ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
УЧРЕЖДЕНИЕ НАУКИ
ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ
СИБИРСКОГО ОТДЕЛЕНИЯ РАН

На правах рукописи

ФИШМАН ВЕНИАМИН СЕМЕНОВИЧ

Сравнение пространственной организации геномов фибробластов и сперматозоидов мыши методом Hi-C

03.02.07 – генетика

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель:
доктор биологических наук,
профессор Серов Олег Леонидович

Новосибирск 2015

Оглавление

Список использованных сокращений

п.о. – пара оснований

Hi-C - High-Throughput Chromosome conformation capture, высоко эффективный захват конформации хромосом

нм – нанометры

мкм – микрометры

FISH - Fluorescent *in situ* Hybridisation, флуоресцентная гибридизация *in situ*

ЛАД - ламин-ассоциированный домен

3C - chromosome conformation capture, метод захвата конформации хромосом

4C - Circularized Chromosome Conformation Capture, метод «кольцевого» захвата конформации хромосом

ChIA-PET - Chromatin Interaction Analysis by Paired-End Tag Sequencing, анализ взаимодействий хроматина на основе секвенирования спаренных концов

TAD, TAD-домены - topologically associated domains, топологические ассоциированные домены, или просто топологические домены

ЭСК - эмбриональные стволовые клетки

kb – “kilobase”, тысяча пар оснований

Mb – “megabase”, один миллион пар оснований

Введение

Актуальность

Создание высоко производительных технологий секвенирования геномов позволило получить исчерпывающую информацию о первичной структуре ДНК многих видов млекопитающих и человека (Hattori, 2005). Зная последовательность геномной ДНК, исследователь может с высокой точностью предсказать аминокислотную последовательность большинства белков и, используя данные сравнительной геномики, охарактеризовать структуру и даже функции этих белков. Однако геном не просто представляет собой информацию о структуре белков и РНК, но и обладает функцией избирательной реализации этой информации. Активность генома сложно и точно регулируется, обеспечивая работу тех или иных генов только в определенных органах, тканях или типах клеток, на точно заданном уровне и в строго заданный момент времени. Жизнедеятельность организма определяется и тем, какую генетическую информацию он несет, и тем, как регулируется реализация этой информации. И если в эпоху массового секвенирования геномов расшифровка составляющей их нуклеотидной последовательности не представляет большой сложности, то наше понимание механизмов регуляции и функционирования геномов ещё далеко от полного.

С биологической точки зрения, регуляция функционирования генома организована очень сложно и осуществляется множеством разных механизмов, действующих на разных уровнях и, зачастую, сложно связанных между собой. К настоящему времени, достаточно хорошо изучены базовые механизмы транскрипционной и трансляционной активности генома, но менее понятно значение эпигенетических модификацией геномной ДНК, например, метилирования (Smith et al., 2013) или роль разнообразных модификаций белков хроматина в функционировании генома (Sneppen et al., 2012).

Однако, даже понимание этих механизмов не дает ответ на целый ряд фундаментальных и прикладных вопросов. Как осуществляется взаимодействие между генами и удаленными от них на многие миллионы нуклеотидов регуляторными элементами? Как эти регуляторные элементы «находят» именно

«свой» ген и почему, при этом, оказывают слабое влияния на активность близлежащих генов? Почему удаление крупных, некодирующих фрагментов генома, не содержащих известных регуляторных областей, приводит к аномалиям (Lupiáñez et al., 2015) и даже остановке развития?

Попытки ответить на эти и сходные вопросы, позволили показать связь между пространственной организацией ДНК в ядре и функционированием генома. Хотя известно, что пространственная конформация небольших фрагментов ДНК (до 200 п.о.; размер персистенции порядка 35-50 нм) определяется её первичной структурой (Brinkers et al., 2009), большие по размеру молекулы ДНК способны с равной вероятностью формировать огромное множество пространственных (конформационных) структур. Несмотря на такое потенциальное разнообразие, в геномах эукариот существуют некие предпочтительные варианты укладки молекулы ДНК, характеризующиеся определенной архитектурой (Cremer et al., 2001a). Более того, была показана связь пространственной организации определенных участков генома и их активности в клетке (Cremer et al., 2001a).

В последнее десятилетие был разработан ряд молекулярно-биологических методов, позволивших значительно расширить и детализировать информацию о пространственной укладке ДНК. Среди таких методов следует особо выделить Hi-C (High-Throughput Chromosome Conformation Capture, высоко эффективный захват конформации хромосом), позволяющий в одном эксперименте получить данные о пространственной укладке всего генома с высоким разрешением (Lieberman-Aiden et al., 2009). Используя этот и подобный методы было показано, что трехмерная архитектура генома действительно является важнейшим уровнем регуляции большинства базовых молекулярно-генетических процессов: транскрипции, репликации, молекулярной эволюции и т.д. (Баттулин и др., 2012; de Wit et al., 2012; Vietri Rudan et al., 2015; Rao et al., 2014).

Между процессами, которые регулируются пространственной структурой генома, и самой пространственной структурой, существуют сложные прямые и обратные связи. Например, пространственная организация генома влияет на транскрипцию, поскольку, за счет формирования крупномасштабных петель ДНК, обеспечивается взаимодействие удаленных промоторов и энхансеров (Rao et al.,

2014; Cope et al., 2010). С другой стороны, сам процесс транскрипции связан с поддержанием пространственной структуры генома: активно транскрибирующиеся гены участвуют в формировании пространственных доменов (Dixon et al., 2012); модификации хроматина, связанные с активацией или репрессией транскрипции, изменяют положение соответствующих участков ДНК относительно центра и периферии ядра (Therizols et al., 2014), и так далее (Cope et al., 2010).

Помимо функции регуляции, у компактизации ДНК, очевидно, есть прямая функция – организация генома в пространстве ядра. Задача размещения ДНК в пространстве ядре является сложной – необходимо упаковать геном млекопитающих, линейный размер которого составляет более метра, внутрь ядра размером в несколько десятков или сотен микрометров. Условия этой задачи – соотношение параметров геометрии ядра и размера генома - сходны практически для всех типов клеток организма млекопитающих. Однако ряд клеток является особенными с точки зрения пространственной организации генетического материала в ядре – это клетки сперматозоидов. Во-первых, ДНК сперматозоидов упакована протаминами, тогда как ДНК других типов клеток млекопитающих упакована гистонами (Balhorn et al., 1999). Во-вторых, размер ядра сперматозоида на порядок меньше, чем ядер соматических клеток (Lee et al., 1997). По этому параметру, уровень конденсации ДНК в сперматозоидах сходен с конденсацией ДНК в митотической хромосоме. В-третьих, в зрелых сперматозоидах отсутствует процесс транскрипции (Mudrak et al., 2011). Как было указано выше, существует прямая и обратная связь между транскрипцией и пространственной организацией генома. Поэтому, пространственная организация транскрипционно-неактивных клеток представляет особый интерес. Наконец, в-четвертых, основной функцией сперматозоидов является передача генетической информации от родителей к потомкам. Если механизм передачи первичной структуры генома сперматозоидами известен, то происходит ли передача информации о трехмерной организации геномной ДНК, неизвестно.

Учитывая значение трехмерной организации для регуляции активности генома, а также вышеперечисленные особенности сперматозоидов, несомненно, актуальным представляется исследование пространственной организации генома половых

клеток и сравнение пространственной укладки ДНК сперматозоидов и соматических клеток, например, фибробластов.

Цели и задачи исследования

Целью данной работы является сравнение пространственной организации геномов сперматозоидов и фибробластов мыши

Для выполнения данной цели были поставлены следующие задачи.

1. Построить карту пространственных контактов геномной ДНК сперматозоидов и фибробластов мыши.
2. Провести анализ и сравнение пространственных доменов геномов фибробластов и сперматозоидов.
3. Сравнить пространственную укладку и частоту контактов индивидуальных локусов в этих типах клеток.
4. Оценить зависимость частоты контактов участков генома от их удаленности в линейной молекуле ДНК для сперматозоидов и фибробластов.
5. Оценить влияние компактизации генома сперматозоидов на особенности укладки ДНК этих клеток.

Научная новизна работы

На основе новейшего метода Hi-C впервые получены пространственные карты геномов фибробластов и сперматозоидов мыши. Важно отметить, что в данной работе впервые исследованы пространственные контакты транскрипционно-неактивных клеток. Впервые показано наличие пространственных доменов в геномах этих клеток. В работе проведено сравнение трехмерной организации геномов соматических и половых клеток с использованием как ранее описанных, так и оригинальных, разработанных автором методов. Впервые получен список локусов генома сперматозоидов, пространственная укладка которых наиболее значительно отличается от укладки соматических клеток. Более того, в ходе работы разработана методика нормализации частот контактов, учитывающая эффект компактизации генома. Разработанный автором алгоритм статистического сравнения частот контактов, в сочетании с методикой нормализации, позволил адекватно оценить влияние компактизации генома сперматозоидов и других особенностей этих клеток

на специфику трехмерной организации ДНК.

Теоретическая и практическая значимость исследования

С теоретической точки зрения, анализ пространственной архитектуры генома сперматозоидов способствует расширению наших знаний о таких фундаментальных характеристиках эукариотического генома, как роль процессов транскрипции в трехмерной организации ДНК, влияние компактизации генома на структуру пространственных доменов и влияние белков-упаковщиков ДНК (таких как гистоны и протамины) на организацию этой молекулы на макроуровне.

Работа расширяет список типов клеток мыши, для которых были получены полногеномные карты пространственных контактов. На сегодняшний день, такие карты получены только для эмбриональных стволовых клеток (ЭСК), клеток кортекса (Dixon et al., 2012) и клеток печени (Vietri Rudan et al., 2015) мыши. Поэтому, исследование двух новых типов клеток, фибробластов и сперматозоидов, и их сравнение с ранее изученными клетками, позволяет лучше понять связь клеточной специализации и пространственной структуры генома.

Понимание теоретических механизмов, лежащих в основе формирования и поддержания структуры пространственных доменов, помогает объяснить причины заболеваний, связанных с нарушениями пространственной укладки генома и, в перспективе, может способствовать разработкам методов их прогнозирования и лечения.

