

Федеральное государственное бюджетное научное учреждение
Федеральный исследовательский центр «Институт цитологии и генетики
Сибирского отделения Российской академии наук»



На правах рукописи

ФИШМАН ВЕНИАМИН СЕМЕНОВИЧ

**Трехмерная организация хроматина у животных и ее
нарушения при хромосомных перестройках**

1.5.22. клеточная биология

1.5.8. математическая биология, биоинформатика

Диссертация в виде научного доклада на соискание учёной
степени доктора биологических наук

Новосибирск – 2024

Работа выполнена на базе Отдела молекулярных механизмов онтогенеза Федерального исследовательского центра “Институт цитологии и генетики СО РАН”

Научный консультант:	Серов Олег Леонидович , профессор, доктор биологических наук, заведующий отделом Молекулярных механизмов онтогенеза Федерального исследовательского центра “Институт цитологии и генетики СО РАН”.
Официальные оппоненты:	Графодатский Александр Сергеевич , профессор, доктор биологических наук, член-корреспондент РАН, заведующий Отделом разнообразия и эволюции геномов ФГБУН “Институт молекулярной и клеточной биологии Сибирского отделения Российской академии наук”
	Поляков Александр Владимирович , профессор, доктор биологических наук, член-корреспондент РАН, заведующий лабораторией ДНК-диагностики ФГБУН “Медико-генетический научный центр имени академика Н.П. Бочкова“
	Степанов Вадим Анатольевич , профессор, доктор биологических наук, академик РАН, директор ФГБУН “Томский национальный исследовательский медицинский центр Российской академии наук”
Ведущая организация:	ФГБУН “Институт общей генетики им. Н.И. Вавилова Российской академии наук”

Защита диссертации состоится «__» ноября 2024 г. на утреннем заседании Диссертационного совета 24.1.239.01 на базе ФГБНУ «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения РАН» в конференц-зале Института по адресу: 630090, г. Новосибирск, проспект ак. Лаврентьева, 10, т. (383)363-49-06, факс (383) 333-12-78, e-mail: dissov@bionet.nsc.ru. С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте Института <http://www.icgbio.ru/>

Научный доклад разослан «__» _____ 2024 г.

Ученый секретарь
диссертационного совета,
доктор биологических наук

Хлебодарова Т.М.

Оглавление

1. Введение, цель и задачи исследования.....	2
Актуальность исследования.....	5
Научная новизна.....	6
Практическая значимость.....	7
Методология и методы исследования.....	8
Положения, выносимые на защиту.....	10
Апробация работы.....	11
Степень достоверности полученных результатов.....	13
Публикации по теме научного доклада.....	13
Личный вклад автора.....	19
2. ОСНОВНОЕ СОДЕРЖАНИЕ ДОКЛАДА.....	20
2.1. Сравнительный анализ укладки хроматина в разных типах клеток и у разных видов позвоночных животных.....	20
2.2. Сравнительный анализ укладки хроматина у комаров рода <i>Anopheles</i>	26
2.3. Разработка биоинформационных методов для предсказания трехмерной архитектуры хроматина и эпигенетических характеристик хроматина в клетках млекопитающих.....	39
2.4. Использование технологий захвата конформации хромосом для поиска хромосомных перестроек у человека.....	54
2.4.1. Разработка технологии Ехо-С.....	54
2.4.2. Применение технологии Ехо-С для идентификации и характеристики хромосомных перестроек у пациентов.....	58
2.5. Применение методов геномики и биоинформатики для решения задач в области генетики развития и медицинской генетики.....	66
3. ЗАКЛЮЧЕНИЕ И ВЫВОДЫ.....	67
БЛАГОДАРНОСТИ.....	72
СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ.....	72

1. Введение, цель и задачи исследования

Развитие многоклеточного организма происходит за счет процессов дифференцировки и специализации клеток. Поскольку подавляющее большинство клеток организма несет практически идентичную генетическую информацию, формирование различных клеточных типов в первую очередь обуславливается эпигенетическими механизмами регуляции экспрессии генов. У высших животных система эпигенетической регуляции устроена сложно и включает многочисленные, часто взаимодействующие между собой молекулярные пути. Тонкая настройка систем эпигенетической регуляции критически важна для развития и жизнедеятельности организма, поскольку многие белки, например факторы транскрипции, чувствительны даже к небольшим отклонениям в концентрации.

Ещё в прошлом веке были сформулированы представления о том, что трехмерная укладка ДНК в ядре является неслучайной, а формирование специфических трехмерных структур хроматина может быть важным элементом регуляции экспрессии генов. С появлением новых молекулярных и цитологических методов появилась возможность детально описать укладку ДНК в ядре ¹. Важнейшим прорывом стала разработка полногеномных вариантов захвата конформации хромосом ² - семейства биохимических методов, позволяющих охарактеризовать ранее недоступный исследователям уровень укладки ДНК в масштабах от тысячи до сотен тысяч пар оснований ¹. Был открыт ряд специфических структур интерфазного хроматина, таких как компартменты ² и топологически ассоциированные домены ³. Компартменты - пространственно-разделенные районы генома, характеризующиеся разным профилем связывания белков хроматина. Топологически-ассоциированные домены (ТАДы) - локальные, инсулированные (т.е. пространственно-изолированные) от окружающих последовательностей участки генома, внутри которых наблюдается повышенная частота контактов хроматина.

Неслучайное распределение функциональных элементов генома, в частности промоторов генов и регуляторных элементов, относительно ТАДов и компартментов позволило сформулировать предположение о роли этих трехмерных структур в

регуляции различных геномных процессов, в первую очередь, в регуляции экспрессии генов. В то же время, работы нашей группы ⁴ и других авторов ³ показали, что границы ТАДов часто совпадают в разных клеточных типах мыши и человека, инициировав дискуссию о том, как могут структуры хроматина, в значительной степени инвариантные, обеспечивать тонкую клеточно-специфичную настройку уровня транскрипции генов.

Параллельно с уточнением функциональной роли ТАДов, активно изучались механизмы формирования этих структур. Оказалось, что расположение ТАДов в клетках человека хорошо объясняется взаимодействием между комплексом когезин, осуществляющим АТФ-зависимое протягивание (экструзию) петель, с белком CTCF. В то же время для дрозофилы было показано, что связывание белка CTCF не играет существенной роли в формировании ТАДов ⁵. Принципиальное различие в механизмах формирования ТАДов привело к тому, что некоторые исследователи предложили альтернативные варианты названий: петлевые домены (петлевые ТАДы), домены хроматина, физические домены и т.д. В рамках этой работы, мы будем называть ТАДами любые участки генома, соответствующие определению выше, независимо от механизма их формирования. Однако важно отметить, что помимо ткане-специфичности актуальным является вопрос видо-специфичности ТАДов, момента их появления в филогенезе, универсальности механизмов, обуславливающих их формирование, и значения этих структур с точки зрения эволюции животных.

Помимо локус-специфичных особенностей организации хроматина использование методов захвата конформации хромосом позволило оценить важные биофизические параметры укладки генома. Благодаря этому, появились возможности использования широкого арсенала методов из области физики полимеров и сопоставления биофизических моделей с экспериментальными данными ⁶. Важно отметить, что зависимость частоты трехмерных контактов хроматина от геномного расстояния, которая хорошо описывается степенной функцией, согласуется с моделью “фрактальной глобулы” - специфической формы укладки ДНК, которая противопоставляется модели “случайных блужданий” полимера ². Помимо

фундаментального значения с точки зрения физики полимеров, описание степенной зависимости контактов хроматина от геномного расстояния открыло возможности использования методов захвата конформации хромосом для сборки геномов, гаплотипирования, а также детекции хромосомных перестроек^{7,8}.

Вскоре стало ясно, что изменения трехмерных контактов, вызванные хромосомными перестройками и другими геномными вариантами, не только являются удобным индикатором, позволяющим детектировать хромосомную перестройку, но и могут играть важное функциональное значение с точки зрения развития патологий⁹. Ряд работ убедительно показал, что нарушения укладки отдельных локусов может приводить к патологическим изменениям генной экспрессии, сопровождающихся дефектами эмбрионального развития или связанными с онкологическими заболеваниями. В то же время, было показано, что далеко не все хромосомные перестройки приводят к существенным нарушениям активности генов. Учитывая важное прикладное значение этих исследований и сложность анализа, связанного со статистическим характером данных о трехмерных контактах хроматина, сформировалась потребность в разработке специализированных экспериментальных методов и биоинформационных инструментов для изучения укладки генома и ее изменений, связанных с возникновением хромосомных перестроек.

В рамках вышеперечисленных проблем, нами была поставлена **цель**: исследовать изменения трехмерных контактов хроматина, возникающие в процессах клеточной дифференцировки, эволюции и вследствие эффектов хромосомных перестроек у животных. Для достижения этой цели были поставлены следующие задачи:

1. Исследовать взаимосвязь между распределением эпигенетических характеристик и трехмерных контактов хроматина в фибробластах и эритроцитах курицы домашней *Gallus gallus* и провести сравнительный анализ укладки генома в этих клетках.

2. Сравнить распределение трехмерных контактов хроматина в соматических клетках птиц и млекопитающих.

3. Исследовать трехмерные контакты хроматина в клетках личинок и имаго пяти видов комаров рода *Anopheles*. Провести как внутривидовое сравнение разных типов клеток, так и межвидовое сравнение укладки хроматина.

4. Оценить возможность использования технологий захвата конформации хромосом для детекции хромосомных перестроек и однонуклеотидных вариантов у пациентов с пороками развития.

5. Разработать инструменты для аннотации эпигенетических профилей в различных типах клеток человека на основе анализа последовательностей ДНК

Для решения этих задач

6. Разработать экспериментально-компьютерную платформу для получения и анализа данных о трехмерной укладке хроматина, включающую модули для

- сравнения контактов хроматина в клетках разных видов и в разных типах клеток одного вида;
- предсказания контактов хроматина в клетках млекопитающих на основе эпигенетических характеристик и оценки точности полученных предсказаний;
- предсказания эффекта геномных вариантов на трехмерную укладку хроматина и другие эпигенетические характеристики в клетках человека;
- детекции и функциональной аннотации хромосомных перестроек в клетках человека;

Актуальность исследования

Изучение трехмерной укладки хроматина предоставляют одно из наиболее активно развивающихся фундаментальных направлений геномики. Интерес к этой области связан с представлением о том, что укладка хроматина в ядре клетки играет большое, а в некоторых случаях даже определяющее значение для процессов транскрипции, репликации и репарации ДНК. В то же время, за пределами модельных клеточных линий и видов организмов, таких как человек или мышь, укладка хроматина остается малоизученной. Это не позволяет представить масштаб ее изменчивости в норме и патологии, определить, насколько консервативны

определяющие укладку ДНК механизмы, как и когда они были сформированы в ходе эволюции животных и к каким функциональным последствиям могут вести их нарушения. Актуальность работы связана с исследованием уникальных биологических объектов - с точки зрения их филогенетического положения относительно ранее изученных видов, необычной организации ядра, или наличия редких хромосомных перестроек..

Следует отметить, что благодаря развитию технологий секвенирования и появлению других высокопроизводительных технологий, процесс накопления экспериментальных данных зачастую опережает возможности их анализа. Подобная ситуация особенно характерна для полногеномных исследований, в которых из-за большого объема и статистической природы экспериментальных данных разработка качественных инструментов анализа критически важна для получения надежных результатов. Актуальность части работы, связанной с созданием и применением биоинформатических инструментов, обусловлена сформировавшимся разрывом между скоростью накопления экспериментальных данных в геномике и способами их анализа.

Научная новизна

Научная новизна работы связана с исследованием ранее не охарактеризованных с точки зрения трехмерной укладки хроматина таксономических групп, специфических клеточных типов и редких генетических вариантов, ассоциированных с различными патологиями развития, а также с разработкой и применением новых экспериментальных и вычислительных инструментов для анализа геномных данных.

Впервые показано радикальное изменение укладки ТАДов в ходе созревания эритроцитов. Впервые показано формирование ТАДов и компартментов в соматических клетках птиц. Впервые проведено количественное сравнение частот трехмерных контактов хроматина в клетках птиц и млекопитающих и получена количественная оценка различий, наблюдаемых в укладке хроматина у этих видов.

Впервые получены данные о трехмерной укладке хроматина на уровне ТАДов и компартментов для комаров рода *Anopheles*, а также высококонтинуальные сборки

геномов этих видов. Предложена гипотеза, позволяющая объяснить эволюционную консервативность ТАДов у комаров рода *Anopheles*. Впервые обнаружены протяженные (до 10 миллионов пар оснований), эволюционно-консервативные хроматиновые петли в геномах комаров рода *Anopheles*.

Предложен новый экспериментальный подход для одновременной полногеномной детекции хромосомных перестроек и поиска однонуклеотидных вариантов в экзOME человека.

Разработаны новые вычислительные инструменты для межвидового сравнения данных Hi-C (High-throughput chromosome conformation capture, высокопроизводительный захват конформации хромосом), предсказания контактов хроматина и других эпигенетических характеристик в норме и при патологии.

Практическая значимость

Изучение механизмов укладки хроматина и роли контактов ДНК в различных геномных процессах важно для решения различных прикладных задач. В контексте этой работы акцент делается на задачи из области медицинской генетики, в частности на применение методов 3D-геномики для детекции и функциональной интерпретации геномных вариантов у пациентов с различными наследственными и онкологическими заболеваниями.

Опубликованные и полученные в рамках работы данные убедительно показывают, что без глубокого понимания эпигенетических механизмов регуляции экспрессии генов невозможно объяснить возникновение целого спектра патологий человека. Однако механизмы эпигенетической регуляции настолько сложны и многообразны, что сложно представить использование этих знаний врачом в рутинной клинической практике. В связи с этим, большое практическое значение имеют созданные в рамках данной работы вычислительные инструменты, способные предсказывать влияние геномных вариантов на контакты хроматина и другие эпигенетические характеристики. Их использование может быть автоматизировано и не требует глубокого погружения в теоретические основы современной геномики.

Важное прикладное значение имеют разработанные экспериментальные методики и вычислительные инструменты для поиска хромосомных перестроек. Детекция сбалансированных хромосомных перестроек субмикроскопического размера до сих пор является одной из самых сложных задач медицинской генетики. Разработанные нами экспериментальные методы позволяют с высокой чувствительностью детектировать хромосомные перестройки размером более 100 тысяч пар оснований, что является заметным вкладом в решении этой сложной проблемы. Крайне важно, что созданные в рамках работы вычислительные инструменты предоставляют возможность автоматизировать процесс поиска геномных вариантов, поскольку автоматизация анализа является критически важным условием для внедрения любой методики в клинику.

Методология и методы исследования

Большинство экспериментальных методов исследования укладки ДНК, использованных в работе, основаны на различных модификациях технологии захвата конформации хромосом. В наиболее часто используемом полногеномном варианте этой технологии, который называется Hi-C, предполагается проведение следующих этапов. Сначала клетки фиксируют формальдегидом, после чего проводят мягкий лизис, что позволяет обеспечить доступ ферментов к ДНК, сохраняя при этом укладку хроматина. Хроматин обрабатывают эндонуклеазами рестрикции и проводят лигирование в условиях низкой концентрации ДНК. Низкая концентрация субстрата и фиксация материала, которая сохраняется на протяжении всего эксперимента, обеспечивают преимущественное лигирование концов ДНК, которые на момент фиксации находились в клетке близко друг к другу. Секвенирование позволяет количественно определить частоту возникновения продуктов лигазной реакции и, таким образом, оценить вероятность сближения соответствующих участков генома. В литературе эту величину принято называть частотой контактов хроматина. Парные частоты контактов хроматина для всего генома или отдельных его участков графически представляют в виде тепловых карт (или Hi-C-карт).

В работе используется как классический метод Hi-C, так и его модификации. Например, для достижения более равномерного распределения сайтов рестрикции на этапе гидролиза хроматина использовались ферменты ДНКазы I и S1-нуклеаза. Для исследования ооцитов куриц, на основе сравнительно анализа существовавших протоколов, разработан наиболее эффективный метод Hi-C индивидуальных клеток (single-cell Hi-C).

В работе используются и другие методы геномики, включая приготовление геномных, транскриптомных и метиломных библиотек для секвенирования, поиск сайтов связывания белков хроматина методом иммунопреципитации и секвенирования (ChIP-seq), прочтение длинных фрагментов ДНК технологиями секвенирования третьего поколения и т.д.

Для создания моделей и объектов исследования был использован ряд подходов клеточной и молекулярной биологии, включая получение и дифференцировку индуцированных плюрипотентных стволовых клеток (ИПСК), направленную модификацию геномов животных при помощи технологии CRISPR/Cas9.

Для анализа полученных в работе данных используется широкий арсенал методов биоинформатики, включающий:

- программные инструменты доступа к базам данных (Ensemble REST API, BioMart, ENCODE batch download и др.),
- стандартные инструменты анализа и визуализации геномных данных (bowtie2, bwa, minimap2, bedtools, deeptools, samtools, bcftools, IGV, juicebox и т.д.),
- комбинации (пайплайны) биоинформатических инструментов (Hi-C Pro, juicer, cooler, GATK toolkit, Ensemble VEP, AnnoVar и т.д.),

Кроме того, в ходе работы проведена разработка собственных инструментов, преимущественно на языке Python с использованием широкого спектра специализированных библиотек, включая модули, обеспечивающие хранение, фильтрацию и статистический анализ данных (pandas, scipy, numpy, sklearn, h5py), а также модули для работы со специфическими форматами биологических данных (pysam, pybedtools, pyvcf, biopython). Для разработки предсказательных моделей с

использованием методов машинного обучения использовалась стандартная методология, включающая параметризацию предикторов, разбиение наборов данных на тренировочную, валидационную и тестовую часть, подбор гиперпараметров моделей. Для разработки таких инструментов использовались модули pytorch, tensorflow, xgboost и другие специализированные пакеты для машинного обучения.

При анализе клинического материала автор строго руководствовался Российским руководством по интерпретации данных последовательности ДНК человека, полученных методами массового параллельного секвенирования. Исследования были одобрены локальной биоэтической комиссией и проведены после получения информированного добровольного согласия участников.

Положения, выносимые на защиту

1. Принципы укладки хроматина в ядрах эритроцитов курицы являются эволюционно-консервативными среди позвоночных животных и отличаются от характерных для других соматических клеток принципов укладки ДНК отсутствием CTCF-опосредованных топологически ассоциированных доменов и усиленной компартментализацией генома.

2. Особенностью трехмерной организации хроматина в клетках комаров *Anopheles* определяется наличие характерных, консервативных в пределах рода, петлевых взаимодействии между локусами X-хромосомы, расположенными на расстояниях от нескольких десятков тысяч до нескольких миллионов пар оснований друг от друга.

3. Новая архитектура нейронной сети DeepCT, включающая модули для анализа нуклеотидной последовательности и клеточного типа, позволяет представлять клеточно-специфичные особенности эпигенетической регуляции генома человека в форме многомерных векторов и предсказывать функциональные последствия изменений нуклеотидной последовательности ДНК для различных типов клеток.

4. Компьютерно-экспериментальная методика Echo-C на основе технологий Hi-C и экзоминого обогащения позволяет детектировать большинство сбалансированных и несбалансированных хромосомных перестроек размером более

сотен тысяч пар оснований, а также однонуклеотидные мутации в экзонах генов, в том числе клинически-значимых.

5. Экспериментально-компьютерная платформа, разработанная для анализа и реконструкции 3D-архитектуры генома, включает программные продукты 3DGenBench, 3DPredictor, C-InterSecture, ABCE, FastContext, Exoclasma и Exo-C, которые позволяют на основе данных Hi-C (High-throughput chromosome conformation capture) с высокой точностью выявлять клеточно-специфические особенности трехмерной организации хроматина у различных видов организмов, в том числе возникающие в процессе клеточной дифференцировки и/или в результате хромосомных перестроек.

Апробация работы

Результаты работы были представлены на следующих международных и российских конференциях и научно-практических школах:

1. Всероссийская цитогеномная конференция, онлайн, 21-24 июня 2023
2. CRISPR 2023, Новосибирск, Россия, 11-13 сентября 2023
3. Хромосома-2023, Новосибирск, Россия, 5-9 сентября 2023
4. Зимняя школа ПИЯФ 2023, Санкт-Петербург, Россия, 13–18 марта, 2023
5. GBB-2022: Школа “Тены. Мозг. Поведение”. Сочи, Россия, 14-16 декабря 2022
6. Генетика человека и патология, XIII научная конференция, Томск, Россия, 20-22 ноября 2022
7. Юбилейная научная конференция «Николай Константинович Кольцов и биология XXI века», Москва, Россия, 3-8 октября 2022
8. Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS/SB-2022): The Thirteenth International Multiconference, Новосибирск, Россия, 4-8 июля 2022
9. 13-ая Школа «Системная биология и Биоинформатика» SBB-2021 («Systems biology and Bioinformatics»), Новосибирск, Россия, 4-9 октября 2021
10. 13th European Cytogenomics Conference, online, 3-5 июля 2021

11. IX Съезд Российского общества медицинских генетиков. Москва, Россия, 3-4 июля, 2021
12. Диалоги о геномике: лучшие практики лабораторий РФ и Европы, Сочи, Россия, 7–10 июля, 2021
13. MGNGS'21 - NGS в медицинской генетике, Суздаль, Россия, 28–30 апреля 2021
14. GBB-2021: Школа “Гены. Мозг. Поведение”. Сочи, Россия, 13-14 декабря 2021
15. MCCMB-2021: 10ая Московская конференция по вычислительной молекулярной биологии, Москва, Россия, 30 июля - 2 августа 2021
16. GBB-2021: Школа “Гены. Мозг. Поведение”, онлайн, 14-16 декабря 2020
17. The 7th International Symposium on 3D Genomics, онлайн, 6-7 декабря 2020
18. BGRS/SB-2020: 12th International Multiconference “Bioinformatics of Genome Regulation and Structure/Systems Biology”, онлайн, 06-10 июля 2020
19. XII Научная конференция «Генетика человека и патология: Актуальные проблемы клинической и молекулярной цитогенетики», Томск, Россия, 20-22 ноября 2019 года, г. Томск
20. Четвёртая международная научно-практическая конференция «NGS в медицинской генетике 2019», Суздаль, Россия, 24–26 апреля 2019
- 21.4 Съезд Ассоциации специалистов медицины плода «Национального общества пренатальной медицины» с научно-практической конференцией «Пренатальная диагностика в междисциплинарной помощи матери и ребенку» при международном участии, Москва, Россия, 27-29 сентября 2019
22. 11-я Международная школа молодых ученых «Системная биология и биоинформатика» - SBB-2019, Новосибирск, Россия, 24-28 июня 2019
23. XVIII Конференция-школа с международным участием “Актуальные проблемы биологии развития”, Москва, Россия, 14-19 октября 2019
24. VI Съезд биохимиков России (2019), Сочи (Дагомыс), Россия, 1-6 октября 2019

25. EMBO Workshop “The genome in three dimensions”, Kyllini, Греция, 20 – 24 мая 2019
26. 11-ая Международная мультikonференция по биоинформатике регуляции и структуры геномов и системной биологии / Bioinformatics of Genome Regulation and Structure\ Systems Biology - BGRS\SB-2018, Новосибирск, Россия, 20-25 августа 2018
27. Хромосома 2018, 20-24 августа 2018
28. Международная мини-конференция «Хромосомы и митоз», Новосибирск, Россия, 8 декабря 2017
29. 25th Wilhelm Bernhard Workshop on the Cell Nucleus, Нижний Новгород, Россия, 19-22 июня 2017.

Степень достоверности полученных результатов

Достоверность результатов определяется использованием современных экспериментальных и вычислительных подходов с необходимым количеством контрольных экспериментов с рационально обоснованными размерами выборки и подтверждается публикацией результатов в высокорейтинговых рецензируемых журналах.

