Аннотация:

Диагностика и лечение пациентов с наследственными заболеваниями требует создания эффективных методов исследования индивидуальных геномов. Существующие подходы либо нацелены на поиск узкого набора геномных вариантов, либо слишком дороги для применения в рутинной практике. Мы исследовали возможность детекции точечных мутаций и межхромосомных транслокаций при помощи секвенирования обогащенных 3C-библиотек. Данный метод позволил бы диагностировать широкий круг хромосомных перестроек, при стоимости за анализ приемлемой для введения в повседневную практику в клинике.

Тема работы: Разработка подходов для поиска и оценки клинической значимости хромосомных перестроек с использованием методов 3D-геномики

Гранты: Работа выполняется в рамках грантов РФФИ № 18-29-13021 и 20-34-90110.

Состав коллектива:

Исполнитель: аспирант ИЦиГ, младший научный сотрудник сектора постгеномной нейробиологии Можейко Евгений Александрович. Данная работа проводится в рамках написания диссертации и гранта РФФИ 20-34-90110. Научный руководитель гранта с.н.с. лаборатории молекулярной генетики человека ИЦиГ СО РАН Бабенко Владимир Николаевич bob@bionet.nsc.ru.

Научное содержание работы

1. **Постановка задачи**

Регуляция активности генома эукариот, и в том числе человека, осуществляется крайне сложной, многоуровневой молекулярной системой, многие компоненты которой на сегодняшний день плохо изучены. Благодаря появлению новых молекулярно-генетических подходов в последние годы были существенно расширены наши представления о роли трехмерной организации хроматина в регуляции генома. Анализ значительного массива данных, описывающих трехмерную организацию хроматина разных видов организмов и типов клеток с высоким разрешением, позволил сформулировать два важных для этой области фундаментальных вопроса. Во-первых, необходимо понять, какие молекулярные механизмы отвечают за укладку хроматина в ядре и как первичная структура ДНК связана с её трехмерной организацией. Во-вторых, неизвестно, насколько велик вклад трехмерной укладки генома, в сравнении с другими механизмами регуляции, в формирование паттерна генной экспрессии, задание клеточной идентичности в ходе развития и в манифестацию генетических патологий у человека. Для ответа на эти вопросы, наш Проект

предлагает использование подхода, основанного на принципах обратной генетики, позволяющего связать изменения первичной структуры ДНК с изменениями её трехмерной организацией и оценить последствий этих изменений с точки зрения генной экспрессии.

2. Современное состояние проблемы

Геномная диагностика является одной из активно развивающихся областей современной медицины. Геном каждого человека содержит 4-5 миллионов точечных мутаций, а также 1-2 тысячи крупных структурных перестроек, затрагивающих, в совокупности, около 20 миллионов пар оснований. Хотя большинство вариаций встречается в некодирующих областях, часть из них является клинически значимой и используется для диагностики и выбора подходов к лечению наследственных и онкологических заболеваний. Диагностика и лечение пациентов с наследственными заболеваниями требует создания эффективных методов исследования индивидуальных геномов. На сегодняшний день, диагностическая эффективность различных генетических тестов лежит в диапазоне 30-70%. Причина столь невысокой диагностической эффективности связана, в первую очередь, с принципиальными ограничениями используемых методов. Среди таких ограничений: узкий набор анализируемых геномных вариантов(в случае экзомного секвенирования или использования генных панелей), в том числе только в районах генов, не включая некодирующие области;низкая разрешающая способность методов (в случае микрочипового анализа). Даже наиболее эффективный из существующих методов, полногеномное секвенирование, не может быть использован в рутинной клинической практике, из-за слишком высокой стоимости исследования. Таким образом актуальной задачей в современной медицинской генетике является создание эффективного метода геномной диагностики.

