

Тема работы. Поиск хромосомных перестроек в геноме человека на основе технологии захвата конформации хромосом

Состав коллектива (на момент выполнения работы):

1. к.б.н. Фишман Вениамин Семенович (НГУ, ИЦиГ СО РАН)
2. к.б.н. Белокопытова Полина Станиславовна (НГУ)
3. к.б.н. Нуриддинов Мирослав Абдурахимович (ИЦГ СО РАН)

Финансовая поддержка. Грант РНФ 22-14-00247 (руководитель Фишман В.С.), гос. задание ИЦиГ СО РАН (руководитель Груntenко Н.Е.), грант Министерства Высшего Образования и Науки FSUS--2024--0018

Научное содержание работы.

Постановка задачи

Современные методы генетического анализа, такие как секвенирование экзома (WES), хромосомный микроматричный анализ (aCGH) и методы захвата конформации хромосом (Hi-C), имеют ограничения в определении геномных вариаций. Для преодоления этих ограничений разработана новая экспериментальная методика Echo-C, которая объединяет анализ хромосомной конформации с обогащением экзомного покрытия. Основная задача исследования — создать вычислительные инструменты для анализа данных Echo-C и оценить возможности этого метода для одновременного выявления структурных вариаций (SVs) и однонуклеотидных замен (SNVs) в геноме человека.

Современное состояние проблемы (на момент начала работы)

Традиционные подходы к анализу геномных вариаций либо ограничиваются конкретными типами мутаций, либо требуют значительных затрат и времени. Например, WES эффективно определяет SNVs, но не способен обнаружить крупные структурные перестройки за пределами экзома. Hi-C, напротив, подходит для выявления структурных вариаций, но имеет низкую чувствительность к SNVs. Таким образом, назрела необходимость интеграции методов для одновременного анализа различных типов вариаций.

Подробное описание работы, включая используемые алгоритмы

Для оценки эффективности метода Echo-C мы применили его к когорте из 66 человеческих образцов, включая 28 женщин, 37 мужчин и один образец клеточной линии K562. До проведения анализа Echo-C большинство этих образцов прошло молекулярное и/или цитогенетическое исследование. В состав когорты входили пациенты с разнообразными генетическими аномалиями: 23 сбалансированные транслокации, 6 инверсий (за исключением клинически незначимых перичентрических инверсий в гетерохроматиновых регионах, не определяемых методами секвенирования с коротким прочтением), 86 CNV, а

также несколько экзонных SNV с клинической значимостью. Кроме того, в исследование были включены 6 здоровых доноров с нормальным кариотипом и клетки K562 в качестве контрольных образцов. У части пробандов наблюдались сложные структурные вариации, ранее охарактеризованные методами FISH и классического кариотипирования. Совокупный набор данных Ехо-С, включающий все образцы, получен из примерно 2 миллиардов прочтений и содержит около 1 миллиарда Hi-C контактов, что делает его одним из самых масштабных наборов данных по захвату Hi-C у человека на сегодняшний день.

Для автоматического определения транслокаций мы разработали два взаимодополняющих метода: градиентный анализ паттернов (gradient pattern search) и линейный анализ паттернов (line pattern search).

Градиентный метод.

На каждом уровне разрешения карты Ехо-С (4096 кб, 256 кб, 16 кб, в порядке убывания) мы определяем «точки интереса» (DOI — dots of interest) как элементы матрицы контактов Ехо-С, которые удовлетворяют следующим условиям:

1. Контакт в позиции (i, j) не равен нулю ни в образце, ни в контрольной выборке;
2. Разность логарифмов биномиальных вероятностей контакта между образцом и контролем меньше определённого порогового значения, зависящего от разрешения карты.

Расчёт метрики:

$$DAr_{ij} = \log_{10}(P_{\text{sample}}) - \log_{10}(P_{\text{control}}) < Thr(res)$$

где P_{sample} и P_{control} — биномиальные вероятности получения Hi-C-контакта в точке (i, j) при соответствующих глубинах секвенирования.

Значения порогов (Thr):

- -4.0 при разрешении 4096 кб
- -2.0 при разрешении 256 кб
- -1.0 при разрешении 16 кб

Затем для каждой найденной точки DOI проверяется гипотеза, что она соответствует разрыву при транслокации. Предполагается, что DOI находится на расстоянии 1 бина от разрыва, её соседи — на расстоянии 2 бина и далее. Под эту гипотезу рассчитывается вероятность цис-контакта. В альтернативной гипотезе (никакой перестройки нет), используется вероятность, соответствующая межхромосомным контактам.

Для каждой DOI выбираются четыре прямоугольных окна (верхнеправое, верхнелевое, нижнеправое, нижнелевое), и в каждом окне рассчитывается метрика ArLG — уровень

"артефакта", говорящего в пользу транслокации, как агрегированная (суммированием) разность логарифмов вероятностей вышописанных гипотез по всем точкам окна. Если сумма положительна, то окно с DOI содержит аномальные контакты, более характерные для внутрихромосомного взаимодействия, чем для фоновых межхромосомных.

Метод линейного поиска паттернов (Line pattern search) также начинается с определения «точек интереса» (DOI) на контактной карте Eho-C, как и в предыдущем методе. Анализ проводится на более высоком разрешении: 256 кб, 16 кб и 1 кб.

Затем для каждой точки вычисляется специальная метрика (ArLL), которая показывает, насколько вероятно, что наблюдаемые контакты между хромосомными участками принадлежат образцу, а не контролю. Если контактные данные больше похожи на сигнал из образца, чем на контрольный шум, то значение ArLL положительно.

