- **Тема работы**
  Разработка вычислительной платформы для оценки точности алгоритмов, предсказывающих пространственную архитектуруы генома.
- **Состав коллектива: ФИО без сокращений, место работы/учёбы, учёные степени и звания. Опционально контактный адрес электронной почты.**
  Белокопытова Полина Станиславовна, ИЦиГ СО РАН, НГУ, belka2195@mail.ru
  Фишман Вениамин Семенович, ИЦиГ СО РАН, НГУ, кбн
  Валеев Эмиль
- **Научное содержание работы:**
  1. *Постановка задачи.*

Целью работы является создание вычислительной платформы для сравнения алгоритмов, предсказывающих пространственную организацию хроматина. Для этого было необходимо проанализировать литературу и единообразно обработать большой набор cHi-C и ChIP-seq данных. Далее необходимо разработать и реализовать набор метрик для сравнения точности предсказания алгоритмов между собой и оформить полученные метрики в виде вычислительной платформы.

2. *Современное состояние проблемы (на момент начала работы).*

По мере развития экспериментальных техник изучения трёхмерной укладки хроматина и увеличения количества работ, связанных с функциональным изучением трёхмерной организации хроматина, начало появляться все больше алгоритмов для предсказания 3D архитектуры генома, основанных на физических или статистических методах моделирования. Многие модели имеют возможность предсказывать не только трёхмерную архитектуру хроматина в норме, но также изменения, происходящие в ней при хромосомных перестройках, что является особенно актуальным для поиска причин патологий, опосредованных генетикой. Однако среди большого количества разработанных алгоритмов достаточно трудно выбрать самый точный, поскольку все опубликованные работы по моделированию пространственной организации хроматина используют свои методы и свой набор данных и примеров для оценки качества предсказания моделей. Возможность выбрать лучшую модель для предсказания 3D архитектуры генома может быть актуальна не только с точки зрения медицинской генетики, но также для лучшего понимания биологических закономерностей, лежащих в основе трёхмерной организации генома. Поскольку в основе предсказательных моделей лежат разные биологические данные, заложены разные алгоритмы, сравнение моделей между собой позволяет выявить именно те биологические паттерны, которые приводят к формированию различных пространственных структур хроматина. Таким образом, создание платформы, где можно было бы сравнить разные алгоритмы между собой на одном наборе данных, является особенно актуальным.

3. *Подробное описание работы, включая используемые алгоритмы.*

На первом этапе был проведён анализ литературы для создания набора необходимых Hi-C данных. Поскольку некоторые алгоритмы имеют возможность предсказывать трехмерную организацию хроматина только для нормальных клеток, мы сделали два базовых набора данных и два типа оценки точности работы алгоритмов. Первый набор данных включает в себя Hi-C данные для 2 человеческих линий клеток K562, GM12878, мышиной линии эмбриональных стволовых клеток и дрозофилиной эмбриональной линии клеток Kc167 (набор данных доступен по адресу https://github.com/genomech/3DGenBench/blob/stable/whole_genome_regions.txt). В этом

варианте анализа предлагается предсказывать трехмерную организацию локусов размером около 20 Мб в нормальных клетках без учета эффектов хромосомных перестроек. Поскольку многим алгоритмам необходимы эпигенетические данные для работы, была создана сводная таблица со ссылками для скачивания наиболее используемых эпигенетических меток и сайтов связывания транскрипционных факторов в форматах bed и bigwig для используемых типов клеток (https://github.com/genomech/3DGenBench/blob/stable/epigenetics_data.txt).

Второй набор данных включает в себя пары Hi-C карт для нормальных и перестроенных геномов, что позволяет оценить возможность алгоритмов предсказывать изменения трёхмерной структуры хроматина при хромосомных перестройках. Мы включали в этот набор только capture Hi-C (cHi-C) данные, так как такие данные имеют высокое разрешение и чаще всего исследователи предпочитают проводить именно такой вариант эксперимента для описания архитектуры хроматина перестроенных районов. В результате был создан набор данных, состоящий из 49 парных cHi-C карт, описывающих хроматин в клетках дикого типа и после различных мутаций. Собранные данные основаны на 9 исследованиях [Bianco и др., 2018; Despang и др., 2019; Franke и др., 2016; Hanssen и др., 2017; Kragesteen и др., 2018; Paliou и др., 2019; Rodríguez-Carballo и др., 2017] проведенных с 2016 по 2019 годы, и описывают 16 клеточных линий (https://github.com/genomech/3DGenBench/blob/stable/rearrangements_table.tsv). Для всех типов клеток были обработаны ChIP-seq данные, описывающие профиль связывания белка CTCF. Каждый локус для предсказания трехмерной архитектуры хроматина имеет размер около 3 Мб или больше, в зависимости от размеров хромосомной перестройки. Все Hi-C данные были обработаны Валеевым Эмилем в соответствии со стандартным протоколом обработки Hi-C данных Juicer [Durand и др., 2016] и с использованием нормализации C-TALE [Golov и др., 2020] для cHi-C данных.

