материалы и методы

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

**Контроль качества NGS-данных.** Для контроля качества прочтений мы использовали утилиту FastQC [[61](#_bookmark91)], способную оценивать наличие адаптер- ных последовательностей, распределение прочтений по длине, GC-состав про- чтений, а также производить анализ зависимости нуклеотидного состава от позиции в прочтении. Критерии качества были использованы согласно про- токолу разработчика [[61](#_bookmark91)].

**Удаление адаптерных последовательностей.** Удаление адаптерных после- довательностей производилось с помощью утилиты cutadapt [[44](#_bookmark74)]. В [[46](#_bookmark76)] реко- мендуется 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 сде- лан 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. Пребразо- вание FastQ-файлов в uBAM не предотвращает потерю метаданных, но зна- чительно 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 [[6](#_bookmark92) и BWA [[63](#_bookmark93)]. BWA показал лучшие результаты; кроме того, он значительно

более эффективно работает с химерными ридами, что немаловажно для ис- пользуемого нами метода Exo-C.

Для картирования был взят геном GRCh37/hg19, предоставленный NCBI. Из него были удалены так называемые неканоничные хромосомы (некартиро- ванные/вариативные референсные последовательности), что позволило улуч- шить качество выравнивания и значительно упростить работу с готовыми данными.

Кроме того, для правильного функционирования инструментов на даль- нейших 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, осно- ванные на следующих требованиях [[46](#_bookmark76)]:

* + 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 по- мещать в него информацию о чипе секвенирования (баркод чипа), ячей- ке и баркоде (номере) образца. Во время BQSR поле PU является прио- ритетным по отношению к ID.
  + Поле LB является уникальным для каждой библиотеки, приготовлен- ной 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 инструмен- та samtools flagstat [[64](#_bookmark94)].

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

Удаление дубликатов производилось инструментом MarkDuplicates от

Picard [[65](#_bookmark95)], интегрированным в GATK. Оптимальные показатели скорости MarkDupl достигаются при запуске Java с параллелизацией сборщиков мусора и количе-

ством сборщиков мусора равным двум [[66](#_bookmark96)]. Также, согласно рекомендациям разработчиков, прочтения были предварительно отсортированы по именам, чтобы удалению подверглись не только первичные, но и добавочные вырав- нивания [[46](#_bookmark76)].

**Рекалибровка качества прочтений (BQSR).** Рекалибровка производилась с помощью инструментов GATK –– BaseRecalibrator и ApplyBQSR. Для обу- чения 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, предостав- ленную NCBI [[67](#_bookmark97)]. 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 [[66](#_bookmark96)] было показано, что оптимальные показатели скорости BaseRecalibrator достигаются, как и в случае с MarkDuplicates, запуском Java с двумя парал- лельными сборщиками мусора; кроме того, BaseRecalibrator поддаётся внеш- нему распараллеливанию путём разделения картированных прочтений на хро-

мосомные группы. Хромосомные группы формировались вручную для ис- пользуемой сборки генома, каждая запускалась с помощью bash-скрипта. Нам удалось усовершенствовать данный этап –– запуск BaseRecalibrator произво- дился с помощью библиотеки Python3 subprocess, а параллелизация осуществ- лялась библиотекой multiprocessing, таким образом, можно было делить файл с картированными прочтениями по хромосомам и обрабатывать их отдельно, так как multiprocessing автоматически распределяет процессы по имеющим-

ся потокам. Также для повышения отказоустойчивости скрипта у BaseRecalibrator и ApplyBQSR была устранена разница в фильтрации прочтений, из-за кото- рой при малых размерах библиотек пайплайн экстренно завершал работу.

**Оценка покрытия и обогащения.** Покрытие и обогащение в экзоме оце- нивались с помощью скрипта на основе bedtools [[68](#_bookmark98)].

**Поиск вариантов.** Поиск вариантов производился с помощью инструмен- та HaplotypeCaller от GATK. Инструмент запускался с дополнительным па- раметром --dont-use-soft-clipped-bases, который не позволял исполь-

зовать для поиска генетических вариантов клипированные химерные части и адаптеры.

Как и в случае с BaseRecalibrator, HaplotypeCaller поддаётся внешне- му распараллеливанию [[66](#_bookmark96)]. Мы также осуществили параллелизацию с помо- щью сочетания subprocess и multiprocessing, достигнув 10–12-кратного уско- рения по сравнению с запуском на одном потоке.

**Рекалибровка и ранжирование вариантов.** В GATK также присутству- ют инструменты для рекалибровки и ранжирования вариантов, с использо- ванием 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 ис- ключают необходимость фильтрации генетических вариантов. Таким обра- зом, от этого этапа решено было отказаться.

**Аннотация вариантов.** Аннотация вариантов производилась вначале с по-

мощью инструмента Ensembl VEP [[69](#_bookmark99)], затем мы мигрировали на ANNOVAR [[70](#_bookmark100)].