Далее алгоритм выполняет следующие шаги:

1. Для каждой пары колонок (то есть участков двух хромосом) оценивается метрика ArLL.
2. Если метрика превышает порог 6 и расстояние между колонками меньше 5 бинов, они объединяются в одну транслокацию.
3. После объединения колонок ArLL пересчитывается с учётом новых границ.
4. Аналогичная процедура выполняется и по строкам матрицы, чтобы уточнить точную позицию вставки (ближайшую к точке разрыва).
5. В конце применяется дополнительный фильтр, зависящий от общей суммы контактов в области: значение ArLL должно быть больше, чем экспонента от логарифма суммы контактов, с параметром, подобранным по моделированным данным.

Для поиска CNV мы адаптировали существующие инструменты для анализа WES — CNVkit, CoNIFER и GATK — на Eho-C данных. Выравнивание было выполнено на сборку hg19, с последующим использованием BAM-файлов и фильтрацией на основе базы DGV Gold для исключения распространённых вариантов. Дополнительно CNV были классифицированы на высоконадёжные (подтверждённые qPCR или другими методами) и низконнадёжные (не подтверждённые).

Полученные результаты

В результате экспериментов мы получили один из крупнейших на сегодняшний день наборов Hi-C-данных для человека, включающий более 2 миллиардов прочтений и около 1 миллиарда контактов. Использование Eho-C позволило визуализировать крупные хромосомные перестройки даже несмотря на смещение покрытия в сторону экзона — хромосомные перестройки оказывали влияние на паттерны контактов далеко за пределами своих точек разрыва.

Традиционные инструменты анализа Hi-C, такие как EagleC, оказались недостаточно чувствительными для поиска хромосомных перестроек в полученных нами данных: F1-score составил 0.15 для транслокаций, 0.03 для инверсий, 0 для CNV. Наша система автоматического поиска транслокаций продемонстрировала 100% полноту (recall) и 36% точность (precision) на необработанных данных. После визуальной валидации и фильтрации по метрике cis/trans (58%–75%) точность повысилась до 73%, сохраняя полноту в 100%. Большинство ложно-положительных событий представляют собой реальные сигналы, не видимые при стандартной микроскопии, и часть из них была подтверждена методами PCR или FISH.

Моделирование 330 синтетических транслокаций с помощью алгоритма Charm (<https://www.biorxiv.org/content/10.1101/2023.11.22.568374>) показало ожидаемую зависимость чувствительности от размера фрагмента: 5% для вставок 5–50 кб, 45% для 200–500 кб, 100% — для >1 Мб. Общий F1-score составил 0.77 (точность 87%, полнота 68%).

Для инверсий, при анализе шести клинических случаев, алгоритм корректно идентифицировал три, но с высоким числом ложных срабатываний (220). После фильтрации по cis/trans соотношению число FP снизилось до 33 при сохранении чувствительности 50%. Одна из инверсий, изначально посчитанная FP, была подтверждена секвенированием по технологии ONT.

На синтетических данных, для 240 инверсий средний F1-score составил 0.49 (точность 56%, полнота 45%), с ухудшением распознавания крупных инверсий (>90 Мб) из-за особенностей распределения контактов на концах хромосом.

Для CNV наиболее стабильные результаты показал инструмент CNVkit, определяя >60% известных CNV даже без доработки под специфику Echo-C-данных. Однако только высоконадёжные CNV хорошо воспроизводились всеми тремя инструментами. Фильтрация по Jaccard Index >50% и исключение распространённых вариантов из публичной базы DGV заметно снизила FP, особенно у GATK (на 70% для CNV <100 кб), но не дала значительного прироста точности в абсолютных значениях.

Таким образом, нами продемонстрировано, что Echo-C данные пригодны для автоматического одновременного анализа широкого спектра структурных перестроек, включая транслокации, инверсии и CNV. Разработанные нами алгоритмы обеспечивают высокую чувствительность и удовлетворительную точность, особенно в случае транслокаций, с возможностью дальнейшей валидации обнаруженных событий с помощью ортогональных методов.

Эффект от использования кластера в достижении целей работы

Для обработки больших объемов данных Echo-C использовались ресурсы высокопроизводительного вычислительного кластера Новосибирского Государственного Университета. Использование узлов кластеры обеспечило:

- Быструю обработку карт контактов хроматина и секвенированных данных.
- Моделирование структурных вариаций и автоматизацию алгоритмов их обнаружения.
- Анализ больших наборов данных, включая валидацию результатов на 66 клинических образцах.

Применение параллельных вычислений позволило существенно сократить время обработки данных.

Перечень публикаций, содержащих результаты работы.

1. Gridina et al., 2024. A novel approach for simultaneous detection of structural and single-nucleotide variants based on a combination of chromosome conformation capture and exome sequencing. Препринт (bioRxiv). <https://doi.org/10.1101/2024.01.26.577292>

Аннотация для сайта

Мы разработали метод Ехо-С, который представляет собой инновационное решение в области генетического анализа, объединяющее технологии секвенирования экзома и анализа хромосомной конформации. Это позволяет одновременно выявлять однонуклеотидные замены и структурные перестройки генома, что значительно ускоряет и удешевляет диагностику генетических заболеваний. Использование вычислительного кластера НГУ для обработки данных сделало возможным анализ генетических образцов с высокой скоростью и выявления патогенных вариантов, включая редкие мутации, которые ранее оставались незамеченными. Исследование продемонстрировало успешное применение Ехо-С для выявления сложных генетических аномалий, подтверждая его потенциал для широкого использования в клинической практике.