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

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

УДК: 616-074+57.087.1

# Разработка инструментов для поиска клинически значимых полиморфизмов в геноме человека на основе данных секвенирования 3C-библиотек

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

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

***Аннотация.*** Здесь приводится краткое содержание статьи. Обязательно должны быть сформулированы основные результаты работы. Текст аннотации должен быть самодостаточным, без ссылок на список литературы, с понятными обозначениями, без аббревиатур. Обращаем внимание русскоязычных авторов на необходимость качественного перевода описания статьи на английский язык, поскольку весь зарубежный научный мир будет иметь представление о работе именно по этому описанию. Аннотация на английском языке не должна быть точным переводом русскоязычной аннотации. Она призвана отражать содержание статьи, которая недоступна для читателей, не владеющих русским языком. Длина аннотации должна быть 200–250 слов (не менее 25 строк).

***Ключевые слова:*** *Exo-C, ...*

## ВВЕДЕНИЕ

Наследственные заболевания являются одной из основных причин младенческой и детской смертности в развитых странах[1]. Взрослые люди с такими патологиями требуют огромных затрат средств на медикаменты, оперативные вмешательства, специальный уход и социальные льготы. Таким образом, доступные и точные методы диагностики наследственных заболеваний могут помочь в сокращении заболеваемости и смертности, а также повысить экономическое благополучие населения.

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

Наиболее перспективными в диагностическом и исследовательском плане в настоящее время являются методы секвенирования — например, полногеномное и полноэкзомное секвенирование. В Секторе геномных механизмов онтогенеза ИЦиГ СО РАН был разработан новейший метод секвенирования — Exo-C, сочетающий технологии экзомного обогащения с захватом конформации хромосом. Потенциальным преимуществом данного метода может быть возможность поиска как крупных перестроек, так и точечных полиморфизмов в экзоме при относительно небольшой

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

глубине секвенирования, от которой напрямую зависит цена секвенирования. Широкий спектр применения метода и доступность в финансовом аспекте делают метод Exo-C привлекательным как для медико-биологических научных исследований, так и для внедрения в клиническую практику.

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

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

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

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

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

## МАТЕРИАЛЫ И МЕТОДЫ

### 1. Данные секвенирования

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

### 2. Контроль качества NGS-данных

Для контроля качества прочтений мы использовали утилиту FastQC [2], способную оценивать наличие адаптерных последовательностей, распределение прочтений по длине, GC-состав прочтений, а также производить анализ зависимости нуклеотидного состава от позиции в прочтении. Критерии качества были использованы согласно протоколу разработчика [2].

### 3. Удаление адаптерных последовательностей

Удаление адаптерных последовательностей производилось с помощью утилиты cutadapt [3]. В [4] рекомендуется использовать в качестве входных данных некартированный BAM-файл (англ. *Unmapped Binary sequence AlignmentMap, uBAM*), а для удаления адаптеров использовать их собственный инструмент — MarkIlluminaAdapters, так как это позволяет сохранить важные метаданные. Тем не менее, был сделан акцент на том, что uBAM должен использоваться как выходной формат на уровне секвенатора, что не является общепринятой практикой.

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

## 4. Картирование

Картирование производилось с помощью инструментов Bowtie2 [5] и BWA [6]. BWA показал лучшие результаты; кроме того, он значительно более эффективно работает с химерными ридами, что немаловажно для используемого нами метода Exo-C.

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

Кроме того, для правильного функционирования инструментов на дальнейших этапах был разработан скрипт, создающий метку группы прочтений (англ. *Read Group tag, RG*) для каждого файла. Конкретных рекомендаций по составлению RG не существует, поэтому мы разработали собственные, основанные на требованиях Broad Institute [4].

Объединение BAM-файлов производилось инструментом MergeSamFiles. Сбор статистики по картированию мы осуществляли с помощью инструмента SAMTools flagstat [7].

## 5. Удаление ПЦР-дубликатов

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

Удаление дубликатов производилось инструментом MarkDuplicates от Picard [8], интегрированным в GATK. Оптимальные показатели скорости MarkDuplicates достигаются при запуске Java с параллелизацией сборщиков мусора и количеством сборщиков мусора равным двум [9]. Также, согласно рекомендациям разработчиков, прочтения были предварительно отсортированы по именам, чтобы удалению подверглись не только первичные, но и добавочные выравнивания [4].