Публикации по теме научного доклада

Общий список публикаций по теме научного доклада включает 53 работы, опубликованные в 2015-2024 гг. в журналах из БД WOS, Scopus или РИНЦ. Из них 35 публикаций в журналах, индексируемых базами данных WOS и Scopus и относящихся к категории Q1 или Q2:

1. Belokopytova, P., **Fishman, V.**, 2020. Predicting Genome Architecture: Challenges and Solutions. *Front. Genet.* 11, 617202. <https://doi.org/10.3389/fgene.2020.617202>. **WOS, Scopus, Q2**
2. Belokopytova, P., Viesná, E., Chiliński, M., Qi, Y., Salari, H., Di Stefano, M., Esposito, 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 Res.* 50, W4–W12. <https://doi.org/10.1093/nar/gkac396>. **WOS, Scopus, Q1**

3. Belokopytova, P.S., Nuriddinov, M.A., Mozheiko, E.A., Fishman, D., **Fishman, V.**, 2020. Quantitative prediction of enhancer–promoter interactions. *Genome Res.* 30, 72–84. <https://doi.org/10.1101/gr.249367.119>. **WOS, Scopus, Q1**
4. da Costa, S.S., **Fishman, V.**, Pinheiro, M., Rodrigueiro, A., Sanseverino, M.T., Zielinsky, P., Carvalho, C.M.B., Rosenberg, C., Krepischi, A.C.V., 2024. A germline chimeric KANK1-DMRT1 transcript derived from a complex structural variant is associated with a congenital heart defect segregating across five generations. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* 32, 6. <https://doi.org/10.1007/s10577-024-09750-2>. **WOS, Scopus, Q2**
5. Filatova, A., Reveguk, I., Piatkova, M., Bessonova, D., Kuziakova, O., Demakova, V., Romanishin, A., **Fishman, V.**, Imanmalik, Y., Chekanov, N., Skitchenko, R., Barbitoff, Y., Kardymon, O., Skoblov, M., 2023. Annotation of uORFs in the OMIM genes allows to reveal pathogenic variants in 5'UTRs. *Nucleic Acids Res.* 51, 1229–1244. <https://doi.org/10.1093/nar/gkac1247>. **WOS, Scopus, Q1**
6. **Fishman, V.**, Battulin, N., Nuriddinov, M., Maslova, A., Zlotina, A., Strunov, A., Chervyakova, D., Korablev, A., Serov, O., Krasikova, A., 2019. 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. *Nucleic Acids Res.* 47, 648–665. <https://doi.org/10.1093/nar/gky1103>. **WOS, Scopus, Q1**
7. **Fishman, V.**, Pindyurin, AV, 2022. Editorial: The role of high-order chromatin organization in gene regulation. *Front. Genet.* 13. <https://doi.org/10.3389/fgene.2022.1045787>
8. **Fishman, V.S.**, Salnikov, P.A., Battulin, N.R., 2018. Interpreting Chromosomal Rearrangements in the Context of 3-Dimensional Genome Organization: A Practical Guide for Medical Genetics. *Biochem. Biokhimiia* 83, 393–401. <https://doi.org/10.1134/S0006297918040107>. **WOS, Scopus, Q2**
9. **Fishman, V.S.**, Shnayder, T.A., Orishchenko, K.E., Bader, M., Alenina, N., Serov, O.L., 2015. Cell divisions are not essential for the direct conversion of fibroblasts into neuronal cells. *Cell Cycle* 14, 1188–1196. <https://doi.org/10.1080/15384101.2015.1012875>. **WOS, Scopus, Q1**
10. Gridina, M., **Fishman, V.**, 2022. Multilevel view on chromatin architecture alterations in cancer. *Front. Genet.* 13, 1059617. <https://doi.org/10.3389/fgene.2022.1059617>. **WOS, Scopus, Q2**
11. Gridina, M., Mozheiko, E., Valeev, E., Nazarenko, L.P., Lopatkina, M.E., Markova, Z.G., Yablonskaya, M.I., Voinova, V.Y., Shilova, N.V., Lebedev, I.N., **Fishman, V.**, 2021. A cookbook for DNase Hi-C. *Epigenetics Chromatin* 14, 15. <https://doi.org/10.1186/s13072-021-00389-5>. **WOS, Scopus, Q1**
12. Gridina, M., Taskina, A., Lagunov, T., Nurislamov, A., Kulikova, T., Krasikova, A., **Fishman, V.**, 2022. Comparison and critical assessment of single-cell Hi-C protocols. *Heliyon* 8. <https://doi.org/10.1016/j.heliyon.2022.e11023>. **WOS, Scopus, Q1**
13. Gridina, M.M., Matveeva, N.M., **Fishman, V.S.**, Menzorov, A.G., Kizilova, H.A., Beregovoy, N.A., Kovrigin, I.I., Pristyazhnyuk, I.E., Oscorbin, I.P., Filipenko, M.L., Kashevarova, A.A., Skryabin, N.A., Nikitina, T.V., Sazhenova, E.A., Nazarenko, L.P.,

- Lebedev, I.N., Serov, O.L., 2018. Allele-Specific Biased Expression of the CNTN6 Gene in iPS Cell-Derived Neurons from a Patient with Intellectual Disability and 3p26.3 Microduplication Involving the CNTN6 Gene. *Mol. Neurobiol.* 55, 6533–6546. <https://doi.org/10.1007/s12035-017-0851-5>. **WOS, Scopus, Q1**
14. Ivanoshchuk, D.E., Shakhtshneider, E.V., Rymar, O.D., Ovsyannikova, A.K., Mikhailova, S.V., **Fishman, V.S.**, Valeev, E.S., Orlov, P.S., Voevoda, M.I., 2021. The mutation spectrum of maturity onset diabetes of the young (Mody)-associated genes among western siberia patients. *J. Pers. Med.* 11. <https://doi.org/10.3390/jpm11010057>. **WOS, Scopus, Q2**
 15. Ivanov, A.A., Leonova, O.N., Wiebe, D.S., Krutko, A.V., Gridina, M.M., **Fishman, V.S.**, Aulchenko, Y.S., Tsepilov, Y.A., Golubeva, T.S., 2022. Method for the Isolation of “RNA-seq-Quality” RNA from Human Intervertebral Discs after Mortar and Pestle Homogenization. *Cells* 11, 3578. <https://doi.org/10.3390/cells11223578>. **WOS, Scopus, Q1**
 16. Kabirova, E., Nurislamov, A., Shadskiy, A., Smirnov, A., Popov, A., Salnikov, P., Battulin, N., **Fishman, V.**, 2023. Function and Evolution of the Loop Extrusion Machinery in Animals. *Int. J. Mol. Sci.* 24, 5017. <https://doi.org/10.3390/ijms24055017>. **WOS, Scopus, Q1**
 17. Kashevarova, A.A., Nazarenko, L.P., Skryabin, N.A., Nikitina, T.V., Vasilyev, S.A., Tolmacheva, E.N., Lopatkina, M.E., Salyukova, O.A., Chechetkina, N.N., Vorotelyak, E.A., Kalabusheva, E.P., **Fishman, V.S.**, Kzhyshkowska, J., Graziano, C., Magini, P., Romeo, G., Lebedev, I.N., 2018. A mosaic intragenic microduplication of LAMA1 and a constitutional 18p11.32 microduplication in a patient with keratosis pilaris and intellectual disability. *Am. J. Med. Genet. A.* 176, 2395–2403. <https://doi.org/10.1002/ajmg.a.40478>. **WOS, Scopus, Q2**
 18. Khabarova, A., Koksharova, G., Salnikov, P., Belokopytova, P., Mungalov, R., Pristyazhnuk, I., Nurislamov, A., Gridina, M., **Fishman, V.**, 2023. A Cre-LoxP-based approach for combinatorial chromosome rearrangements in human HAP1 cells. *Chromosome Res.* 31, 11. <https://doi.org/10.1007/s10577-023-09719-7>. **WOS, Scopus, Q2**
 19. Koksharova, G., Kokh, N., Gridina, M., Khapaev, R., Nimaev, V., **Fishman, V.**, 2024. A HGF Mutation in the Familial Case of Primary Lymphedema: A Report. *Int. J. Mol. Sci.* 25, 5464. <https://doi.org/10.3390/ijms25105464>. **WOS, Scopus, Q1**
 20. Krasheninina, O.A., **Fishman, V.S.**, Lomzov, A.A., Ustinov, A.V., Venyaminova, A.G., 2020. Postsynthetic On-Column 2' Functionalization of RNA by Convenient Versatile Method. *Int. J. Mol. Sci.* 21, 5127. <https://doi.org/10.3390/ijms21145127>. **WOS, Scopus, Q1**
 21. Krasheninina, O.A., Lomzov, A.A., **Fishman, V.S.**, Novopashina, D.S., Venyaminova, A.G., 2017. Rational design and studies of excimer forming novel dual probes to target RNA. *Bioorg. Med. Chem.* <https://doi.org/10.1016/j.bmc.2017.02.042>. **WOS, Scopus, Q2**
 22. Krasikova, A., **Fishman, V.**, Kulikova, T., 2023. Lampbrush chromosome studies in the post-genomic era. *BioEssays* 2200250. <https://doi.org/10.1002/bies.202200250>. **WOS,**

Scopus, Q1

23. Lukyanchikova, V., Nuriddinov, M., Belokopytova, P., Taskina, A., Liang, J., Reijnders, M.J.M.F., Ruzzante, L., Feron, R., Waterhouse, R.M., Wu, Y., Mao, C., Tu, Z., Sharakhov, I.V., **Fishman, V.**, 2022. Anopheles mosquitoes reveal new principles of 3D genome organization in insects. *Nat. Commun.* 13, 1960. <https://doi.org/10.1038/s41467-022-29599-5>. **WOS, Scopus, Q1**
24. Maslova, A., Plotnikov, V., Nuriddinov, M., Gridina, M., **Fishman, V.**, Krasikova, A., 2023. Hi-C analysis of genomic contacts revealed karyotype abnormalities in chicken HD3 cell line. *BMC Genomics* 24, 66. <https://doi.org/10.1186/s12864-023-09158-y>. **WOS, Scopus, Q1**
25. Matveeva, N.M., **Fishman, V.S.**, Zakharova, I.S., Shevchenko, A.I., Pristyazhnyuk, I.E., Menzorov, A.G., Serov, O.L., 2017. Alternative dominance of the parental genomes in hybrid cells generated through the fusion of mouse embryonic stem cells with fibroblasts. *Sci. Rep.* 7. <https://doi.org/10.1038/s41598-017-18352-4>. **WOS, Scopus, Q1**
26. Menzorov, A.G., Matveeva, N.M., Markakis, M.N., **Fishman, V.S.**, Christensen, K., Khabarova, A.A., Pristyazhnyuk, I.E., Kizilova, E.A., Cirera, S., Anistoroaei, R., Serov, O.L., 2015. Comparison of American mink embryonic stem and induced pluripotent stem cell transcriptomes. *BMC Genomics* 16, S6. <https://doi.org/10.1186/1471-2164-16-S13-S6>. **WOS, Scopus, Q1**
27. Menzorov, A.G., Orishchenko, K.E., **Fishman, V.S.**, Shevtsova, A.A., Mungalov, R.V., Pristyazhnyuk, I.E., Kizilova, E.A., Matveeva, N.M., Alenina, N., Bader, M., Rubtsov, N.B., Serov, O.L., 2019. Targeted genomic integration of EGFP under tubulin beta 3 class III promoter and mEos2 under tryptophan hydroxylase 2 promoter does not produce sufficient levels of reporter gene expression. *J. Cell. Biochem.* 120, 17208–17218. <https://doi.org/10.1002/jcb.28981>. **WOS, Scopus, Q2**
28. Nuriddinov, M., **Fishman, V.**, 2019. C-InterSection—a computational tool for interspecies comparison of genome architecture. *Bioinformatics* 35, 4912–4921. <https://doi.org/10.1093/bioinformatics/btz415>. **WOS, Scopus, Q1**
29. Nurislamov, A., Lagunov, T., Gridina, M., Krasikova, A., **Fishman, V.**, 2022. Single-Cell DNA Methylation Analysis of Chicken Lampbrush Chromosomes. *Int. J. Mol. Sci.* 23, 12601. <https://doi.org/10.3390/ijms232012601>. **WOS, Scopus, Q1**
30. Pristyazhnyuk, I.E., Minina, J., Korablev, A., Serova, I., **Fishman, V.**, Gridina, M., Rozhdestvensky, T.S., Gubar, L., Skryabin, B.V., Serov, O.L., 2019. Time origin and structural analysis of the induced CRISPR/cas9 megabase-sized deletions and duplications involving the Cntn6 gene in mice. *Sci. Rep.* 9, 14161. <https://doi.org/10.1038/s41598-019-50649-4>. **WOS, Scopus, Q1**
31. Ryzhkova, A., Taskina, A., Khabarova, A., **Fishman, V.**, Battulin, N., 2021. Erythrocytes 3D genome organization in vertebrates. *Sci. Rep.* 11, 4414. <https://doi.org/10.1038/s41598-021-83903-9>. **WOS, Scopus, Q1**
32. Sindeeva, M., Chekanov, N., Avetisian, M., Shashkova, T.I., Baranov, N., Malkin, E., Lapin, A., Kardymon, O., **Fishman, V.**, 2023. Cell type-specific interpretation of noncoding variants using deep learning-based methods. *GigaScience* 12, giad015. <https://doi.org/10.1093/gigascience/giad015>. **WOS, Scopus, Q1**

33. Smirnov, A., **Fishman, V.**, Yunusova, A., Korablev, A., Serova, I., Skryabin, B.V., Rozhdestvensky, T.S., Battulin, N., 2020. DNA barcoding reveals that injected transgenes are predominantly processed by homologous recombination in mouse zygote. *Nucleic Acids Res.* 48, 719–735. <https://doi.org/10.1093/nar/gkz1085>. **WOS, Scopus, Q1**
34. Yan, A.P., Salnikov, P.A., Gridina, M.M., Belokopytova, P.S., **Fishman, V.S.**, 2024. Towards Development of the 4C-Based Method Detecting Interactions of Plasmid DNA with Host Genome. *Biochem. Biokhimiia* 89, 653–662. <https://doi.org/10.1134/S0006297924040059>. **WOS, Scopus, Q2**
35. Yunusova, A.M., **Fishman, V.S.**, Vasiliev, G.V., Battulin, N.R., 2017. Deterministic versus stochastic model of reprogramming: new evidence from cellular barcoding technique. *Open Biol.* 7, 160311. <https://doi.org/10.1098/rsob.160311>. **WOS, Scopus, Q1**

А также 18 публикаций в журналах, индексируемых БД РИНЦ, WOS или Scopus, не относящиеся к категориям Q1 и Q2 по версии WOS и Scopus:

36. Beklemisheva, V.R., Belokopytova, P.S., **Fishman, V.S.**, Menzorov, A.G., 2021. Derivation of Ringed Seal (*Phoca hispida*) Induced Multipotent Stem Cells. *Cell. Reprogramming*. <https://doi.org/10.1089/cell.2021.0037>. **WOS, Scopus**
37. Fishman, V., Salnikov, P.A., Khabarova, A.A., Koksharova, G.S., Mungalov, R.V., Belokopytova, P.S., Pristyazhnyuk, I.E., Nurislamov, A.R., Somatich, P., Gridina, M.M., **Fishman, V.S.**, 2021. Here and there: the double-side transgene localization. *Vavilov J. Genet. Breed.* <https://doi.org/10.18699/VJ21.068>. **WOS, Scopus, РИНЦ**
38. Gridina, M.M., Vesna, E., Minzhenkova, M.E., Shilova, N.V., Ryzhkova, O.P., Nazarenko, L.P., Belyaeva, E.O., Lebedev, I.N., **Fishman, V.S.**, 2023. Influence of human peripheral blood samples preprocessing on the quality of Hi-C libraries. *Vavilovskii Zhurnal Genet. Sel.* 27, 83–87. <https://doi.org/10.18699/VJGB-23-11>. **WOS, Scopus, РИНЦ**
39. Mozheiko, E.A., **Fishman, V.S.**, 2019. Detection of Point Mutations and Chromosomal Translocations Based on Massive Parallel Sequencing of Enriched 3C Libraries. *Russ. J. Genet.* <https://doi.org/10.1134/S1022795419100089>. **WOS, Scopus, РИНЦ**
40. Pristyazhnyuk, I.E., Shnayder, T.A., **Fishman, V.S.**, Matveeva, N.M., Serov, O.L., 2015. Direct conversion of somatic cells to neuronal precursors: Problems and outlooks. *Russ. J. Genet. Appl. Res.* 5, 543–551. <https://doi.org/10.1134/s2079059715060106>. **WOS, Scopus, РИНЦ**
41. Savchenko, R.R., Vasilyev, S.A., Lebedev, I.N., **Fishman, V.S.**, Sukhikh, E.S., Sukhikh, L.G., Murashkina, A.A., 2020. Effect of the THBS1 Gene Knockout on the Radiation-Induced Cellular Response in a Model System In Vitro. *Russ. J. Genet.* 56, 618–626. <https://doi.org/10.1134/S1022795420050129>. **WOS, Scopus, РИНЦ**
42. Vasilyev, S.A., Savchenko, R.R., Belenko, A.A., Skryabin, N.A., Sleptsov, A.A., **Fishman, V.S.**, Murashkina, A.A., Gribova, O.V., Startseva, Z.A., Sukhikh, E.S., Vertinskiy, A.V., Sukhikh, L.G., Serov, O.L., Lebedev, I.N., 2022. ADAMTS1 Is

- Differentially Expressed in Human Lymphocytes with Various Frequencies of Endogenous γ H2AX Foci and Radiation-Induced Micronuclei. *Russ. J. Genet.* 58, 1235–1244. <https://doi.org/10.1134/s102279542210012x>. **WOS, Scopus, РИНЦ**
43. Vasilyev, S.A., Skryabin, N.A., Kashevarova, A.A., Tolmacheva, E.N., Savchenko, R.R., Vasilyeva, O.Y., Lopatkina, M.E., Zarubin, A.A., **Fishman, V.S.**, Belyaeva, E.O., Filippova, M.O., Shorina, A.R., Maslennikov, A.B., Shestovskikh, O.L., Gayner, T.A., Čulić, V., Vulić, R., Nazarenko, L.P., Lebedev, I.N., 2021. Differential DNA Methylation of the IMMP2L Gene in Families with Maternally Inherited 7q31.1 Microdeletions is Associated with Intellectual Disability and Developmental Delay. *Cytogenet. Genome Res.* 161, 105–119. <https://doi.org/10.1159/000514491>. **WOS, Scopus**
 44. Viesná, E., **Fishman, V.**, 2022. FastContext: A tool for identification of adapters and other sequence patterns in next generation sequencing (NGS) data. *Vavilovskii Zhurnal Genet. Sel.* 26, 806–809. <https://doi.org/10.18699/VJGB-22-97>. **WOS, Scopus, РИНЦ**
 45. Крашенинина, О.А., **Фишман, В.С.**, Новопашина, Д.С., Веняминова, А.Г., 2017. 5'-биспирентильные зонды типа “молекулярный маяк” для детекции РНК. *Биоорганическая Химия* 43, 261–272. <https://doi.org/10.7868/S0132342317030095>. **РИНЦ**
 46. Курочкина, Ю.Д., Королев, М.А., Летягина, Е.А., **Фишман, В.С.**, Гридина, М.М., Валеева, Э.С., 2023. Семейный случай редкого аутовоспалительного заболевания, ассоциированного с мутациями в генах *NLRP3* и *TNFRSF1A*, в практике ревматолога. *Бюллетень Сибирской Медицины* 22, 170–175. <https://doi.org/10.20538/1682-0363-2023-2-170-175>. **РИНЦ**
 47. Лопаткина, М.Е., **Фишман, В.С.**, Гридина, М.М., Скрыбин, Н.А., Никитина, Т.В., Кашеварова, А.А., Назаренко, Л.П., Серов, О.Л., Лебедев, И.Н., 2020. Транскриптомное профилирование нейронов, дифференцированных из индуцированных плюрипотентных стволовых клеток пациентов с реципрокными моногенными CNV в субсегменте 3p26.3. *Медицинская Генетика* 19, 10–11. <https://doi.org/10.25557/2073-7998.2020.03.10-11>. **РИНЦ**
 48. Мензоров, А.Г., Лукьянчикова, В.А., Кораблев, А.Н., Серова, И.А., **Фишман, В.С.**, 2016. Практическое руководство по редактированию геномов системой CRISPR/Cas9. *Вавиловский Журнал Генетики И Селекции* 20, 930–944. <https://doi.org/10.18699/VJ16.214>. **РИНЦ**
 49. Миньженкова, М.Е., Маркова, Ж.Г., Юрченко, Д.А., Тарлычева, А.А., Маркова, Т.В., Семенова, Н.А., Муртазина, А.Ф., Кадышев, В.В., Гридина, М.М., Весна, Э., **Фишман, В.С.**, Шилова, Н.В., 2022. Комплексные геномные перестройки в этиологии “хромосомного фенотипа.” *Медицинская Генетика* 21, 44–47. <https://doi.org/10.25557/2073-7998.2022.11.44-47>. **РИНЦ**
 50. Савченко, Р.Р., Васильев, С.А., **Фишман, В.С.**, Сухих, Е.С., Грибова, О.В., Старцева, Ж.А., Мурашкина, А.А., Дорофеева, А.В., Шункова, Д.М., Лебедев, И.Н., 2020. Влияние дифференциальной экспрессии генов ADAMTS1, RBFOX2, THBS1 и WHSC1 на формирование радиационно-индуцированного клеточного ответа. *Медицинская Генетика* 19, 85–87. <https://doi.org/10.25557/2073-7998.2020.09.85-87>. **РИНЦ**

51. Смирнов, А.В., Юнусова, А.М., Муравьева, А.А., Валеев, Э.С., **Фишман, В.С.**, Баттулин, Н.Р., 2021. Создание библиотек баркодированных плазмид с помощью метода клонирования по Гибсону. Письма В Вавиловский Журнал Генетики И Селекции 7, 34–45. <https://doi.org/10.18699/LettersVJ2021-7-05>. **РИНЦ**
52. Таскина, А.К., Муравьёва, А.А., Ельсукова, А.С., **Фишман, В.С.**, 2020. Методы машинного обучения в биологии. Природа 3–17. <https://doi.org/10.7868/S0032874X2009001X>. **РИНЦ**
53. Шахтштейндер, Е.В., Иванощук, Д.Е., Рагино, Ю.И., **Фишман, В.С.**, Полонская, Я.В., Каштанова, Е.В., Чернявский, А.М., Мурашов, И.С., Воевода, М.И., 2021. Анализ дифференциальной экспрессии генов липидного обмена в атеросклеротических бляшках у пациентов с коронарным атеросклерозом. Сибирский Журнал Клинической И Экспериментальной Медицины 36, 156–163. <https://doi.org/10.29001/2073-8552-2021-36-4-156-163>. **РИНЦ**

Личный вклад автора

Автор принимал непосредственное участие в формулировке направлений исследований, планировании и дизайне всех экспериментальных и биоинформатических работ, а также в анализе, систематизации, обобщении, интерпретации и публикации результатов. Основные результаты, изложенные в работе, получены непосредственно автором или членами научного коллектива под его руководством. Ряд молекулярно-генетических экспериментальных данных, необходимых для проведения анализа, был взят из публичных источников или передан автору коллегами: Н.Р. Баттулиным и М.М. Гридиной. В рамках выполнения работ под руководством автора были защищены 9 бакалаврских и 5 магистерских диссертации, одна работа специалиста и 3 диссертации кандидата биологических наук. Работы по теме диссертации были поддержаны фондами РФФИ и РНФ и проводились в сотрудничестве с группами Н.Р. Баттулина (ИЦиГ СО РАН), А.В. Красиковой (СПбГУ), И.Н. Лебедева (Томский НИИ Медицинской генетики), М.Ю. Скоблова и Н.В. Шиловой (МГНЦ им. Бочкова), И.В. Шарахова (Политехнический институт Вирджинии, США) и О.Л. Кардымон (Институт искусственного интеллекта).

2. ОСНОВНОЕ СОДЕРЖАНИЕ ДОКЛАДА

2.1. Сравнительный анализ укладки хроматина в разных типах клеток и у разных видов позвоночных животных

Для различных типов клеток млекопитающих нами ^{4,10} и нашими коллегами ³ был показан высокий уровень сходства трехмерной укладки хроматина. Для того чтобы понять, насколько консервативны принципы укладки ДНК за пределами этой таксономической группы, мы решили исследовать архитектуру ядра в клетках птиц. Мы выбрали в качестве объекта исследований домашнюю курицу, *Gallus gallus* - этот организм является модельным, что позволяет использовать широкий спектр накопленных другими исследователями геномных данных и, что важно, геном курицы, по сравнению с геномами других птиц, собран наиболее полно.

Мы выбрали для исследования два типа клеток курицы: эмбриональные фибробласты и эритроциты взрослых животных (рис. 1). Эмбриональные фибробласты были выбраны в качестве репрезентативного соматического клеточного типа, поскольку данные об укладке генома эмбриональных фибробластов разных животных доступны в литературе ⁴. Зрелые эритроциты представляют интерес, поскольку у птиц, в отличие от млекопитающих, эти клетки содержат ядро - однако ДНК в ядрах эритроцитов высококомпактизована и связана большим количеством гистона H5 ¹¹. Мы предположили, что эти особенности строения ядра эритроцитов могут отразиться на укладке генома. В качестве материала для сравнительного анализа, мы использовали полученные ранее данные нашей группой методом Hi-C данные о частотах контактов хроматина в мышинных эмбриональных фибробластах, а также доступные для других млекопитающих публичные данные.

