На правах рукописи

БЕЛОКОПЫТОВА ПОЛИНА СТАНИСЛАВОВНА

РАЗРАБОТКА И ОЦЕНКА ТОЧНОСТИ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ ТРЕХМЕРНОЙ УКЛАДКИ ХРОМАТИНА МЛЕКОПИТАЮЩИХ

1.5.8 – математическая биология, биоинформатика
(биологические науки)

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата биологических наук

Новосибирск - 2023

Работа выполнена в секторе геномных механизмов онтогенеза ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения российской академии наук» и в лаборатории структурно-функциональной организации генома «Новосибирский национальный исследовательский государственный университет», г. Новосибирск.

Научный руководитель: **Фишман Вениамин Семенович**, кандидат биологических наук, заведующий сектором геномных механизмов онтогенеза, ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск

Официальные оппоненты: **Кулаковский Иван Владимирович**, доктор биологических наук, ведущий научный сотрудник группы регуляции биосинтеза белка, Институт белка РАН г.Москва

Попцова Мария Сергеевна, кандидат физико-математических наук, доцент департамента больших данных и информационного поиска ФКН ВШЭ, заведующая международной лабораторией биоинформатики НИУ ВШЭ, г. Москва

Ведущая организация: Институт биологии гена РАН, г. Москва

Защита диссертации состоится «___» _____ 20___ г. на утреннем заседании Диссертационного совета 24.1.239.01 на базе ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики СО РАН» в конференц-зале Института по адресу: 630090, г. Новосибирск, проспект ак. Лаврентьева, 10, т. (383)363-49-06, факс (383) 333-12-78, e-mail: dissov@bionet.nsc.ru.

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института http://www.icgbio.ru/

Автореферат разослан «___»_____20___г.

Ученый секретарь
диссертационного совета,
доктор биологических наук Т.М. Хлебодарова

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы. Для обеспечения функционирования генома требуется точная работа различных регуляторных механизмов. В частности, нужны механизмы для поддержания необходимого уровня экспрессии генов. Многочисленные исследования показывают, что укладка хроматина в пространстве ядра вносит важный вклад в регуляцию геномных процессов. Исследования, связанные с изучением трёхмерной архитектуры хроматина, сейчас являются актуальными и активно развивающимися. Например, было показано, что хромосомные перестройки, приводящие к нарушению пространственных контактов хроматина, могут служить причиной развития патологий. К настоящему времени было опубликовано достаточное количество работ, показывающих важность механизмов, обеспечивающих пространственную организацию генома, в регуляции экспрессии генов [Barutcu и др., 2018; Franke и др., 2016a; Hanssen и др., 2017; Kraft и др., 2019; Kragesteen и др., 2018a; Rodríguez-Carballo и др., 2017].

В основном пространственную организацию хроматина изучают при помощи экспериментальных методов, основанных на технологии захвата хромосом, которые позволяют получить информацию о локусах генома, находящихся близко в пространстве. Кроме того, на сегодняшний день активно применяются методы машинного обучения и физического моделирования как для предсказания 3D структуры генома, так и для изучения биологических закономерностей, лежащих в ее основе. Таким образом, использование современных методов из разных областей знания позволяет с разных сторон взглянуть на процессы и механизмы, происходящие с хроматином внутри ядра клетки.

Данная работа состоит из двух основных частей. Первая часть посвящена использованию методов машинного обучения для предсказания трёхмерной организации геномов млекопитающих на основе эпигенетических данных, а также применению этого метода для предсказания последствий хромосомных перестроек.

Вторая часть работы посвящена разработке вычислительной платформы для оценки точности алгоритмов по предсказанию 3D архитектуры хроматина в нормальных и перестроенных геномах.

Основной **целью** представленной диссертационной работы является разработка алгоритма для предсказания пространственной организации хроматина и создание инструмента для оценки точности таких предсказательных моделей. Для достижения поставленной цели были сформулированы следующие **задачи**:

1. Оценить возможность применения алгоритма TargetFinder для предсказания промотор-энхансерных взаимодействий.
2. Разработать инструмент 3DPredictor, основанный на машинном обучении, для предсказания пространственной архитектуры генома млекопитающих.
3. Оценить точность реконструкции пространственной архитектуры хроматина алгоритмом 3DPredictor для разных типов клеток человека и мыши.
4. Оценить точность моделирования изменений трехмерной организации хроматина, вызванных хромосомными перестройками, на основе алгоритма 3DPredictor.
5. На основе анализа опубликованных экспериментальных работ создать набор Hi-C данных для модельных клеточных линий животных дикого типа и с различными хромосомными перестройками.
6. Разработать метрики для единообразной оценки точности предсказаний 3D организации хроматина.
7. Разработать программное обеспечение, позволяющее оценить точность предсказания 3D архитектуры генома.

Научная новизна. Нами был разработан инструмент 3DPredictor для предсказания пространственной организации хроматина. Таким образом, впервые был предложен алгоритм, основанный на градиентном бустинге, который способен предсказывать Hi-C карту контактов, используя в качестве входных данных такие характеристики хроматина как ChIP-seq белка CTCF, информацию о транскрипционной активности и расстояние между геномными локусами. Кроме того, мы впервые показали, что такой инструмент можно использовать для предсказания изменений в пространственной организации хроматина, произошедших в результате хромосомных перестроек.

