ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НОВОСИБИРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ» (НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ, НГУ)

На правах рукописи

БЕЛОКОПЫТОВА ПОЛИНА СТАНИСЛАВОВНА

РАЗРАБОТКА И ОЦЕНКА ТОЧНОСТИ ПРЕДСКАЗАТЕЛЬНЫХ МОДЕЛЕЙ ТРЕХМЕРНОЙ УКЛАДКИ ХРОМАТИНА МЛЕКОПИТАЮЩИХ

1.5.8. – математическая биология, биоинформатика
(биологические науки)

ДИССЕРТАЦИЯ
на соискание ученой степени кандидата биологических наук

Научный руководитель: к.б.н. Фишман В.С.

Новосибирск - 2023

ОГЛАВЛЕНИЕ

СПИСОК СОКРАЩЕНИЙ. КРАТКАЯ ХАРАКТЕРИСТИКА ТЕРМИНОВ.

AUC (от англ. **A**rea **U**nder **C**urve) – площадь под кривой.

3C-технологии (от англ. **c**hromosome **c**onformation **c**apture) – молекулярно-биологический метод захвата конформации хромосом, позволяющий получить информацию о пространственных контактах хроматина.

CAGE (от англ. **c**ap **a**nalysis **g**ene **e**xpression) – кэп-анализ экспрессии генов. Молекулярно-биологический метод, позволяющий получить информацию о транскрипционном профиле эукариотических клеток.

ChIA-drop (от англ. **C**hromatin **I**nteraction **A**nalysis by droplets**) –** молекулярно-биологический метод для получения информации о пространственных взаимодействиях хроматина на уровне одной клетки.

ChIA-PET (от англ. **C**hromatin **I**nteraction **A**nalysis by **P**aired-**E**nd **T**ag Sequencing) – молекулярно-биологический метод, позволяющий получить информацию о пространственной организации хроматина для тех участков генома, с которыми связан определённый белок.

ChIP-seq (от англ. **ch**romatin **i**mmuno**p**recipitation followed with deep **seq**uencing) – молекулярно-биологический метод, позволяющий получить информацию об участках генома, с которыми связан определённый белок.

cHi-C (от англ. Capture **c**hromosome conformation capture with **hi**gh-throughput sequencing) - молекулярно-биологический метод, позволяющий получить информацию о пространственных контактах хроматина для отдельного локуса генома.

FISH (от англ. **f**luorescence ***in situ*** **h**ybridization) флуоресцентная *in situ* гибридизация.

FPKM (от англ. **f**ragments **p**er **k**ilobase of exons per **m**illion mapped reads) – Количество фрагментов на 1000 пар нуклеотидов на миллион выровненных прочтений. Значение, характеризующее уровень экспрессии определённого гена.

GAM (от англ. **g**enome **a**rchitecture **m**apping) - метод криосрезов для картирования совместно локализованных участков ДНК способом без лигирования концов ДНК.

Hi-C (от англ. **c**hromosome conformation capture with **hi**gh-throughput sequencing) – молекулярно-биологический метод, позволяющий получить информацию о пространственных контактах хроматина для всех локусов генома.

MAE (от англ. **M**ean **A**bsolute **E**rror) **–** средняя абсолютная ошибка.

MRE (от англ. **M**ean **R**elative **E**rror) – средняя относительная ошибка.

MSE (от англ. **M**ean **S**quared **E**rror) – средняя квадратичная ошибка.

OoE (от англ. **O**bserved over **e**xpected) - отношение наблюдаемой частоты контактов к ожидаемой частоте контактов на определённом расстоянии на Hi-C карте.

SCC (от англ. **S**tratum-adjusted **c**orrelation **c**oefficient) – метрика, предложенная в [1] для сравнения Hi-C карт.

SPRITE (от англ. **S**plit-**P**ool **R**ecognition Of **I**nteractions By **T**ag **E**xtension) – молекулярно-биологический метод для картирования пространственных взаимодействий хроматина. Позволяет получить информацию сразу о нескольких взаимодействующих локусах.

TSS (от англ. **t**ranscription **s**tart **s**ite) – сайт начала транскрипции.

Кб – килобаза (1000 пар нуклеотидов).

Мб – мегабаза (1000000 пар нуклеотидов).

п.н. – пары нуклеотидов.

ПЦР – полимеразная цепная реакция.

ТАД – топологически ассоциированный домен.

ВВЕДЕНИЕ

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 [2–7].

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 ядра. Однако 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 генома млекопитающих на основе эпигенетических данных. Отдельная часть работы связана с применением этого метода для предсказания последствий хромосомных перестроек.

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

Таким образом, **цель** нашей работы заключается в **разработке алгоритма для предсказания пространственной организации хроматина и создании инструмента для оценки точности таких предсказательных моделей.**

Для достижения цели были поставлены 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 архитектуры 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, RNA-seq данные и расстояние между геномными локусами. Кроме того, мы впервые показали, что такой инструмент можно использовать для предсказания изменений в трёхмерной архитектуре генома, произошедших в следствие хромосомных перестроек.

В последние годы появилось несколько алгоритмов, способных предсказывать трёхмерную организацию хроматина в норме и при различных мутациях. Такие инструменты могут быть полезны в медицинской генетике, однако нужно 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 белка CTCF и информация об экспрессии генов, являются доступными и широко распространёнными для разных типов клеток. Разработанный нами алгоритм 3DPredictor позволяет предсказывать 3D организацию генома для таких типов клеток. Кроме того,

наш алгоритм может предсказать изменения трёхмерной организации хроматина, произошедшие при хромосомной перестройке. Это позволяет предположить, как изменится экспрессия генов, что может быть интересно медицинским генетикам для объяснения патологий, вызванных хромосомными перестройками. В связи с тем, что за последние два года появилось уже несколько подобных алгоритмов, 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 производительность 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.

Методы диссертационной работы.

Для подготовки данной работы использовались различные биоинформатические 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.

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.**

Апробация результатов и публикации.

Научные результаты, изложенные в данной работе, были представлены на нескольких международных конференциях в виде стендовых и устных докладов. А именно:

1. **Belokopytova PS**, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. «3DPredictor: an algorithm for predicting spatial chromatin interactions», Interdisciplinary school in 3D genomics: from experiments to models and back, Lyon, France (online), 23.11.2020 - 25.11.2020
2. **Belokopytova PS**, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. «3DPredictor: machine learning-based algorithm for prediction of 3D chromatin structure», Системная биология и биоинформатика (SBB - 2020), Ялта, РФ, 14.09.2020 - 20.09.2020
3. **Белокопытова ПС,** «Разработка и экспериментальная валидация модели для предсказания пространственных контактов хроматина на основе эпигенетических характеристик геномов мыши и человека», МНСК-2019, Новосибирск, РФ, 14.04.2019-19.04.2019
4. **Белокопытова ПС**, Нуриддинов МА, Можейко ЕА, Фишман ДС, Фишман ВС, «Разработка модели для предсказания пространственных контактов хроматина на основе эпигенетических характеристик геномов мыши и человека», XVIII Конференция - школа

с международным участием "Актуальные проблемы биологии развития" Москва, РФ, 14.10.2019 - 19.10.2019

5. **Belokopytova PS**, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. «Design of algorithm for 3D chromatin interactions prediction based on epigenetic genomic features», Chromosomes and mitosis. International mini-conference, Новосибирск, РФ, 21.11.2019 -21.11.2019

6. **Белокопытова ПС**, Фишман ДС «Оценка последствий хромосомных перестроек с точки зрения трёхмерной организации генома», МНСК-2018, Новосибирск, РФ, 22.04.2018

По теме диссертации было опубликовано 3 работы. Основные результаты были изложены в рецензируемых журналах *Genome Research* и *Nucleic Acid Research*.

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

Вклад автора. Автором была написана большая часть кода для работы алгоритмы 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.

Структура и объем диссертационной работы.

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.

Благодарности.

Автор диссертационной работы в первую очередь выражает огромную благодарность своему научному руководителю к.б.н. В.С. Фишману (ИЦиГ СО РАН, Новосибирск) за поддержку и вдохновение в работе, а также за создание той атмосферы, в которой хочется развиваться, узнавать что-то новое и двигаться вперёд. Кроме того, автор благодарит всех своих коллег из отдела молекулярных механизмов онтогенеза (ИЦиГ СО РАН, Новосибирск) за продуктивные научные дискуссии и умение поддержать, когда всё кажется бессмысленным. Ещё хочется поблагодарить д.т.н. профессора М.В. Первухина (СФУ, Красноярск), вдохновившего автора заняться наукой ещё в школьные годы. И автор выражает благодарность всем своим друзьям, способствовавших написанию этой работы в самых трудных жизненных ситуациях, особенно А.Стрельнику за помощь в эффективном планировании времени.

ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ.

Известно, что молекула ДНК человека имеет длину около 2м [8], а ядро клеток млекопитающих имеет сферическую форму объёмом до 2500 мкм³ [9]. Таким образом получается, что геном должен быть очень плотно упакован в пространстве клеточного ядра. Тем не менее даже при такой сильной компактизации обеспечивается клеточно-специфическая экспрессия и репликация генетического материала. Это приводит к мысли, что компактизованный хроматин имеет организованную структуру, которая позволяет эффективно реализовывать молекулярно-генетические процессы транскрипции, репликации и репарации.

Пространственная организация хроматина достаточно хорошо совпадает между разными типами клеток. Различия в основном наблюдаются в локусах с разной экспрессией генов. Геномный материал во всех клетках одинаковый и клеточная специфичность достигается за счёт различной регуляции клеточных процессов. Исследование механизмов, лежащих в основе этой регуляции является «горячей» темой биологии. Регуляции экспрессии генов осуществляется на нескольких уровнях и одними из важных регуляторов транскрипции являются такие регуляторные цис-элементы как промоторы и энхансеры. Взаимодействие этих элементов опосредовано взаимодействием транскрипционных факторов и физическим расположение промоторов и энхансеров в пространстве ядра. Таким образом изучение организации хроматина в пространстве ядра клеток является актуальной областью биологических исследований, для которой на сегодняшний день сложились свои экспериментальные и вычислительные методы исследования.

1.1 Трехмерная организация хроматина. Основные методы изучения.

Изучение 3D архитектуры хроматина проводилось в основном двумя подходами [10]. К первому подходу относится визуализация пространственных контактов генома различными методами микроскопии как,

например, FISH (fluorescent *in situ* hybridization). Этот подход появился первым и позволяет визуализировать ограниченное количество локусов генома в пространстве.

Альтернативным методом являются технологии 3C (**c**hromosome **c**onformation **c**apture), основанные на захвате конформации хромосом [11]. Известно, что с ДНК связано большое количество белков, и если зафиксировать их в пространстве ядра, то зафиксируется и связанная с ними ДНК. В методе 3C было предложено фиксировать хроматин, для чего обычно используют формальдегид. После фиксации ДНК фрагментируют – этого можно добиться обработкой эндонуклеазой рестрикции, другой нуклеазой (например, MNaseI) или ультразвуком. Чаще всего для фрагментации используют рестриктазы, узнающие последовательности из 4-6 пар оснований, такие как DpnII, HindIII, MboI. В результате получается геном, разбитый на небольшие фрагменты ДНК, которые связаны с белками. Если затем провести лигирование в условиях сильного разбавления, то участки ДНК, находящиеся близко в пространстве, залигируются в одну химерную молекулу. Затем полученные фрагменты ДНК выделяют и очищают от белков. Получается так называемая 3C-библиотека, которая состоит из смеси химерных молекул ДНК, полученных в соответствии с их расположением в ядре клетки (Рис. 1). Эта библиотека служит основой для множества методов, основанных на 3C-технологии [12,13]. В классической 3C-технологии анализируются контакты между специально выбранными участками генома, поэтому проводится ПЦР с праймерами, специфичными к исследуемым локусам генома, и оценивается количество получившихся ПЦР продуктов относительно контроля.

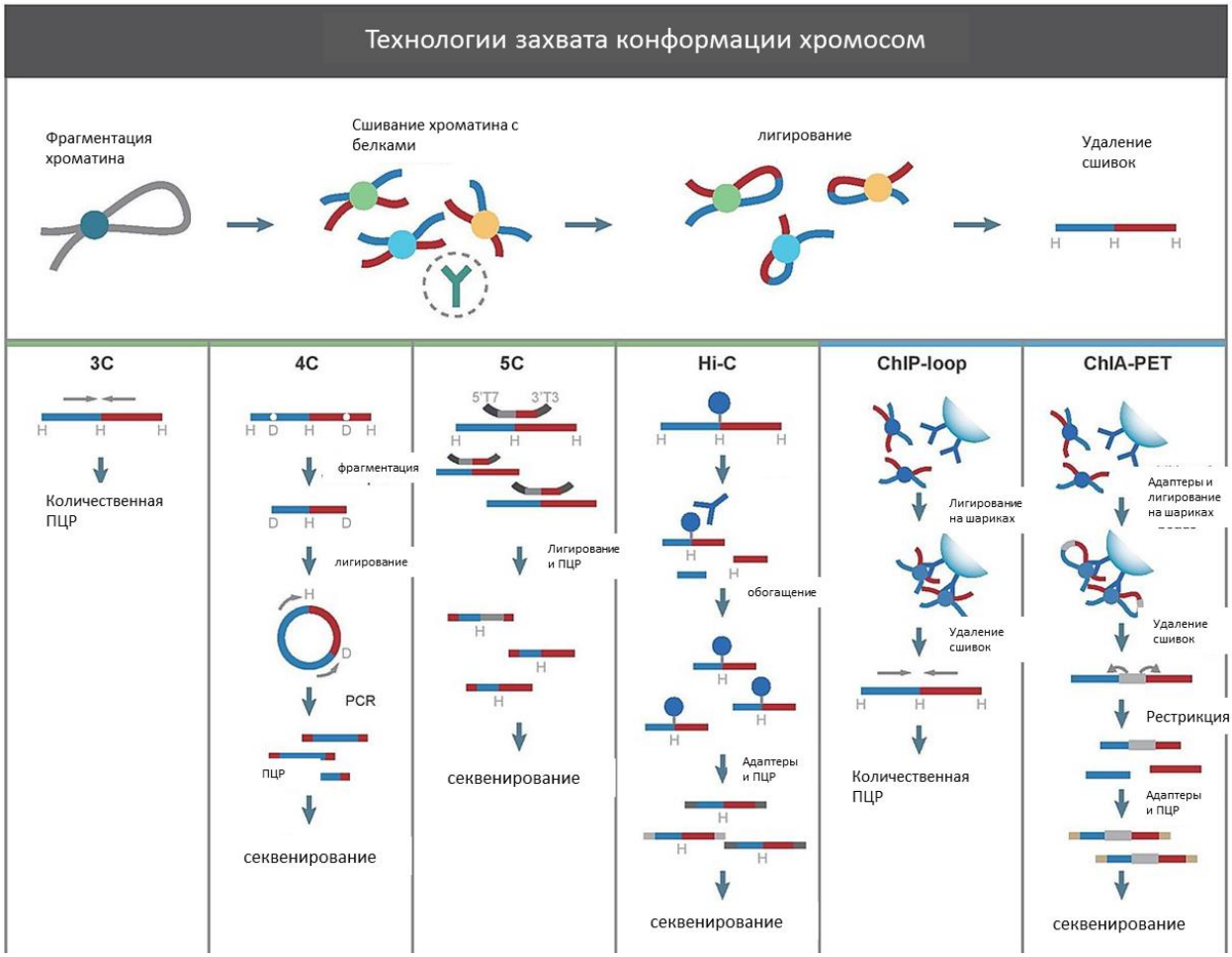


Рис. 1. Методы, основанные на технологии 3C. Рисунок адаптирован из [14]

Другие методы, основанные на технологии захвата хромосом, такие как 4C и 5C позволяют анализировать пространственный взаимодействия одного выбранного участка генома с множеством других геномных локусов. Комбинация 3C-методов с методами ChIP-seq позволяет исследовать пространственную организацию хроматина, с которым связан конкретным белком. Особую популярность приобрёл метод Hi-C, с помощью него можно исследовать пространственные контакты всех возможных пар локусов в геноме. Библиотеку, сделанную на основании 3C-технологии, секвенируют с помощью методов секвенирования нового поколения. Анализ полученных прочтений позволяет определить какие участки генома оказались ковалентно соединены друг с другом, и насколько часто такие сшивки наблюдались. Результатом такого эксперимента является матрица попарных частот

взаимодействий для всех локусов генома. По-другому такие матрицы называют Hi-C картами или тепловыми картами пространственных контактов генома (Рис. 2) [12,13,15].

Каждая карта имеет своё разрешение – размер бинов, то есть равных частей, на которые разбит геном. Предположим разрешение Hi-C карты равно 5 Кб, значит один бин равен 5 Кб, как на Рис. 2. Тогда точка на Hi-C карте, отражающая частоту контактов между координатами генома 154300 Кб и 155500 Кб, представляет собой усредненное количество контактов между участками генома, попадающими в локусы 154300-154305 Кб и 155500-155505 Кб. Разрешение Hi-C карты ограничено размером рестрикционных фрагментов и глубиной секвенирования. Не так давно были получены Hi-C карты человека с разрешением 1 Кб [16,17], это на настоящий момент Hi-C карты пространственных контактов генома человека с самым высоким разрешением. На Hi-C карте можно увидеть структуры, представляющие собой треугольники, яркие красные точки, линии и т.д. Более подробно об этих структурах и механизмах, лежащих в их основе, написано в следующем разделе.

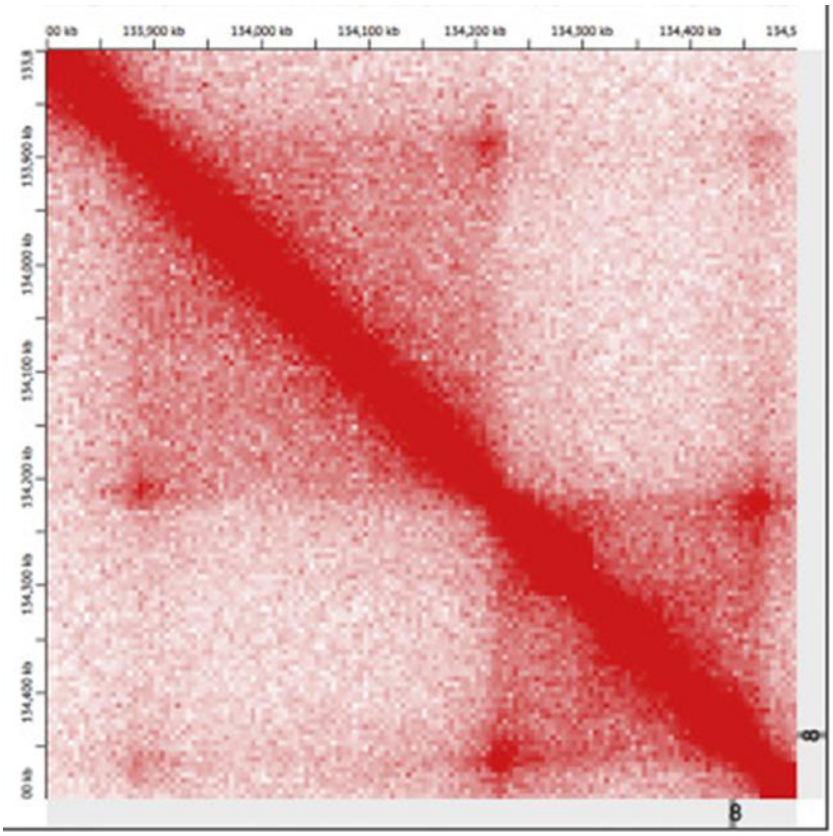


Рис. 2. Пример Hi-C карты из [18]. Каждая точка на карте отражает частоту взаимодействий между двумя локусами генома, чем точка краснее, тем частота контактов выше. Сверху и слева отмечены координаты на хромосоме 8.

Достаточно часто бывает, что исследователи изучают пространственную организацию хроматина конкретного локуса генома и исследуют как она меняется при различных условиях. В этом случае 3C-библиотека обогащается фрагментами ДНК, чья последовательность комплементарна интересующей области генома. Это делается при помощи биотинилированных олигонуклеотидов, комплементарных исследуемым локусам генома. Такой метод называется capture Hi-C (cHi-C) [19][20]. Hi-C карты, полученные данным методом, позволяют даже при неглубоком секвенировании исследовать интересующий участок генома на довольно высоком разрешении (до 1-5 Кб).

Недавно появились новые методы исследования пространственных контактов генома, такие как GAM (**g**enome **a**rchitecture **m**apping) [21], SPRITE (**S**plit-**P**ool **R**ecognition Of **I**nteractions By **T**ag **E**xtension) [22] и ChIA-Drop

(**C**hromatin **I**nteraction **A**nalysis by droplets) [23]. В этих методах не используется лигирование участков ДНК в условиях сильного разбавления, что снимает некоторые ограничения метода Hi-C. Так, например, они позволяют обнаруживать те контакты хроматина, которые включают три или более участков ДНК, а также позволяют выявлять контакты локусов генома, расположенных на расстоянии десятков миллионов пар оснований [10].

1.2 Трехмерная организация хроматина. Основные структуры и механизмы.

Различные методы изучения пространственной архитектуры хроматина позволяют получить понимание того, какие физические и биологические механизмы участвуют в процессах, определяющих пространственную организацию генома. Поскольку линейный размер молекулы ДНК многократно превышает размеры интерфазного ядра, ДНК должна быть эффективно упакована, для того чтобы поместиться в малом объеме ядра [24]. С точки зрения физики процесса компактизации хроматина были предложены две основные модели укладки: стохастическая [25] и фрактальная [26] модели глобулы.

Стохастическая модель глобулы предполагает случайное выпетливание, которое приводит к компактизации, однако проблема этой модели заключается в том, что, согласно её предсказаниям, в ДНК должны были бы возникать узлы, которые с биологической точки зрения могут препятствовать реализации биологических процессов.

Модель фрактальной глобулы наиболее вероятно соответствует биологическим представлениям. Результаты Hi-C экспериментов также подтверждают именно модель фрактальной глобулы [15,27].

На хромосомном уровне было замечено, что каждая хромосома занимает свою территорию в интерфазном ядре [28,29]. Гетерохроматин обычно находится на периферии ядра, а эухроматин кластеризуется ближе к центру [30,31].

Благодаря методам, основанным на 3С-технологии, удалось изучить пространственную организацию генома более детально. На картах Hi-C можно наблюдать различные структуры (Рис. 3), о которых далее будет рассказано более подробно.

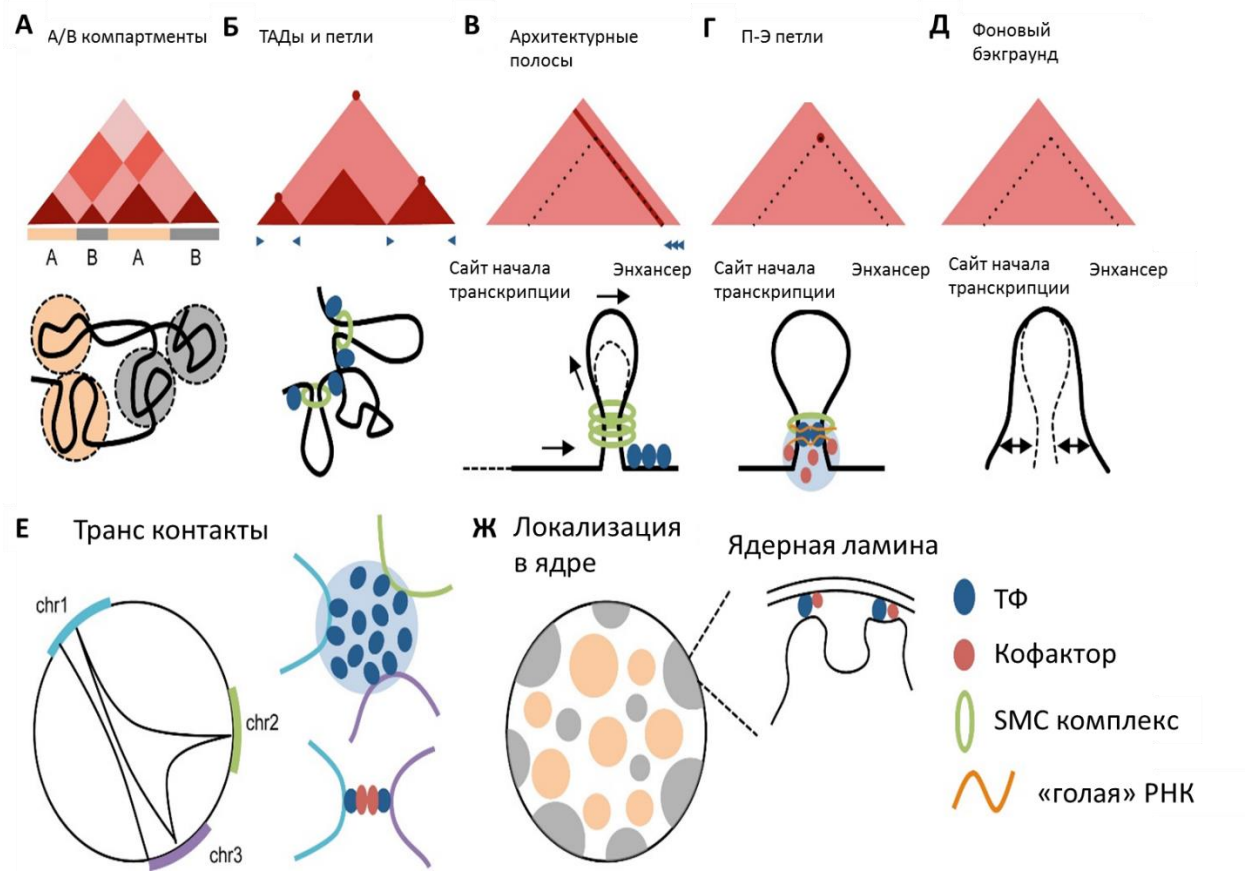


Рис. 3. Основные структуры пространственной организации хроматина. Рисунок адаптирован из [32]. ТФ – транскрипционный фактор, П-Э – промотор-энхансерные. Схемы Hi-C карт генома (вверху) и механизмов, которые их формируют (внизу). (А) Области активного или неактивного хроматина (желтые и серые полосы под контактной картой) связывают с образованием конденсатов.

(Б) ТАДы и петли образуются в результате «протягивания петли» когезином (зеленые кольца) и их блокирования конвергентно ориентированными белками CTCF (синие стрелки указывают ориентацию мотива).

(В) Архитектурные полосы образуются в результате частой загрузки когезина и однонаправленных сайтов посадки CTCF.

(Г) Промотор-энхансерные петли могут быть вызваны образованием конденсатов, прямой или непрямой олигомеризацией, механизмом «протягивания петли» и/или взаимодействиями белок-РНК.

(Д) В некоторых случаях частоты контактов пары энхансер-TSS (transcription start site) могут быть на уровне локального фонового контакта.

(Е) Транс-контакты между различными хромосомами могут быть опосредованы конденсатами (вверху справа) или олигомеризацией (внизу справа).

(Ж) Ядерная локализация, такая как локализация В-компартмента (выделен серым цветом) на периферии ядра, образуется в результате прямых или косвенных взаимодействий транскрипционных факторов с закреплёнными локусами.

1.2.1 А- и В- компартменты.

Одной из особенностей Hi-C карт является «пледчатый» паттерн контактов (Рис. 3А). Есть участки генома, которые контактируют друг с другом чаще, чем с другими, находящимися на таком же расстоянии. Основываясь на этом наблюдении, весь геном разделили на две категории, которые назвали А- и В-компартментами [15]. Для участков из А-компартмента характерно наличие контактов с другими регионами А-компартмента, при чём локусы могут быть удалены на большое расстояние. Участки В-компартмента предпочтительно взаимодействуют с ближайшими регионами из В-компартмента, и для них взаимодействия между сильно удаленными локусами генома не характерны. Это позволило предположить, что В-компартмент более плотно упакован, что сходится с данными FISH [33]. Изучение корреляции различных эпигенетических характеристик с компартментами показало, что А-компартмент коррелирует с присутствием генов, более высокой экспрессией и открытым хроматином [15,34]. Таким образом, А- и В- компартменты, видимые на Hi-C картах, соответствуют открытой и закрытой конформации хроматина.

Физическое разделение хроматина объясняется взаимодействиями нуклеиновых кислот и белков между собой. Например, транскрипционные факторы способны связывать ДНК, а также взаимодействовать с другими белками. Кроме того, некоторые транскрипционные факторы могут олигомеризовываться, напрямую взаимодействуя друг с другом, другие связываются при участии дополнительных белков-кофакторов. Основным физическим процессом, участвующим в разделении хроматина, считается механизм жидкость-жидкостной фазовой сепарации [32]. Жидкость-жидкостная фазовая сепарация представляет собой процесс, примером которого является образование капель масла в водной среде. В ядре клетки растворы белков и нуклеиновых кислот конденсируются в плотную фазу, которая сосуществует с жидкой фазой [35]. Движущей силой процесса фазовой сепарации является обмен взаимодействий макромолекула-вода на макромолекула-макромолекула и вода-вода в условиях, где такой процесс является энергетически выгодным [36].

1.2.2 ТАДы

Кроме «пледчатой» структуры на Hi-C карте можно увидеть яркие треугольники (Рис. 3Б), которые охватывают районы генома размером около 0,1-1 Мб. Такие структуры называются ТАДами (топологически ассоциированными доменами) и физически представляют собой участки хроматина, упакованные в виде клубка. Домены наблюдаются в различных типах клеток и являются консервативными для разных видов [37]. Было показано, что даже в таких необычных с точки зрения организации хроматина клетках как сперматозоиды, где ДНК сверхплотно упакована в ядре, есть ТАДы, которые имеют схожее строение с ТАДами других клеточных типов [38]. Было замечено, что границы ТАДов обогащены белками CTCF и промоторами генов домашнего хозяйства [37]. Кроме того, в вершинах ТАДов в большинстве случаев наблюдается увеличение частоты контактов, что свидетельствует о пространственной близости локусов на границах ТАДа.

Такие структуры называют петлями. Основной механизм, участвующий в образовании этих структур - это механизм «протягивания петли» (Рис. 4).

