Тема работы:
Разработка алгоритма поиска инверсий в данных Hi-C данных с экзомным обогащением.

Состав коллектива:
Белокопытова Полина Станиславовна, к.б.н. м.н.с. ФЕН НГУ, н.с. ИЦиГ СО РАН
Нуриддинов Мирослав Абдурахимович, к.б.н. н.с. ИЦиГ СО РАН
Лагунов Тимофей Аркадьевич, аспирант ФФ НГУ, инженер-исследователь ИЦиГ СО РАН
Фишман Вениамин Семёнович, д.б.н., в.н.с. ИЦиГ СО РАН, доцент НГУ

Научное содержание работы:

1) Постановка задачи:
Целью работы является разработка алгоритма поиска инверсий с использованием данных о пространственной организации хроматина. Ставились следующие задачи:

- Анализ данных Hi-C с экзомным обогащением;
- Тестирование алгоритма EagleC для поиска хромосомных перестроек;
- Нормализация данных Hi-C с экзомным обогащением;
- Разработка алгоритма поиска инверсий в нормализованных данных Hi-C;
- Тестирование алгоритма поиска инверсий на экспериментальных и смоделированных данных Hi-C с экзомным обогащением.

2) Современное состояние проблемы:
Хромосомные перестройки - генетические мутации, которые заключаются в изменении порядка, удалении или дупликации определенных фрагментов генома. Поиск хромосомных перестроек необходим для диагностики наследственных и онкологических заболеваний человека. На сегодняшний день не существует идеального метода для детекции хромосомных перестроек в геноме человека.

Самые передовые подходы анализа генома человека, такие как полногеномное секвенирование, являются дорогостоящими и, следовательно, не могут быть широко внедрены в клиническую практику. Обычные методы генетической диагностики включают молекулярно-цитогенетический анализ (например, сравнительную геномную гибридизацию на основе массива (aCGH)), кариотипирование и секвенирование экзома [1]. Однако каждый из этих методов ограничен по разрешающей способности и спектру обнаруживаемых типов вариантов.

Семейство методов определения конформации хромосом, включая широко используемый анализ Hi-C, использует методы бесконтактного лигирования для анализа контактов хроматина по всему геному [2]. Результатом Hi-C-эксперимента является список частот трехмерных контактов для каждой пары участков генома человека. Использование Hi-C-данных для поиска хромосомных перестроек основывается на том факте, что порядок расположения участков в геноме влияет на частоту контактов между ними. Исследования других авторов [3] и наше собственное исследование [4] продемонстрировали, что структурные варианты, включая инверсии и транслокации, визуально видны на картах Hi-C. Однако, как полногеномный анализ, Hi-C требует значительной глубины секвенирования для идентификации структурных перестроек, что является дорогостоящим. Чтобы решить эту проблему нашим коллективом был разработан метод Hi-C с экзомным обогащением, который может быть использован для одновременного выявления структурных вариантов и точечных мутаций в геноме человека. Этот метод позволяет получать информацию о пространственной архитектуре хроматина с хорошим разрешением в кодирующих участках генома. Однако текущие алгоритма поиска хромосомных перестроек по картам Hi-C [5, 6] не работают на данных такого типа, что ставит задачу разработки автоматического пайплайна поиска хромосомных перестроек на картах Hi-C с экзомным обогащением.

1. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. Nat Rev Genet. 2013;14:415–26.
2. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
3. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018;50:1388–98.
4. Lukyanchikova V, Nuriddinov M, Belokopytova P, Taskina A, Liang J, Reijnders MJMF, et al. Anopheles mosquitoes reveal new principles of 3D genome organization in insects. Nat Commun. 2022;13:1960.
5. Li J, Gao L, Ye Y HiSV: a control-free method for structural variation detection from Hi-C data. PLoS Comput Biol. 2023; 19:e1010760. 10.1371/journal.pcbi.1010760. - DOI - PMC - PubMed
6. Wang X, Luan Y, Yue F EagleC: a deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. Sci Adv. 2022; 8:eabn9215. 10.1126/sciadv.abn9215.