В последние годы количество алгоритмов, способных предсказывать трёхмерную организацию хроматина в норме и при различных мутациях, значительно выросло. Такие инструменты могут быть полезны в медицинской генетике, однако для этого необходимо иметь возможность сравнивать алгоритмы между собой, чтобы выбрать наиболее подходящий для поставленных задач. Для этой цели нами впервые был собран и единообразно процессирован большой набор cHi-C данных для нормальных и перестроенных геномов, включающий 49 различных случаев хромосомных перестроек. Такой набор данных может служить референсом для сравнения алгоритмов между собой. Для того чтобы было удобно сравнивать алгоритмы между собой, нами была разработана вычислительная платформа 3DGenBench, аналогов которой не существует на текущий момент.

Теоретическая и практическая значимость исследования. На сегодняшний день далеко не для всех типов клеток получена информация о пространственной организации хроматина, однако эпигенетические данные, в том числе ChIP-seq за белок CTCF и информация о транскрипционной активности генов, являются доступными и широко распространёнными для разных типов клеток. Разработанный нами алгоритм 3DPredictor позволяет предсказывать 3D организацию хроматина для таких типов клеток. Кроме того, наш алгоритм имеет возможность предсказывать изменения трёхмерной организации хроматина, произошедшие при хромосомной перестройке. Это позволяет предположить, как изменится экспрессия генов, что может быть интересно в клинике для объяснения патологий, вызванных хромосомными перестройками. В связи с тем, что за последние годы число таких моделей растёт, мы разработали платформу 3DGenBench, которая может быть полезна при выборе алгоритма для предсказания пространственной архитектуры генома в норме и при мутации. Кроме того, возможность единым образом оценивать точность алгоритмов для предсказания трёхмерной укладки хроматина позволяет обнаружить слабые места каждого алгоритма и понять какие признаки и механизмы являются наиболее значимыми для пространственной организации генома.

Основные положения, выносимые на защиту:

1. Разработан инструмент 3DPredictor, который позволяет, на основе информации о транскрипционной активности, распределении белка CTCF и локализации его сайтов связывания в геноме, выявлять клеточно-специфичные особенности трёхмерной архитектуры генома и предсказывать изменения пространственных контактов хроматина, вызванные хромосомными перестройками.

2. Вычислительная платформа 3DGenBench, разработанная на основе сравнения матриц пространственных контактов хроматина, позволяет проводить оценку точности предсказательных моделей укладки хроматина в клетках животных.

Личный вклад автора. Автором была написана большая часть кода для работы алгоритма 3DPredictor на языке python. Некоторые скрипты для алгоритма 3DPredictor были написаны Фишманом В.С. (ИЦиГ СО РАН). Все ChIP-seq и RNA-seq данные были обработаны автором. Набор промотор-энхансерных взаимодействий был подготовлен Нуриддиновым Мирославом (ИЦиГ СО РАН). cHi-C данные для базы данных платформы 3DGenBench были обработаны Валеевым Эмилем (ИЦиГ СО РАН, Новосибирск). Вся серверная часть для сайта 3DGenBench была написана автором на языке python. Часть кода,

необходимая непосредственно для работы сайта, была написана Валеевым Эмилем (ИЦиГ СО РАН, Новосибирск) на языке php.

Апробация работы и публикации. Научные результаты, изложенные в данной работе, были представлены на нескольких международных конференциях в виде стендовых и устных докладов. А именно: Interdisciplinary school in 3D genomics: from experiments to models and back, Lyon, France (online), 23 - 25 ноября 2020; SBB - 2020, Ялта, РФ, 14 - 20 сентября 2020; МНСК-2019, Новосибирск, РФ, 14 - 19 апреля 2019; XVIII Конференция - школа с международным участием "Актуальные проблемы биологии развития", Москва, РФ, 14 – 19 октября 2019; Chromosomes and mitosis. International mini-conference, Новосибирск, РФ, 21 ноября 2019; МНСК-2018, Новосибирск, РФ, 22 апреля 2018. По теме диссертации было опубликовано 3 работы. Основные результаты были изложены в рецензируемых журналах Genome Research и Nucleic Acid Research.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, выводов, списка литературы и приложений. Работа изложена на 118 страницах, проиллюстрирована 30 рисунками, содержит 4 таблицы и 4 приложения.

СОДЕРЖАНИЕ РАБОТЫ

Глава 1. Обзор литературы

Данная глава подробно освещает основные механизмы, лежащие в основе трёхмерной организации хроматина, и экспериментальные методы ее исследования. Также разобраны основные подходы и принципы *in silico* моделирования трехмерной архитектуры генома. Отдельное внимание уделено функциональной роли 3D укладки хроматина и алгоритмам, способным предсказывать последствия хромосомных перестроек, обусловленные изменениями в пространственной организации хроматина. Уже сейчас открытия, сделанные в сфере 3D геномики, имеют интерес не только для фундаментальной науки, но также имеют прикладной характер и применяются, например, в медицинской генетике.

Глава 2. Материалы и методы.

Для подготовки данной работы использовались различные биоинформационные программы и языки программирования. Весь основной код написан на языке python, однако для части задач, в частности для анализа RNA-seq данных использовался язык программирования R. Были освоены и

использованы различные пайплайны, программы и методы для анализа таких данных как RNA-seq, ChIP-seq, Hi-C. Кроме того, активно использовались python библиотеки для машинного обучения и анализа больших данных. Множество скриптов для анализа данных было написано самостоятельно на языке программирования python.

Главы 3-4. Результаты и обсуждение.

Применение и анализ алгоритма TargetFinder для предсказания промотор-энхансерных взаимодействий.

