Отчёт о проделанной работе с использованием оборудования ИВЦ НГУ
1. Аннотация.
Проведена разработка алгоритма для моделирования данных 3D-геномики с предопределёнными хромосомными перестройками и реализованного в виде ПО Charm (https://github.com/genomech/Charm/) В основе данного ПО лежит пересчёт пространственных контактов с референсного генома на геном, несущий перестройки, с учётом влияние перестроек на геномных расстояний и копийности отдельных локусов генома. Программа Charm включает в себя инструментарий для описания перестроек таких как инверсии, транслокации и вариации числа копий, создание псевдореплик, а также тестирования различных моделей зависимости числа контактов от известных характеристик генома.
Проведённые тесты показывают, что построенные с помощью Charm модели контактов имеют сходства с реальными данными на уровне биологических реплик (коэффициент корреляции Пирсона ~0.6 – 0.9). Также разработанное ПО одинаково эффективно работает как не обогащёнными данными Hi-C эксперимента, так и с его различными модификациями, такими как exone capture и promoter capture Hi-C.
С помощью программы Charm была успешно создана базы перестроек для тестирования алгоритмов поиска перестроек.

2. Тема работы Моделирование 3D геномных данных с предопределёнными хромосомными перестройками

3. Состав коллектива Нуридинов Мирослав Абдурахимович, канд.биол. наук 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, №22-14-00247, Разработка новых подходов для исследования механизмов пространственной организации хроматина и их функционального значения в регуляции генной экспрессии у животных, 2022-2024, руководитель: Фишман Вениамин Семенович.

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. Большое количество нулевых контактов на картах Hi-C, особенно на разрешениях глубже 50 тысяч п.о. создаёт ряд проблем для моделирования. Во-первых, из-за дискретного характера данных Hi-C наличие нулевого контакта не означает, что вероятность формирования контактов между целевыми локусами равна нулю. Однако оценить эту вероятность напрямую из данных нельзя, что особенно критично для межхромосомных перестроек. Во-вторых, моделирование нулевых контактов значительно увеличивает число пар локусов для которых нужно провести расчёты, требуемые ресурсы и процессорное время растут квадратично от глубины разрешения моделируемой карты. Так моделирование перестройки на разрешение 5 тысяч п.о. требует в 100 раз больше времени, чем моделировании на разрешение 50 тысяч п.о. В-третьих, отказ от моделирования нулевых контактов приводит к «утечке» референсных данных в модель, что приводит к неоправданному завышению её качества. Указанные проблемы были решены следующим образом. Во-первых, а основе статистик известных по референсу, проводится оценка вероятности контакта для нулевых контактов, на основе которой и проводится моделирование. Во-вторых, само моделировании происходит в две стадии: сначала создаётся карта контактов на крупном масштабе, после чего для всех пар локусов, для которых предсказано не нулевое число контактов, проводится перемоделирование на более мелком масштабе. Данный подход позволил эффективно решить задачи по моделированию всех классов перестроек, в том числе межхромосомных.

5.4. Полученные результаты.

Разработанная программа была валидирована с использованием данных whole genome Hi-C (wgHi-C) для клеточных линий IMR90 и K562; promoter capture Hi-C (pcHi-C) для IMR90, LG1 и LG2; и exone capture Hi-C для K562 и клеток переферической крови человека.

Сравнение карт контактов для IMR90, LG1 и LG2 и моделей для IMR90 и LG1 показало, что корреляция (по Пирсону) карт Hi-C для разных линий клеток находится на уровне ~0.4-0.5, для биологических реплик на уровне ~0.5-0.7 и ~0.7-0.9 для технических реплик. Модели карт контактов, построенные на основе разных линий клеток отличаются друг от друга также сильно, как референсные карты контактов друг от друга, при этом они демонстрируют сходство с моделируемым клеточным типом на уровне биологических и технических реплик.

На следующем шаге была проведено моделирование подтверждённых экспериментально структурных вариаций присутствующих в геноме клеток линии K562. В качестве референса выступила другая клеточная линия — IMR90 – не имеющей существенного количества перестроек. Модели «дикого типа» имели корреляцию с наблюдаемыми перестройками около ~0.3-0.5, что даже меньше, чем между разными клеточными типами. В то же время моделированные перестройки показали корреляцию с реальными на уровне ~0.6-0.7, что соответствует уровню различий между разными клеточными типами и подтверждает точность предсказания паттерна контактов, возникающем при перестройка. Стоит отметить, что эти результаты были получены при сравнении карт контактов на разрешении 50kb, что находится на нижнем уровне чувствительности наиболее популярных алгоритмов предсказании перестроек.
Таким образом, разработанное ПО Charm позволяет с высокой точностью моделировать структурные вариации, и использовать полученные модели для тестирования алгоритмов предсказания структурных вариаций.

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) 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.
2) Maria M Gridina, Yana K Stepanchuk, Miroslav A Nurridinov, et al. Modification of the Hi-C Technology for Molecular Genetic Analysis of Formalin-Fixed Paraffin-Embedded Sections of Tumor Tissues // Biochemistry (Mosc) - 2024 – V. 89(4). – P. 637-652.