Анализ Hi-C-данных, полученных Н.Р. Баттулиным с коллегами, показал существенное различие в структуре контактов фибробластов и эритроцитов курицы. На Hi-C-картах фибробластов были обнаружены типичные паттерны, свидетельствующие о формировании ТАДов по механизму экструзии петли. Визуально эти паттерны, представленные на рисунке 2, А, представляют собой красные квадраты с яркой точкой в вершине. Согласно модели экструзии петли, такой паттерн

объясняется за счет взаимодействия белков когезинового комплекса и CTCF⁵. Поскольку белок CTCF способен блокировать продвижение белкового комплекса когезин, который, в свою очередь, осуществляет протягивание (экструзию) петли ДНК, яркая точка на карте Hi-C соответствует петле, сформированной между удаленными друг от друга в геноме участкам связывания белков CTCF. При этом сайт связывания белка CTCF является ассиметричным, и его ориентация играет важную роль для взаимодействия с когезином: CTCF связанный с сайтом в прямой ориентации блокирует продвижение когезина в “обратном” направлении (3’-5’), а CTCF связанный с сайтом в обратной ориентации блокирует транслокацию комплекса в направлении 5’-3’. Таким образом, на участке ДНК между двумя сайтами связывания CTCF в конвергентной ориентации может формироваться множество когезин-опосредованных петель, а комплексы когезина, заблокированные на сайтах связывания CTCF, образуют стабильную петлю.

Механизм формирования ТАДов за счет взаимодействия белков CTCF и когезин был предложен на основе анализа данных об укладке хроматина в клетках млекопитающих, при этом формируются ли ТАДы у птиц и каков механизм их формирования было не известно. Для того чтобы проверить, соответствуют ли структуры, выявленные на Hi-C-картах эмбриональных фибробластов курицы, когезин-опосредованным ТАДам, мы сравнили расположение границ ТАДов и сайтов связывания белка CTCF. Анализ показал обогащение границ сайтами связывания белка CTCF, причем сайты связывания в границах оказались расположенными преимущественно в конвергентной ориентации (рис. 2, А, Б). Эти результаты позволяют утверждать, что формирование ТАДов в фибробластах птиц и млекопитающих происходит согласно одному эволюционно-консервативному механизму, связанному с экструзией петель ДНК. С этим утверждением хорошо согласуются результаты анализа распределения участков активной транскрипции относительно границ ТАДов. Хотя, по сравнению с активностью когезина, процесс транскрипции играет существенно меньшее значение в формировании ТАДов у млекопитающих, их границы у мыши и человека обогащены активным хроматином и

характеризуются высокой плотностью генов, а изменение транскрипционной активности в некоторых случаях может влиять на геномное расположение границ ТАДов.

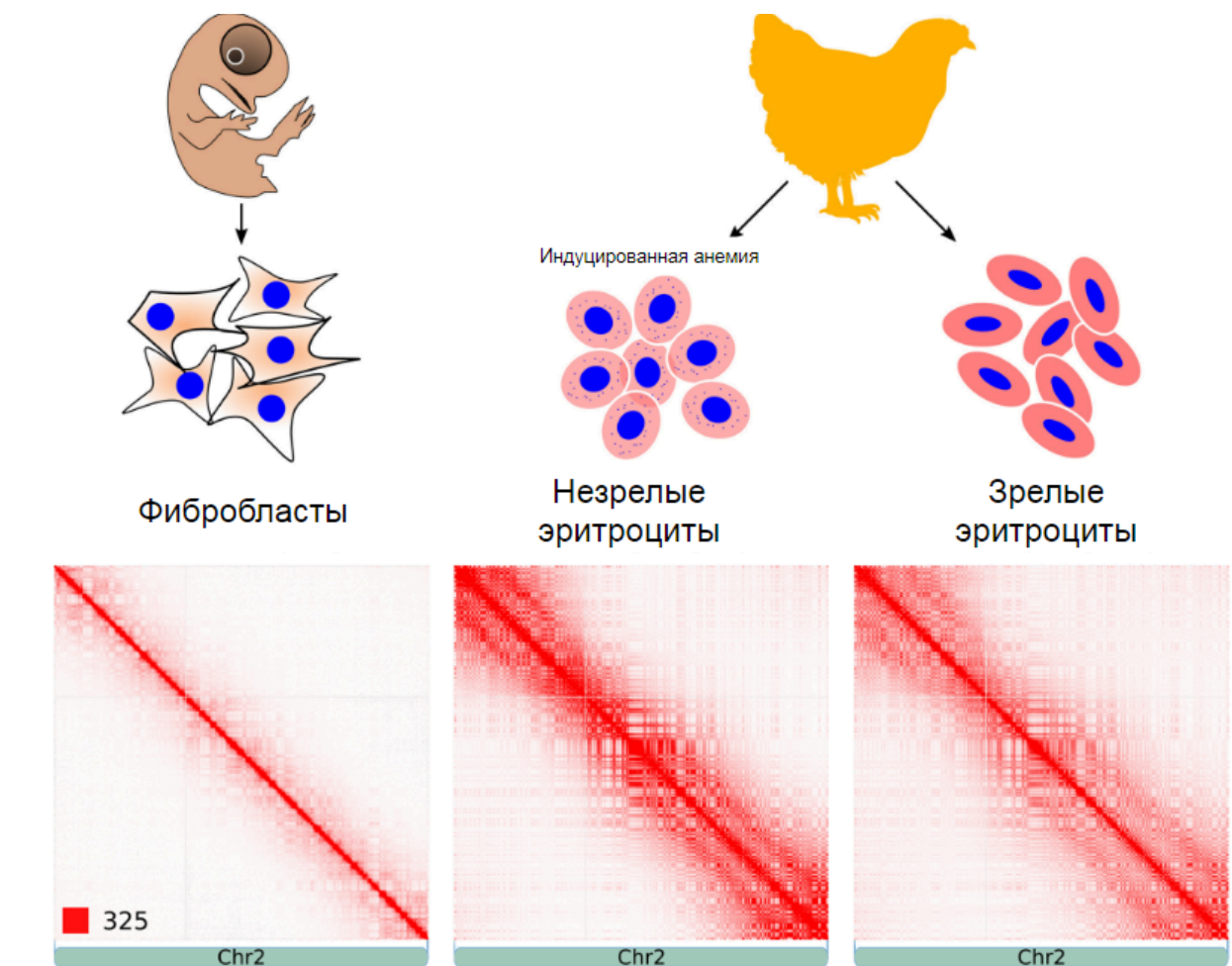


Рисунок 1. Исследование 3D-организации хроматина в клетках курицы ¹². Вверху схематично показаны исследованные в работе типы клеток, внизу - репрезентативные примеры карты контактов хроматина для фибробластов, зрелых и незрелых эритроцитов курицы. Здесь и далее на картах цветом обозначена частота трехмерных контактов между соответствующими локусами ДНК, более насыщенный цвет соответствует большей частоте контактов.

В фибробластах курицы мы наблюдали распределение, сходное с описанным ранее для фибробластов и других соматических клеток млекопитающих: границы ТАДов оказались обогащены генами, а средний уровень транскрипции вблизи границы ТАДов выше, чем внутри ТАДа (рис. 2, Г). Наконец, средний размер ТАДов

фибробластов курицы (200-300 т.п.о) оказался сопоставимым с размером ТАДов, выделенных в геномах млекопитающих. Эти данные также свидетельствуют о сходстве ТАДов, выявленных в геномах эмбриональных фибробластов курицы и описанных ранее для млекопитающих.

Сходство укладки хроматина эмбриональных фибробластов курицы и клеток млекопитающих не ограничивается свойствами границ ТАДов. В обоих случаях наблюдается паттерн “шахматной доски”, который свидетельствует о компартментализации хроматина. Также мы наблюдали сходную зависимость частоты контактов хроматина от расстояния между локусами. В эмбриональных фибробластах курицы для геномных расстояний в диапазоне 100 000 - 5 000 000 п.о. эта зависимость характеризуется степенной функцией с показателем степени -1.1 (рис. 2, Д), что соответствует фрактальной укладке генома. Показатель степени близкий к -1.1 наблюдался ранее и при анализе Hi-C-данных клеток млекопитающих (например, в мышинных сперматозоидах и фибробластах показатель степени составляет -1.07 и -1.27, соответственно ⁴). Все это говорит о сходстве основных принципов, лежащих в основе укладки генома эмбриональных фибробластов курицы и различных ранее исследованных соматических клеток млекопитающих. Для более детального анализа контактов хроматина мы разработали инструмент C-InterSecure ¹⁰, который позволил сравнить особенности укладки отдельных локусов ДНК у разных видов и, в том числе, сравнивать расположение границ отдельных ТАДов в геномах разных видов. Сравнивая индивидуальные границы ТАДов эмбриональных фибробластов курицы и мыши в синтенных областях генома, мы обнаружили что значительная часть ТАДов сохраняет своё расположение (рис. 2, Е) несмотря на значительное время дивергенции этих видов (более 300 млн лет). Согласуется с этим и наше наблюдение о том, что хромосомные перестройки, которые закрепились в геномах птиц, преимущественно проходят в границах ТАДов - а не внутри ТАДов, т.е. ТАДы в ходе эволюции предпочтительно перемещаются целиком, не разрываясь ¹². Наконец, следует отметить что разработанный нами метод межвидового сравнения Hi-C-карт показал сходство между видами и на уровне индивидуальных контактов хроматина внутри ТАДов ¹⁰.

Основываясь на этих данных, можно предположить, что многие ТАДы играют важную функциональную роль, и их границы поддерживаются отбором в ходе эволюции.

Проведя анализ Hi-C-данных эмбриональных фибробластов, мы решили сравнить архитектуру хроматина этих клеток с эритроцитами, обладающими гораздо более компактным ядром. Сравнительный анализ выявил ряд особенностей в укладке генома у эритроидных клеток. Несмотря на то, что анализ Hi-C-карт эритроцитов позволяет выделить локальные области генома, инсулированные от окружающих участков, границы этих областей не формируют петлю, характерную для образованных по экструзионному механизму ТАДов (рис 2, А). Кроме того, расположение участков инсуляции в геномах эритроцитов не соответствует расположению границ ТАДов, выделенных в эмбриональных фибробластах, не показывает обогащение сайтами связывания белка CTCF, участками активной транскрипции и метками открытого хроматина. Наконец, мы не обнаружили следов формирования когезин-опосредованных ТАДов при анализе кривой зависимости частот контактов хроматина от расстояния. Процесс экструзии приводит к увеличению контактов хроматина на геномных расстояниях, сопоставимых с характерным размером экструзионной петли (до 1 миллиона пар оснований) ⁵. На кривой, соответствующей данным Hi-C эмбриональных фибробластов курицы в этом диапазоне наблюдается характерный прогиб (рис. 2, Д), в то время как у эритроцитов кривая в этом диапазоне гладкая. Таким образом, можно утверждать, что образование этих структур хроматина в эритроцитах курицы происходит по механизму, отличному от когезин-опосредованных ТАДов фибробластов.

Каков механизм образования обнаруженных в геноме эритроцитов структур хроматина? Для ответа на этот вопрос, мы провели анализ компартментализации генома. Hi-C-данные позволяют разделить геном на два пространственно-сегрегированных компартмента, А и В, которые по своим характеристикам соответствуют активному эу- и неактивному гетерохроматину. Мы обнаружили, что выявленные в геномах эритроцитов области инсуляции совпадают с границами А- и В-компартментов (рис. 2, Ж, З). Интересно, что в эритроцитах сила

компарментализации, т.е. превышение частот гомотипных (внутри компартмента) контактов хроматина над гетеротипными (между компартментами), оказывается существенно выше, чем в фибробластах. Усиленная компарментализация хроматина хорошо согласуется с цитологическими данными о большей сегрегации эу- и гетерохроматина в этих клетках по сравнению с другими соматическими клетками.

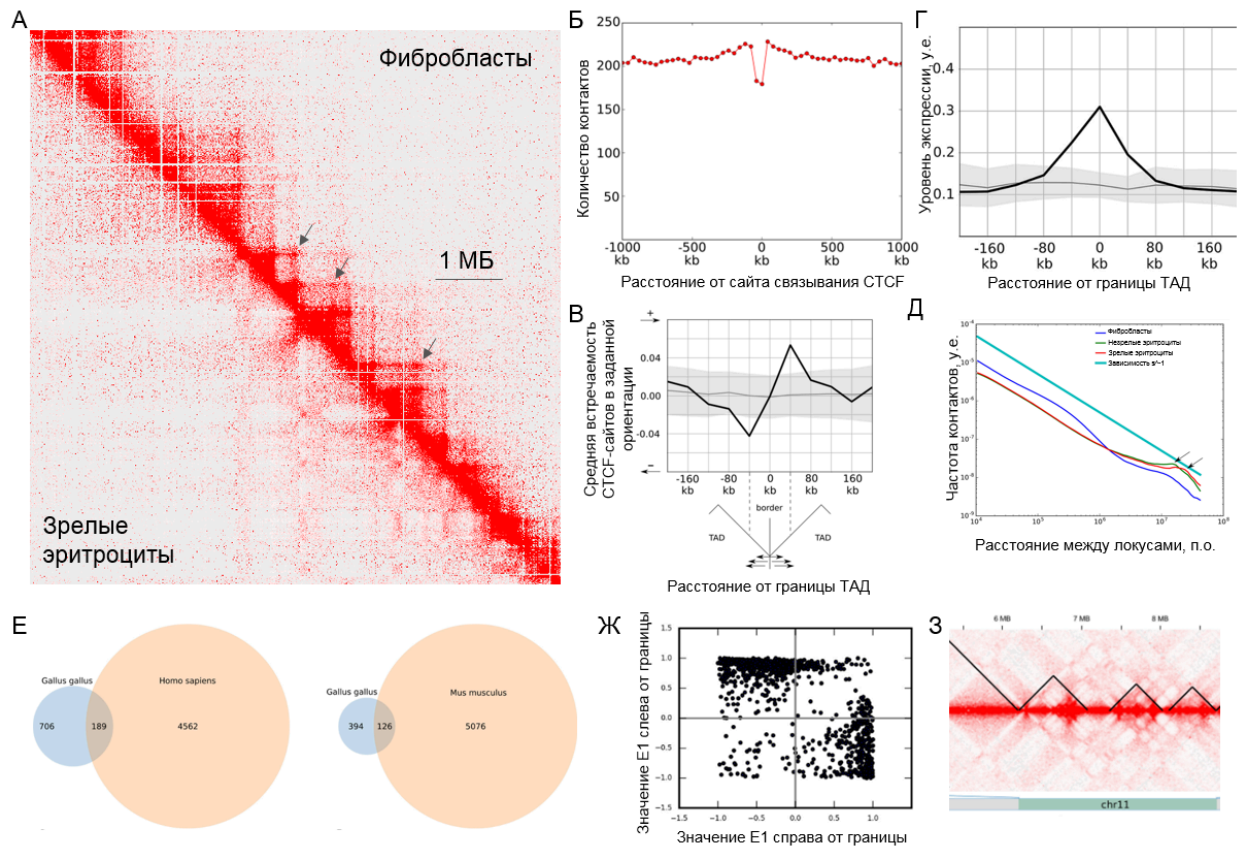


Рисунок 2. Архитектура хроматина в фибробластах и эритроцитах курицы домашней. А. Репрезентативный фрагмент внутрихромосомной карты контактов хроматина для фибробластов (выше диагонали) и эритроцитов (ниже диагонали). Стрелки показывают участки карты, соответствующие CTCF-опосредованным петлям в границах ТАДов. Б и В. Зависимость частоты трехмерных контактов от расположения (Б) и ориентации (В) сайтов связывания белка CTCF. Зависимость построена на основе данных, полученных для фибробластов курицы. Г. Зависимость уровня транскрипции и расположения границ ТАДов в фибробластах курицы. Д. Зависимость частоты трехмерных контактов от геномного расстояния для фибробластов курицы. Е. Сравнение геномной локализации границ ТАДов в фибробластах курицы, мыши и человека. Ж. Значения первого собственного вектора (E1) в границах ТАДов. Значения E1 отражают компарментализацию генома - положительные значения соответствуют А-компарменту, отрицательные -

В-компарменту. 3. Фрагмент внутривхромосомной карты контактов хроматина зрелых эритроцитов курицы.

Таким образом, можно заключить что в геномах эритроцитов отсутствуют когезин-опосредованные ТАДы или существенно снижено их количество, а основным фактором укладки хроматина является сегрегация компарментов (эу- и гетерохроматина), которая более выражена в этих клетках по сравнению с эмбриональными фибробластами. Интересно, что сходный паттерн контактов был обнаружен нами и в ретикулоцитах (незрелых эритроцитах) курицы, которых от стадии зрелого эритроцита отделяет одно клеточное деление. В то же время, в других соматических типах клеток курицы, которые мы изучили позднее (например - в эритробластоидных клетках линии HD3), наблюдалось формирование когезин-опосредованных ТАДов ¹³. На основе полученных данных можно сделать вывод о том, что отсутствие когезин-опосредованных ТАДов является специфической особенностью эритроцитарной линии клеток курицы, возникающей в определенный момент эритропоэза, но не на терминальной стадии дифференцировки. Интересно, что согласно полученным Н.Р. Баттулиным экспериментальным данным о контактах хроматина в терминальных эритроблестах мыши и нашему анализу, проведенному на публичных данных проекта DNAAzoо, описанные выше особенности организации хроматина эритроцитов являются эволюционно-консервативными и наблюдаются у всех исследованных позвоночных животных ¹². Вероятно, эти особенности организации хроматина эритроцитов важны для экстремальной компактизации ядра, которая происходит в ходе дифференцировки клеток.

2.2. Сравнительный анализ укладки хроматина у комаров рода *Anopheles*

Приведенные выше результаты убедительно показывают, что укладка хроматина позвоночных животных, за исключением клеток эритроцитарного ряда, демонстрирует значительный уровень сходства, причем важнейшим фактором формирования ТАДов является процесс экструзии петли. В то же время в серии работ С.В. Разина и С.В. Ульянова на клеточных линиях ¹⁴ и в последующих исследованиях *in vivo* ¹⁵ было

показано, что белок CTCF не играет существенной роли в образовании ТАДов дрозофилы. Таким образом, сложилось мнение о том, что механизмы формирования ТАДов дрозофилы отличаются от описанных для позвоночных животных, однако за пределами группы дрозофилид этот вопрос оказался мало изученным. Кроме того, у позвоночных животных наблюдалось сходство расположения границ ТАДов в синтенных районах, в то время как для беспозвоночных животных вопрос межвидовых различий границ ТАДов оставался дискуссионным: был опубликован ряд работ, результаты которых противоречили друг другу¹⁶⁻¹⁸. Эти результаты мотивировали нас исследовать укладку хроматина у представителей рода *Anopheles*. Эта таксономическая группа показалась нам интересна, поскольку для комаров-анофелесов укладка хроматина ранее не исследовалась методом Hi-C, а также потому что многие представители этой группы являются переносчиками возбудителя малярии - опасного заболевания, от которого ежегодно в мире умирают сотни тысяч людей. Потому новые данные по 3D-организации генома комаров-анофелесов могут быть потенциально полезными в медицинских исследованиях.

В качестве экспериментального материала мы собрали личинки 4ой стадии развития для 5 видов комаров, *An. coluzzii*, *An. merus*, *An. stephensi*, *An. atroparvus* и *An. albimanus* (рис. 3, А-Г). Мы также провели эксперимент Hi-C на линии клеток MSQ43, полученной из имагинальных дисков комаров *An. stephensi*, и на материале взрослых комаров *An. merus*. Следует отметить что выбранные виды имеют широкий разброс времени дивергенции (рис. 3, А), от ~0.5 млн лет (*An. coluzzii* vs *An. merus*) до более чем 100 млн лет (*An. coluzzii* vs *An. albimanus*).

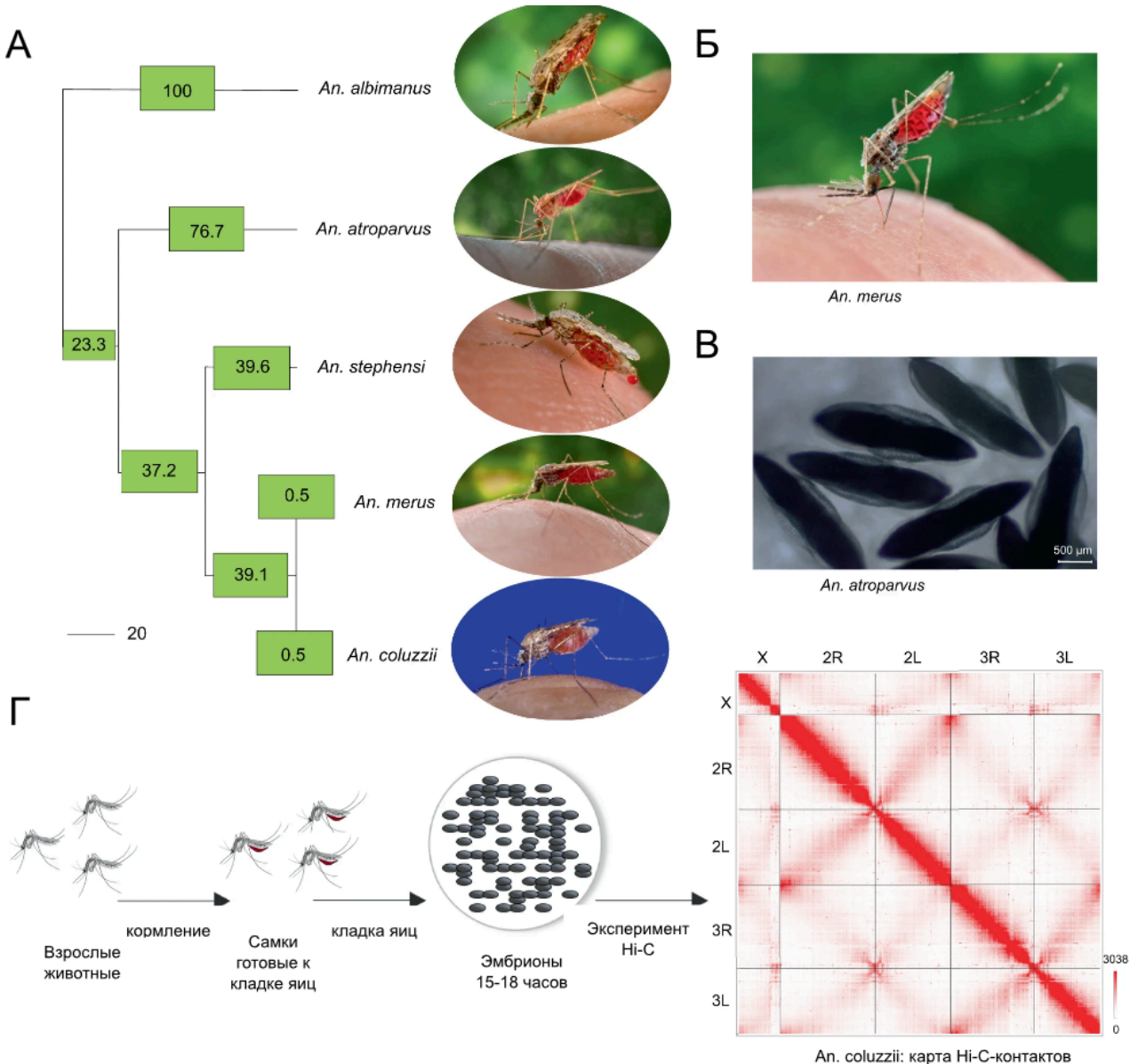


Рисунок 3. Дизайн эксперимента для исследования архитектуры хроматина у комаров рода *Anopheles*. А. Филогенетическое дерево пяти видов комаров, взятых в анализ. Б и В. Фотографии, демонстрирующие морфологию взрослых комаров (Б) и их яиц (В). Г. Схема эксперимента: самки взрослых животных после кормления синхронно откладывают яйца, из которых спустя 15-18 часов формируются эмбрионы. На этой стадии материал фиксируется и используется для проведения Hi-C-эксперимента, а также экспериментов по анализу транскриптома и связывания белков хроматина.