Механизм «протягивания петли» был предложен параллельного в лабораториях Мирного [39] и Эрез-Либермана Айдена [40]. Согласно предложенной модели, главную роль в создании петель хроматина в интерфазе играет белок когезин. Он выступает как мотор, способный наращивать петли, когда хромосомы «распущены». Эта активность когезина в дальнейшем была подтверждена *in vitro* [41]. Предполагается, что когезин, имеющий форму кольца, протягивает петли ДНК и останавливается, когда встречает белок CTCF. Сайт посадки CTCF асимметричен, т.е. может иметь прямую (→) или обратную (←) ориентацию. CTCF сайты в основании петель, таким образом, могут быть ориентированы конвергентно (→….←), дивергентно (←….→) или сонаправленно (→….→). Это определяет возможность формирования петель: петли образуются преимущественно между двумя конвергентно направленными сайтами посадки CTCF. Как предполагается в модели Мирного и коллег, белки CTCF — это знаки «стоп» для когезина. Если когезин доходит до определенным образом ориентированного CTCF с каждой стороны растущей петли, то эти белки оказываются вместе, и когезин прекращает протягивать хроматин. Такой метод компактизации генома не только придает форму и структуру хромосомам, но и сближает нужные участки ДНК друг с другом.

Рис. 4. Механизм петлеобразования согласно модели «протягивания петли». Рисунок адаптирован из [42].

1.2.3 Петли.

В предыдущем разделе было отмечено, что часто области с повышенной частотой контактов (петли) возникают между локусами генома, которые находятся в границах топологических доменов. В этом случае образование

этих структур хорошо объясняется механизмом «протягивания петли».

Однако есть петли, в основаниях которых не содержится сайтов посадки белков CTCF. Механизм их образования, по крайней мере частично, независим от протягивания петель когезином, так как при деградации когезина часть петель сохраняется [43].

Такие петли обогащены промотор-энхансерными взаимодействиями (Рис. 3Г). Далеко не все механизмы и белки, участвующие в этом взаимодействии изучены в настоящее время. Но было показано, что петли могут образовываться посредством взаимодействия регуляторных белковых комплексов, таких как, например, комплекс белка поликомб [44] и других транскрипционных факторов [45]. Кроме того, Abraham S Weintraub с коллегами было показано, что белок YY1 связывается с активными энхансерами и регионами рядом с промоторами, и образует димеры, которые облегчают взаимодействие локусов ДНК. Делеция сайтов связывания YY1 или разрушение белка YY1 нарушает образование энхансер-промоторных петель и экспрессию генов [46,47].

1.2.4 Архитектурные «полосы».

Некоторые ТАДы имеют так называемы архитектурные «полосы» в своих границах (Рис. 3В). Эти полосы на Hi-C карте возникают из-за того, что участок генома в границе ТАДа часто оказывается близко в пространстве со всеми остальными регионами генома, находящимися внутри ТАДа. Возникновение такого паттерна связано с тканеспецифическими суперэнхансерами. Как полагают, эти структуры образуются в результате протягивания петель за счет загрузки когезина в места, где есть кластер однонаправленных сайтов связывания CTCF (основание полосы), который блокирует продвижение когезина в одном из направлений [48]. В результате, когезин движется от своего места загрузки однонаправленно, внутрь ТАДа, в результате чего участок ДНК, соответствующий месту загрузки когезина, последовательно взаимодействует со всеми остальными локусами ТАДа. Это

приводит к паттерну «полосы» на Hi-C карте.

1.3 Функциональная роль 3D архитектуры хроматина

Трёхмерная организация хроматина является достаточно консервативной внутри отрядов и даже классов животных. Эволюционная консервативность структур 3D архитектуры генома позволяет выдвинуть гипотезу о функциональной значимости пространственной организации хроматина.

В первую очередь значение пространственной организации хроматина связывают с участием во взаимодействии таких регуляторных элементов генома как энхансеры и промоторы. Именно таргетное взаимодействие промоторов и энхансеров определяет программу развития и клеточную специфичность. Считается, что энхансеры регулируют транскрипцию только когда промотор и энхансер находятся физически близко в пространстве ядра [49]. Однако, промотор и специфический к нему энхансер могут находится на расстоянии до нескольких Мб [50] в геномных координатах. Классический пример удаленных промотор-энхансерных взаимодействий регуляции - это регуляция экспрессии генов в локусе β-глобиновых генов [51,52]. Локус β-глобиновых генов у человека содержит пять кодирующих генов, которые специфически экспрессируются в эритроидных клетках. Активная транскрипция этих генов обеспечивается физическим взаимодействием промоторов этих генов с одним и тем же сильным энхансером на разных стадиях развития. Таким образом, 3D организация генома этого локуса определяет то, какие из возможных промотор-энхансерных взаимодействий будут реализованы в клетке и, соответственно, какой из глобиновых генов будет экспрессироваться.

Сейчас уже ясно, что активация транскрипции зависит от совокупности факторов, включающих не только близость промотора и энхансеров в пространстве, но также действия различных транскрипционных факторов и

других молекул, в том числе и самой энхансерной РНК (Рис. 5) [53]. Механизм участия энхансера в активации транскрипции может объясняться различными процессами, включающими ремоделлинг хроматина, привлечение комплекса инициации транскрипции, привлечение РНК полимеразы II, влияние на белки репрессоры транскрипции, снятие РНК полимеразы с паузы и т.д. Однако точный молекулярный механизм, описывающий роль энхансеров в активации транскрипции, до сих пор полностью не исследован. Тем не менее, ясно, что важно именно физическое сближение регуляторных элементов. В частности, было показано, что контакт энхансера и промотора увеличивает частоту «транскрипционных взрывов» [54].

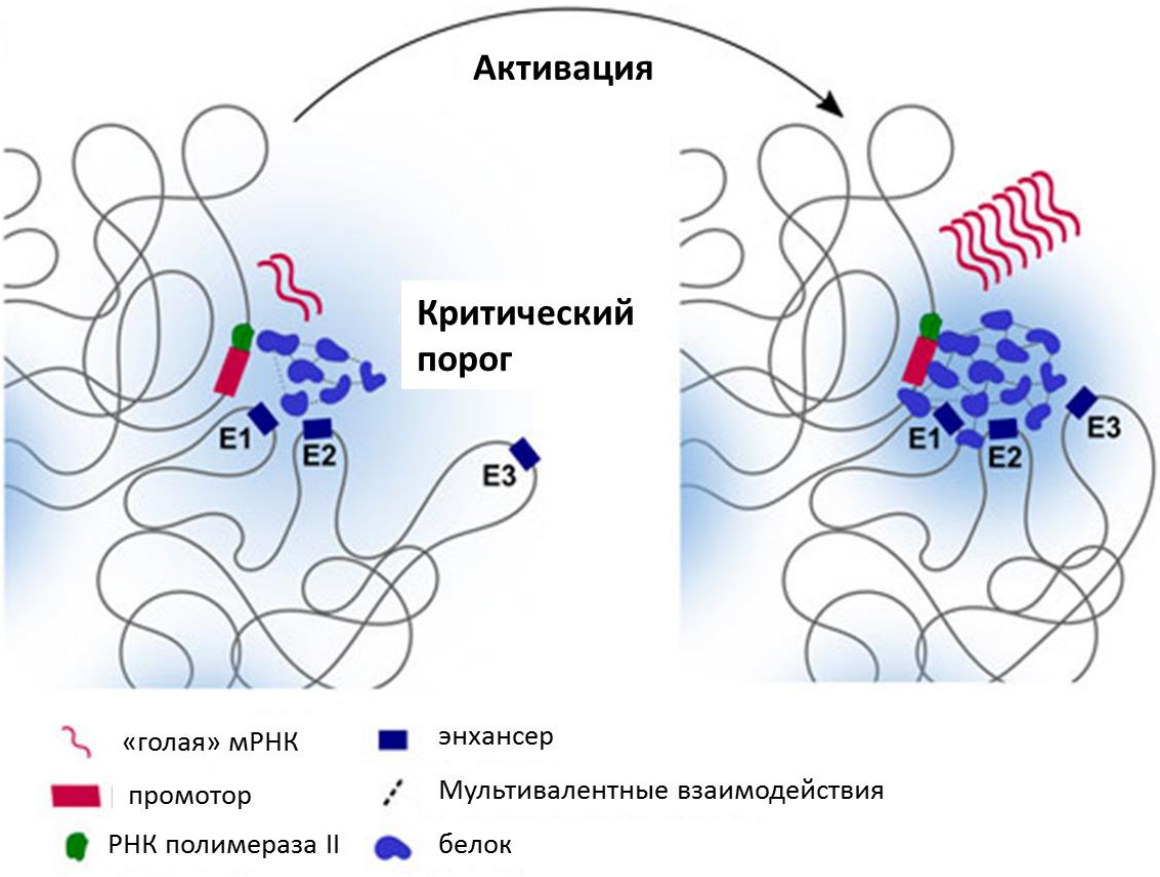


Рис. 5. Мультифакторная модель регуляции транскрипции. Компартментализация регуляторов транскрипции, энхансеров и промоторов способствует активации транскрипции. Рисунок адаптирован из [53].

1.3.1 Роль пространственной организации хроматина при хромосомных перестройках

Создание доступных методов редактирования генома сделало возможным тестирование функциональной значимости тех или иных элементов пространственной организации хроматина путем создания генетически-модифицированных клеточных линий и организмов. Давно известно, что мутации внутри генов могут приводить к синтезу нефункционального белка, что в свою очередь, может являться причиной заболеваний. Однако изменение экспрессии генов может произойти даже в случае мутаций в некодирующей области генома, в частности, за счет изменения пространственной организации хроматина. Был проведён ряд работ, в которых показали, как изменения в пространственной организации генома, могут влиять на экспрессию генов [2,4–6,55–57]. Ниже приведены несколько примеров, иллюстрирующих роль 3D организации хроматина в регуляции правильного функционирования генов.

В одной из первых таких работ было показано, что хромосомные перестройки, затрагивающие границы ТАДов, могут приводить к патологиям развития конечностей [58] (Рис. 6).

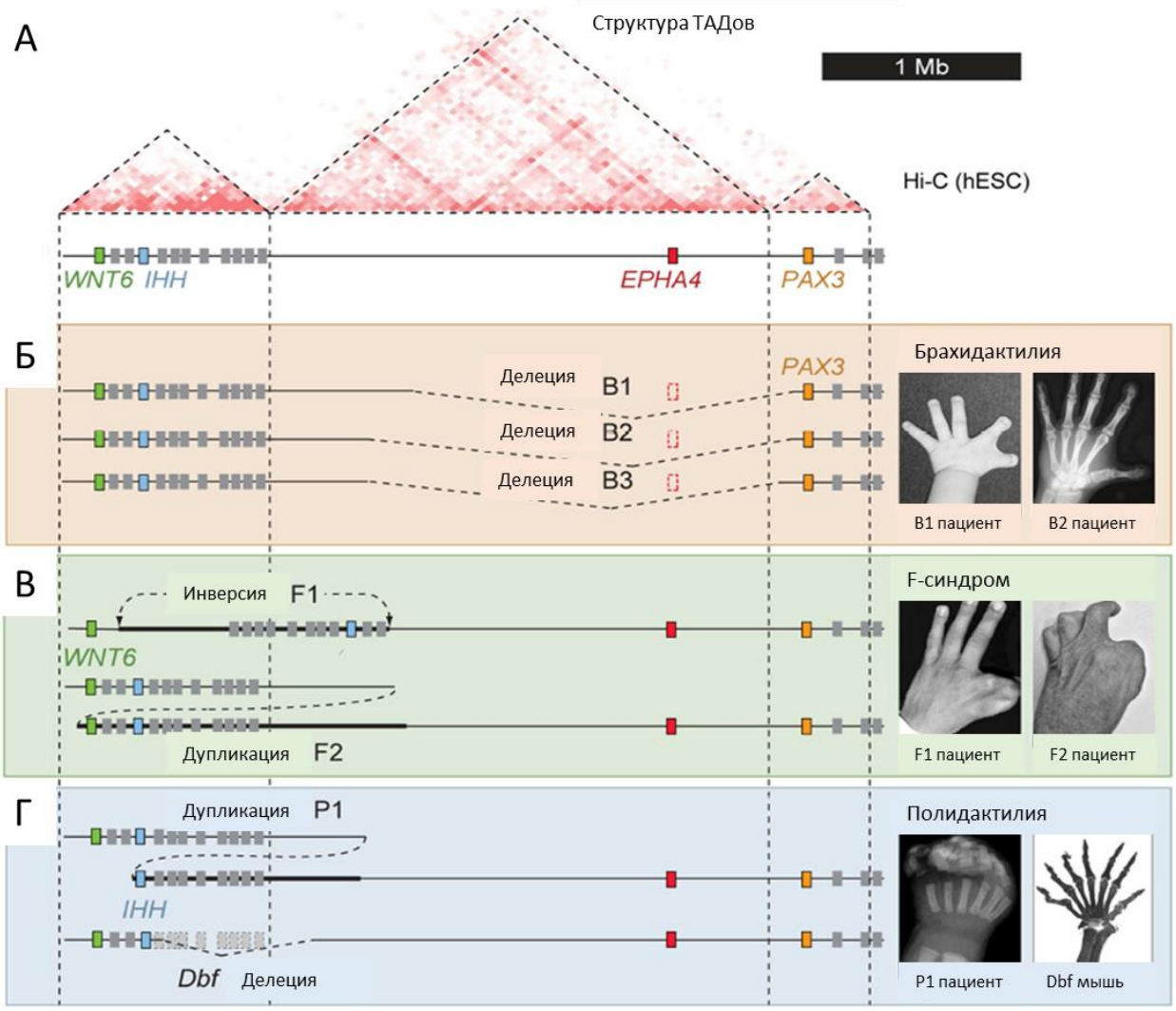


Рис. 6. Различные патогенные варианты хромосомных перестроек в локусе гена *EPHA4* из [58]. (А) Hi-C карта контактов локуса *EPHA4* в эмбриональных стволовых клетках человека. Пунктирные линии обозначают границы ТАДов. (Б–Г) Схема хромосомных перестроек (слева) и связанных с ними фенотипов (справа). (Б) Делеции B1, B2, B3 приводят к брахидактилии у человека. (В) Инверсия F1 и дупликация F2 приводит к F-синдрому. (Г) Дупликация P1 и делеция Dbf приводит к полидактилии у человека и мыши соответственно.

В качестве объекта для исследования авторы использовали локус *WNT6/IHH/EPHA4/PAX3*, в котором можно наблюдать три выраженных ТАДа. В литературе были описаны мутации в этом локусе, которые приводят к патологиям развития конечностей у людей. При помощи технологии CRISPR-

Cas9 авторы получили линии мышей, в которых были воспроизведены аналогичные хромосомные перестройки, затрагивающие границы между ТАДами. Исследователи показали, что патологии возникают из-за повышенной экспрессии генов *PAX3* или *WNT6* и *IHH*. Причина этого – увеличение частоты пространственных взаимодействий между промоторами этих генов с энхансерами из ТАДа *EPHA4*, что происходит из-за нарушения инсуляции ТАДов (Рис. 7).

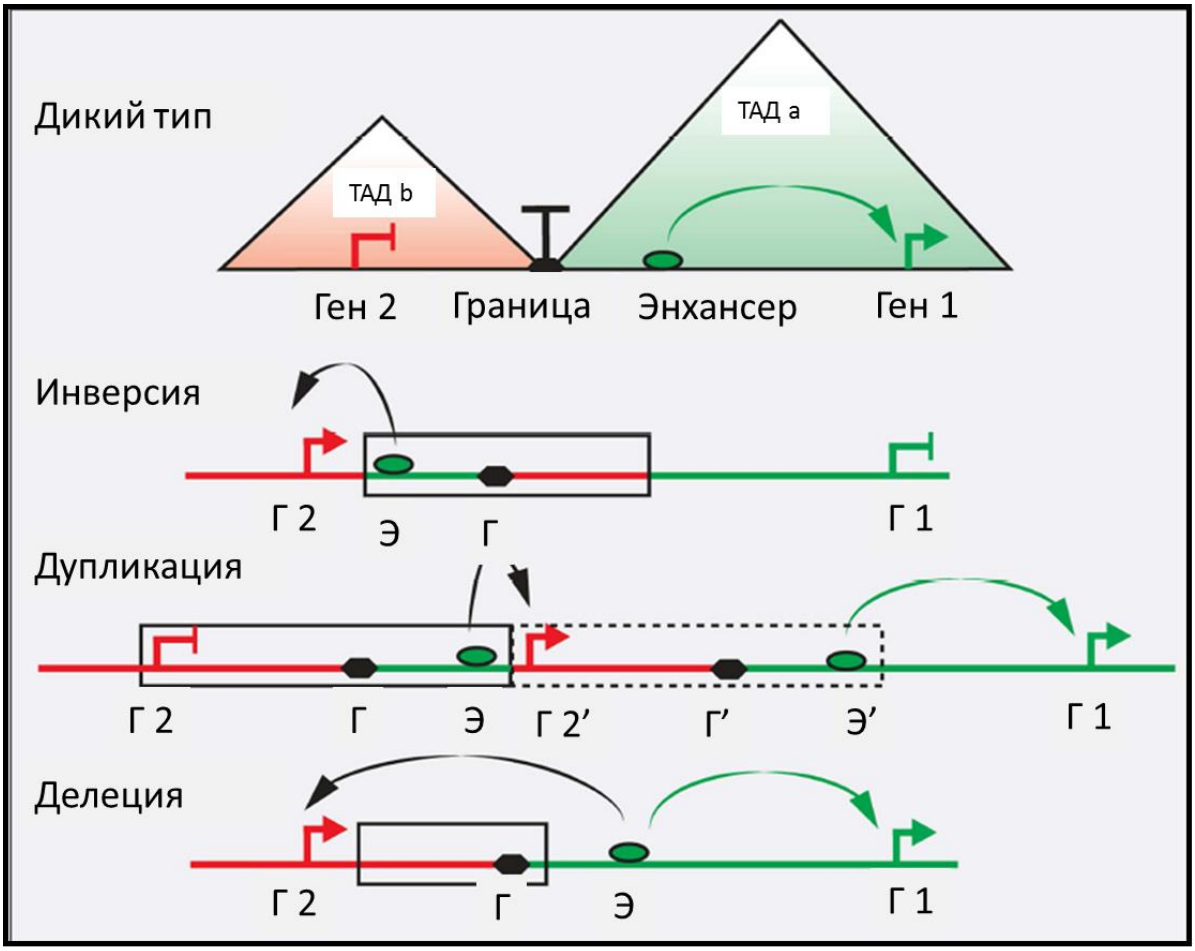


Рис. 7. Эктопические контакты, возникающие при различных хромосомных перестройках. Рисунок адаптирован из [58].

В другой работе тестировали влияние дупликаций некодирующих регионов внутри и на границе ТАДов. Дупликации некодирующей ДНК в ТАДе, содержащем ген *SOX9*, который участвует в процессе переключения с женского типа развития на мужской у человека, привели к увеличению

контактов дублицированных областей внутри домена, но не изменили общую структуру ТАДа. Напротив, дуплицированные области, попавшие в соседний ТАД, привели к образованию нового домена хроматина, который был отделен он остальной части генома. Если в этот новый ТАД не попадали никакие гены, это не вызывало фенотипического эффекта. Однако включение гена *KCNJ2* в новый ТАД привело к возникновению контактов гена *KCNJ2* с регуляторными элементами области гена *SOX9*, и соответственно к изменению экспрессии гена *KCNJ2* и нарушению развития конечностей. Эти результаты свидетельствуют о том, что ТАДы являются геномными регуляторными единицами, и к тому же обладают высокой степенью внутренней стабильности [3].

В ещё одной работе был сделан ряд инверсий, в результате которых кластер энхансеров из одного ТАДа переносили в другой домен, чтобы изучить патогенные последствия таких хромосомных перестроек [6]. Исследователи обнаружили возникновение эктопических контактов между энхансером и другими генами, локусы которых на Hi-C карте соответствовали локусам архитектурных полос. Авторы пришли к выводу, что архитектурные полосы, по-видимому, возникают благодаря энхансерной активности. Кроме того, они показали, что архитектурные полосы часто формируются в клетках в ходе эмбрионального развития.

Недавние исследования показали, что нарушение границ ТАДов часто встречается в раковых клетках и способствует онкогенезу при помощи двух механизмов. Один механизм приводит к локальному разрушению доменов и связан с удалением или мутацией границы ТАДа, что приводит к слиянию двух соседних доменов. Другой механизм включает в себя геномные перестройки, которые приводят к появлению новых доменов, не влияя непосредственно на уже существующие границы ТАДов [59]. Предполагается, что такие перестройки приводят к раку из-за активации онкогенов, так как

гены начинают взаимодействовать с энхансерами в новом геномном контексте, благодаря чему их уровень экспрессии увеличивается.

Однако иногда удаление сайтов CTCF на границе топологических доменов может не приводить к ожидаемым изменениям пространственной организации хроматина и фенотипическим последствиям. В одной из работ исследователи делали последовательные делеции сайтов посадки CTCF на границе ТАДов [60]. Они пришли к выводу, что часто на границе доменов наблюдается избыточность сайтов посадки CTCF и делеция отдельных сайтов посадки CTCF не приводит к изменениям пространственной архитектуры хроматина. Кроме того, даже при слиянии ТАДов взаимодействие регуляторных элементов и экспрессия соответствующих генов изменяется незначительно, что говорит о присутствии других механизмов регуляции. Другая группа на примере кластера *HoxD* генов также показала устойчивость границ ТАДов к делециям CTCF сайтов, и только большая (около 400 Кб) делеция привела к слиянию ТАДов [2].

Таким образом, безусловно пространственная архитектура хроматина играет важную роль в регуляции экспрессии генов, однако механизмы, участвующие в процессах регуляции, изучены не полностью и требуют дальнейших исследований.

1.4 Моделирование в области 3D-геномики

С момента открытия метода Hi-C было проведено много экспериментальных работ, в которых исследовали законы и механизмы, лежащие в основе пространственной архитектуры генома. Однако такие методы являются не единственным способом изучать принципы упаковки ДНК. Одним из подходов, активно используемых для этой цели, является моделирование. Моделирование позволяет проверять актуальность существующих гипотез и механизмов, предсказывать пространственную организацию генома для тех типов клеток, для которых нет прямых экспериментальных данных о пространственных контактах хроматина,

предсказывать последствия хромосомных перестроек и т.д. Более подробно области применимости моделирования в 3D-геномике обсуждаются далее по тексту, после раздела о методах моделирования.

1.4.1 Принципы и подходы *in silico* моделирования трехмерной архитектуры генома

Алгоритмы, существующие на сегодняшний день для моделирования трехмерной структуры хроматина, можно разделить на 2 основные группы: физическое моделирование, основанное на свойствах хроматина как полимера, и статистическое моделирование, основанное на поиске закономерностей между эпигенетическими и геномными характеристиками.

Физическое моделирование 3D архитектуры хроматина

Физическое моделирование основано на представлении хроматина как полимера. Законы, описывающие поведение полимеров в различных условиях, были сформулированы ещё в 1980 году в работе [61]. Было показано, что при достаточно больших размерах полимера его поведение в растворе не зависит от химической структуры мономеров, а зависит от таких параметров как концентрация мономеров, свойства растворителя и температура. При определенных параметрах системы полимер находится в одном из равновесных состояний, таких как случайный клубок, равновесная глобула и т.д. [62]. Было проведено множество исследований, в результате которых предположили, что хроматин в ядре представляет собой фрактальную глобулу. Это предположение было высказано А. Гроссбергом, С. Нечаевым и Е. Шахновичем ещё в 1988 году [63]. Этой гипотезе долгое время противостояла модель равновесной глобулы [25]. В дальнейшем, уже в 2009-2011 годах, исследователи из лаборатории Мирного наконец подтвердили, что хроматин в ядре имеет структуру фрактальной глобулы [27]. Согласно модели фрактальной глобулы, близко расположенные участки генома вероятнее всего взаимодействуют друг с другом и формируют глобулу первого порядка. Далее

из близлежащих глобул первого порядка собираются глобулы второго порядка и так далее. Такая модель позволяет избежать возникновения узлов при компактизации ДНК, а также позволяет быстро декомпактизовывать отдельные участки хроматина, не изменяя укладку соседних районов.

Развитие и модификация физических моделей шло параллельно с развитием экспериментальных технологий для исследования хроматина. Таким образом, физические модели постоянно меняются и актуализируются в соответствии с текущими представлениями. Несмотря на то, что модель фрактальной глобулы достаточно хорошо описывает пространственную структуру хроматина в ядре клетки, она не является достаточно точной, чтобы описать все взаимодействия, происходящие в ядре. Например, она не описывает локус-специфические механизмы.

Одним из таких механизмов является механизм «протягивания петли» [39]. Добавление эпигенетической информации о местах посадки белков CTCF в физические модели [39] и [40] позволяет моделировать барьеры при протягивании хроматиновой петли когезином (Рис. 8). Такие модели являются более точными и дают возможность предсказывать, как меняется компактизация хроматина при мутациях в основаниях петель, опосредованных CTCF.

Другой подход моделирования локус-специфических особенностей основан на описании взаимодействия разных типов хроматина. В этом случае полимер описывается как последовательность блоков (мономеров), где каждый блок обладает своими физическими свойствами (Рис. 8). Модели, учитывающие предпочтительные взаимодействия между мономерами одного типа, описывают как происходит пространственная сегрегации хроматина на различные компартменты [64–66]. В клетке этот процесс представляет собой жидкость-жидкостную фазовую сепарацию.

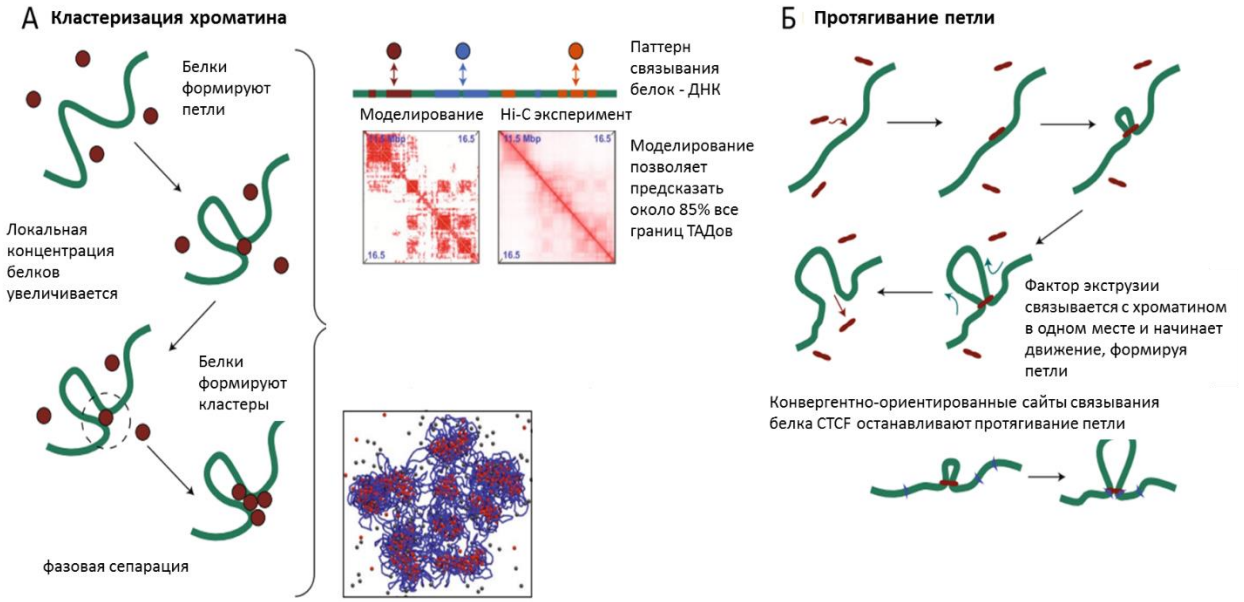


Рис. 8. Основные принципы, на которых строятся модели пространственной организации хроматина[67]. (А) Кластеризация разных типов хроматина. (Б) Механизм протягивания петли. Рисунок адаптирован из [67].

Дальнейшее усложнение физических моделей привело к добавлению в систему так называемых биндеров, молекул, которые опосредуют взаимодействия между блоками хроматина. Существуют модели, где биндеры – это абстрактные молекулы, не соответствующие реальным белкам в клетке [68,69]. Однако углубление знаний об эпигенетической информации и белках, участвующих во взаимодействиях хроматина, позволяет использовать свойства конкретных белков в физических моделях. Например, есть модели, где в качестве биндеров выступает белок HP1 [70] или белки ламины [71,72]. Однако в любом случае положения биндеров выводятся из эпигенетических данных, даже в случае использования в модели абстрактных биндеров. Например, модель Hip-Hop [73] выводит положения биндеров на основе данных о модификации H3K27ac гистонов и/или доступности хроматина. Авторы показали, что этой эпигенетической информации достаточно для предсказания пространственных взаимодействий хроматина. А в модели

PRISMR [55] данные Hi-C, полученные из клеток дикого типа, используются для определения количества типов биндеров и их аффинности. Эта информация может быть дополнительно использована для моделирования изменений 3D укладки хроматина, произошедших в результате хромосомной перестройки.

Таким образом, физическое моделирование может быть мощным инструментом как для проверки предполагаемых молекулярных механизмов, лежащих в основе 3D укладки хроматина, так и для предсказания пространственных взаимодействий хроматина на основе эпигенетических данных.

Моделирование 3D организации генома при помощи статистических методов

Другой способ моделирования пространственной организации хроматина основан на статистических методах. Известно, что паттерн определенных эпигенетических модификаций и профиль связывания факторов транскрипции коррелирует с расположением регуляторных элементов, состояниями хроматина и другими эпигенетическими характеристиками. Например, гистоновая метка H3K9me3 хорошо коррелирует с конститутивным гетерохроматином, который коррелирует с B-компартментом [74], границы ТАДов обогащены белком CTCF [37,75], а открытый активный хроматин обогащен специфическими модификациями гистонов. Таким образом, в самом простом варианте можно просто использовать линейную регрессию для прогнозирования пространственной организации генома на основе эпигенетических данных. Основанные на корреляции методы используются, например, для предсказания энхансер-промоторных взаимодействий с использованием гистоновых модификаций, CAGE, ChIP-seq и других характеристик хроматина в качестве входных данных [76].

Хотя линейные модели могут до некоторой степени объяснить особенности трехмерной организации генома, ясно, что зависимости между эпигенетическими характеристиками и трехмерной архитектурой хроматина не являются линейными. Например, известно, что падение частоты контактов между локусами с увеличением геномного расстояния не описывается линейной функцией, а наиболее близко к степенной зависимости. Нелинейной может быть также связь компартментализации, контактов хроматина, локус-специфического эпигенетического окружения и других особенностей генома.

Нелинейные взаимодействия могут быть описаны эмпирически. При помощи подбора соответствующих функций, можно описать взаимодействия между разными характеристиками генома в форме алгебраических выражений с некоторыми свободными параметрами. В работе [77] предложили алгебраическое выражение, объединяющее линейные и экспоненциальные члены, для предсказания геномных контактов на основе транскрипционных данных GRO-seq, данных о связывании белка CTCF и геномного расстояния. Были предсказаны такие основные структуры хроматина как ТАДы и петли.

