Отчёт о проделанной работе с использованием оборудования ИВЦ НГУ

1. Аннотация

Проводится разработка программного обеспечения для моделирования данных 3D-геномики с предопределёнными хромосомными перестройками. В основе данного ПО лежит использование алгоритма перекартирования пространственных контактов с референсного генома на геном, несущий перестройки.

Использование в качестве списка заранее предопределённых перестроек карт синтении позволяет проводить межвидовое сравнение архитектуры хроматина на уровне отдельных контактов. Исследование, проведённое на разных видах позвоночных и комарах рода *Anopheles,* показало, что внутри указанных таксономических групп наблюдается высокая консервативность архитектуры хроматина. Во всех случаях, б*о*льшая консервативность наблюдается для наиболее сильных и наиболее слабых контактов, что соответствует представлениям о механизмах формирования пространственной организации хроматина.

В текущий момент проводится расширения данного алгоритма для моделирования пространственной организации хроматина с заранее определёнными пользователем хромосомными перестройками. Первичные результаты показывают, что достигается высокое сходство модели и реальных перестроек для разных типов данных 3D-геномики.

2. Тема работы

Моделирование 3D геномных данных с предопределёнными хромосомными перестройками

3. Состав коллектива

Нуридинов Мирослав Абдурахимович, аспирант, м.н.с., Институт Цитологии и Генетики СО РАН

Фишман Вениамин Сергеевич, канд. биол. наук, в.н.с. Институт Цитологии и Генетики СО РАН

4. Информация о грантах

РНФ, №21-14-00182, Механизмы хромосомной пластичности малярийных комаров, 2021-2023, руководитель - Коханенко А.А.

РНФ, №21-65-00017, Патогенетика наследственных форм умственной отсталости: клеточные, молекулярные и онтогенетические аспекты, 2021-2023, руководитель - Лебедев Игорь Николаевич

5. Научное содержание работы

5.1. Постановка задачи

Разработка методов моделирования данных 3D-геномики с заранее предопределёнными перестройками. Проверка эффективности разработанных методов для разных типов данных по 3D-геномике.

5.2. Современное состояние проблемы

В настоящее время, благодаря развитию экспериментальных методик семейства захвата конформации хромосом показано, что хроматин в пространстве клеточного ядра уложен не произвольно, а формируя сложные, динамические, структуры [1]. Показано, что архитектура

хроматина демонстрирует эволюционную консервативность [2,3] для млекопитающих. Консервативным оказывается и организация хроматина и между разными типами клеток [2,4], при этом наблюдаемые различия соответствуют разнице в профиле экспрессии [5].

Более детальные исследования показывают, что поддержание правильной укладки хроматина в ядре непосредственно связано с реализацией генетической информации. Нарушение архитектуры хроматина в следствии, например, хромосомных перестроек, является одним из факторов развития врождённых заболеваний [6] и отклонений в онтогенезе [7,8]. Таким образом, детекция хромосомных аббераций и микроперестроек является клинически значимой. Показано, что одним из наиболее точных способов обнаружения перестроек являются методы, основанные на анализе пространственной организации хроматина с помощью Hi-C [9]. Использование машинного поиска для обнаружения перестроек в Hi-C подобных данных, кажется полезным; к сожалению, обилие популяционных вариаций и множество неизвестных ещё перестроек мешает настройке алгоритмов поиска. Чтобы решить эту проблему, разрабатывается алгоритм для моделирования Hi-C-подобной контактной карты с предварительно заданными хромосомными перестройками.

1. Lieberman-Aiden E. et al. Comprehensive mapping of long range interactions reveals folding principles of the human genome. // Science. – 2009. – N. 326. – P. 289-293.

2. Dixon J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. // Nature. – 2012. – N. 485. – P. 376–380.

3. Vietri Rudan M. et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture // Cell Rep. – 2015. – N. 10. – P. 1297-1309.

4. Battulin N. et al. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. // Genome Biology. – 2015. – N. 16. – S. 77.

5. Fraser J. et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation // Molecular Systems Biology. – 2015. – N. 11. – S 852.

6. Jackson M. et al. The genetic basis of disease // Essays in Biochemessry. – 2018. – V. 62. – I. 5. – P. 643–723.

7. Lupiáñez D.G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions // Cell. – 2015. – V. 161 – I. 5. – P. 1012-1025.