Используемые базы данных:

1. Human Gene Mutation Database (HGMD®) [[56](#_bookmark86)]
2. Online Mendelian Inheritance in Man (OMIM®) [[9](#_bookmark39)]
3. GeneCards®: The Human Gene Database [[71](#_bookmark101)]
4. CLINVAR [[40](#_bookmark70)]
5. dbSNP [[67](#_bookmark97)]
6. Genome Aggregation Database (gnomAD) [[52](#_bookmark82)]
7. 1000 Genomes Project [[72](#_bookmark102)]
8. Great Middle East allele frequencies (GME) [[53](#_bookmark83)]
9. dbNSFP: Exome Predictions [[73](#_bookmark103)]
10. dbscSNV: Splice site prediction [[74](#_bookmark104)]
11. RegSNPIntron: intronic SNVs prediction [[75](#_bookmark105)]

**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 несо- ответствий между ними;
  + Ранги инструментов, предсказывающих патогенность генетического ва- рианта. Трёхранговые системы (патогенный, вероятно патогенный и безвредный) были сведены к двухранговой (патогенный и безвредный). Отдельно были агрегированы предсказательные инструменты для эк- зонов, инструменты для интронов и сплайс-вариантов также учитыва- лись отдельно;
  + Ранги инструментов, предсказывающих консервативность нуклеотида. Эмпирическим путём было подобрано пороговое значение 0.7 –– нук- леотид считался консервативным, если его предсказанная консерватив- ность была выше, чем у 70% всех нуклеотидов. Это максимальное поро- говое 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 кон- кретному генетическому варианту была выбрана максимальная часто- та.

Фильтрация происходила в две стадии:

* 1. Фильтрация отдельных генетических вариантов на основе имеющихся аннотаций. Самая жёсткая фильтрация, которой подвергались все вари- анты:
     + По глубине покрытия. Генетический вариант считался существу- ющим, если он присутствовал в двух перекрывающихся парных прочтениях, либо в чётырёх независимых прочтениях;
     + Частота генетического варианта в популяции не более 3% [[57](#_bookmark87)].

Прочие фильтры были мягкими –– генетический вариант отсеивался толь- ко 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 [[57](#_bookmark87)];
    - 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.
  1. 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 [[76](#_bookmark106)];
     + 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 производи- лось в соответствии с рекомендациями Американского колледжа медицин- ской генетики и геномики (*англ.* American College of Medical Genetics, Bethesda, MD, USA) и Ассоциации молекулярной патологии [[77](#_bookmark107)].

# 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 поис- ку генетических вариантов;
  2. Проверка качества данных, полученных в результате массового парал- лельного секвенирования Exo-C-библиотек;
  3. Разработка, отладка и тестирование автоматизированного инструмента для обработки данных секвенирования Exo-C-библиотек.

## Результаты секвенирования Exo-C-библиотек

Несмотря на то, что составляющие протокола Exo-C –– таргетное обогаще- ние и Hi-C –– в настоящее время достаточно отработаны, сочетание этих ме- тодик имеет свои подводные камни. Было разработано две вариации прото- кола Exo-C (ExoC-19 и ExoC-20), обе этих вариации были использованы для приготовления библиотек клеточной линии K562 [[78](#_bookmark108), [79](#_bookmark109)] (Gridina et al. в пе- чати). Критическим различием протоколов является использование дополни- тельных адаптеров в протоколе ExoC-19. 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 про- чтений);
* Отношение среднего покрытия вне и внутри экзома, которое можно считать показателем качества таргетного обогащения.

Данные по качеству Exo-C-библиотек представлены в Табл. [1](#_bookmark19).

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 реги- онов 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 |

## Автоматизация обработки данных секвенирования

При обработке данных секвенирования приходится сталкиваться с пробле- мами различного характера. Одними из ключевых являются проблемы ис- пользования ресурсов компьютера. Результаты секвенирования даже в сжа- том виде занимают десятки и сотни гигабайт дискового пространства, и мно- гие 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 зна- чительное количество, и не все стадии нужны при обработке конкретного бло- ка данных секвенирования. Ручная настройка и контроль процесса отнимают значительное количество времени исследователя; таким образом, встаёт во- прос стандартизации и автоматизации процесса обработки данных секвени- рования.

Существующие инструменты для обработки данных секвенирования были разработаны независимыми группами людей. Эти инструменты разли- чаются по многим аспектам. Так как разработка каждого отдельного инстру- мента является сложным и трудоёмким процессом, целесообразно использо- вать их как есть, а несоответствия устранять с помощью специально разрабо- танной надстройки. Таким образом, для нами был создан пайплайн, интегри-

рующий все стадии обработки данных секвенирования. Блок-схема пайплай- на представлена на Рис. [3](#_bookmark21).



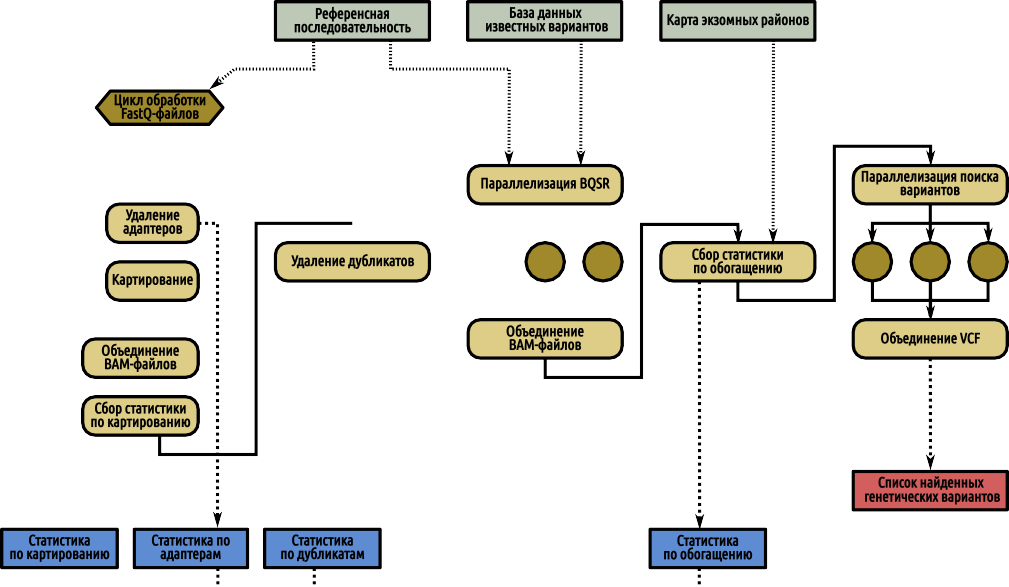
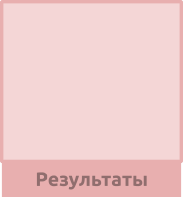