Однако есть нелинейные зависимости между гистоновыми модификациями, связыванием факторов транскрипции и 3D взаимодействиями хроматина, которые не могут быть определены аналитически как алгебраические выражения. Такие зависимости возможно найти с помощью алгоритмов машинного обучения, таких как градиентный спуск, регрессия случайного леса, нейронные сети и другие [78].

Алгоритмы машинного обучения работают с численным представлением входных данных (признаков), таких как нуклеотидная последовательность, геномное расстояние или эпигенетические модификации, а также значениями целевых характеристик, таких как частота пространственных контактов между локусами, основания петель и т. д. Основным результатом алгоритма машинного обучения является функция,

которая преобразует входные данные в предсказания целевых значений. Сходство между предсказаниями и экспериментальными данными измеряется с помощью определяемой пользователем функции потерь. На этапе обучения часть данных, называемая обучающей выборкой, используется для обучения и оптимизации алгоритма так, чтобы функция потерь была минимальной. Так алгоритм находит взаимозависимости между входными и целевыми значениями. Найденные зависимости могут отражать общие биологические механизмы, а могут являться артефактами, специфичными для обучающей выборки. Кроме того, функция преобразования входных данных в предсказание целевых значений обычно имеет множество настраиваемых параметров. Иногда это приводит к тому, что выходные значения в точности повторяют обучающие данные до такой степени, что это отрицательно влияет на точность модели, использующей в качестве входа валидационные данные. В таком случае разработанный алгоритм бесполезен, поскольку найденные им зависимости не могут быть обобщены на те выборки данных, которые никогда не встречались при подборе параметров (т.е. при обучении) алгоритма.

Чтобы убедиться, что любое повышение точности по обучающей подвыборке является обобщаемым, оценку алгоритма выполняют с использованием тестовых данных (тех, которые алгоритм до этого никогда «не видел»). Очень важно, чтобы тестовая выборка не содержала образцов, представленных в обучающей выборке. Однако здесь следует отметить, что геномные объекты, не эквивалентные с математической точки зрения, могут содержать большой объем пересекающейся биологической информации. Например, вложенные друг в друга основания хроматиновых петель могут содержать большую часть общей эпигенетической информации, находящейся в «окне» между основаниями петель, хотя сами по себе основания петель не перекрываются и формально представляют собой непересекающиеся объекты. Такое косвенное перекрытие приводит к пересечению информации между наборами обучающих и тестовых данных, что приводит к переоценке точности

предсказания. Чтобы решить эту проблему, можно, например, использовать разные хромосомы для набора тестовой и обучающей выборки.

Считается, что алгоритмы на основе машинного обучения могут находить сложные нелинейные закономерности. Машинное обучение позволяет предсказывать структуры, начиная от взаимодействия двух локусов и заканчивая Hi-C картами. Не так давно было разработано несколько алгоритмов, использующих эти методы для предсказания промотор-энхансерных взаимодействий: TargetFinder [79], DeepTACT [80] и HiC-Reg [81]. Наиболее полно алгоритмы и их различия для предсказания промотор-энхансерных взаимодействия описаны в обзоре [76]. Другие пространственные структуры хроматина, такие как петли [82–86] и частоты пространственных контактов [81,87,88] также можно предсказывать с помощью алгоритмов, основанных на машинном обучении. Кроме того, подход, основанный на машинном обучении, позволяет выявить те биологические признаки, которые дают наибольший вклад в предсказание трехмерной укладки хроматина, что позволяет лучше понять биологические механизмы, лежащие в основе. Например, извлечение весов из конкретных слоев сверточных нейронных сетей помогает найти признаки, в частности, последовательности, дающие основной вклад в предсказание и, следовательно, в трехмерную структуру хроматина. Другой пример - анализ вклада всех признаков в предсказание в случае метода градиентного спуска. В этом случае алгоритм выдает ранжированный список признаков, которые внесли наибольший вклад в уменьшение функции потерь. В любом случае, анализ признаков, дающих наибольший вклад в предсказание, и параметров алгоритма может натолкнуть на гипотезу о функции каждого конкретного белка в организации пространственной архитектуры хроматина.

1.4.2 Области применения методов моделирования в 3D геномике

С помощью экспериментальных методов можно получить достаточно много информации о трёхмерной архитектуре генома, однако различные методы моделирования также позволяют изучать и выдвигать гипотезы о механизмах, лежащих в основе 3D структуры хроматина [89]. Кроме того, моделирование позволяет предсказывать пространственные контакты хроматина для типов клеток, не имеющих доступных экспериментальных данных о пространственной организации генома. Одной из прикладных целей моделирования является предсказание изменений пространственных контактов хроматина, происходящих в результате хромосомных перестроек.

Моделирование для проверки гипотез и поиска молекулярных механизмов.

Можно использовать моделирование для того, чтобы получить новые теории или проверить уже существующие гипотезы о молекулярных механизмах. Для этой цели чаще используется физическое моделирование. В течение последних нескольких лет был получен значительный объем данных, описывающих основные особенности 3D архитектуры генома и молекулярные механизмы, лежащие в основе. Открытию таких механизмов как «протягивание петли» и фазовая сепарация значительно поспособствовал такой способ моделирования [90,91]. Однако известные механизмы не объясняют все особенности 3D организации хроматина и соответственно модели, основанные на этих гипотезах, также не идеальны, поэтому требуются дальнейшие исследования.

Статистические методы, как сверточные нейронные сети, такие как, например, Akita [88] и DeepC [87] и нейронные сети трансформеры [92] позволяют находить основные характеристики нуклеотидной последовательности и эпигенетические метки, дающие вклад в 3D структуру

генома, что позволяет строить гипотезы о биологических механизмах, связанных с найденными последовательностями.

Предсказание функциональных последствий хромосомных перестроек.

Моделирование 3D архитектуры генома можно использовать для прогнозирования функциональных последствий, вызванных изменениями в трехмерной организации генома. В этих случаях моделирование трехмерной архитектуры хроматина необходимо для точного предсказания последствий геномных мутаций.

Подходы к автоматизированному анализу последствий хромосомных перестроек в контексте трехмерной архитектуры генома начали развиваться ещё в 2016 году [93]. Был предложен алгоритм, предсказывающий, какие хромосомные перестройки могут быть ассоциированы с болезнями за счет изменений в экспрессии генов, опосредованных нарушениями 3D-структуры хроматина [93]. Исследователи использовали оценку SI (Structure Influence), которая количественно определяет степень влияния хромосомной перестройки на пространственную структуру хроматина. Оценивались такие перестройки, как делеции, инверсии и дупликации.

Другая группа исследователей провела оценку количества патологий, связанных с изменением пространственной архитектуры хроматина. Они использовали базу данных Human Phenotype Ontology, чтобы соотнести фенотипы у 922 пациентов с различными делециями, зарегистрированными в базе данных DECIPHER, с фенотипами, наблюдаемыми при моногенных заболеваниях. Они обнаружили, что до 11,8% делеций лучше всего объясняются взаимодействием гена с дополнительными энхансерами или комбинацией эффекта дозы гена с эффектом новых промотор-энхансерных взаимодействий, а соответственно изменением пространственной структуры хроматина [94].

Эти открытия важны для медицинской генетики, так как интерпретация хромосомных перестроек в некодирующих областях генома всё ещё остаётся проблемой. Активно разрабатываются конвейеры программ, которые медицинские генетики могли бы использовать для анализа мутаций пациентов. Например, [95] предлагают подробные инструкции о том, как запустить последовательность программ, который идентифицирует варианты-кандидаты, способные вызвать нарушения экспрессии за счет эффектов положения. Эта последовательность программ в том числе включает в себя анализ ТАДов и возможность изменений энхансер-промоторных взаимодействий, произошедших из-за хромосомной перестройки. Благодаря таким работам анализ последствий хромосомных перестроек в контексте трехмерной структуры генома становится рутинным методом анализа. Недавно опубликованный алгоритм машинного обучения TADA [96] может ранжировать крупные хромосомные перестройки в зависимости от их патогенности.

Помимо предсказаний силы эффекта мутаций можно предсказывать изменения в трехмерных структурах генома, таких как ТАДы и петли. Алгоритм 3D-GNOME [97][98] генерирует 3D структуры хроматина с использованием подхода Монте-Карло, который использует данные захвата конформации хроматина (3C). Этот алгоритм использует ChIA-PET данные белка CTCF или РНК-полимеразы II для описания исходного паттерна контактов хроматина. Затем используется ряд простых правил, на основе которых алгоритм 3D-GNOME прогнозирует изменения, произошедшие в трехмерной структуре хроматина в связи с хромосомной перестройкой. Другой подход заключается в предсказании петель хроматина и того, как они изменятся с помощью алгоритма DeepMILO, основанного на машинном обучении [82]. Алгоритм использует непосредственно последовательности ДНК, находящиеся в основании петель, не используя экспериментальную информацию о связывании белка CTCF. Это позволяет предсказывать

образование петель, не имеющих ChIP-seq пиков CTCF в своём основании. DeepMILO может предсказывать эффекты даже небольших мутаций, и благодаря этому авторы нашли те важные петли, участвующие в инсуляции ТАДов, которые разрушаются у некоторых больных раком, что приводит к изменению работы генов, контролируемых данными структурами.

Алгоритмы, описанные выше, предсказывают изменение определенных структур хроматина, таких как петли и ТАДы. Однако есть другой тип алгоритмов, которые могут предсказывать всю Hi-C-карту трехмерных контактов мутированного локуса. Такие алгоритмы, как Akita [88], DeepC [87], PRISMR [55], 3DPolyS-Fit [99] и другие, могут предсказывать изменения в трехмерной организации хроматина, вызванные хромосомными перестройками.

Область растущего интереса и активных исследований — это влияние коротких вставок или делеций (инделов) или однонуклеотидных замен на 3D архитектуру хроматина. Известно, что даже замена одного нуклеотида может приводить к изменениям в трехмерной структуре генома, например, путем модификации сайтов связывания белка CTCF [100,101]. Таким образом, отдельная задача алгоритмов - предвидеть последствия таких мутаций. Некоторые алгоритмы, такие как DeepMILO [82] Akita [88] и DeepC [87] используют нуклеотидную последовательность в качестве основного входного параметра для предсказания. И в этом случае, такие алгоритмы очень эффективны для предсказания изменений, вызванных небольшими мутациями, поскольку мутации напрямую влияют на входные характеристики. С другой стороны, для обучения этим алгоритмам требуется знание о трехмерной организации хроматина в интактных клетках, так как нуклеотидная последовательность ничего не говорит об эпигенетических характеристиках, специфичных для конкретного клеточного типа.

Таким образом, ограничением алгоритмов, использующих нуклеотидные последовательности для предсказания трехмерной архитектуры генома, является клеточная специфичность, в то время как для алгоритмов, использующих для предсказания эпигенетические свойства локусов, это ограничение снимается. В то же время для использования этого типа алгоритмов необходимо вначале предсказать, как хромосомная перестройка повлияет на свойства генома, используемые в качестве входных данных модели, что часто является не менее сложной задачей, чем собственно предсказание изменений в трехмерных контактах хроматина.

Предсказание архитектуры хроматина как альтернатива проведения 3C-эксперимента в ранее не исследованных типах клеток

Наконец, можно использовать моделирование для предсказания трехмерной архитектуры генома для тех типов клеток, для которых ещё нет экспериментальных данных Hi-C. Для этой цели подходят алгоритмы, которым для предсказаний требуется лишь небольшой набор эпигенетических данных, доступных для многих типов клеток. В основном это алгоритмы, основанные на машинном обучении.

Таким образом, 3D геномика является активно развивающейся областью науки. Методы, которые используются для изучения механизмов, лежащих в основе пространственной архитектуры хроматина, включают в себя как традиционные «мокрые» биологические методы, так и математические и физические методы исследования. Уже сейчас открытия, сделанные в этой сфере, имеют интерес не только для фундаментальной науки, но также имеют прикладной характер и применяются, например, в медицинской генетике.

ГЛАВА 2. МАТЕРИАЛЫ И МЕТОДЫ

2.1 Подготовка данных для запуска алгоритма TargetFinder

2.1.1 Выбор регуляторных элементов (энхансеров и промоторов) для алгоритма TargetFinder

Промоторы были определены как интервал +\- 5 КБ от сайтов начала транскрипции (TSS). Для некоторых экспериментов использовали дополнительно альтернативный вариант определения промоторов: +/- 1 КБ от сайта начала транскрипции. TSS были взяты с базы данных UCSC для мышиного генома сборки mm10.

Энхансеры были взяты из статьи [102][103]. Мы выбрали энхансеры, характеризованные авторами как «poised» и «strong» и конвертировали их координаты из сборки генома мыши версии mm9 в координаты, соответствующие mm10 сборке генома с помощью инструмента LiftOver базы данных UCSC.

Альтернативные энхансеры и промоторы были взяты из статьи [103].

Были проанализировано три типа клеток мыши: эмбриональные стволовые, нейральные предшественники и кортикальные нейроны.

2.1.2 Взаимодействующие и невзаимодействующие энхансер-промоторные пары для алгоритма TargetFinder

Взаимодействующие пары энхансер-промотор были выделены с использованием данных Hi-C для трех типов клеток мыши: эмбриональных стволовых клеток (ЭСК), нейральных предшественников, зрелых нейронов [104]. Пара энхансер-промотор нами считалась взаимодействующей, если промотор и энхансер были расположены в основаниях одной и той же хроматиновой петли. Петли были выделены на основе Hi-C данных программой Juicer HiCCUPS [105] с рекомендованными авторами параметрами (-m 512 -r 5000,10000 -k KR -f .1,.1 -p 4,2 -i 7,5 -t 0.02,1.5,1.75,2 -d 20000,20000,50000). Затем, взаимодействующие пары были разбиты на 5 бинов (равных групп) в зависимости от расстояния между энхансером и

промотором (прил. 1). На бины разбивали так, чтобы в каждый бин попадало одинаковое количество взаимодействующих пар. Невзаимодействующие пары были набраны в тех же бинах так, чтобы в каждом бине невзаимодействующих пар было в 20 раз больше, чем взаимодействующих. Альтернативные взаимодействующие пары энхансер-промотор были взяты из статьи [103].

2.1.3 Выбор взаимодействующих пар энхансер-промотор на основе баз данных SlideBase и GeneHancer

База данных SlideBase (http://slidebase.binf.ku.dk) поддерживается консорциумом FANTOM5 (http://fantom.gsc.riken.jp/data/) [106] и представляет собой карту регуляторных элементов для разных типов клеток человека. В базе данных содержится информация об уровне экспрессии энхансеров, полученная методом CAGE для более чем 200 линий раковых клеток и более чем 200 первичных линий клеток. Мы выбирали энхансеры для моноцитов человека, которые присутствовали в более чем 25% образцах, относящихся к выбранной линии клеток.

Базу данных GeneHancer (https://www.genecards.org) мы использовали для выбора промоторов, регулируемых конкретными энхансерами. GeneHancer — это база данных полногеномных ассоциаций энхансеров и промоторов с генами.

Мы получали список взаимодействующих промотор-энхансерных пар путем объединения энхансеров из базы данных SlideBase и промоторов из списка ассоциаций энхансер-ген, полученного из базы данных GeneHancer.

2.1.4 Параметризация признаков для алгоритма TargetFinder

Для каждой пары промотор-энхансер генерировался вектор признаков. Признаки включали следующую информацию:

1) Расстояние между промотором и энхансером.

2) ChIP-seq и DNAse-seq данные для используемого типа клеток (все наборы данных были взяты из [79] https://github.com/shwhalen/targetfinder). Для каждого типа ChIP-seq данных генерировалось число, отражающее суммарное распределение ChIP-seq пиков между промотором и энхансером. В качестве значения, характеризующего ChIP-seq пик, мы использовали значение signalValue (мера обогащения прочтениями участка ДНК) из bed файла, затем все значения суммировались в «окне» между промотором и энхансером.
3) RNA-seq данные. Для RNA-seq данных также генерировался суммарный сигнал, отражающий уровень транскрипции между промотором и энхансером. Мы использовали значения FPKM (fragments per kilobase of exons per million mapped reads) для всех генов, находящихся в «окне» между промотором и энхансером и суммировали их. Более подробно то, как генерировались признаки, описано в результатах.

2.2 Визуализация и анализ Hi-C данных

Hi-C матрицы представляют собой тепловые карты, где каждая точка отражает частоту контактов между двумя бинами (локусами генома) (Рис. 9). Чем точка краснее, тем частота контактов выше. Бины – это равные участки, на которые поделен геном, внутри которых все прочтения суммируются. Размер бина определяет разрешение Hi-C карты. Известно, что частота контактов между локусами экспоненциально падает в зависимости от линейного расстояния между ними. Часто бывает важно выделить локальное повышение частоты контактов, которое не может быть объяснено геномным расстоянием между ними. В этом случае для каждой диагонали матрицы считается ожидаемая частота контактов на данном расстоянии (expected) и каждое наблюдаемое (observed) значение частоты контакта делится на эту величину. В результате получается так называемая OoE (observed over expected) карта контактов (Рис.

9Б). Hi-C матрицы могут быть представлены в виде симметричной матрицы, как на Рис. 9. В этом случае матрицы контактов выше и ниже главной диагонали абсолютно идентичны и отражают одну и ту же информацию, по оси X и Y отложены геномные координаты. Ещё один способ визуализации Hi-C карт – это представление их в виде треугольников, как, например, на Рис. 6А. При таком способе визуализации используется только половина симметричной матрицы, которая поворачивается на 45 градусов.

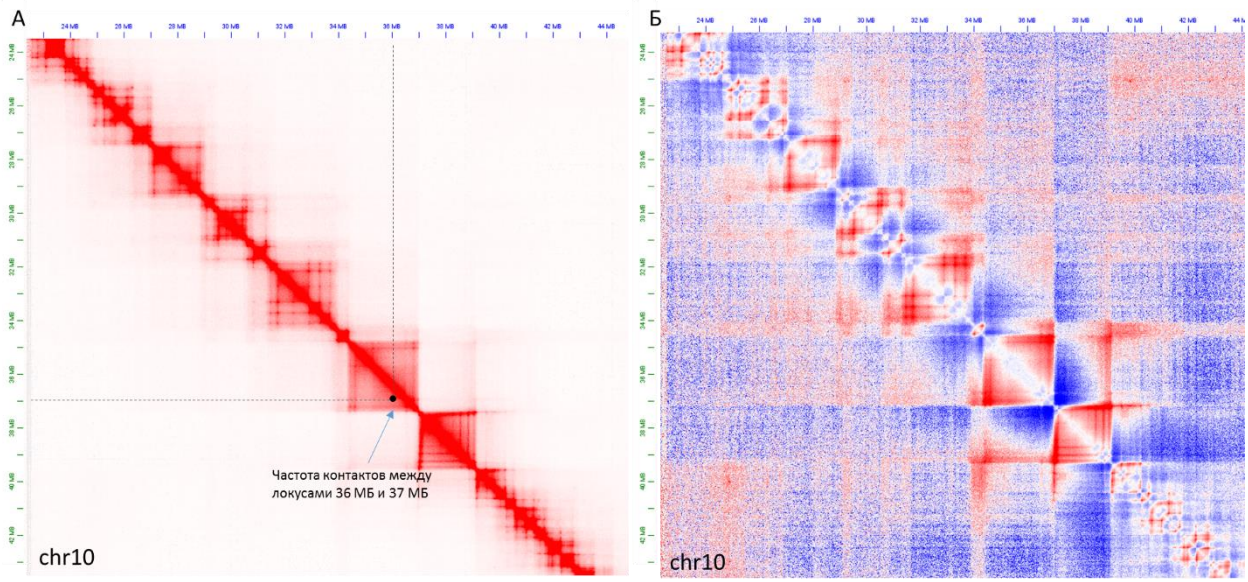


Рис. 9. Пример визуализации Hi-C данных для участка chr10:23000000-44000000 клеток нейральных предшественников мыши из [104] на разрешении 20 Кб. (А) Наблюдаемые частоты контактов (Б) OoE (observed over expected) значения контактов.

2.3 Параметризация признаков для алгоритма 3DPredictor

Для того чтобы использовать вычислительные алгоритмы предсказания контактов хроматина на основе эпигенетических и других биологических данных, необходимо сначала представить биологическую информацию в виде численных параметров, т.е. параметризовать входные данные. Мы использовали следующую параметризацию.

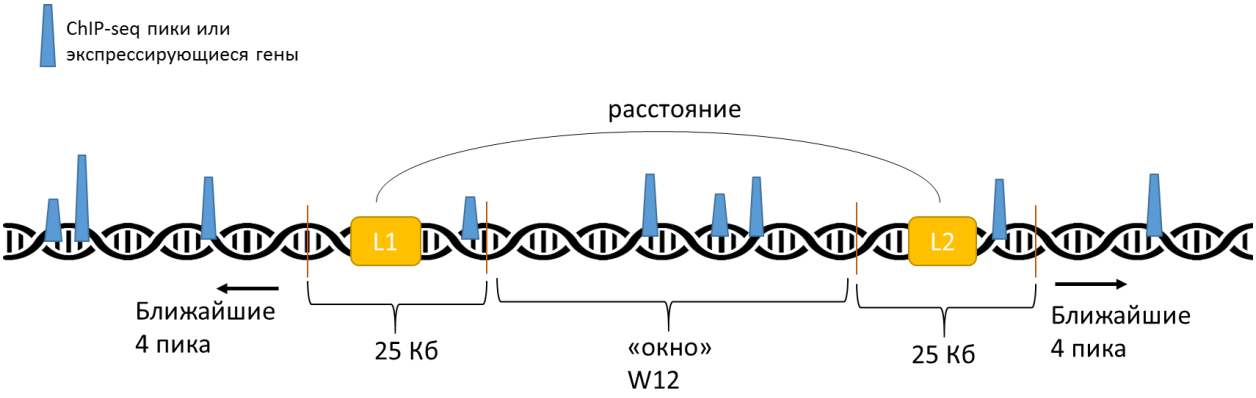
Для каждого контакта локусов L1 и L2 мы использовали геномные данные, находящиеся в окне W12 между локусами и по разные стороны от них (Рис. 10). Для параметризации данных о связывании белка CTCF с ДНК были сгенерированы несколько групп признаков. Признаки, описывающие связывание снаружи от локусов L1 и L2: 1) расстояние до L1 и величины signal value (стандартная характеристика интенсивности связывания белка с ДНК в ChIP-seq данных) для 4-х пиков (участков связывания), расположенных по левую сторону от региона L1 (8 действительных чисел); 2) расстояние до L2 и signal value для 4-х пиков, расположенных по правую сторону от L2 региона (8 действительных чисел).

Признаки, описывающие связывание белка CTCF с ДНК в окне между локусами L1 и L2: 3) сумма signal value для всех пиков внутри W12 между локусами L1 и L2 (одно действительное значение).

Признаки, описывающие локусы L1 и L2: 4) сумма всех значений signal value внутри 25-КБ региона, центрированного вокруг локусов L1 и L2 (два действительных значения). Также мы использовали расстояние между L1 и L2 как отдельный признак (одно целое число).

Для RNA-seq данных использовалась такая же параметризация, как и для ChIP-seq данных, но вместо значения signal value использовалась стандартная мера оценки уровня экспрессии генов FPKM (fragments per kilobase of exons per million mapped reads).

Также мы использовали данные об ориентации мотивов связывания белка CTCF. Мы параметризовали эту информацию следующим образом: I) Для каждого участка связывания белка CTCF, параметризованного как в 1) и 2), мы добавляли значение идентичности мотива связывания белка CTCF (степень уверенности, что в данном месте есть мотив связывание белка), вычисленное при помощи инструмента GimmeMotifs [76], для прямого и обратного направления цепи ДНК; II) количество конвергентно



ориентированных сайтов связывания белка CTCF в окне между L1 и L2.

Рис. 10. Схематичное описание параметризации признаков для алгоритма 3DPredictor. L1 и L2 – отдельные локусы на хромосоме. W12 – геномный регион между двумя локусами. Высота ChIP-seq пиков определяется значением signal value для данного пика.

2.4 Обработка ChIP-seq данных

Все ChIP-seq данные были обработаны с использованием ENCODE aquas-pipeline (https://github.com/kundajelab/chipseq_pipeline). Ссылки на сырые данные в базе данных SRA представлены по адресу https://github.com/genomech/3DGenBench/blob/stable/rearrangements_table.tsv.

2.5 Обработка RNA-seq данных

Все данные были обработаны стандартными протоколами с использованием HISAT2 [107], deeptools bamCoverage [108], Stringtie [109].

Некоторые данные были обработаны с использованием web-платформы Galaxy (usegalaxy.org).

2.6 Процессирование данных для web-платформы 3DGenBench

2.6.1 Генерирование данных с разным уровнем шума для тестирования метрик платформы 3DGenBench

Данные с разным уровнем шума были сгенерированы путём добавления случайного значения δ_{i_j} к каждому значению частоты контакта w_{ij} из экспериментальной матрицы Hi-C контактов:

$$\tilde{w_{ij}} = w_{ij} + \delta_{i_j}, \quad (1)$$

где $\tilde{w_{ij}}$ - новое значение частоты контакта. Случайное значение δ_{i_j} было выбрано из нормального распределения со средним, равным наблюдаемой частоте контакта и стандартным отклонением в диапазоне от 0.1% от наблюдаемой частоты контакта (самый низкий уровень шума) до 5000% от наблюдаемой частоту контакта (самый высокий уровень шума).

$$\delta_{i_j} \in N(w_{ij}, \mathrm{k}^* w_{ij}), \quad (2)$$

$$k = [0.0001, 0.001, 0.01, 0.5, 1, 2, 10, 20, 50]$$

Для тестирования метрик «горизонтального» типа сравнения мы генерировали модели с разным уровнем шума только для мутантной матрицы частот контактов.

2.7 Метрики для оценки качества предсказаний пространственной архитектуры хроматина

Для оценки точности предсказания трёхмерной архитектуры генома алгоритмами мы использовали 2 типа сравнения. Метрики из первого типа сравнения («горизонтального») показывают, насколько предсказанная матрица частот контактов похожа на экспериментальную. Метрики для второго типа сравнения («вертикального») отражают, насколько хорошо

алгоритмы предсказывают изменения в 3D организации хроматина, произошедшие в результате мутации.

Для «горизонтального» типа сравнения использовались:

- Корреляция Спирмана между предсказанной и экспериментальной Hi-C матрицей
- SCC [80], функция для корреляции Hi-C карт на отдельных геномных расстояниях, реализованная в пакете hicreppy (https://github.com/cmdoret/hicreppy)
- Корреляция Спирмана инсуляции каждого бина для предсказанных и экспериментальных данных. Инсуляторный профиль посчитан функцией *calculate_insulation_score* из пакета coolltools.api.insulation (версия 0.5.0)
- Сила компартментализации посчитана как в [110]. Сила компартментализации CS_{b_i} для каждого бина b_i определяется как:

$$CS_{b_i} = \frac{\frac{1}{n_c}\sum_{j=1}^{n_c} OoE_j,}{\frac{1}{n}\sum_{g=1}^{n} OoE_g} \quad (3)$$

Гдеj обозначает номер бина, принадлежащий тому же компартменту, что и b_i, n_c - количество бинов, принадлежащих тому же компартменту, что и бин b_i, а n – общее количество бинов. CS_{b_i} <=1 отражает случай, когда компартментализации нет, CS_{b_i}>1 означает, что компартментализация есть. Затем мы использовали коэффициент корреляции Спирмана между полученными треками силы компартментализации для всех бинов в качестве результирующей метрики.

- Метрика оценки частоты контактов на разных геномных расстояниях. Мы подсчитали среднюю частоту контактов на отдельных геномных расстояниях (каждой диагонали матрицы)

для экспериментальных и предсказанных данных. Для полученных массивов значений считали корреляцию Спирмана.

Для «вертикального» типа сравнения использовали:

- Изменение профиля инсуляции. Профиль инсуляции отражает расположение границ ТАДов. Изменение инсуляции ΔIS считали следующим образом:

$$\Delta IS = \frac{IS_{mut}}{IS_{wt}} \quad (4)$$

Где IS_{mut} профиль инсуляции для мутантного фенотипа, посчитанный функцией *calculate_insulation_score* из пакета coolltools.api.insulation (версия 0.5.0), а IS_{wt} профиль инсуляции для дикого типа, посчитанный так же.

Полученные массивы значений для предсказанных и экспериментальных данных коррелировали методом Спирмана.

- Эктопические взаимодействия для экспериментальных и предсказанных данных были рассчитаны как в [55]. А именно:

I. Мы нормализовали мутантную матрицу частот контактов по покрытию, не используя области делеции и дупликации для расчёта коэффициента нормализации. Пусть M это Hi-C матрица для случая мутации с элементами m_{ij}, тогда W это матрица для дикого типа с элементами w_{ij}. Мы посчитали нормализованную матрицу для мутации таким образом:

$$\check{M} = \frac{M * \sum_{i,jdel,dup} w_{ij}}{\sum_{i,jdel,dup} m_{ij}} \quad (5)$$

В случае дупликации мы рассчитывали дополнительный коэффициент нормализации:

$$D_{coef} = \frac{\sum_{i \in dup} w_{ij}}{\sum_{\substack{i \in dup \\ j \in dup}} \check{m}_{ij}} \quad (6)$$

Где $\check{m}_{ij}$ являются элементами матрицы $\check{M}$.

Далее мы получили нормализованную мутантную матрицу:

$$\tilde{M} = \check{M} * D_{coef} \quad (7)$$

Где $D_{coef} = 1$для делетированных и инвертированных регионов.

Эти этапы нормализации были сделаны как для экспериментальных, так и для предсказанных данных.

II. На втором шаге мы вычли матрицу W из матрицы $\tilde{M}$.

III. Затем мы нормализовали полученную матрицу различий, где каждое значение разделили на среднее значение частоты контакта на том же геномном расстоянии.

IV. Наконец мы посчитали Z-оценки на каждой диагонали матрицы. Мы не использовали значения, превышающие 96 перцентиль для вычисления стандартных отклонений на каждой диагонали.

Мы определили эктопические взаимодействия как те, которые превышают Z-оценку, равную двум. Эти эктопические взаимодействия мы подавали на вход функции *precision_recall_curve* из python библиотеки sckit-learn как истинные значения. Предсказанную матрицу с Z-оценками подавали вторым аргументом функции *precision_recall_curve*. Показатель AUC также был рассчитан функцией из библиотеки sckit-learn.

2.8 Программное обеспечение

Весь код написан на языке python (версия 3.6). Для высокопроизводительных вычислений использовался информационный вычислительный комплекс Новосибирского Государственного Университета (http://www.nusc.ru/).