## 6. Рекалибровка качества прочтений (BQSR)

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

К сожалению, предоставленная Broad Institute база данных оказалась сильно устаревшей и не вполне подходила для сделанной нами геномной сборки, поэтому было решено подвергнуть обработке dbSNP v150, предоставленную NCBI [10]. 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 GATK IndexFeatureFile (этот же инструмент одновременно проверяет БД на пригодность для BQSR).

В [9] было показано, что оптимальные показатели скорости BaseRecalibrator достигаются, как и в случае с MarkDuplicates, запуском Java с двумя параллельными сборщиками мусора; кроме того, BaseRecalibrator поддаётся внешнему распараллеливанию путём разделения картированных прочтений на хромосомные группы. Хромосомные группы формировались вручную для используемой сборки генома, каждая запускалась с помощью bash-скрипта. Нам удалось усовершенствовать данный этап — запуск BaseRecalibrator производился с помощью библиотеки Python subprocess, а параллелизация осуществлялась библиотекой multiprocessing, таким образом, можно было делить файл с картированными прочтениями по хромосомам

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

## 7. Оценка покрытия и обогащения

Покрытие и обогащение в экзоме оценивались с помощью скрипта на основе BEDTools [11].

## 8. Поиск вариантов

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

Как и в случае с BaseRecalibrator, HaplotypeCaller поддаётся внешнему распараллеливанию [9]. Мы также осуществили параллелизацию с помощью сочетания subprocess и multiprocessing, достигнув 10–12-кратного ускорения по сравнению с запуском на одном потоке.

## 9. Аннотация вариантов

Аннотация вариантов производилась вначале с помощью инструмента ANNOVAR [12].

Используемые базы данных:

1. Human Gene Mutation Database (HGMD®) [13]
2. Online Mendelian Inheritance in Man (OMIM®) [14]
3. GeneCards®: The Human Gene Database [15]
4. CLINVAR [16]
5. dbSNP [10]
6. Genome Aggregation Database (gnomAD) [17]
7. 1000Genomes Project [18]
8. Great Middle East allele frequencies (GME) [19]
9. dbNSFP: Exome Predictions [20]
10. dbscSNV: Splice site prediction [21]
11. RegSNPIntron: intronic SNVs prediction [22]

## 10. Фильтрация генетических вариантов

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

- Имена генов по разным БД — для облегчения поиска;
- Описания функциональных классов из разных БД — для устранения несоответствий между ними;

- Ранги инструментов, предсказывающих патогенность генетического варианта. Трёхранговые системы (патогенный, вероятно патогенный и безвредный) были сведены к двухранговой (патогенный и безвредный). Отдельно были агрегированы предсказательные инструменты для экзонов, инструменты для интронов и сплайс-вариантов также учитывались отдельно;

- Ранги инструментов, предсказывающих консервативность нуклеотида. Эмпирическим путём было подобрано пороговое значение 0.7 — нуклеотид считался консервативным, если его предсказанная консервативность была выше, чем у 70 % всех нуклеотидов. Это максимальное пороговое значение, которое обеспечивает распределение балла агрегатора от минимального до максимального (от 0 до 7 баз данных, считающих данный нуклеотид консервативным);

- Популяционные частоты — из всех имеющихся в базах данных по конкретному генетическому варианту была выбрана максимальная частота.

Фильтрация происходила в две стадии:

1. Фильтрация отдельных генетических вариантов на основе имеющихся аннотаций. Самая жёсткая фильтрация, которой подвергались все варианты:

    - По глубине покрытия. Генетический вариант считался существующим, если он присутствовал в двух перекрывающихся парных прочтениях, либо в чётырёх независимых прочтениях;

    - Частота генетического варианта в популяции не более 3 % [23].