Положения, выносимые на защиту

На защиту выносятся следующие положения и результаты:

1.Общность принципов пространственной организации геномов соматических и половых клеток выражается в характерной для обоих типов клеток степенной зависимости распределения частот контактов локусов от расстояния между ними в линейной молекуле и наличии топологических доменов в геномах фибробластов и сперматозоидов

2. Особенностью пространственной укладки генома сперматозоидов является увеличение частот как внутрихромосомных контактов между удаленными локусами ДНК, так и межхромосомных взаимодействий.

3. Различия в частотах пространственных контактов в геномах фибробластов и сперматозоидов в 25% случаев объясняются более высокой компактизацией генома последних, а в 75% – другими причинами.

Вклад автора

Автором самостоятельно получены основные результаты: построены и проанализированы матрицы пространственных контактов геномов фибробластов и сперматозоидов, идентифицированы и проанализированы пространственные домены, проведено сравнение пространственной укладки и частоты контактов индивидуальных локусов в геномах половых и соматических клеток, разработана модель «компрессии» генома, оценена роль компактизации генома сперматозоидов в возникновении различий половых и соматических клеток.

Hi-C библиотеки фибробластов и сперматозоидов мыши были получены Баттулиным Н.Р. и Хабаровой А.А., при участии автора. Алгоритм анализа зависимости частоты взаимодействий от расстояния между локусами в линейной молекуле был предложен Помазным М.Ю. и адаптирован при участии автора.

Апробация работы

Работа была доложена на следующих научных конференциях:

1. N. Battulin, **V.S. Fishman**, A.M. Mazur, M. Pomaznoy, A.A. Khabarova, D.A. Afonnikov, E.B. Prokhortchouk , O.L. Serov. **Comparison of the 3D organization of sperm and fibroblast genomes using the Hi-C approach**, IV Международная научно-практическая конференция «Постгеномные методы анализа в биологии, лабораторной и клинической медицине». Казань, Россия, 29 Октября – 1 Ноября, 2014
2. N. Battulin, **V.S. Fishman**, A.M. Mazur, M. Pomaznoy, A.A. Khabarova, D.A. Afonnikov, E.B. Prokhortchouk, O.L. Serov. **Comparison of the 3D-organizaitoin of sperm and fibroblasts genomes by Hi-C approach**, Международная биомедицинская конференция “Терапия будущего”. Сколтех, Россия, 26 - 28 мая 2014.

По материалам работы опубликованы следующие статьи:

1. Н.Р. Баттулин, **В.С. Фишман**., Ю.Л. Орлов, А.Г. Мензоров, Д.А. Афонников, О.Л. Серов **3С-методы в исследованиях пространственной организации генома.** // Вавиловский журнал генетики и селекции. – 2012. – Т 16. - № 4/2. - с 872-878.
2. N.R. Battulin, **V.S. Fishman**, A.A. Khabarova, M.Yu. Pomaznoy, T.A. Shnaider, D.A. Afonnikov, O.L. Serov, **Investigation of the spatial genome organization of mouse sperm and fibroblasts by the Hi-C method** // Russian Journal of Genetics: Applied Research. – 2014. - V 4. - № 6. - p 556-560.
3. N. Battulin, **V.S. Fishman**, A.M. Mazur, M. Pomaznoy, A.A. Khabarova, D.A. Afonnikov, E.B. Prokhortchouk, O.L. Serov. **Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach** // Genome Biology. – 2015. - V 16. - I 1. - doi: 10.1186/s13059-015-0642-0.

Структура и объем работы

Диссертация состоит из оглавления, списка сокращений, введения, обзора литературы, описания используемых материалов и методов, результатов, обсуждения, выводов и списка литературы. Работа изложена на 113 страницах, содержит 14 рисунков, 2 таблицы и 3 приложения.

Обзор литературы

Пространственная организация ДНК млекопитающих

Несмотря на то, что основой геномов большинства организмов, живущих на земле, является молекула двухцепочечной ДНК, способы её организации значительно различаются. В ходе эволюции, особенно при переходе от бактерий к многоклеточным эукариотам, произошло увеличение числа генов (менее чем на порядок), и огромное – на 2-3 порядка – увеличение размера генома (от миллионов до миллиардов пар нуклеотидов). Увеличение числа нуклеотидов в эукариотическом геноме привело к увеличению физической длины молекул ДНК. Для упаковки и хранения отдельных молекул ДНК в эукариотических клетках появились специальные функциональные структуры – хромосомы, кардинально отличающиеся от бактериальных хромосом.

Помимо решения основной функции – уменьшения длины молекулы ДНК за счет формирования многочисленных петель, белки, участвующие в формировании петель, а также сами петли, приобрели важные регуляторные функции. Практически все клетки многоклеточного организма несут молекулы ДНК с одинаковой первичной последовательностью, однако каждый тип клеток обладает уникальным профилем экспрессии генов. Различия в генной экспрессии обеспечиваются эпигенетическими механизмами – в частности, особенностями пространственной организации ДНК. Таким образом, пространственная укладка ДНК играет роль в таких фундаментальных процессах, как клеточная специализация, дифференцировка и реализации программы развития.

Мы ещё далеки от полного понимания механизмов, связывающих пространственную структуру ДНК и функционирование генома. Известно, что пространственная структура ДНК является многоуровневой – её крупнейшей единицей является хромосома, содержащая миллиарды нуклеотидов, тогда как наименьший единицей можно считать фрагмент ДНК в несколько сотен пар оснований, организованный в пространстве специальным комплексом белков-гистонов - нуклеосомой. На каждом из этих уровней пространственной организации ДНК оказывает влияние на функционирование генов. Однако это влияние

реализуется на каждом уровне по-разному, и, поэтому, требует для изучения различных методологических подходов.

Изучение организации ДНК на нуклеосомном уровне ведется методами молекулярной биологии, в то время как изучение ДНК на хромосомном уровне долгое время было исключительной прерогативой методов световой и электронной микроскопии. Малоизученным являлись промежуточные, субхромосмные уровни организации ДНК – фрагментов длиной от десятков до сотен тысяч пар оснований. В последние десятилетия, в связи с развитием новых молекулярно-биологических методов и технологий массового параллельного секвенирования, эта брешь в наших знаниях постепенно заполняется.

В последующих главах будут кратко представлены основные принципы пространственной организации ДНК на хромосомном и субхромосомном уровнях. Основное внимание будет уделено новейшим данным о принципах организации ДНК на субхромосмном уровне. В ходе описания пространственных структур, будет подчеркиваться их роль в регуляции экспрессии. Кроме того, будут коротко описаны существующие методы построения моделей пространственной организации ДНК в ядрах клеток. Обзор литературы будет завершен главной, описывающей известные на сегодняшний день данные о специфике пространственной организации генома объекта данного исследования – сперматозоидов млекопитающих.

Хромосома как единица пространственной организации ДНК в клетке

Первые исследования пространственного расположения хромосом в клетке на микроскопическом уровне были проведены ещё в конце 19-го века (Rabl, 1885; Boveri, 1909). Эти наблюдения отчетливо указывали на определенное положение хромосомы в ходе митоза, однако, в силу ограничений методов микроскопии, не позволяли однозначно установить пространственную организацию хромосом в ходе интерфазы. Тем не менее, на основании косвенных данных, Теодором Бовери (Boveri, 1909) было сделано предположение о том, что в ходе интерфазы индивидуальные хромосомы занимают определенную часть пространства ядра, не смешиваясь друг с другом. Для описания участка ядра, занимаемого определенной хромосомой, им был введен термин «хромосомная территория».

Наличие хромосомных территорий было подтверждено экспериментально только в 70-ые годы XX века. Например, небольшие участки ядра локально облучались лазером, что приводило к повреждениям ДНК в облученном участке, которые репарировались в присутствии ^{3}H-тимидина (Zorn et al., 1979; Zorn et al., 1976; Cremer et al., 1982a; Cremer et al., 1982b). В большинстве случаев, метку (^{3}H-тимидин) после репарации регистрировали преимущественно в одной из хромосом, что свидетельствует о наличие только одной хромосомы в дискретном, облученном лазером, участке ядра.

Позднее, развитие методов гибридизации *in situ* (Fluorescent *in situ* Hybridisation, FISH) позволило подтвердить наличие хромосомных территорий более надежным, прямым методом (Manuelidis, 1985; Schardin et al., 1985). Суть данного метода заключается в том, что фрагмент ДНК (зонд), меченый флюорофором, гибридизуется с геномной ДНК (Manuelidis, 1985; Schardin et al., 1985). После этого, детектируя сигнал флюорофора методами микроскопии, можно определить локализацию в пространстве участка ДНК, комплементарного зонду.

Использование метода FISH позволило не только показать наличие хромосомных территорий, но и выявить определенные закономерности в их распределении. Оказалось, что богатая генами хромосома 19 человека имеет тенденцию к расположению ближе к центру ядра, по сравнению с обедненной генами хромосомой 18, которая располагается ближе к периферии ядра (Cremer et al., 2003; Croft et al., 1999; Cremer et al., 2001b). Более того, данная особенность расположения хромосом оказалась эволюционно-консервативной: участки, ортологичные хромосоме 18 у приматов также располагались ближе к периферии ядра, в то время как ортологичные хромосоме 19 – ближе к центру (Tanabe et al., 2002).

Связь близости хромосомы к центру ядра и её обогащением генами была показана впоследствии и для других хромосом у грызунов (Mayer et al., 2005; Neusser et al., 2007), парнокопытных (Koehler et al., 2009), и птиц (Habermann et al., 2001). Важно отметить, что речь в вышеприведённых примерах идет уже не только о целых хромосомах, а о расположении отдельных участков хромосом внутри хромосомных территорий. При этом общая закономерность сохраняется: богатые генами регионы располагаются ближе к центру ядра, в то время как бедные генами участки ближе к

периферии (Kupper et al., 2007).

Кроме того, следует отметить, что наблюдается связь между удаленностью того или иного участка хромосомы от центра ядра и рядом факторов, таких как: GC-состав, время репликации и активности генов в данном участке (Mayer et al., 2005; Federico et al., 2006; Goetze et al., 2007; Grasser et al., 2008; Hepperger et al., 2008), а также типом исследуемых клеток (Hepperger et al., 2008). Связь активности генов с расположением кодирующих их участков ближе к центру ядра, а также различия деталей пространственной организации в разных типах клеток (Kosak et al., 2002; Andrulis et al., 1998), позволяют предположить, что организация хромосомных территорий играет определенную роль в регуляции экспрессии. Считается, что взаимодействие хромосом и располагающейся на периферии ядра ядерной ламины сопровождается подавлением экспрессии генов (Cope et al., 2010). Для обозначения протяженных участков хромосом, взаимодействующих с ядреной ламиной, применяется термин ламин-ассоциированный домен (ЛАД).