8. Anania C. et al. In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation // Nature Genetics. – 2022. – V. 54. – P. 1026-1036.

9. Harewood L. et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours // Genome Biology. – 2017. – V. 18. – I. 1. - P. 125.

5.3. Подробное описание работы, включая используемые алгоритмы

В основе данного ПО лежит алгоритм перекартирования индивидуальных контактов (или иной величины, вычисленной на их основе) между геномами. На первом этапе, на основе референсных карт контактов, происходит статистическое описание пространственной организации хроматина, которая включает в себя определение зависимости частоты контактов от геномного расстояния между ними, обогащённость индивидуальных контактов по сравнению со среднем и обогащённость отдельных локусов контактами. На втором этапе

алгоритму подаётся карта синтении, указывающая соответствие между локусами референсного и перестроенного генома. На третьем этапе с использованием карт синтении и статистики референсной карты контактов строится модель.

Основная сложность проведения такого рода перекартирования является биновый характер данных Hi-C: для получения достоверных результатов необходимо объединять все контакты на протяжённом участке генома — бине, чья длина для большей части экспериментов составляет тысячи и десятки тысяч пар оснований.

Чтобы учесть эти особенности, сравниваемые геномы разбиваются на короткие фрагменты (~200 п.н.), между которыми и устанавливается соответствие. Разбиение происходит таким образом, чтобы полученные фрагменты не пересекали границы бинов и не перекрывались друг с другом. Предполагается, что каждый такой фрагмент вносит свой вклад в наблюдаемую частоту контактов, на основе чего и рассчитывается моделированный контакт. В случае необходимости, с помощью биномиального распределения рассчитанная величина контакта модифицируется, для имитации шума.

5.4. Полученные результаты

На первом этапе данный алгоритм был использован для сравнения пространственной организации хроматина *Homo sapiens, Mus musculus* и *Gallus gallus,* описанных методом Hi-C. Результаты сравнения показали высокую консервативность архитектуры хроматина, прослеживаемую на уровне отдельных контактов. Так коэффициент корреляции при сравнении *Homo sapiens* и *Mus musculus* достигает 0.56 и 0.38 для сравнения *Homo sapiens* и *Gallus gallus*.

Аналогичные результаты были получены при исследовании архитектуры хроматина комаров рода *Anopheles.* Примечательно, что даже между самыми далёкими видами, эволюционное расстояние между которыми сравнимо с расстоянием между *Homo sapiens* и *Mus musculus,* коэффициент корреляции между картами контактов составлял от 0.8 до 0.95.

Применение данного алгоритма для моделирования известных хромосомных перестроек с имитацией шума показало, что уровень сходства/различия моделированных и реальных данных соответствует уровню сходства/различия между репликами. Таким образом, разрабатываемый алгоритм может быть использован для генерирования карт контактов с заранее определёнными хромосомными перестройками для последующего тестирования и оптимизация программ, ответственных за автоматизированный поиск перестроек.

6. Эффект от использования кластера в достижении целей работы

Обработка данных по пространственной организации хроматина требует больших объёмов физической и оперативной памяти, что делает её невозможной на обычных рабочих станциях. Использование ресурсов кластера является определяющим для достижения целей работы.

7. Перечень публикаций, содержащих результаты работы

1) Miroslav Nuriddinov, Veniamin Fishman. C-InterSecture—a computational tool for interspecies comparison of genome architecture// Bioinformatics. – 2019. – V. 35. – I. 23. – P. 4912–4921. https://doi.org/10.1093/bioinformatics/btz415

2) Varvara Lukyanchikova, Miroslav Nuriddinov, Polina Belokopytova, Jiangtao Liang, Maarten J.M.F. Reijnders, Livio Ruzzante, Robert M. Waterhouse, Zhijian Tu, Igor V. Sharakhov, Veniamin Fishman. Anopheles mosquitoes revealed new principles of 3D genome organization in insects // Nature Communications – 2022. – V. 13. – I. 1 – P. 1960. https://doi.org/10.1038/s41467-022-29599-5. https://doi.org/10.1038/s41467-022-29599-5

3) Nuriddinov M., Mozheiko E., Fishman V. Simulating of 3D genome data with predefined chromosomal rearrangements // Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022): The Thirteenth International Multiconference (04–08 July 2022, Novosibirsk, Russia); Abstracts / Institute of Cytology and Genetics, the Siberian Branch of the Russian Academy of Sciences. - 2022. – P. 122.