Понимание того, как изменится экспрессия генов при разных мутациях и к каким последствиям это приведет – один из ключевых вопросов генетики. Как известно, экспрессия генов обусловлена взаимодействиями промоторов и энхансеров, и наша первоначальная задача состояла в создании инструмента, способного предсказывать изменения трехмерных контактов хроматина, в первую очередь промотор-энхансерных взаимодействий, сопровождающих хромосомные перестройки. На момент начала работы для этой задачи подходил только один алгоритм TargetFinder [Whalen, Truty, Pollard, 2016], опубликованный в журнале Nature Genetics. Алгоритм TargetFinder является бинарным классификатором, который используя информация о белках, связанных с промотором, энхансером и «окном» (участком ДНК) между ними качественно отвечает на вопрос о взаимодействии конкретной пары энхансер-промотор.

Мы применили алгоритм TargetFinder, использую доступные данные для различных типов клеток мыши. Однако обнаружили, что точность работы алгоритма намного ниже, чем было представлено в оригинальной работе. Проведя серию экспериментов, варьируя параметры входных данных и параметры алгоритма мы обнаружили, что различия в точности работы алгоритма обусловлены разными способами создания обучающей и валидационной выборки. В своей работе мы использовали разные хромосомы для генерирования обучающей и валидационной выборки. Таким образом наборы данных для обучения и валидации никогда не пересекались. В оригинальной же работе авторы делили весь набор энхансер-промоторных пар случайным образом на выборку для обучения (90% от всего набора) и валидации (10% от всего набора). Это является логичным с математической точки зрения, однако такой способ генерирования наборов для тренировки и теста приводит к переобучению алгоритма, так как в таком дизайне косвенно признаки между выборками пересекаются.

В итоге стало ясно, что алгоритм TargetFinder в существующем дизайне не

способен улавливать комплексные биологические закономерности между эпигенетическими характеристиками и трёхмерной архитектурой генома. Используя некоторые идеи, предложенные в работе об алгоритме TargetFinder, мы задались целью разработать альтернативный алгоритм для предсказания пространственной организации хроматина.

Разработка алгоритма 3DPredictor.

При разработке нового алгоритма, мы пересмотрели определение промотор-энхансерных взаимодействий и решили предсказывать их не качественно, а количественно. В связи с этим, целью следующего этапа нашей работы стала разработка инструмента, предсказывающего частоты контактов всех пар локусов, расположенных на расстояние до 1.5 Мб друг от друга. Расстояние 1.5 Мб было выбрано исходя из того, что промоторы и энхансеры, расположенные на больших расстояниях, как правило не взаимодействуют и основные топологические структуры такие как ТАДы и петли находятся в пределах этого расстояния.

Наша идея заключалась в том, чтобы, сопоставив эпигенетические данные и частоту контактов локусов, определить закономерности, на основе которых можно было бы количественно предсказывать вероятность взаимодействия локусов генома. В качестве алгоритма машинного обучения мы выбрали алгоритм XGBoost, основанный на ансамблях деревьев решений, где ошибка минимизируется алгоритмом градиентного спуска. В качестве функции потерь использовалась среднеквадратичная ошибка (MSE).

Мы обучали алгоритм, используя Hi-C карты на разном разрешении (5 Кб и 25 Кб). Для того, чтобы избежать переобучения алгоритма, были использованы непересекающиеся выборки для обучения и валидации модели. Чтобы выбрать лучший набор признаков для предсказания, мы варьировали набор используемых эпигенетических данных и оценивали точность работы алгоритма. В результате многих тестов, мы пришли к выводу, что точность алгоритма незначительно растет при добавлении к трем основным признакам (ChIP-seq для белка CTCF, включая ориентацию сайтов посадки CTCF, RNA-seq и геномное расстояние) каких-либо дополнительных. Конечная версия алгоритма включает только эти признаки (Рис. 1).