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

    - Присутствие описания связанной с геном патологии в базе данных OMIM;

    - Присутствие генетического варианта в базе данных HGMD;

    - Балл агрегатора патогенности экзомных вариантов не менее 3 [23];

    - Ранг «патогенный» у агрегаторов интронных или сплайс-вариантов;

    - Ранги «патогенный» и «возможно патогенный» по базе данных CLINVAR;

    - По функциональному классу: сдвиги рамки считывания, потери стоп- и старт-кодонов, нонсенс- и сплайс-варианты.

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

    - Значение pLI более 0.9, согласно рекомендациям в оригинальной статье [24];

    - 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).

### 11. Интерпретация

Интерпретация данных и составление отчёта производилось в соответствии с рекомендациями Американского колледжа медицинской генетики и геномики (англ. *American College of Medical Genetics, Bethesda, MD, USA*) и Ассоциации молекулярной патологии [25].

## РЕЗУЛЬТАТЫ

### 12. Результаты секвенирования Exo-C-библиотек

Несмотря на то, что составляющие протокола Exo-C — таргетное обогащение и Hi-C — в настоящее время достаточно отработаны, сочетание этих методик имеет свои подводные камни. Было разработано две вариации протокола Exo-C (ExoC-19 и ExoC-20), обе этих вариации были использованы для приготовления библиотек клеточной линии K562 [26, 27, 28]. Критическим различием протоколов является использование дополнительных 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 прочтений);
- Отношение среднего покрытия вне и внутри экзома, которое можно считать показателем качества таргетного обогащения.

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

**Таблица 1.** Данные по обогащению Exo-C-библиотек

| **Название** | Глубина, прочтений | **Доля дубликатов, %** | Доля экзома с глубиной покрытия более 10, % | Среднее покрытие в экзоме | Среднее покрытие вне экзома | Обогащение экзома, раз | Доля непокрытых регионов в экзоме, % | Доля непокрытых регионов вне экзома, % |
|---|---|---|---|---|---|---|---|---|
| ExoC-19 | 136 609 179 | 18.86 | 91.68 | 60.51 | 5.56 | 10.89 | 1.75 | 28.12 |
| ExoC-20 | 109 486 529 | 15.00 | 72.58 | 14.88 | 7.74 | 1.92 | 1.66 | 11.62 |

### 13. Автоматизация обработки данных секвенирования

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

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

Существующие инструменты для обработки данных секвенирования были разработаны независимыми группами людей. Эти инструменты различаются по многим

аспектам. Так как разработка каждого отдельного инструмента является сложным и трудоёмким процессом, целесообразно использовать их как есть, а несоответствия устранять с помощью специально разработанной надстройки. Таким образом, для нами был создан пайплайн, интегрирующий все стадии обработки данных секвенирования. Блок-схема пайплайна представлена на рис. 1.