Следующий этап работы включал в себя разработку метрик для оценки точности предсказания для двух типов сравнения. Все скрипты были написаны на python.

Метрики для «горизонтального» типа сравнения

Для «горизонтального» типа сравнения мы использовали коэффициент корреляции Спирмана, посчитанный между предсказанными и экспериментальными Hi-C частотами контактов. Однако Hi-C матрицы контактов обладают своими особенностями, в частности во всех Hi-C картах очень хорошо прослеживается тенденция к падению частоты контактов в зависимости от геномного расстояния, что в итоге приводит к высокой корреляции. Поэтому другой метрикой, которую мы используем для оценки точности предсказания, является SCC [80].

Кроме общего сравнения двух матриц частот контактов друг с другом, важно понимать, насколько хорошо предсказываются конкретные биологические структуры, такие как, например, ТАДы. Для этой цели мы получили профиль инсуляции для экспериментальной Hi-C карты и предсказанной Hi-C карты, затем использовали корреляцию Спирмана для корреляции этих величин.

Другой важной структурой Hi-C карт являются компартменты. Мы использовали метрику, отражающую силу компартментализации каждого бина, и считали корреляцию Спирмана между силой компартментализации каждого бина в предсказанной и экспериментальной Hi-C матрице контактов. Сила компартментализации считалась также, как было предложено в [Falk и др., 2019].

И последняя метрика для «горизонтального» типа сравнения оценивает то, насколько хорошо модели улавливают зависимость частоты контактов от геномного расстояния. Для этого считается средняя частота контактов на отдельных геномных расстояниях для экспериментальных и предсказанных данных. Полученные массивы значений сравниваются с использованием корреляции Спирмана.

Метрики для «вертикального» типа сравнения

Метрики для «вертикального» типа сравнения – это метрики, необходимые для оценки точности предсказания изменений, произошедших в пространственной организации генома вследствие хромосомной перестройки.

Во-первых, мы оценивали насколько меняется профиль инсуляции в случае мутации по сравнению с диким типом. Во-вторых, мы оценивали, насколько точно были предсказаны те частоты контактов, которые изменились больше всего вследствие мутации. Эктопические взаимодействия определялись как в [Bianco и др., 2018].

Для того чтобы оценить правильность работы предложенных метрик, нами был сгенерирован набор Hi-C карт, который смог бы являться некоторым базисом для сравнения. Мы выбрали несколько образцов из подготовленного набора данных и протестировали, как меняются значения метрик в зависимости от количества шума в данных (Рис. 2 А, Б).

Создание такого базиса для сравнения является особенно полезным, поскольку для большинства типов клеток имеется только одна реплика и сравнить сходство эксперимента и предсказания с уровнем схожести Hi-C карт между репликами невозможно. Такой набор данных позволяет оценить, насколько значения метрик, полученные при сравнении предсказания алгоритма и экспериментальных данных, высокие или низкие.