Рис. 3: Принципиальная схема пайплайна для обработки Exo-C-данных Решённые задачи:

* Отказоустойчивость: максимально устранены несоответствия форма-

тов входных и выходных данных; процесс разделён на стадии, и в слу- чае экстренного прерывания вычислений (программного или аппарат- ного) предусмотрен автоматический откат.

* Оптимизация, параллелизация и масштабируемость: все процессы, ко- торые способны использовать стандартные потоки ввода/вывода, объ- единены вместе, поддающиеся внешнему распараллеливанию были рас- параллелены, также были подобраны оптимальные параметры запуска приложений, использующих машину Java. Пайплайн может быть ис- пользован как на кластерах с большим количеством ядер и оператив- ной памяти, так и на относительно небольших мощностях офисных компьютеров;
* Значительно упрощены процессы развёртки и использования пайплай- на: автоматизировано индексирование референсной последовательно- сти, настройки вынесены в специальный конфигурационный файл, есть возможность обработки пула данных, используя один короткий сцена- рий;

Код пайплайна доступен на GitHub [[80](#_bookmark110)].

## Сравнение данных секвенирования клеточной линии K562

Следующим важным этапом работы была проверка эффективности поиска ге- нетических вариантов в Exo-C-библиотеках. Было решено использовать для этого распространённую иммортализованную клеточную линию K562, по- лученную от пациентки с хроническим миелолейкозом [[81](#_bookmark111)]. Данная клеточ- ная линия была многократно секвенирована различными лабораториями с ис- пользованием различных методик приготовления библиотек. Таким образом, несмотря на то, что в этой клеточной линии наблюдается некоторая гетероген- ность между лабораториями из-за большого количества пассажей, несмотря на наличие систематических ошибок при использовании разных методов се- квенирования 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 пуб- личных источников [[15](#_bookmark45), [82](#_bookmark112), [83](#_bookmark113), [84](#_bookmark114), [85](#_bookmark115), [86](#_bookmark116), [87](#_bookmark117), [88](#_bookmark118)]. Использованные в этих ста- тьях методики включают WGS, WES, Hi-C и Repli-seq. Из данных полноэкзо- много секвенирования в дальнейшем были исключены все генетические ва- рианты в интервале chr2:25455845-25565459 с фланкированием 1 тыс. п.о. (ген DNMT3A), так как в одной из работ использовали генетически модифи- цированную линию с вариантами в данном гене [[82](#_bookmark112)]. В качестве тестовых Exo-C-образцов мы использовали данные, полученные на основе клеточной линии K562, имеющейся в Институте Цитологии и Генетики СО РАН. Тех- нические данные контроля качества по тестовым и контрольным образцам представлены в Табл. [4](#_bookmark119) и Табл. [5](#_bookmark120).

В общей сложности, объединив варианты из всех контрольных образ- цов, мы получили 5 496 486 различных генетических вариантов. Также в биб- лиотеках было найдено некоторое количество уникальных генетических ва- риантов, встречающихся в одной библиотеке и не встречающихся в осталь- ных (Табл. [2](#_bookmark23)). 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. [[82](#_bookmark112)] | WES | 254 983 225 | 408 008 | 41 830 | 10,25 |
| Belaghzal et al. [[83](#_bookmark113)] | Hi-C | 72 914 268 | 1 399 457 | 27 365 | 1,95 |
| Dixon et al. [[84](#_bookmark114)] | WGS | 366 291 496 | 4 649 012 | 327 184 | 7,03 |
| Moquin et al. [[85](#_bookmark115)] | Hi-C | 256 500 659 | 2 365 361 | 67 678 | 2,86 |
| Rao et al. [[15](#_bookmark45)] | Hi-C | 1 366 228 845 | 4 218 233 | 320 508 | 7,59 |
| Ray et al. [[86](#_bookmark116)] | Hi-C | 428 306 794 | 1 789 324 | 89 624 | 5,00 |
| Wang et al. [[87](#_bookmark117)] | Repli-seq | 301 663 640 | 2 207 451 | 37 578 | 1,70 |
| Zhou et al. [[88](#_bookmark118)] | 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 вариан- тов, но не пересекаются с результатами других исследований;
    2. В одной или нескольких работах не найдено подавляющее большин- ство вариантов, найденных во всех остальных работах;
    3. Распределение уникальных вариантов и число общих вариантов между парами работ 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 вари- антов с особенностями какого-то одного из использованных наборов данных, мы протестировали все комбинации из семи и шести работ. Результаты пред- ставлены на Рис. [4](#_bookmark24).