2.9 Доступность кода

Скрипты, необходимые для запуска алгоритма 3DPredictor, доступны по ссылке https://github.com/labdevgen/3Dpredictor. Исходный код, разработанный для web-платформы 3DGenBench находится по адресу https://github.com/genomech/3DGenBench, адрес разработанной нами платформы - https://inc-cost.cytogen.ru:8230/ или https://inc-cost.eu/benchmarking .

ГЛАВА 3. РЕЗУЛЬТАТЫ

В этой работе мы использовали методы машинного обучения как основной инструмент для изучения трёхмерной архитектуры генома. Первая часть главы результатов посвящена использованию этого подхода для предсказания карт 3D контактов хроматина с целью изучения последствий хромосомных перестроек и предсказания трехмерной архитектуры генома в норме.

Второй раздел главы результатов посвящён разработке web-платформы для удобной и унифицированной оценки точности алгоритмов, предсказывающих трёхмерную архитектуру хроматина.

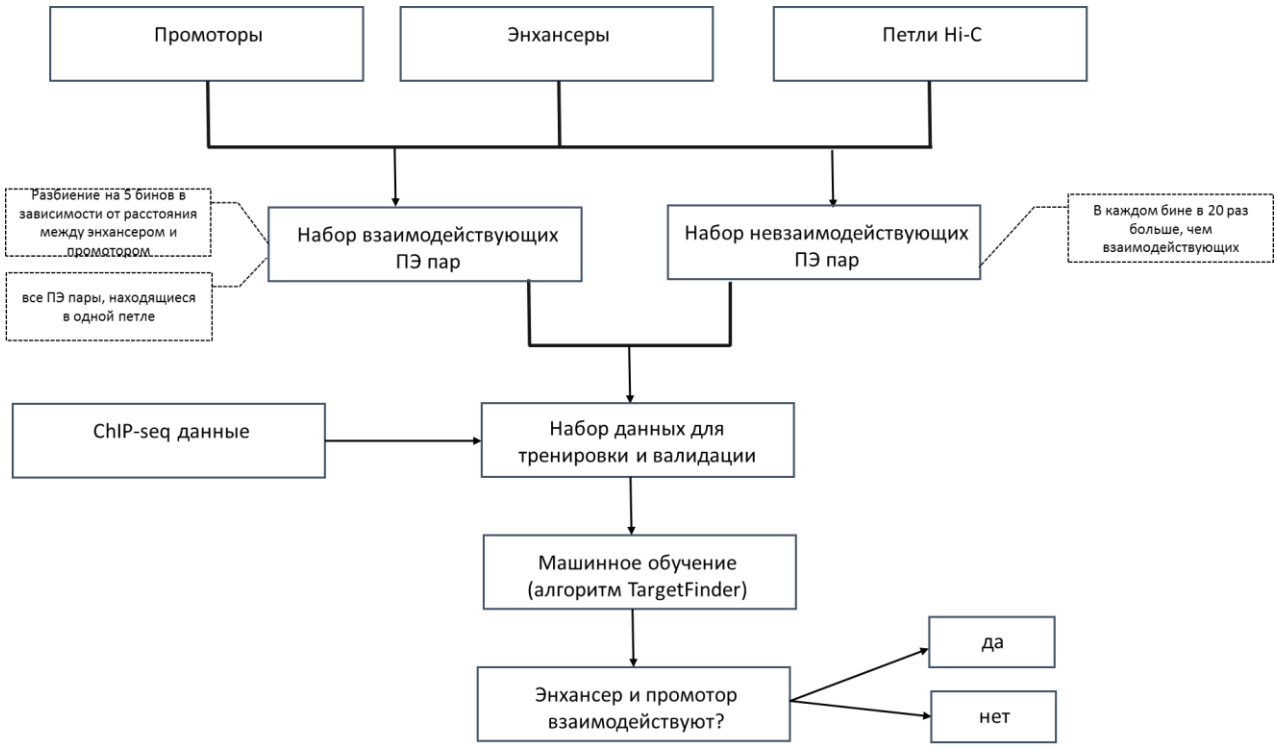
3.1 Применение и анализ алгоритма TargetFinder для предсказания промотор-энхансерных взаимодействий.

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 пространственных контактов хроматина, в первую очередь промотор-энхансерных взаимодействий, сопровождающих хромосомные перестройки. Хотя в обзоре литературы перечислено достаточное количество алгоритмов [81,82,85–88] прямо или косвенно решающих поставленную задачу, на момент начала работы для этой задачи подходил только один алгоритм TargetFinder [79], опубликованный в журнале *Nature Genetics*.

Алгоритм TargetFinder является бинарным классификатором, основанным на градиентном бустинге, таким образом, данный инструмент качественно (не количественно) отвечает на вопрос о взаимодействии конкретной пары энхансер-промотор. На вход алгоритму подаётся информация о белках, связанных с промотором, энхансером и «окном»

(участком ДНК) между ними (Рис. 11, раздел «Материалы и Методы» 2.1.4). Для обучения и предсказания используются эпигенетические характеристики, доступные для данного типа клеток, такие как ChIP-seq данные, отражающие распределение сайтов связывания различных белков с ДНК, информация о доступности хроматина (DNAse-seq анализ) и данные об экспрессии генов. TargetFinder был разработан для предсказания промотор-энхансерных взаимодействий для таких активно изучаемых линий клеток человека как GM12878, K562, NHEK и т.д. Мы решили адаптировать этот алгоритм для мышиной модели, чтобы иметь возможность проверить полученные предсказания экспериментально *in vivo*. Было выбрано три типа клеток мыши с наиболее представленными данными ChIP-seq и Hi-C картами высокого разрешения: эмбриональные стволовые клетки, клетки предшественники нейронов и кортикальные нейроны [104].

Мы сделали обучающую и валидационную выборку из взаимодействующих и невзаимодействующих энхансер-промоторных пар, определяя взаимодействующие регуляторные элементы таким же образом, как это описано в статье [111] (раздел «Материалы и Методы» 2.1.2). Если кратко,



то энхансер и промотор считались взаимодействующими, если они находились в основании одной петли, выделенной на основе Hi-C данных.

Рис. 11. Краткая схема подготовки данных для алгоритма машинного обучения TargetFinder. ПЭ – промотор-энхансерные пары

Для оценки точности предсказания алгоритма использовались такие стандартные оценки качества работы бинарных классификаторов как доля верных ответов, полнота, точность и f-мера. Обучив алгоритм TargetFinder на данных для мышиных клеток, мы обнаружили, что точность работы алгоритма достаточно низкая (f-мера~0.15-0.2) по сравнению с оригинальной статьёй (f-мера~0.7-0.85). Переобучив алгоритм для линии клеток GM12878, мы также наблюдали разницу в точности алгоритма по сравнению с оригиналом (Табл. 1).

Табл. 1. Алгоритм TargetFinder показывает низкую точность предсказаний по сравнению с оригинальной работой. В таблице представлена оценка точности работы алгоритма TargetFinder для разных типов клеток, f-мера посчитана для разных наборов обучающих выборок, выборка, где классы представлены в соотношение 1:1 или 1:20.

Клеточный тип	Количество признаков	Количество петель	Количество взаимодействующих Э-П пар	Способ разбиения данных для обучения и валидации
Мышиные эмбриональные клетки	24	9091	1602	Эта работа
				Оригинальная статья
Мышиные клетки кортекса	10	9972	625	Эта работа
				Оригинальная статья
Мышиные клетки нейральные предшественники	10	9360	635	Эта работа
				Оригинальная статья
GM12878	100	9448	2113	Эта работа
				Оригинальная статья

Для того чтобы увеличить точность предсказания, мы варьировали входные параметры. Например, использовали промоторы разной длины: длиной не более 2 Кб или промоторы не более 10 Кб, также использовались как промоторы активные в данном типе клеток, так и все известные промоторы. Пробовали использовать альтернативные определения энхансеров (координаты энхансеров из статьи [79] или из результатов сегментации генома

алгоритмом HMM ([102]; https://github.com/gireeshkbogu/chromatin_states_chromHMM_mm9). Варьировали размер обучающей и тестовой выборок, для того чтобы количество образцов в обучающей выборке было как можно больше. Кроме того, мы увеличили количество взаимодействующих пар. Для этого мы использовали не только петли, которые выделяются программой Juicer [105], но также пары локусов, для которых на Hi-C картах наблюдается обогащение контактами. Однако ни один из этих вариантов значительно не улучшил точность предсказаний. Балансировка классов (взаимодействующие и не взаимодействующие пары) в обучающей и валидационной выборке позволила увеличить f-меру (Табл. 1), однако точность алгоритма все ещё была ниже, чем в оригинальной работе.

Тщательное исследование различий в методе создания выборки для обучения и тестирования показало, что разная точность предсказания алгоритма TargetFinder в оригинальной работе и у нас связана с особенностью распределения энхансер-промоторных пар в обучающей и валидационной выборке. 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 тестирования (10% от всего набора). Это является логичным с математической точки зрения, однако такой способ генерирования наборов данных для обучения и теста приводит к переоценке обобщающей способности алгоритма, так как в таком дизайне косвенно признаки между выборками пересекаются.

Во-первых, среди всего набора промотор-энхансерных пар существует достаточно большое количество таких случаев, где промотор является общим для нескольких пар. Наличие такие пар обусловлено тем, что с одним

промотором взаимодействует множество расположенных рядом друг с другом энхансеров и наоборот. С точки зрения принципа составления выборки взаимодействующих промотор-энхансерных пар, это логично: размер основания Hi-C петли определяется разрешением Hi-C данных и составляет в нашем случае 5 Кб, а размер энхансера — около 250 п.н., поэтому все энхансеры, лежащие на расстоянии менее 5 Кб друг от друга, будут иметь одинаковый паттерн взаимодействий. Если в оригинальных данных для клеток человека убрать дубликаты промоторов и энхансеров, то алгоритм предсказывает промотор-энхансерные взаимодействия с меньшей точностью (Рис. 12). Такая тенденция прослеживается как для линии клеток человека GM12878, так и для мышиной эмбриональной линии клеток.

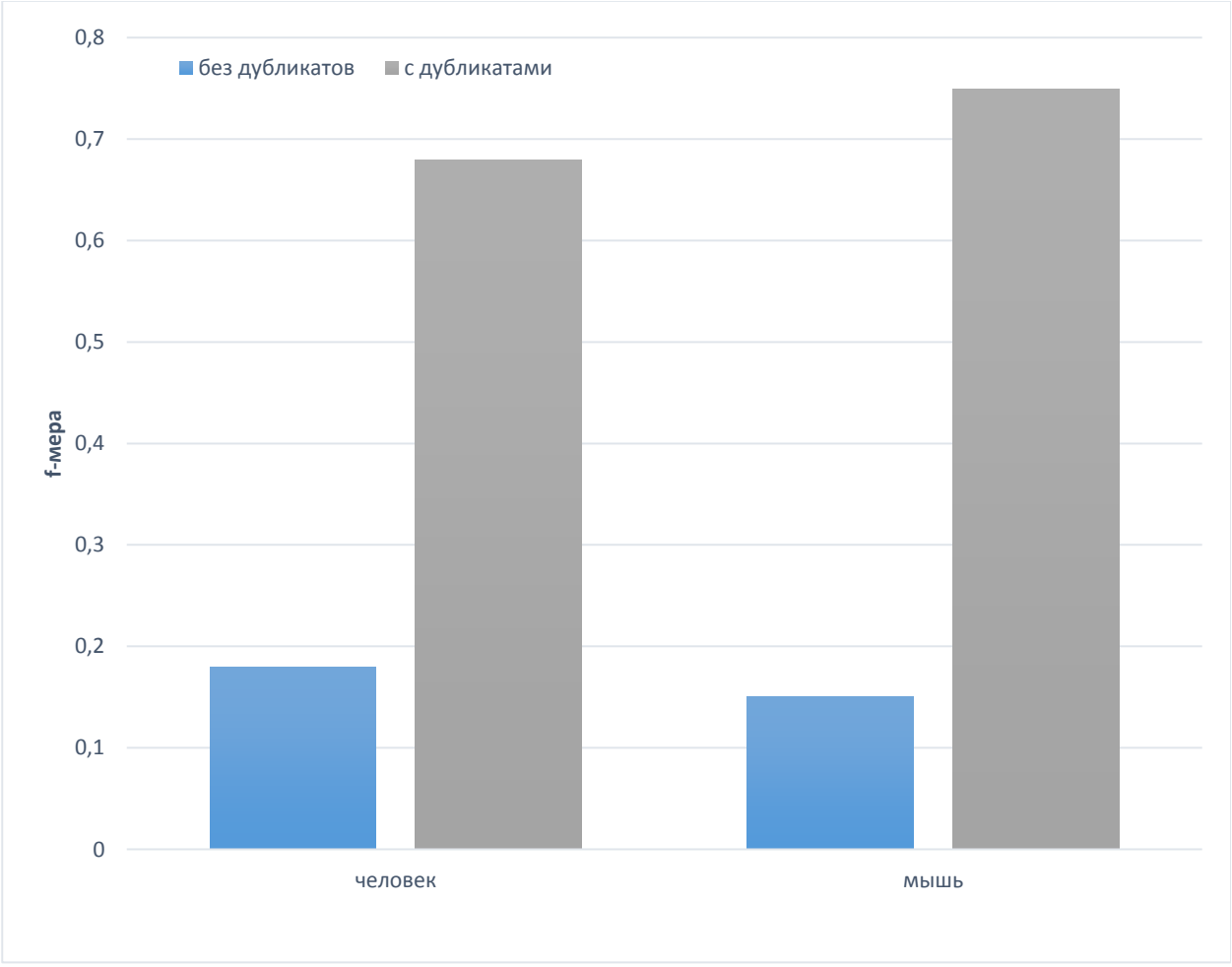


Рис. 12. f-мера уменьшается при наличии повторяющихся промоторов и энхансеров в обучающей и тестовых выборках.

Одна из причин снижения точности предсказания могла бы быть связана с тем, что количество образцов в обучающей выборке уменьшилось и алгоритм на таком небольшом наборе данных не может выявить закономерности. Однако мы проверили, что при использовании обучающей и тестовой выборки одинакового размера у выборки с дубликатами точность предсказаний существенно выше (Рис. 13), что говорит о том, что уменьшение размера обучающей выборки не является определяющим фактором в снижении точности алгоритма. Также мы отметили, что точность предсказания зависит ещё и от количества признаков, участвующих в обучении, при увеличении количества признаков f-мера растёт (Рис. 13).

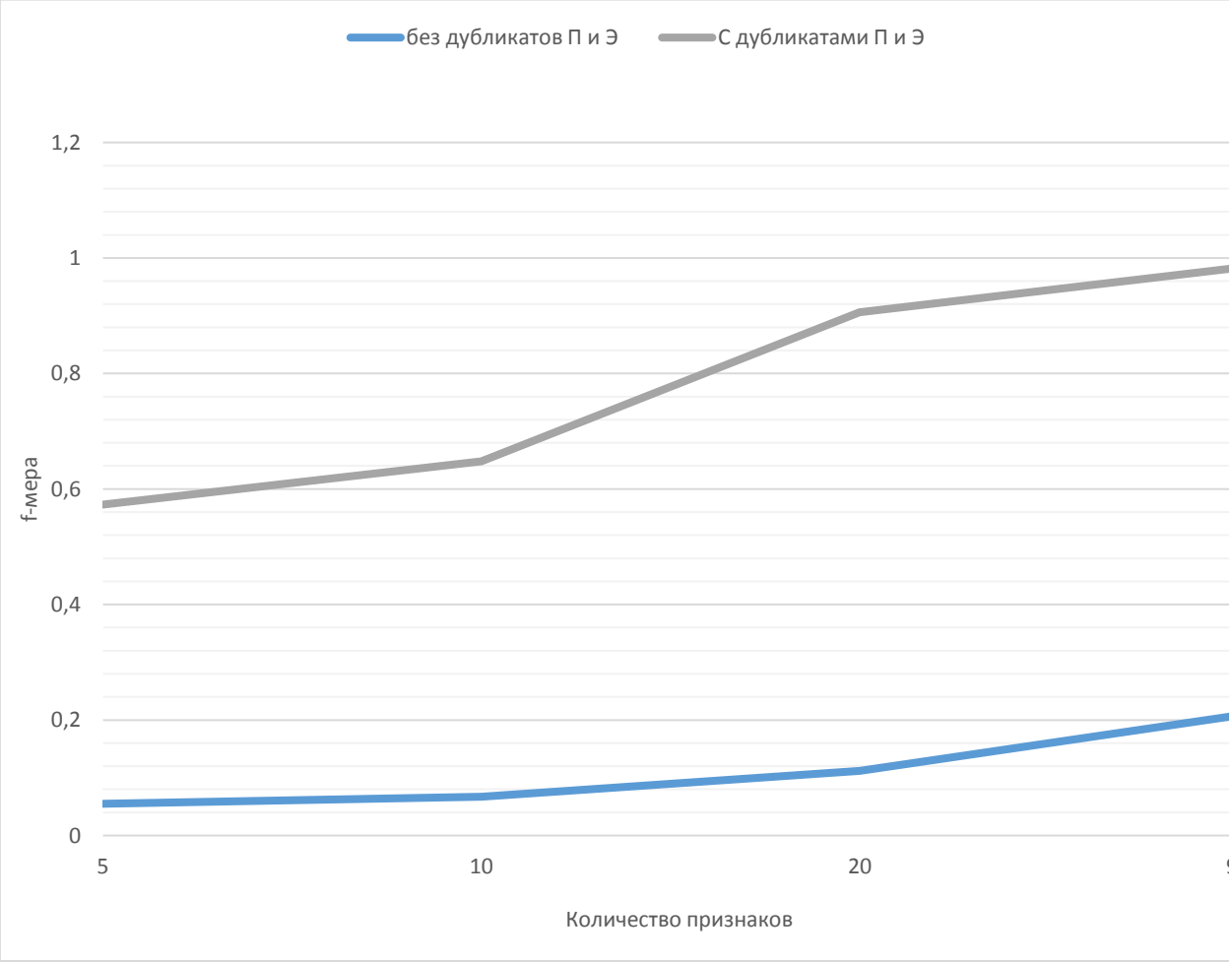


Рис. 13 f-мера изменяется в зависимости от количества признаков для обучающих выборок одного размера, содержащих (серая линия) или не содержащих (синяя линия) дубликаты промоторов и энхансеров.

Во-вторых, кроме прямых пересечений промоторов и энхансеров в разных парах, существуют также косвенные пересечения признаков между парами энхансер-промотор. Энхансеры и промоторы часто находятся рядом друг с другом, и соответственно такие пары имеют общий геномный регион между ними (общую часть «окна» между ними). А поскольку в качестве признаков мы используем ChIP-seq данные, описывающие связывание белков в «окне», то вектора признаков для разных энхансер-промоторных пар в этом случае сильно пересекаются. Таким образом, эпигенетические признаки, характеризующие «окно» для этих пар, не являются независимыми, и энхансер-промоторные пары с пересекающимся «окном» не должны быть включены в выборки для обучения и валидации одновременно.

К таким же выводам параллельно с нашей группой пришли коллеги из университета Вашингтона, опубликовавшие заметку об алгоритме TargetFinder в журнале *Plos Computational biology* [112], где они также отметили проблему пересекающихся признаков для алгоритмов машинного обучения, используемых в 3D-геномике.

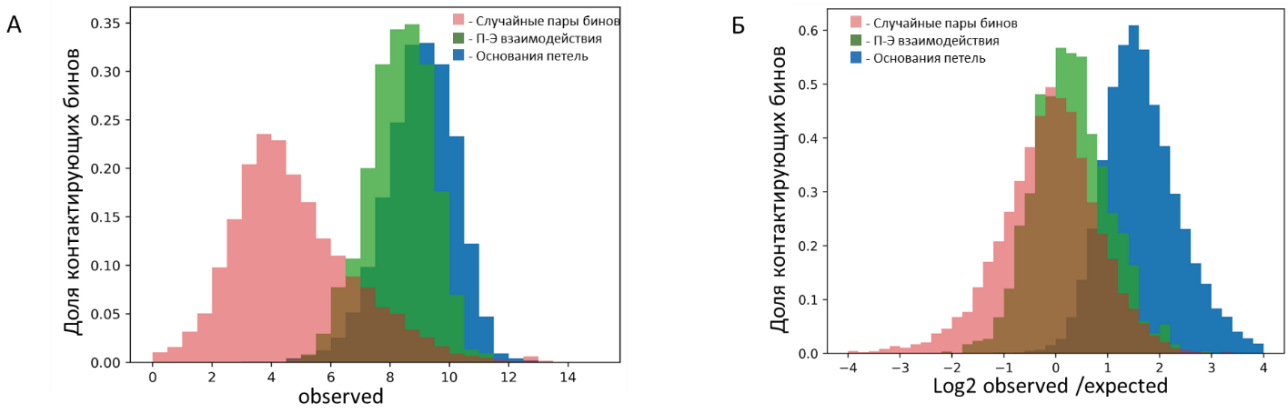
В итоге стало ясно, что алгоритм TargetFinder в существующем дизайне не способен улавливать комплексные биологические закономерности между эпигенетическими характеристиками и трёхмерной архитектурой генома, и соответственно он не подходит для предсказания промотор-энхансерных взаимодействий. Используя часть идей, предложенных в работе об алгоритме TargetFinder, мы задались целью разработать альтернативный алгоритм для предсказания пространственной организации генома.

3.2 Разработка алгоритма 3DPredictor для предсказания Hi-C карт пространственных контактов хроматина.

3.2.1 Схема работы алгоритма 3DPredictor.

При разработке нового алгоритма, мы пересмотрели определение промотор-энхансерных взаимодействий. Как в случае алгоритма TargetFinder, так и во многих других работах [80,113,114], энхансер-промоторные пары определялись как взаимодействующие, если находились в основание одной Hi-C петли. Мы решили проверить насколько справедливо это предположение. Для этого мы собрали все взаимодействующие промотор-энхансерные пары для моноцитов человека на основе баз данных SlideBase и GeneHancer (раздел «Материалы и Методы» 2.1.3. Мы выбрали моноциты человека, так для этого типа клеток доступны как Hi-C карты контактов с хорошим разрешении [115], так и достаточно полная информация о взаимодействии регуляторных элементов в базах данных SlideBase и GeneHancer. Для определения взаимодействующих регуляторных элементов базы данных SlideBase и GeneHancer используют информацию о совместной экспрессии промоторов и регуляторных элементов. Таким образом, мы могли сравнить два метода определения взаимодействующих промотор-энхансерных пар: регуляторные элементы, находящиеся в основаниях петель Hi-C, и регуляторные элементы, взаимодействующие на основе данных о коэкспрессии.

Оказалось, что большинство взаимодействующих (по базам данных SlideBase и GeneHancer) энхансер-промоторных пар часто не находятся в основании одной петли. В то же время частоты контактов между взаимодействующими (по данным SlideBase и GeneHancer) энхансер-промоторными парами, а также между основаниями Hi-C петель превышают среднюю частоту контактов по геному (Рис. 14 А, Б). Таким образом, количественная характеристика – частота Hi-C-контактов - позволяет лучше описать промотор-энхансерные взаимодействия, чем качественная характеристика (наличие или отсутствие петли). Кроме того, само определение петли, используемое в работах по поиску промотор-энхансерных взаимодействий, прямо связано с распределением Hi-C-контактов: основания петли выделяются специальным алгоритмом как участки, обогащенные Hi-C-



контактами. Таким образом, информация о распределении контактов является более полной, чем информация о расположении петель.

Рис. 14. Промотор-энхансерные взаимодействия, выделенные на основе SlideBase и GeneHancer, далеко не всегда находятся в основании одной петли. (А) Данные для наблюдаемых частот контактов в бинах, которые содержат взаимодействующие (по версии SlideBase и GeneHancer) промоторы и энхансеры (зелёная гистограмма), бины, которые пересекаются с основаниями петель (синяя гистограмма) и частоты контактов для случайных бинов (розовый). (Б) Те же данные для значений OoE (Observed over Expected ratio), отражающих превышение частоты контактов над средним для данного

геномного расстояния (см. раздел «Материалы и Методы» 2.4 для более подробной информации об OoE)

Сделанные наблюдения привели нас к выводу о том, что результатом работы предсказательного алгоритма должна стать частота контактов для каждой пары энхансер-промотор, что уже является решением задачи регрессии, а не классификации. В связи с этим, целью следующего этапа нашей работы стала разработка алгоритма, 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.

3.2.2 Разработка алгоритма 3DPredictor

Наша идея заключалась в том, чтобы, сопоставив эпигенетические данные и значение частоты пространственных контактов, определить закономерности, на основе которых можно было бы количественно предсказывать вероятность взаимодействия промотора и энхансера. 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). Мы создали непересекающиеся выборки для обучения и валидации модели. Так, например, мы использовали четные хромосомы для обучения, а нечетные для валидации и наоборот, либо обучали алгоритм на одной хромосоме, а предсказывали частоты контактов на всех оставшихся.

Нами был сгенерирован набор данных для обучения и валидации, где каждой паре локусов генома соответствовал свой вектор признаков. В

качестве признаков мы использовали эпигенетические характеристики, находящиеся между двумя контактирующими локусами генома. Кроме того, мы добавили эпигенетические характеристики, находящиеся на некотором отдалении от контактирующих бинов (подробная параметризация признаков представлена в разделе «Материалы и методы» 2.5). Такой выбор признаков обоснован механизмами, обуславливающими формирование трёхмерной организации хроматина. Например, в соответствии с механизмом «протягивания петли» [39] образование петель обусловлено остановкой белка когезина на местах посадки белка CTCF, находящихся в конвергентной ориентации, что в свою очередь приводит к пространственному сближению локусов генома, находящихся между ними. Поэтому важно учитывать эпигенетическую информацию не только внутри двух контактирующих локусов генома (бинов размером 5-25 Кб), но также между ними и снаружи от них. Мы варьировали параметризацию признаков и остановились на том, что для предсказания используются три ближайших ChIP-seq пика снаружи от двух контактирующих локусов и все ChIP-seq пики в «окне» между ними. Кроме того, в качестве информации, важной для предсказания, мы добавили признаки, отражающие ориентацию сайтов посадки CTCF (Рис. 15).

Мы назвали набор разработанных нами программ для подготовки данных, обучения моделей и предсказания частот контактов хроматина с использованием обученной модели инструментом ***3DPredictor***.

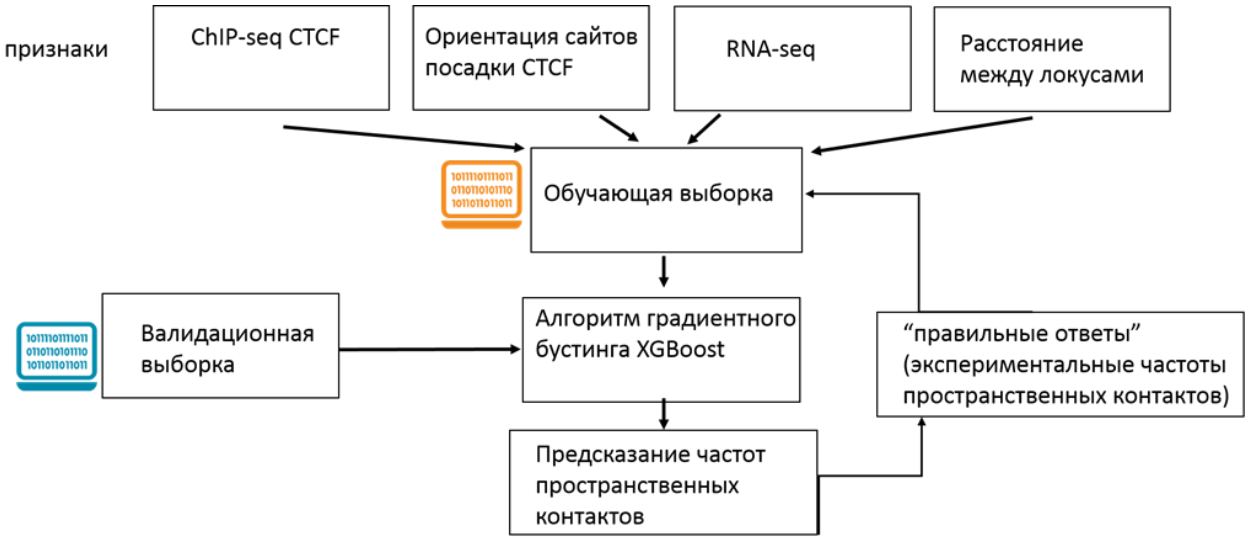


Рис. 15. Схематическое представление работы алгоритма 3DPredictor.

Чтобы выбрать лучший набор признаков, мы варьировали набор используемых эпигенетических характеристик и оценивали точность работы алгоритма. Для оценки точности работы алгоритма использовались такие метрики как корреляция Пирсона, средне квадратичное отклонение (MSE), средняя абсолютная ошибка (MAE), средняя относительная ошибка (MRE), а также SCC – метрика, специфичная для сравнения Hi-C карт (в разделе «Результаты» 3.2.3 более подробно описаны метрики для оценки качества предсказаний). Мы использовали линии клеток человека и мыши с наиболее представленными эпигенетическими данными для тестирования набора признаков для обучения. Для линии клеток человека GM12878 мы взяли все доступные данные с портала ENCODE, содержащие ChIP-seq, RNA-seq, DNAse-seq данные и данные о метилировании ДНК. Несмотря на то, что модель, обученная на наборе из всех 96 признаков, показала наилучший результат, анализ вклада каждого признака (анализ важности признаков из python библиотеки *sckit-learn*) показал, что ChIP-seq CTCF, RNA-seq и расстояние между двумя локусами являются наиболее важными признаками, влияющими на предсказание (Табл. 2). Это согласуется с биологическими представлениями, так как белок CTCF играет главную роль в формировании доменов и петель, расстояние между локусами хорошо коррелирует с частотой

контактов, а RNA-seq отражает экспрессию генов, что косвенно указывает на то, к каком компартменту (активному или неактивному) относится данный локус генома. Таким образом, не имея технических возможностей перебрать все возможные комбинации признаков, мы выбрали наиболее важные с биологической точки зрения и с точки зрения оценки вклада каждого признака (Рис. 16).