3) Подробное описание работы, включая используемые алгоритмы:

Нашими коллегами был сделан Hi-C эксперимент с экзомным обогащением для 66 образцов человека, включающих 28 женщин, 37 мужчин и два образца иммортализованных клеточных линий: K562 и A549.

Обработка данных Hi-C-эксперимента

Первичная обработка сырых данных секвенирования включала удаление адаптеров Illumina с использованием программы Cutadapt версии 4.1 [1]. Для выравнивания парных чтений использовался скрипт juicer версии 1.5.6 [2]. Мы использовали сборку генома человека hg19. При фильтрации оставлялись только пары чтений без дупликатов и с качеством картирования не ниже 30. Далее были получены .mcool файлы при помощи утилиты cooler [3].

Разработка алгоритма поиска инверсий

Для поиска инверсии мы использовали контрольный образец, созданный как сумма всех доступных образцов. Мы искали инверсии с разным разрешением (1 МБ, 250 Кб, 100 Кб, 10 Кб) независимо друг от друга, а затем объединяли все эти предсказания. Чтобы ускорить вычисление, мы определили минимальный и максимальный размеры инверсии на каждом требуемом разрешении (Таблица 1).

Таблица 1. Размер инверсий на разном разрешении (bin sizes)

Разрешение (bp)	**Мин. Размер инверсии (bp)**	**Макс. Размер инверсии (bp)**	**Sweet size (bins)**
1000000	25000000	Длина хромосомы	5
250000	10000000	30000000	10
100000	2000000	12000000	10
10000	50000	3000000	20

Для каждого разрешения на первом этапе мы считали матрицу различий между контрольным и экспериментальным образцом (уравнение 1).

$$DAr_{ij} = log_{10}\left(Bin\left(C^s_{ij}, D^s, \frac{C^c_{ij}}{D^c}\right)\right) - log_{10}\left(Bin\left(C^c_{ij}, D^c, \frac{C^c_{ij}}{D^c}\right)\right) < Thr(res) \quad (1)$$

Где Bin функция вероятности биномиального распределения, C_{ij} вероятность частоты контактов хроматина в точке (i, j) для образца (s) и контроля (c), D глубина секвенирования для образца и контроля. Мы исключили как из контроля, так и из образца точки, содержащие нулевое значение. Thr(res) это порог, который зависит от разрешения матрицы.

$$Thr\,(res) = \begin{cases} -4.0, & res = 4096kb \\ -2.0, & res = 256kb \\ -1.0, & res = 16kb \end{cases}$$

Далее, для каждой точки мы считали метрику, которая показывает вероятность того, что эта точка является точкой инверсии. Эта метрика позволяет найти паттерн инверсии в виде бабочки на карте Hi-C. Мы назвали ее “sweet metric”:

$$SweetMetric = LeftSum + RightSum \quad (2)$$

Где

$$LeftSum = \sum_{i,j \in W_{left}} DAr_{ij} \quad (3)$$

$$RightSum = \sum_{i,j \in W_{right}} DAr_{ij} \quad (4)$$

Где DAr_{ij} - разность логарифмов биномиальной функции вероятности образца и контроля (уравнение 1). W_{right} и W_{left} – равнобедренные треугольники с вершиной в точке (i, j) (Рис.1) и ребрами, равными размеру треугольника. Мы использовали разные размеры треугольника для разных разрешений (Таблица 1).

Затем вычислялось значение z-score для каждой точки. Мы установили пороговое значение для каждого разрешения и выбрали все значения z-score, которые были ниже этого порога. После этого мы объединили все пересекающиеся координаты с текущим разрешением, используя координаты точки с наименьшим значением z-score в качестве финальных координат инверсии. Наконец, все предсказанные инверсии с разными разрешениями объединялись и уточнялись точные координаты инверсии на разрешении 10 Кб. Для этого мы итеративно считали “sweet metric” для точек примерно в 10 бинах от изначальных координат.

1. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads//EMBnet.journal, 2011, T. 17, N 1, C. 10.
2. Durand N.C., Shamim M.S., Machol I., Rao S.S.P., Huntley M.H., Lander E.S., Aiden E.L. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments//Cell Systems, 2016, Vol. 3, No. 1, P. 95-98.
3. Abdennur N., Mirny L.A. Cooler: scalable storage for Hi-C data and other genomically labeled arrays//Bioinformatics, 2020, Vol. 36, Cooler, No. 1, P. 311-316

4) Полученные результаты:

Из шести инверсий, найденных в клинических образцах другими экспериментальными методами, наш инструмент поиска инверсий успешно выявил три, наряду с 220 ложноположительными инверсиями, что соответствует точности 10% и полноте 50%. Стоит отметить, что чувствительность обнаружения инверсии тесно связана с качеством экспериментальных данных, количественно определяемым соотношением цис- и транс-контактов. Интересно, что количество ложноположительных результатов значительно снижается с 220 до 33 в образцах похожего соотношения цис- и транс-контактов, в то время как полнота остаётся около 50% для этой подгруппы. Дальнейшее исследование показало, что некоторые ложноположительные результаты соответствуют инверсиям, которые невозможно обнаружить с помощью цитогенетического анализа, но которые можно проверить альтернативными методами. Например, инверсия размером 2 Мб, выявленная в данных образца P10, первоначально была найдена как ложноположительная. Последующая проверка с использованием секвенирования с длинными прочтениями, подтвердила эту инверсию и ее корреляцию с наблюдаемым фенотипом в образце P10. Следовательно, фактическая точность алгоритма обнаружения инверсии, вероятно, выше, чем предполагалось изначально, что свидетельствует о завышенной оценке ложноположительных результатов.

Затем мы оценили производительность алгоритма в диапазоне размеров инверсий, от 100 килобаз до нескольких мегабаз, используя 240 инверсий, сгенерированных *in silico* (Рис. 2 и Рис. 3). Этот сравнительный анализ, проведенный с использованием серии сгенерированных карт Hi-C, даёт f1-score равный 0,49, для всех длин инверсий (точность: 56%, полнота: 45%, истинных положительных результатов: 107, ложноположительных результатов: 85, ложноотрицательных результатов: 133). Анализ выявил положительную корреляцию между точностью работы алгоритма и размером инверсии, причем более крупные инверсии дают более высокий f1-score. Было отмечено, что алгоритм плохо работает для поиска инверсий размером более 90 Мб, в первую очередь из-за того, что их координаты расположены на концах хромосом, что затрудняет распознавание паттернов инверсий на картах Hi-C.