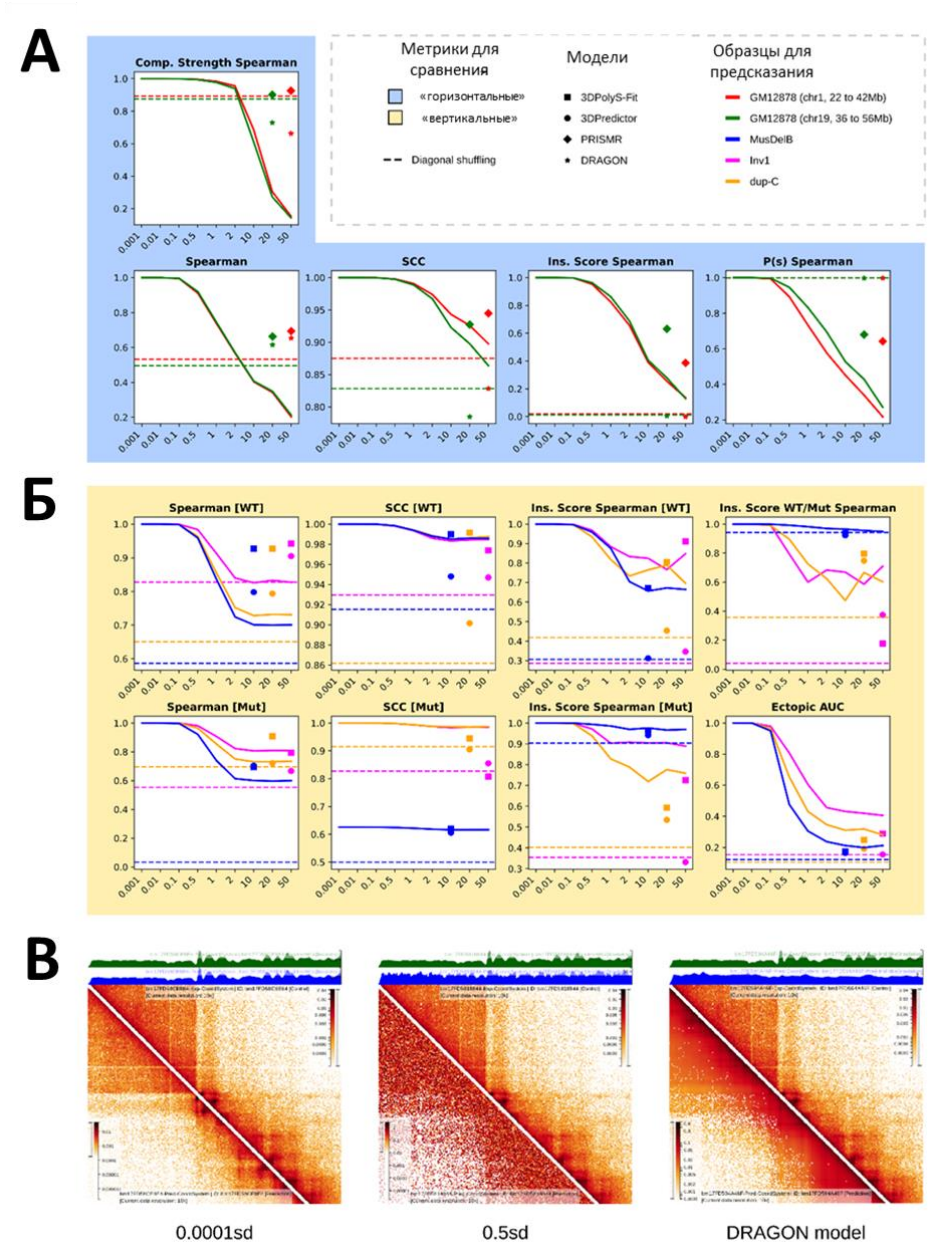

Рис. 2. Разработанные метрики отражают различия данных, предсказанных разными алгоритмами. (А) Все метрики «горизонтального» типа сравнения, посчитанные для 2 разных локусов (красный и зелёный цвет) для 2 моделей (PRISMR и DRAGON) (круг и ромб). Кривые отражают зависимость значения метрик от уровня шума в данных. Чем значение по оси X выше, тем больший уровень шума присутствует в Hi-C данных.. Прерывистые линии отражают значение метрики для Hi-C карты с перемешанными значения на диагоналях (Б) Все метрики «вертикального» типа сравнения, посчитанные для 3 разных типов перестроек для 2 моделей (3DPredictor и 3DPolyS-Fit). Кривые обозначают то же самое, что и в (А). (В) Визуализация данных с низким уровнем шума, с высоким уровнем шума, предсказание модели DRAGON. Везде снизу предсказание, сверху экспериментальные данные. Сверху зеленый трек отражает профиль инсуляции экспериментальных данных, синиий трек – профиль инсуляции предсказанной Hi-C карты.

На рисунке 2 А, Б видно, что значения всех метрик снижаются в соответствии с уровнем сгенерированного шума, что является ожидаемым и показывает, что предложенные метрики адекватно отражают сходство Hi-C карт.

Мы проверили применимость разработанных метрик на конкретных примерах с использованием таких алгоритмов как PRISMR [Bianco и др., 2018], DRAGON [Qi, Zhang, 2019], 3DPolyS-Fit [Szabo и др., 2018] и разработанного нами инструмента 3DPredictor. Для демонстрации возможности использования метрик, разработанных для «вертикального» сравнения, мы получили от коллег из группы Daniel Jost предсказания архитектуры хроматина для трех разных типов хромосомных перестроек (инверсия, делеция и дупликация), сделанных алгоритмом 3DPolyS-Fit [Szabo и др., 2018]. Мы сгенерировали предсказания архитектуры хроматина для этих же локусов, используя разработанный нами алгоритм 3DPredictor, и сравнили полученные модели. Приведенные примеры наглядно показывают, что созданная система метрик и набор данных являются хорошим инструментом для сравнения разных алгоритмов между собой.