Существует целый ряд примеров, подтверждающий подавление экспрессии генов, входящих в ЛАД. Например, хорошо изучено подавление транскрипции в тепломерном локусе дрожжей при его контактах с ядерной мембраной, или отдаление от ядерной мембраны Ig-локуса В-лимфоцитов перед его активацией и структурной перестройкой (Kosak et al., 2002; Andrulis et al., 1998). Показано также, что для ряда активно транскрибирующихся локусов, в частности *MHC*, *EDC* или локусов *HOX*-генов, характерно выпетливание их ДНК к центру ядра из областей основных хромосомных территорий.

Недавно было проведено целенаправленное исследование связи транскрипции, компактизации хроматина и положения генов в ядре (Therizols et al., 2014). В этом исследовании было убедительно показано, что активация транскрипции неактивных генов, расположенных в ЛАД, сопровождается деконденсацией хроматина и перемещением соответствующих локусов из периферии ядра в центр. Интересно, что в случае, если исследователи проводили деконденсация хроматина в области тех же генов, не активируя при этом их транскрипцию, также наблюдался эффект перемещения соответствующих локусов к центру ядра (Therizols et al., 2014).

Помимо закономерностей, связанных с расположением хромосом относительно

центра и периферии ядра, в ряде работ отмечаются контакты территорий различных хромосом друг с другом (Brianna Caddle et al., 2007; Khalil et al., 2007). Такие межхромосомные контакты, однако, во-первых являются редкими (т.е. описаны только для некоторых хромосом) и, во-вторых, не являются обязательными, т.е. наблюдаются только в части популяции клеток (Zeitz et al., 2009). Собственно, более корректным представляется интерпретация этого феномена как статистически достоверное увеличение частоты соседства территорий определенных хромосом друг с другом.

Такое представление о хромосомных контактах подтверждается и данными 3C-методов (методов захвата конформации хромосом, **C**hromosome **C**onformation **C**apture, 3C). Не вдаваясь в детали этих методов, подробно разъясняемых ниже, следует отметить, что, по данным Hi-C, число межхромосомных контактов на несколько порядков меньше, чем число внутрихромосмоных. Среди закономерностей, обнаруженных при анализе межхромосомных контактов методом Hi-C, следует отметить следующие: во-первых, короткие аутосомы (номер 10-22 у человека) имеют тенденцию контактировать друг с другом больше, чем с остальными хромосомами (Kalhor et al., 2012). Во-вторых, участки, расположенные на периферии хромосомных территорий, чаще участвуют в межхромосомных контактах, чем участки в центре территории (Kalhor et al., 2012). И, в-третьих, локусы, содержащие активно транскрибируемые гены, чаще участвуют в контактах, чем неактивные участки генома (Kalhor et al., 2012).

Итак, отдельные хромосомы в ядре занимают дискретные участки, хромосомные территории, которые могут контактировать друг с другом, но не перекрываются. Пространственная организация хромосомы в целом подчиняется ряду закономерностей, наиболее явной из которых является расположение обогащенных и обедненных генами участков генома ближе к центру и периферии ядра, соответственно. Расположение в центре или на периферии ядра может играть роль в регуляции генной экспрессии и различаться в зависимости от типа клеток и стадии дифференцировки. Кроме того, специфика пространственной структуры хромосомных территорий может нести и другие функции, не связанные с известными нам механизмами регуляции работы генов (Solovei et al., 2009).

Организация ДНК на субхромосомном уровне

Из описанных выше закономерностей организации пространственной структуры хромосомных территорий естественным образом вытекает предположение о петлевой организации ДНК. Действительно, бедные генами участки хромосомы могут чередоваться с обогащенными генами, при этом один обогащенный генами участок может быть отделен многими миллионами нуклеотидов от другого. Для того чтобы, несмотря на такое удаление в линейной молекуле, в пространстве ядра обогащенные генами участки располагались ближе к центру, чем обедненные генами, логично предположить наличие петель большого (порядка миллиона нуклеотидов) масштаба, сближающих их друг с другом. Наличие петель меньшего масштаба логично предположить для объяснения того, как происходят взаимодействия энхансеров с промоторами генов, удаленных от них на расстояние десятков и сотен тысяч нуклеотидов.

Несмотря на то, что сближение тех или иных участков ДНК может быть детектировано при помощи метода FISH (Osborne et al., 2004), ряд ограничений лимитирует применение этого метода для изучения петлевой организации ДНК. Во-первых, в эксперименте можно визуализировать лишь небольшое количество участков ДНК. Во-вторых, исследование можно провести на ограниченном количестве (порядка нескольких сотен) клеток. В-третьих, пространственное разрешение метода ограниченно. Для того чтобы два локуса были разрешены в ядре клетки, в геноме они должны быть разделены не менее чем 100 тысячами пар нуклеотидов (Gilbert et al., 2004; Morey et al., 2007), и даже применение высокоразрешающей микроскопии, скорее всего, не позволит улучшить информативность данного метода, поскольку необходимость стадии денатурации ДНК ставит вопрос о сохранности наноразмерных хроматиновых структур при приготовлении препаратов для FISH. Альтернативный метод прижизненной локализации ДНК в пространстве ядра был недавно предложен на основе системы CRISPR/Cas9 (Chen et al., 2013; Anton et al., 2014; Ma et al., 2015) однако этот метод также лимитирован низким разрешением и позволяет визуализировать не более чем 3 участка единовременно.

В связи с вышеперечисленными ограничениями, для изучения пространственных контактов ДНК с высоким разрешением в последние десятилетие был разработан ряд молекулярные методов под общим названием «захват конформации хроматина» (методы 3С). Поскольку понимание молекулярных механизмов, лежащие в основе этих методов, важно для правильной интерпретации полученных с их помощью данных, в следующей главе будет подробно разобрана методология 3С, и коротко описаны основные модификации этой базовой методики.

Технология захвата конформации хроматина

Все 3С методики основываются на следующем экспериментальном подходе (рис. 1, А-Г). Сначала, сближенные участки ДНК ковалентно связывают (фиксируют) с окружающими белками и другими макромолекулами, что препятствует изменению имеющейся в клетке пространственной структуры ДНК (Этап 1). За счет взаимодействий с белками, пространственно сближенные фрагменты ДНК оказываются связанными друг с другом в одном мульти-молекулярном комплексе. Затем, проводят расщепление ДНК на фрагменты небольшой длины, каждый из которых по-прежнему ковалентно связан с белками и фрагментами ДНК, находившимися рядом с ними изначально (Этап 2). После этого, концы фрагментов лигируют так, что изначально находившиеся близко друг другу в пространстве ядра фрагменты ДНК оказываются с высокой вероятностью сшитыми друг с другом (Этап 3). Для того, чтобы подчеркнуть, что полученные в результате реакции лигирования молекулы состоят из двух участков ДНК, разделенных в линейной молекуле тысячами или даже миллионами пар оснований, мы будем называть их «гибридными».

Полученный набор фрагментов, который называется 3С-библиотекой, может быть проанализирован различными молекулярными методами, определяющими качественно или количественно присутствие всех или только определенных гибридных фрагментов (Этап 4). Рассмотрим каждый из этапов методики 3С более подробно.