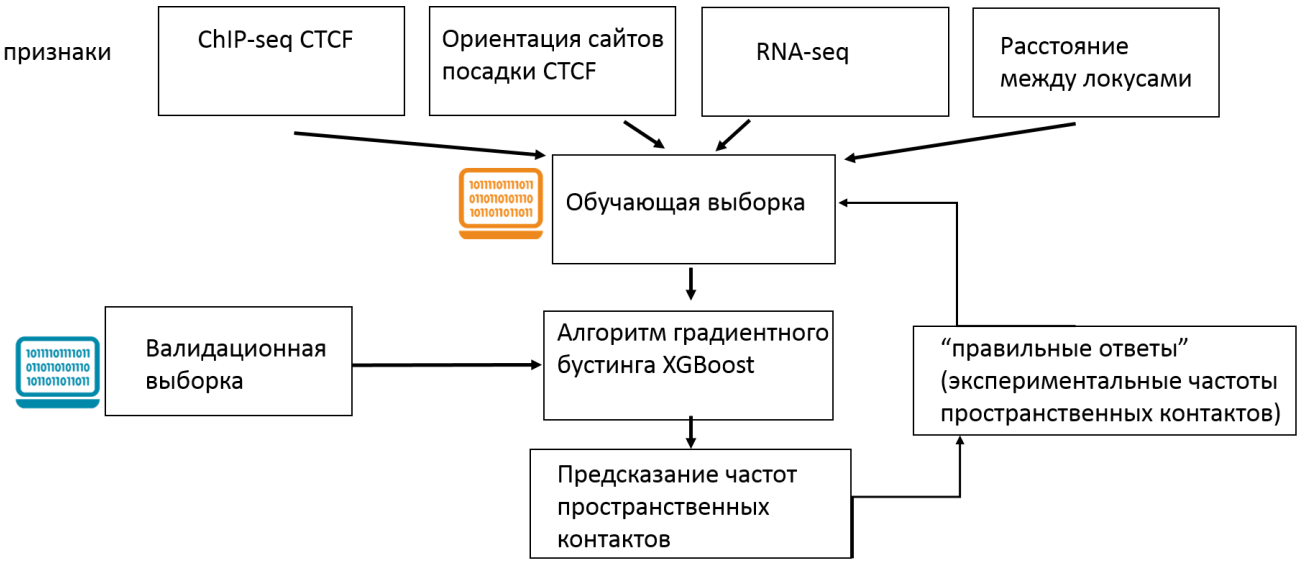

Рис. 1. Краткая схема работы алгоритма 3DPredictor.

Для того чтобы оценить точность работы алгоритма 3DPredictor нужно было выбрать и реализовать подходящие метрики. Задача состояла в том, чтобы понять насколько одна Hi-C карта «похожа» на другую. Метрика, применяемая для сравнения карт пространственных контактов хроматина – коэффициент корреляция Пирсона. Однако, коэффициент корреляции, посчитанный для всей предсказанной матрицы, не отражает точность предсказания конкретных топологических структур, таких как ТАДы или петли. Основная метрика, которую мы использовали – это метрика SCC (stratum-adjusted correlation coefficient) предложенная в работе [Yang и др., 2017] специально для сравнения Hi-C карт между собой. Она устраняет эффект зависимости частоты контактов от расстояния путем стратификации данных Hi-C в соответствии с их геномным расстоянием. Кроме того, для оценки точности алгоритма мы использовали такие стандартные метрики как средняя абсолютная ошибка (MAE), средняя квадратичная ошибка (MSE) и средняя относительная ошибка (MRE).

Абсолютные значения метрик мало о чём говорят, поэтому важно использовать некоторые базовые значения, относительно которых оценивается качество предсказания. Например, можно оценивать различия в предсказанной и экспериментальной Hi-C картах, относительно уровня различий между репликами. Однако реплики не всегда доступны. Другим базисом для сравнения может быть сравнение Hi-C карт между разными типами клеток.

Для обучения и оценки точности алгоритма мы использовали те типы клеток, для которых имелись в хорошем качестве Hi-C, ChIP-seq CTCF и RNA-seq данные. На рисунке 2 представлено предсказание алгоритма 3DPredictor, обученного на данных, сгенерированных для гепатоцитов мыши.

Алгоритм 3DPredictor даёт достаточно точные предсказания со значениями

метрик: корреляция Пирсона 0.92-0.95, SCC 0.53-0.72, MSE 0.0017-0.0082, MAE 0.0010-0.0015, MRE 0.52-1.74 (Рис. 2). В качестве базиса для сравнения использовались реплики гепатоцитов и другие типы клеток.