4. *Полученные результаты.*

Мы получили 2 набора данных для 2 основных типов сравнения алгоритмов. Мы определили тип сравнения, отвечающий на вопрос, насколько хорошо алгоритмы предсказывают Hi-C карту контактов по сравнению с экспериментальными данными, как «горизонтальный». Тип сравнения, показывающий насколько хорошо модели предсказывают изменения в трёхмерной организации генома, произошедшие при хромосомной перестройке, мы определили как «вертикальный» (Рис. 1).

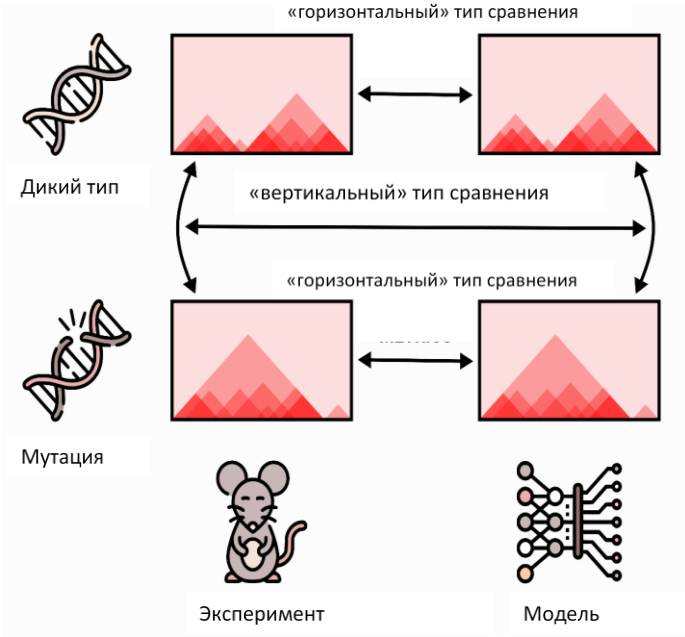

Рис. 1. Разные типы сравнения предсказанных и экспериментальных Hi-C карт.

Все метрики были реализованы на python (https://github.com/genomech/3DGenBench). Для того, чтобы разработчики моделей по предсказанию трёхмерной архитектуры генома могли использовать унифицированные метрики для оценки качества предсказаний, мы

разработали web-платформу 3DGenBench, которую сможет использовать любой желающий. Разработанный онлайн-ресурс позволяет получить значения всех метрик, описанных выше, в удобном для пользователя формате с визуализацией предсказанных и экспериментальных данных в HiGlass (Рис. 3).