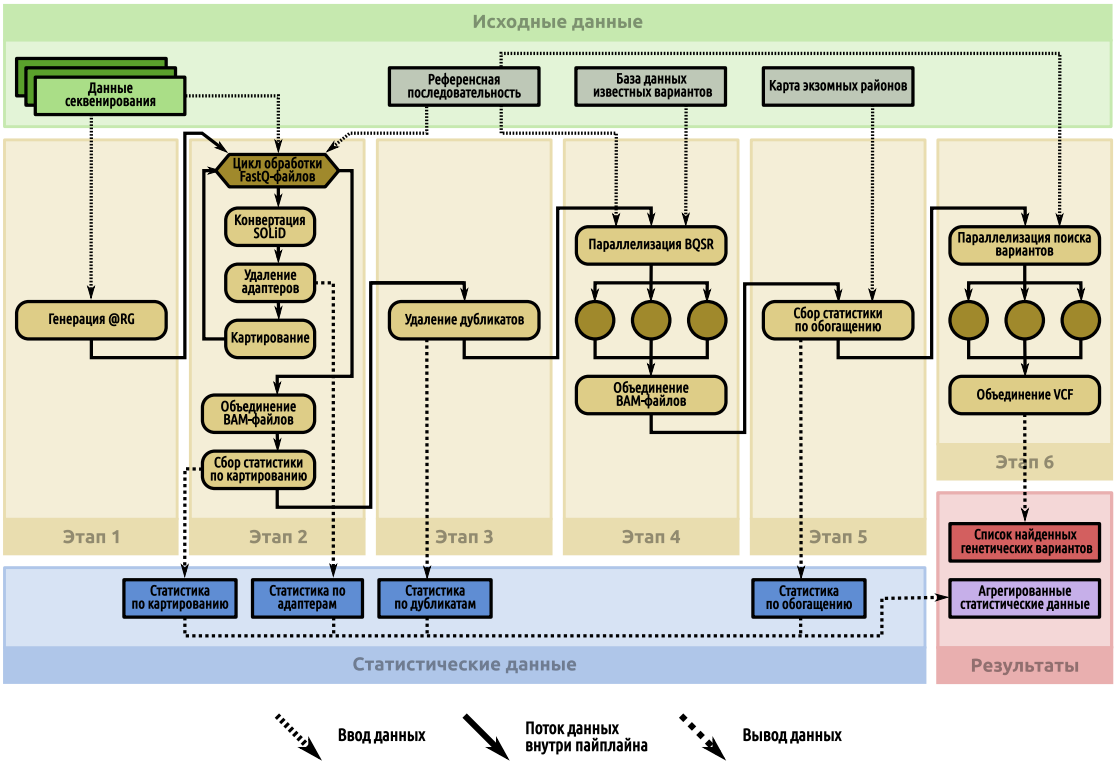


**Рис. 1.** Принципиальная схема пайплайна для обработки Exo-C-данных

Решённые задачи:

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

Код пайплайна доступен на GitHub [29].

## 14. Сравнение данных секвенирования клеточной линии K562

Следующим важным этапом работы была проверка эффективности поиска генетических вариантов в Exo-C-библиотеках. Было решено использовать для этого распространённую иммортализованную клеточную линию K562, полученную от пациентки с хроническим миелолейкозом [30]. Данная клеточная линия была многократно секвенирована различными лабораториями с использованием различных методик приготовления библиотек. Таким образом, несмотря на то, что в этой клеточной линии наблюдается некоторая гетерогенность между лабораториями из-за большого количества пассажей, несмотря на наличие систематических ошибок при использовании разных методов секвенирования и приготовления библиотек, по K562 существует достаточное количество данных, чтобы использовать эту клеточную линию как стандарт для поиска генетических вариантов.

Результаты секвенирования клеточной линии K562 были взяты из публичных источников [31, 32, 33, 34, 35, 36, 37, 38]. Использованные в этих статьях методики включают WGS, WES, Hi-C и Repli-seq. Из данных полноэкзомного секвенирования в дальнейшем были исключены все генетические варианты в интервале chr2:25455845-25565459 с фланкированием 1 kbp (ген *DNMT3A*), так как в одной из работ использовали генетически модифицированную линию с вариантами в данном гене [31]. В качестве тестовых Exo-C-образцов мы использовали данные, полученные на основе клеточной линии K562, имеющейся в Институте Цитологии и Генетики СО РАН. Технические данные контроля качества по тестовым и контрольным образцам представлены в табл. 5 и табл. 6.

В общей сложности, объединив варианты из всех контрольных образцов, мы получили 5 496 486 различных генетических вариантов. Также в библиотеках было найдено некоторое количество уникальных генетических вариантов, встречающихся в одной библиотеке и не встречающихся в остальных (табл. 2). Наибольший процент уникальных вариантов найден в данных Banaszak et al. [31]

**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. [31] | WES | 254 983 225 | 408 008 | 41 830 | 10.25 |
| Belaghzal et al. [32] | Hi-C | 72 914 268 | 1 399 457 | 27 365 | 1.95 |
| Dixon et al. [33] | WGS | 366 291 496 | 4 649 012 | 327 184 | 7.03 |
| Moquin et al. [34] | Hi-C | 256 500 659 | 2 365 361 | 67 678 | 2.86 |
| Rao et al. [35] | Hi-C | 1 366 228 845 | 4 218 233 | 320 508 | 7.59 |
| Ray et al. [36] | Hi-C | 428 306 794 | 1 789 324 | 89 624 | 5.00 |
| Wang et al. [37] | Repli-seq | 301 663 640 | 2 207 451 | 37 578 | 1.70 |
| Zhou et al. [38] | WGS | 2 621 311 293 | 4 412 455 | 166 451 | 3.77 |

75 328 генетических вариантов были найдены в данных из всех восьми статей — их было решено использовать как «золотой стандарт». Сразу можно внимание на то, что это составляет лишь 1.37 % геномных SNV клеток K562. Такая ситуация может возникнуть в следующих случаях:

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

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

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

Чтобы проверить, не связана ли низкая доля общих генетических вариантов с особенностями какого-то одного из использованных наборов данных, мы протестировали все комбинации из семи и шести работ. Результаты представлены на рис. 2.