3. Подходы для решения задачи

Основой разработки метода для геномной диагностики били данные секвенирования 3C-библиотек. Этапы разработки метода:

1. Анализ качества 3C-библиотек
2. Разработка метода для поиска межхромосомных и внутрехромосомных транслокаций
3. Оценка чувствительности и специфичности метода для поиска транслокаций на модельных данных
4. Разработка ПО для автоматического поиска перестроек по Hi-C данным

Для анализа качества 3C-библиотек, мы использовали программное обеспечение Hi-C pro. Это программа, которая позволяет получить из прочтений готовую матрицу контактов в формате .hic, а также вычисляет некоторые показатели качества 3C-библиотеки. В процессе выполнения

проекта, нам пришлось отказаться от Hi-C pro так, как для картирования прочтений Hi-C pro использует программу Bowtie2, а Bowtie2 показал существенно худшие результаты в сравнении с BWA. Поэтому мы приняли решение разработать собственный код для обработки выходных данных BWA и получения всей необходимой нам статистики по качеству 3C-библиотек. Программа была реализована на языке Shell, код программы досутпен в репотизтории гитхаб https://github.com/evgeniy240294/ExoC/tree/master.

Для поиска изменений количества копий участков генома, мы анализировали трек покрытия генома прочтениями 3C-бибилиотек. Трек покрытия исследуемого образца сравнивался с треком покрытия контрольного образца без перестроек. Величины отклонения значений покрытия исследуемого образца от контрольного были показателем наличия перестройки. Для сравнения треков покрытия мы использовали программное обеспечение BIC-seq2.

Для поиска межхромосомных транслокаций мы использовали особенности трехмерной организации, а именно то, что частота внутрихромосомных (cis) контактов в среднем существенно выше, чем частота межхромосомных (trans) контактов [Lieberman-Aiden et. al 2009]. Благодаря этой особенности трехмерной организации, межхромосомные транслокации легче детектировать на основе Hi-C-данных, чем другие виды хромосомных перестроек. Эта особенность известна из литературы[Lieberman-Aiden et. al 2009], и существуют публикации, в которых описаны различные подходы к детекции межхромосомных транслокаций по Hi-C карте [Harewood et. al 2017, Chakraborty et. al 2018, Ma et. al 2015, Wang et. al 2020]. Однако все эти подходы предназначены для полной, не обогащенной последовательностями экзонов, Hi-C карты. При обогащении необходимо учитывать повышенную неравномерность покрытия, и полное отсутствие покрытия в некоторых участках генома. Эти факторы могут оказать влияние на точность и специфичность детекции транслокаций.

Суть разработанного нами метода для поиска транслокаций заключается в следующем: мы рассматриваем вероятность наблюдать изменение соотношения частот cis/trans-контактов фиксированного локуса относительно контрольных значений cis/trans-контактов. Таким образом, при транслокации, за счет того, что частота cis контактов существенно больше частоты trans контактов, мы увидим отклонение отношения cis/trans от контрольного в данном локусе.

Для поиска внутрехромосомных транслокаций мы использовали аналогичных подход, что и для межхромосомных: ведь при внутрехромосомных транслокация одно плечо хромосомы оказывается ближе к транслоцированному участку, чем до транслокации, в то время как другое плечо, наоборот оказывается дальше. Таким образом для детекции подобной транслокации мы предложили использовать абсолютную разницу в контактах фиксированного локуса со всей хромосомой, для исследуемого образца и контрольного.

4. **Полученные результаты**

Разработаны подходы для анализа качества 3C-библиотек, обогащенных последовательностями экзонов. С помощью этих подходов были проанализированы данные пятидесяти обогащенных 3C-библиотек. Эффективный анализ позволил выявить недостатки используемого протокола сборки 3C-библиотек, в и процессе помог существенно улучшить используемый протокол. На рисунке 1 изображены ключевые показатели качества 3C-библиотек в зависимости от используемого протокола сборки 3C-библиотек. На рисунке видно, что для некоторых протоколов сборки, все показатели качества в среднем лучше, что говорит о нашей успешной модификации протоколов.