Хотя исследования комаров-анофелесов привлекают большое внимание, геномы многих видов имели низкое качество сборки на момент начала исследования. Для

улучшения сборки генома мы использовали Hi-C-данные и результаты секвенирования третьего поколения (Oxford Nanopore и PacBio). Использование данных Hi-C для сборки геномов основано на степенной зависимости между частотой контактов хроматина и геномным расстоянием. Благодаря этой зависимости, можно определять расположение и ориентацию геномных скаффолдов - чем больше контактов наблюдается между скаффолдами, тем больше вероятность, что они граничат друг с другом в геноме ⁸. Аналогичным образом можно определить ориентацию скаффолдов друг относительно друга. Этот подход работает наиболее эффективно, если скаффолды достаточно большие - в этом случае число контактов скаффолда достаточно, чтобы уверенно определить его геномных соседей. Поэтому, наиболее высокая континуальность оказалась у тех сборок генома, для которых были доступны не только Hi-C-данные, но и данные секвенирования 3-го поколения ⁷. Тем не менее следует отметить, что качественная сборка центромерного и теломерного гетерохроматина невозможна даже с использованием всех доступных нам данных, поэтому в дальнейшем анализе мы удаляли эти районы из рассмотрения.

Полученные нами сборки генома оказались существенно лучше, чем ранее доступные (на основе критерия N50, отражающего континуальность сборки), причем для трех видов из пяти мы впервые получили сборки генома хромосомного уровня ⁷. Эти результаты согласуются с ранее опубликованными данными о том, что технология Hi-C является эффективной не только для исследования контактов хроматина, но и для сборки целевых геномов организмов. Более того, на основе Hi-C-данных мы смогли обнаружить в геномах комаров ряд полиморфных инверсий (рис. 4), часть из которых была описана ранее на основе анализа цитологических данных, а часть обнаружена нами впервые. Этот результат представляет интерес в контексте того, что инверсии играют важную роль в эволюции комаров рода *Anopheles*. Кроме того, этот результат показывает эффективность применения технологии Hi-C как инструмента для поиска структурных перестроек, в том числе - сбалансированных. Этот вопрос будет более подробно освещен в разделе 2.4.

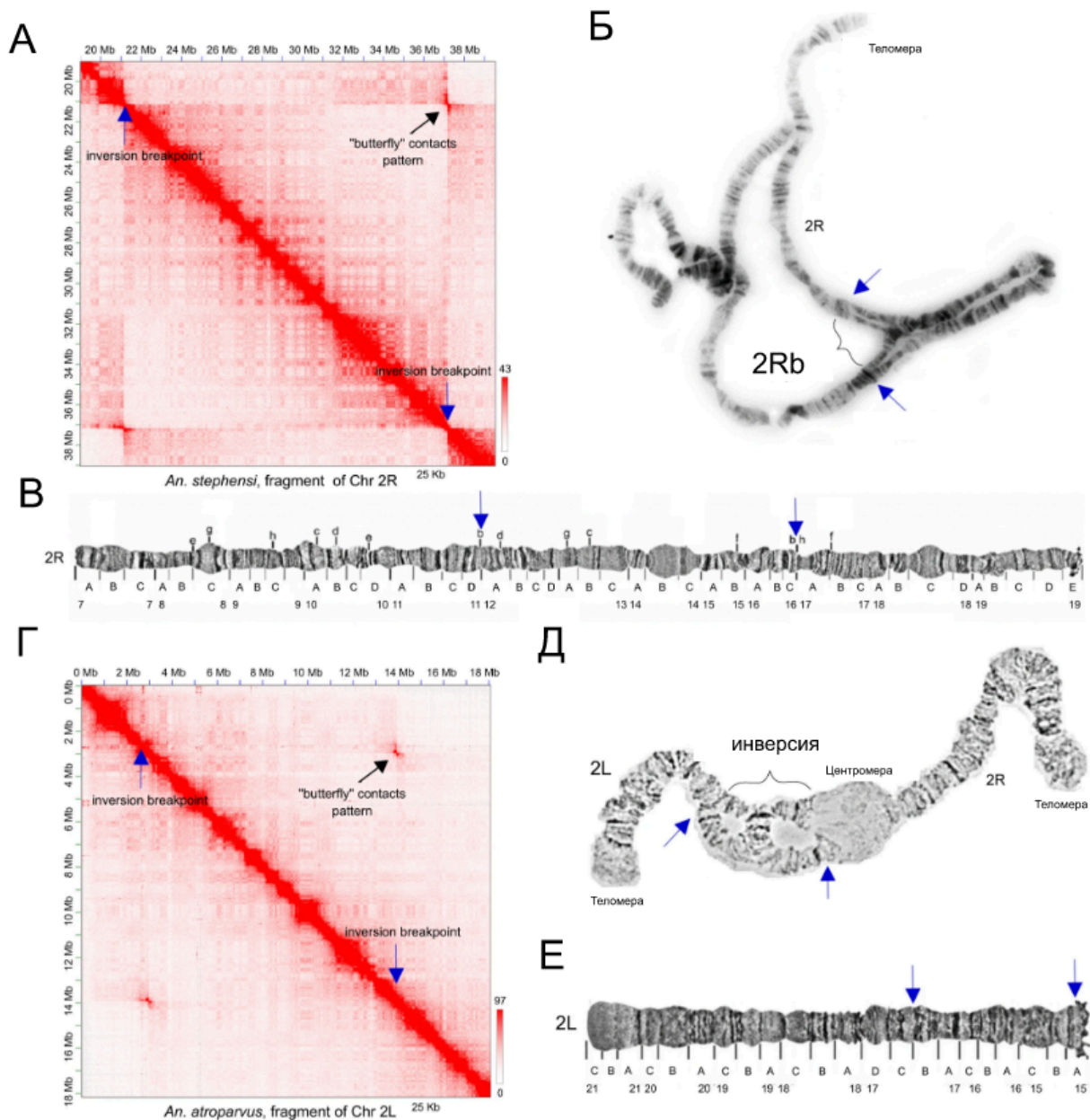


Рисунок 4. Идентификация инверсий в геномах комаров рода *Anopheles* на основе Hi-C-данных. Панели А-В демонстрируют инверсию 2Rb, описанную ранее на основе цитологических данных. Панели Г-Е демонстрируют инверсию 2L01, также описанную в литературе ранее на основе литературных данных. Показаны примеры фрагментов Hi-C-карт, демонстрирующих "инверсионный паттерн" (А, Г), соответствующие инверсиям микрофотографии политенных хромосом (Б, Д) и цитогенетических карт (В, Е), опубликованных ранее.

Основываясь на полученных геномных сборках, мы построили карты трехмерных контактов и провели их анализ (рис. 5, А-Д). На уровне межхромосомных

взаимодействий, мы обнаружили признаки организации хромосом “по Раблю” - сближение центромерных и теломерных участков разных хромосом друг с другом (рис. 5, А, Д). Кроме того, мы наблюдали увеличение частот контактов хроматина для локусов, расположенных периферически на одинаковом расстоянии от центромеры - что также является следствием организации хромосомы “по Раблю” (рис. 6). Полученные данные хорошо согласуются с опубликованным позднее исследованием ¹⁹ группы Либерман-Айдена, в котором была показана связь между потерей субъединиц комплекса конденсина II и организацией хромосом “по Раблю”: в геномах анофелесов отсутствует субъединица CapG2 конденсина II ²⁰. Интересно, что у взрослых комаров признаки организации хромосом “по Раблю” были более выражены, чем у личинок (рис. 6) - это говорит о том что помимо активности конденсина II другие клеточно-специфические факторы могут влиять на форму хромосомных территорий.

Визуализируя карты контактов хроматина внутри отдельных хромосом, мы обнаружили паттерн “шахматной доски”, сходный с ранее наблюдавшимся в клетках млекопитающих и птиц и свидетельствующий о сегрегации компартментов хроматина (рис. 5, Б). Стандартные инструменты, разработанные для анализа хроматина позвоночных, не позволили в автоматическом режиме получить разметку компартментов, соответствующую наблюдаемому паттерну. Мы показали, что это связано с влиянием на профиль контактов конфигурации хромосом по Раблю и разработали собственный инструмент “ABCE”, которые позволили решить эту проблему.

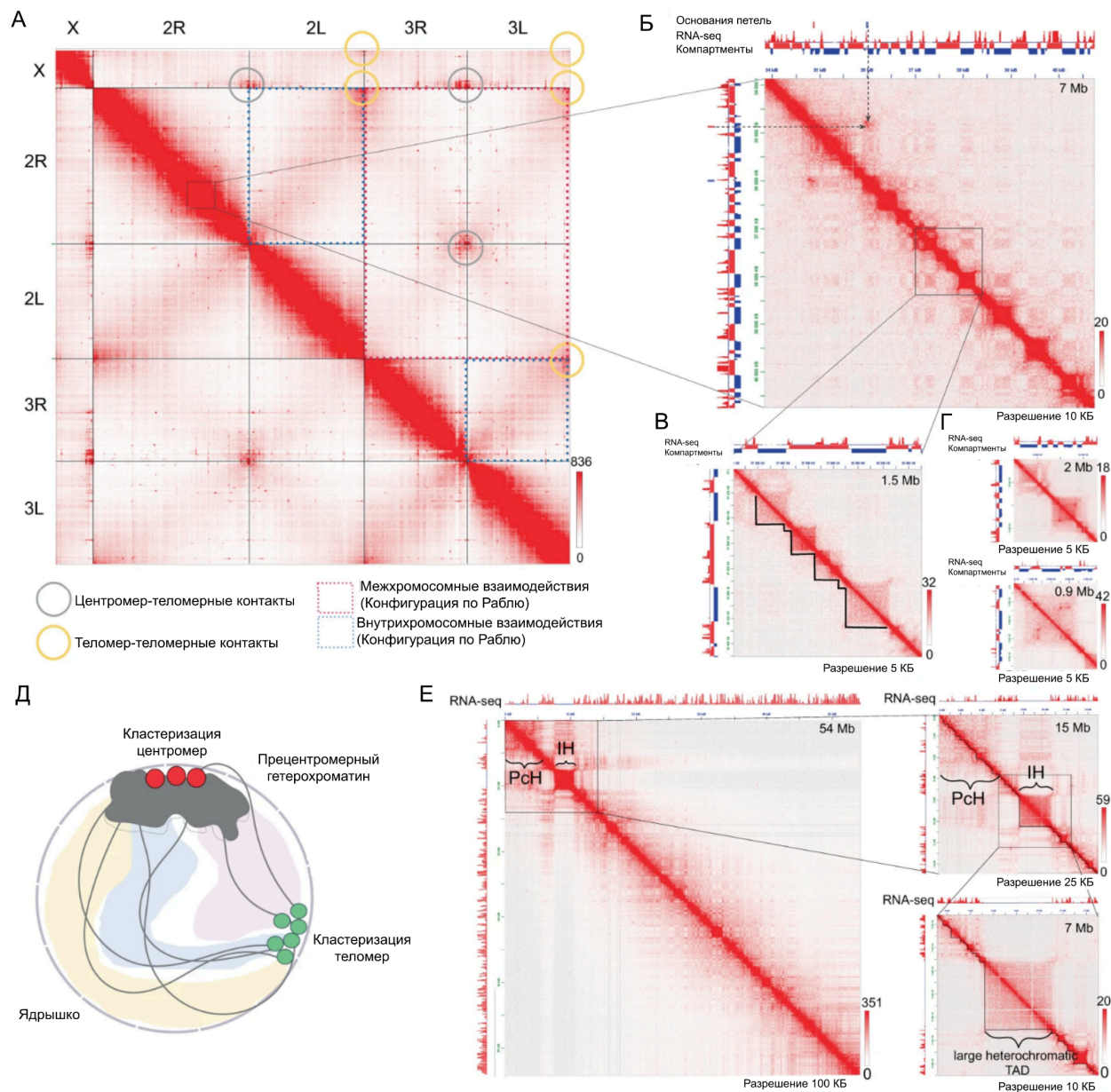


Рисунок 5. Анализ паттернов трехмерных контактов хроматина у комаров рода *Anopheles*. А. Общий вид Hi-C-карты. Для межхромосомных контактов характерно обогащение взаимодействиями центромерных и теломерных участков, что свидетельствует в пользу организации хромосом по Раблю. Б. Фрагмент внутрихромосомной карты контактов. Можно наблюдать признаки компарментализации хроматина (паттерн шахматной доски), причем выделенные на основе данных о контактах хроматина компартменты показывают существенное различие по уровню транскрипции (треки над картой контактов). Также наблюдаются паттерны, свидетельствующие о наличии длинных хроматиновых петель в геномах комаров рода *Anopheles*. В и Г. Фрагмент внутрихромосомной карты контактов

хроматина на большом разрешении позволяет визуализировать паттерны около диагонали, свидетельствующие о разделении генома комаров на ТАДы. Е. Фрагменты прицентромерного (РН) и интерколярного (ИН) гетерохроматина выделяются на Hi-C картах как отдельные ТАДы. Д. Схема организации хроматина в ядрах личинок комаров рода *Anopheles*, построенная на основе полученных Hi-C данных.

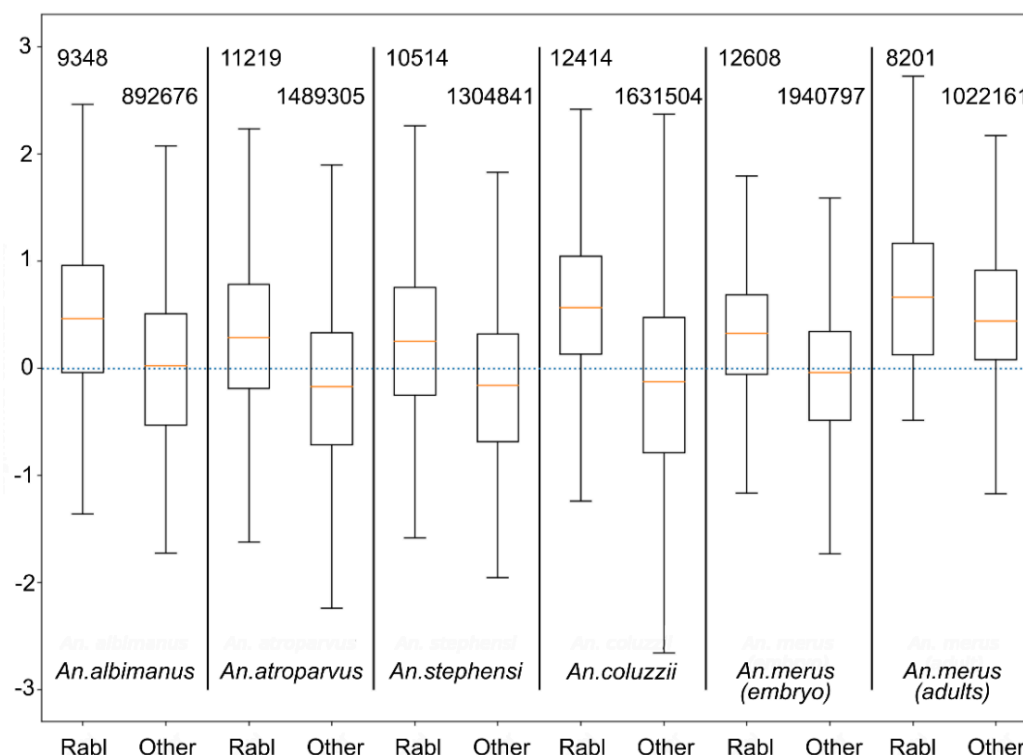


Рисунок 6. Анализ Hi-C-контактов согласуется с укладкой хромосом по Раблю у комаров рода *Anopheles*. Данные о частотах контактов хроматина представлены в виде “ящичков с усами”. Контакты разделены на две группы: в первую группу (Rabl) входят пары локусов, расположенные на разных плечах одной хромосомы на одинаковом расстоянии от центромеры (разница в расстояниях до центромеры <100 т.п.о.). Во вторую группу отнесены контакты между парами локусов, расположенные на разных плечах одной хромосомы на разных расстояниях до центромеры (разница в расстояниях до центромеры >300 т.п.о.). В качестве величины частоты контакта используется логарифм отношения наблюдаемой для пары локусов частоты контакта к средней частоте контактов локусов, расположенных на данном геномном расстоянии. Число над каждым графиком показывает размер выборки. Для комаров рода *An. merus* был проведен анализ двух типов образцов - эмбриональных (embryo) и взрослых (adults). Для остальных видов анализировали только эмбрионы. Парное сравнение выборок контактов для образцов одного вида во всех случаях выявило статистически-значимые различия (q-value < 0.01, U-критерий Манна — Уитни)

Чтобы понять, чем обусловлено формирование выявленных геномных компартментов, мы провели эксперимент по РНК-секвенированию и, используя ранее отработанные методы сборки и анализа транскриптомов ²¹⁻²³, провели анализ транскрипции. Оказалось, что, как и для дрозофилы и позвоночных животных, разделение генома комаров-анофелесов на компартменты связано с различием в транскрипционной активности локусов: А-компартмент имеет большую плотность генов и преимущественно обогащен активно-экспрессирующимися генами, в то время как В-компартмент показывает заметно меньший средний уровень экспрессии (рис. 5, Б). В то же время, различия в уровне транскрипции не способны полностью объяснить дисперсию контактов хроматина на уровне компартментов - можно предположить, что, как и в клетках дрозофилы, формирование компартментов в геномах комаров-анофелесов связано не только с уровнем транскрипции, но и с распределением различных архитектурных белков хроматина. Для подтверждения этой гипотезы мы провели серию ChIP-seq-экспериментов на клетках MSQ43 и установили распределения эпигенетических меток активного хроматина (H3K4me3, H3K27ac), гетерохроматина (H3K9me3) и маркеров белков группы поликомб (H2A119ub, H3K27me3). Анализ полученных данных показал, что на границе компартментов наблюдается смена эпигенетического состояния хроматина, при этом внутри А-компартмента характерно повышенное количество меток активного хроматина, тогда как внутри В-компартмента наблюдается больше меток белков группы поликомб и сигналов конститутивного гетерохроматина.

Интересно, что в клетках комаров, как и в эритроцитах позвоночных животных, наблюдалось хорошее совпадение между границами ТАДов и компартментов (рис. 7, А,Б,Е). При анализе Hi-C-карт мы не обнаружили признаков формирования когезин-опосредованных ТАДов (см. рисунок 7, Д, а также подробный анализ, приведенный в ⁷). Медианный размер ТАДа составлял около 130 тысяч п.о., при этом ТАДы, соответствующие А-компартменту оказались в среднем мельче, чем ТАДы, соответствующие В-компартменту (рис. 7, В, Г). Мы выделили несколько особенно крупных ТАДов, размером более 1 миллиона п.о., которые при сопоставлении с

цитологической картой соответствовали блокам интеркалярного гетерохроматина (рис. 7, Д). Таким образом, можно сделать вывод о том, что границы ТАДов у комаров-анофелесов не задаются процессом экструзии, а связаны с компартиментализацией эу- и гетерохроматина.

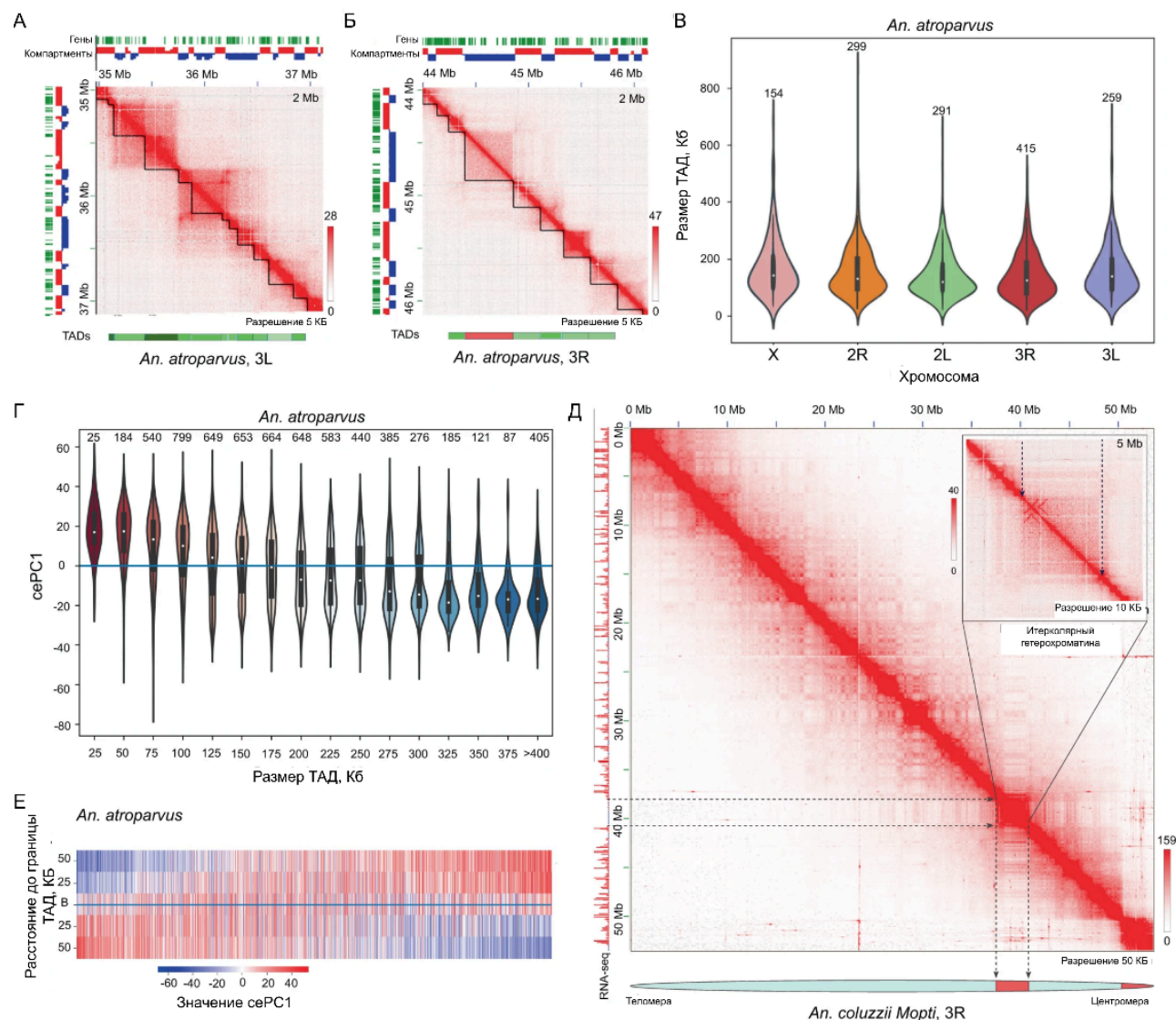


Рисунок 7. Анализ ТАДов в геномах комаров рода *Anopheles*. А и Б. Фрагменты карт внутрихромосомных контактов хроматина у комаров рода *Anopheles*. На этих и других панелях значения sePC1 (трек “компарменты”) отражают компартиментализацию хроматина - положительные значения соответствуют А-компарменту, отрицательные - В-компарменту. В - распределение длин ТАДов показано на примере данных, полученных для *An. albimanus*. Число над скрипичными графиками показывает число ТАДов для каждой из хромосом. Г. Распределение размеров ТАДов стратифицированное по среднему значению sePC1 локусов внутри

ТАД (отражает принадлежность ТАДа к А- или В-компарменту). Д. Фрагменты карт внутрихромосомных контактов хроматина у комаров рода *Anopheles* для участка хромосомы, содержащего блок интерколярного гетерохроматина (выделяется в отдельный ТАД). Е. Сравнение значения *sePC1* (отражает принадлежность локуса к А- или В-компарменту) слева и справа от границы ТАД. Каждая вертикальная полоса соответствует одной границе ТАДа (положение границы отмечено горизонтальной линией). Видно, что граница ТАДа в подавляющем большинстве случаев разделяет участки хроматина, относящиеся к разным компарментам.

Анализируя Hi-C-карты комаров, мы сделали интересное наблюдение: оказалось, что у всех исследованных видов на X-хромосоме формируется длинная петля, основания которой удалены на расстояние до 30 миллионов п.о. (рис. 8, А-Ж). Более того, анализ синтении показал, что петли четырех из пяти видов формируются между синтенными участками, хотя геномное расстояние, которое их разделяет, может существенно варьировать (в случае *An. albimanus*, из-за очень большого времени дивергенции с другими видами, нам не удалось получить однозначный ответ на вопрос о синтении в области оснований петель). В литературе описано два механизма образования столь длинных петель в геномах насекомых: взаимодействия, опосредованные белками группы поликомб и формирование транскрипционных фабрик с участием высоко-транскрибируемых генов. Однако анализ транскриптомных данных и ChIP-seq-профилей показал, что лежащие в основаниях петель участки по уровню транскрипции и обогащению метками белков группы поликомб находятся на уровне, сопоставимом с участками генома, которые таких петель не формируют (рис. 8, А-Ж). Это позволяет предположить, что выявленные длинные петлевые взаимодействия сформированы новым, ранее не известным механизмом.