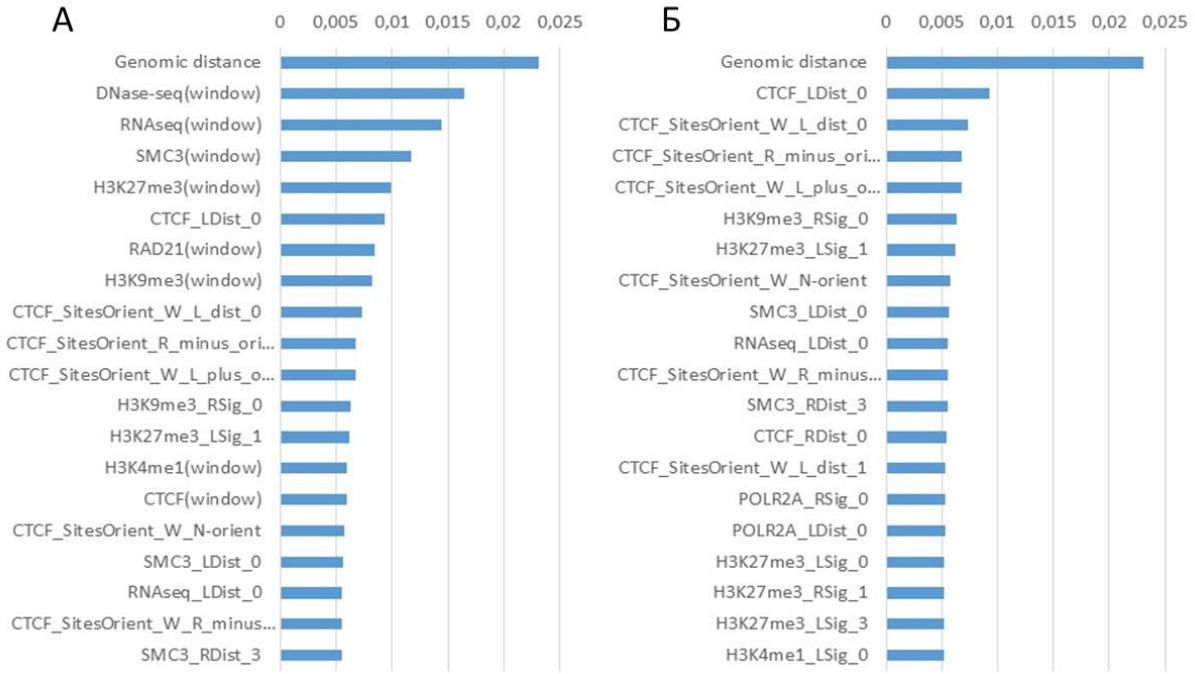


Рис. 16. Оценка вклада признаков в предсказание для линии клеток K562. Получена с использованием стандартных функций оценки вклада признаков (feature importances) из python библиотеки XGBoost. (А) 20 признаков с наибольшим значением важности признака, при этом признаки, отражающие информацию в «окне» между локусами обозначены «window». (Б) 20 признаков с наибольшим значением важности признака, за исключением признаков, кодирующих информацию в «окне».

Табл. 2. Использование комбинаций разных признаков указывает на то, что CTCF, расстояние между локусами и RNA-seq играют наибольшее значение для предсказания. Предсказания сделаны для линии клеток человека GM12878.

Размер обучающей выборки	Районы для обучения	Признаки	Районы для предсказания	Pearson correlation	SCC	MSE	MAE	MRE
1 850 471	Chr2	all	Chr3	0,9707	0,6403	0,0013	0,0007	0,3288
1 850 471	Chr2	CTCF, CTCF_orientation, RNA-seq, contact_distance	Chr3	0,9685	0,6068	0,0013	0,0007	0,3387
1 850 471	Chr2	NFYB	Chr3	0,7912	0,0139	0,0041	0,0025	1,3259
1 850 471	Chr2	CTCF, CTCF_orientation, contact_distance	Chr3	0,9673	0,5769	0,0014	0,0008	0,3488
1 850 471	Chr2	FAIRE-seq	Chr3	0,8932	0,1903	0,0024	0,0014	0,5813
1 850 471	Chr2	DNAse and methylation	Chr3	0,8778	0,2362	0,0025	0,0015	0,7037

На примере линии клеток человека K562 для Hi-C карт на разрешении 5 Кб мы более систематически сравнили точность алгоритма для разных наборов признаков (Рис. 17). Стоит отметить, что предсказание модели, обученной на всех доступных эпигенетических данных с ENCODE, практически не отличается от предсказания алгоритма, обученного на 3 основных, выбранных нами признаках (CTCF, RNA-seq, геномное расстояние), что также наблюдалось и для линии клеток GM12878 (Табл.2). Предсказания, сделанные с помощью модели, обученной на каждом из этих признаков по отдельности, не отличаются высокой точностью, а белок CTCF играет ключевую роль в точности предсказания частот контактов.

В результате многих тестов, варьируя признаки в обучающей выборке, 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. Таким образом, в конечной версии алгоритма 3DPredictor используются три этих признака, параметризованные так, как это описано в разделе «Материалы и Методы» 2.4.

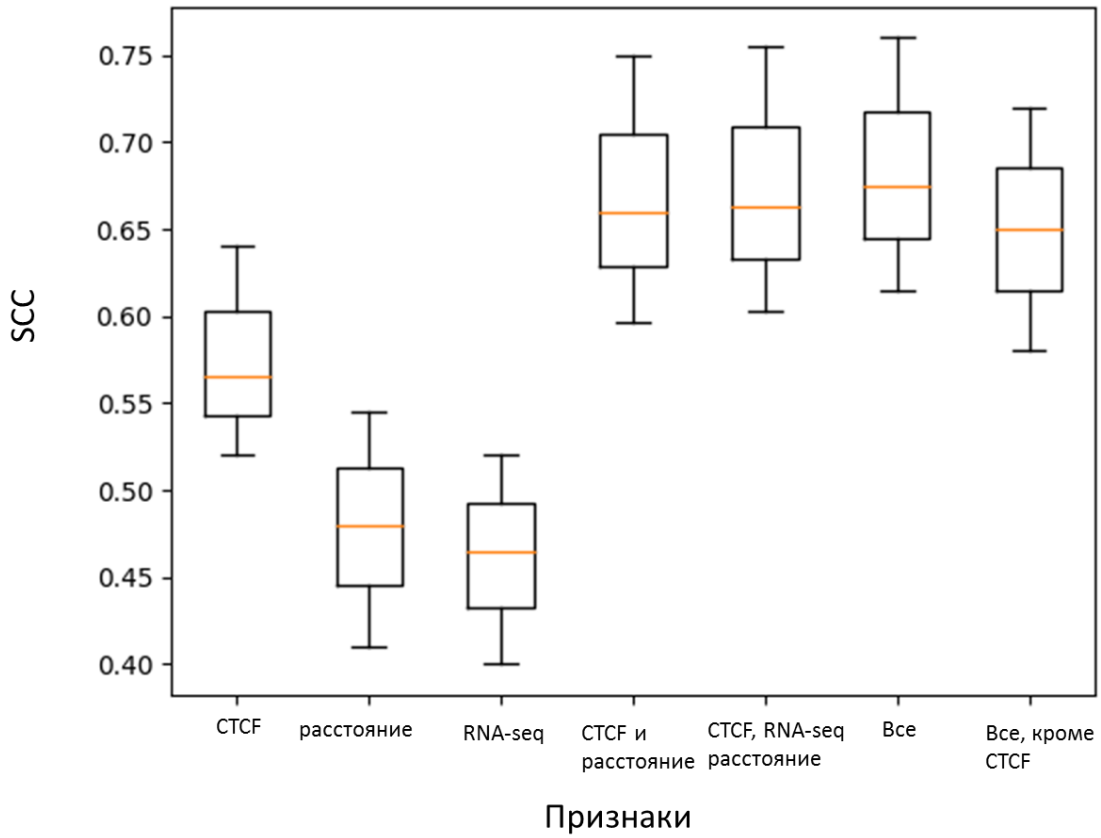


Рис.17. 3DPredictor показывает высокую точность (SCC=0.67) для модели, обученной на трех основных признаках. (RNA-seq, CTCF, геномное расстояние). На графике показана зависимость метрики SCC от набора использованных при обучении признаков. Чем выше SCC, тем лучшим считается предсказание. Боксплоты отражают предсказания, сделанные для разных хромосом (chr4, chr10, chr14) человека линии клеток K562.

Кроме изменения набора признаков в обучающей выборке, мы пытались улучшить точность предсказания путем увеличения размера обучающей

выборки (Табл. 3). Можно заметить, что точность алгоритма растет при увеличении размера обучающей выборки, однако при размере обучающей выборки больше 250 000 контактов точность предсказания значительно не увеличивается. Кроме размера обучающей выборки, также важен набор хромосом и вариабельность данных, используемых при обучении. Так предсказание, сделанное на модели, обученной на одной хромосоме хуже, чем предсказание, сделанное с использованием такого же размера обучающей выборки, но содержащее несколько хромосом для обучения (Табл. 3).

Табл. 3. Размер и характер обучающей выборки влияют на точность работы алгоритма. Предсказания были сделаны для хромосомы 2 линии клеток K562.

Размер обучающей выборки	Хромосомы для обучения	Признаки	Хромосомы для валидации	Корреляция Пирсона	SCC	MSE	MAE	MRE
25 000	Хромосомы 1,5,9,13	CTCF, CTCF_orientation, RNA-seq, contact_distance	Хромосома 2	0,9389	0,7235	0,0022	0,0011	0,7051
25 000	Хромосома 10	CTCF, CTCF_orientation, RNA-seq, contact_distance	Хромосома 2	0,9378	0,6610	0,0022	0,0012	0,7127
250 000	Хромосомы 1,5,9,13	CTCF, CTCF_orientation, RNA-seq, contact_distance	Хромосома 2	0,9451	0,7359	0,0021	0,0011	0,6746
250 000	Хромосома 10	CTCF, CTCF_orientation, RNA-seq, contact_distance	Хромосома 2	0,9257	0,7148	0,0000	0,0012	0,713
500 000	Хромосомы 1,5,9,13	CTCF, CTCF_orientation, RNA-seq, contact_distance	Хромосома 2	0,9459	0,7356	0,0020	0,0011	0,6702

Ещё один способ, с помощью которого мы пытались улучшить точность предсказания, связан с изменением типа значений частот контактов, которые алгоритм предсказывает на выходе. Так в приведенных выше экспериментах мы обучали 3DPredictor на наблюдаемой частоте контактов между двумя локусами (контакты были нормализованы с учетом количества прочтений), и предсказывали соответственно тоже наблюдаемую частоту контактов. Известно, что частоты контактов хорошо коррелируют с геномным расстоянием. Для того, чтобы учесть паттерны частот контактов, опосредованные другими факторами, можно оценить, насколько наблюдаемые частоты контактов отклоняются от ожидаемой для данного геномного расстояния частоты (отношение наблюдаемой частоты к

ожидаемой, или observed over expected ratio, OoE). Поскольку, зная геномное расстояние между локусами, значения OoE могут быть трансформированы в значения наблюдаемых частот контактов простым линейным преобразованием, мы смогли сравнить модели между собой. Эксперименты показали, что при обучении на OoE данных точность алгоритма не увеличилась, и даже была ниже, чем при обучении на наблюдаемых частотах контактов (Рис. 18).

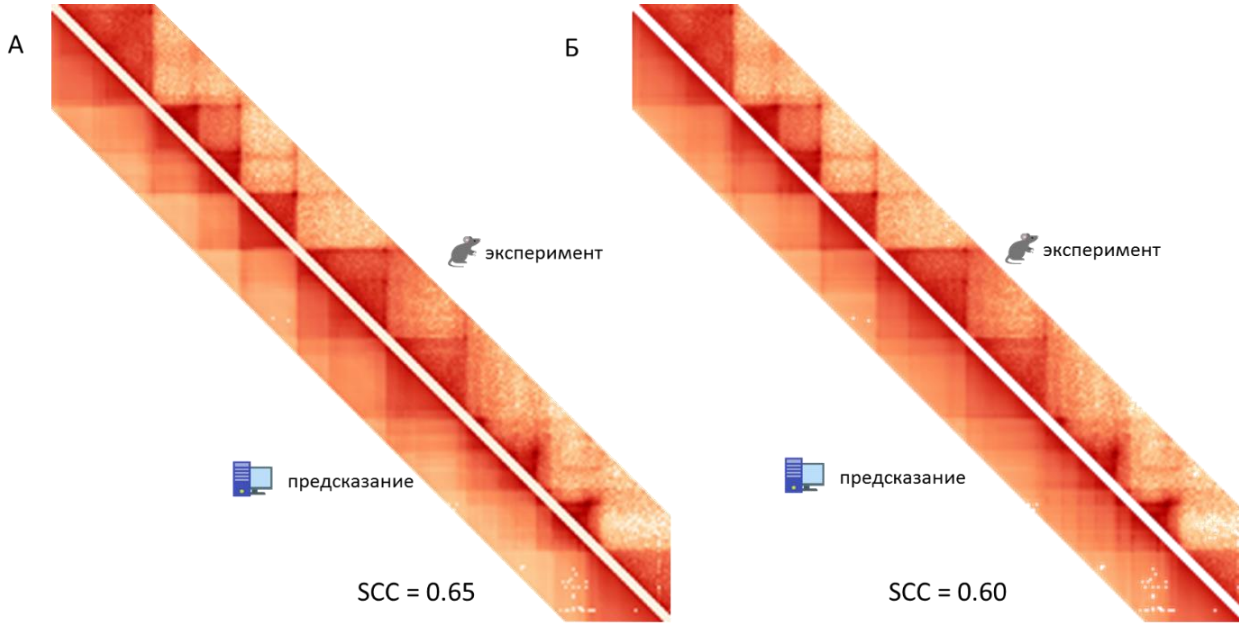


Рис. 18. Изменения типа данных, используемых для обучения, не приводит к увеличению точности алгоритма: (А) модель, обученная на значениях наблюдаемых контактов; (Б) модель, обученная на значениях контактов OoE. Модель обучена на 5 хромосоме с использованием таких признаков как CTCF, RNA-seq, геномное расстояние. Предсказание сделано для хромосомы 2 гепатоцитов мыши на разрешении 5Кб. На обоих рисунках сверху представлена карта Hi-C, полученная в результате эксперимента, снизу – предсказание.

3.2.3 Оценка точности 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. Поскольку карта Hi-C – это матрица чисел, можно использовать любые метрики, применимые для сравнения матриц. Однако лыюбые данные имеют свои специфические особенности, Hi-C данные в том числе. Важно, чтобы выбранный способ оценки точности алгоритма отражал качество предсказания основных структур пространственной архитектуры хроматина.

Основная 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. Кроме того, абсолютные значения метрик мало о чём говорят, поэтому важно использовать некоторые базовые значения, относительно которых оценивается точность предсказания. Например, можно сравнить, насколько корреляция между предсказанными и экспериментальными данными отличается от корреляции между экспериментальными репликами. В случае высокой точности алгоритма, предсказание должно быть настолько похоже на экспериментальные данные, насколько реплики между собой. Однако реплики не всегда доступны. Кроме того, в работе [1] показано, что корреляция Пирсона между случайными образцами имеет тот же порядок, что и различия между репликами (Рис. 3 в [1]).

Другим базисом для сравнения может быть сравнение Hi-C карт между разными типами клеток. Вообще 3D организация хроматина достаточно консервативна для разных типов клеток [37,38], и даже между разными видами млекопитающих Hi-C карты очень похожи [116]. Поэтому достаточно сложно с

хорошей точностью предсказывать клеточно-специфичные пространственные структуры хроматина, поскольку алгоритм должен улавливать даже небольшие различия. Для хорошего алгоритма стоит ожидать, что разница между предсказанными и экспериментальными данными для целевого типа клеток будет меньше, чем разница между Hi-C картами разных типов клеток.

Чтобы преодолеть ограничения, связанные с использованием стандартной корреляционной метрики при сравнении Hi-C карт, в работе [1] была предложена метрика, которая минимизирует влияние шума и отклонений путем сглаживания матрицы Hi-C. Кроме того, она устраняет эффект зависимости частоты контактов от расстояния путем стратификации данных Hi-C в соответствии с их геномным расстоянием. Как было показано, эта метрика, названная SCC (stratum-adjusted correlation coefficient), улавливает даже небольшие различия в 3D организации между близкородственными клеточными линиями, биологическими репликами и псевдорепликами (Рис. 3 в [1]).

Кроме вышеупомянутых способов сравнения Hi-C матриц, мы использовали такие стандартные метрики как средняя абсолютная ошибка (MAE), средняя квадратичная ошибка (MSE) и средняя относительная ошибка (MRE). Как и в случае корреляции Пирсона, эти метрики требуют сравнения с базовыми значениями. И, наконец, мы визуализировали предсказанную Hi-C карту, чтобы быть уверенными, что выбранная метрика действительно отражает различия между картами контактов.

Ещё один способ оценить производительность алгоритма – это оценка точности предсказания конкретных трехмерных структур хроматина, таких как ТАДы и петли.

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 19 представлено предсказание инструмента 3DPredictor, обученного на данных, сгенерированных для гепатоцитов мыши.

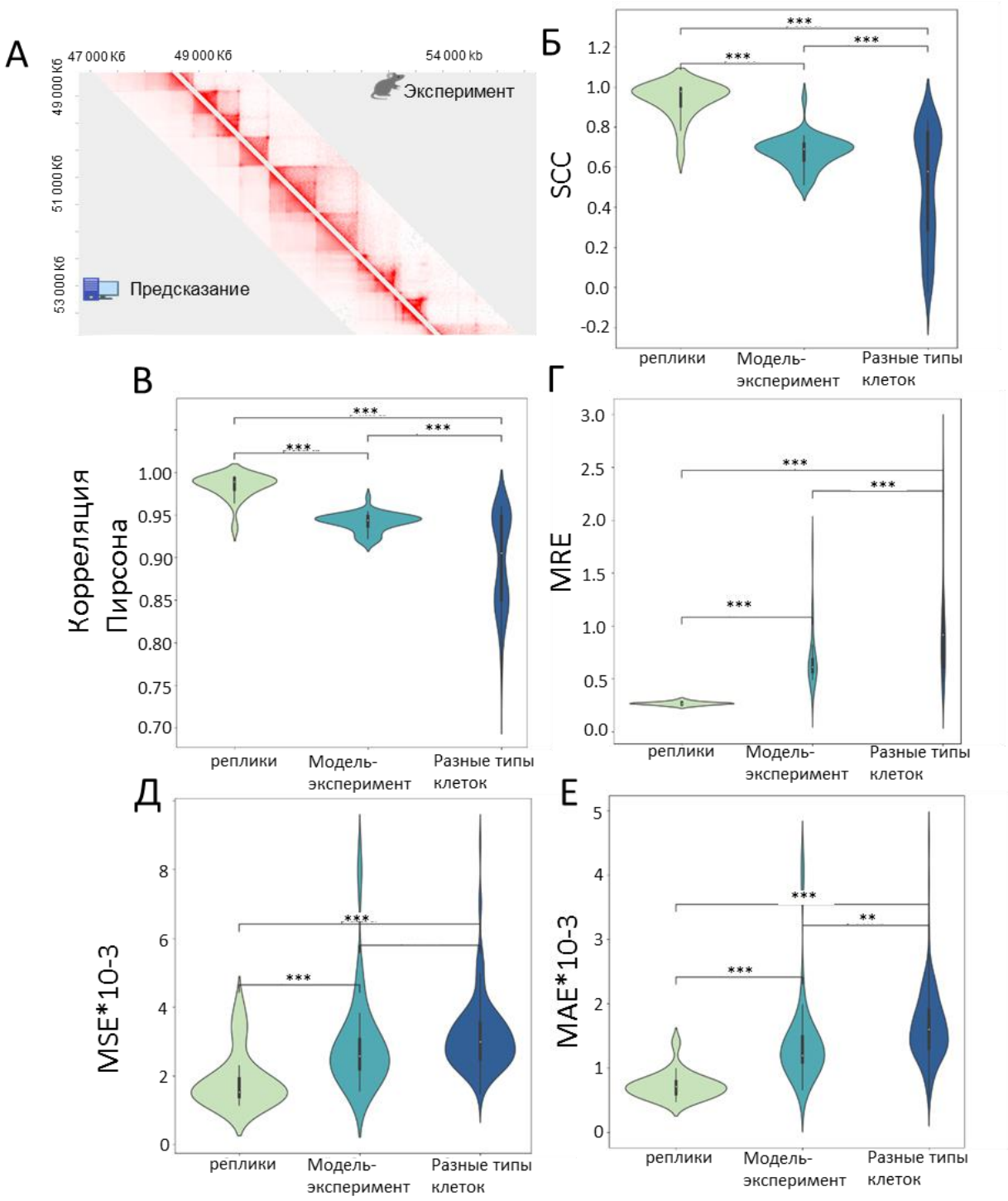


Рис. 19. Алгоритм 3DPredictor демонстрирует высокую точность предсказания пространственной организации генома для гепатоцитов мыши.

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.

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. 19). В качестве базиса для сравнения использовались реплики Hi-C данных гепатоцитов и Hi-C данные других типов клеток. Стоит отметить, что предсказанные значения находятся на уровне различий между разными типами клеток, а по некоторым метрикам, таким как MSE, MRE и MAE, приближаются к различиям между репликами (Прил. 4). Также мы проверили точность работы алгоритма 3DPredictor на данных для человеческой линии клеток GM12878 (Прил. 2). Для этой линии клеток все метрики, за исключением SCC, показывали хорошие результаты, по значению приближающиеся к уровню различий между репликами, чем различия Hi-C карт между разными клеточными типами, даже при обучении только на одной хромосоме. В то же время результаты по метрике SCC оказались хуже для этого типа клеток по сравнению с данными для мыши.

Известно, что млекопитающие имеют консервативные механизмы, участвующие в организации пространственной архитектуры генома. Соответственно если мы будем знать закономерности между распределением сайтов посадки транскрипционных факторов, эпигенетическими модификациями гистонов и пространственной укладкой ДНК, можно будет предсказывать паттерн пространственных контактов хроматина для тех типов

клеток, для которых не был проведен Hi-C эксперимент. Для того чтобы проверить, способен ли 3DPredictor улавливать эти закономерности, мы обучили наш алгоритм на данных для мышиных гепатоцитов и предсказали пространственную укладку генома для клеток предшественников нейронов мыши (Рис. 20). По основной метрике SCC, использующейся нами для оценки качества предсказания, предсказание карты частот контактов для клеток предшественников нейронов находится на уровне предсказания частот контактов для гепатоцитов, а для некоторых хромосом даже выше. Таким образом, даже не имея данных Hi-C, а имея информацию только лишь о сайтах связывания белка CTCF, транскрипции и геномном расстоянии, мы можем предсказать паттерн контактов для многих типов клеток.

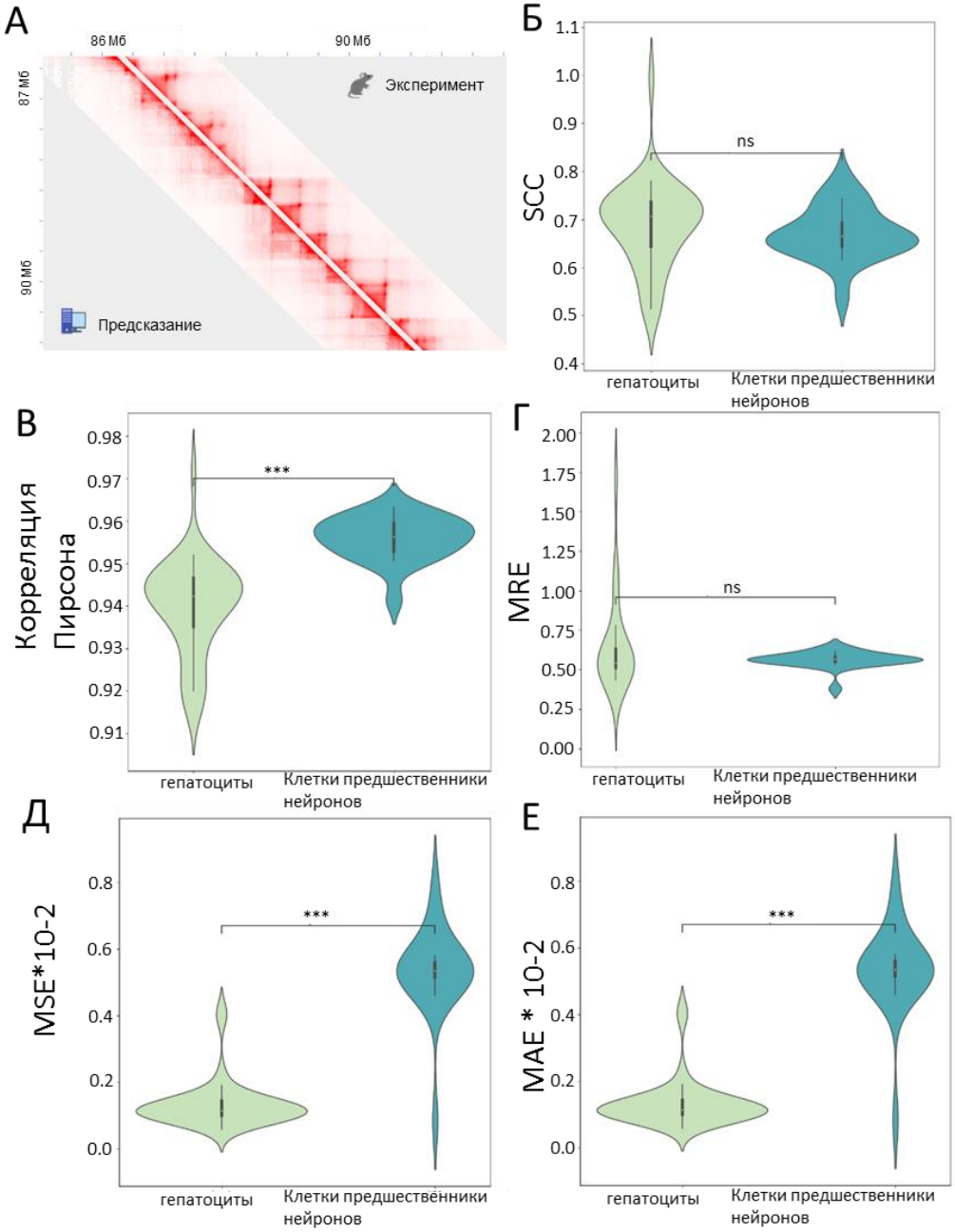


Рис. 20. Предсказание пространственной организации генома для клеток предшественников нейронов мыши имеет сходную точность с предсказанием карты частот контактов для того же типа клеток, на котором обучали (метрики SCC и MRE). Модель обучена на данных для гепатоцитов мыши на четных и нечетных хромосомах. (А) Карта частот контактов предсказанная (снизу) и полученная экспериментально (сверху) для локуса chr2:85000000-92000000 клеток предшественников нейронов на разрешение 5 Кб. (Б-Е) График зелёного цвета отражает значения метрик, полученные при сравнении предсказанных контактов клеток гепатоцитов мыши с

экспериментальными данными. График синего цвета отражает значения метрик, полученные при сравнении предсказанных контактов клеток предшественников нейронов мыши с экспериментальными данными. Среднее значение получено как среднее из значений метрики для каждой хромосомы. В качестве статистического критерия использовался t-критерий Стьюдента.

Для того чтобы избежать переобучения алгоритма, мы использовали разные хромосомы в обучающей и тестовой выборке. Например, мы обучали алгоритм на одной хромосоме и предсказывали контакты на других хромосомах, а также пробовали обучать алгоритм на половине хромосом и предсказывать контакты для локусов другой половины генома. Было замечено, что точность алгоритма при обучении на разных хромосомах немного различается, то есть некоторые хромосомы использовать для обучения лучше, чем другие.

Для более широкого и удобного применения алгоритма 3DPredictor мы разработали web версию инструмента (https://github.com/genomech/Web_3DPredictor). Для того, чтобы сгенерировать предсказания контактов хроматина, необходимо загрузить результаты RNA-seq эксперимента и ChIP-seq данные о связывании белка CTCF для интересующего типа клеток в bed формате, либо можно использовать ссылку на данные с таких популярных источников, как портал ENCODE. Созданный web-ресурс позволяет исследователям легко воспользоваться нашим инструментом и быстро получить предсказание Hi-C контактов для того типа клеток, который их интересует.

3.2.4 Предсказание основных структур трёхмерной организации хроматина алгоритмом 3DPredictor.

Как упоминалось ранее, в случае предсказания Hi-C карт важно оценить, насколько точно предсказываются различные структуры пространственной организации хроматина. Алгоритм 3DPredictor предсказывает такие

структуры как ТАДы с хорошим соответствием, часто предсказанные границы доменов совпадают с экспериментальными данными, как видно на рисунках 19, 20 (А).

Другая структура трехмерной организации генома, предсказания которой интересны с биологической точки зрения – это петли. Мы аннотировали петли с помощью инструмента HICCUPS (https://github.com/aidenlab/juicer/wiki/HiCCUPS), а также провели ручную аннотацию петель, поскольку далеко не все петли выделяются методом HICCUPS, и, кроме того, по техническим причинам этот инструмент нельзя применить к картам, полученным алгоритмом 3DPredictor. В большинстве случаев аннотированные вручную петли согласуются с аннотациями HiCCUPS (Рис. 21 Б). Различия между петлями, аннотированными вручную и методом HiCCUPS связаны с вычислительными артефактами инструмента HiCCUPS, что можно увидеть на Рис. 21 А. Примерно половина петель, аннотированных в экспериментальных данных, также аннотированы на предсказаниях 3DPredictor (Рис. 21 Б). Кроме того, количественный анализ петлевых взаимодействий показал, что предсказанные частоты контактов между основаниями петель значительно превышают среднюю частоту контактов, наблюдаемую на соответствующем геномном расстоянии (OoE>1) (Рис. 21 В).