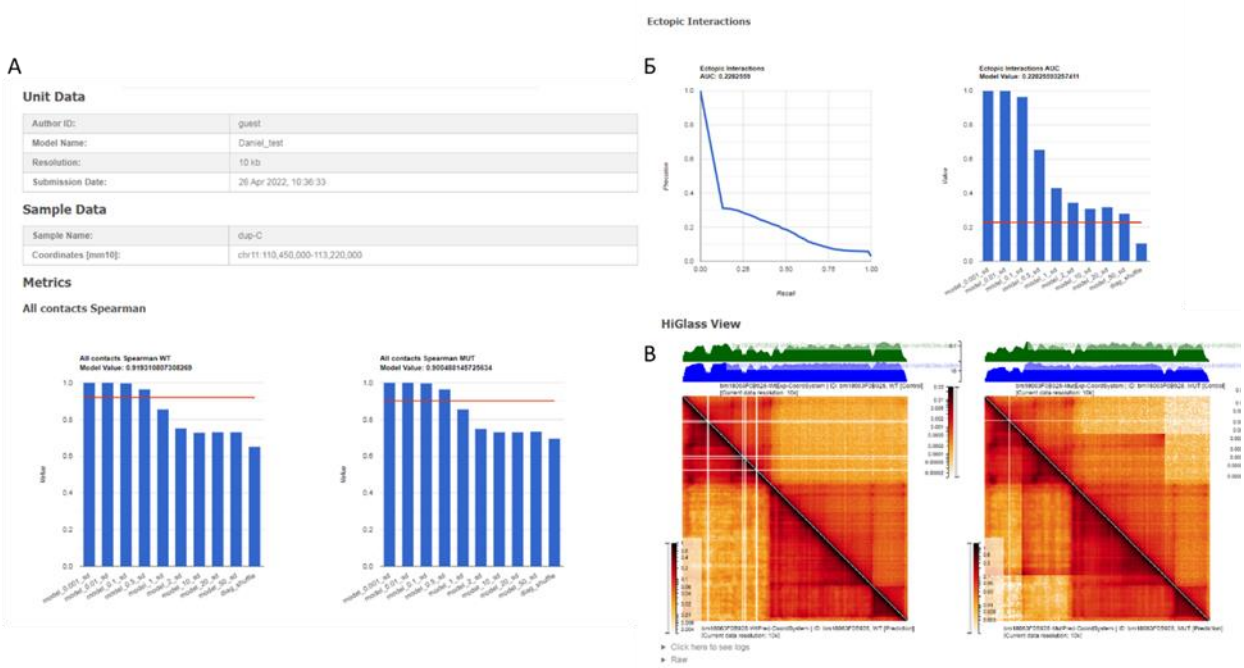

Рис. 3. Фрагменты визуализации метрик с сайта 3DGenBench. (А) Описание выбранного локуса и значение корреляции Спирмана по сравнению с базисом (данные с разным уровнем шума). (Б) Пример визуализации метрики, отражающей точность предсказанных эктопических взаимодействий. (В) Пример визуализации предсказанных и экспериментальных данных для мутантного и дикого типа на сайте.

Созданная вычислительная платформа 3DGenBench является инструментом для сравнения моделей, предсказывающих 3D архитектуру генома между собой. Появляются новые алгоритмы и идёт активное изучение механизмов, лежащих в основе пространственной организации генома, появляются новые случаи, доказывающие функциональную значимость пространственной организации хроматина. В этих условиях платформа для унифицированной оценки точности работы алгоритмов является особенно актуальной.

Список, процитированной литературы в разделах 3,4:

1. Bianco S. и др. Polymer physics predicts the effects of structural variants on chromatin architecture // Nat. Genet. 2018. Т. 50. № 5. С. 662–667.

2. Despang A. и др. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture // Nat. Genet. 2019. Т. 51. № 8. С. 1263–1271.

3. Durand N.C. и др. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments Tool Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments // Cell Syst. 2016. Т. 3. С. 95–98.

4. Falk M. и др. Heterochromatin drives compartmentalization of inverted and conventional

nuclei // Nature. 2019. Т. 570. № 7761. С. 395–399.

5. Franke M. и др. Formation of new chromatin domains determines pathogenicity of genomic duplications // Nature. 2016. Т. 538. № 7624. С. 265–269.

6. Golov A.K. и др. C-TALE, a new cost-effective method for targeted enrichment of Hi-C/3C-seq libraries // Methods. 2020. Т. 170. С. 48–60.

7. Hanssen L.L.P. и др. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo // Nat. Cell Biol. 2017. Т. 19. № 8. С. 952–961.

8. Kragesteen B.K. и др. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis // Nat. Genet. 2018. Т. 50. № 10. С. 1463–1473.

9. Paliou C. и др. Preformed chromatin topology assists transcriptional robustness of Shh during limb development // Proc. Natl. Acad. Sci. 2019. Т. 116. № 25. С. 12390–12399.

10. Qi Y., Zhang B. Predicting three-dimensional genome organization with chromatin states // PLoS Comput. Biol. 2019. Т. 15. № 6.

11. Rodríguez-Carballo E. и др. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes // Genes Dev. 2017. Т. 31. № 22. С. 2264–2281.

12. Szabo Q. и др. TADs are 3D structural units of higher-order chromosome organization in Drosophila // Sci. Adv. 2018. Т. 4. № 2.

- **Эффект от использования кластера в достижении целей работы.**

Многие биоинформационные программы для обработки сырых Hi-C данных требуют большое количество оперативной памяти. Кроме того, обработка большого количества данных требует большого количества вычислительных ресурсов для эффективной работы. Таким образом, работа с таким массивом данных на персональных компьютерах является практически невозможной..

- **Перечень публикаций, содержащих результаты работы**

1. Belokopytova P, Fishman V. Predicting Genome Architecture: Challenges and Solutions. Front Genet. 2021 Jan 22;11:617202. doi: 10.3389/fgene.2020.617202. PMID: 33552135; PMCID: PMC7862721.

2. Belokopytova, P., Viesná, E., Chiliński, M., Qi, Y., Salari, H., di Stefano, M., Esposito, A., Conte, M., Chiariello, A. M., Teif, V. B., Plewczynski, D., Zhang, B., Jost, D., & Fishman, V. (2022). 3DGenBench: a web-server to benchmark computational models for 3D Genomics. Nucleic Acids Research, 50(W1), W4–W12. https://doi.org/10.1093/nar/gkac396