Рис. 4: Исключение отдельных образцов из контрольной выборки позволило увеличить количество генетических вариантов, которые можно использовать как стандарт. На рисунке показано суммарное количество вариантов в выбор- ке (светло-зелёный) и процент общих для этой выборки вариантов (тёмно- зелёный) для выборок размером в 6 и 7 образцов. Слева указаны названия исключённых из выборки образцов.

При исключении из выборки данных Banaszak et al. и Belaghzal et al. общими являются 1 091 331 (19,85%) вариантов. Их решено было использо- вать как добавочный («серебряный») стандарт. Далее мы использовали ва- рианты «серебряного» и «золотого» стандартов для того, чтобы определить точность поиска генетических вариантов в наших Exo-C-библиотеках. Для этого мы оценили количество генетических вариантов, являющихся общими для «серебряного» и «золотого» стандартов и наших Exo-C-библиотек, их долю от общего числа вариантов в Exo-C-библиотеках, а также количество и долю ложноположительных (отсутствующих в контрольных образцах) гене- тических вариантов. Также было решено проверить эффективность исполь- зованного нами базового фильтра –– удаление всех генетических вариантов, в которых глубина альтернативного аллеля составляет менее 4. Поиск вари- антов «серебряного» и «золотого» стандартов в наших библиотеках был про- изведён до и после фильтрации. Результаты показаны в Табл. [3](#_bookmark25).

Таблица 3: Параметры Exo-C-библиотек. (F–) –– до фильтрации по глубине

альтернативного аллеля, (F+) –– после фильтрации, (∆) –– изменение пара- метра после фильтрации в процентах

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Параметр** | **ExoC-19** | | | **ExoC-20** | | | **В обеих** | | | **Ни в одной** | | |
| **F–** | **F+** | ∆**, %** | **F–** | **F+** | ∆**, %** | **F–** | **F+** | ∆**, %** | **F–** | **F+** | ∆**, %** |
| 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 |
| Доля вариантов «золотого стандарта», % | 82,75 | 70,00 | — | 96,52 | 89,30 | — | 80,62 | 64,84 | — | 1,35 | 5,53 | — |
| 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 «серебряного стандарта», % | 56,48 | 35,85 | — | 90,06 | 75,32 | — | 53,18 | 31,23 | — | 6,64 | 20,06 | — |
| Количество вариантов библиотеки, отсутствующих в контрольных образцах | 1 130 049 | 84 770 | –92,50 | 354 044 | 41 719 | –88,22 | 14 455 | 2 981 | –79,38 | — | — | — |
| Доля вариантов, отсутствующих в контроль- ных образцах, от вариантов библиотеки, % | 35,61 | 6,07 | –82,95 | 9,44 | 1,62 | –82,86 | — | — | — | — | — | — |

# Обсуждение результатов

## Контрольные образцы

«Золотой стандарт» с учётом подбора библиотек скорее всего является набо- ром генетических вариантов, относящихся к экзомным регионам, так как од- на из библиотек представляла собой результаты WES. Их было обнаружено

75 тыс., что соответствует оценкам среднего количества генетических вариан- тов в кодирующих регионах у человека –– 100 тыс. [[48](#_bookmark78)]. Общее число несоот- ветствий с референсным геномом у среднего человека составляет от 4,1 до 5 млн [[72](#_bookmark102)], 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. со- держит 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 использо- вались линии клеток, в значительной степени отличающиеся от классической линии K562.

Прослеживается ожидаемая положительная связь между глубиной се- квенирования 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.

## Оценка результатов секвенирования Exo-C-библиотек

В Exo-C-библиотеках глубина секвенирования составляет 136,6 млн прочте- ний (2*,* 05 *·* 1010 п.о.) и 109,4 млн прочтений (1*,* 64 *·* 1010 п.о.), а среднее покры- тие в экзоме –– 60,51 и 14,88 прочтений для ExoC-19 и ExoC-20 соответствен- но. Глубину покрытия более 10 прочтений имеют 91,68% и 72,58% экзома для ExoC-19 и ExoC-20 соответственно. Согласно [[59](#_bookmark89)], для репрезентативных ре- зультатов экзомного секвенирования необходима глубина секвенирования не менее чем в 1010 п.о., а для Hi-C –– не менее чем 100 млн прочтений. Мини- мальным порогом глубины для возможности поиска генетических вариантов считается 10 прочтений, практически все гомозиготные SNV могут быть най- дены при глубине в 15 прочтений, а гетерозиготные требуют глубину прочте- ний 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, при- годна 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 обога- щение 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 явля- ется 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 «зо- лотого» и «серебряного» стандартов.
    - Доля ложноположительных (отсутствующих в контрольных образцах) генетических вариантов снизилась в 5 раз.

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 яв- ляется эффективным способом улучшения данных низкого качества.

# Предварительные выводы

Таким образом, из приведённых нами данных можно сделать следующие вы- воды:

1. Пайплайн, созданный нами с учётом актуальных рекомендаций для био- информационной обработки, позволяет обрабатывать данные Exo-C-се- квенирования, а также находить в этих данных SNV.
2. Использование разработанного конвейера биоинформационных инстру- ментов позволило обнаружить около 5,5 млн генетических вариантов в

контрольных данных клеточной линии K562 (что сопоставимо со сред- ним количеством точечных полиморфизмов в геноме человека), из кото- рых наличие 75 тыс. подтвердилось в восьми независимых исследова- ниях, а 1 млн –– в шести независимых исследованиях, не включающих экзомные данные.