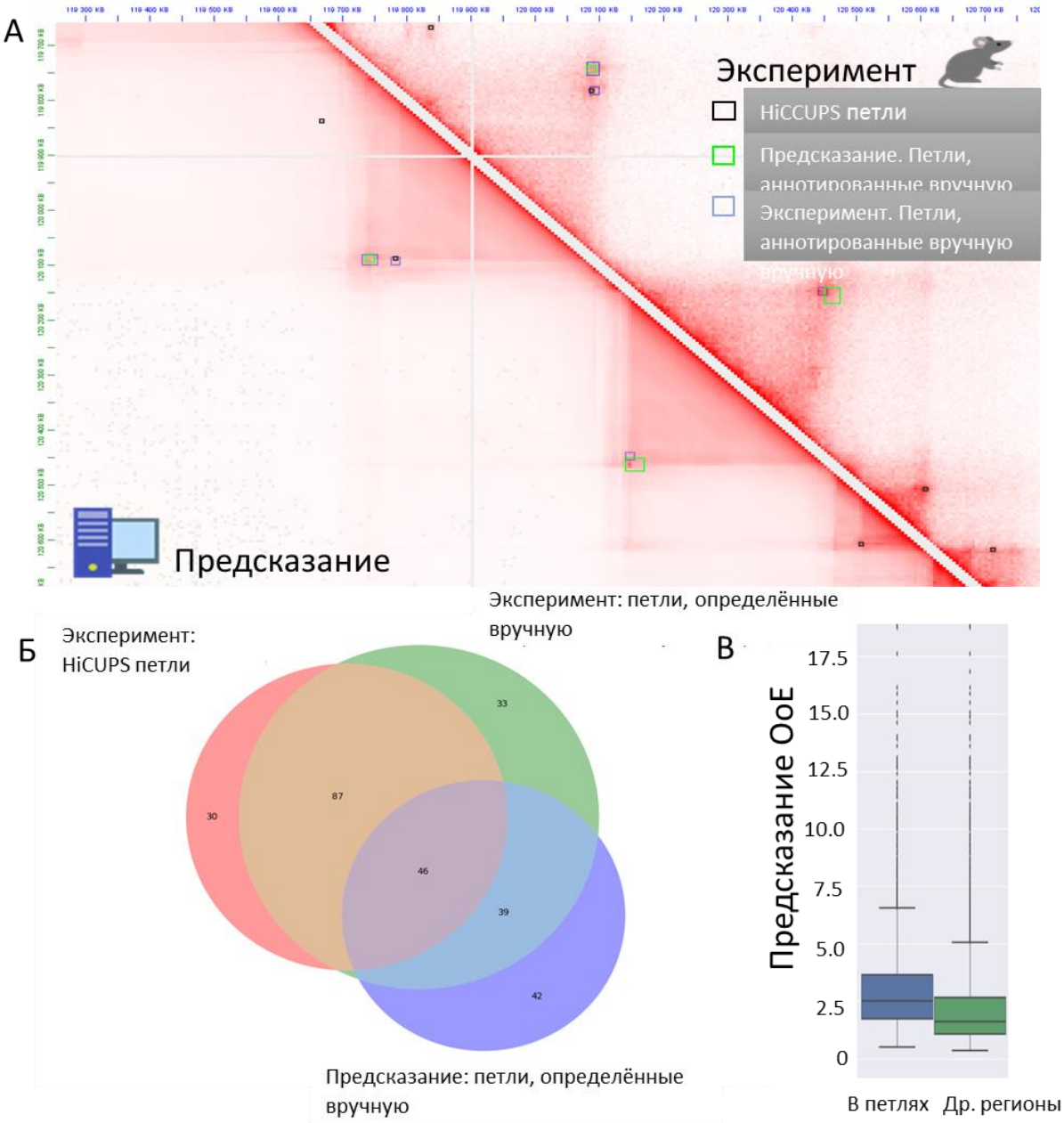


Рис. 21. 3DPredictor предсказывает примерно половину петель, детектированных в экспериментальных данных. (А) Снизу предсказанная Hi-C карта, сверху экспериментальная; петли аннотированы либо вручную, либо инструментом HiCCPUS. (Б) Пересечение петель, аннотированных разными способами в экспериментальных и предсказанных данных. Цифрой указано количество аннотированных петель в каждом из секторов диаграммы. (В) Предсказанные значения контактов (наблюдаемое/ожидаемое) в петлях (синий) и в других регионах (зелёный).

3.2.5 Предсказание инструмента 3DPredictor является клеточно-специфичным.

Основные структуры трёхмерной организации генома достаточно консервативны между разными типами клеток, но есть небольшое количество локусов, являющихся клеточно-специфичными. В этих локусах границы топологических доменов значительно отличаются, как, например, на хромосоме 3 34000-36000 Кб для мышиных гепатоцитов и клеток предшественников нейронов мыши (Рис. 22 А). Такие различия в трёхмерной организации генома связаны также с разной экспрессией генов, поэтому для предсказательного алгоритма важно улавливать эту разницу. Предсказания 3DPredictor действительно различаются для разных типов клеток. На Рис. 22 Б,В видно, что предсказанные границы доменов с хорошей точностью совпадают с экспериментальными данными. Полногеномный анализ точности предсказания клеточно-специфичных границ ТАДов был проведен Можейко Евгением и представлен в статье [117]. Результаты этого анализа подтверждают приведенный выше вывод о том, что 3DPredictor способен предсказывать клеточно-специфичные изменения границ ТАДов.

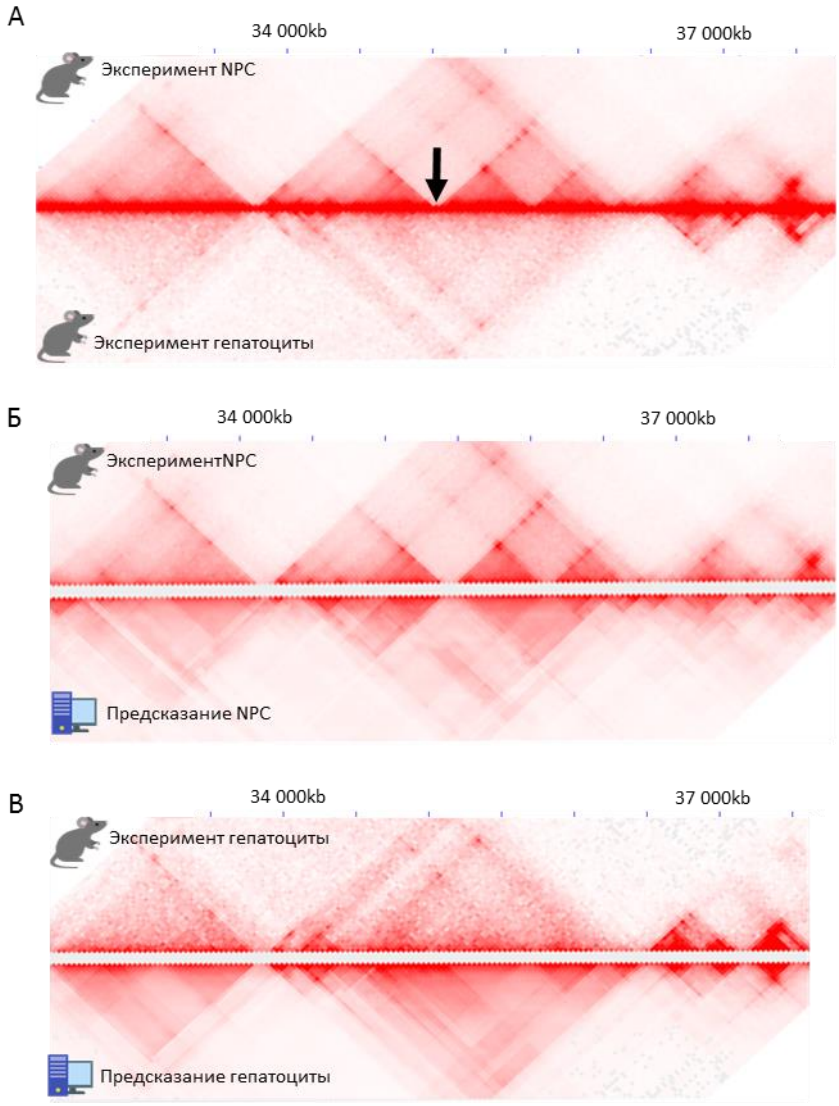


Рис. 22. Предсказание инструмента 3DPredictor является клеточно-специфичным. (А) Сверху экспериментальная Hi-C карта для клеток предшественников нейронов мыши, снизу экспериментальная Hi-C карта для гепатоцитов для локуса chr3:33000000-37500000 на разрешении 5 Кб. (Б) Сверху эксперимент, снизу предсказание для нейральных предшественников мыши. (В) Сверху эксперимент, снизу предсказание для гепатоцитов мыши.

3.2.6 Предсказание функциональных последствий хромосомных перестроек при помощи инструмента 3DPredictor.

Одной из областей применения инструмента 3DPredictor является предсказание последствий хромосомных перестроек. Это позволяет понять,

как меняются пространственные взаимодействия промотор-энхансерных пар после хромосомной перестройки.

Мы использовали данные cHi-C [58], описывающие хромосомные перестройки в локусе *Epha4* мыши, чтобы выяснить, сможет ли 3DPredictor предсказать эктопические взаимодействия в мутированном 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 взаимодействий 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. 23 А). Мы использовали метод, описанный в [55], 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. 23 А, Б). Реальные и предсказанные эктопические взаимодействиями достаточно хорошо перекрываются и их пересечение значительно отличается от случайного (P-значение $< 5 \times 10\text{-}6$) (Рис. 23 В). Это доказывает, что наша модель успешно предсказывает эктопические взаимодействия в перестроенном геноме.

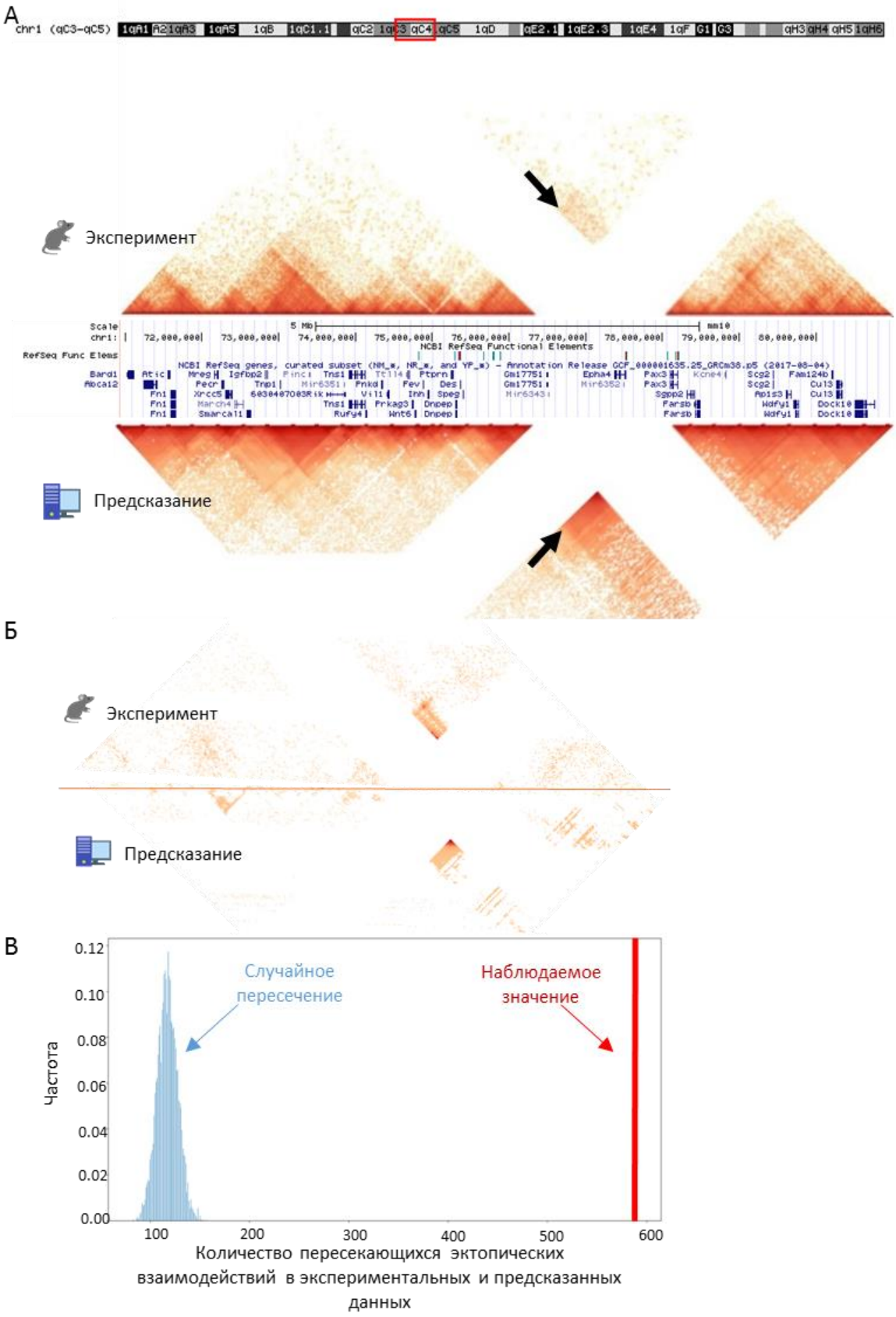


Рис. 23. Алгоритм 3DPredictor предсказывает эктопические взаимодействия в перестроенном геноме. (А) Сверху экспериментальная Hi-C

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 «Материалы и методы» 2.8 и [55] (В) Предсказанные эктопические взаимодействия были пересечены с эктопическими контактами, полученными в эксперименте. Количество пересекающихся эктопических взаимодействий представлено красной прямой. Также предсказанные эктопические взаимодействия были пересечены со случайными взаимодействиями из экспериментальных данных (количество случайно выбранных взаимодействий соответствует количеству экспериментальных эктопических взаимодействий) несколько раз. Распределение пересекающихся взаимодействий представлено синим цветом.

Ещё один случай хромосомной перестройки, меняющей пространственную организацию хроматина – это делеция фрагмента хромосомы 5 в тучных клетках мыши. Делеция затрагивает область между двумя ТАДами, включающую четыре сайта связывания белка CTCF, детектированных в эксперименте ChIP-seq, проведенном на тучных клетках (Рис. 24 А) (координаты делеции chr5:75852814-75881252). В одном из доменов находится ген *Kit*, в другом *Kdr*. Наши коллеги провели серию ChIP-seq и cHi-C-экспериментов, которые позволили охарактеризовать эпигенетический ландшафт и профиль контактов хроматина в тучных клетках мыши дикого типа и у мышей с вышеописанной делецией [118]. Мы сравнили полученные коллегами экспериментальные данные и результаты предсказания, полученные при помощи 3DPredictor. Инструмент 3DPredictor предсказывает явно видимую границу между двумя доменами в клетках дикого типа (Рис. 24 Б). Предсказание, сделанное для генома клеток, имеющих

делецию, указывает на то, что ТАДы сливаются и граница исчезает (Рис. 24 В), что подтверждается экспериментальными данными.

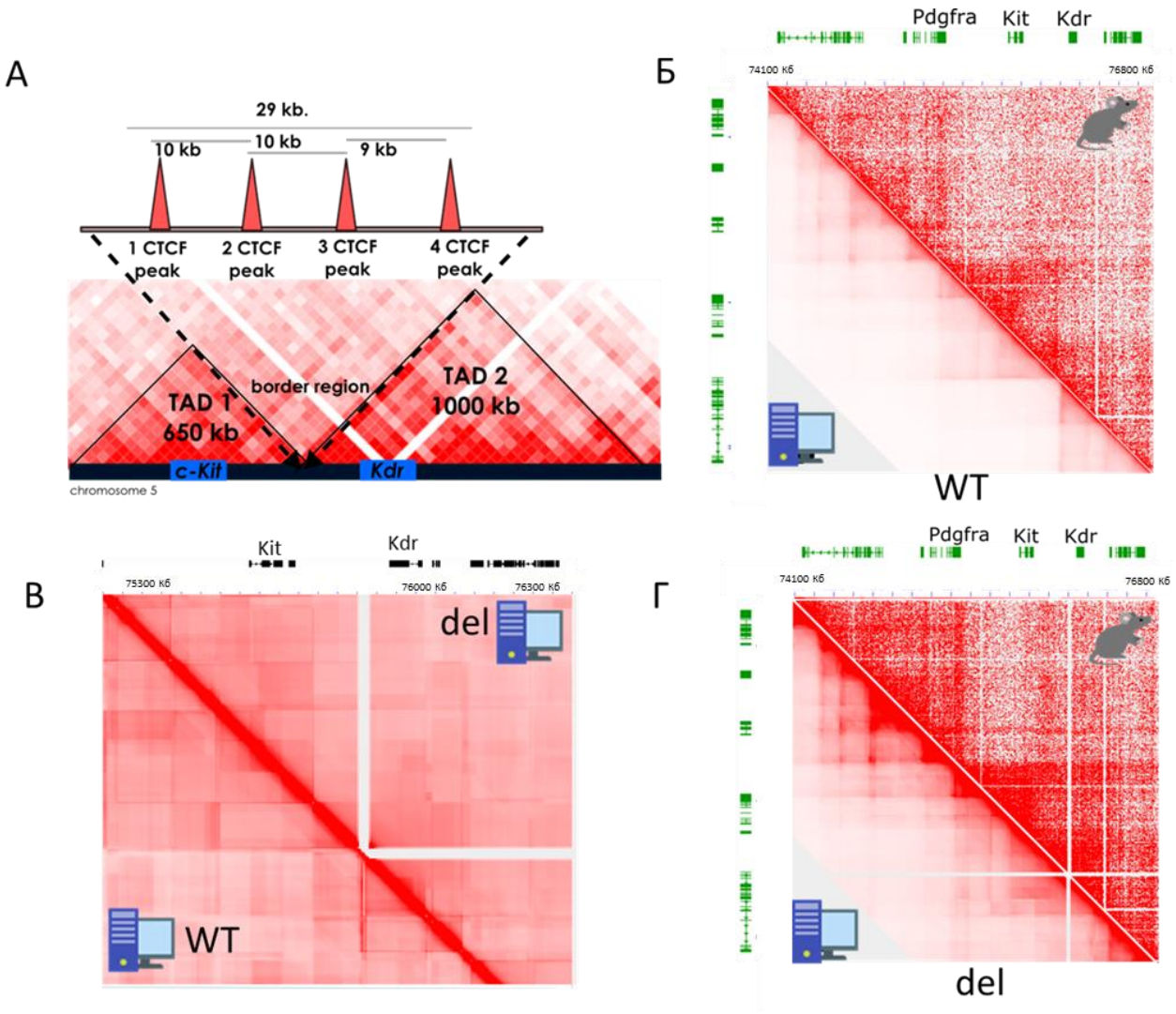


Рис. 24. 3DPredictor предсказывает слияние ТАДов при хромосомной перестройке. (А) Схематичное представление исследуемого локуса chr5:74100000-76800000 в тучных клетках мыши. (Б) Экспериментальная Hi-C карта для локуса chr5:74100000-76800000 на разрешении 5 Кб для тучных клеток мыши дикого типа (сверху) и предсказанная Hi-C карта для того же локуса (снизу). (В) Предсказанная Hi-C карта для локуса chr5:75200000-76400000 на разрешение 5 Кб для тучных клеток мыши с делецией chr5:75852814-75881252 (сверху) и предсказанная Hi-C карта для того же

локуса дикого типа (снизу). (Г) Экспериментальная Hi-C карта для локуса chr5:74100000-76800000 на разрешении 5 Кб для тучных клеток мыши с делецией chr5:75852814-75881252 (сверху) и предсказанная Hi-C карта для того же локуса (снизу).

3.2.7 Сравнение алгоритма 3DPredictor с другими моделями

На момент публикации алгоритма 3DPredictor существовали разные методы предсказания различных структур пространственной архитектуры хроматина. Так, например, были алгоритмы, которые предсказывают хроматиновые петли, границы ТАДов и компартменты [20,84–86,119], а также промотор-энхансерные взаимодействия [111,120]. Однако, в отличие от 3DPredictor, эти подходы дают качественные, а не количественные предсказания, и для большинства из них требуется значительно больше входной информации, чем для алгоритма 3DPredictor. Таким образом, в контексте сравнения 3DPredictor с другими методами, нас больше интересуют модели, которые также предсказывают пространственные взаимодействия хроматина количественно. На момент публикации алгоритма 3DPredictor существовало несколько таких инструментов (Прил. 3).

Например, модель MEGABASE+MiChroM [121] предсказывает пространственные взаимодействия хроматина на разрешении 50 Кб, используя информацию об эпигенетических модификациях и петлях, опосредованных белком CTCF. Петли, опосредованные CTCF, модель извлекает из Hi-C данных, таким образом, нельзя использовать только эпигенетические характеристики для предсказания пространственных взаимодействий хроматина, всегда нужны экспериментальные данные Hi-C. Мы сравнили 3DPredictor с этой моделью и обнаружили, что 3DPredictor значительно превосходит ее, демонстрируя гораздо более высокий уровень SCC (0.6 для 3DPredictor, 0.27 для MEGABASE+MiChroM) (Табл. 4; Рис. 25).

Табл. 4. Сравнение инструмента 3DPredictor с другими моделями на основе значений разных метрик. В сравнении участвуют алгоритмы 3DPredictor, модель MEGABASE+MichroM [121], модель Qi с соавторами. [122] и модель Rowley с соавторами. [77].

Модель	Регион для предсказания	Корреляция Пирсона	SCC	MSE	MAE	MRE
3DPredictor	chr4:53Мб-58Мб	0.95	0.60	39904.52	35600.39	0.97
MEGABASE+MichroM	chr4:53Мб-58Мб	0.95	0.27	49782.26	37368.94	1.41
3DPredictor	chr1:22Мб-23Мб	0.97	0.70	0.00061	0.00033	0.99
Модель Qi и соавт.	chr1:22Мб-23Мб	0.96	0.61	0.04024	0.01870	23.83
3DPredictor	chr4:71Мб-76Мб	0.91	0.54	0.0069	0.00046	1.92
Модель Rowley и соавт.	chr4:71Мб-76Мб	0.93	0.75	0.00053	0.00024	0.57

Тем не менее, стоит отметить, что модель MEGABASE+MiChroM изначально была разработана для предсказания дальних взаимодействий на уровне компартментов хроматина, и отсутствие информации о петлях, опосредованных CTCF, может объяснить, по крайней мере частично, плохую точность предсказания ближних взаимодействий.

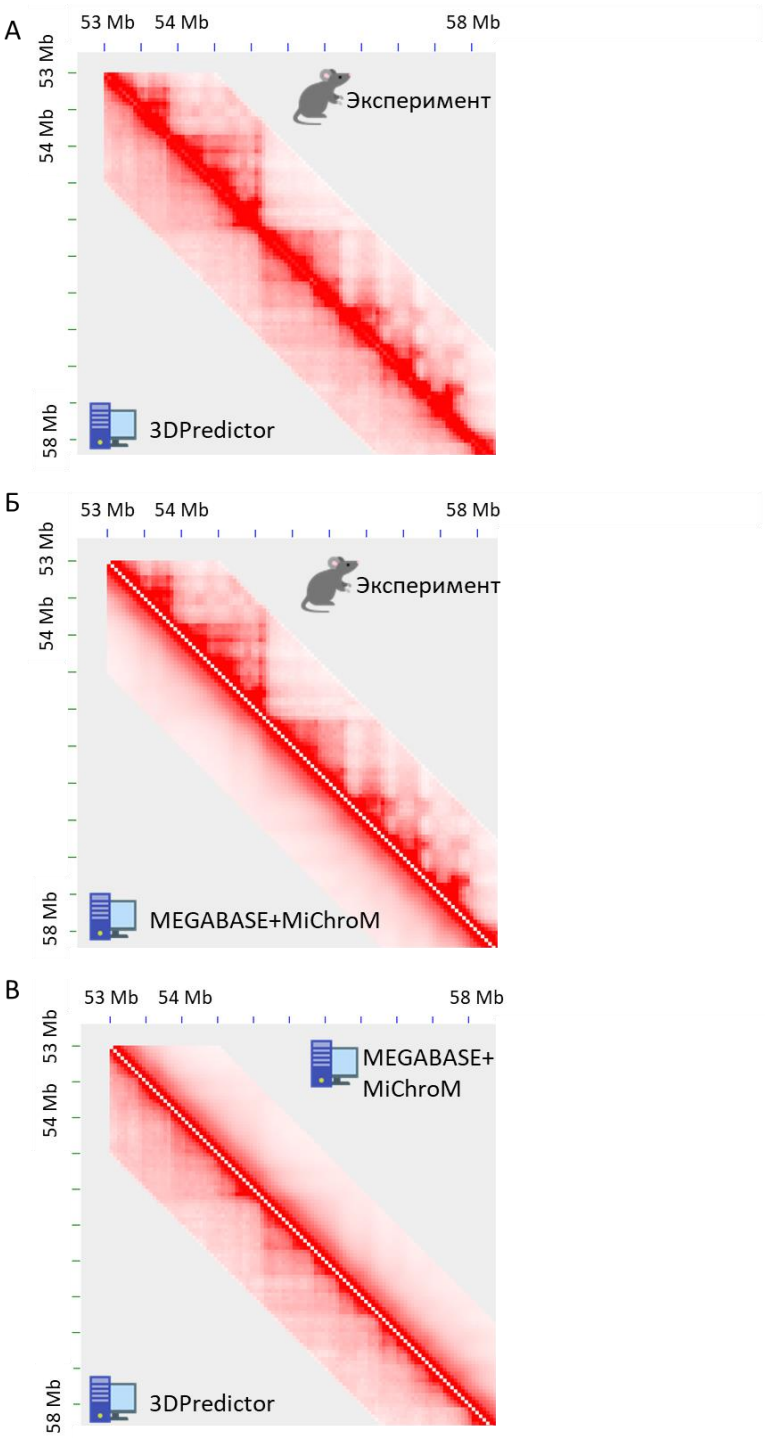


Рис. 25. Сравнение алгоритма 3DPredictor и модели MEGABASE+MiChroM. Клетки GM12878, Hi-C карты представлены для локуса chr5:53-58 Мб на разрешении 50 Кб.

Ещё один алгоритм, основанный на физическом моделировании полимеров, был предложен Qi с соавторами [122]. Эта модель использует

эпигенетические данные и геномную последовательность, доступные для сотен типов клеток, и позволяет предсказывать структуры хроматина *de novo* на разрешении до 5 Кб. При использовании этого подхода точность предсказания была выше, чем для модели MEGABASE+MiChroM (Табл. 4). Однако, для того же набора данных предсказания инструмента 3DPredictor были точнее, если судить по метрикам SCC, MSE и корреляции Пирсона (Рис. 26, Табл. 4).

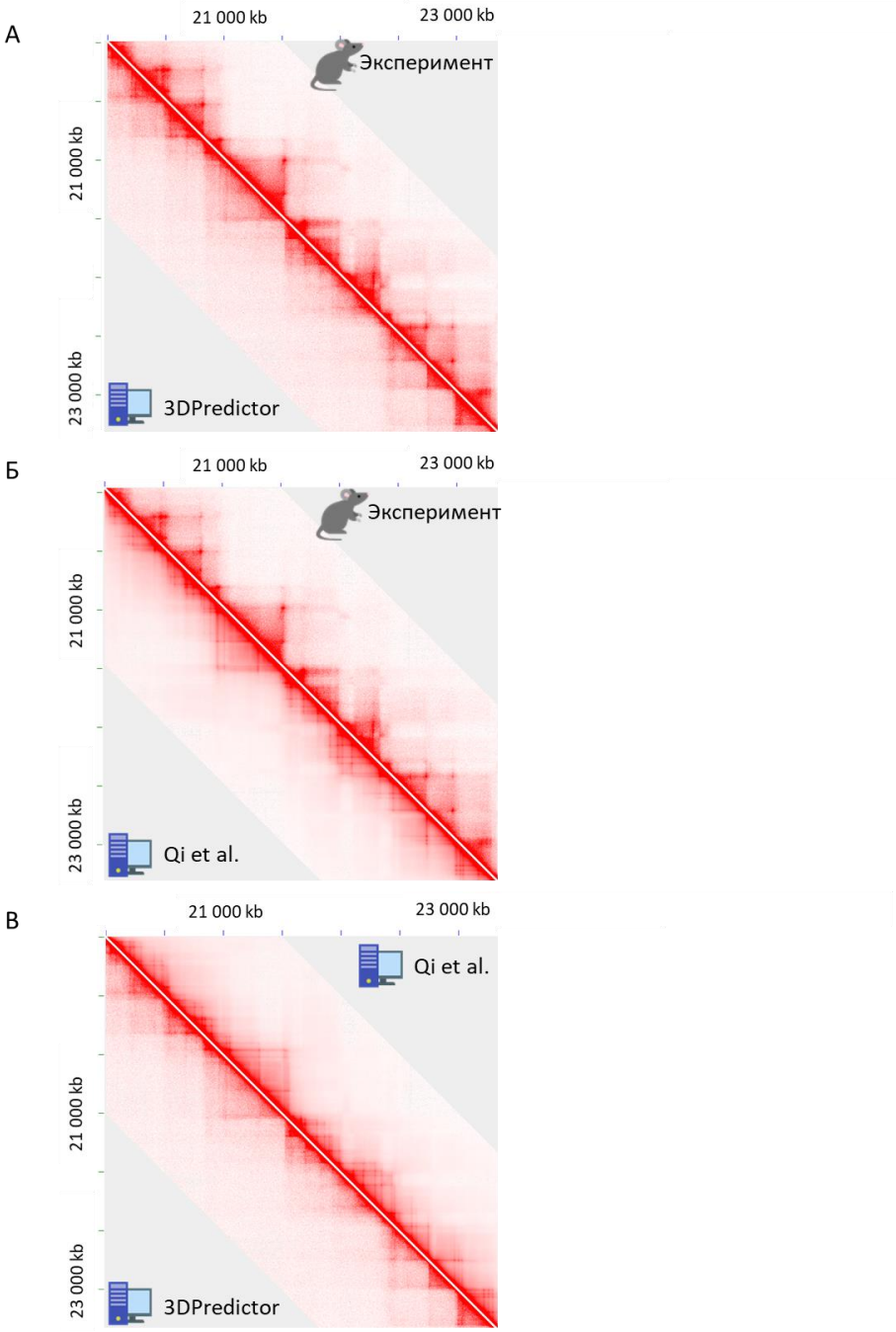


Рис. 26. Сравнение алгоритма 3DPredictor и модели Qi с соавторами [122]. Клетки GM12878, Hi-C карты представлены для локуса chr1:20-23 Мб на разрешении 5 Кб.

Ещё один подход для предсказания карты пространственных контактов хроматина основан на статистических методах и требует очень ограниченного количества информации в качестве входных данных, используя информацию о петлях, опосредованных CTCF [77]. Однако, подобно модели MEGABASE + MiChroM, этот алгоритм нельзя использовать для предсказания неизвестных взаимодействий хроматина, поскольку информация о профиле связывания белка CTCF, участвующего в образовании петель, извлекается из экспериментальных данных Hi-C. Например, на хромосоме 4 в клетках GM12878 в модели Rowley с соавторами используются только 63 выбранных вручную сайта CTCF, которые составляют примерно 35,4% всех сайтов CTCF в этом локусе. Для работы этой модели нужно получить информацию о тех сайтах CTCF, которые взаимодействуют друг с другом. Эту информацию нельзя легко получить из данных ChIP-seq, потому что в некоторых случаях петли образуются не между каждыми сайтами посадки CTCF в конвергентной ориентации [85], и для этого авторами используются экспериментальные Hi-C данные. Тем не менее, мы сравнили 3DPredictor с моделью Rowley с соавторами и обнаружили, что последняя дает значительно лучшие результаты (Табл. 4; Рис. 27).