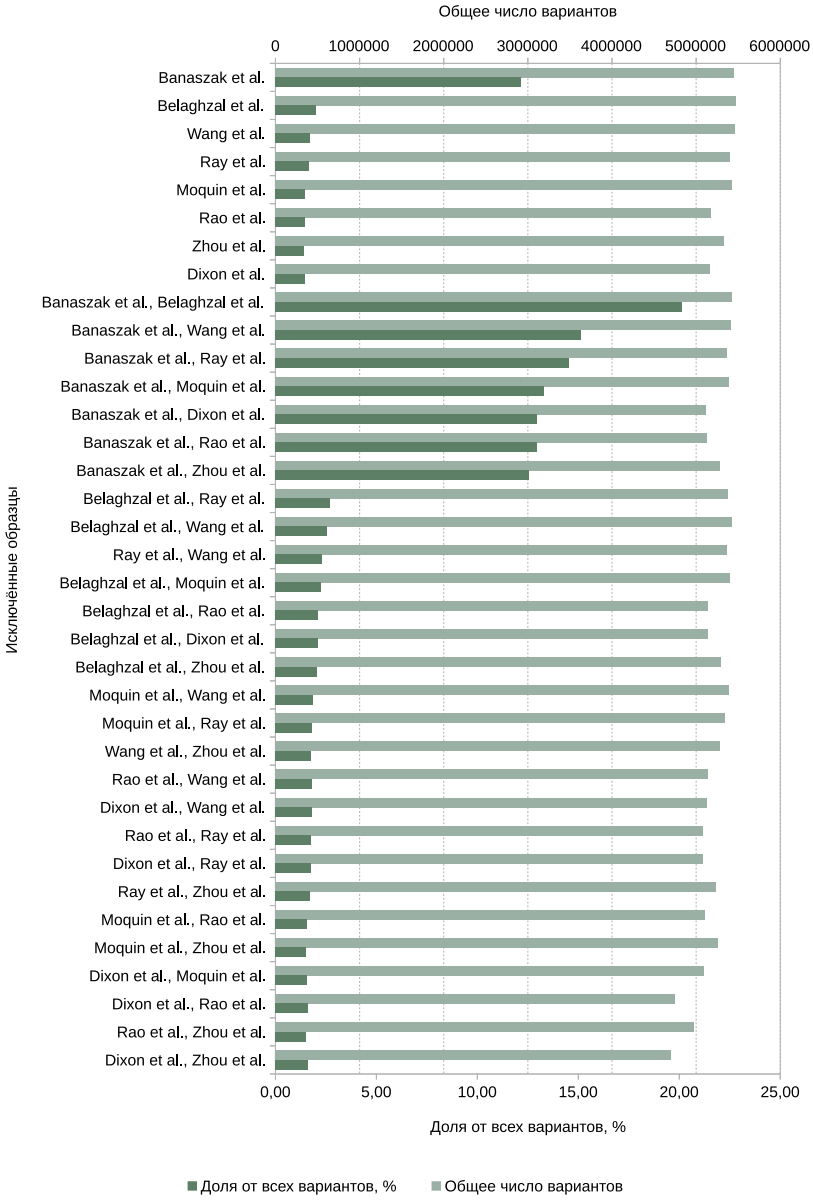


**Рис. 2.** Исключение отдельных образцов из контрольной выборки позволило увеличить количество генетических вариантов, которые можно использовать как стандарт. На рисунке показано суммарное количество вариантов в выборке (светло-зелёный) и процент общих для этой выборки вариантов (тёмно-зелёный) для выборок размером в 6 и 7 образцов. Слева указаны названия исключённых из выборки образцов.