1. Сравнение генетических вариантов, полученных из контрольных образ- цов и Exo-C-библиотек, позволяет утверждать, что метод Exo-C спосо- бен детектировать около 75–90% SNV, обнаруживаемых другими мето- дами.

# План работы

В следующем семестре мы планируем:

1. Произвести анализ генетических вариантов в контрольных и наших об- разцах по следующим параметрам:
   1. тип;
   2. количество альтернативных аллелей;
   3. распределение в геноме (в том числе с учётом проблемных регио- нов);
   4. глубина покрытия;
   5. зиготность.
2. Произвести анализ данных Exo-C на предмет систематических ошибок поиска генетических вариантов.
3. Произвести анализ результатов секвенирования Exo-C-библиотек у ре- альных пациентов.

# Список литературы

1. Hiroi, N. & Agatsuma, S. Genetic susceptibility to substance dependence. *Mol Psychiatry* **10**, 336–344 (2004). URL [https://doi.org/10.1038/sj. mp.4001622](https://doi.org/10.1038/sj.mp.4001622).
2. Galanello, R. & Origa, R. Beta-thalassemia. *Orphanet J Rare Dis* **5** (2010). URL <https://doi.org/10.1186/1750-1172-5-11>.
3. Jett, K. & Friedman, J. M. Clinical and genetic aspects of neurofibromatosis 1. *Genet Med* **12**, 1–11 (2009). URL [https://doi.org/10.1097/gim. 0b013e3181bf15e3](https://doi.org/10.1097/gim.0b013e3181bf15e3).
4. Field, M. J. & Behrman, R. E. (eds.) *When Children Die: Improving Palliative and End-of-Life Care for Children and Their Families* (National Academies Press (US), Washington (DC), 2003). URL [https://pubmed.ncbi.nlm. nih.gov/25057608](https://pubmed.ncbi.nlm.nih.gov/25057608). NBK220818[bookaccession].
5. Herder, M. What is the purpose of the orphan drug act? *PLoS Med* **14**, e1002191 (2017). URL [https://doi.org/10.1371/journal.pmed. 1002191](https://doi.org/10.1371/journal.pmed.1002191).
6. Richter, T. *et al.* Rare disease terminology and definitions—a systematic global review: Report of the ISPOR Rare Disease Special Interest Group. *Value in Health* **18**, 906–914 (2015). URL [https://doi.org/10.1016/ j.jval.2015.05.008](https://doi.org/10.1016/j.jval.2015.05.008).
7. Bomar, J. M. *et al.* Mutations in a novel gene encoding a CRAL-TRIO domain cause human Cayman ataxia and ataxia/dystonia in the jittery mouse. *Nat Genet* **35**, 264–269 (2003). URL <https://doi.org/10.1038/ng1255>.
8. The Lancet Neurology. Rare neurological diseases: a united approach is needed. *The Lancet Neurology* **10**, 109 (2011). URL [https://doi.org/ 10.1016/S1474-4422(11)70001-1](https://doi.org/10.1016/S1474-4422(11)70001-1).
9. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh,

A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online

catalog of human genes and genetic disorders. *Nucleic Acids Research* **43**, D789–D798 (2014). URL <https://doi.org/10.1093/nar/gku1205>.

1. French National Institute of Health and Medical Research (INSERM). Orphanet: an online database of rare diseases and orphan drugs. Available online at: [http://www.orpha.net](http://www.orpha.net/) (1997). Accessed 2020/12/08.
2. Winklhofer, K. F., Tatzelt, J. & Haass, C. The two faces of protein misfolding: gain- and loss-of-function in neurodegenerative diseases. *EMBO J* **27**, 336– 349 (2008). URL <https://doi.org/10.1038/sj.emboj.7601930>.
3. Handy, D. E., Castro, R. & Loscalzo, J. Epigenetic modifications. *Circulation* **123**, 2145–2156 (2011). URL [https://doi.org/10.1161/ circulationaha.110.956839](https://doi.org/10.1161/circulationaha.110.956839).
4. Jin, Z. & Liu, Y. DNA methylation in human diseases. *Genes & Diseases* **5**, 1–8 (2018). URL <https://doi.org/10.1016/j.gendis.2018.01.002>.
5. Tang, B., Dean, B. & Thomas, E. A. Disease- and age-related changes in histone acetylation at gene promoters in psychiatric disorders. *Transl Psychiatry* **1**, e64–e64 (2011). URL [https://doi.org/10.1038/tp.2011. 61](https://doi.org/10.1038/tp.2011.61).
6. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). URL <https://doi.org/10.1016/j.cell.2014.11.021>.
7. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat Rev Genet* **19**, 453–467 (2018). URL [https://doi.org/ 10.1038/s41576-018-0007-0](https://doi.org/10.1038/s41576-018-0007-0).
8. Rizvi, A. A. 46, XX man with SRY gene translocation: Cytogenetic characteristics, clinical features and management. *The American Journal of the Medical Sciences* **335**, 307–309 (2008). URL [https://doi.org/10. 1097/maj.0b013e31811ec1b4](https://doi.org/10.1097/maj.0b013e31811ec1b4).
9. O’Connor, C. Human chromosome translocations and cancer. *Nature Education* **1**, 56 (2008). URL

[https://www.nature.com/scitable/topicpage/](https://www.nature.com/scitable/topicpage/human-chromosome-translocations-and-cancer-23487/)

[human-chromosome-translocations-and-cancer-23487/](https://www.nature.com/scitable/topicpage/human-chromosome-translocations-and-cancer-23487/).