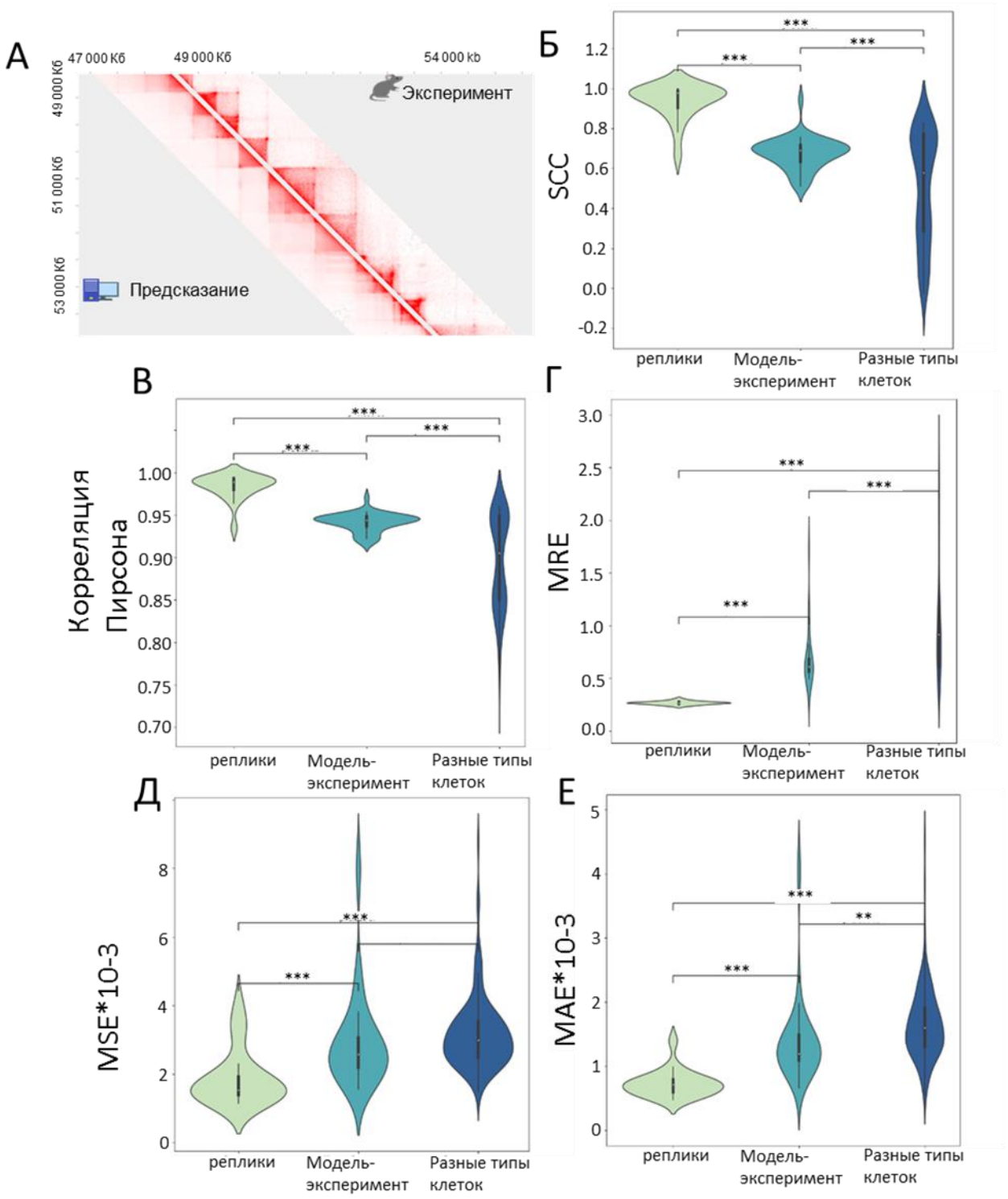

Рис. 2. Алгоритм 3DPredictor демонстрирует высокую точность предсказания пространственной организации генома для гепатоцитов мыши. Модель обучена на четных и нечетных хромосомах, выборка для предсказания не пересекается с обучающей выборкой. (А) Карта контактов предсказанная (снизу) и полученная экспериментально (сверху) для локуса chr2:47000000-55000000 на разрешении 5

Кб. (Б, В, Г, Д, Е) Значения метрик (MRE, SCC, MAE, корреляция Пирсона, MSE), полученные при сравнении Hi-C карт реплик между собой, предсказанных контактов гепатоцитов мыши с экспериментальными данными и пространственных контактов хроматина для разных типов клеток. Среднее значение получено как среднее из значений метрики для каждой хромосомы. В качестве статистического критерия использовался t-критерий Стьюдента.

Основные структуры трёхмерной организации генома достаточно консервативны между разными типами клеток, но есть небольшое количество регионов, являющихся клеточно-специфичными. В этих локусах границы топологических доменов значительно отличаются, как, например, в регионе 34000-36000 Кб на хромосоме 3 для мышиных гепатоцитов и клеток нейральных предшественников. Мы показали, что алгоритм 3DPredictor способен улавливать эти различия и предсказывать клеточно-специфичные изменения границ ТАДов.

Одной из областей применения алгоритма 3DPredictor является предсказание последствий хромосомных перестроек. Это позволяет понять, как меняются пространственные взаимодействия промотор-энхансерных пар после перестройки.

Мы использовали данные cHi-C [Lupiáñez и др., 2015], описывающие хромосомные перестройки в локусе *Epha4* мыши, чтобы выяснить, сможет ли 3DPredictor предсказать эктопические взаимодействия в перестроенном геноме. Мы анализировали данные, полученные на клетках дикого типа, а также на клетках, несущих гомозиготную делецию ~1,5 Мб, охватывающую ген *Epha4*. Эта делеция приводит к появлению эктопических контактов между геном *Pax3* и кластером энхансеров *Epha4*, что приводит к неправильной экспрессии гена *Pax3*, проявляясь в фенотипе как брахидактилия.

Мы использовали инструмент 3DPredictor, обученный на эпигенетических данных гепатоцитов мыши, чтобы предсказать трехмерную организацию перестроенного локуса *Epha4* для клеток задней конечности мыши. Мы не использовали какие-либо априорные знания о трехмерной организации локуса *Epha4* дикого типа в клетках задних конечностей, но результаты 3DPredictor были очень похожи на экспериментальные данные (Рис. 3 А). Мы использовали метод, описанный в [Bianco и др., 2018], чтобы найти эктопические контакты в перестроенном локусе на основе экспериментальных данных или предсказанной карты контактов хроматина. Из 1561 эктопических контактов, полученных на основе экспериментальных данных, 589 были предсказаны 3DPredictor, включая большинство взаимодействий между геном *Pax3* и энхансерами *Epha4* (Рис. 3 А, Б). Реальные и предсказанные эктопические взаимодействиями достаточно хорошо перекрываются и их пересечение значительно отличается от случайного

(P-значение < 5 × 10-6).

Подводя итог, мы разработали инструмент 3DPredictor, основанный на машинном обучении, способный количественно предсказывать пространственные взаимодействия локусов генома, включая взаимодействия промоторов и энхансеров. Для более широкого и удобного применения алгоритма 3DPredictor, мы разработали web версию инструмента (https://github.com/genomech/Web_3DPredictor).