5) Иллюстрации

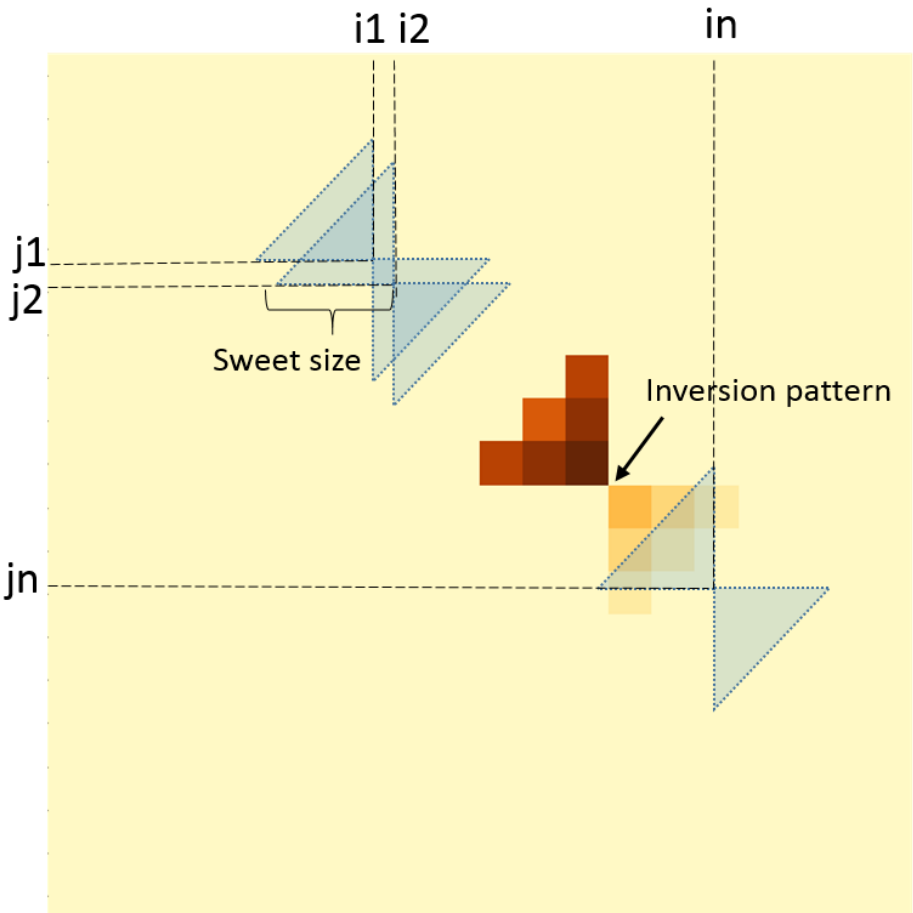

Рис. 1. Схематическое представление “sweet metric” для поиска инверсий. (i, j) - координаты в матрице Hi-C. Считается сумма значений вероятности контакта, расположенных внутри синих треугольников, для каждой точки в матрице.

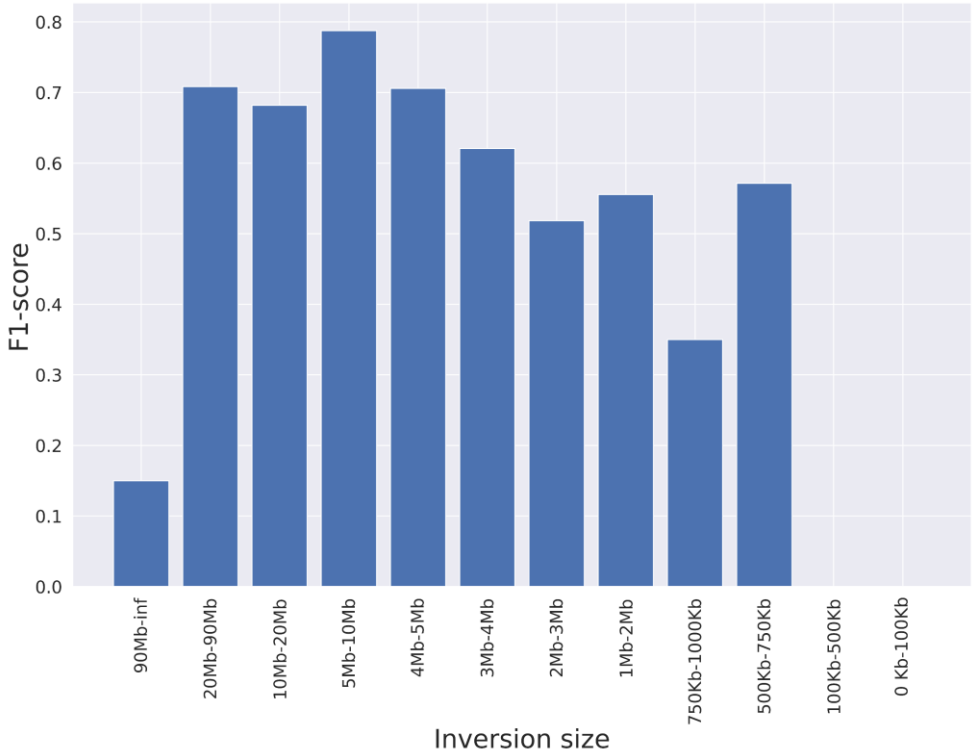

Рис. 2 f1-score алгоритма поиска инверсий для набора смоделированных Hi-C карт с экзомным обогащением и инверсиями различной длины.