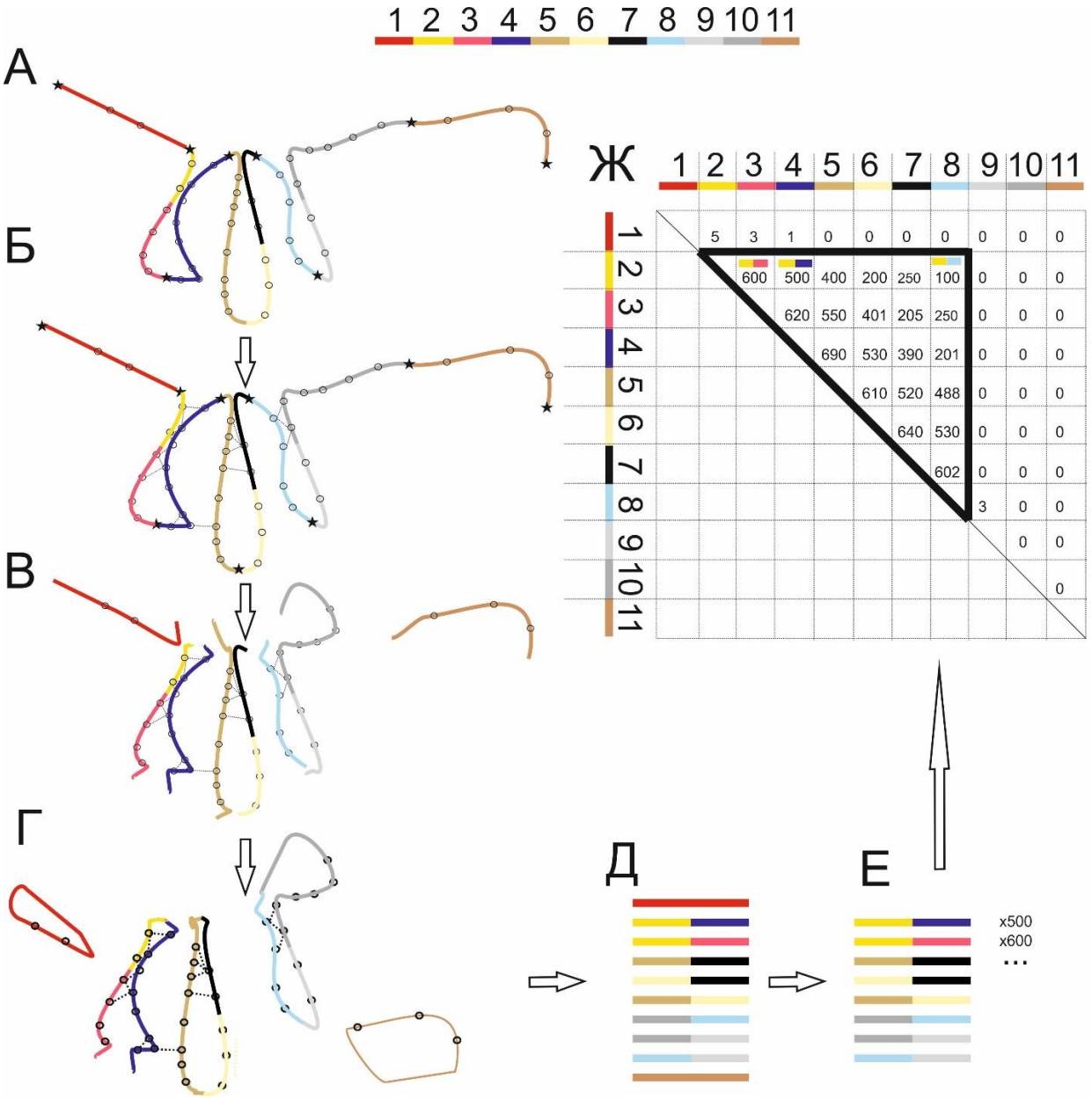


Рисунок 1. Схематичное изображение метода Hi-C. Сверху показан фрагмент линейной молекулы ДНК, различные участки которого показаны разным цветом. А. Трехмерная структура указанного выше участка ДНК в клетке. Звездочкой отмечены сайты рестрикции, кружками изображены белки хроматина. Б. Фиксация формальдегидом комплекса ДНК-белок (этап 1). Ковалентные сшивки между белками показаны пунктиром. В. Гидролиз ДНК ферментом рестрикции (этап 2). Г. Лигирование (этап 3). Д. Библиотека 3C, состоящая из гибридных молекул. Е. Фильтрация полученных результатов секвенирования позволяет удалить продукты лигирования концов одного и того же рестрикционного фрагмента друг с другом и другие неспецифические продукты (обсуждается в главе «Материалы и методы»). Ж. Матрица пространственных контактов, специфическая графическая форма интерпретации результатов Hi-C эксперимента. Треугольником выделен участок ДНК, обогащенный контактами (топологический домен, обсуждается ниже).

1. Фиксация ДНК. В подавляющем большинстве 3C методов используется формальдегид для фиксации хроматин-опосредованных пространственных контактов ДНК (Lieberman-Aiden et al., 2009), хотя использование соединений со сходными химическими свойствами не влияет на результат существенно (Lin et al., 2012). Формальдегид фиксирует белок-белковые, белок-ДНК и белок-РНК-взаимодействия за счет образования ковалентных связей между первичной аминогруппой белка и нуклеиновой кислотой. В 3C экспериментах чаще всего применяется именно формальдегид, поскольку поперечные сшивки, которые он образует, имеют наименьший размер (2 ангстрема) среди других фиксирующих агентов, таких как, например, глутаровый альдегид. Эта особенность позволяет повысить пространственное разрешение метода.

Кроме того, фиксация формальдегидом обратима: ковалентные связи, опосредованные формальдегидом, могут быть разрушены при нагревании. (Jackson, 1999; Dekker et al., 2002; Fujita et al., 2004; Orlando et al., 1997). Эта особенность позволяет удалить формальдегид и ДНК-связывающие белки после стадии лигирования.

2. Фрагментация ДНК. В оригинальной методике 3C, предложенной в 2002 году группой Деккера (Dekker et al., 2002), для фрагментации ДНК использовалась эндонуклеаза рестрикции EcoRI. В этой и последующих (Dixon et al., 2012) работах, было показано, что выбор фермента для гидролиза ДНК не имеет принципиального значения. Однако, следует отметить, что от частоты, с которой встречается в ДНК сайт узнавания фермента, зависит средний размер образующихся фрагментов ДНК. Например, при гидролизе ДНК ферментом рестрикции, узнающим 6 определенных нуклеотидов, можно ожидать фрагменты длиной приблизительно 4000 пар оснований и, следовательно, исследовать пространственные взаимодействия фрагментов ДНК, разделенных в линейной молекуле меньшим количеством нуклеотидов, будет невозможно. Кроме того, необходимо учитывать, что, хотя в масштабе генома распределение сайтов узнавания ферментом рестрикции можно считать относительно равномерным, для небольших локусов это может быть не так. Поэтому, фермент рестрикции при анализе специфический районов ДНК необходимо выбирать с учетом первичной последовательности этих районов.

Альтернативой ферментативному гидролизу является фрагментация ДНК ультразвуком. Параметры фрагментации подбираются таким образом, чтобы обеспечить формирование фрагментов ДНК подходящей длины. Для некоторых 3С-методов, в частности 4С (Circularized Chromosome Conformation Capture), было показано незначительно более высокое качество данных (Gao et al., 2013) при использовании такого подхода, однако, в большинстве работ для фрагментации ДНК используются ферменты рестрикции.

3. Лигирование фрагментированной ДНК. Преимущественное лигирование пространственно сближенных участков ДНК является основой любой 3С-методики, и обуславливается двумя факторами: проведением реакции лигирования в разбавленном растворе и наличием ковалентных, формальдегид-опосредованных сшивок между молекулами ДНК и белков (Simonis et al., 2007).

Для небольших, растворимых фрагментов хроматина, преимущественное лигирование фрагментов ДНК, которые были сближены в клетке исходно, достигается за счет лигирования в разбавленном растворе. При этом чем ближе изначально находились участки ДНК, тем больше для них вероятность оказаться после фиксации в одном мульти-молекулярного комплексе. Поскольку при разбавлении вероятность межмолекулярного взаимодействия намного ниже, чем вероятность внутримолекулярного, сближенные изначально фрагменты ДНК лигируются чаще.

Крупные, нерастворимые комплексы хроматина, могут с высокой вероятностью внутри одного комплекса содержать фрагменты ДНК, изначально удаленные на большое расстояние друг от друга в ядре. Однако, такой хроматиновый комплекс представляет собой ковалентно связанную формальдегидом трехмерную «сеть» из молекул ДНК и белков, поэтому сближение изначально удаленных фрагментов внутри комплекса после фиксации невозможно. Таким образом, внутри крупных комплексов также осуществляется лигирование только сближенных изначально (до фиксации) молекул ДНК. Лигирование фрагментов, входящих в состав двух крупных комплексов маловероятно в разбавленном растворе по описанным выше причинам.

Крупные нерастворимые комплексы, образующиеся в 3С-библиотеке, представляют собой фиксированные ядра или их крупные фрагменты (Gavrilov et al.,

2013). Интересно, что такие комплексы содержат большую часть ДНК библиотеки и, по видимому, позволяют детектировать пространственные взаимодействия гораздо более эффективно, чем растворимая фракция библиотеки (Gavrilov et al., 2013). Это предположение подтверждает и тот факт, что недавно предложенная методика приготовления 3С-библиотек *in situ,* в которой проведение этапов 1-3 приготовления библиотеки проводят на целых, нефрагментированных ядрах клеток, фиксированных на подложке, позволяет значительно снизить уровень шума (продуктов неспецифического лигирования) в получаемых материалах (Rao et al., 2014).

Поскольку лигирование в условиях разбавления является одним из ключевых условий успешного приготовления 3С библиотек, был предложен ряд методов, направленных на оптимизацию этого этапа. Так, в работе группы Калхор было предложено иммобилизовать ДНК-белковые комплексы на магнитных шариках, что позволяет существенно увеличить площадь поверхности реакции и частично предотвращает межмолекулярное взаимодействие между хроматиновыми комплексами (Kalhor et al., 2012). Иммобилизация происходит за счет образования связи между биотином, который неспецифически ковалентно присоединяется к белкам хроматина перед лигированием, и стрептавидином, связанным с поверхностью шариков. Таким образом, иммобилизация позволяет дополнительно очистить библиотеку от фрагментов ДНК, не связанных с белками (поскольку такие фрагменты не содержат биотина, они не будут иммобилизованы на шариках). Эта модификация метода Hi-C называется TCC (tethered chromosome conformation capture), именно она была использована в работе группы Калхор. Кроме того, группой Калхора был предложен метод отделения лигированных фрагментов ДНК от непровзаимодействовавшей фракции молекул (Kalhor et al., 2012).

4. Количественный и качественный анализ контактов 3С библиотеки. В оригинальном методе, разработанном группой Деккера (Dekker et al., 2002), наличие определенных гибридных фрагментов в 3С-библиотеке детектировались методом ПЦР. При этом частоты контактов оценивались методом полуколичественного ПЦР. Такой подход имеет два существенных недостатка. Во-первых, для изучения каждого пространственного контакта необходимо проведение индивидуальной реакции ПЦР

со специфическими праймерами. Таким образом, сравнение частот большого количества контактов между собой в одном эксперименте затруднительно. Во-вторых, метод ПЦР является полуколичественным, и, поэтому, не позволяет оценить различия в частотах контактов с высокой точностью.

Метод 3C был вскоре модифицирован: для оценки частот отдельных контактов с большей точностью был использован количественный метод ПЦР в реальном времени (3C-qPCR) (Hagège et al., 2007). С развитием методов массового параллельного секвенирования, появилась возможность полного секвенирования 3C-библиотеки. При секвенировании определяется первичная последовательность всех гибридных молекул ДНК 3C библиотеки, каждая из которых отражает пространственный контакт двух участков ДНК в исходной клетке. При этом, в отличие от классического метода 3C, в одном эксперименте генерируются данные о множестве контактов. Фактором, лимитирующем число проанализированных в эксперименте контактов, является соотношение частоты контактов и глубины секвенирования (т.е. общего числа гибридных молекул, последовательность которых определяется в эксперименте). Так, даже при небольшой глубине секвенирования, высока вероятность выявить и сравнить частоту встречаемости наиболее репрезентированных контактов, тогда как для идентификации крайне редких контактов необходимо глубокое секвенирование. Вариант методики 3C, в котором 3C-библиотека содержит контакты всего генома, и все эти контакты анализируются методом массового параллельного секвенирования, называется Hi-C (Lieberman-Aiden et al., 2009) (рис. 1, Д-Ж).