1. Gonzalez, E. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005). URL <https://doi.org/10.1126/science.1101160>.
2. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007). URL [https://doi.org/10.1126/ science.1138659](https://doi.org/10.1126/science.1138659).
3. Dong, Y. *et al.* Impact of chromosomal translocations on male infertility, semen quality, testicular volume and reproductive hormone levels. *J Int Med Res* **40**, 2274–2283 (2012). URL [https://doi.org/10.1177/ 030006051204000625](https://doi.org/10.1177/030006051204000625).
4. Asim, A., Kumar, A., Muthuswamy, S., Jain, S. & Agarwal, S. Down syndrome: an insight of the disease. *J Biomed Sci* **22** (2015). URL [https:](https://doi.org/10.1186/s12929-015-0138-y)

[//doi.org/10.1186/s12929-015-0138-y](https://doi.org/10.1186/s12929-015-0138-y).

1. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of *de novo* mutations in health and disease. *Genome Biol* **17** (2016). URL <https://doi.org/10.1186/s13059-016-1110-1>.
2. Brea-Fernandez, A. *et al.* An update of *in silico* tools for the prediction of pathogenesis in missense variants. *CBIO* **6**, 185–198 (2011). URL [https:](https://doi.org/10.2174/1574893611106020185)

[//doi.org/10.2174/1574893611106020185](https://doi.org/10.2174/1574893611106020185).

1. Kozak, M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl Acids Res* **15**, 8125–8148 (1987). URL [https:](https://doi.org/10.1093/nar/15.20.8125)

[//doi.org/10.1093/nar/15.20.8125](https://doi.org/10.1093/nar/15.20.8125).

1. Young, S. K. & Wek, R. C. Upstream open reading frames differentially regulate gene-specific translation in the integrated stress response. *Journal of Biological Chemistry* **291**, 16927–16935 (2016). URL [https://doi.org/ 10.1074/jbc.r116.733899](https://doi.org/10.1074/jbc.r116.733899).
2. Abramowicz, A. & Gos, M. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genetics* **59**, 253–268 (2018). URL <https://doi.org/10.1007/s13353-018-0444-7>.
3. Schreck, R. R. & Distèche, C. M. Chromosome banding techniques. *Current protocols in human genetics* **Chapter 4**, Unit4.2–Unit4.2 (2001). URL <https://pubmed.ncbi.nlm.nih.gov/18428280>. 18428280[pmid].
4. Sampson, B. & McGuire, A. Genetics and the molecular autopsy. In

*Pathobiology of Human Disease*, 3459–3467 (Elsevier, 2014). URL [https:](https://doi.org/10.1016/b978-0-12-386456-7.06707-1)

[//doi.org/10.1016/b978-0-12-386456-7.06707-1](https://doi.org/10.1016/b978-0-12-386456-7.06707-1).

1. Guo, B., Han, X., Wu, Z., Da, W. & Zhu, H. Spectral karyotyping: an unique technique for the detection of complex genomic rearrangements in leukemia. *Translational pediatrics* **3**, 135–139 (2014). URL [https://pubmed.ncbi. nlm.nih.gov/26835331](https://pubmed.ncbi.nlm.nih.gov/26835331). 26835331[pmid].
2. Huber, D., von Voithenberg, L. V. & Kaigala, G. Fluorescence *in situ* hybridization (FISH): History, limitations and what to expect from micro- scale FISH? *Micro and Nano Engineering* **1**, 15–24 (2018). URL [https:](https://doi.org/10.1016/j.mne.2018.10.006)

[//doi.org/10.1016/j.mne.2018.10.006](https://doi.org/10.1016/j.mne.2018.10.006).

1. Theisen, A. Microarray-based Comparative Genomic Hybridization (aCGH). *Nature Education* **1**, 45 (2008). URL [https://www.nature.com/scitable/topicpage/ microarray-based-comparative-genomic-hybridization-acgh-45432/](https://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432/).
2. Gresham, D., Dunham, M. J. & Botstein, D. Comparing whole genomes using DNA microarrays. *Nat Rev Genet* **9**, 291–302 (2008). URL [https:](https://doi.org/10.1038/nrg2335)

[//doi.org/10.1038/nrg2335](https://doi.org/10.1038/nrg2335).

1. Stuppia, L., Antonucci, I., Palka, G. & Gatta, V. Use of the MLPA assay in the molecular diagnosis of gene copy number alterations in human genetic diseases. *IJMS* **13**, 3245–3276 (2012). URL [https://doi.org/10.3390/ ijms13033245](https://doi.org/10.3390/ijms13033245).
2. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain- terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977). URL <https://doi.org/10.1073/pnas.74.12.5463>.
3. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333– 351 (2016). URL <https://doi.org/10.1038/nrg.2016.49>.
4. Yu, D. *et al.* Noninvasive prenatal testing for fetal subchromosomal copy number variations and chromosomal aneuploidy by low-pass whole-genome sequencing. *Mol Genet Genomic Med* **7** (2019). URL [https://doi.org/ 10.1002/mgg3.674](https://doi.org/10.1002/mgg3.674).
5. Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Human Molecular Genetics* **19**, R145–R151 (2010). URL [https:](https://doi.org/10.1093/hmg/ddq333)

[//doi.org/10.1093/hmg/ddq333](https://doi.org/10.1093/hmg/ddq333).