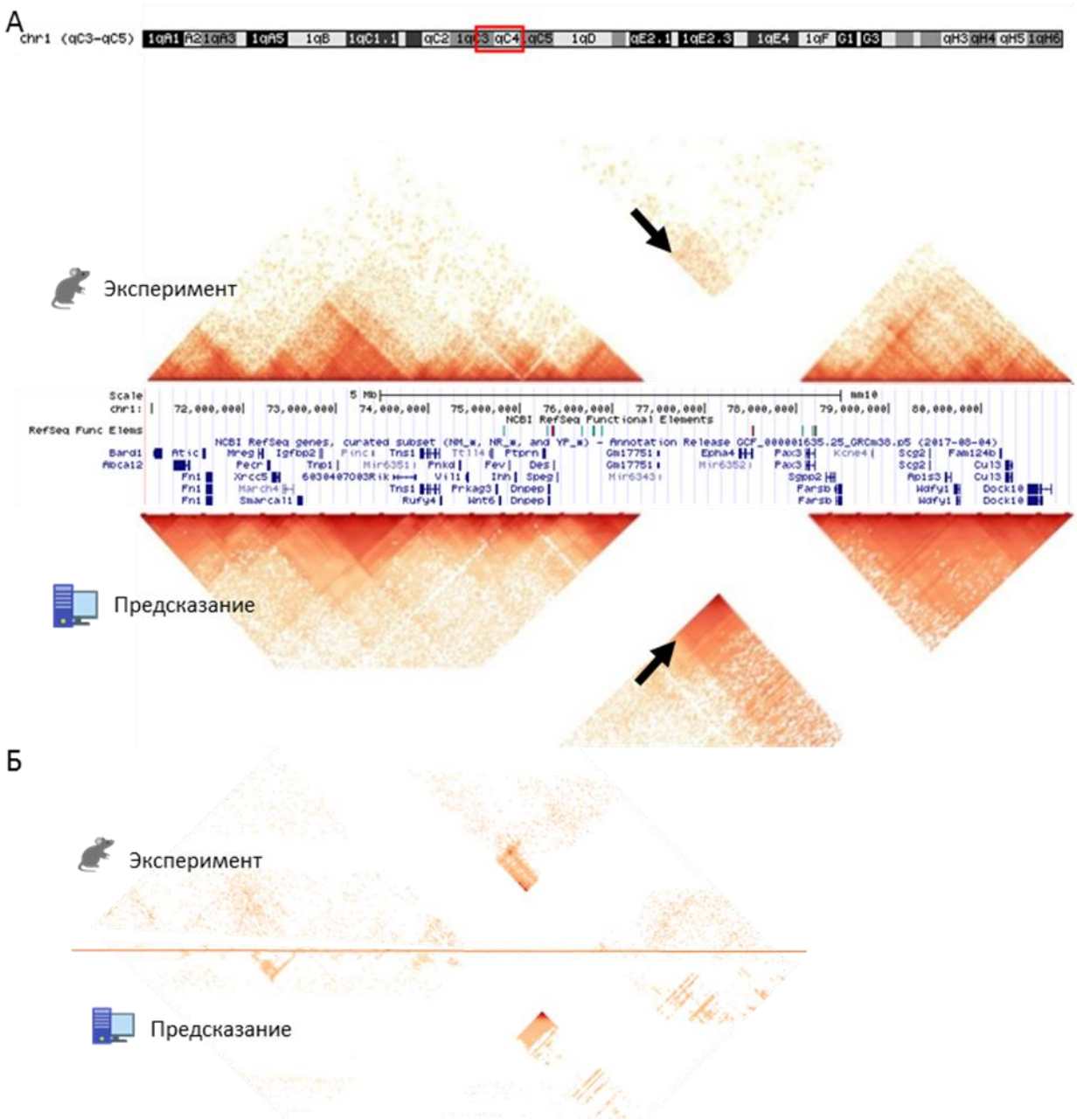

Рисунок 3. Алгоритм 3DPredictor предсказывает эктопические взаимодействия в перестроенном геноме. (А) Сверху экспериментальная Hi-C карта для локуса 71-81 Мб хромосомы 1 на разрешении 5 Кб с делецией региона 75,5-78,5 Мб. Снизу предсказание для этого же локуса для клеток задней конечности мыши. Стрелкой указан регион с повышенной частотой контактов относительно дикого типа. (Б) Эктопические взаимодействия для эксперимента (сверху) и предсказания (снизу) для

того же локуса. Эктопические взаимодействия получены, как описано в разделе в [Bianco и др., 2018].

Разработка вычислительной платформы 3DGenBench для оценки точности алгоритмов для предсказания 3D архитектуры генома.

По мере развития экспериментальных техник изучения трёхмерной укладки хроматина и увеличением количества работ, связанных с функциональным изучением трёхмерной организации хроматина, появляется все больше алгоритмов для предсказания 3D архитектуры генома, основанных на физических или статистических методах моделирования [Bianco и др., 2018; Fudenberg, Kelley, Pollard, 2020; Schwessinger и др., 2020; Szabo, Bantignies, Cavalli, 2019; Zhang и др., 2019]. 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 3D архитектуры генома используют свои методы и свой набор данных и примеров для оценки точности предсказания моделей. Таким образом, создание платформы, где можно было бы сравнить разные алгоритмы между собой на одном наборе данных является особенно актуальным.

Для того чтобы создать такую платформу, необходимо было сформировать большой набор данных, используемый для обучения и тестирования алгоритмов. На первом этапе был проведён анализ литературы для создания набора необходимых Hi-C данных. Поскольку некоторые алгоритмы умеют 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). 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. 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., 2016b; Hanssen и др., 2017; Kragesteen и др., 2018b; 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/epigenetics_data.txt). Для всех типов клеток были обработаны ChIP-seq данные, описывающие профиль связывания белка CTCF.

Таким образом, 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 определил как «вертикальный». Для каждого типа сравнения были реализованы свои метрики.