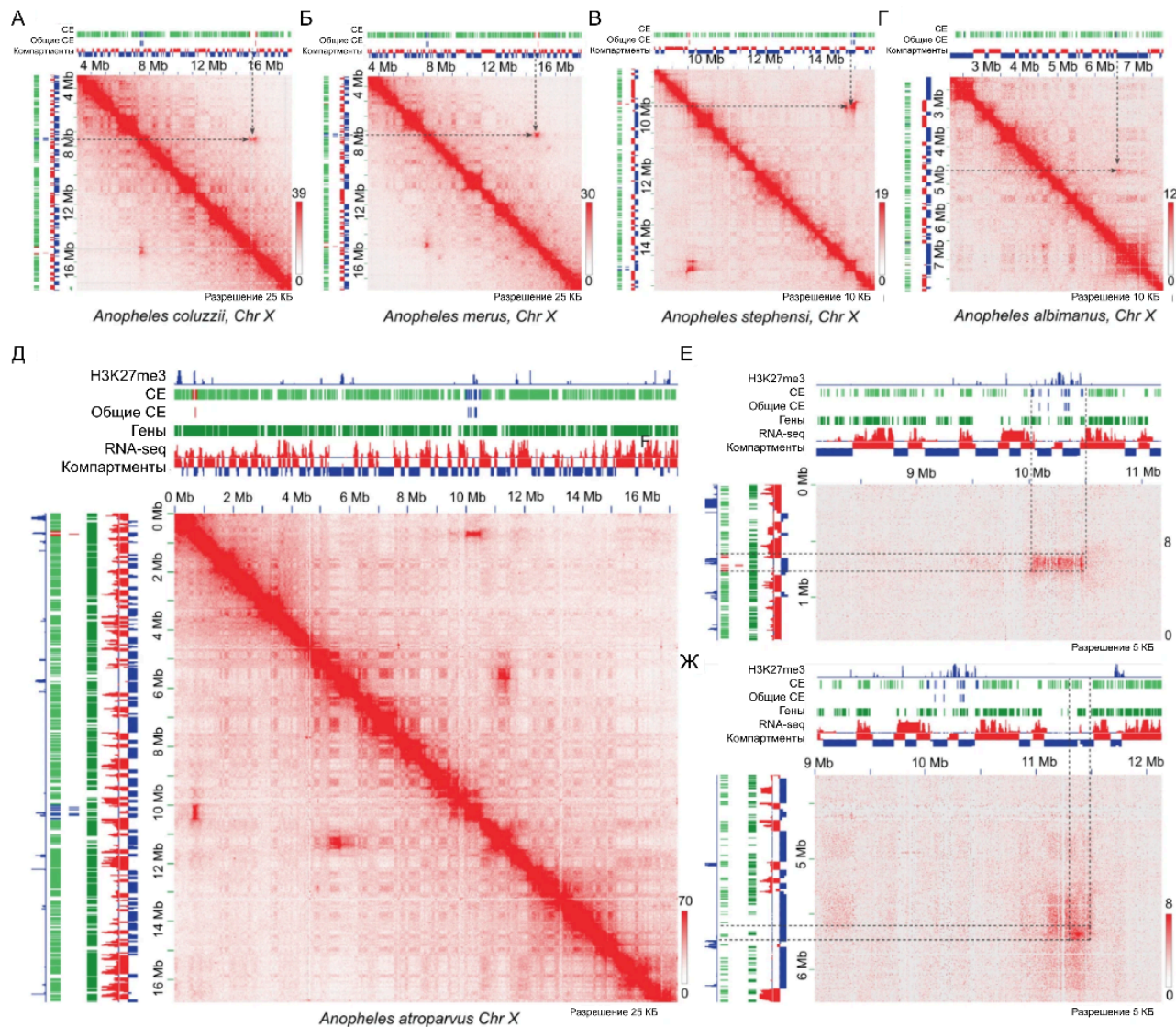


Рисунок 8. Анализ петлевых взаимодействий в геномах комаров рода *Anopheles*. А-Ж. Показаны фрагменты карт контактов хроматина хромосомы X, свидетельствующие о формировании взаимодействий между удаленными на большое расстояние участками хроматина. Рядом с картами (сверху и сбоку) показано распределение различных геномных элементов: эволюционно-консервативных последовательностей (CE), консервативных последовательностей общих для оснований петель у всех исследованных видов комаров кроме *An. albimanus* (общие CE), профиль транскрипционной активности (RNA-seq), компартиментализации (“Компартменты”, значения больше нуля соответствуют А-компартменту, меньше нуля - В-компартменту).

Детальное описание контактов хроматина у пяти видов комаров рода *Anopheles* с использованием разработанного нами инструмента C-InterStructure позволило нам

исследовать как меняется укладка генома в ходе эволюции. Центральный вопрос, который мы поставили, заключался в том, являются ли границы ТАДов консервативными в ходе эволюции. Для ответа на этот вопрос, мы выделили синтенные области в геномах комаров и, используя разработанные нами методы межвидового сравнения контактов хроматина ¹⁰, сопоставили границы ТАДов и индивидуальные контакты хроматина внутри ТАДов для каждой пары исследованных видов. Основным результатом, полученный в данном анализе, показал, что закрепившиеся в ходе эволюции хромосомные перестройки предпочтительно проходят в границах ТАДов (т.е. не “разрывают” ТАД), а контакты хроматина внутри синтенных областей у разных видов демонстрируют высокое сходство.

Полученные нами данные предполагают, что структура ТАДов поддерживается отбором, а при изменениях геномного расположения локусов, вызванных хромосомными перестройками, ТАД предпочтительно меняет позицию единым блоком. В то же время, согласно нашему анализу границы ТАДов у комаров-анофелесов являются особенными регионами, причем среди их свойств следует особенно выделить обогащение генами и метками активного хроматина. Известно, что частота хромосомных перестроек пропорциональна плотности генов ²⁴, поэтому мы предположили что именно это свойство границ ТАДов, а не их способность инсулировать контакты хроматина, объясняют предпочтительное совпадение с границами синтенных блоков. Чтобы проверить эту гипотезу, мы сравнили инсуляцию в участках закрепившихся в эволюции хромосомных перестроек с наборами геномных участков, имеющих такую же плотность генов. Оказалось, что при использовании выровненного по плотности генов контроля повышенная инсуляция в районах закрепившихся хромосомных перестроек не наблюдается. Таким образом, ТАДы в ходе эволюции предпочтительно сохраняются как единые блоки и не разрываются хромосомными перестройками, но это связано не с ролью границ ТАДов в инсуляции контактов хроматина, а является побочным следствием обогащения границ ТАДов генами.

В совокупности, результаты, полученные нами при исследовании трехмерной организации геномов животных, позволяют сделать следующие заключения. Во-первых, ряд принципов организации генома является общим как для позвоночных, так и для беспозвоночных животных, в первую очередь - компартментализация хроматина, связанная с распределением специфических эпигенетических сигналов и интенсивности транскрипции. В то же время, некоторые механизмы являются специфичными для отдельных таксонов или даже типов клеток. Так, формирование когезин-опосредованных ТАДов характерно для большинства соматических типов клеток позвоночных (за исключением дифференцированных клеток эритроидного ряда), но, по всей видимости, не наблюдается у беспозвоночных животных. В геномах комаров рода *Anopheles* наблюдаются уникальные петлевые взаимодействия, которые не обнаружены у других исследованных таксонов. Во-вторых, закрепившиеся в ходе эволюции хромосомные перестройки чаще происходят в окрестности границ ТАДов, а контакты хроматина в синтенных областях генома, как правило, демонстрируют сходство между видами. Эти закономерности, согласно нашим данным, наблюдаются как для позвоночных ¹², так и для беспозвоночных ⁷ животных, что свидетельствует в пользу функциональной роли ТАДов. В то же время, в некоторых случаях особенности взаимного расположения “горячих точек” хромосомных перестроек и границ ТАДов может быть объяснено расположением генов, и поэтому не обязательно свидетельствует о действии отбора на поддержание структуры ТАДов.

2.3. Разработка биоинформационных методов для предсказания трехмерной архитектуры хроматина и эпигенетических характеристик хроматина в клетках млекопитающих

Из приведенных выше данных о сохранении укладки генома неизменной на протяжении сотен миллионов лет следует предположение о функциональной роли контактов хроматина. Вскоре это предположение было экспериментально подтверждено в серии работ Штефана Мундлоса, а позднее и в исследованиях других групп ⁹. Было убедительно показано, что структурные перестройки, затрагивающие

границы ТАДов, могут приводить к изменению профиля экспрессии генов и вызывать различные нарушения развития у человека. Одновременно были получены сходные данные о роли ТАДов в развитии онкологических заболеваний ²⁵. Эти работы вызвали значительный интерес специалистов в области медицинской генетики, поскольку нередко структурные перестройки, границы которых удалены на сотни тысяч пар оснований от гена, могут вызывать нарушения его регуляции. Учитывая, что на частоту контактов хроматина влияют многие факторы - связывание белка CTCF, расположение сайтов загрузки когезина, распределение модификаций гистоновых белков, процесс транскрипции - оценить как именно хромосомная перестройка повлияет на профиль контактов хроматина достаточно сложно. Например, далеко не все делеции сайтов связывания белка CTCF приводят к полному слиянию ТАДов, и даже полное слияние ТАДов не всегда приводит к развитию заболеваний ⁹. Все это делает крайне сложной задачу интерпретации пациент-специфичных геномных вариантов, в первую очередь структурных перестроек, в контексте 3D-геномики.

В связи с этим, мы решили разработать инструменты автоматического предсказания профиля трехмерных контактов хроматина на основе эпигенетических данных. Поскольку для систем регуляции экспрессии генов характерны нелинейные зависимости, а эпигенетические механизмы часто взаимосвязаны между собой, мы решили использовать методы машинного обучения, которые способны аппроксимировать сложные зависимости.

В разработанном нами подходе, реализованном в виде биоинформатического инструмента 3DPredictor ²⁶, на вход подается информация об эпигенетических характеристиках пары локусов, а на выходе требуется получить частоту контактов между этими локусами. Процесс машинного обучения заключается в поиске математического преобразования, способного на основе входной информации получать значения, наиболее точно соответствующие искомым частотам контактов хроматина. Для обучения необходимы экспериментальные данные о частотах контактов хроматина и эпигенетических характеристиках локусов, которые используются в процессе подбора математического преобразования. После обучения

созданный инструмент можно применять в ситуациях, когда экспериментальные Hi-C данные отсутствуют - например, для типов клеток, на которых Hi-C-эксперимент не проводился, или для геномов, содержащих хромосомные перестройки. Поскольку процесс вычисления полностью автоматизирован, его можно повторить для каждой пары локусов в геноме и, таким образом, “предсказать” результат Hi-C-эксперимента (на практике мы анализировали локусы на расстоянии 1-2 миллионов п.о., поскольку подавляющее большинство регуляторных взаимодействий в геноме человека ограничено этим расстоянием).

Мы использовали для обучения и тестирования алгоритма опубликованные Hi-C-данные нескольких типов клеток мыши и человека, для которых был доступен широкий набор эпигенетических профилей, а также данные секвенирования транскриптомов. Для оценки точности предсказаний мы разработали специализированную вычислительную платформу 3DGenBench ²⁷. Мы обнаружили, что параметризация, или, другими словами, математическое представление этих биологических данных, является критически важным этапом для создания предсказательного алгоритма. Параметризация, которую мы использовали в финальной версии алгоритма, включала несколько статистических параметров распределения эпигенетических факторов внутри локусов, частоту контактов хроматина между которыми необходимо предсказать, а также вне этих локусов и в участке генома между ними. Для более детального описания профиля связывания CTCF мы дополнительно разработали способы параметризации, численно выражающие направление сайтов связывания ²⁶.

Работа с методами машинного обучения как правило предполагает разделение доступного массива данных на несколько наборов. Тренировочный набор используется для автоматического поиска функции, отображающей входные данные в искомое значение - в данном случае, параметризованную информацию об эпигенетическом состоянии локусов в частоту контактов хроматина между ними. Валидационный и тестовый наборы данных используется для подбора параметров обучения и для оценки того, насколько хорошо найденные в ходе обучения закономерности обобщаются на

данных, не использованных для обучения. Для адекватной оценки точности алгоритма необходимо, чтобы образцы в созданных наборах данных были независимыми. В ходе этого исследования мы сделали важное наблюдение - при использовании методов машинного обучения для решения задач геномики, наборы данных не должны содержать не только идентичных образцов, но и образцов с пересекающимися геномными интервалами. Мы обнаружили, что при наличии в наборах данных пересекающихся геномных интервалов точность алгоритма может быть существенно переоценена и показали, как в этом случае может происходить “утечка” (“data leak”) информации из тренировочного в валидационный и/или тестовый набор данных ²⁶.

Используя корректное разбиение данных и оптимальную схему параметризации эпигенетической информации, мы обучили модель на основе ансамбля решающих деревьев для предсказания частот контактов хроматина. Анализ важности отдельных эпигенетических характеристик локусов показал, что информация о расположении сайтов связывания белка CTCF, уровне транскрипции, а также расстоянии между локусами, является наиболее важной для точности предсказания (рис. 9). Использование дополнительной информации о состоянии хроматина (например, распределения гистоновых модификаций или участков открытого хроматина) позволяет достичь лишь небольшого повышения точности, в пределах нескольких процентов.

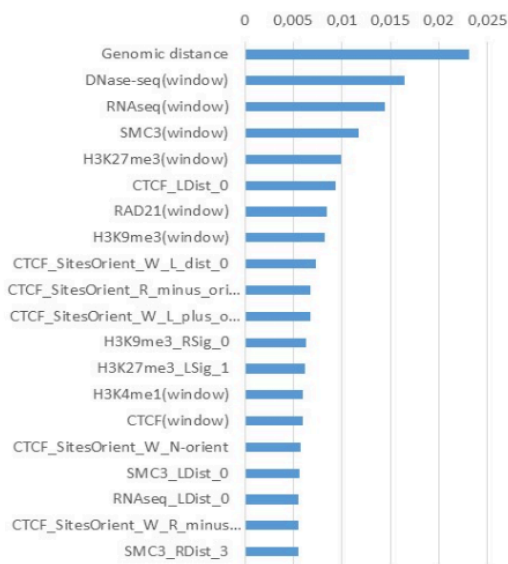


Рисунок 9. Значимость различных факторов для предсказания трехмерных контактов хроматина инструментом 3DPredictor X (подписана сверху) показывает значимость фактора в условных единицах (чем больше значение, тем больший вклад фактор в успешное предсказание). Первый по значимости фактор (genomic distance) - геномное расстояние между локусами. Обозначение “window” (означает, что использовалась информация в геномном окне между двумя точками контакта, которых необходимо предсказать).

Мы также показали, что модель, обученная на данных одного типа клеток, может быть использована для предсказания контактов хроматина в других типах клеток - для этого на вход уже обученной модели нужно подать эпигенетическую информацию для нового типа клеток (рис. 10, А-Г). Способность модели обобщать закономерности на типы клеток, не использованные в обучении, является крайне важной. В противном случае созданный алгоритм был бы, по существу, бесполезен - ведь для обучения необходимо знать результат Hi-C-эксперимента, а значит предсказания, сделанные для этого типа клеток, не имеют никакой практической ценности.

Используя разработанный инструмент, мы предсказали, как изменятся трехмерные контакты хроматина в случае хромосомных перестроек. Для этого, мы модифицировали входные данные в соответствии со структурой перестройки: например, в случае делеции, мы удаляли попавшие в область делеции сайты связывания белка CTCF, при дупликации - дублировали фрагмент генома, содержащимися в нем сайтами связывания и т.п. Мы предложили несколько различных параметров, которые могут быть использованы для сравнения карт трехмерных контактов хроматина в нормальных и мутантных геномах. Например, мы оценили

сходство предсказанных и экспериментально-измеренных эктопических контактов хроматина - контактов, вызванных хромосомной перестройкой. Оказалось, что предсказанные эктопические контакты в значительной степени пересекаются с экспериментально-измеренными и статистически отличаются от сгенерированных случайно.

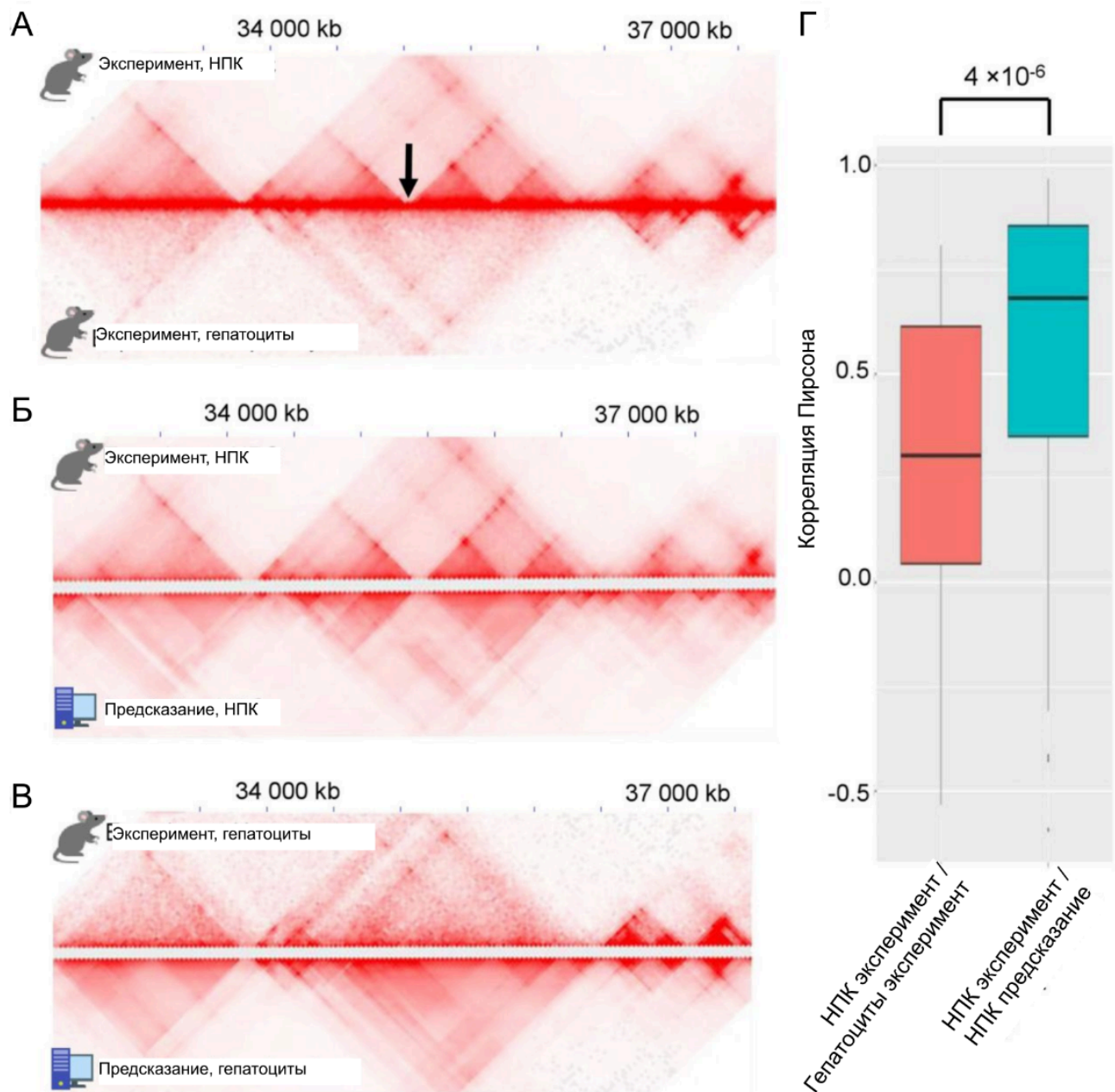


Рисунок 10. Компьютерное предсказание контактов хроматина на основе эпигенетических данных. А-В. Для одного и того же фрагмента генома показаны

экспериментально-измеренные контакты хроматина в гепатоцитах, нейральных предшественниках (НПК), а также компьютерные предсказания этих контактов сделанные на основе соответствующих клеточным типам эпигенетических данных. Стрелкой отмечено место клеточно-специфичных различий. Г. Показатели инсуляции для 88 участков генома, укладка хроматина в которых специфична для НПК, коррелируют с показателями инсуляции, рассчитанными на основе предсказаний модели значительно лучше, чем с показателями инсуляции, полученными на основе экспериментальных данных гепатоцитов ($p\text{-value } 4 \times 10^{-6}$).

Создание метода для предсказания трехмерных контактов хроматина в перестроенных геномах может иметь важное прикладное значение. Однако у предложенного нами подхода есть важное ограничение. В процессе предсказания мы предполагаем, что изменения эпигенетических характеристик локуса происходит только локально, в месте хромосомной перестройки. Это предположение не учитывает транс-эффекты, которые могут быть вызваны мутацией. Кроме того, для предсказания необходимо знать распределение сайтов связывания CTCF и профиль транскрипции в клетках дикого типа - эти данные не всегда доступны. На сегодняшний день, для огромного количества клеток мыши и человека проведены исследования состояния хроматина, однако в каждом случае используется свой набор экспериментальных методик. Например, в базе данных Encode (крупнейшая база эпигенетических профилей для клеток человека) по состоянию на 2020 год было зарегистрировано 858 типов клеток человека и более 40 эпигенетических характеристик (распределение модификацией гистонов, сайтов посадки транскрипционных факторов, доступности хроматина и т.д.), что составляет 34 320 различных возможных измерений. Хотя только 3026 (8%) из этих измерений были проведены, для каждого из 858 типов клеток было доступно более одного измерения (рис. 11, А-В). Это поставило перед нами важный вопрос: возможно ли создать универсальный метод, который на основе любого из 40 измеряемых параметров клетки сможет определить оставшиеся?

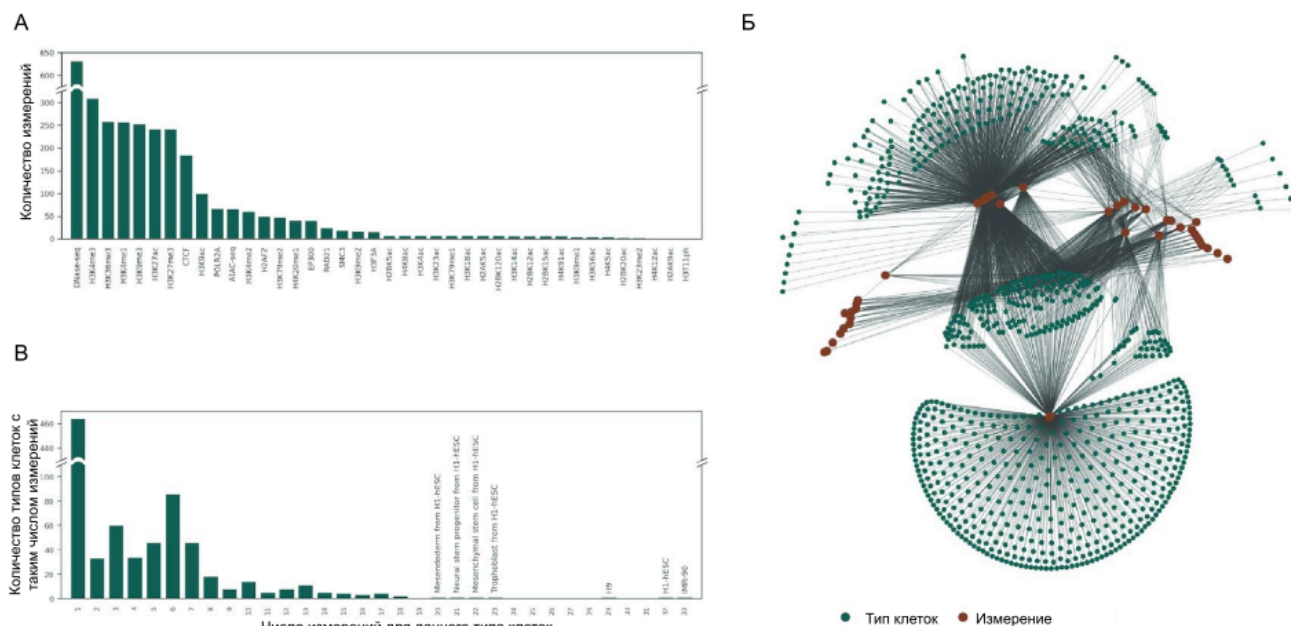


Рисунок 11. Обзор наборов эпигенетических данных ENCODE, доступных для разных типов клеток человека. (А) Количество эпигенетических профилей, доступных для каждого из 40 взятых в работу эпигенетических признаков. (С) Графическое представление измеренных эпигенетических данных. Каждый узел представляет тип клеток (зеленые узлы) или эпигенетические особенности (коричневые узлы), а каждая линия показывает измерение данного признака в данном типе клеток. (В) Гистограмма, показывающая распределение числа измерений эпигенетических признаков для различных типов клеток.