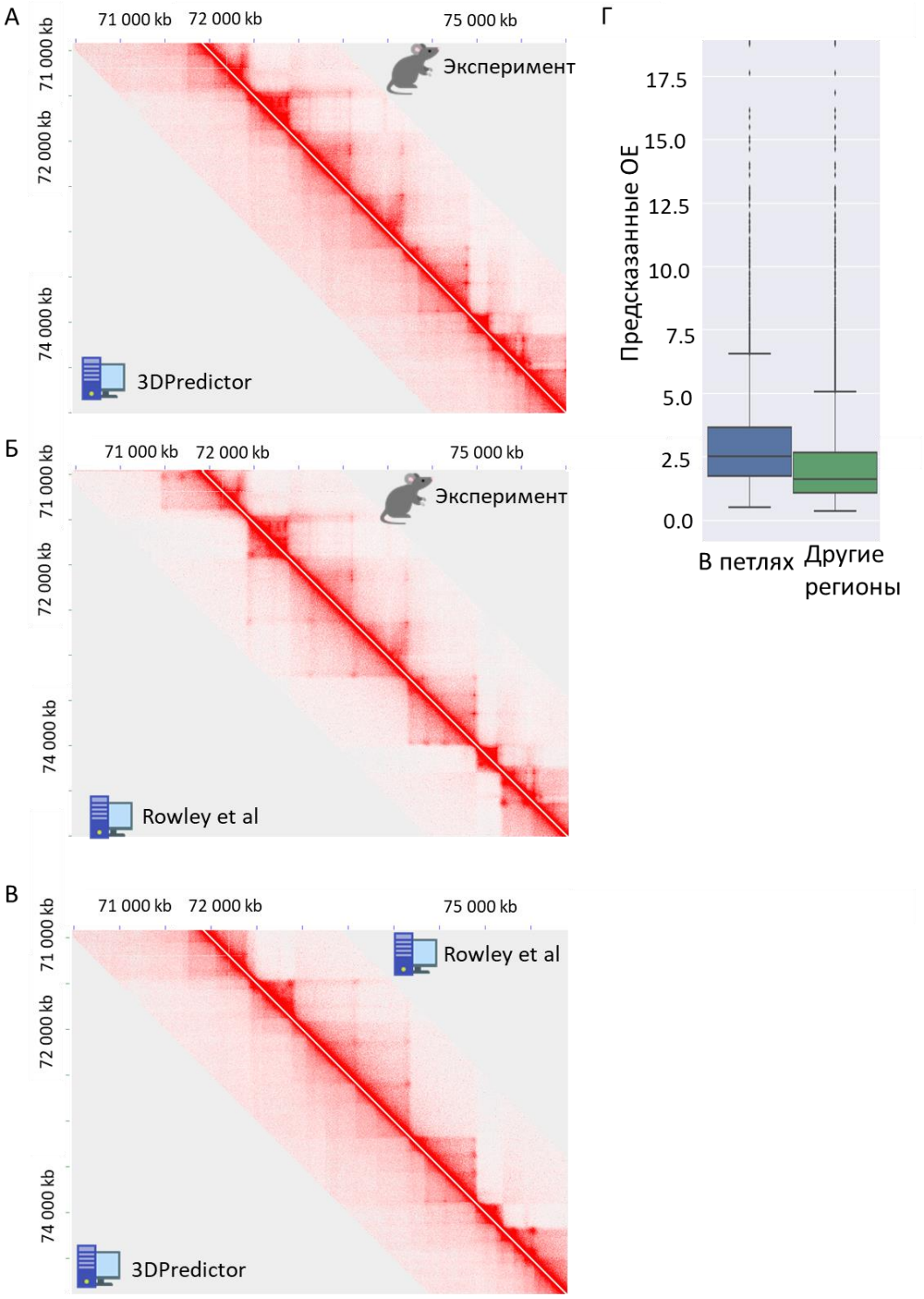


Рис. 27. Сравнение алгоритма 3DPredictor и модели Rowley с соавторами (А-В) Клетки GM12878. Hi-C карты представлена для локуса chr4:70-76 Мб на разрешении 5 Кб. (Г) Предсказание OoE значений контактов алгоритмом 3DPredictor в петлях и других локусах генома.

Поскольку Rowley с соавторами получали информацию о CTCF, участвующих в образовании петель, из экспериментальных данных, их модель особенно хорошо предсказывает петлевые взаимодействия (Рис. 27). И хотя 3DPredictor не требует какой-либо экспериментальной Hi-C информации, он также предсказывает хроматиновые петли, улавливая примерно половину петлевых взаимодействий (Рис. 21). Кроме того, мы проверили OoE частоту предсказанных контактов в основаниях петель, и оказалось, что она выше в петлях, чем между другими областями генома (Рис. 27 Г).

Подводя итог этой части работы, мы разработали инструмент 3DPredictor, основанный на машинном обучении, способный количественно предсказывать пространственные взаимодействия локусов генома, включая взаимодействия промоторов и энхансеров. Кроме того, инструмент 3DPredictor может предсказывать эктопические взаимодействия, возникающие в результате хромосомной перестройки, что является важным шагом на пути к предсказанию изменений экспрессии генов и фенотипических проявлений, связанных с этими изменениями.

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

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 [55,81,87,88,123]. 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.

3.3.1 Создание набора данных для платформы 3DGenBench.

Для того чтобы создать платформу с вышеописанными характеристиками, 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 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 [2,4,55–57,60,124] 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 [105] и с использованием нормализации C-TALE [125] для cHi-C данных.

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 определили как «вертикальный» (Рис. 28).

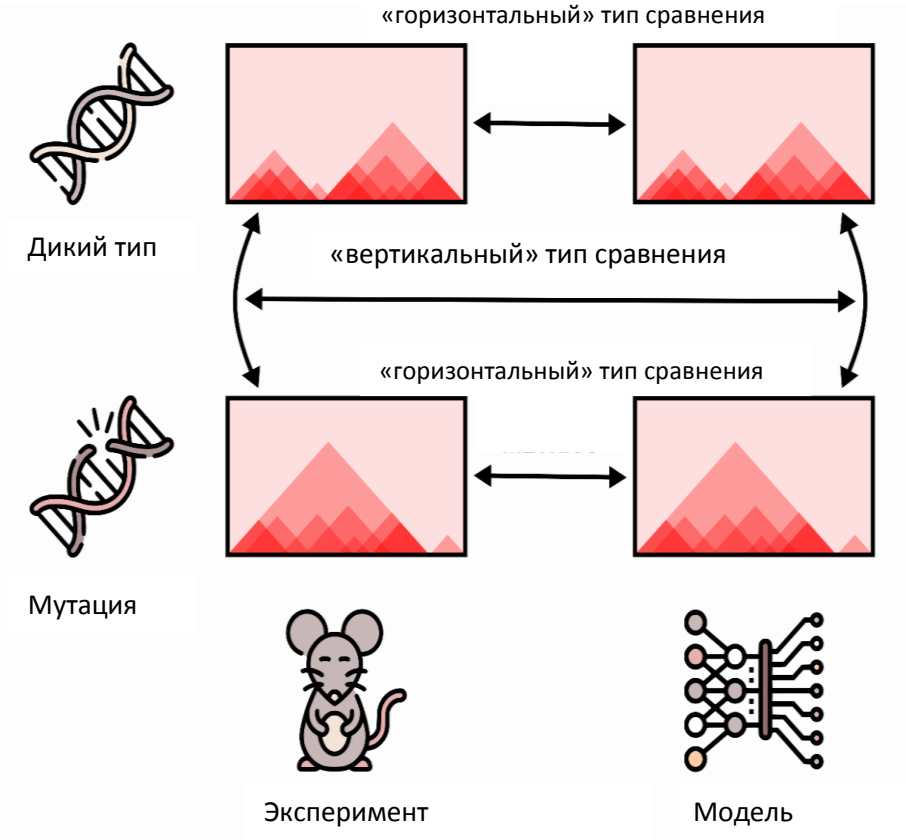


Рис. 28. Разные типы сравнения предсказанных и экспериментальных Hi-C карт.

3.3.2 Разработка метрик для оценки точности алгоритмов, предсказывающих пространственную архитектуру хроматина

Следующий этап работы включал в себя разработку метрик для оценки точности предсказания для двух типов сравнения.

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], которая более подробно была описана выше. Напомним, что при использовании этой метрики на первом этапе сглаживают матрицу контактов, для того чтобы уменьшить влияние шума на корреляцию, а затем считается обычная корреляция Пирсона по стратам для контактов, находящихся на одинаковом геномном расстоянии друг от друга.

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 (раздел «Материалы и методы» 2.7).

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 [110] (раздел «Материалы и методы» 2.7).

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.

Эти метрики мы использовали как для Hi-C, так и для cHi-C данных дикого типа и мутации соответственно. С тем лишь ограничением, что метрика силы компартментализации и зависимости частот контактов от геномного расстояния в наборе cHi-C данных не использовались, так как локусы для предсказаний являются слишком маленькими для подсчёта этих метрик.

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. Для этого мы делили значения профиля инсуляции, посчитанного по Hi-C карте с мутацией, на значения инсуляторного профиля для Hi-C карты дикого типа. Затем считалась корреляция Спирмана изменений профиля инсуляции для экспериментальных и предсказанных данных (раздел «Материалы и методы» 2.7).

Во-вторых, мы оценивали, насколько точно были предсказаны те частоты контактов, которые изменились больше всего вследствие мутации. Эктопические взаимодействия определялись как в [55], более подробное описание подсчёта эктопических взаимодействий описано в разделе «Материалы и методы» 2.7.

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 (раздел «Материалы и методы» 2.6.1). Мы выбрали несколько образцов из

подготовленного набора данных и протестировали, как меняются значения метрик в зависимости от количества шума в данных (Рис. 29 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.

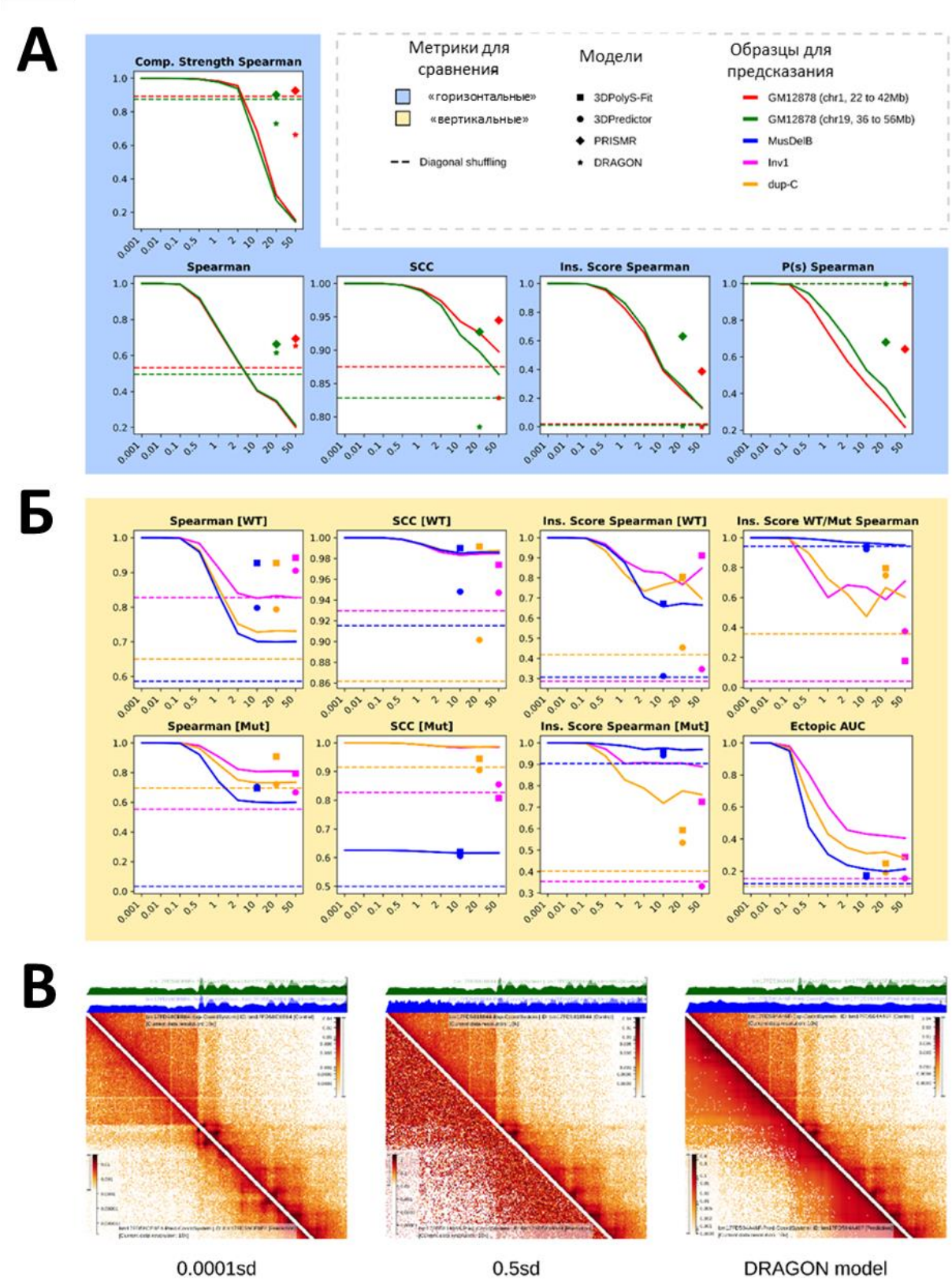


Рис. 29. Разработанные метрики отражают различия данных, предсказанных разными алгоритмами. (А) Все метрики «горизонтального» типа сравнения,

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. (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.

На рисунке 29 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 [55], DRAGON [122], 3DPolyS-Fit [99] и разработанного нами инструмента 3DPredictor. Для «горизонтального» типа сравнения, наши колабораторы предоставили предсказанные Hi-C карты контактов хроматина для локусов 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. 29 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. Однако опираясь на результаты другой метрики стоит отметить, что 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.

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 [99]. Мы сгенерировали предсказания архитектуры хроматина для этих же локусов, используя разработанный нами алгоритм 3DPredictor, и сравнили полученные модели. В первую очередь видно, что оба алгоритма с разной точностью предсказывают контакты для разных типов перестроек (Рис. 29 Б). Изменения инсуляторного профиля, вызванные хромосомной перестройкой, оба алгоритма предсказывают примерно на одинаковом уровне, но модель 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.

3.3.3 Разработка web-платформы для оценки точности работы алгоритмов, предсказывающих 3D организацию генома

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. 30 А, Б).

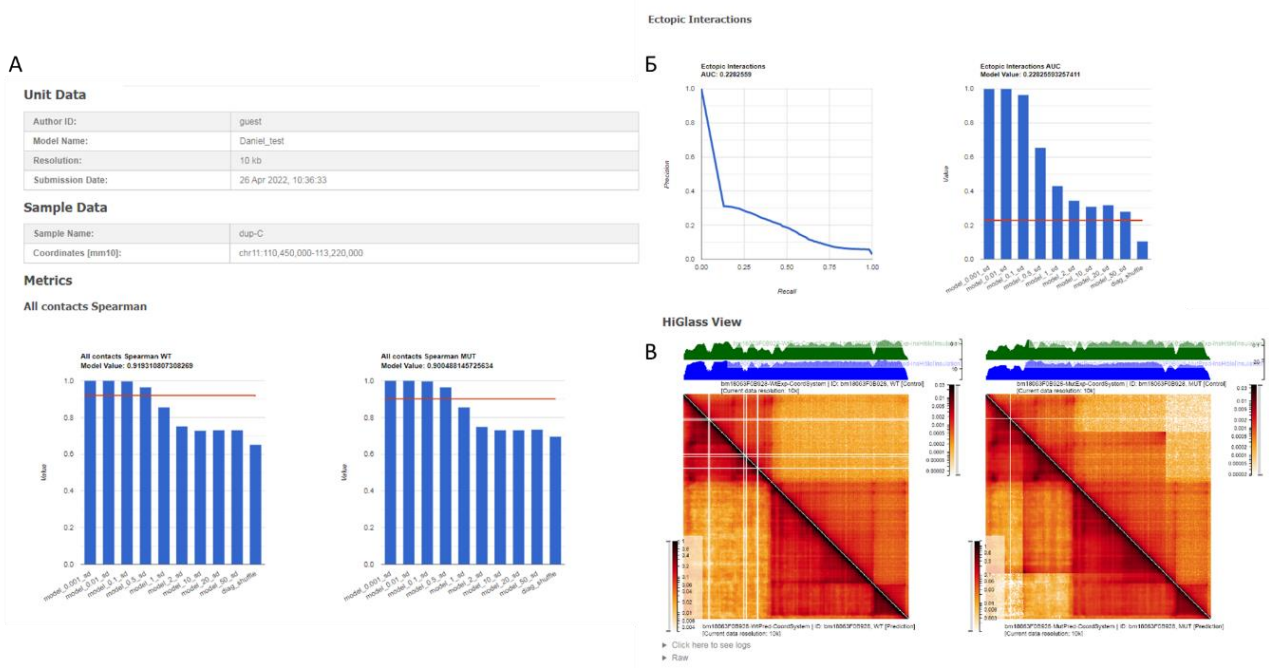


Рис. 30. Фрагменты визуализации метрик с сайта 3DGenBench. (А) Описание выбранного локуса и значение корреляции Спирмана по сравнению с базисом (данные с разным уровнем шума). (Б) Пример визуализации метрики, отражающей точность предсказанных эктопических взаимодействий. (В) Пример визуализации предсказанных и экспериментальных данных для мутантного и дикого типа на сайте.

Кроме того, пользователи имеют возможность использовать разработанную вычислительную 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 (5 Кб, 105 Кб, 205 Кб, 255 Кб, 505 Кб), которые легко можно скачать. Обработанные ChIP-seq данные о распределении белка CTCF на ДНК и информацию об ориентации сайтов связывания CTCF в bed формате также можно скачать с сайта.

Часто бывает полезно визуально оценить предсказанные карты контактов, для того чтобы удостовериться, что сгенерированные данные и метрики сходятся с нашими представлениями о «хороших» и «плохих» предсказаниях. Такие представления предсказанных Hi-C карт показаны на Рис. 29, 30 В.

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.

ГЛАВА 4. ОБСУЖДЕНИЕ

4.1 Проблема создания несвязанных выборок для обучения и валидации моделей машинного обучения

Известно, что при тренировке алгоритмов машинного обучения, нужно использовать как минимум две непересекающиеся выборки, одну для обучения и другую для валидации. Это важно, поскольку такой метод позволяет удостовериться, что алгоритм действительно основан на обобщающих закономерностях между входными и выходными данными, а не просто «выучил» обучающую выборку. Важно проверять, что с увеличением точности алгоритма на данных для обучения увеличивается и точность работы алгоритма на валидационных данных. Считается, что в тот момент, когда метрика точности для валидационных данных начинает падать, а та же метрика для тренировочных данных продолжает расти, следует остановить обучение алгоритма, поскольку дальше начинается «подгонка» параметров алгоритма под тренировочные данные. Те же самые принципы должны применяться и при работе с геномными данными. Однако понятие пресекающихся выборок для геномных данных является более сложным.

Часто алгоритмы машинного обучения в геномике используются для предсказания каких-либо свойств двух локусов (взаимодействие промоторов и энахансеров, пространственное взаимодействие двух локусов, предсказание Hi-C петель и т.д.) [79,81,84,86]. Для таких предсказаний в качестве признаков используются доступные биологические данные, например, ChIP-seq данные, характеризующие область генома между ними или в некой окрестности интересуемых локусов. Это является биологически обоснованным, поскольку ясно, что эпигенетическое окружение влияет на взаимодействие локусов генома между собой, и нужно рассматривать эпигенетические характеристики не точечно, а в некоторой окрестности. Таким образом, при составлении непересекающихся выборок для обучения и валидации важно отследить,

чтобы наборы признаков двух пар геномных локусов, используемых для обучения и валидации, не пересекались между собой. Например, можно использовать данные, полученные с разных хромосом. Выше в разделе «Результаты» 3.1 мы более подробно разбирали этот вопрос на примере выбора промотор-энхансерных пар для алгоритма TargetFinder [79]. Однако это не единственный пример такого варианта набора тренировочных и валидационных данных с косвенно пересекающимися признаками [113,126]. И здесь стоит отметить, что следует очень внимательно подходить к оценке точности работы алгоритмов, учитывая специфику геномных данных.

4.2 Ограничения алгоритма 3DPredictor

Алгоритм 3DPredictor работает на основе градиентного бустинга и использует эпигенетические характеристики для предсказания пространственной карты контактов. Мы выбрали ограниченный набор эпигинетических характеристик для финальной версии алгоритма на основе вклада каждого признака в предсказание. В финальной версии алгоритма среди таких признаков оказались ChIP-seq данные белка CTCF, ориентация сайтов посадки белка CTCF, RNA-seq данные и геномное расстояние между локусами. Большой вклад этих признаков в предсказание также обоснован биологически. Так, например, белок CTCF является основным архитектурным белком хроматина у млекопитающих. Однако исследования пространственной организации хроматина показали, что CTCF не единственный архитектурный белок, и далеко не все петли обеспечиваются им.

Было показано, что есть другой архитектурный белок, YY1, который участвует в организации промотор-энхансерных взаимодействийы [46,47]. Исследователи показали, что при дифференцировке наивных плюрипотентных клеток в нейральные предшественники белком CTCF обеспечиваются промотор-энхансерные взаимодействия для генов, работающих только в плюрипотентных клетках. Однако новые промотор-

энхансерные взаимодействия, возникающие при дифференцировке, не колокализуются с белком CTCF, а демонстрируют сильное обогащение белком YY1.

Ещё один случай, когда транскрипционный фактор участвует в организации специфических пространственных взаимодействий хроматина, связан с белком MyoD [127]. Авторы показали, что в клетках мышц этот белок активно участвует в организации хроматиновых петель между энхансерами и промоторами активных генов. При чём есть петли, опосредованные MyoD и CTCF, а есть хроматиновые петли, где в основании петель находятся только белки MyoD.

Всё это приводит к мысли, что механизмы регуляции экспрессии генов, опосредованные пространственными контактами хроматина, очень разнообразны. И механизм «протягивания петли» является не единственным механизмом, регулирующим пространственную организацию генома. Продолжающиеся исследования в области 3D геномики открывают всё больше новых фактов о регуляции работы генов, и, вероятно, многие из них ещё не известны. В этом и состоит одно из ограничений инструмента 3DPredictor. Среди признаков, на которых обучался 3DPredictor, нет ChIP-seq данных многих других архитектурных белков, таким образом, предполагается, что алгоритм машинного обучения может находить только закономерности, которые объясняют возникновение CTCF-опосредованных петель и ТАДов. Соответственно, предсказание петель, опосредованных другими белками, не должно быть успешным, для этого необходимую информацию нужно добавлять в вектор признаков при обучении алгоритма.

Наконец, следует отметить, что для ряда специализированных типов клеток и некоторых клеточных состояний укладка хроматина не связана с белком CTCF и/или транскрипцией. Среди таких клеток можно выделить зрелые гаметы млекопитающих [38,128], зрелые эритроциты позвоночных

[129], клетки в стадии митоза или мейоза [130], клетки раннего эмбриона позвоночных и насекомых [131]. Очевидно, что обученная на данных соматических клеток модель 3DPredictor будет неприменима для предсказания пространственной организации хроматина в этих клетках.

Экспериментальные методы изучения хроматина постоянно совершенствуются, на сегодняшний день появились Hi-C карты на высоком разрешении около 1 Кб [16]. Исследователи отмечают, что на таких Hi-C картах представлена более полная информация о пространственной архитектуре генома, выявляются новые структуры субдоменов и больше петель. Алгоритм 3DPredictor обучался на Hi-C картах с разрешением от 5 Кб, таким образом, для того чтобы получить более точные предсказания, необходимо переобучать алгоритм на новых Hi-C данных и совершенствовать его в соответствии с актуальными данными.

4.1 Причинно-следственная связь между пространственной организацией хроматина и экспрессией генов

Ещё один неразрешённый вопрос заключается в том, что до сих пор в научном сообществе идут дебаты о том, что является причиной, а что следствием: пространственная организация генома регулирует экспрессию генов, или хроматин уложен в пространстве ядра именно так из-за установившейся транскрипционной активности, и являются ли ТАДы некими регуляторными единицами генома. На этот вопрос нельзя ответить однозначно. Так, например, во многих работах показали, что разрушения границ ТАДов приводит к изменению пространственной организации хроматина и часто к изменению взаимодействий между энхансером и промотором ([3,58,132] раздел «Обзор литературы» 1.3), что отражается на экспрессии генов. Однако, стоит отметить, что эти результаты были получены для конкретных локусов генома, и эти выводы нельзя обобщать на весь геном. Кроме того, в ряде работ ([3,60] «Обзор литературы» 1.3) было показано, что

удаление сайтов посадки CTCF на границе доменов не всегда приводит к слиянию ТАДов. А если даже ТАДы все таки сливаются в результате делеции сайтов посадки CTCF на границе, не всегда это отражается на экспрессии генов [60]. Однако, в той же работе исследователи показали, что внесение нового CTCF сайта способствует возникновению новой границы между ТАДами, что приводит к изменению пространственных контактов между промотором и энхансером, и, как следствие, к снижению экспрессии регулируемого гена [60].

Неопределенность следственных связей между разными эпигенетическими характеристиками клеток и транскрипцией приводит к сложностям в моделировании хромосомных перестроек. В нашей работе при создании моделей хромосомных перестроек мы исходили из того, что удаление, удвоение или инверсия участков генома прямо отражается на соответствующих эпигенетических метках. Например, часть эпигенетических меток удаляется, удваивается или меняет своё геномное расположение строго в соответствии со структурой переместившихся локусов генома. Однако такой «наивный» подход к моделированию может не отражать всю сложность взаимосвязей между регуляторными системами, в которых изменения геномных разметок, прямо связанные с изменением в копийности и расположении локусов, могут за счет сложного регуляторного каскада вторично влиять на эти же разметки. Так, изменения в трехмерной организации хроматина, смоделированные нами, могут влиять на уровень транскрипции, который, в свою очередь скажется на прогстранственной организации хроматина – такой вторичный эффект не может быть учтен в нашей модели.

Таким образом, важность 3D архитектуры хроматина для регуляции экспрессии генов зависит от локуса и контекста.

4.2 Моделирование в 3D-геномике

Стоит отметить, что с момента публикации алгоритма 3DPredictor появились новые модели, основанные на методах глубокого обучения [82,87,88], которые способны предсказывать пространственную архитектуру хроматина и изменения в ней, вызванные хромосомными перестройками на высоком уровне. Кроме того, растёт количество экспериментальных данных, и появляются новые методы и гипотезы, что также увеличивает точность новых моделей как физических, так и основанных на статистических методах.

Улучшение точности предсказательных моделей позволяет применять их не только для предсказания последствий крупных хромосомных перестроек, но также и для предсказания изменений, вызванных однонуклеотидными заменами. Другой областью применения моделей является предсказание неизвестных экспериментальных данных. Так, недавно алгоритм Akita [88] был применён для предсказания пространственной организации хроматина неандертальцев. Это стало возможно, поскольку на вход алгоритму требуется только нуклеотидная последовательность. Эта работа позволила обнаружить те локусы генома, в которых пространственная организация хроматина изменилась больше всего в течение эволюции. Таким образом, моделирование в 3D геномике является многосторонним инструментом, применимым в том числе и для изучения эволюционных изменений.

4.3 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 после хромосомных перестроек, 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. Разработанная нами платформа 3DGenBench и единообразный набор данных для предсказания оказываются действительно 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 [133–135].

Сейчас моделирование переходит к более комплексному предсказанию свойств. Например, исследователи предсказывают не только места посадки транскрипционных факторов в разных типах клеток, но и регуляторные взаимодействия, возникающие между ними и промоторами соответствующих генов [136]. Также есть подходы, в которых пытаются извлечь закономерности, характеризующие разные типы клеток, что позволяет предсказывать эпигенетические характеристики в разных клеточных типах, не обучаясь на данных для конкретного типа клеток [137]. Использование и совершенствование идей, заложенных в первых моделях, позволяет создавать более сложные и комплексные алгоритмы. Таким образом, алгоритм 3DPredictor занял своё место в наборе существующих моделей для предсказания пространственной архитектуры генома. Платформа 3DGenBench может расширяться в соответствии с развитием новых моделей и методов, что позволит ысравнивать 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 литературы

1. Yang T. et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient // Genome Res. 2017. Vol. 27, № 11. P. 1939–1949.
2. Rodríguez-Carballo E. et al. The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes // Genes Dev. Cold Spring Harbor Laboratory Press, 2017. Vol. 31, № 22. P. 2264–2281.
3. Franke M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications // Nature. Nature Publishing Group, 2016. Vol. 538, № 7624. P. 265–269.
4. Hanssen L.L.P. et al. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo // Nat. Cell Biol. 2017. Vol. 19, № 8. P. 952–961.
5. Kragesteen B.K. et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis // Nat. Genet. Nature Publishing Group, 2018. Vol. 50, № 10. P. 1463–1473.
6. Kraft K. et al. Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations // Nature Cell Biology. Nature Publishing Group, 2019. Vol. 21, № 3. P. 305–310.
7. Barutcu A.R. et al. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus // Nat. Commun. 2018. Vol. 9, № 1. P. 1444.
8. Piovesan A. et al. On the length, weight and GC content of the human genome // BMC Res. Notes. 2019. Vol. 12, № 1. P. 106.
9. Edgar B.A., Kim K.J. Sizing Up the Cell // Science (80-.). 2009. Vol. 325, № 5937. P. 158–159.
10. Kempfer R., Pombo A. Methods for mapping 3D chromosome architecture // Nature Reviews Genetics. Nature Research, 2020. Vol. 21, № 4. P. 207–226.
11. Dekker J. et al. Capturing Chromosome Conformation // Science (80-.). 2002. Vol. 295, № 5558. P. 1306–1311.
12. Баттулин Н.Р. et al. 3С-Методы В Исследованиях Пространственной Организации Генома // Вавиловский Журнал Генетики И Селекции. 2012. Vol. 16, № 4/2. P. 872–878.
13. Denker A., De Laat W. The second decade of 3C technologies: Detailed

insights into nuclear organization // Genes and Development. 2016. Vol. 30, № 12. P. 1357–1382.