1. Yohe, S. & Thyagarajan, B. Review of clinical next-generation sequencing. *Archives of Pathology & Laboratory Medicine* **141**, 1544–1557 (2017). URL <https://doi.org/10.5858/arpa.2016-0501-ra>.
2. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research* **46**, D1062–D1067 (2017). URL <https://doi.org/10.1093/nar/gkx1153>.
3. Barbitoff, Y. A. *et al.* Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci Rep* **10** (2020). URL [https://doi.org/10.1038/ s41598-020-59026-y](https://doi.org/10.1038/s41598-020-59026-y).
4. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009). URL <https://doi.org/10.1126/science.1181369>.
5. Oluwadare, O., Highsmith, M. & Cheng, J. An overview of methods for reconstructing 3D chromosome and genome structures from Hi-C

data. *Biol Proced Online* **21** (2019). URL [https://doi.org/10.1186/](https://doi.org/10.1186/s12575-019-0094-0) [s12575-019-0094-0](https://doi.org/10.1186/s12575-019-0094-0).

1. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011). URL [https://doi.org/10. 14806/ej.17.1.200](https://doi.org/10.14806/ej.17.1.200).
2. Burrows, M. & Wheeler, D. A block-sorting lossless data compression algorithm. Tech. Rep., Palo Alto, CA: Digital Equipment Corporation (1994).
3. Auwera, G. A. *et al.* From FastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics* **43** (2013). URL [https://doi.org/10.1002/0471250953. bi1110s43](https://doi.org/10.1002/0471250953.bi1110s43).
4. Ebbert, M. T. W. *et al.* Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics* **17** (2016). URL [https://doi.org/10.1186/ s12859-016-1097-3](https://doi.org/10.1186/s12859-016-1097-3).
5. Supernat, A., Vidarsson, O. V., Steen, V. M. & Stokowy, T. Comparison of three variant callers for human whole genome sequencing. *Sci Rep* **8** (2018). URL <https://doi.org/10.1038/s41598-018-36177-7>.
6. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine* **6**, 26 (2014). URL [https://doi. org/10.1186/gm543](https://doi.org/10.1186/gm543).
7. Jesaitis, A. The state of variant annotation: A comparison of AnnoVar, snpEff and VEP. Tech. Rep., The Golden Helix Blog (GHB), [https://blog.goldenhelix.com/](https://blog.goldenhelix.com/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/)

[the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-an](https://blog.goldenhelix.com/the-sate-of-variant-annotation-a-comparison-of-annovar-snpeff-and-vep/)

(2014).

1. Ziegler, A., Colin, E., Goudenège, D. & Bonneau, D. A snapshot of some pLI score pitfalls. *Human Mutation* (2019). URL [https://doi.org/10. 1002/humu.23763](https://doi.org/10.1002/humu.23763).
2. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). URL [https:](https://doi.org/10.1038/s41586-020-2308-7)

[//doi.org/10.1038/s41586-020-2308-7](https://doi.org/10.1038/s41586-020-2308-7).

1. Scott, E. M. *et al.* Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet* **48**, 1071–1076 (2016). URL <https://doi.org/10.1038/ng.3592>.
2. Naslavsky, M. S. *et al.* Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Human Mutation* **38**, 751–763 (2017). URL [https:](https://doi.org/10.1002/humu.23220)

[//doi.org/10.1002/humu.23220](https://doi.org/10.1002/humu.23220).

1. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017). URL [https:](https://doi.org/10.1038/nature22034)

[//doi.org/10.1038/nature22034](https://doi.org/10.1038/nature22034).

1. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* **136**, 665–677 (2017). URL <https://doi.org/10.1007/s00439-017-1779-6>.
2. Ryzhkova, O. *et al.* Guidelines for the interpretation of massive parallel sequencing variants. *Medical Genetics* **16**, 4–17 (2017). URL [https:](https://www.medgen-journal.ru/jour/article/view/308/224)

[//www.medgen-journal.ru/jour/article/view/308/224](https://www.medgen-journal.ru/jour/article/view/308/224).

1. Melo, U. S. *et al.* Hi-C identifies complex genomic rearrangements and TAD-shuffling in developmental diseases. *The American Journal of Human Genetics* **106**, 872–884 (2020). URL [https://doi.org/10.1016/j.ajhg. 2020.04.016](https://doi.org/10.1016/j.ajhg.2020.04.016).
2. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet* **15**, 121–132 (2014). URL <https://doi.org/10.1038/nrg3642>.
3. Mozheiko, E. A. & Fishman, V. S. Detection of point mutations and chromosomal translocations based on massive parallel sequencing of

enriched 3C libraries. *Russ J Genet* **55**, 1273–1281 (2019). URL [https:](https://doi.org/10.1134/s1022795419100089)

[//doi.org/10.1134/s1022795419100089](https://doi.org/10.1134/s1022795419100089).