Существует ряд разновидностей метода 3C, позволяющих обогатить 3C-библиотеку контактами одного локуса со всем остальным геномом (4C) или нескольких локусов между собой (5C) (более подробный анализ этих технологий проведен в обзорах (Simonis et al., 2007) и (Баттулин et al., 2012)). При этом, для той же глубины секвенирования, методы 4C и 5C потенциально позволяют выявить большее число различных пространственных контактов в исследуемых локусах, однако при этом теряется информация о контактах остального генома. Более того, ряд методов под общим названием ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing) (Fullwood et al., 2009), позволяет обогащать 3C-

библиотеки контактами, опосредованными только каким-то определенным белком.

Таким образом, 3C-методы представляют собой набор технологий, позволяющих исследовать пространственную организацию генома с высоким разрешением за счет комбинации молекулярно-биологических методов. Все 3C-методы методы активно применяются для изучения пространственной организации как относительно небольших фрагментов ДНК размером в десятки тысяч нуклеотидов (например, b-глобинового локуса (Gavrilov et al., 2013)), так и крупных пространственных доменов в масштабе всего генома (Rao et al., 2014; Lieberman-Aiden et al., 2009; Dixon et al., 2012).

A- и B-домены хроматина

Первый полногеномный анализ пространственных контактов ДНК млекопитающих методом Hi-C показал, что в рамках каждой хромосомы можно выделить два крупных домена (A- и B-домены) таким образом, что контакты локусов внутри каждого домена сходны между собой и отличаются от контактов локусов, принадлежащих разным доменами (Lieberman-Aiden et al., 2009). Под сходством контактов двух локусов в данном случае понимается высокая позитивная корреляция частот их взаимодействий с остальными участками генома (а под различием – высокая негативная корреляция). Используя метод главных компонент, авторы подтвердил разделение генома на A- и B-домены, которое отражалось в близости значений первого собственного вектора матрицы контактов для локусов внутри домена (Lieberman-Aiden et al., 2009).

В отличие от коэффициента корреляции, биофизический смысл данной величины (значения первого собственного вектора матрицы контактов) не является тривиальным. Попробуем пояснить его следующим образом. Матрица контактов представляет собой форму записи о частотах контактов (таблицу NxN, где в строке i записаны контакты участка i со всеми остальными участками генома, по порядку). Таким образом, каждый элемент такой матрицы A_{ij} представляет собой число, отражающее частоту контактов локусов i и j в геноме. Разложение такой матрицы на собственные вектора позволяет представить каждый её элемент A_{ij} как

$$A_{ij} = \sum \lambda_k * E_i^k * E_j^k \quad (1)$$

Где λ_k – k-ое собственное число, E_i^k и E_j^k – значения k-ого собственного вектора (E^k) (читатель может более подробно ознакомиться с методом главных компонент и применением этого метода в биологии в обзоре (Shlens, 2005)).

Как видно из (1), для выбранного k, значение λ_k одинаково для всех локусов, поэтому специфичность контактов локусов (при выбранном k) определяется только E_i^k и E_j^k. Поэтому, биологический смысл этих величин можно интерпретировать следующим образом. E_i^k и E_j^k отражают свойства локусов i и j, определяющие их потенциал к формирования пространственных контактов. λ_k – это коэффициент, влияющий на этот потенциал.

Понятно, что чем больше λ_k, тем больше вклад E_i^k и E_j^k в значение A_{ij}. Поэтому, если отсортировать λ_k по убыванию и выбрать первое (наибольшее) значение, то именно соответствующие этому λ_1 значения E_i^1 и E_j^1 будут основными фактором, определяющими частоту контактов между локусами i и j. Такие значения – значения E^1 и составляют первый собственный вектор. Таким образом, значение первого собственного вектора для локуса i представляет собой число, которое характеризует потенциал этого локуса к формированию пространственных контактов в большой степени.

Следует отметить, что значения E^1 определены с точностью до знака. Действительно, если в формуле (1) оба значения E_i^k и E_j^k поменяют знак, результат произведения окажется неизменным. Поэтому, корректно говорить о разнице значений E^1, а не об их абсолютной величине.

В работе Либермана-Айдена и коллег все локусы с положительным значением собственного вектора были отнесены к одному домену, а с отрицательным – к другому. Поскольку, как уже было отмечено, знак E^1 можно выбрать произвольно, авторы выбирали его так, чтобы корреляция значений E^1 с GC-составом была положительной. Оказалось, что участки ДНК, принадлежащие одному из доменов, преимущественно контактировали с участками этого же домена, т.е. число контактов внутри доменов было значительно больше, чем контактов между ними. Это наблюдение позволяет предположить, что локусы внутри доменов пространственно сближены друг с другом и отдалены от локусов другого домена. Действительно, авторы показали, используя технологию FISH, что физическое расстояние между

локусами ДНК, принадлежащими одному домену, меньше, чем между локусами, лежащими в разных доменах, даже если исследуемые локусы разных доменов располагались ближе в линейной последовательности ДНК.

Локусы B-домена показывали, в среднем, большую частоту контактов, чем локуса A-домена, т.е. B-домен был упакован более плотно, чем A-домен. Более плотная упаковка B-домена также подтверждалась данными FISH. Детальный анализ A- и B-доменов показал, что первый обогащен генами и участками открытого хроматина с гистоновыми метками, ассоциированными с активацией транскрипцией. На основе этого анализа авторами был сделан вывод о том, что A-домен соответствует открытому, активно транскрибирующемуся хроматину (Lieberman-Aiden et al., 2009).

Участки, принадлежащие A- или B-доменами были выявлены при анализе данных Hi-C с разрешением 1 миллион пар оснований (Mb). Это означает, что весь геном был разделен на участки размером 1Mb, и контакты всех фрагментов внутри такого участка суммировались. Естественно, что протяженность регионов ДНК, входящих в A- или B-домен, исчислялась как минимум несколькими миллионами нуклеотидов. В дальнейшем, при анализе данных с более высоким разрешением (100 тысяч пар оснований (kb) и выше) было показано, что непрерывное распределение лучше описывает структуру пространственных доменов, чем дискретное разделение на два компартмента (Imakaev et al., 2012). Так, было показано, что значения первого собственного вектора непрерывно переходят из положительной в отрицательную область, без дискретно выраженной границы. Это означает, что ряд локусов в геноме имеет характеристики A-домена (и имеют большие положительные значения E^1) или B-домена (большие по модулю отрицательные значения E^1), однако ряд локусов занимает промежуточное состояние между ними.

В дополнение к вышеперечисленному, была показана корреляция значений E^1 со многими геномными элементами, в частности, GC-составом, временем репликации (причем более раннее время репликации соответствовало A-домену, более позднее – B-домену) и распределением различных гистоновых меток (при этом присутствие гистоновых меток характерных для активной транскрипции являлось

атрибутом A-домена) (Imakaev et al., 2012).

Таким образом, внутри хромосом можно выделить крупные домены двух типов (A и B) со средним размером порядка нескольких миллионов нуклеотидов. Пространственные контакты преимущественно наблюдаются между участками ДНК, принадлежащими доменам одного типа. При этом первый тип пространственных доменов лучше соответствует активно экспрессирующемуся, открытому хроматину, а второй – закрытому, неэкспрессирующемуся.

Топологические домены

Увеличение глубины секвенирования играет решающую роль для повышения уровня разрешения, с которым проводится анализ данных эксперимента Hi-C. A- и B-домены были обнаружены при анализе данных с разрешением от 100 kb до 1Mb. Увеличение глубины секвенирования и переход к разрешению 40-100kb, в сочетании с применением альтернативного биоинформационного подхода, позволил выявить пространственные домены меньшего размера – топологически ассоциированные домены (**T**opologically **A**ssociated **D**omains; TADs) (Dixon et al., 2012). Эти домены представляют собой участки ДНК, активно контактирующей внутри домена, и имеющие значительно меньшее количество контактов со всем остальным геномом. Средний размер TAD составляет ~1Mb (в несколько раз меньше, чем размер A/B-доменов).

Следует отметить, что размер TAD зависит от разрешения, на котором ведется анализ, и математического алгоритма, который используется для этого анализа. В оригинальной работе Диксона с коллегами, посвященной выявлению TAD, анализ проводился с разрешением 40kb, а в качестве математического алгоритма использовались скрытые модели Маркова (Dixon et al., 2012). Эти условия анализа в дальнейшем использовались в большинстве работ (Symmons et al., 2014; Trimarchi et al., 2014; Tark-Dame et al., 2014; Phillips-Cremins et al., 2013; Pope et al., 2014). Однако, были разработаны и альтернативные алгоритмы идентификации топологических доменов (Hou et al., 2012; Filippova et al., 2014), которые, при применении их на тех же данных, выявляют несколько отличные от опубликованных Диксоном с коллегами домены (Filippova et al., 2014). Более того, вариации в математических параметрах

алгоритма, предложенного Диксоном с коллегами, приводят к различиям в результатах анализа (Dixon et al., 2012). В дальнейшем, упоминая TAD мы будем иметь в виду домены, выявление которых проводилось строго в соответствии с алгоритмом, описанным Диксоном с коллегами (Dixon et al., 2012).

В геноме эмбриональных стволовых клеток (ЭСК), первого типа клеток, в котором было проведено исследование TAD, насчитывается порядка 2200 доменов со средним размером около 1Mb. TAD содержат порядка 91% всего генома ЭСК, а оставшиеся 9% приходятся на границы между доменами. Согласно данным FISH анализа, внутридоменные контактирующие фрагменты ДНК находятся ближе друг к другу, чем междоменные.

В этой и последующих работах было показано, что топологические домены являются пространственной структурой, являющейся единицей организации важнейших биологических процессов. Границы топологические доменов разделяют зоны эу- и гетерохроматина (Dixon et al., 2012), ранней и поздней репликации (Dixon et al., 2012; Pope et al., 2014) и A/B-доменов (Dixon et al., 2012). Гены, расположенные внутри одного TAD, имеют тенденцию к синхронному и однонаправленному изменению уровня транскрипции (Le Dily et al., 2014; Symmons et al., 2014). Более того, границы TAD доменов часто разделяют другие пространственные домены ядра, в частности ЛАД-домены (Dixon et al., 2012). Геномные мутации, такие как хромосомные перестройки, также чаще проходят по границам TAD доменов (Vietri Rudan et al., 2015).