При исключении из выборки данных Banaszak et al. [31] и Belaghzal et al. [32] общими являются 1 091 331 (19.85 %) вариантов. Их решено было использовать как добавочный («серебряный») стандарт. Далее мы использовали варианты «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) генетических вариантов. Также было решено проверить эффективность использованного нами базового фильтра — удаление всех генетических вариантов, в которых глубина альтернативного аллеля составляет менее 4. Поиск вариантов «серебряного» и «золотого» стандартов в наших библиотеках был произведён до и после фильтрации. Результаты показаны в табл. 3.

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

| Параметр | ExoC-19 | | | ExoC-20 | | | В обеих | | | Ни в одной | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F– | F+ | Δ, % | F– | F+ | Δ, % | F– | F+ | Δ, % | F– | F+ | Δ, % |
| Общее число вариантов в библиотеке | 3 173 343 | 1 396 525 | −55.99 | 3 750 319 | 2 577 934 | −31.26 | — | — | — | — | — | — |
| Вариантов «золотого стандарта» | 62 335 | 52 732 | −15.41 | 72 705 | 67 270 | −7.48 | 60 728 | 48 840 | −19.58 | 1 016 | 4 166 | +310.04 |
| 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 контрольных образцах, от вариантов библиотеки, % | 35.61 | 6.07 | −82.95 | 9.44 | 1.62 | −82.86 | — | — | — | — | — | — |

## ОБСУЖДЕНИЕ

### 15. Контрольные образцы

«Золотой стандарт» с учётом подбора библиотек скорее всего является набором генетических вариантов, относящихся к экзомным регионам, так как одна из библиотек представляла собой результаты WES. Их было обнаружено 75 тыс., что соответствует оценкам среднего количества генетических вариантов в кодирующих регионах у человека — 100 тыс. [39]. Общее число несоответствий с референсным геномом у среднего человека составляет 4.1–5 млн [18], 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. [31] содержит наибольшее число уникальных вариантов (10.25 %). Это может быть связано с тем, что это данные полноэкзомного секвенирования, с высоким покрытием в экзонах, где и были найдены уникальные варианты. В качестве дополнительной гипотезы можно предположить, что в этой работе использовались линии клеток, в значительной степени отличающиеся от классической линии K562.

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

### 16. Оценка результатов секвенирования Exo-C-библиотек

В Exo-C-библиотеках глубина секвенирования составляет 136.6 млн прочтений ($2.05 \cdot 10^{10}$ bp) и 109.4 млн прочтений ($1.64 \cdot 10^{10}$ bp), а среднее покрытие в экзоме — 60.51 и

14.88 прочтений для ExoC-19 и ExoC-20 соответственно. Глубину покрытия более 10 прочтений имеют 91.68 % и 72.58 % экзома для ExoC-19 и ExoC-20 соответственно. Согласно [40], для репрезентативных результатов экзомного секвенирования необходима глубина секвенирования не менее чем в $10^{10}$ bp, а для Hi-C — не менее чем 100 млн прочтений. Минимальным порогом глубины для возможности поиска генетических вариантов считается 10 прочтений, практически все гомозиготные SNV могут быть найдены при глубине в 15 прочтений, а гетерозиготные требуют глубину прочтений не менее 33. Приемлемая доля экзома с репрезентативным покрытием (более 10 прочтений) составляет 90 %. Таким образом, можно утверждать, что ExoC-19 отвечает требованиям для поиска SNV, а ExoC-20, во-первых, пригодна к поиску только гомозиготных генетических вариантов, а во-вторых, имеет недостаточно хорошее покрытие в экзоме.

«Золотой стандарт» покрыт нашими библиотеками на 82.75 % и 96.52 %, «серебряный стандарт» — на 56.48 % и 90.06 % (библиотеки ExoC-19 и ExoC-20 соответственно). Различия объясняются протоколами приготовления: у библиотеки ExoC-20 выше глубина покрытия в экзоме, в 6 раз выше обогащение в экзомных районах (критерий Манна—Уитни $p = 0.0003$). Кроме того, в библиотеке ExoC-19 были использованы адаптерные последовательности, дающие большое количество шума.