В процессе разработки метода для поиска изменений количества копий генома, была обнаружена зависимость паттерна трека покрытия генома прочтениями от степени обогащения 3C-библиотеки экзомными последовательностями. Таким образом паттерн трека покрытия оказался чувствительным к степени обогащение 3C-библиотеки, что в свою очередь стало причиной существенных различий в покрытии генома прочтениями между двумя экспериментами. На рисунке 2A, для двух с близкой степенью обогащения образцов, гистограмма распределения отношения покрытий 50 kb участков генома прочтениями. На рисунке 2B аналогичная гистограмма для двух образцов с существенно различным обогащением.

Разработаны алгоритмы для поиска внутрехромосомных и межхромосомных транслокаций. Показана неэффективность поиска инверсий и изменений числа копий участков генома по данным о трехмерной организации. Проведена оценка эффективности поиска межхромосомных транслокаций на раковой клеточной линии K562 (рисунок 3). Так как другой наиболее современный и точный метод для поиска транслокаций HiNT [Su Wang et. al 2020] обладает меньшей эффективностью AUC=0.85, то мы показали преимущества разработанного метода над существующими.

Чувствительность и специфичность поиска транслокаций по Hi-C данным мы оценили на симулированных данных. Всего мы смоделировали около 5*(10^4) различных транслокаций, что позволило оценить не только расположение в геноме, но и зависимость от покрытия прочтениями транслоцированного локуса (рисунок 4).

Разработанный пайплайн для оценки качества Hi-C данных и поиска транслокаций был опубликован на github https://github.com/eamozheiko/HiLocus.

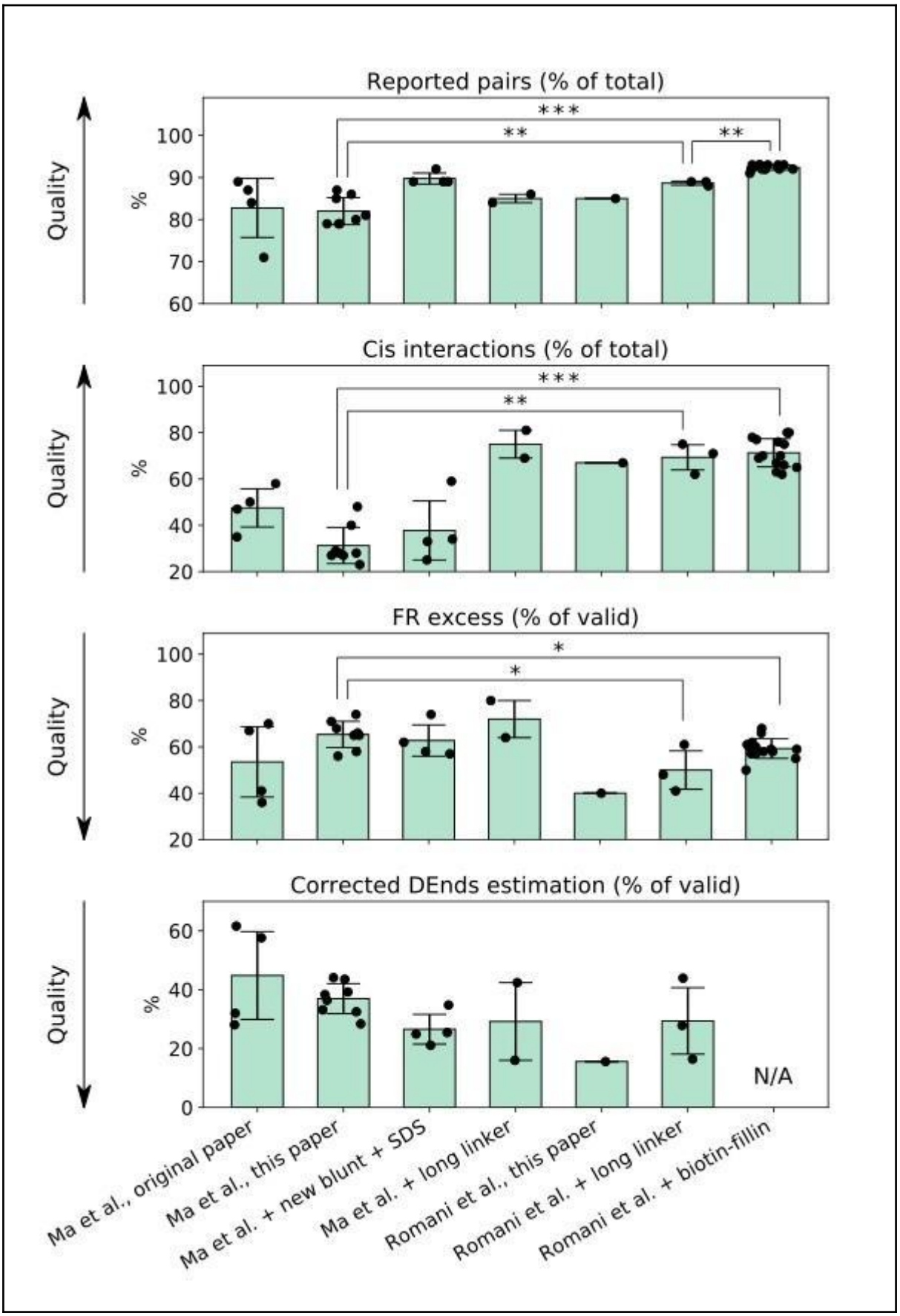
Reported pairs (% of total)
Cis interactions (% of total)
FR excess (% of valid)
Corrected DEnds estimation (% of valid)
Quality
%
***
**
*
N/A
Ma et al., original paper
Ma et al., this paper
Ma et al. + new blunt + SDS
Ma et al. + long linker
Romani et al., this paper
Romani et al. + long linker
Romani et al. + biotin-fillin