Следует особо отметить поразительное сходство TAD-доменов в различных типах клеток. При сравнении клеток кортекса и ЭСК мыши друг с другом, оказалось, что более половины TAD в этих клетках имеют одинаковые границы (Dixon et al., 2012). Аналогичный результат наблюдался при сравнении ЭСК и фибробластов человека. Такое сходство TAD разных клеточных типов, а также вышеперечисленные свойства этих доменов, позволяет предположить, что они являются единицами регуляции генома. При этом в ходе дифференцировки, для формирования клеточной идентичности, активация и репрессия транскрипции осуществляется не на уровне отдельных генов, а на уровне целых геномных блоков, представленным в пространстве ядра топологическими доменами.

Функциональную важность TAD для организации генома (по крайней мере, млекопитающих) подтверждает консервативность этих доменов в ходе эволюции. Например, 53% междоменных границ в геноме ЭСК человека выявляются в тех же позициях генома при анализе участков гомологичной синтении ЭСК мыши, а 75,9% границ ЭСК мыши выявляются в ЭСК человека (Dixon et al., 2012). Высокая консервативность топологических доменов была также показана при сравнении пространственной организации генома в клетках печени мыши, макаки, кролика и собаки (Vietri Rudan et al., 2015).

Однако следует отметить, что столь высокий уровень консерватизма характерен, видимо, только для млекопитающих. Например, в геноме насекомых были обнаружены домены, аналогичные TAD, но их размер приблизительно на порядок меньше размера доменов млекопитающих (средний размер домена составляет ~100kb для дрозофилы и ~1Mb для мыши) (Sexton et al., 2012). Структуры, сходные с TAD были обнаружены у дрожжей, однако только у вида *S. Pombe* (Mizuguchi et al., 2014); у вида *S. Cerevisiae* домены обнаружены не были (Duan et al., 2010). Интересно, что у растений (*A. Thaliana*) TAD обнаружены не были (Grob et al., 2014; Feng et al., 2014), а у простейших (*P. fulciparum*) были обнаружены структуры отдаленно напоминающие TAD, но только в области активно работающих генов (Ay et al., 2014b).

Связь с важнейшими геномными процессами и эволюционный консерватизм TAD среди млекопитающих, подчеркивающий их роль для функционирования генома, остро ставит вопрос о механизмах формирования и поддержания топологических доменов.

Механизмы формирования и поддержания топологических доменов

Первые гипотезы о том, за счет чего поддерживается в геноме архитектура топологических доменов, были построены на основе анализа элементов, формирующих междоменные границы. Как уже упоминалось выше, границы доменов разделяют области активного и неактивного хроматина. При этом границы обогащены активно транскрибирующимися генами домашнего хозяйства и генами тРНК (Dixon et al., 2012). Логично предположить, что ДНК содержащая активно

транскрибирующиеся локусы с конститутивными промоторами находится в деконденсированном состоянии, что соответствует участкам между двумя более компактными, обогащенными контактами топологическими доменами.

Другим фактором, формирующим топологическую структуру доменов, являются белки, обеспечивающие пространственное сближение локусов ДНК или, наоборот, препятствующие сближению специфических локусов (Tark-Dame et al., 2014; Zuin et al., 2014; Mizuguchi et al., 2014). К таким белкам относятся, в частности, белок CTCF (CCCTC binding factor) (Johannes et al., 2013). На сегодняшний день накопилось достаточно фактов, свидетельствующих о том, что именно CTCF играет важную роль в формировании и поддержании структуры топологических доменов. Во-первых, границы доменов обогащены сайтами посадки CTCF по сравнению с внутридоменными участками (Dixon et al., 2012). При этом обогащение границ пространственных доменов фактором CTCF показано не только для млекопитающих, но и для насекомых (Hou et al., 2012; Sexton et al., 2012), и подтверждается при использовании альтернативных биоинформационных алгоритмов для определения позиций топологических доменов (Hou et al., 2012; Filippova et al., 2014). Во-вторых, прямой эксперимент по уменьшению экспрессии CTCF в клетках человека с использованием методики РНК интерференции показал, что при этом число междоменных контактов ДНК увеличивается, а число внутридоменных контактов – уменьшается (Zuin et al., 2014). В-третьих, при сравнении TAD мыши, кролика, шимпанзе и собаки, была показана зависимость эволюционных изменений TAD и сайтов связывания CTCF (Vietri Rudan et al., 2015).

Белок CTCF обладает целым рядом различных функций. Он может играть роль инсулятора, препятствовать связыванию энхансеров и промоторов, блокировать транскрипцию или участвовать в формировании петель за счет физического сближения участков ДНК (Johannes et al., 2013). Выполнение множества функций объясняется, по крайней мере частично, формированием комплексов между CTCF и другими белками (Johannes et al., 2013). Среди известных партнеров CTCF следует отметить белок медиатор и белковый комплекс когезин, причем и когезин, и медиатор могут связывать с ДНК и независимо от CTCF (Johannes et al., 2013).

Основываясь на данных CHIP-Seq, можно утверждать, что когезин, как и CTCF,

преимущественно связывается с ДНК в границах пространственных доменов (Zuin et al., 2014; Mizuguchi et al., 2014). Проведенное недавно исследование предложило элегантный метод для выяснения роли когезина в организации интерфазных хромосом. Для этого использовались клетки, несущие мутацию в гене одной из субъединиц когезина (RAD21). Мутация создавала в белке RAD21 сайт узнавания для протеазы HRV, но не влияла на функцию когезина. После экспрессии протеазы HRV, когезиновый комплекс разрушался, и эффект такого воздействия на топологию пространственных доменов анализировался методом Hi-C. Оказалось, что разрушение когезиновых комплексов приводит к уменьшению количества контактов, причем уменьшение контактов было особенно выражено в случае локусов, удаленных друг от друга на 100-200 kb и более, до 2Mb. Интересно, что уменьшение частот контактов касалось как контактов внутри, так и между пространственными доменами.

Проведенное группой английский ученых в 2013 году независимое исследование также подтвердило совместную функцию CTCF и когезина как инсуляторов, влияющих на структуру пространственных доменов (Sofueva et al., 2013). Авторами было показано, что большая часть сайтов связывания комплекса когезин совпадает с сайтами связывания белка CTCF. Такими совместными CTCF\когезиновыми сайтами были обогащены границы TAD доменов. Более того, TAD домены, содержащие активно транскрибирующиеся гены и принадлежащие к участкам A-доменов, были обогащены CTCF\когезиновыми сайтами по сравнению с TAD из неактивных B-доменов. Авторами был показан инсуляторный эффект сайтов связывания CTCF\когезинового комплекса, который заключался в приблизительно полуторакратном уменьшении контактов между локусами ДНК, расположенными по разным сторонам этого комплекса. В то же время, участки ДНК, расположенные по одну сторону от таких сайтов в непосредственной близости от них, показывали более высокую частоту контактов, чем в среднем по геному.

Интересно, что подобный инсуляторный эффект не наблюдался для сайтов, с которыми связывался только белок CTCF или только комплекс когезинов, что подчеркивает важность совместного действия двух этих регуляторов для формирования пространственных контактов.

Аналогично описанной выше работе, авторы для более детального выяснения роли когезина в пространственных контактах использовали клетки, лишенные RAD21. Однако проводилось не расщепление уже синтезированного белка RAD21 (как в предыдущей работе), а направленное удаление кодирующей части этого белка из генома при помощи системы LoxP/Cre. Как и в описанной выше работе, авторы наблюдали общее уменьшение числа пространственных контактов после удаления белка RAD21, особенно ярко выраженное для участков внутри TAD доменов. В частности, после удаления RAD21 не наблюдалось увеличение пространственных контактов для локусов, лежащих по одну сторону и в непосредственной близости от CTCF\когезиновых сайтов относительной средней частоты контактов в геноме. Кроме того, исследуя контакты в области CTCF\когезиновых сайтов методом 4C, авторы показали, что при отсутствии RAD21 исчезают специфические внутридоменные контакты между удаленными на большое расстояние (~1Mb) локусами. Помимо снижения уровня внутридоменных взаимодействий, в данном исследовании, после удаления белка RAD21, наблюдалось и увеличение междоменных взаимодействий (Sofueva et al., 2013).

Важно отметить, что полногеномный анализ транскрипционной активности показал, что в клетках, не экспрессирующих RAD21, меняется уровень экспрессии сотен генов (Sofueva et al., 2013), причем среди генов, изменивших уровень экспрессии при отсутствии RAD21, обогащены были такие, рядом с которыми (не далее чем в 10kb) находился сайт CTCF\когезина. Этот факт ещё раз подчеркивает связь экспрессии генов и пространственной укладки ДНК.

Наконец, модификация метода Hi-C, разработанная группой Либермана-Айдена в 2014 году, и позволившая изучить организацию топологических доменов с более высоким разрешением, показала, что пространственное взаимодействие между удаленными фрагментами ДНК часто опосредовано двумя сайтами узнавания белка CTCF одинаковой ориентации, связанных с когезиновым комплексом (Rao et al., 2014). Эта работа будет подробно обсуждаться ниже.

Если вышеперечисленные работы о роли факторов CTCF и когезина позволяют частично объяснить, как структура TAD поддерживается в интерфазном ядре, то проведенное недавно исследование пространственной организации митотических

хромосом делает вопрос формирования TAD после выхода клетки из митоза совершенно загадочным. Группа Деккера, исследуя хромосомы клеток HeLa в ходе митоза, показала, что в них отсутствуют пространственные домены (Naumova et al., 2013). Структура контактов митотических хромосом выглядит абсолютно однородной и зависящей только от расстояния между участками в линейной молекуле, без признаков каких-либо локус-специфичных петель. Такой результат был показан как с использованием метода Hi-C, так и метода с большим разрешением - 4C (Naumova et al., 2013). Конечно, можно предполагать, что TAD все-таки сохраняются в клетке во время митоза, но не выявляются из-за высокой компактизации митотических хромосом. Однако, вопрос формирования TAD в интерфазном ядре после выхода из митоза остается на сегодняшний день открытым.

Ещё одним открытым вопросом является то, как происходит формирование TAD-доменов после (или в ходе) процесса репликации. Репликация сопровождается деконденсацией ДНК и, следовательно, изменением структуры TAD-доменов. При этом на сегодняшний день не показано никаких механизмов, объясняющих последующего восстановления этой структуры.