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

- В библиотеке ExoC-19 потеряна большая доля вариантов, чем в ExoC-20 — как относительно общего числа, так и относительно вариантов «золотого» и «серебряного» стандартов.

- 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 является эффективным способом улучшения данных низкого качества.

## ВЫВОДЫ

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

1. Пайплайн, созданный нами с учётом актуальных рекомендаций для биоинформационной обработки, позволяет обрабатывать данные Exo-C-секвенирования, а также находить в этих данных SNV.

2. Использование разработанного конвейера биоинформационных инструментов позволило обнаружить около 5.5 млн генетических вариантов в контрольных данных клеточной линии K562 (что сопоставимо со средним количеством точечных полиморфизмов в геноме человека), из которых наличие 75 тыс. подтвердилось в восьми независимых исследованиях, а 1 млн — в шести независимых исследованиях, не включающих экзомные данные.

3. Сравнение генетических вариантов, полученных из контрольных образцов и Exo-C-библиотек, позволяет утверждать, что метод Exo-C способен детектировать около 75–90 % SNV, обнаруживаемых другими методами.

## КОНФЛИКТ ИНТЕРЕСОВ

Авторы декларируют отсутствие конфликта интересов.

Благодарности и ссылки на гранты размещаются здесь. Thanks and references to fundings are written here.

## Список литературы


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

2. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010. Available online at: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`. Accessed 2020/12/08.

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

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

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

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

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

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

9. Heldenbrand, J. R. *et al.* Recommendations for performance optimizations when using GATK3.8 and GATK4. *BMC Bioinformatics*. 2019. No. 20. doi: 10.1186/s12859-019-3169-7.

10. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*. 2001. No. 29. P. 308–311. doi: 10.1093/nar/29.1.308.

11. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010. No. 26. P. 841–842. doi: 10.1093/bioinformatics/btq033.

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

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

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

15. Stelzer, G. *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics*. 2016. No. 54. doi: 10.1002/cpbi.5.

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

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

18. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015. No. 526. P. 68–74. doi: 10.1038/nature15393.

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

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

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

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

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

24. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016. No. 536. P. 285–291. doi: 10.1038/nature19057.

25. 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*. 2015. No. 17. P. 405–423. doi: 10.1038/gim.2015.30.

26. Ma, W. *et al.* Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. *Methods*. 2018. No. 142. P. 59–73. doi: 10.1016/j.ymeth.2018.01.014.

27. Ramani, V. *et al.* Mapping 3D genome architecture through in situ DNase Hi-C. *Nat Protoc*. 2016. No. 11. P. 2104–2121. doi: 10.1038/nprot.2016.126.

28. Gridina, M., Mozheiko, E. & Fishman, V. A cookbook of DNase Hi-C. *Nat Protoc*. in press. .

29. Valeev, E. *Scissors: Exo-C Variants Search Pipeline*. 2020. GitHub repository: `https://github.com/regnveig/labjournal/tree/master/tools/Scissors`.

30. Lozzio, C. B. & Lozzio, B. B. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood*. 1975. No. 45. P. 321–334. URL: https://pubmed.ncbi.nlm.nih.gov/163658.

31. 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*. 2018. No. 69. P. 10–22. doi: 10.1016/j.bcmd.2017.12.002.

32. 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*. 2017. No. 123. P. 56–65. doi: 10.1016/j.ymeth.2017.04.004.

33. Dixon, J. R. *et al.* Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet*. 2018. No. 50. P. 1388–1398. doi: 10.1038/s41588-018-0195-8.

34. Moquin, S. A. *et al.* The Epstein–Barr virus episome maneuvers between nuclear chromatin compartments during reactivation. *J Virol*. 2017. No. 92. doi: 10.1128/jvi.01413-17.

35. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. 2014. No. 159. P. 1665–1680. doi: 10.1016/j.cell.2014.11.021.

36. Ray, J. *et al.* Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proc Natl Acad Sci USA*. 2019. No. 116. P. 19431–19439. doi: 10.1073/pnas.1901244116.