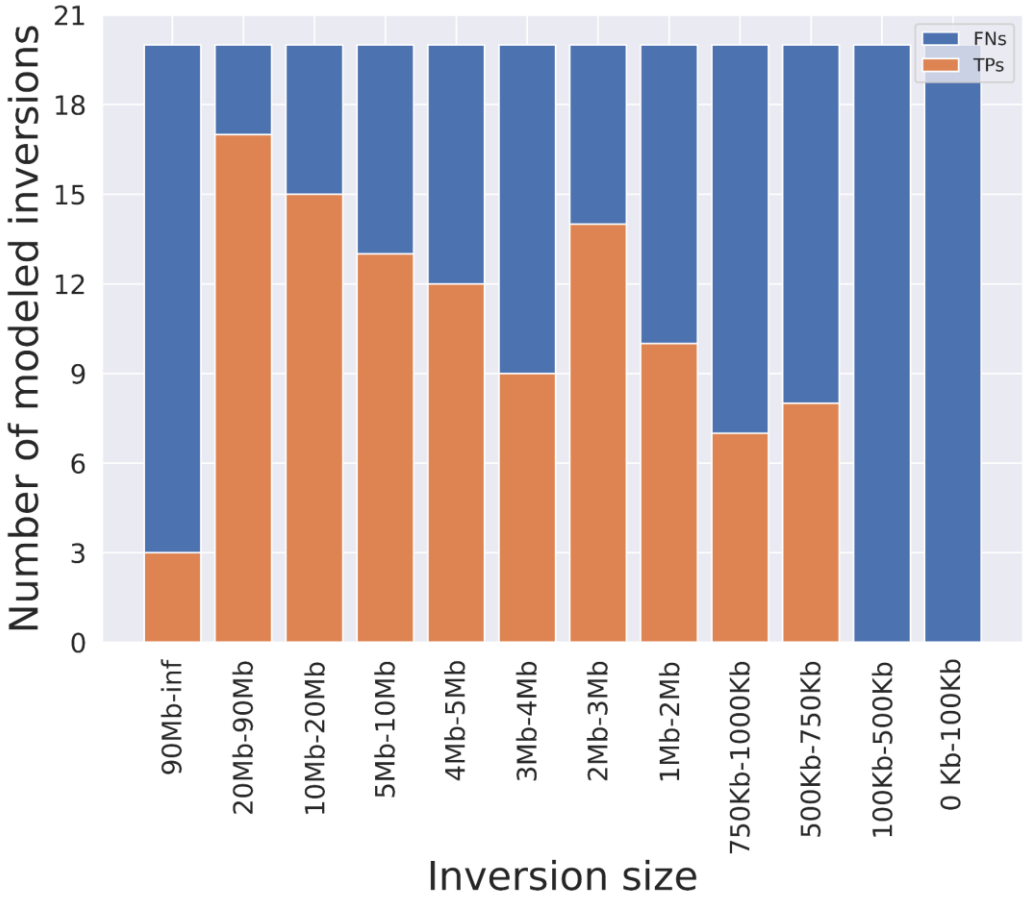

Рис. 3 Количество TP значений для смоделированных инверсий, сгруппированных по размеру, где в каждой группе содержится 20 инверсий.

Эффект от использования кластера в достижении целей работы:
Распараллеливание и большой объём оперативной памяти позволяет эффективно работать с такими большими данными, как данные секвенирования. Это позволяет увеличить скорость выравнивания данных на геном, а также позволяет искать инверсии параллельно на нескольких хромосомах.

Перечень публикаций, содержащих результаты работы.

1. Nuriddinov M, Belokopytova P, Fishman V. Charm is a flexible pipeline to simulate chromosomal rearrangements on Hi-C-like data. NAR Genom Bioinform. 2025 Jun 19;7(2):lqaf081. doi: 10.1093/nargab/lqaf081. PMID: 40585301; PMCID: PMC12204402.
2. Gridina, M., Lagunov, T., Belokopytova, P. *et al.* Combining chromosome conformation capture and exome sequencing for simultaneous detection of structural and single-nucleotide variants. *Genome Med* **17**, 47 (2025). https://doi.org/10.1186/s13073-025-01471-3