Рисунок 1. Метрики качества обогащенных 3C-библиотек для различных вариантов протокола сборки библиотек. Черными точками показаны значения для отдельных экспериментов. Столбцами показано среднее значение параметра качества 3C-библиотеки для конкретного протокола сборки. Стрелками показано положительное направления изменения качества. В качестве метрик качества использовались соответственно: процент уникально картированных пар прочтений 3C-библиотек(**Reported pairs**), процент трехмерных пар контактов на одной молекуле(**Cis**), процент превышения forward-reverse ориентаций пар прочтений(**FR excess**), процентное содержание нелигированных фрагментов(**Corrected DEnds estimation**).

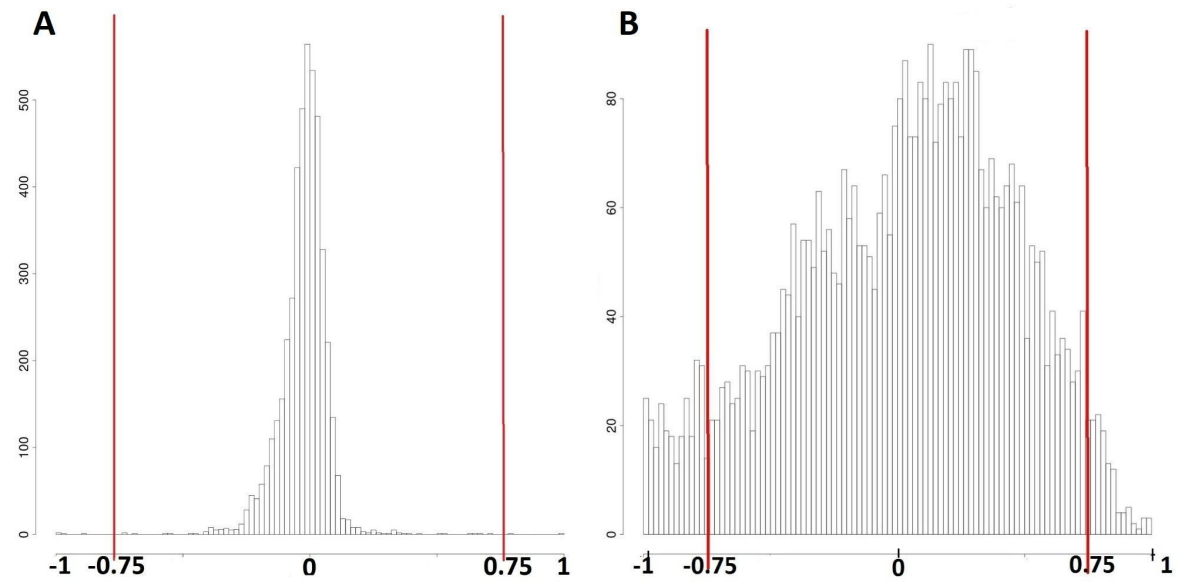

Рисунок 2. Гистограмма логарифма отношения треков покрытия прочтениями 50kb участков генома. Треки прочтения построены по данным обогащенных последовательностями экзонов 3C-библиотек. Слева для образцов со сходной степенью обогащения. Справа для образцов с различной степенью обогащения. Красными линиями обозначены пороги, за которыми значение считается вызванным перестройкой.

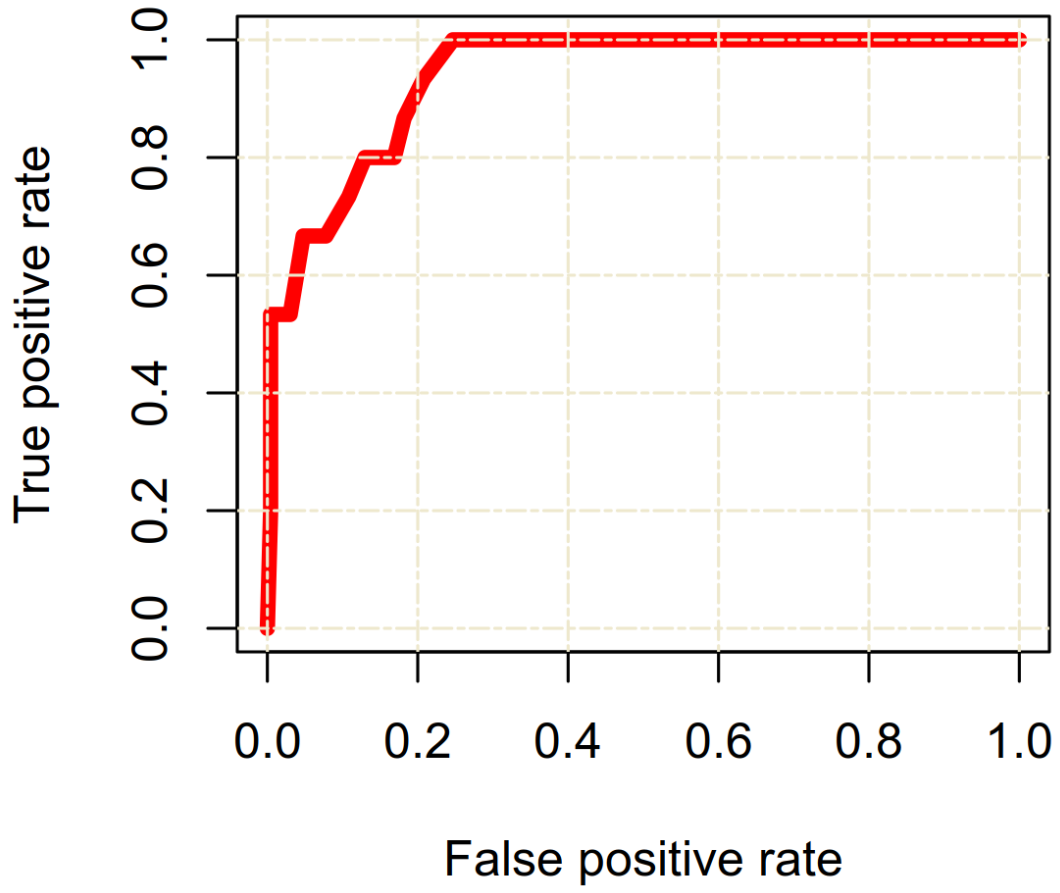

Рисунок 3. ROC для поиска транслоцированных пар хромосом для клеточной линии K562. AUC = 0.93