37. Wang, Y. *et al.* SPIN reveals genome-wide landscape of nuclear compartmentalization. *Cold Spring Harbor Laboratory*. 2020. doi: 10.1101/2020.03.09.982967.

38. Zhou, B. *et al.* Comprehensive, integrated, and phased whole-genome analysis of the primary ENCODE cell line K562. *Genome Res.*. 2019. No. 29. P. 472–484. doi: 10.1101/gr.234948.118.

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

40. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014. No. 15. P. 121–132. doi: 10.1038/nrg3642.

Рукопись поступила в редакцию 01.01.2020.

Переработанный вариант поступил 07.03.2020.

Дата опубликования 07.05.2020.

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

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

UDC: 616-074+57.087.1

# This is a manuscript template for ”Mathematical Biology and Bioinformatics”

**Ivanov I.***[1,2], **Petrov I.****[1]

[1]*Institution 1, Sity1, Country1*
[2]*Institution 2, Sity2, Country2*

***Abstract.*** The abstract should be a single subsection in the length of about 200–250 words (at least 25 lines). It should contain a brief description of used approaches. The main results of the work should be formulated here. URLs must be included only in one case. These must be the links to the websites where data, software or tools referred to in the article are hosted. References should not be included in the abstract.

***Key words:*** *the first one, the second, ..., the last.*

**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. [35] | 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. [35] | 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. [35] | 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. [35] | 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. [35] | 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. [35] | 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. [38] | 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. [38] | 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. [38] | 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. [33] | 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. [37] | 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. [37] | 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. [37] | 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. [37] | 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. [37] | 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. [37] | 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. [37] | 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. [37] | 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. [36] | 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. [36] | 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. [36] | 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. [36] | 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. [34] | 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. [34] | 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. [34] | 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. [34] | 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. [32] | 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. [31] | GEO | SRR5345331 | WES* | 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. [31] | GEO | SRR5345332 | WES* | 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. [31] | GEO | SRR5345333 | WES* | 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. [31] | GEO | SRR5345334 | WES* | 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. [31] | GEO | SRR5345335 | WES* | 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. [31] | GEO | SRR5345336 | WES* | PE | 40 807 904 | 81 649 292 | 99.59 | 0.041 | 40 559 969 | 163 957 | 8 801 283 | 91 545 | 21.77 | 79 383 290 |
| Тестовые данные | | | | | | | | | | | | | | | |
| FG_ExoCBel-001 | ExoC-19 | — | — | Exo-C | PE | 136 609 179 | 359 215 777 | 99.31 | 23.940 | 135 150 334 | 443 409 | 25 453 568 | 159 152 | 18.86 | 319 784 450 |
| FG_Quarantine-A | ExoC-20 | — | — | Exo-C | PE | 53 598 130 | 140 214 460 | 99.79 | 23.150 | 53 598 130 | 259 561 | 7 809 282 | 68 779 | 14.60 | 193 853 459 |
| FG_Quarantine-B | ExoC-20 | — | — | Exo-C | PE | 55 279 173 | 144 641 130 | 99.76 | 23.108 | 55 279 173 | 310 369 | 8 808 307 | 90 489 | 15.97 | 177 375 163 |

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

**Таблица 5.** Образцы данных секвенирования клеточной линии K562

| Образец | Тип данных | Тип прочтений | Глубина, прочтений | Общее число прочтений | Доля картированных, % от общего числа | Доля добавочных, % от общего числа | FR PE прочтения, % от картированных | Картированные PE прочтения | Картированные синглетоны | Картированные на разные хромосомы пары, % от картированных | Картированные на разные хромосомы пары (QMAP 4+), % от картированных на разные хромосомы |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Контрольные данные | | | | | | | | | | | |
| Rao et al. [35] | 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. [38] | WGS | SE | 2 621 311 293 | 2 752 058 305 | 93.43 | 4.751 | — | — | — | — | — |
| Dixon et al. [33] | 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. [37] | Repli-seq | SE | 301 663 640 | 301 669 278 | 98.09 | 0.002 | — | — | — | — | — |
| Ray et al. [36] | 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. [34] | 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. [32] | 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. [31] | 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 |