Таким образом, функциональной и регуляторной единицей генома млекопитающих являются пространственные (топологические) домены размером около 1Mb. Эти домены имеют сходную структуру в разных типах клеток и высоко консервативны среди млекопитающих. Структура TAD поддерживается за счет действия белков-архитекторов хроматина. В частности, большую роль в их поддержании имеют когезин и CTCF. При этом детали механизмов, поддерживающих структуру TAD, а также процессов их формирования остаются неизвестными.

Петлевой уровень организации ДНК

Поскольку TAD сходны для разных типов клеток, связь клеточной идентичности с пространственной структурой её ДНК должна обеспечиваться либо за счет активации или репрессии доменов целиком, либо за счет различий пространственных контактов внутридоменных участков. Примеры регуляции на уровне целых TAD были описаны выше. Кроме этого, известен ряд примеров того,

как происходят пространственные взаимодействия участков ДНК внутри одного TAD, и как они связаны с регуляцией активности генов.

В ряде работ, посвященных изучению пространственной организации ДНК, отмечается сложная внутренняя структура TAD и выделяются субдомены – компактные участки внутри одного TAD, обогащенные пространственными контактами внутри участка (Pope et al., 2014; Phillips-Cremins et al., 2013; Zuin et al., 2014). В классическом Hi-C эксперименте, разрешение не позволяет достоверно идентифицировать такие структуры. В недавней работе, группа Либермана-Айдена предложила модифицированный метод *in situ* Hi-C, позволяющий, за счет проведения этапов фиксации, рестрикции и лигирования ДНК в интактных ядрах, существенно уменьшить количество артефактных взаимодействий и повысить соотношение сигнала и шума (Rao et al., 2014). Используя этот метод, а также за счет огромного масштаба эксперимента (в работе было проанализировано более сотни библиотек) и глубины секвенирования, авторами была получена пространственная карта генома человека с беспрецедентно высоким разрешением в несколько тысяч нуклеотидов. Авторы подтвердили, что при анализе данных на низком разрешении (~1Mb), в геноме можно выделить крупные компартменты, ассоциированные с открытым или закрытым хроматином (аналогичные A- и B-доменам). На более высоком разрешении, внутри таких крупных компартментов выделяются различные домены, размером от 40 kb до 3 Mb (средний размер ~185 kb). Эти структуры были названы «контактными доменами». Нужно отметить, что, хотя часть контактных доменов совпадает по размерам и локализации с TAD, многие из них (размером ~40-200kb) значительно меньше, чем топологические домены. Локусы, находящиеся внутри контактных доменов, имели сходные гистоновые модификации.

При выделении контактных доменов авторы использовали алгоритм, отличный от описанного выше для выявления TAD. Однако, общая идея определения контактных доменов и TAD является сходной: и те, и другие представляют собой участки, контактирующие с регионами внутри домена много больше, чем с окружением. Таким образом, для определения границ TAD и компактных доменов, основную роль играют частоты пространственных контактов участков, располагающихся внутри домена и в его небольшой окрестности. Информация о

пространственных контактах участков внутри домена с остальным геномом, в частности, с участками других хромосом, не учитываются.

Для того, чтобы выделить пространственные домены, используя информацию о контактах фрагментов ДНК со всем геномом, в частности, межхромосомных контактах, можно использовать другую логику. Можно выделить в один домен участки, имеющие сходный паттерн контактов с геномом, не учитывая при этом то, сколько эти локусы контактируют между собой. Например, все локусы, имеющие много больше удаленных контактов, чем в среднем по геному, можно выделить как отдельный домен – даже если такие локусы расположены далеко друг от друга в первичной структуре, и мало контактируют друг с другом. По своей логике аналогом такого разделения являются не TAD, а A- и B-компартменты, которые выделяются как группы локусов, имеющих сходное значение E^1 (которое, в свою очередь, отражает некое свойство пространственных контактов данного локуса со всем остальным геномом).

Руководствуясь вышеописанной логикой, авторы разделили паттерны контактов каждого из локусов с остальным геномом на ряд категорий, используя при этом различные алгоритмы кластерного анализа. Для того, чтобы в классификации локусов учитывался именно паттерн контактов с геномом, а не особенности, связанные с близостью локусов в линейной молекуле ДНК, авторы использовали для классификации межхромосомные контакты, которые значительно меньше зависят от расстояния между локусами в линейной молекуле, чем внутрихромосомные контакты. Оказалось, что весь геном состоит из участков шести типов: A1, A2 и B1, B2, B3, B4. Протяженность участка, относящегося к одной категории, составляет ~300kb.

Участки A1 и A2 располагались преимущественно внутри A-доменов и характеризовались большим количеством генов, высоким уровнем экспрессии и активными хроматиновыми метками. Локусы A1 и A2 относились к участкам ранней репликации, но репликация участков A1 заканчивалась в среднем раньше, чем участков A2. Кроме того, участки A2 характеризовались большей частотой встречаемости триметилированного лизина в девятом положении гистона H3 (хроматиновая метка, связанная с репрессией транскрипции, (Hahn et al., 2011;

Stewart et al., 2005)), более низким GC-составом и большей средней длиной гена, чем участки A1.

Участки B1-B3, в свою очередь, включали локусы из компартмента B, и имели следующие свойства. B1 по своим гистоновым меткам соответствовал факультативному гетерохроматину (повышена частота встречаемости гистоновой модификации H3K9me3, уменьшена – модификации H3K36me3) и реплицировался в середине S-фазы. Участки B2 включали большую долю (62%) перицентромерного гетерохроматина и были обогащены ДНК, ассоциированной с ядерной ламиной и ДНК, ассоциированной с ядрышком. Участки B3 были сходны по характеристикам с участками B2, но не содержали ДНК, ассоциированной с ядрышком. Шестой тип контактов (B4) был характерен исключительно для небольшого участка хромосомы 19 человека, содержащего около половины всех генов семейства KRAB-ZNF. Выделение этого участка в особый тип согласуется с проведенным ранее исследованием, показавшем особую, отличную от остального генома, укладку хроматина в этом регионе (Hahn et al., 2011). Важно отметить, что переходы между участками разного типа (A1, A2 и B1-B4) чаще всего наблюдались в границах контактных доменов.

Помимо классификации локусов по типу пространственных контактов, в описываемой работе (Rao et al., 2014) был также проведен анализ контактов на уровне единичных «хроматиновых петель» - пар локусов ДНК, контактирующих друг с другом с более высокой частотой, чем ожидаемая для локусов, разделенных данным количеством нуклеотидов. В клетках человека линии GM12878 авторы выделили около 10 000 таких контактов, большинство из которых (98%) было представлено локусами, расположенными на расстоянии 2 Mb друг от друга. Функциональную значимость таких контактов подтверждает то, что среди них наблюдалось обогащение взаимодействиями между известными промоторами и энхансерами. Более того, промоторы активно экспрессирующихся генов приблизительно в 6 раз чаще встречались среди идентифицированных контактов. Было показано, что часть таких контактов совпадает в лимфобластоидных клетках GM12878 и других типах клеток (фибробластах, эпителиальных клетках и др.). Однако, 25-45% контактов различались, что может служить основой для клеточно-

специфичной регуляции генной экспрессии. Подтверждением этому является то, что для промоторов генов, экспрессирующихся в клетках на разном уровне, часто наблюдались специфичные только для этих клеток пространственные взаимодействия. Таким образом, была показана связь между формированием специфических для клеточного типа «хроматиновых петель» с участием промоторов генов и увеличением экспрессии этих генов в данном типе клеток (Rao et al., 2014).

Выявление специфичных для клеточного типа пространственных контактов методом Hi-C хорошо согласуется с опубликованными ранее данными о связи пространственной структуры генома и регуляции генной экспрессии. В качестве одного из наиболее изученных примеров можно отметить реорганизацию b-глобинового локуса человека при эритроидной дифференцировке. Показанное в одной из первых работ, использующих метод 3C, в 2002 году взаимодействие между b-глобиновым геном и его энхансером, удаленным на 50 kb, является специфичным для эритроидной дифференцировке и необходимо для корректной экспрессии b-глобинового гена (Carter et al., 2002; Tolhuis et al., 2002). На полногеномном уровне также было показано, что энхансеры часто находятся в одном TAD с генами, которые они регулируют (Duggal et al., 2014; Ay et al., 2014a).

Таким образом, максимально доступным разрешением для полногеномного анализа пространственных контактов на сегодняшний день является ~1 kb. На этом уровне, ДНК организована в компактные домены размером ~185 kb. Локусы, которые входят в состав одного компактного домена, характеризуются специфической организацией контактов с остальным геномом. Можно выделить шесть типов компактных доменов, каждый из которых, оказывается, по своим характеристикам близок к определенному типу организации хроматина (эухроматин, факультативный гетерохроматин и т.д.). Внутри компактных доменов можно различить отдельные пространственные взаимодействия, некоторые из которых являются специфическими для определенного типа клеток и соединяют промоторы и энхансеры активно работающих генов. Общие для разных типов клеток контакты, по-видимому, принимают участие в построении консервативных доменов, которые являются, как описывалось выше, единицей базовых клеточных процессов, таких как, например, репликации или регуляции транскрипции на уровне целого домена.

Моделирование пространственной организации биополимеров

Помимо выявления пространственной организации отдельных локусов, методы полногеномного анализа, такие как Hi-C, предоставляют богатый материал для моделирования пространственной организации ДНК. Такое моделирование необходимо по двум причинам.

Во-первых, сами по себе 3C-методы позволяют получить информацию о частотах попарных контактов исследуемых участков ДНК. В то же время, полным описанием пространственной структуры ДНК является выяснение топологии этой молекулы в пространстве, т.е. получение физических координат каждого из участков относительно некого заданного начала отсчета координат. Построение трехмерной структуры на основе частот попарных контактов является нетривиальной задачей, которая может не иметь однозначного решения. В связи с этим, разрабатываются компьютерные модели, рассчитывающие наиболее вероятную трехмерную конформацию ДНК на основе результатов Hi-C экспериментов (Hu et al., 2013).