14. de Wit E., de Laat W. A decade of 3C technologies: insights into nuclear organization // Genes Dev. 2012. Vol. 26, № 1. P. 11–24.
15. Lieberman-Aiden E. et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome // Science (80-.). 2009. Vol. 326, № 5950. P. 289–293.
16. Krietenstein N. et al. Ultrastructural Details of Mammalian Chromosome Architecture // Mol. Cell. Cell Press, 2020. Vol. 78, № 3. P. 554-565.e7.
17. Hsieh T.H.S. et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding // Mol. Cell. Cell Press, 2020. Vol. 78, № 3. P. 539-553.e8.
18. Robinson J.T. et al. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. 2018.
19. Mifsud B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C // Nat. Genet. 2015. Vol. 47, № 6. P. 598–606.
20. Huang J. et al. Predicting chromatin organization using histone marks // Genome Biol. BioMed Central Ltd., 2015. Vol. 16, № 1. P. 162.
21. Beagrie R.A. et al. Complex multi-enhancer contacts captured by genome architecture mapping // Nature. Nature Publishing Group, 2017. Vol. 543, № 7646. P. 519–524.
22. Quinodoz S.A. et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus // Cell. Cell Press, 2018. Vol. 174, № 3. P. 744-757.e24.
23. Zheng M. et al. Multiplex chromatin interactions with single-molecule precision // Nature. Nature Publishing Group, 2019. Vol. 566, № 7745. P. 558–562.
24. Gillooly J.F., Hein A., Damiani R. Nuclear DNA content varies with cell size across human cell types // Cold Spring Harb. Perspect. Biol. Cold Spring Harbor Laboratory Press, 2015. Vol. 7, № 7. P. 1–27.
25. Münkel C., Langowski J. Chromosome structure predicted by a polymer model // Phys. Rev. E. 1998. Vol. 57, № 5. P. 5888–5896.
26. Grosberg A.Y., Nechaev S.K., Shakhnovich E.I. The role of topological constraints in the kinetics of collapse of macromolecules // J. Phys. 1988.

Vol. 49, № 12. P. 2095–2100.

27. Mirny L.A. The fractal globule as a model of chromatin architecture in the cell. 2011.

28. Sivakumar A., de las Heras J.I., Schirmer E.C. Spatial genome organization: From development to disease // Frontiers in Cell and Developmental Biology. Frontiers Media S.A., 2019. Vol. 7, № MAR. P. 18.

29. Tavares-Cadete F. et al. Multi-contact 3C reveals that the human genome during interphase is largely not entangled // Nat. Struct. Mol. Biol. 2020. Vol. 27, № 12. P. 1105–1114.

30. Cremer T. et al. Role of chromosome territories in the functional compartmentalization of the cell nucleus // Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory Press, 1993. Vol. 58. P. 777–792.

31. Gilbert N., Gilchrist S., Bickmore W.A. Chromatin organization in the mammalian nucleus // Int. Rev. Cytol. Academic Press Inc., 2004. Vol. 242. P. 283–336.

32. Kim S., Shendure J. Mechanisms of Interplay between Transcription Factors and the 3D Genome // Molecular Cell. Cell Press, 2019. Vol. 76, № 2. P. 306–319.

33. Lieberman-Aiden E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome // Science (80-.). NIH Public Access, 2009. Vol. 326, № 5950. P. 289–293.

34. Rao S.S.P. et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping // Cell. 2014. Vol. 159, № 7. P. 1665–1680.

35. Kantidze O.L., Razin S. V. Weak interactions in higher-order chromatin organization // Nucleic acids research. NLM (Medline), 2020. Vol. 48, № 9. P. 4614–4626.

36. Alberti S., Gladfelter A., Mittag T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates // Cell. 2019. Vol. 176, № 3. P. 419–434.

37. Dixon J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions // Nature. 2012. Vol. 485, № 7398. P. 376–380.

38. Battulin N. et al. Comparison of the three-dimensional organization of

sperm and fibroblast genomes using the Hi-C approach // Genome Biol. BioMed Central Ltd., 2015. Vol. 16, № 1. P. 77.

39. Fudenberg G. et al. Formation of Chromosomal Domains by Loop Extrusion // Cell Rep. Elsevier B.V., 2016. Vol. 15, № 9. P. 2038–2049.

40. Sanborn A.L. et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2015. Vol. 112, № 47. P. E6456–E6465.

41. Brini E., Simmerling C., Dill K. Protein storytelling through physics // Science (80-.). 2020. Vol. 370, № 6520.

42. Dolgin E. DNA's secret weapon against knots and tangles // Nature. 2017. Vol. 544, № 7650. P. 284–286.

43. Rao S.S.P. et al. Cohesin Loss Eliminates All Loop Domains // Cell. Cell Press, 2017. Vol. 171, № 2. P. 305-320.e24.

44. Eagen K.P., Aiden E.L., Kornberg R.D. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2017. Vol. 114, № 33. P. 8764–8769.

45. Petrovic J. et al. Oncogenic Notch Promotes Long-Range Regulatory Interactions within Hyperconnected 3D Cliques // Mol. Cell. Cell Press, 2019. Vol. 73, № 6. P. 1174-1190.e12.

46. Weintraub A.S. et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops // Cell. Cell Press, 2017. Vol. 171, № 7. P. 1573-1588.e28.

47. Beagan J.A. et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment // Genome Res. 2017. Vol. 27, № 7. P. 1139–1152.

48. Vian L. et al. The Energetics and Physiological Impact of Cohesin Extrusion // Cell. Cell Press, 2018. Vol. 173, № 5. P. 1165-1178.e20.

49. Furlong E.E.M., Levine M. Developmental enhancers and chromosome topology // Science (80-.). 2018. Vol. 361, № 6409. P. 1341–1345.

50. Lettice L.A. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly // Hum. Mol. Genet. 2003. Vol. 12, № 14. P. 1725–1735.

51. Tuan D., Kong S., Hu K. Transcription of the hypersensitive site HS2 enhancer

in erythroid cells. // Proc. Natl. Acad. Sci. 1992. Vol. 89, № 23. P. 11219–11223.

52. Palstra R.-J. et al. The β-globin nuclear compartment in development and erythroid differentiation // Nat. Genet. 2003. Vol. 35, № 2. P. 190–194.

53. Wurmser A., Basu S. Enhancer-Promoter Communication: It's Not Just About Contact // Front. Mol. Biosci. 2022. Vol. 9.

54. Bartman C.R. et al. Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping // Mol. Cell. 2016. Vol. 62, № 2. P. 237–247.

55. Bianco S. et al. Polymer physics predicts the effects of structural variants on chromatin architecture // Nat. Genet. Springer US, 2018. Vol. 50, № 5. P. 662–667.

56. Franke M. et al. Formation of new chromatin domains determines pathogenicity of genomic duplications // Nature. 2016. Vol. 538, № 7624. P. 265–269.

57. Paliou C. et al. Preformed chromatin topology assists transcriptional robustness of Shh during limb development // Proc. Natl. Acad. Sci. 2019. Vol. 116, № 25. P. 12390–12399.

58. Lupiáñez D.G. et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions // Cell. 2015. Vol. 161, № 5. P. 1012–1025.

59. Valton A.-L., Dekker J. TAD disruption as oncogenic driver Long-range gene regulation occurs within Topologically Associating.

60. Despang A. et al. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture // Nat. Genet. Nature Publishing Group, 2019. Vol. 51, № 8. P. 1263–1271.

61. de Gennes P.G., Witten T.A. Scaling Concepts in Polymer Physics // Phys. Today. Cornell University Press, 1980. Vol. 33, № 6. P. 51–54.

62. Fudenberg G., Mirny L.A. Higher-order chromatin structure: Bridging physics and biology // Current Opinion in Genetics and Development. NIH Public Access, 2012. Vol. 22, № 2. P. 115–124.

63. Grosberg A.Y., Nechaev S.K., Shakhnovich E.I. The role of topological constraints in the kinetics of collapse of macromolecules The role of topological constraints in the kinetics of collapse of macromolecules The role of topological constraints in the kinetics of collapse of macromolecules

// J. Phys. 1988. Vol. 49, № 12. P. 2095–2100.

64. Di Pierro M. et al. Transferable model for chromosome architecture. // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2016. Vol. 113, № 43. P. 12168–12173.

65. Jost D. et al. Modeling epigenome folding: Formation and dynamics of topologically associated chromatin domains // Nucleic Acids Res. Oxford University Press, 2014. Vol. 42, № 15. P. 9553–9561.

66. Ulianov S. V. et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains // Genome Res. Cold Spring Harbor Laboratory Press, 2016. Vol. 26, № 1. P. 70–84.

67. Brackey C.A., Marenduzzo D., Gilbert N. Mechanistic modeling of chromatin folding to understand function // Nat. Methods. Nature Research, 2020. Vol. 17, № 8. P. 767–775.

68. Chiariello A.M. et al. Polymer physics of chromosome large-scale 3D organisation // Sci. Rep. Nature Publishing Group, 2016. Vol. 6, № 1. P. 1–8.

69. Brackley C.A. et al. Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization // Proc. Natl. Acad. Sci. U. S. A. Proc Natl Acad Sci U S A, 2013. Vol. 110, № 38.

70. MacPherson Q., Beltran B., Spakowitz A.J. Bottom–up modeling of chromatin segregation due to epigenetic modifications // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2018. Vol. 115, № 50. P. 12739–12744.

71. Chiang M. et al. Polymer Modeling Predicts Chromosome Reorganization in Senescence // Cell Rep. Elsevier B.V., 2019. Vol. 28, № 12. P. 3212-3223.e6.

72. Ulianov S. V. et al. Nuclear lamina integrity is required for proper spatial organization of chromatin in Drosophila // Nat. Commun. Nature Publishing Group, 2019. Vol. 10, № 1. P. 1–11.

73. Buckle A. et al. Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci Molecular Cell Technology Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci // Mol. Cell. 2018. Vol. 72.

74. Strom A.R. et al. Phase separation drives heterochromatin domain formation // Nature. Nature Publishing Group, 2017. Vol. 547, № 7662. P. 241–245.

75. Rao S.S.P. et al. A 3D map of the human genome at kilobase resolution

reveals principles of chromatin looping // Cell. Cell Press, 2014. Vol. 159, № 7. P. 1665–1680.

76. Xu H. et al. Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction // Computational and Structural Biotechnology Journal. Elsevier B.V., 2020. Vol. 18. P. 558–570.

77. Rowley M.J. et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization // Mol. Cell. Cell Press, 2017. Vol. 67, № 5. P. 837-852.e7.

78. Eraslan G. et al. Deep learning: new computational modelling techniques for genomics // Nature Reviews Genetics. Nature Publishing Group, 2019. Vol. 20, № 7. P. 389–403.

79. Whalen S., Truty R.M., Pollard K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin // Nat. Genet. 2016. Vol. 48, № 5. P. 488–496.

80. Li W., Wong W.H., Jiang R. DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning // Nucleic Acids Res. 2019. Vol. 47, № 10. P. e60–e60.

81. Zhang S. et al. In silico prediction of high-resolution Hi-C interaction matrices // Nat. Commun. Nature Research, 2019. Vol. 10, № 1. P. 5449.

82. Trieu T., Martinez-Fundichely A., Khurana E. DeepMILO: A deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure // Genome Biol. BioMed Central Ltd., 2020. Vol. 21, № 1. P. 79.

83. Zhu Y. et al. Constructing 3D interaction maps from 1D epigenomes // Nat. Commun. Nature Publishing Group, 2016. Vol. 7, № 1. P. 1–11.

84. Al Bkhetan Z., Plewczynski D. Three-dimensional Epigenome Statistical Model: Genome-wide Chromatin Looping Prediction // Sci. Rep. Nature Publishing Group, 2018. Vol. 8, № 1. P. 5217.

85. Kai Y. et al. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features // Nat. Commun. Nature Publishing Group, 2018. Vol. 9, № 1. P. 4221.

86. Zhang R. et al. Predicting CTCF-mediated chromatin loops using CTCF-MP // Bioinformatics. Oxford University Press, 2018. Vol. 34, № 13. P. i133–i141.

87. Schwessinger R. et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning // Nat. Methods. Cold Spring Harbor

Laboratory, 2020. Vol. 17, № 11. P. 1118–1124.

88. Fudenberg G., Kelley D.R., Pollard K.S. Predicting 3D genome folding from DNA sequence with Akita // Nat. Methods. Nature Research, 2020. Vol. 17, № 11. P. 1111–1117.

89. Dekker J., Marti-Renom M.A., Mirny L.A. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data // Nature Reviews Genetics. NIH Public Access, 2013. Vol. 14, № 6. P. 390–403.

90. Imakaev M. V., Fudenberg G., Mirny L.A. Modeling chromosomes: Beyond pretty pictures // FEBS Letters. Elsevier, 2015. Vol. 589, № 20. P. 3031–3036.

91. Lin D. et al. Computational methods for analyzing and modeling genome structure and organization // Wiley Interdiscip. Rev. Syst. Biol. Med. Wiley-Blackwell, 2019. Vol. 11, № 1. P. e1435.

92. Tan J. et al. Cell type-specific prediction of 3D chromatin architecture // bioRxiv. Cold Spring Harbor Laboratory, 2022. P. 2022.03.05.483136.

93. Li R. et al. 3Disease Browser: A Web server for integrating 3D genome and disease-associated chromosome rearrangement data // Nat. Publ. Gr. 2016.

94. Ibn-Salem J. et al. Deletions of chromosomal regulatory boundaries are associated with congenital disease // Genome Biol. 2014. Vol. 15.

95. Zepeda-Mendoza C.J., Menon S., Morton C.C. Computational Prediction of Position Effects of Human Chromosome Rearrangements // Curr. Protoc. Hum. Genet. NLM (Medline), 2018. Vol. 97, № 1.

96. Hertzberg J. et al. TADA – a Machine Learning Tool for Functional Annotation based Prioritisation of Putative Pathogenic CNVs // bioRxiv. Cold Spring Harbor Laboratory, 2020. P. 2020.06.30.180711.

97. Sadowski M. et al. Spatial chromatin architecture alteration by structural variations in human genomes at the population scale // Genome Biol. BioMed Central Ltd., 2019. Vol. 20, № 1.

98. Wlasnowolski M. et al. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome // Nucleic Acids Res. NLM (Medline), 2020. Vol. 48, № W1. P. W170–W176.

99. Szabo Q. et al. TADs are 3D structural units of higher-order chromosome organization in Drosophila // Sci. Adv. Sci Adv, 2018. Vol. 4, № 2.

100. Schmiedel B.J. et al. 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells // Nat. Commun. Nature Publishing Group, 2016. Vol. 7, № 1. P. 1–14.

101. Sun Y. et al. 3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize // Genome Biol. BioMed Central Ltd., 2020. Vol. 21, № 1. P. 1–25.

102. Bogu G.K. et al. Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. // Mol. Cell. Biol. American Society for Microbiology, 2015. Vol. 36, № 5. P. 809–819.

103. Shen Y. et al. A map of the cis-regulatory sequences in the mouse genome // Nature. 2012. Vol. 488, № 7409. P. 116–120.

104. Bonev B. et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. // Cell. Elsevier, 2017. Vol. 171, № 3. P. 557-572.e24.

105. Durand N.C. et al. 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. Vol. 3. P. 95–98.

106. Andersson R. et al. An atlas of active enhancers across human cell types and tissues // Nature. Nature Publishing Group, 2014. Vol. 507, № 7493. P. 455–461.

107. Zhang Y. et al. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N // Genome Res. 2021. Vol. 31, № 7. P. 1290–1295.

108. Ramírez F. et al. deepTools2: a next generation web server for deep-sequencing data analysis // Nucleic Acids Res. 2016. Vol. 44, № W1. P. W160–W165.

109. Kovaka S. et al. Transcriptome assembly from long-read RNA-seq alignments with StringTie2 // Genome Biol. 2019. Vol. 20, № 1. P. 278.

110. Falk M. et al. Heterochromatin drives compartmentalization of inverted and conventional nuclei // Nature. 2019. Vol. 570, № 7761. P. 395–399.

111. Whalen S., Truty R.M., Pollard K.S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin // Nat. Genet. Nature Publishing Group, 2016. Vol. 48, № 5. P. 488–496.

112. Xi W., Beer M.A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy // PLOS

Comput. Biol. 2018. Vol. 14, № 12. P. e1006625.

113. Yang Y. et al. Exploiting sequence-based features for predicting enhancer–promoter interactions // Bioinformatics. 2017. Vol. 33, № 14. P. i252–i260.
114. Singh S. et al. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks // Quant. Biol. 2019. Vol. 7, № 2. P. 122–137.
115. Phanstiel D.H. et al. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development // Mol. Cell. 2017. Vol. 67, № 6. P. 1037-1048.e6.
116. Nuriddinov M., Fishman V. C-InterSecture-a computational tool for interspecies comparison of genome architecture // Bioinformatics. Oxford University Press, 2019. Vol. 35, № 23. P. 4912–4921.
117. Belokopytova P.S. et al. Quantitative prediction of enhancer–promoter interactions // Genome Res. Cold Spring Harbor Laboratory Press, 2020. Vol. 30, № 1. P. 72–84.
118. Evelyn Kabirova, Anastasiya Ryzhkova, Varvara Lukyanchikova, Anna Khabarova, Alexey Korablev, Tatyana Shnaider, Miroslav Nuriddinov, Polina Belokopytova, Galina Kontsevaya, Irina Serova N.B. TAD border deletion at the Kit locus causes tissue-specific ectopic activation of a neighboring gene // bioRxiv. 2022.
119. Fortin J.-P., Hansen K.D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data // Genome Biol. 2015. Vol. 16, № 1. P. 180.
120. Zeng W., Wu M., Jiang R. Prediction of enhancer-promoter interactions via natural language processing // BMC Genomics. 2018. Vol. 19, № S2. P. 84.
121. Di Pierro M. et al. De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture // Proc. Natl. Acad. Sci. U. S. A. National Academy of Sciences, 2017. Vol. 114, № 46. P. 12126–12131.
122. Qi Y., Zhang B. Predicting three-dimensional genome organization with chromatin states // PLoS Comput. Biol. Public Library of Science, 2019. Vol. 15, № 6.
123. Szabo Q., Bantignies F., Cavalli G. Principles of genome folding into topologically associating domains // Science Advances. American Association for the Advancement of Science, 2019. Vol. 5, № 4. P.

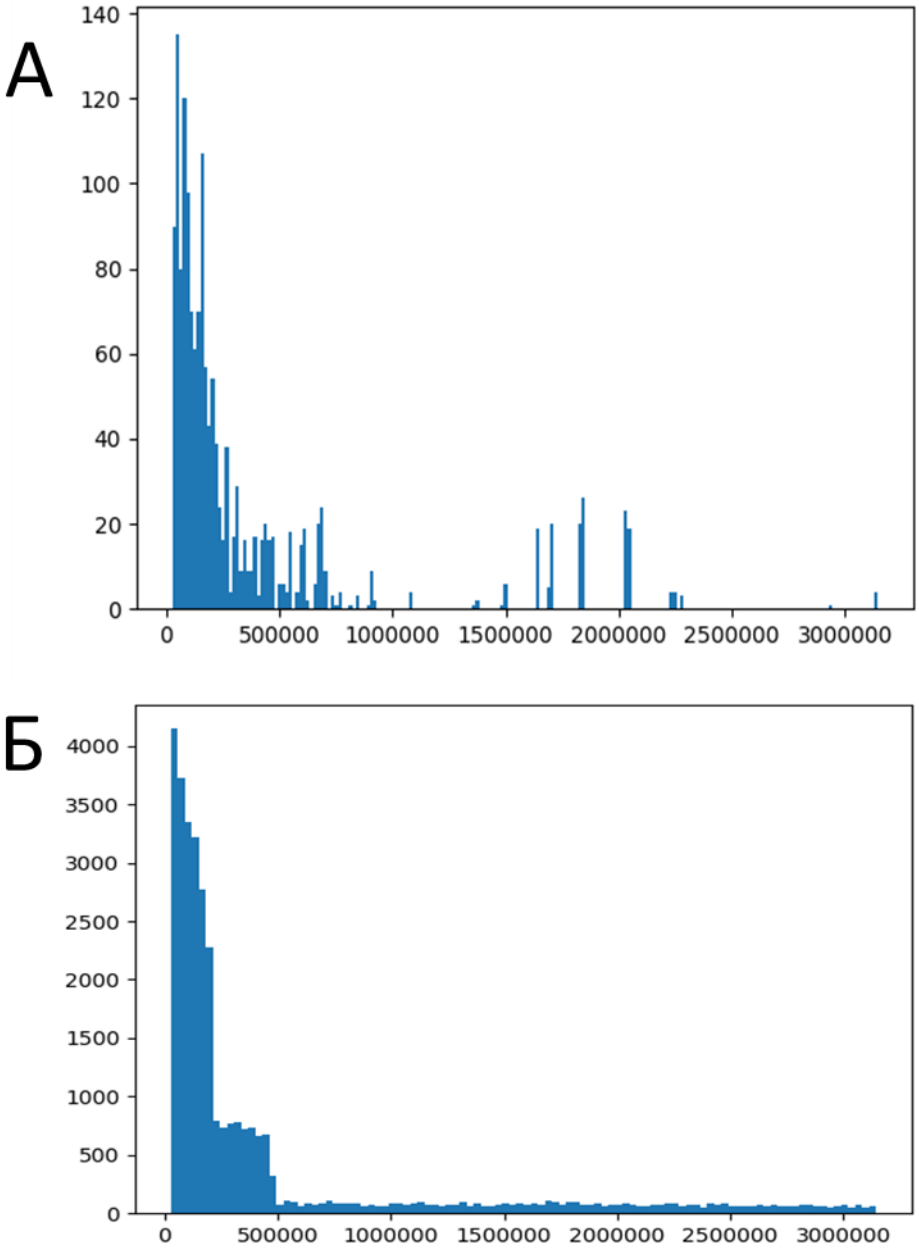
eaaw1668.

124. Kragesteen B.K. et al. Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis // Nat. Genet. 2018. Vol. 50, № 10. P. 1463–1473.
125. Golov A.K. et al. C-TALE, a new cost-effective method for targeted enrichment of Hi-C/3C-seq libraries // Methods. 2020. Vol. 170. P. 48–60.
126. Cao Q. et al. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. 2017.
127. Wang R. et al. MyoD is a 3D genome structure organizer for muscle cell identity // Nat. Commun. 2022. Vol. 13, № 1. P. 205.
128. Zhang S., Tao W., Han J.-D.J. 3D chromatin structure changes during spermatogenesis and oogenesis // Comput. Struct. Biotechnol. J. 2022. Vol. 20. P. 2434–2441.
129. Ryzhkova A. et al. Erythrocytes 3D genome organization in vertebrates // Sci. Rep. 2021. Vol. 11, № 1. P. 4414.
130. Miura H., Hiratani I. Cell cycle dynamics and developmental dynamics of the 3D genome: toward linking the two timescales // Curr. Opin. Genet. Dev. 2022. Vol. 73. P. 101898.
131. Lukyanchikova V. et al. Anopheles mosquitoes reveal new principles of 3D genome organization in insects // Nat. Commun. Cold Spring Harbor Laboratory, 2022. Vol. 13, № 1. P. 1960.
132. Spielmann M., Lupiáñez D.G., Mundlos S. Structural variation in the 3D genome // Nat. Rev. Genet. 2018. Vol. 19, № 7. P. 453–467.
133. Avsec Ž. et al. Effective gene expression prediction from sequence by integrating long-range interactions // Nat. Methods. 2021. Vol. 18, № 10. P. 1196–1203.
134. Keilwagen J., Posch S., Grau J. Accurate prediction of cell type-specific transcription factor binding // Genome Biol. 2019. Vol. 20, № 1. P. 9.
135. Wong A.K. et al. Decoding disease: from genomes to networks to phenotypes // Nat. Rev. Genet. 2021. Vol. 22, № 12. P. 774–790.
136. Carmen Bravo González-Blas, Seppe De Winter, Gert Hulselmans, Nikolai Hecker, Irina Matetovici, Valerie Christiaens, Suresh Poovathingal, Jasper Wouters, Sara Aibar. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks // biorxiv.org. 2022.

137. Maria Sindeeva, Nikolay Chekanov, Manvel Avetisian, Nikita Baranov, Elian Malkin, Alexander Lapin, Olga Kardymon V.F. Cell type-specific interpretation of noncoding variants using deep learning-based methods // bioRxiv. 2021.

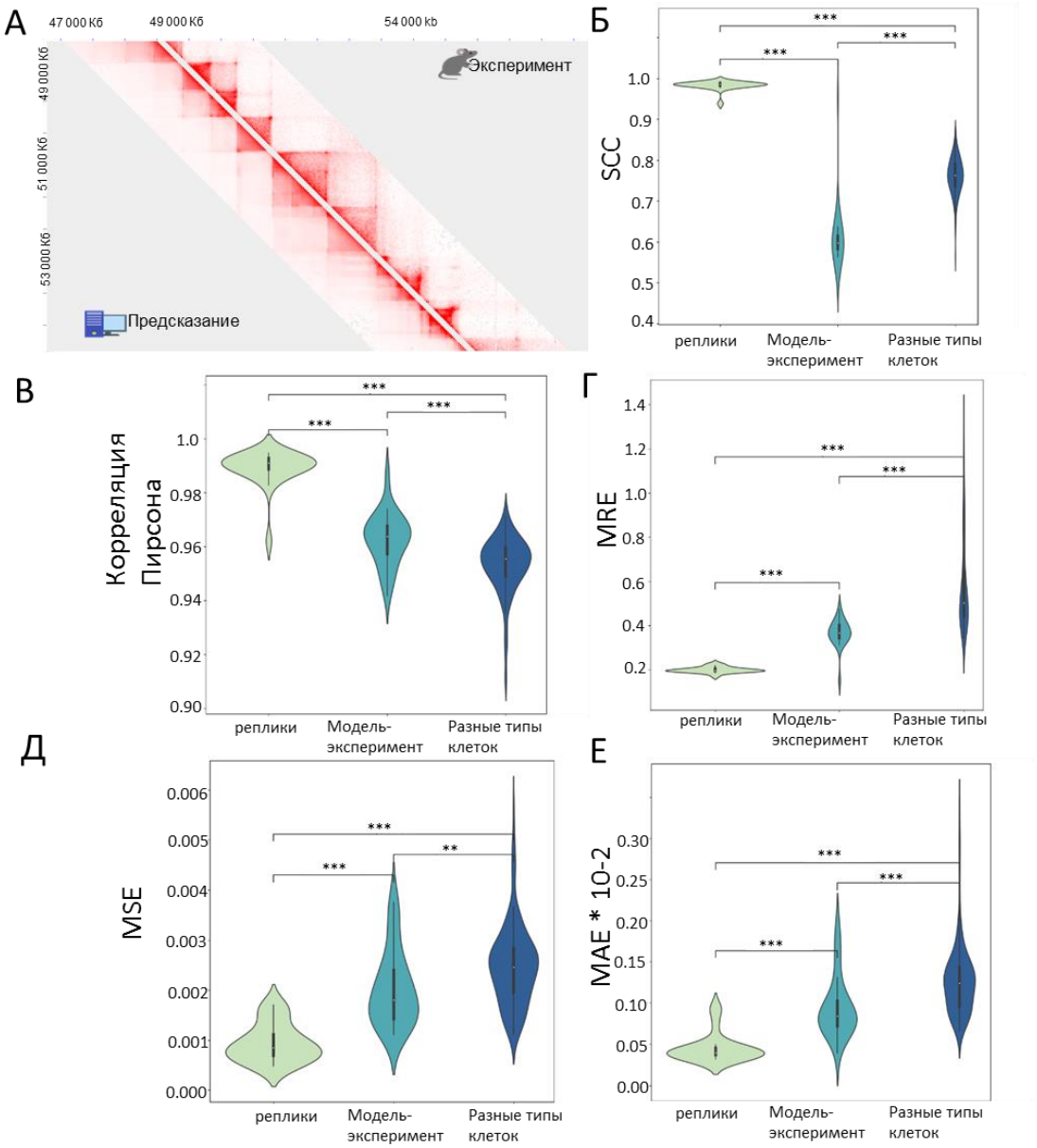
ПРИЛОЖЕНИЕ 1

Распределение расстояний между энхансером и промотором: А для взаимодействующих пар; Б для невзаимодействующих пар для выборки с взаимодействующими парами, определенными из Hi-C данных. По оси X отмечено расстояние между промотором и энхансером в п.н.



ПРИЛОЖЕНИЕ 2

Предсказание пространственной карты контактов для линии клеток человека GM12878. 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 хромосомы 14 на разрешении 5 Кб. (Б-Е) Значения метрик (MRE, SCC, MAE, корреляция Пирсона, MSE), полученные при сравнении Hi-C карт реплик между собой, предсказанных контактов линии клеток GM12878 с экспериментальными данными и пространственных контактов хроматина для разных типов клеток. Среднее значение получено как среднее из значений метрики для каждой хромосомы. Среднее SCC для сравнения модели и реплик (0.76) отличается от сравнения различных типов клеток (0.61), p-value 3.9e-21 (t-критерий Стьюдента). Среднее значение корреляции Пирсона для сравнения модели и реплик (0.96) отличается от сравнения различных типов клеток (0.95), p-value 7.8e-05 (t-критерий Стьюдента).


А
47 000 Кб
49 000 Кб
54 000 kb
49 000 Кб
51 000 Кб
53 000 Кб
Эксперимент
Предсказание
Б
SCC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
***
***
***
реплики
Модель-эксперимент
Разные типы клеток
В
Корреляция Пирсона
1.0
0.98
0.96
0.94
0.92
0.90
***
***
***
реплики
Модель-эксперимент
Разные типы клеток
Г
MRE
1.4
1.2
1.0
0.8
0.6
0.4
0.2
***
***
***
реплики
Модель-эксперимент
Разные типы клеток
Д
MSE
0.006
0.005
0.004
0.003
0.002
0.001
0.000
***
**
***
реплики
Модель-эксперимент
Разные типы клеток
Е
MAE * 10-2
0.30
0.25
0.20
0.15
0.10
0.05
0.00
***
***
***
реплики
Модель-эксперимент
Разные типы клеток

ПРИЛОЖЕНИЕ 3

Краткая характеристика инструментов, которые использовались для сравнения в работе.

Алгоритм	Входные Данные	Метод моделирования
3DPredictor	ChIP-seq CTCF, RNA-seq, геномное расстяние	Машинное обучение
3DPolyS-Fit [99]	Hi-C	Физическое моделирование
DRAGON [122]	ChIP-seq CTCF, гистоновые модификации	Физическое моделирование
PRISMR [55]	Hi-C	Физическое моделирование
MEGABASE+MiChroM [121]	гистоновые модификации	Машинное обучения + физическое моделирование
Модель Rowley с соавт. [77]	GRO-seq ChIP-seq различных транскрипционных факторов	Физическое моделирование

ПРИЛОЖЕНИЕ 4

Уровень различий между моделью и экспериментальными данными ближе к уровню различий между репликами, чем между разными типами клеток. (А-Г) Зеленый график: из значения метрики для различий между репликами вычиталось среднее значение метрики для различий между экспериментальными и предсказанными данными. Синий график: из значения метрики для различий между репликами вычиталось среднее значение метрики для различий между разными типами клеток. Среднее значение получено как среднее из описанных выше вычислений для каждой хромосомы. В качестве статистического критерия использовался t-критерий Стьюдента.