Для того чтобы оценить правильность работы предложенных метрик, нами был сгенерирован набор Hi-C карт, который смог бы являться некоторым базисом для сравнения. Это набор данных с разным уровнем шума в Hi-C матрице, а также одна Hi-C карта с частотами контактов, полученными случайной перестановкой значений на диагоналях матрицы. Мы выбрали несколько образцов из подготовленного набора данных и протестировали, как меняются значения метрик в зависимости от количества шума в данных (Рис. 4 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. На рисунке 4 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 алгоритма 3DPredictor. Для «горизонтального» типа сравнения, наши колабораторы предоставили предсказанные Hi-C карты контактов хроматина для регионов размером 20 Мб для клеточной линии GM12878 (chr1:22000000-42000000, chr19:36000000-56000000), полученные с использованием разработанных ими ранее алгоритмов PRISMR и DRAGON. Из проведенной нами оценки точности моделей видно, что оба эти алгоритма генерируют предсказания, характеризующиеся примерно одинаковым коэффициентом

корреляции Спирмана между экспериментальными и предсказанными матрицами (Рис. 4 А). Однако более высокие значения SCC и корреляции значений инсуляторного профиля для алгоритма PRISMR указывают на то, что этот алгоритм лучше улавливает такие структуры, как ТАДы. А зависимость частоты контактов от геномного расстояния, наоборот, лучше предсказывает алгоритм DRAGON. Этот пример является отличной иллюстрацией того, как можно использовать метрики и подготовленный нами набор данных для сравнения разных алгоритмов между собой.

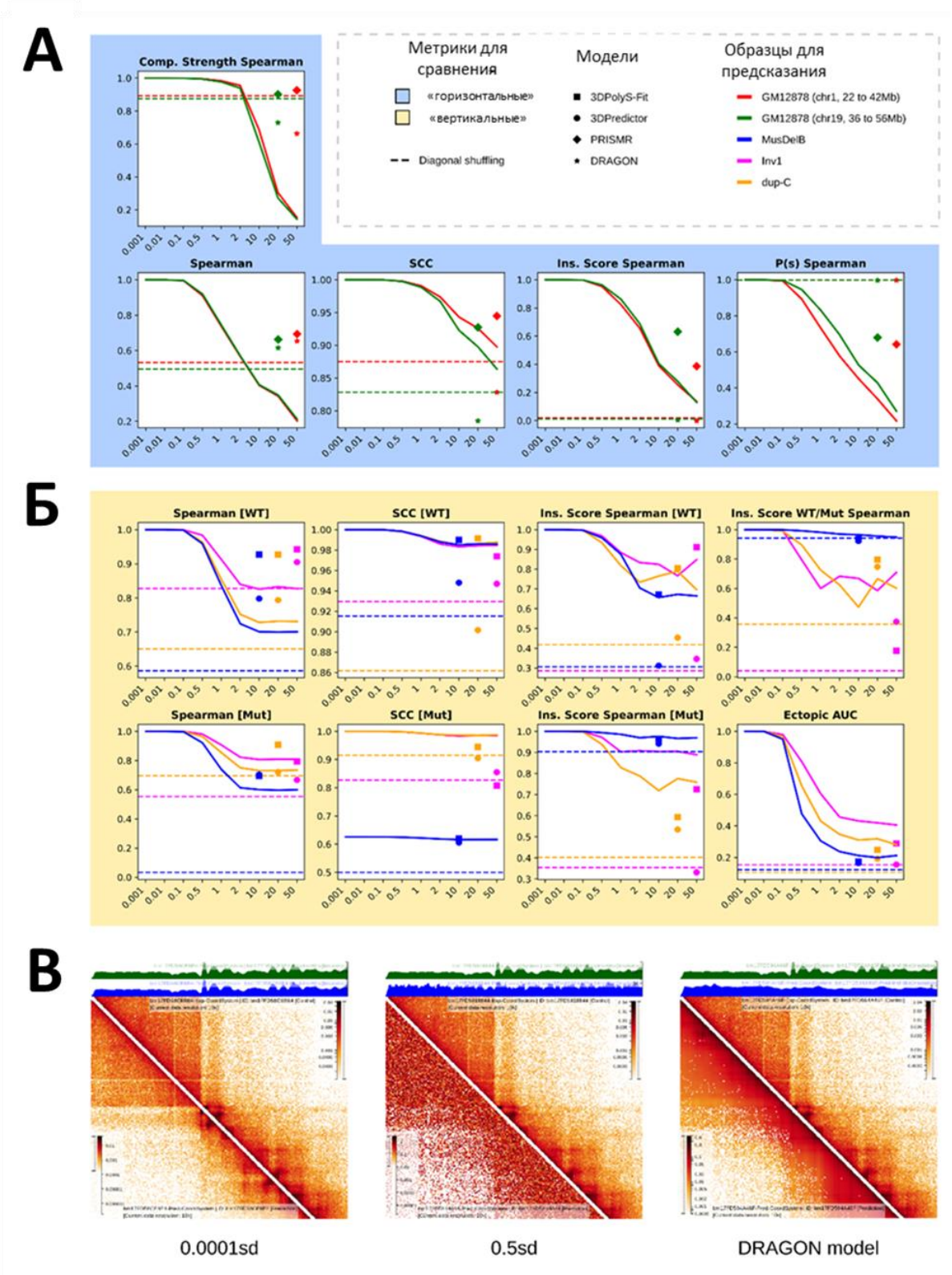

Рис. 4. Разработанные метрики отражают различия предсказанных данных разными алгоритмами. (А) Все метрики «горизонтального» типа сравнения, посчитанные для 2 разных регионов (красный и зелёный цвет) для 2 моделей (PRISMR и DRAGON) (круг и ромб). Кривые отражают зависимость значения метрик от уровня шума в данных. Чем значение по оси X выше, тем больший уровень шума присутствует в Hi-C данных. (Подробная расшифровка значений по оси X представлена в разделе «Материалы и методы» 6.1). Прерывистые линии отражают значение метрики для Hi-C карты с

перемешанными значения на диагоналях («Материалы и методы» 6.1) (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.