1. Andrews, S. FastQC: A quality control tool for high throughput sequence data. Available online at: [http://www.bioinformatics.babraham.ac. uk/projects/fastqc/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (2010). Accessed 2020/12/08.
2. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie2. *Nat Methods* **9**, 357–359 (2012). URL [https://doi.org/10.1038/nmeth. 1923](https://doi.org/10.1038/nmeth.1923).
3. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows– Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). URL [https:// doi.org/10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324).
4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). URL [https://doi.org/10.1093/ bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
5. Broad Institute. Picard Toolkit. GitHub repository: [http:](http://broadinstitute.github.io/picard/)

[//broadinstitute.github.io/picard/](http://broadinstitute.github.io/picard/) (2019). Accessed 2021/01/10.

1. Heldenbrand, J. R. *et al.* Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinformatics* **20** (2019). URL <https://doi.org/10.1186/s12859-019-3169-7>.
2. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**, 308–311 (2001). URL [https://doi.org/10.1093/nar/29. 1.308](https://doi.org/10.1093/nar/29.1.308).
3. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). URL <https://doi.org/10.1093/bioinformatics/btq033>.
4. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**

(2016). URL <https://doi.org/10.1186/s13059-016-0974-4>.

1. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* **38**, e164–e164 (2010). URL [https://doi.org/10.1093/nar/ gkq603](https://doi.org/10.1093/nar/gkq603).
2. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics* **54** (2016). URL <https://doi.org/10.1002/cpbi.5>.
3. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL [https://doi.org/10. 1038%2Fnature15393](https://doi.org/10.1038%2Fnature15393).
4. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Human Mutation* **37**, 235–241 (2016). URL [https:](https://doi.org/10.1002/humu.22932)

[//doi.org/10.1002/humu.22932](https://doi.org/10.1002/humu.22932).

1. Jian, X., Boerwinkle, E. & Liu, X. *In silico* tools for splicing defect prediction: a survey from the viewpoint of end users. *Genet Med* **16**, 497–503 (2013). URL <https://doi.org/10.1038/gim.2013.176>.
2. Lin, H. *et al.* RegSNPs-intron: a computational framework for predicting pathogenic impact of intronic single nucleotide variants. *Genome Biol* **20** (2019). URL <https://doi.org/10.1186/s13059-019-1847-4>.
3. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016). URL [https://doi.org/10.1038/ nature19057](https://doi.org/10.1038/nature19057).
4. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–423 (2015). URL [https://doi.org/10. 1038/gim.2015.30](https://doi.org/10.1038/gim.2015.30).
5. Ma, W. *et al.* Using DNase Hi-C techniques to map global and local three- dimensional genome architecture at high resolution. *Methods* **142**, 59–73 (2018). URL <https://doi.org/10.1016/j.ymeth.2018.01.014>.
6. Ramani, V. *et al.* Mapping 3D genome architecture through in situ DNase Hi- C. *Nat Protoc* **11**, 2104–2121 (2016). URL [https://doi.org/10.1038/ nprot.2016.126](https://doi.org/10.1038/nprot.2016.126).
7. Valeev, E. Scissors: Exo-C variants search pipeline. GitHub repository: [https://github.com/regnveig/labjournal/tree/master/tools/ Scissors](https://github.com/regnveig/labjournal/tree/master/tools/Scissors) (2020).
8. Lozzio, C. B. & Lozzio, B. B. Human chronic myelogenous leukemia cell- line with positive Philadelphia chromosome. *Blood* **45**, 321–334 (1975). URL <https://pubmed.ncbi.nlm.nih.gov/163658>.
9. Banaszak, L. G. *et al.* Abnormal RNA splicing and genomic instability after induction of DNMT3A mutations by CRISPR/Cas9 gene editing. *Blood Cells, Molecules, and Diseases* **69**, 10–22 (2018). URL [https://doi.org/ 10.1016/j.bcmd.2017.12.002](https://doi.org/10.1016/j.bcmd.2017.12.002).
10. Belaghzal, H., Dekker, J. & Gibcus, J. H. Hi-C 2.0: An optimized Hi- C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods* **123**, 56–65 (2017). URL [https://doi.org/10. 1016/j.ymeth.2017.04.004](https://doi.org/10.1016/j.ymeth.2017.04.004).
11. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet* **50**, 1388–1398 (2018). URL [https://doi. org/10.1038/s41588-018-0195-8](https://doi.org/10.1038/s41588-018-0195-8).
12. Moquin, S. A. *et al.* The Epstein–Barr virus episome maneuvers between nuclear chromatin compartments during reactivation. *J Virol* **92** (2017). URL <https://doi.org/10.1128/jvi.01413-17>.
13. Ray, J. *et al.* Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proc Natl Acad Sci*

*USA* **116**, 19431–19439 (2019). URL [https://doi.org/10.1073/pnas.](https://doi.org/10.1073/pnas.1901244116)

[1901244116](https://doi.org/10.1073/pnas.1901244116).

1. Wang, Y. *et al.* SPIN reveals genome-wide landscape of nuclear compartmentalization (2020). URL [https://doi.org/10.1101/2020.03. 09.982967](https://doi.org/10.1101/2020.03.09.982967).
2. Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.* **29**, 472–484 (2019). URL <https://doi.org/10.1101/gr.234948.118>.

61

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 | WES1 | 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 | WES1 | 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 | WES1 | 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 | WES1 | 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 | WES1 | 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 | WES1 | 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.

62

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 |