Для ответа на этот вопрос, необходимо рассмотреть, как формируется распределение эпигенетических сигналов в геноме. Хроматин формируется за счет взаимодействия белков - транс-факторов - с последовательностями ДНК. Поэтому, свойства хроматина в первую очередь обусловлены последовательностями ДНК, которые влияют на аффинность связывания белков. Однако очевидно, что одна лишь последовательность ДНК не может полностью задать свойства хроматина: различные клеточные типы, несмотря на идентичную последовательность генома, значительно отличаются друг от друга эпигенетически. Это связано с двумя причинами: во-первых, в разных клеточных типах транс-факторы могут иметь разный уровень экспрессии. Например, если в клетке не экспрессируется ген, кодирующий транскрипционный фактор, все потенциальный сайты связывания этого фактора будут свободными.

Во-вторых, эпигенетические модификации локуса, унаследованные от клетки-предшественника, могут влиять на доступность ДНК для связывания с транс-фактором. Таким образом, эпигенетическая история клетки, набор работающих в ней транс-факторов и последовательность ДНК вместе определяют эпигенетические свойства конкретного локуса (рис. 12, А).

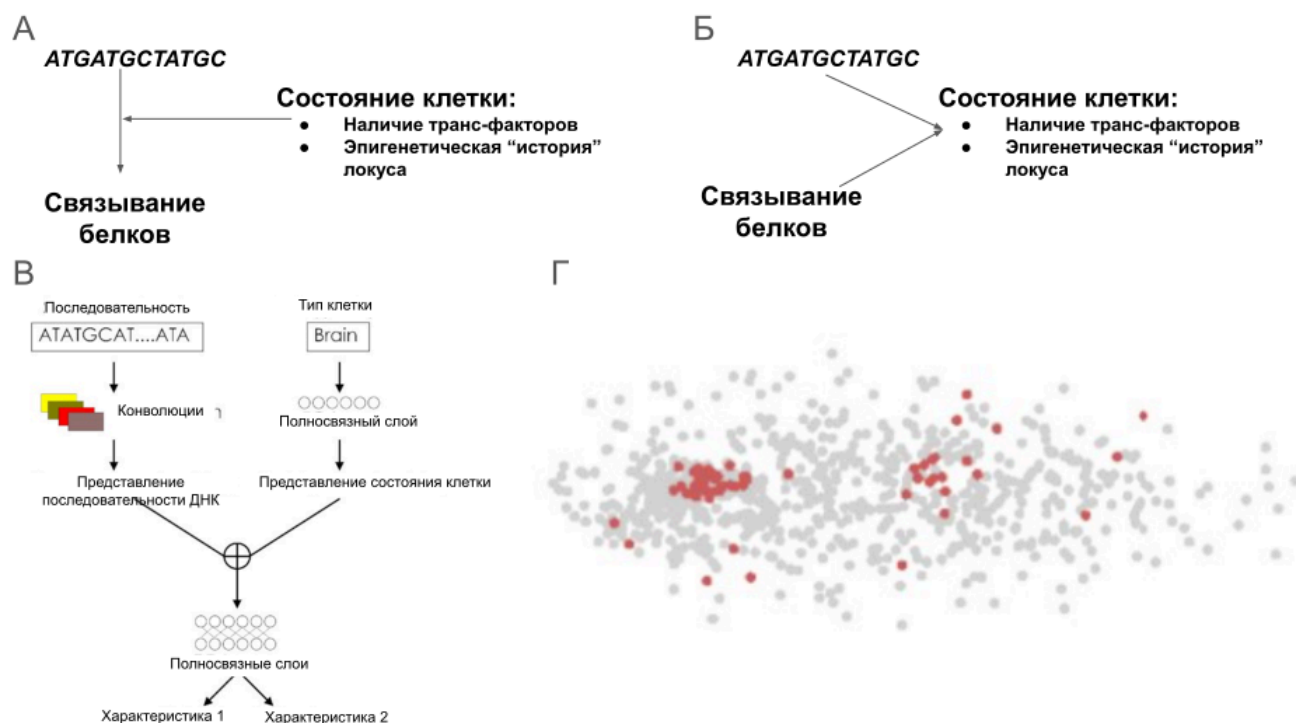


Рисунок 12. Принцип работы нейросети ДеерСТ. А-Б. Идея определения клеточного состояния. А. Белки хроматина связываются с определенными последовательностями ДНК, однако для связывания необходимо наличие белка в ядре клетки и доступность целевого участка ДНК (которая, в свою очередь, определяется эпигенетической "историей" клетки). В совокупности, эти факторы мы определили, как *состояние клетки*. Таким образом, зная состояние клетки и последовательность ДНК можно предсказать связывание белка в данном участке генома. Б. Задача, обратная описанной в схеме - зная информацию о связывании белка и последовательность ДНК можно определить состояние клетки. В. Архитектура нейросети ДеерСТ. Г. Проекция векторных представлений состояний клеток, выученных нейросетью ДеерСТ, в двухмерное пространство. Цветом отмечены подтипы мышечных клеток.

Мы решили ввести понятие *состояния клетки*, которое включает в себя ее эпигенетическую историю и набор работающих транс-факторов. Технически, мы

представляем состояние клетки как вектор в многомерном пространстве, такой, что “похожие” (с биологической точки зрения) клетки представлены близкими геометрически векторами. Мы считаем, что состояния клеток могут быть определены по полногеномному паттерну экспрессии или эпигенетическим меткам, что позволяет группировать сходные типы клеток на основе их геномных свойств (рис. 12, Б). Если состояние клетки действительно отражает ее онтогенетическую историю, то на основе него должно быть возможным описать эпигенетический профиль выбранной последовательности ДНК. Поскольку, многие эпигенетические метки взаимосвязаны, мы предположили, что можно использовать подмножество измеренных эпигенетических характеристик для определения состояния клетки и, в свою очередь, на основе полученного состояния клетки извлечь неизмеренных параметры хроматина.

Идея представления состояния клетки в виде вектора в многомерном пространстве, описанная выше, может быть реализована с использованием различных архитектур нейронных сетей (рис. 12, В). Мы продемонстрировали это, разработав несколько реализаций модели DeepCT²⁸. Модель принимает последовательность ДНК и метку типа клетки в качестве входных данных. ДНК-вход обрабатывает последовательность нуклеотидов и строит ее векторное представление с использованием глубокой сверточной нейронной сети. Второй вход отвечает за интерпретацию состояния клетки. Изначально, каждый вариант состояния клетки (например, каждый тип клеток, в случае анализа базы данных Encode или каждый кластер в случае анализа единичных клеток) кодируется при помощи прямого унитарного кода, т.е. все клеточные состояния представлены ортогональными векторами. Затем, при помощи линейного преобразования происходит переход в пространство меньшей размерности, в котором, согласно нашей концепции, вектора соответствующие биологически “сходным” клеточным типам должны оказаться геометрически близкими. На следующем этапе векторные представления последовательности ДНК и состояния клетки объединяются и передаются в

финальную часть нейросети, которая обрабатывает их и выдает предсказания о наличии каждой эпигенетической метки.

Как было сказано, мы ожидали, что во время обучения сеть будет подбирать свои параметры таким образом, чтобы в скрытом пространстве модели вектора близких клеточных типов кластеризовались вместе. Действительно, небольшой пример из 6 типов клеток, для которых была измерена одна и та же эпигенетическая метка, показывает, как происходит кластеризация похожих (под)-типов во время обучения нейронной сети (рис. 13). Однако возможно ли применение такой модели для более сложных систем?

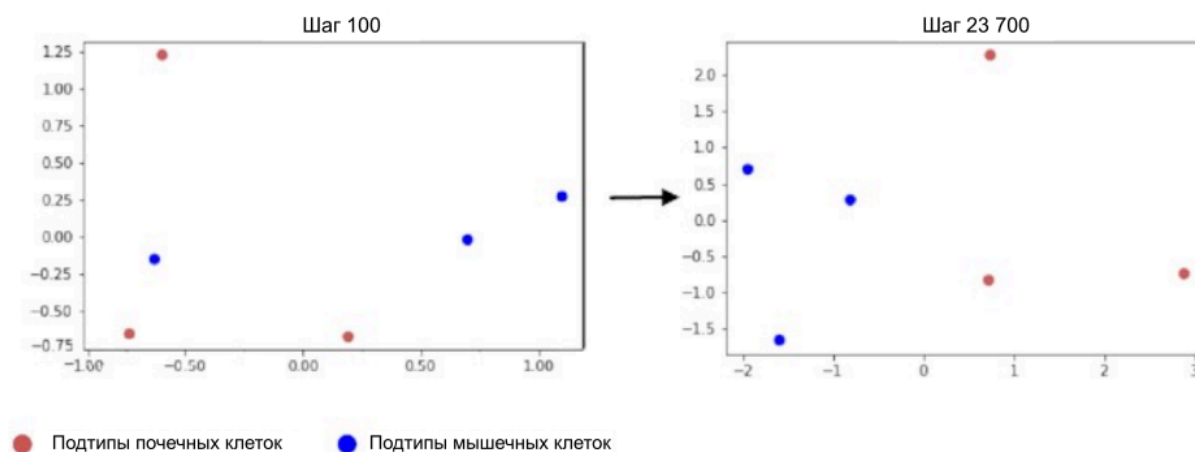


Рисунок 13. Кластеризация векторных представлений для сходных типов клеток во время обучения нейросети DeepCT. Проекция векторных представлений состояний клеток, выученных нейросетью DeepCT, в двухмерное пространство.

Применяя DeepCT к полному набору данных Encode, мы заметили, что кластеризация типов клеток часто соответствовала биологическим ожиданиям (рис. 12, Г). Например, наблюдается колокализация мышечных клеток (рис. 12, Г), а также колокализация клеток тимуса. Однако другие типы клеток, такие как клетки крови или стромальные клетки, не образуют четко выраженных кластеров. Это может быть связано либо с более высокой гетерогенностью этих тканей, либо со специфическими свойствами эпигенетических данных, доступных для этих типов клеток. В целом, сходство векторов, представляющих типы клеток одной и той же ткани, было значительно выше, чем для вложений случайно выбранных типов клеток

(коэффициент Отиаи 0,44 против $0,3917 \pm 0,0029$ для случайных векторов, $p < 1,3e-62$). Этот результат показывает, что модель DeepCT фиксирует биологические свойства клеток во время обучения.

Чтобы оценить точность разработанной модели DeepCT, мы рассмотрели 3 задачи (рис 14, А-В). В первой задаче (задача на различимость клеточных типов, рис. 14, А), мы для каждого геномного локуса случайно выбрали 10% измеренных эпигенетических профилей и исключили их из набора тренировочных данных. Затем, мы оценили точность предсказания модели на исключенных данных. В качестве базового решения (базиса) для сравнения в этой задаче мы использовали значение эпигенетических сигналов локуса, усредненное по всем типам клеток, включенным в обучающую выборку. Биологически этот базисный уровень можно интерпретировать как эпигенетический профиль “среднестатистического” типа клеток. Нужно отметить, что такой базисный уровень имеет высокие показатели качества, значительно превышающие ожидаемые для случайного распределения сигнала (средняя площадь под кривой специфичность-полнота, mAP, равна 0,446), поскольку различия между последовательностями объясняют значительно большую дисперсию эпигенетических данных, чем различия между типами клеток.

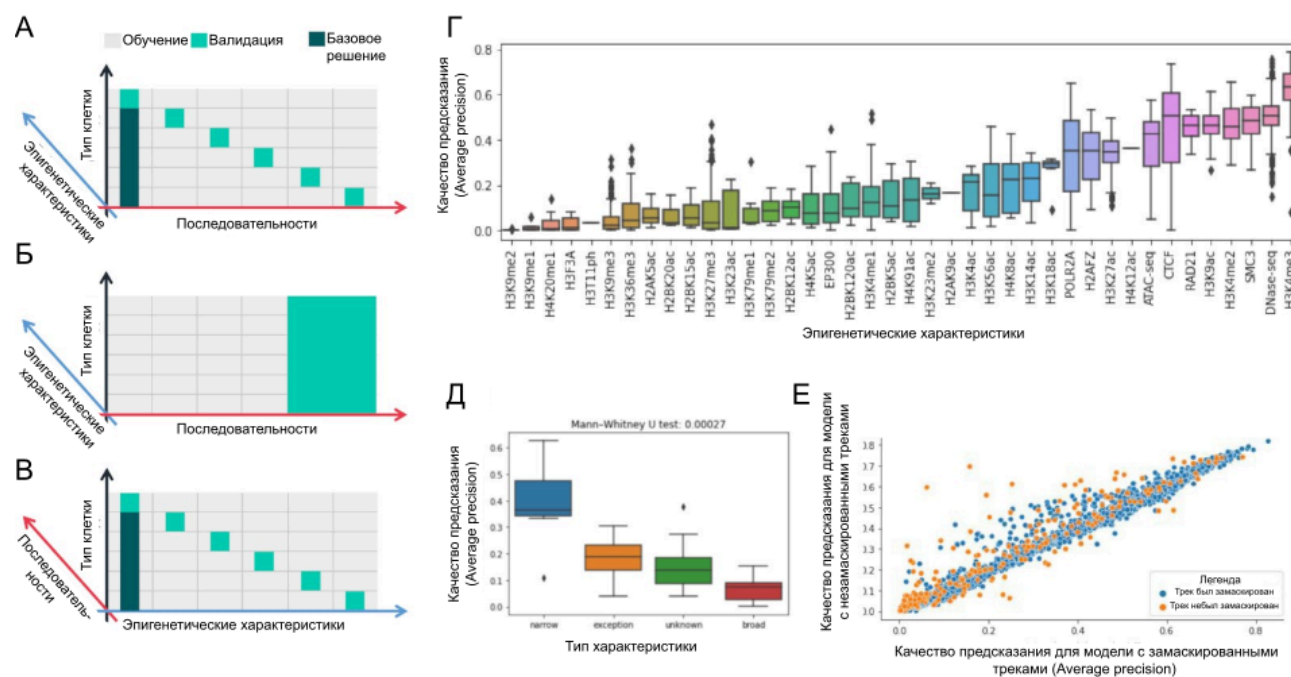


Рисунок 14. Качество предсказания эпигенетических характеристик на основе нейросети DeepCT. А-В. Дизайн трех задач, для решения которых применялась архитектура DeepCT. Каждый набор данных содержал 3 модальности: различные фрагменты последовательностей ДНК, различные типы клеток и различные эпигенетические характеристики. В каждой из задач образцы с определенной комбинацией модальностей используются для обучения и валидации моделей, а также для составления “базового” решения (решения, основанного на простых статистиках). Например, на панели А показано, что в обучающей выборке присутствуют все геномные последовательности, но для каждого участка генома данные об эпигенетических сигналах доступны только для части клеточных типов. Недоступные в обучении (но измеренные экспериментально) данные используются для валидации обученной модели. Базовое значение в этом Г. Точность предсказания различных эпигенетических характеристик моделью DeepCT. Д. Зависимость точности предсказания гистоновых модификаций от типа модификации (тип определялся по классификации консорциума Encode как один из трех: broad, narrow, exception). Е. Сравнение точности предсказания эпигенетических характеристик для моделей, в которых часть “треков” (трек - это комбинация модальностей тип клетки-эпигенетическая характеристика) была замаскирована (недоступна при обучении) или у которой все треки были доступны при обучении. Результаты соответствуют дизайну эксперимента, представленному на панели В. Для всех моделей качество данных измерено как площадь под графиком специфичность-полнота (“Average precision”).

Мы сравнили производительность моделей DeepCT с полученными базовыми показателями и обнаружили, что модель превосходит их: mAP модели составляет $0,4753 \pm 0,0015$. (статистически выше базовых значений, $p < 2,1e-14$). Это показывает, что выученные моделью состояния клеток имеют биологический смысл и что модель предсказывает специфичные для типов клеток различия в эпигенетических профилях с использованием изученных состояний.

Затем мы протестировали, как хорошо модель обобщает закономерности, выявленные из анализа последовательностей ДНК. В этом тесте мы определили подмножество локусов генома для тестовой выборки и полностью исключили любую информацию об этих локусах из набора обучающих данных (рис. 14, Б). Если в предыдущей задаче для каждого геномного локуса в процессе обучения была доступна хотя бы часть эпигенетической информации, то теперь модель вообще не получала никакой информации о тестовых локусах во время обучения. Ожидаемо,

производительность модели в этом случае оказалась ниже ($mAP \sim 0,34$). Следует отметить, что использование ранее опубликованной архитектуры DeepSEA показало сопоставимые оценки ($mAP \sim 0,32$), хотя DeepCT получает возможность делать выводы о неизмеренных эпигенетических профилях, а DeepSEA - нет. Таким образом, добавление в архитектуру модели данных о состоянии клеток по крайней мере не ухудшает ее точность в задаче обобщения на последовательностях ДНК.

Мы обнаружили, что производительность DeepCT значительно различается в зависимости от выбранной эпигенетической метки (рис. 14, Г, Д). Например, предсказание монометилирования гистонов (H3K9 и H4K20) имеют почти нулевую метрику mAP ; с другой стороны, метка H3K4me3 показывает высокий показатель mAP (0,62), а для лучших типов клеток мы достигли $AP \sim 0,8$ на скрытых из обучающей выборки последовательностях генома. Прогностическая сила также была высокой для предсказания районов гиперчувствительности ДНКазы I, связывания белка CTCF, когезина и некоторых других наборов эпигенетических данных.

Чтобы понять наблюдаемую разницу между точностью предсказания различных свойств хроматина, мы классифицировали все эпигенетические профили в соответствии с принципами консорциума ENCODE на профили с «узкими» сигналами, «широкими» сигналами, «исключения» и «неизвестные» (рис. 14, Д). Профили первой категории («узкие») как правило включают факторы транскрипции и области открытого хроматина, содержащиеся в пределах коротких (100–200 п.н.) фрагментов генома, тогда как белки хроматина, профили которых соответствуют «широкой» категории, включают гистоновые метки, распространяющиеся на участки до нескольких сотен тысяч пар оснований - например, сигнал транскрипции H3K36me3. Мы обнаружили, что профили с узкими сигналами демонстрируют значительно (U-критерий Манна-Уитни, $p=0,00027$) более высокие средние значения точности (рис. 14, Д). Это различие, вероятно, связано с относительно короткой входной последовательностью ДНК, используемой в инструменте DeepCT. В этой последовательности не содержится всей информации, необходимой для предсказания «широких» сигналов.

Наконец, наиболее сложной задачей является предсказание ранее неизмеренных эпигенетических характеристик (рис. 14, В). Как и ожидалось, качество прогнозов для треков, полностью исключенных из обучения, было ниже, чем для треков, информация о которых была доступна в ходе обучения по крайней мере в части геномных локусов (рис. 14, Е). Однако падение прогностической способности было относительно небольшим, в среднем всего 1,6% от величины АР. Это указывает на то, что архитектуру DeepCT можно использовать для предсказания не измеренных экспериментально эпигенетических профилей, а также для интерпретации некодирующих вариантов в геноме человека.

Таким образом, можно заключить что использованная нами архитектура нейросети DeepCT позволяет находить взаимосвязи между различными эпигенетическими характеристиками клетки, вычленять функционально-значимую информацию из геномных последовательностей и, на основе этих данных, предсказывать эпигенетические профили целевых последовательностей в различных типах клеток. Уникальность этой разработки заключается в возможности использовать произвольный набор входных данных для запуска модели и характеристики типа клеток. Насколько нам известно, аналогов такого подхода до этого предложено не было. Однако следует отметить, что точность модели все же слишком низкая, чтобы широко использовать ее в практических задачах. Ещё одной уникальной особенностью предложенного подхода является возможность представить в численном виде *клеточные состояния* - т.е. эпигенетическую историю и транскрипционный профиль клетки. Аналогичный подход был предложен лабораторией Д. Келли для кодирования состояний единичных клеток ²⁹.

Помимо этого, алгоритм 3DPredictor позволяет получить высокоточные предсказания трехмерной архитектуры генома для типов клеток, охарактеризованных с точки зрения сайтов связывания белка CTCF и экспрессии генов. Этот алгоритм можно использовать для предсказания эффектов, вызванных хромосомными перестройками, а также для поиска регуляторных элементов, пространственно сближенных в ядрах клеток позвоночных животных.

2.4. Использование технологий захвата конформации хромосом для поиска хромосомных перестроек у человека

Анализ геномов комаров рода *Anopheles* убедительно показал, что методы захвата конформации хроматина могут быть использованы не только для изучения механизмов укладки ДНК, но и для сборки геномов¹³. Для геномов, у которых референсная сборка уже имеет высокие показатели континуальности, сходный подход может использоваться для поиска хромосомных перестроек. В случае исследований на клетках и тканях человека, такая технология может иметь важное биомедицинское значение. Однако существенным ограничением этого метода является то, что он способен выявить лишь ограниченный спектр геномных вариантов: хромосомные перестройки, в первую очередь, транслокации. Для применения в клинической практике, важно обеспечить возможность поиска однонуклеотидных вариантов, в первую очередь, в экзонах клинически-значимых генов, на основе Hi-C-данных.

2.4.1. Разработка технологии Echo-C

В основе нашего подхода лежит простой прием - обогащение Hi-C-библиотек экзомными последовательностями (рис. 15, А)^{8,30}. За счет прочтения обогащенных экзомных фрагментов становится возможным детектировать однонуклеотидные варианты или короткие вставки и делеции. В то же время за счет анализа паттерна контактов экзомных последовательностей сохраняется возможность детектировать хромосомные перестройки. Важно, что хромосомные перестройки оказывают влияние не только на контакты участков, близлежащих к точкам разрывов хромосом, но и на существенно более далеки локусы. Поэтому мы предположили, что, анализируя контакты экзонов, можно детектировать хромосомные перестройки, точки разрывов которых лежат вне генов. Таким образом, при глубине секвенирования ненамного большей, чем при стандартном экзомном секвенировании, мы получим информацию как о структурных хромосомных перестройках, так и о точковых мутациях. Важным дополнительным преимуществом нашего подхода является тот факт, что результаты

секвенирования 3С-библиотек могут быть использованы, в случае обнаружения перестроек, для характеристики изменений границ ТАДов в районе перестройки.

В целях разработки этой технологии, мы протестировали набор протоколов с использованием различных типов нуклеаз. В стандартном 3С-протоколе используются эндонуклеазы рестрикции, узнающие четырехбуквенную последовательность, для фрагментации ДНК при создании Hi-C библиотеки. В этом случае около 50% всего генома будет расположено на расстоянии менее 400 букв от какого-либо сайта рестрикции. Таким образом, при парноконцевом секвенировании библиотеки фрагментов с размером вставки 300-400 п.о. (характерный размер вставки для paired-end библиотеки Illumina) мы имеем возможность исследовать только около 50% потенциальных вариантов в экзOME человека. В связи с этим, нам необходимо было модифицировать 3С-протокол, используя эндонуклеазы рестрикции без фиксированного сайта узнавания - это позволило бы обеспечить равномерное покрытие прочтениями экзонов генов. Мы провели эксперименты с использованием ДНКазы I - этот фермент не имеет специфического сайта узнавания, поэтому можно анализировать фрагменты генома независимо от их удаленности от какого-либо сайта рестрикции и, таким образом, довести теоретическое количество анализируемых полиморфизмов до 100%. Мы показали, что при использовании ДНКазы I наблюдается гораздо более равномерное покрытие прочтениями, чем при использовании рестриктазы DpnII с фиксированным сайтом узнавания (рис. 15, Б) Поэтому, мы поставили задачу совместить протокол Hi-C с использованием ДНКазы I (DNase Hi-C) с этапом экзомного обогащения.