Для демонстрации возможности использования метрик, разработанных для «вертикального» сравнении, мы получили от коллег из группы Daniel Jost предсказания пространственной организации 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 инсрумент 3DPredictor, и сравнили полученные предсказания. В первую очередь видно, что оба алгоритма с разной точностью предсказывают контакты для разных типов перестроек (Рис.4 Б).

Изменения инсуляторного профиля, вызванные мутацией, оба алгоритма предсказывают примерно на одинаковом уровне, но алгоритм 3DPolyS-Fit больше преуспевает в предсказании изменений, произошедших в результате инверсии. Та же тенденция прослеживается и в случае метрики, отражающей точность определения эктопических взаимодействий. Приведенные примеры наглядно показывают, что созданная система метрик и набор данных являются хорошим инструментом для сравнения разных алгоритмов между собой.

Для того, чтобы разработчики моделей по предсказанию трёхмерной архитектуры генома могли использовать унифицированные метрики для оценки качества предсказаний, мы разработали вычислительную платформу 3DGenBench, которую сможет использовать любой желающий. Разработанный онлайн-ресурс позволяет получить значения всех метрик в удобном для пользователя формате с визуализацией предсказанных и экспериментальных данных в HiGlass. Кроме того, пользователи имеют возможность использовать разработанную веб-платформу как базу данных, где собраны единообразно обработанные наборы Hi-C и cHi-C данных в часто используемых hic- и cool-форматах на нескольких разрешениях, которые легко можно скачать.

ЗАКЛЮЧЕНИЕ

Подводя итог, можно сказать, что 3DPredictor — это уникальный инструмент, позволяющий количественно предсказывать пространственные взаимодействия хроматина, в том числе и промотор-энхансерные

взаимодействия, а также изменения, происходящие при хромосомных перестройках, используя лишь небольшое количество входных эпигенетических данных.

Для унифицированного сравнения таких алгоритмов, умеющих предсказывать пространственную организацию хроматина в норме и при хромосомных перестройках, нами была разработана вычислительная платформа 3DGenBench. В последнее время тема предсказания 3D архитектуры хроматина стала особенно популярна, и активно начали появляться новые модели для предсказания. В этой ситуации, разработанная нами платформа 3DGenBench и унифицированный набор данных для предсказания как нельзя важны для сравнения постоянно возрастающего числа моделей. В данный момент эта платформа предназначена только для сравнения моделей по предсказанию 3D организации хроматина. Однако 3D геномика - не единственная область, где активно применяют моделирование. Например, существуют модели, предсказывающие экспрессию генов и всевозможные эпигенетические модификации [Avsec и др., 2021; Keilwagen, Posch, Grau, 2019; Wong и др., 2021].

Сейчас моделирование переходит к более комплексному предсказанию свойств. Например, исследователи предсказывают не только места связывания транскрипционных факторов в разных типах клеток, но и регуляторные взаимодействия, возникающие между ними и промоторами соответствующих генов [Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar, 2022]. Также есть подходы, в которых пытаются извлечь закономерности, характеризующие разные типы клеток, что позволяет предсказывать всевозможные эпигенетические характеристики в разных клеточных типах, не обучаясь на данных для конкретного типа клеток [Maria Sindeeva, Nikolay Chekanov, Manvel Avetisian, Nikita Baranov, Elian Malkin, Alexander Lapin, Olga Kardymon, 2021]. Использование и совершенствование идей, заложенных в первых моделях, позволяет создавать более сложные и комплексные алгоритмы. Таким образом, инструмент 3DPredictor занял своё место в наборе существующих моделей для предсказания пространственной архитектуры генома. А платформа 3DGenBench может расширяться в соответствии с развитием новых моделей и методов, что позволит единообразно сравнивать новые модели между собой.

ВЫВОДЫ

1. Разработанный с использованием методов машинного обучения инструмент 3DPredictor позволяет предсказывать пространственные частоты контактов хроматина в гепатоцитах и клетках-

предшественниках нейронов мыши, а также линиях клеток человека GM12878 и K562 на основе данных о генной экспрессии, геномных расстояниях между контактирующими локусами и распределении белка CTCF и ориентации его сайтов связывания.

2. Сравнительный анализ предсказанных и полученных экспериментально карт пространственных контактов хроматина в гепатоцитах и клетках-предшественниках нейронов мыши показал, что инструмент 3DPredictor улавливает тканеспецифичные особенности трехмерной организации генома.
3. Моделирование последствий хромосомных перестроек на примере локуса Epha4 мыши и локуса Kit мыши с помощью инструмента 3DPredictor позволило предсказать эктопические пространственные взаимодействия хроматина, вызванные делециями, что согласуется с экспериментальными данными.
4. Разработанная платформа 3DGenBench позволяет оценить точность инструментов, моделирующих трёхмерную архитектуру хроматина.

СПИСОК ОСНОВНЫХ ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИОННОЙРАБОТЫ:

1. **Belokopytova PS**, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. Quantitative prediction of enhancer-promoter interactions. Genome Res. 2020 Jan;30(1):72-84. doi: 10.1101/gr.249367.119. Epub 2019 Dec 2. PMID: 31804952; PMCID: PMC6961579.

2. **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.

3. **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