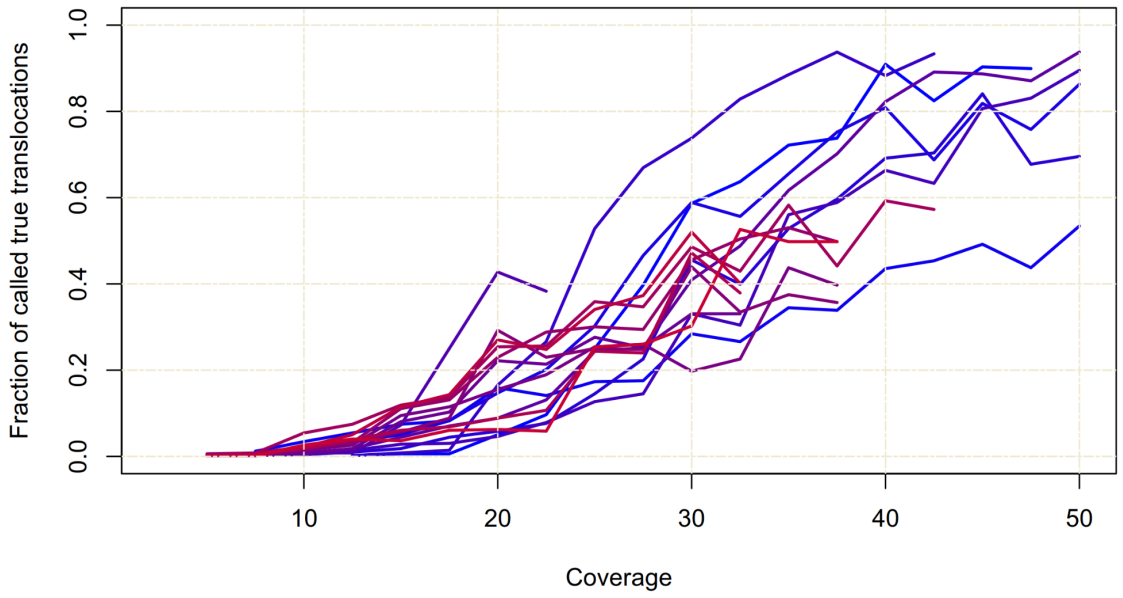

Рисунок 4. Оценка чувствительности и специфичности поиска транслокаций на симулированных транслокациях. Показана точность метода в зависимости от покрытия прочтениями транслоцированных учатсков. Различными цветами показаны различные образцы, которые использовались в качестве набора нормальных (не перестроенных) случаев.

Эффект от использования кластера в достижении целей работы

Использование вычислительного кластера сыграло ключевую роль в достижении целей работы так, как секвенированные 3C-библиотеки представляют из себя огромный объем данных, который крайне затруднительно или невозможно обработать без необходимых вычислительных ресурсов (оперативная память, дисковое пространство, параллельное вычисление)

Публикации содержащие полученные результаты

1. Maria Gridina, Ph.D; **Evgeniy Mozheiko**; Emil Valeev;Ludmila P. Nazarenko; Maria E. Lopatkina; Zhanna G. Markova; Maria I. Yablonskaya; Viktoria Yu. Voinova; Nadezhda V. Shilova; Igor N. Lebedev; Veniamin Fishman. A cookbook of DNase Hi-C. Epigenetics & Chromatin. 2021 Mar 20;14(1):15. https://doi.org/10.1186/s13072-021-00389-5

2. Основные результаты по алгоритмам поиска перестроек находятся в процессе написания статьи