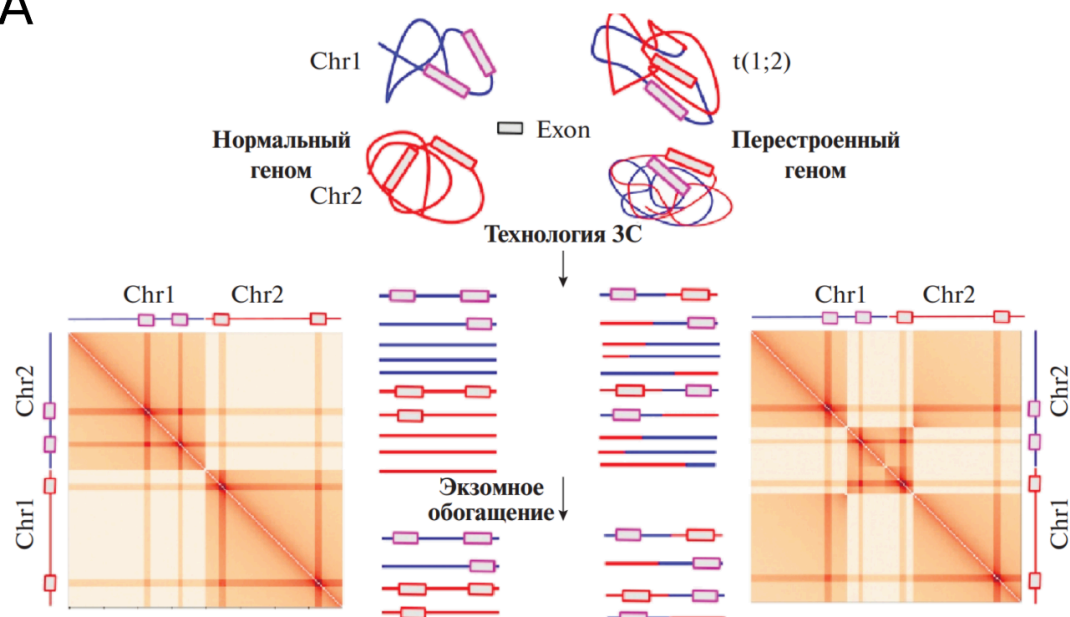
В процессе оптимизации нами было подготовлено более 50 геномных библиотек используя несколько модификаций протокола Hi-C с ДНКазой I (рис. 15; каждый пул на панели Д соответствует определенной модификации протокола) ³¹. По сравнению с опубликованными ранее методами, усовершенствованная нами технология позволяет существенно повысить выход данных, а также добиться высокого уровня экзомного обогащения (рис. 15, Д). Однако для всех версий Hi-C с ДНКазой I наблюдается высокое количество нехимерных фрагментов ДНК (dangling ends, рис. 15, Д). Мы

предположили, что причиной появления таких фрагментов может быть никазная активность ДНКазы I (этот фермент может вносить как одно-, так и двухцепочечные разрывы в ДНК), не совместимая с процедурой селекции химерных фрагментов в протоколе Hi-C (см. более подробное объяснение этой гипотезы в нашей работе ³²). В соответствии с этим предположением, мы провели поиск нуклеаз, обеспечивающий равномерный гидролиз фиксированного хроматина и не демонстрирующих никазную активность. Протестировав несколько ферментов, удовлетворяющих этим условиям, мы обнаружили, что использование S1-нуклеазы позволяет снизить количество нехимерных фрагментов приблизительно в 1.5 раза по сравнению с ДНКазой I (рис. 15, Д, пул 12 и рис. 15, В и Г).

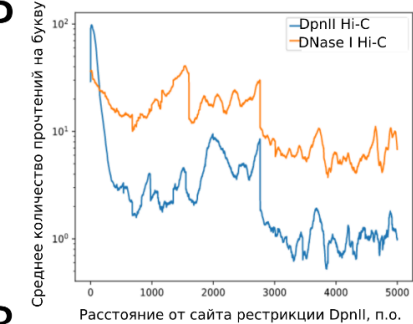
Следует подчеркнуть, что, насколько нам известно, это первый в мире опыт использования S1-нуклеазы в 3C-протоколах. Учитывая, что этот фермент может быть использован и в других экспериментальных технологиях геномного анализа, мы решили провести более детальное исследование профиля гидролиза хроматина S1-нуклеазой. Оказалось, что этот фермент имеет небольшое предпочтение к гидролизу участков активного хроматина, однако распределение сайтов гидролиза S1-нуклеазы более равномерное, чем у ДНКазы I, к тому же S1-нуклеаза не имеет специфичной последовательности узнавания. Таким образом, практически любая последовательность может быть субстратом для S1-нуклеазы, что делает использование этого фермента привлекательным для методов, требующих обеспечить равномерную представленность фрагментов ДНК в результате секвенирования.

Таким образом, нами была впервые показана возможность совмещения технологий захвата конформации хромосом и обогащения экзоными последовательностями. Мы назвали разработанный протокол **Exo-C** (от совмещения **Exome** и **Hi-C**).

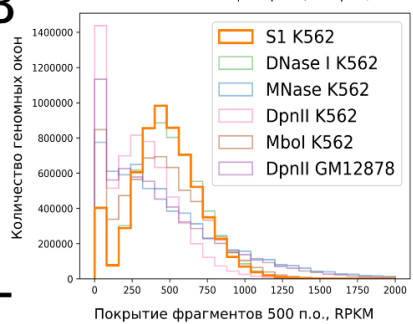
A



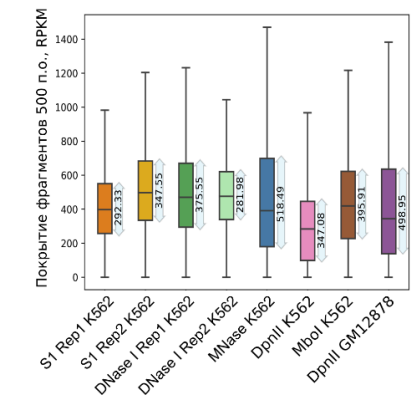
Б



В



Г



Д

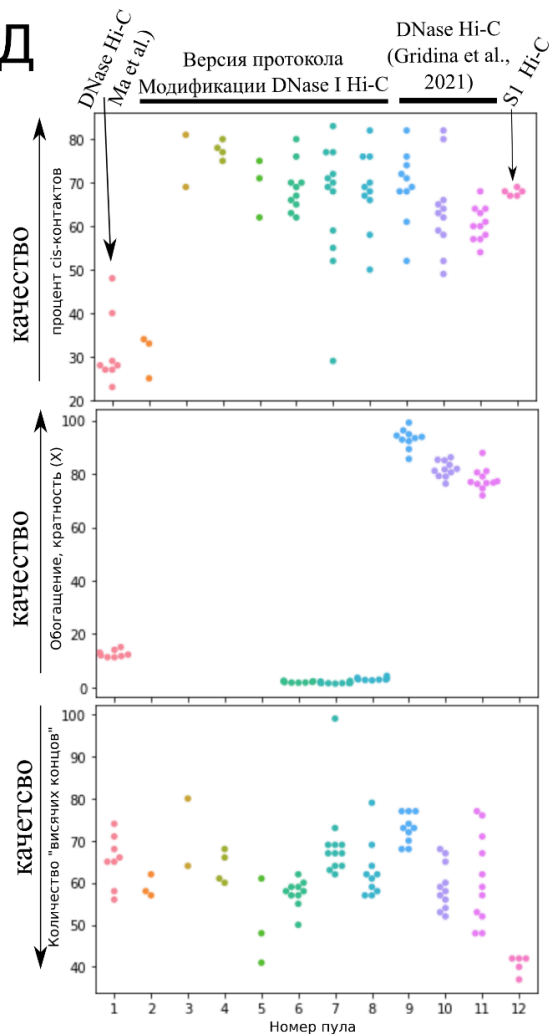


Рисунок 15. Разработка метода Ехо-С. А. Схема Ехо-С-эксперимента. Вверху схематично изображена 3D-конформация нормального (слева) и перестроенного (справа) генома. В качестве примера перестройки приведена реципрокная транслокация между хромосомами 1 (синий цвет) и 2 (красный цвет). Внизу в центре показаны полная и обогащенная экзомными последовательностями библиотеки, каждая полоска соответствует одной химерной молекуле ДНК, образовавшейся после лигирования. Справа и слева от них изображены тепловые карты пространственных контактов, соответствующие обогащенным библиотекам: каждой паре локусов соответствует одна точка на тепловой карте, а ее цвет отражает количество прочтений, содержащих фрагменты ДНК, картирующиеся на данную пару локусов. Красные полосы отражают районы обогащения (экзоны). Видно, что хромосомная перестройка приводит не только к изменению картины в месте соединения фрагментов двух хромосом, но и паттерна контактов, значительно удаленных от места перестройки. Б. Зависимость геномного покрытия от расстояния до сайта рестрикции DpnII при проведении Hi-C-эксперимента с ферментом DpnII или DNКаза I. В и Г. Распределение количества прочтений, приходящихся на 500-буквенные участки, для экспериментов Hi-C проведенных с использованием разных ферментов. Панель В иллюстрирует форму распределения, панель Г отражает то же самое распределение в виде ящиков с усами. На панели Г сбоку от каждого графика подписана разница между значением 75-го и 25-го перцентиля распределения. Чем больше это значение, тем более “широкое” распределение, и, соответственно, тем менее равномерно геномное покрытие. Д. Характеристика качества Ехо-С-библиотек, приготовленных с использованием разных протоколов. Процент cis-контактов отражает шум в 3С-данных (чем выше процент cis-контактов, тем меньше шум), степень обогащения показывает отношение среднего покрытия в экзоне к среднему покрытию в оставшейся части картируемого генома, количество “висящих концов” (dangling ends) показывает число нехимерных фрагментов ДНК, посчитанное на основе пропорции ридов в FF/RR/FR/RF ориентациях и отражает долю прочтений, несущих информацию о трехмерных контактах (чем ниже метрика, тем больше информативных ридов). Следует отметить, что не все образцы доводили до стадии экзомного обогащения и глубокого секвенирования, поэтому для части пулов статистика по обогащению не представлена. Детально каждый из протоколов, а также характеристики, использованные для оценки качества библиотек, описаны в статье³¹.

2.4.2. Применение технологии Ехо-С для идентификации и характеристики хромосомных перестроек у пациентов

Используя разработанный протокол Ехо-С, мы получили карты трехмерных контактов для 66 образцов клеток человека (рис. 16). Мы показали возможность получения Ехо-С библиотек из разного биоматериала: 55 библиотек были получены из первичных лимфоцитов периферической крови без предварительной заморозки; 5

библиотек - из замороженных образцов мононуклеаров крови; 3 библиотеки - из культур перевивных клеточных линий человека: LNCaP, K542, A549; 3 библиотеки - из индуцированных плюрипотентных стволовых клеток, ранее полученных из первичных культур пациентов. Совокупная карта 3С-контактов лимфоцитов, построенная по нашим данным, включает более чем 2 миллиарда прочтений.

Мы подбирали доноров материалов для сборки Ехо-С-библиотек, основываясь на нескольких критериях. Во-первых, мы собрали выборку пациентов с описанными хромосомными перестройками: транслокациями (23 пациента), инверсиями (6 пациентов), делециями и дупликациями (27 пациентов). Эту группу пациентов мы использовали для валидации разрабатываемых подходов к детекции геномных вариантов на основе Ехо-С-данных. Во-вторых, мы взяли в анализ несколько пациентов, у которых был проведен тот или иной генетический тест первой линии (хромосомный микроматричный анализ и кариотипирование в связи с подозрениями на хромосомную аномалию; экзомное секвенирование - при подозрении на моногенное заболевание), но при этом патогенный вариант найден не был.

Все Ехо-С образцы: 66

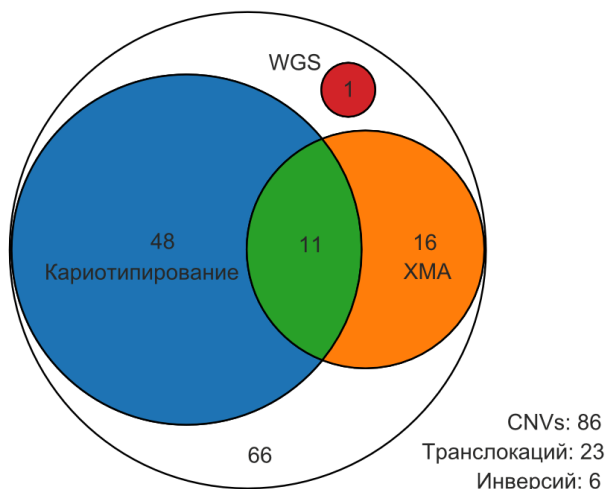


Рисунок 16. Характеристика образцов, использованных для апробации метода Ехо-С. Диаграмма Венна показывает различные тесты, выполненные для выборки из 66 человек, взятых в анализ. ХМА - хромосомный микроматричный анализ. В правом нижнем углу показано общее количество аномалий копийности (CNV), транслокаций и инверсий, обнаруженных у пациентов, взятых в анализ.

Поскольку фокусом исследования была разработка метода детекции генетических вариантов, а не изучение какой-то определенной группы генетических заболеваний, мы не ограничивали набор участников исследования пациентами с

какой-то одной нозологией. В нашей выборке оказались пациенты с различными аномалиями, но большинство из них имело те или иные отставания в психоречевом и/или моторном развитии.

Применение Ехо-С для анализа клеток периферической крови у собранной группы пациентов, а также *in silico* моделирование Hi-C-карт с паттернами, характерными для хромосомных перестроек показали, что совмещение технологий захвата конформации хромосом и экзомного обогащения позволяет эффективно (чувствительность 50% и более, вплоть до 100%; специфичность более 50%) детектировать транслокации и инсерции размером более 100 тысяч п.о. и инверсии размером более 500 тыс. п.о. Более того, в отдельных случаях инверсии и транслокации существенного меньшего размера (вплоть до 1 тысячи п.о. в случае транслокаций) могут быть детектированы при помощи этого метода.

Среди пациентов с неустановленной молекулярной причиной патологии по результатам Ехо-С-эксперимента удалось найти патогенные или вероятно патогенные однонуклеотидные варианты в экзомах трех пациентов. Кроме того, мы смогли подтвердить наличие патогенных хромосомных перестроек у подавляющего большинства пациентов, для которых патогенный вариант уже был аннотирован до начала анализа с использованием классических цитогенетических методов диагностики.

Среди найденных нами патогенных вариантов, некоторые не были описаны ранее в литературе и представляют отдельный интерес для изучения. Мы также обнаружили некоторые варианты, которые могут быть отнесены к вариантам неопределенного клинического значения согласно текущим стандартам диагностики, и провели ряд исследований для уточнения их клинического значения. Ниже приведено два наиболее интересных из найденных нами клинических случаев.

Случай 1. Материал пациентки P10 был предоставлен нам коллегами из группы Н.В. Шиловой (МГНЦ). Согласно фенотипическому описанию, предоставленному врачами, у пациентки (возраст 8 месяцев на момент исследования) наблюдались: задержка психомоторного развития (поворачивается на живот с 6 месяцев, не ползает,

сама не садится), открытое овальное окно (закрывается), мало прибавляет в весе, гипотония, микроцефалия (окружность головы 42 см), большой родничок 1х1 см, широкая переносица, монголоидный разрез глаз, низко посаженные диспластичные ушные раковины, клинодактилия 5 пальцев кистей, неправильное положение пальцев стоп, вальгусные стопы. Цитологический анализ, проведенный группой Н.В. Шиловой, показал наличие сбалансированной транслокации t(5;10).

Нами был проведен Ехо-С-эксперимент, который подтвердил наличие транслокации и позволил уточнить границы разрывов с точностью несколько тысяч п.о. (рис. 17, А). На основе полученных Ехо-С-данных мы провели дизайн праймеров и амплифицировали предполагаемую границу разрыва, а затем секвенировали полученный ампликон. Результаты секвенирования подтвердили наличие разрыва в предполагаемой области, а также позволили с точностью до нуклеотида установить границы разрыва. Выяснилось, что одна из точек разрыва находится в интроне гена *PCDH15*, а другая - недалеко от гена *RNF180* (рис. 17, Б). Оба гена являются клинически-значимыми: мутации в гене *PCDH15* ассоциированы с наследственной глухотой (синдром Ушера, OMIM 605514), а в гене *RNF180* - с определенными формами рака желудка. Однако фенотип пациента не соответствует нарушениям, наблюдаемым при мутациях *RNF180* и *PCDH15*. В то же время, Ехо-С-данные показывают, что транслокация вызывает изменения контактов промотора гена *DKK1*, расположенного приблизительно в 2 миллионах пар оснований от границы перестройки (рис. 17, А). Хотя это расстояние является очень большим для цис-регуляторных взаимодействий, в геномах млекопитающих описаны отдельные случаи функциональных контактов элементов, столь удаленных друг от друга. При этом нарушения экспрессии гена *DKK1* вызывают отклонения в морфогенезе лицевого скелета и аномалии развития пальцев конечностей, что хорошо соответствует фенотипу пациентам. Поэтому мы предположили, что фенотип, наблюдаемый при транслокации, вызван нарушениями в экспрессии *DKK1*, а не мутациями *RNF180* или *PCDH15*.

Для проверки этой гипотезы мы репрограммировали периферические мононуклеары, выделенные из крови пациента, получив индуцированные плюрипотентные стволовые клетки (ИПСК), используя протокол описанный нами ранее (рис. 17, В) ^{33,34}. Мы выбрали 2 клон ИПСК и провели анализ маркеров плюрипотентности, который показал экспрессию генов *OCT4*, *SOX2* и *NANOG*, характерных для плюрипотентных клеток (рис. 17, Д). Затем, мы провели дифференцировку ИПСК в клетки первичной полоски, в которых ген *DKK1* экспрессируется на высоком уровне (рис. 17, Г). Анализ экспрессии гена *DKK1* не показал значимых различий между дифференцированными производными ИПСК пациента и клетками, полученными от здорового донора (рис. 17, Е). Таким образом, мы убедились, что экспрессия гена *DKK1* не изменяется, по крайней мере в клетках первичной полоски, несмотря на появление эктопических контактов промотора этого гена с транслоцированным фрагментом хромосомы 5.

Транслокация между хромосомами 5 и 10 являлась единственным вариантом, обнаруженном при цитологическом кариотипировании пациента P10. Поскольку нам не удалось получить данных, свидетельствующих о патогенности этого варианта, мы предположили, что в геноме пациента могут быть другие нарушения субмикроскопического масштаба, не выявленные при кариотипировании. Анализ однонуклеотидных вариантов в экзоме пациента не выявил патогенных или вероятно-патогенных вариантов. Поэтому, мы провели детальный анализ Ехо-С-карт пациента в поисках субмикроскопических хромосомных перестроек, и обнаружили несколько, наиболее важной из которых с клинической точки зрения оказалась гетерозиготная инверсия участка хромосомы 10 размером около 2 миллионов п.о. (рис. 17, Ж, 3). Инверсионный паттерн виден на картах Ехо-С, но ещё лучше виден на небогащенных экзомными последовательностями картах ДНКазной Hi-C, дополнительно просеквенированных нами для пациента P10 (рис. 17, Ж).

По данным Hi-C и Ехо-С мы смогли установить, с точностью до нескольких тысяч пар оснований, границы инверсии и показали, что они проходят внутри интронов генов *KAT6B* и *OIT3* (рис. 17, Ж, 3). Ген *KAT6B* кодирует лизиновую

ацетилтрансферазу, а его мутации приводят к развитию SBBYS-синдрома (OMIM #603736) по аутосомно-доминантному типу. Характерные для этого синдрома нарушения хорошо согласуются с фенотипом пациента, что позволяет предположить, что гетерозиготная инверсия, обнаруженная нами, приводит к образованию нефункционального транскрипта *KAT6B*. Для проверки этой гипотезы, мы провели транскриптомный анализ ИПСК, полученных от пациента P10 и двух контрольных линий.

Согласно публичным базам данных, для ИПСК характерна экспрессия *KAT6B*, но не *OIT3*. Действительно, в полученных нами транскриптомных данных большое количество ридов картировалось в экзоны *KAT6B*, но не *OIT3*. При этом нам удалось обнаружить пять химерных ридов, соединяющих третий экзон гена *KAT6B* (находится около 5'-границы инверсии) и участки в интронах генов *OIT3* и *MCU* (эти гены фланкируют инвертированный участок с 3'-стороны). Кроме того, мы обнаружили кластер ридов в интроне гена *MCU*, не наблюдающийся в контроле. Эти данные свидетельствуют, что в геноме пациента P10 действительно инвертирован участок между генами *OIT3* и *KAT6B*, что приводит к формированию химерного транскрипта, включающего три первых экзона гена *KAT6B* и, в зависимости от альтернативного варианта сплайсинга, различные участки генов *OIT3* и *MCU*.

Наконец, чтобы прямо подтвердить отсутствие полноразмерных транскриптов *KAT6B* с инвертированного аллеля, мы обратились к анализу однонуклеотидных вариантов в этом гене. Анализируя Hi-C-данные, нам удалось обнаружить в геноме P10 гетерозиготный вариант A/G, расположенный в восьмом экзоне *KAT6B*. Восьмой экзон находится правее обнаруженной нами точки разрыва инверсии, поэтому мы ожидали, что транскрипты, стартующие внутри инвертированного участка с промотора *KAT6B* не будут содержать восьмой экзон. Это предположение подтвердилось, поскольку в транскриптомных данных, несмотря на достаточное покрытие восьмого экзона (28 ридов в сумме в двух репликах) мы обнаружили только один из двух аллелей (рис. 17, И,К). Таким образом, мы показали, что один из аллелей *KAT6B* не экспрессируется в

клетках пациента P10, что объясняет фенотип пациента эффектом гаплонедостаточности по гену *KAT6B*.

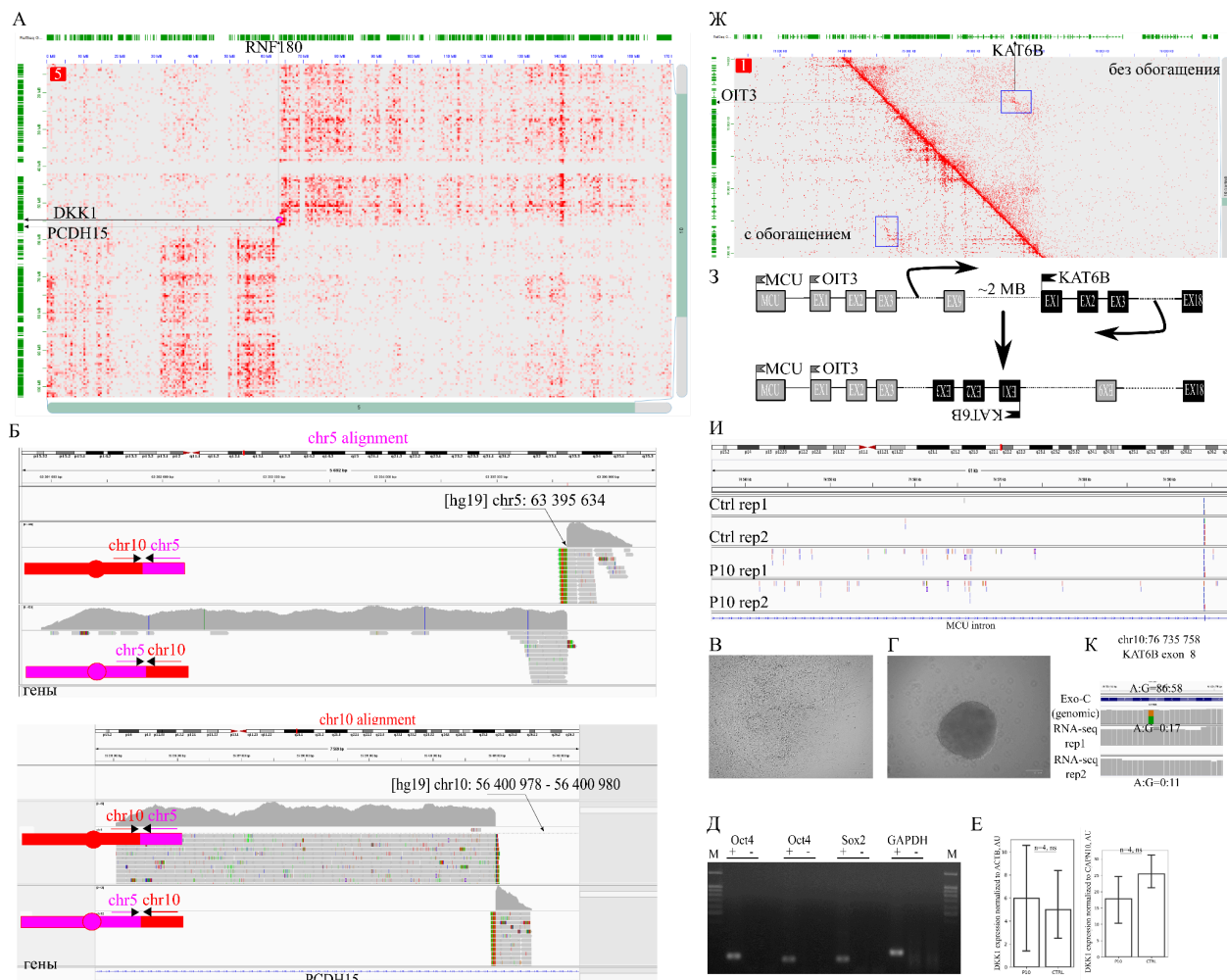


Рисунок 17. А. Фрагмент межхромосомных контактов хромосом 5 и 10 у пациента P10. Б. Результаты секвенирования ПЦР продуктов на стыках 5 и 10 хромосом позволили с точностью до нуклеотида реконструировать границу транслокации у пациента P10. В и Г. Общий вид полученных от пациента P10 ИПСК (В) и их дифференцированных производных (Г). Д. Анализ экспрессии ключевых маркеров плюрипотентности в одном из полученных клонов ИПСК. Е. Анализ экспрессии гена *DKK1* в клетках первичной полоски, полученных путем дифференцировки ИПСК пациента P10 и контрольного фенотипически здорового донора. Ж. Инверсионный паттерн, детектируемый на Hi-C-картах образцов P10 без обогащения (сверху) и после обогащения (снизу). 3. Схема инвертированного участка. И. Скриншот IGV-браузера, демонстрирующий выравнивания RNA-seq ридов в области интрона гена *MCU*. Можно отметить, что в контрольных ИПСК транскрипция в интроне не наблюдается, в то время как у пациента P10 обнаружен кластер ридов, вероятнее всего образовавшихся в результате альтернативного сплайсинга химерного

гена *KAT6B-OIT3-MCU*. К. Количественный анализ транскриптов, содержащих аллельные варианты A/G в восьмом экзоне гена *KAT6B*.

Случай 2. Пациент Р47 обратился в НИИ Медицинской Генетики (группа И.Н. Лебедева) в связи с жалобами на множественные экзостозы, низкорослость. По результатам кариотипирования был выявлен кариотип 46, XY, t(8;11;22)(q23;q22;q21). Нами был проведен Ехо-С-анализ, в результате которого были найдены границы всех цитологически-видимых перестроек с точностью несколько тысяч п.о. и, кроме того, обнаружена инсерция участка 8ой хромосомы между фрагментами 11ой и 21ой хромосом (рис. 17, А, Б). Цитологически эта инсерция не наблюдалась, вероятнее всего, из-за небольшого размера инсцированного участка (8.5 МБ). Важно, что в найденной при Ехо-С-анализе точке инсерции фрагмента 8ой хромосомы находится *EXT1*. Разрыв проходит внутри протяженного первого интрона этого гена, таким образом в результате перестройки ген оказывается разорван и не может кодировать функциональный транскрипт. Мутации в гене *EXT1* приводят к формированию множественных экзостозов, эта патология наследуется по аутосомно-доминантному типу. Таким образом, найденная по результатам Ехо-С-анализа позволяет полностью объяснить фенотип пациент и установить молекулярных механизм развития заболевания.

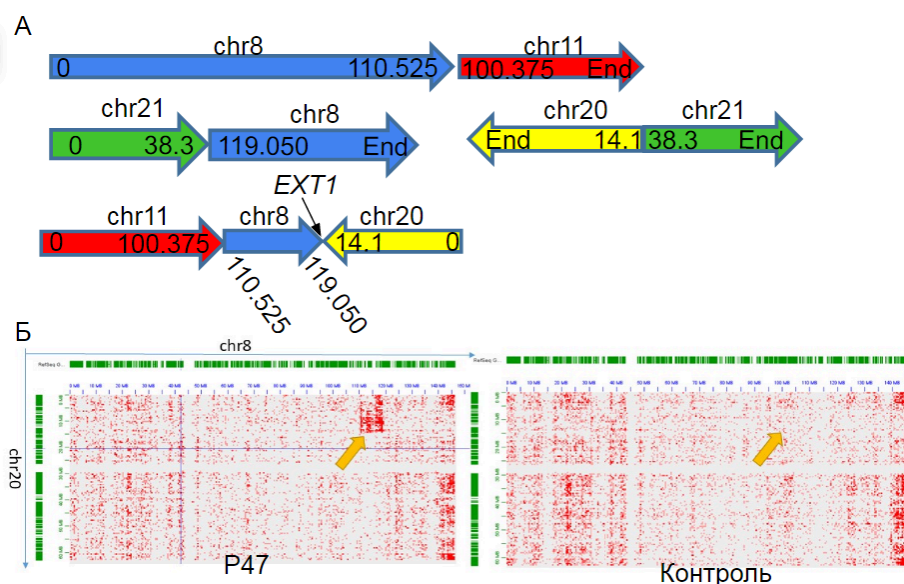


Рисунок 17. Анализ хромосомных перестроек у пациента Р47. Схема четырех хромосомных дериватов, полученная на основе Ехо-С-данных (А), а также фрагмент карты межхромосомных контактов, подтверждающих контакты фрагмента хромосомы 8 с хромосомой 20 (Б) у пациента Р47.

2.5. Применение методов геномики и биоинформатики для решения задач в области генетики развития и медицинской генетики

Наша работа строилась вокруг исследований трехмерной архитектуры ядра и использования методов геномики для изучения роли укладки хроматина в норме и при развитии различных патологий. Однако созданные и освоенные нами подходы оказались востребованы и в смежных областях генетики и геномики. Так, разработанные нами инструменты анализа NGS-данных³⁵ применялись для создания новых подходов к изучению репарационных событий^{36,37}, репрограммированию клеток³⁸, геномной локализации трансгенных конструкций^{39–41}. Вычислительные методы транскриптного анализа и экспериментальные подходы, позволяющие выделять РНК из труднодоступных тканей²¹ оказались востребованы для сборки транскриптомов немодельных видов животных^{7,22} и аннотации транскрипционных изменений в ходе трансдифференцировки³⁴ и перепрограммирования²³. Развитие методов машинного обучения позволило создать ряд важных инструментов для анализа больших медико-генетических данных, связанных не только с организацией укладки ДНК в ядре, но и другими механизмами реализации генетической информации⁴². Наконец, разработка цитологических методов для детекции ДНК-ДНК и ДНК-РНК взаимодействий в ядре клетки, хотя и не оказалась востребована в наших работах в области 3D-геномики, позволила апробировать новых технологии для детекции нуклеиновых кислот, которые могут быть в дальнейшем использованы для изучения биологии РНК^{43,44}. Разработка методов для детекции и интерпретации геномных вариантов на основе технологий массового параллельного секвенирования, реализованных в виде программных пакетов FastContext, Exoclasma и ExoC, позволила описать несколько уникальных генетических вариантов, связанных с развитием генетических патологий^{33,45–47}. В целом, разработка комплекса экспериментальных

методов и биоинформатических инструментов позволила создать на базе ИЦиГ СО РАН центр компетенций в области геномных технологий, который может быть в будущем использован для решения широкого круга биотехнологических и молекулярно-генетических задач.

3. ЗАКЛЮЧЕНИЕ И ВЫВОДЫ

Последнее десятилетие ознаменовалось существенным прогрессом в изучении укладки ДНК в ядре, во многом связанным с появлением новых, высокопроизводительных и доступных экспериментальных и биоинформатических методов. Благодаря нашей работе и смежным исследованиям становится понятно, насколько консервативны механизмы укладки и структуры хроматина между различными типами клеток и видами животных. Становится очевидной роль этих механизмов в различных геномных процессах, в первую очередь - регуляции транскрипции генов. Однако первый энтузиазм, связанный с детальным описанием наднуклеосомного уровня организации ДНК, сменяется пониманием сложности взаимосвязей между различными эпигенетическими процессами и укладкой хроматина. Зачастую, как в самих процессах, так и в их эволюционных изменениях, трудно определить причинно-следственные отношения. Так, до сих пор остается дискуссионным вопрос о взаимосвязях между процессами сближения промоторов и энхансеров и активацией транскрипции. Не до конца понятна природа колокализации границ топологических доменов и границ участков распространения эпигенетических меток. В наших работах, на примере исследования колокализации границ синтенных блоков и границ ТАДов у насекомых, показано, что корреляции в расположении структур 3D-генома и эпигенетических меток могут объясняться не прямыми причинно-следственными связями, но и наличием общего конфаундера, которым является распределение генов. Кроме того, наши работы по предсказанию укладки хроматина у млекопитающих говорят о значительной роли геномного расстояния в установлении профиля контактов между регуляторными элементами, в то время как формирование ТАДов и других структур хроматина объясняет существенно меньшую

долю дисперсии в распределении частот контактов ДНК. Все это свидетельствует о том, что трехмерную организацию хроматина следует считать лишь одним из элементов сложной эпигенетической машины, регулирующей развитие и функционирование клеток. В некоторых случаях, вклад этого уровня регуляции может быть небольшим, но для отдельных процессов роль трехмерной укладки генома является решающей, а ее нарушения могут приводить к развитию патологий.

Сложность взаимосвязей между разными уровнями эпигенетической регуляции подчеркивает значимость второй части нашей работы, в которой мы показываем перспективу современных методов машинного обучения для интерпретации и предсказания геномных характеристик. Предложенные нами методы только начинают использоваться в геномных исследованиях, однако уже очевидно, что во многих практических задачах, а иногда и в фундаментальных исследованиях, они могут заменить классические статистические подходы. Например, предложенная нами концепция клеточных состояний позволяет “оцифровать” совокупность эпигенетических процессов в каждом типе клеток. Это позволяет перейти на совершенно новый уровень изучения процессов клеточной дифференцировки и развития в целом. В диссертации представлена только базовая идея такого подхода, которая ещё нуждается как в доработке технических нюансов, так и в масштабировании и применении к данным описывающим экспрессию единичных клеток. Развитие методов машинного обучения, безусловно, будет одним из перспективных трендов геномных и эпигеномных исследований в ближайшем будущем.

Несмотря на то, что наша работа изначально была нацелена на изучение фундаментальных вопросов, в первую очередь - на изучение механизмов регуляции генома, разработанные методы нашли применения для решения важных практических задач: сборки геномов немодельных видов животных и поиска хромосомных перестроек у человека. Позитивные результаты, полученные с использованием этих методов, позволяют надеяться, что они будут успешно использованы не только в

фундаментальных работах, но внедряться в практику медико-генетических анализов в России и за рубежом.

В совокупности, полученные результаты позволили нам сделать следующие выводы.

1. С использованием технологии захвата конформации хромосом впервые получены полногеномные карты контактов хроматина фибробластов и эритроцитов курицы. Сравнительный анализ частот трехмерных контактов генома в фибробластах птиц и млекопитающих показал, что большинство локус-специфических особенностей укладки хроматина в этих клетках сохраняется неизменным на протяжении сотен миллионов лет. При этом контакты хроматина на масштабах от десятков до сотен тысяч пар оснований в значительной степени обусловлены распределением сайтов связывания белка CTCF.

2. Проведенный анализ трехмерной архитектуры хроматина в эритроцитах курицы и их предшественниках позволил впервые выявить ряд особенностей, связанных с потерей CTCF-опосредованных топологически ассоциированных доменов и усиленной компартментализацией генома. Эти черты строения ядра эритроцитов являются эволюционно-консервативными среди позвоночных животных.

3. Впервые получены данные об организации хроматина для пяти видов комаров рода *Anopheles*: *An. coluzzii*, *An. merus*, *An. stephensi*, *An. atroparvus* и *An. albimanus*, включающие профили контактов ДНК, транскрипционной активности и распределения гистоновых меток. Полученные данные позволили выявить:

1) сближение центромерных и теломерных участков разных хромосом, что соответствует известной модели конфигурации хромосом «по Раблю»;

2) присутствие компартментов А и В, распределение которых в геномах комаров, как у дрозофилы и позвоночных, связано с транскрипционной активностью: компартмент А характеризуется более высоким уровнем транскрипционной активности, чем компартмент В;

3) разделение генома на ТАДы со средним размером 130 тысяч пар оснований и формирование небольшого числа протяженных ТАДов свыше миллиона пар

оснований, включающих участки интеркалярного гетерохроматина;

4) формирование петель ДНК, включающих эволюционно-консервативные контакты между парами локусов, расположенными в хромосоме X на расстоянии до 10 миллионов пар оснований друг от друга;

5) высокий уровень межвидового сходства контактов хроматина внутри синтенных участков генома.

4. Анализ 3026 эпигенетических профилей и соответствующих геномных последовательностей инструментом DeepCT, основанном на методах машинного обучения, позволил выявить клеточно-специфические особенности геномной регуляции для 868 типов клеток человека и представить эти особенности в виде векторов в многомерном пространстве. Использование векторных представлений клеточно-специфических особенностей геномной регуляции позволяет предсказать распределение более чем 300 тысяч ранее не измеренных эпигенетических характеристик клетки и влияние геномных вариантов на связывание белков хроматина с ДНК.

5. Разработана оригинальная технология Echo-C, основанная на совмещении экзомного секвенирования (обогащения) и захвата конформации хромосом, которая позволяет одновременно детектировать короткие нуклеотидные варианты в экзоне человека и хромосомные перестройки. На выборке из 66 образцов клеток человека показано, что технология Echo-C позволяет выявлять хромосомные перестройки размером более 100-500 тысяч пар оснований с чувствительностью, превышающей 50%, независимо от того, проходят ли границы хромосомных перестроек внутри генов или в межгенных участках генома.

6. Создана компьютерно-экспериментальная платформа для функциональной характеристики хромосомных перестроек, включающая определение границ перестройки методом Echo-C, получение пациент-специфических клеточных моделей патологии - набор линий, индуцированных плюрипотентных стволовых клеток, направленную их дифференцировку и аллель-специфический анализ генной экспрессии в дифференцированных производных. Позитивные результаты применения

этой платформы были продемонстрированы при анализе двух пациентов: установлена причинно-следственная связь между инверсией на хромосоме 10 и нарушением структуры гена *KAT6B*, ответственного за формирование множественных аномалий развития, а также выявлено изменение структуры гена *EXT1*, ответственного за развитие множественных экзостозов костной ткани, вследствие комплексной транслокации с участием хромосомы 8. Таким образом, открывается перспектива использования предложенного метода функциональной характеристики хромосомных перестроек в практической медицине.

7. Разработана компьютерно-экспериментальная платформа для анализа данных высокопроизводительного секвенирования геномных библиотек, приготовленных по технологии захвата конформации хромосом (Hi-C). Платформа состоит из семи модулей: FastContext, Exoclasma, Exo-C, ABCE, 3DPredictor, 3DGenBench и C-InterSecture, специализированных для решения различных задач:

А) модули FastContext, Exoclasma и Exo-C позволяют оценить качество Hi-C-данных, выявить контаминации специфическими нуклеотидными последовательностями, детектировать и аннотировать однонуклеотидные варианты в геноме человека;

Б) модуль ABCE дает возможность выделять геномные компартменты в случаях, когда организацию хромосом в ядре соответствует “конфигурации Рабля”;

В) модуль C-InterSecture включает набор инструментов для межвидового сравнения Hi-C-данных;

Г) модули 3DPredictor и 3DGenBench позволяют предсказывать профиль трехмерных контактов ДНК млекопитающих на основе эпигенетических характеристик генома и оценивать точность предсказаний.

В совокупности, разработанные инструменты дают возможность проводить статистический, филогенетический и функциональный анализ карт трехмерных контактов хроматина у животных.

БЛАГОДАРНОСТИ

Автор выражает глубокую признательность коллегам, совместно с которыми были выполнены представленные в этой работе исследования: группам Н.Р. Баттулина (ИЦиГ СО РАН), И.Н. Лебедева (Томский НИМЦ), А.В. Красиковой (СПбГУ), Н.В. Шиловой (МГНЦ), Ольги Кардымон (AIRI), Е.В. Шахтшнейдер (ИЦиГ), Igor Sharakov (Virginia Tech), Michael Bader и Natalia Alenina (MDC), Ana Krepishi (USP), за глубокие научные дискуссии и деятельное участие в совместных исследовательских проектах. Сотрудникам отдела молекулярных механизмов онтогенеза ИЦиГ СО РАН, на базе которого выполнялась большая часть работ, представленных в диссертации, за создание теплой атмосферы на рабочем месте, энтузиазм и интерес к совместным работам. Администрации ИЦиГ СО РАН, за поддержку и организацию комфортных условий для работы. Научному консультанту работы, Олегу Леонидовичу Серову, за продемонстрированный пример приверженности ценностям науки и за возможность реализовывать исследовательские проекты на базе его лаборатории. Рецензентам и оппонентам этой работы, за их время и усилия, потраченные, чтобы сделать эту работу лучше. Своим друзьям, родным и близким, за неустанную поддержку во всех начинаниях и за предоставленную возможность уделять большое количество времени научной работе.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
2. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
3. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
4. Battulin, N. *et al.* Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol.* **16**, 77 (2015).
5. Kabirova, E. *et al.* Function and Evolution of the Loop Extrusion Machinery in Animals. *Int. J. Mol. Sci.* **24**, 5017 (2023).
6. Belokopytova, P. & Fishman, V. Predicting Genome Architecture: Challenges and Solutions.

Front. Genet. **11**, (2021).

7. Lukyanchikova, V. *et al.* Anopheles mosquitoes reveal new principles of 3D genome organization in insects. *Nat. Commun.* **13**, 1960 (2022).
8. Mozheiko, E. A. & Fishman, V. S. Detection of Point Mutations and Chromosomal Translocations Based on Massive Parallel Sequencing of Enriched 3C Libraries. *Russ. J. Genet.* (2019) doi:10.1134/S1022795419100089.
9. Fishman, V. S., Salnikov, P. A. & Battulin, N. R. Interpreting Chromosomal Rearrangements in the Context of 3-Dimensional Genome Organization: A Practical Guide for Medical Genetics. *Biochem. Biokhimiia* **83**, 393–401 (2018).
10. Nuriddinov, M. & Fishman, V. C-InterSequire-a computational tool for interspecies comparison of genome architecture. *Bioinformatics* **35**, 4912–4921 (2019).
11. Kowalski, A. & Palyga, J. Chromatin compaction in terminally differentiated avian blood cells: The role of linker histone H5 and non-histone protein MENT. *Chromosome Res.* **19**, 579–590 (2011).
12. Fishman, V. *et al.* 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. *Nucleic Acids Res.* **47**, 648–665 (2019).
13. Maslova, A. *et al.* Hi-C analysis of genomic contacts revealed karyotype abnormalities in chicken HD3 cell line. *BMC Genomics* **24**, 66 (2023).
14. Ulianov, S. V. *et al.* Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* **26**, 70–84 (2016).
15. Kaushal, A. *et al.* CTCF loss has limited effects on global genome architecture in *Drosophila* despite critical regulatory functions. *Nat. Commun.* **12**, 1011 (2021).
16. Liao, Y., Zhang, X., Chakraborty, M. & Emerson, J. J. Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Res.* **31**, 397–410 (2021).
17. Ghavi-Helm, Y. *et al.* Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282 (2019).
18. Renschler, G. *et al.* Hi-C guided assemblies reveal conserved regulatory topologies on X and autosomes despite extensive genome shuffling. *Genes Dev.* **33**, 1591–1612 (2019).
19. Hoencamp, C. *et al.* 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science* **372**, 984–989 (2021).
20. King, T. D. *et al.* Recurrent Losses and Rapid Evolution of the Condensin II Complex in Insects. *Mol. Biol. Evol.* **36**, 2195–2204 (2019).
21. Ivanov, A. A. *et al.* Method for the Isolation of 'RNA-seq-Quality' RNA from Human Intervertebral Discs after Mortar and Pestle Homogenization. *Cells* **11**, 3578 (2022).
22. Menzorov, A. G. *et al.* Comparison of American mink embryonic stem and induced pluripotent stem cell transcriptomes. *BMC Genomics* **16**, S6 (2015).
23. Matveeva, N. M. *et al.* Alternative dominance of the parental genomes in hybrid cells generated through the fusion of mouse embryonic stem cells with fibroblasts. *Sci. Rep.* **7**, (2017).

24. Berthelot, C., Muffato, M., Abecassis, J. & Roest Crollius, H. The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep.* **10**, 1913–1924 (2015).
25. Gridina, M. & Fishman, V. Multilevel view on chromatin architecture alterations in cancer. *Front. Genet.* **13**, 1059617 (2022).
26. Belokopytova, P. S., Nuriddinov, M. A., Mozheiko, E. A., Fishman, D. & Fishman, V. Quantitative prediction of enhancer–promoter interactions. *Genome Res.* **30**, 72–84 (2020).
27. Belokopytova, P. *et al.* 3DGenBench: a web-server to benchmark computational models for 3D Genomics. *Nucleic Acids Res.* **50**, W4–W12 (2022).
28. Sindeeva, M. *et al.* Cell type-specific interpretation of noncoding variants using deep learning-based methods. *GigaScience* **12**, giad015 (2023).
29. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).
30. Миньженкова, М. Е. *et al.* Комплексные геномные перестройки в этиологии ‘хромосомного фенотипа’. *Медицинская Генетика* **21**, 44–47 (2022).
31. Gridina, M. *et al.* A cookbook for DNase Hi-C. *Epigenetics Chromatin* **14**, 15 (2021).
32. Gridina, M. *et al.* Expanding the list of sequence-agnostic enzymes for chromatin conformation capture assays with S1 nuclease. 2023.06.15.545138 Preprint at <https://doi.org/10.1101/2023.06.15.545138> (2023).
33. Gridina, M. M. *et al.* Allele-Specific Biased Expression of the CNTN6 Gene in iPS Cell-Derived Neurons from a Patient with Intellectual Disability and 3p26.3 Microduplication Involving the CNTN6 Gene. *Mol. Neurobiol.* **55**, 6533–6546 (2018).
34. Fishman, V. *et al.* Cell divisions are not essential for the direct conversion of fibroblasts into neuronal cells. *Cell Cycle* **14**, 1188–1196 (2015).
35. Viesná, E. & Fishman, V. FastContext: A tool for identification of adapters and other sequence patterns in next generation sequencing (NGS) data. *Vavilovskii Zhurnal Genet. Sel.* **26**, 806–809 (2022).
36. Smirnov, A. *et al.* DNA barcoding reveals that injected transgenes are predominantly processed by homologous recombination in mouse zygote. *Nucleic Acids Res.* **48**, 719–735 (2020).
37. Смирнов, А. В. *et al.* Создание библиотек баркодированных плазмид с помощью метода клонирования по Гибсону. *Письма В Вавиловский Журнал Генетики И Селекции* **7**, 34–45 (2021).
38. Yunusova, A. M., Fishman, V. S., Vasiliev, G. V. & Battulin, N. R. Deterministic versus stochastic model of reprogramming: new evidence from cellular barcoding technique. *Open Biol.* **7**, 160311 (2017).
39. Salnikov, P. A. *et al.* Here and there: the double-side transgene localization. *Vavilov J. Genet. Breed.* **25**, 607–612 (2021).
40. Khabarova, A. *et al.* A Cre-LoxP-based approach for combinatorial chromosome rearrangements in human HAP1 cells. *Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol.* **31**, 11 (2023).
41. Pristyazhnyuk, I. E. *et al.* Time origin and structural analysis of the induced CRISPR/cas9

- megabase-sized deletions and duplications involving the *Cntn6* gene in mice. *Sci. Rep.* **9**, 14161 (2019).
42. Filatova, A. *et al.* Annotation of uORFs in the OMIM genes allows to reveal pathogenic variants in 5'UTRs. *Nucleic Acids Res.* **51**, 1229–1244 (2023).
43. Krasheninina, O. A., Lomzov, A. A., Fishman, V. S., Novopashina, D. S. & Venyaminova, A. G. Rational design and studies of excimer forming novel dual probes to target RNA. *Bioorg. Med. Chem.* (2017) doi:10.1016/j.bmc.2017.02.042.
44. Krasheninina, O. A., Fishman, V. S., Lomzov, A. A., Ustinov, A. V. & Venyaminova, A. G. Postsynthetic On-Column 2' Functionalization of RNA by Convenient Versatile Method. *Int. J. Mol. Sci.* **21**, 5127 (2020).
45. Kuppusamy, M. *et al.* A Novel KCNJ5-insT149 Somatic Mutation Close to, but Outside, the Selectivity Filter Causes Resistant Hypertension by Loss of Selectivity for Potassium. *J. Clin. Endocrinol. Metab.* jc20141927 (2014) doi:10.1210/jc.2014-1927.
46. Kashevarova, A. A. *et al.* A mosaic intragenic microduplication of LAMA1 and a constitutional 18p11.32 microduplication in a patient with keratosis pilaris and intellectual disability. *Am. J. Med. Genet. A.* **176**, 2395–2403 (2018).
47. Ivanoshchuk, D. E. *et al.* The Mutation Spectrum of Maturity Onset Diabetes of the Young (MODY)-Associated Genes among Western Siberia Patients. *J. Pers. Med.* **11**, 57 (2021).