Во-вторых, материалом для 3C эксперимента в большинстве случаев является клеточная популяция. Это означает, что наблюдаемая разница в частоте контактов тех или иных локусов может потенциально иметь две причины: либо локусы, контактирующие чаще, находятся в пространстве ближе, чем локусы контактирующие реже, либо локусы находятся в обоих случаях на одинаковом расстоянии, но контакт фиксируется только в части клеток популяции. Проведение анализа единичных клеток показало достаточно высокую вариабельность пространственных контактов в популяции (Nagano et al., 2013). Однако, исследование популяции является на сегодняшний день обязательным условием для полногеномного анализа на высоком разрешении, поскольку анализ единичных клеток не позволяет получить достаточное количество материала (Nagano et al., 2013). Поэтому, разрабатываются модели, способные описывать пространственную структуру ДНК на основе данных 3C-экспериментов с учетом популяционной вариабельности контактов (Kalhor et al., 2012; Hu et al., 2013).

По результатам Hi-C экспериментов, группой Мирного были разработаны две альтернативные модели укладки ДНК: в виде равновесной и фрактальной глобулы

(Lieberman-Aiden et al., 2009; Mirny, 2011; Naumova et al., 2013). Можно представить себе образование равновесной глобулы следующим образом: стартуя из некоторой точки, нить полимера распространяются в определенном направлении, до тех пор, пока не встречает границу области, которой ограничен допустимый объем. После этого, движение полимера продолжается от границы внутрь объема в случайном направлении. Так, постепенно, полимер заполняет доступное пространство (следует отметить, что это объяснение приведено исключительно для пояснения структуры равновесной глобулы и может не иметь ничего общего с реальным физическим механизмом сворачивания равновесной глобулы). Образование фрактальной глобулы можно представить себе так: в каждом участке полимер образуются петли небольшого масштаба, затем полимер (уже состоящий из петель на каждом участке) формирует петли большего масштаба за счет сближения удаленных петель, образовавшихся на первом этапе, после чего образуются еще большие петли и так далее (Mirny, 2011).

Фрактальная и равновесная глобулы характеризуются разной зависимостью вероятности контактов и физического расстояния от расстояния между локусами в линейной молекуле. Если обозначать расстояние между локусами в линейной молекуле за s, то вероятность контактов для двух локусов в пространстве P (s) для фрактальной глобулы пропорциональная s^{-1}, а для равновесной глобулы $s^{-3/2}$. Физическое расстояние между двумя локусами L(s) также зависит от выбора модели упаковки ДНК: $L(s) \sim s^{1/2}$ для равновесной глобулы и $s^{1/3}$ для фрактальной глобулы (Lieberman-Aiden et al., 2009; Mirny, 2011). Эксперименты Hi-C показали, что для геномов многих типов клеток характерна зависимость частоты контактов от удаленности локусов $P(s) \sim s^{-1}$, что соответствует фрактальной укладке ДНК (Mirny, 2011; Naumova et al., 2013; Imakaev et al., 2012; Lieberman-Aiden et al., 2009).

Интересно, что профиль зависимости частоты контактов от удаленности локусов в митотических хромосомах составляет $P(s) \sim s^{-0,5}$ для участков, разделенных 100 kb – 10 Mb, и резко падала для более удаленных участков. Такое поведение не характерно ни для равновесной, ни для фрактальной глобулы. Для описания митотических хромосом, была создана модель, включающая две стадии компактизации: формирование вложенных петель (по типу фрактальной глобулы) на

первом этапе, и последующее линейное сжатие всей структуры, на втором этапе. В сочетании с заданием цилиндрической геометрии, такая модель хорошо воспроизводила особенности укладки метафазных хромосом (Naumova et al., 2013).

Для физического моделирования вышеописанных фрактальной и равновесных глобул, используется следующий подход. Полимер ДНК представляется набором шариков, соединенных пружинками, при этом каждый шарик представляет собой участок ДНК определенной длины. Затем, варьируя условия взаимодействий между шариками (т.е. изменяя их характерный размер и свойства, а также размер и свойства соединяющих их пружинок), и накладывая некие общие ограничения (например, геометрию модели, или тип организации полимера – фрактальный или равновесный), подбираются условия соответствия полученным в эксперименте данным. В частности, в моделях группы Мирного, таким условием было распределение вероятности контактов от удаленности локусов, P(s).

Принципиально другими можно назвать модели, которые позволяют подобрать оптимальный набор координат локусов ДНК в пространстве так, чтобы расстояния между ними максимально соответствовали выявленным в эксперименте частотам пространственных контактов (Hu et al., 2013). В отличие от моделей группы Мирного, в которых пространственное сближение индивидуальных шариков, соответствующих фрагментам ДНК, описывается известными физическими законами, такие модели физические закономерности организации ДНК не учитывают. Можно сказать, что в этих моделях пружинки между шариками выбрасываются, и исследователя интересуют только положения шариков – а не законы, которыми это расположение обеспечивается. Таким образом, эти модели не учитывают физическую природу организации ДНК, однако позволяют получить физические координаты для определенных локусов. При этом современные модификации моделей позволяют учитывать и популяционную вариабельность пространственных контактов среди клеток, а также особенности первичной структуры генома, которые потенциально могут влиять на результаты 3C методов (Hu et al., 2013).

Примерами такой модели можно называть байесовские алгоритмы BACH и BACH-MIX, которые, основываясь на данных о пространственных контактах, а

также локальных особенностях каждого участка генома (таких как число и распределение сайтов рестрикции, GC-состав и вырожденность последовательность ДНК данного региона), определяют наиболее вероятное распределение геномных локусов в пространстве.

Особенности пространственной укладки ДНК сперматозоидов

В отличие от всех других клеточных типов, большая часть геномной ДНК в ядре сперматозоидов связана не с гистонами, а с другими белками – протаминами (Balhorn et al., 1977). Протамины представляют собой относительно небольшие, сильно основные белки. Замещение гистонов на протамины происходит постепенно в ходе созревания сперматозоидов (Carrell et al., 2007). При этом, по мере созревания сперматозоида хроматин переходит из «открытого» активного состояния к очень компактному, электронно-плотному полностью неактивному состоянию. Такая реорганизация и компактизация хроматина, по-видимому, происходит сходным образом у всех млекопитающих.

Все наши современные знания о тонкой организации генома сперматозоидов получены на основе данных электронной и атомно-силовой микроскопии. Так, по данным электронной микроскопии ДНК, в сперматидах, организована типичным для соматических клеток способом, то есть формирует ~11 нм узелки и 30 нм фибриллы (Horowitz et al., 1994). Позднее такая структура преобразуется в фибриллы диаметром 50-100 нм, значительно больше нуклеосом (Fuentes-Mascorro et al., 2000). По мере дальнейшей конденсации хроматина эти фибриллы объединяются и становятся настолько плотными, что не могут быть разрешены методами электронной микроскопии. Более глубокий анализ структуры ДНК зрелых сперматозоидов возможен только при применении специальных методов деконденсации хроматина. Исследование частично деконденсированного хроматина сперматозоидов методами электронной микроскопии показало существование двух типов структурных единиц различного размера. Первый тип имеет характерный размер порядка нуклеосомы (диаметр ~10 нм, толщина ~5 нм), второй имеет форму тора с диметром 60–100 нм, толщиной в 20 нм и с отверстием в центре (Balhorn et al., 1999).

Тем не менее, такие структурные особенности хроматина соматических клеток как, хромосомные территории, петлевые домены и районы прикрепления к матриксу (matrix attachment regions, MAR) по-видимому, сохраняются и в хроматине сперматозоидов даже после замены гистонов на протамины и общей конденсации хроматина. FISH с использованием хромосомо-специфических зондов показала наличие хромосомных территорий в ядре зрелого сперматозоида человека (Zalenskaya et al., 2004). Белковый состав ядерного матрикса меняется по мере дифференцировки сперматид (Chen et al., 2001), однако ДНК все это время остается связанной с матриксом в ~50,000 сайтов. ДНК между сайтами прикрепления к матриксу, по-видимому, сохраняет петлевую организацию свойственную соматическим клеткам (Heng et al., 2004; Heng et al., 2001). Сохранение подобной организации хроматина важно для реактивации генома после оплодотворения и инициации первого цикла репликации ДНК (Shaman et al., 2007). Кроме того, считается, что сохранение петлевых доменов способствует переупаковке ДНК и протаминов в тороиды.

Стоит подчеркнуть, что описанные структуры были открыты методами микроскопии и до сих пор нет данных, позволяющих связать эти структуры с конкретными последовательностями ДНК в масштабе всего генома. Остается неизвестным, сохраняется ли в сперматозоидах типичная для соматических и стволовых клеток доменная организация генома Hi-C (Dixon et al., 2012).

Материалы и методы

Материалы

Hi-C библиотеки

Получение Hi-C библиотек было выполнено Баттулиным Н.Р. и Хабаровой А.А. с участием автора, по протоколу TCC, предложенному Калхором с соавторами (Kalhor et al., 2012). Библиотеки были получены из

1. подвижных сперматозоидов мыши C57/Black, выделенных из придатков семенников и

2. эмбриональных фибробластов мыши, полученных из 12,5-дневных эмбрионов мыши линии C57/Black

При приготовлении библиотек рестрикция ДНК проводилась ферментом HindIII.

Секвенирование образцов проводилось А. М. Мазуром и Е.В. Прохорчуком на базе центра «Биоинженерия» (Москва). Секвенирование проводилось на платформе Illumina GAL II. Длина ридов (прочтений) составляла 100 п.н. Нами было проведено несколько независимых раундов секвенирования библиотек фибробластов и сперматозоидов, результаты которых затем были объединены (см. главу «Результаты» для пояснений).

В работе также использовались Hi-C библиотеки эмбриональных стволовых клеток и кортекса мыши, опубликованные Диксоном с соавторами (Dixon et al., 2012). Данные, полученные после массового секвенирования этих библиотек, были загружены из публичного депозитария NCBI (архивы SRR443883, SRR443884 и SRR443885).

Вычислительные ресурсы

Для запуска скриптов использовался вычислительный кластер Новосибирского Государственного Университета (http://www.nusc.ru/) под управлением ОС SUSE Linux Enterprise Server 11. Необходимое для анализа данных программное обеспечение было написано на языках Python (версии 2.6 и 2.7), Perl (версия 5.0) или R (версия 3.0) и, при необходимости, скомпилировано на операционных системах

Linux. Для работы с картами пространственных контактов с высоким разрешением, использовались очереди вычислительного кластера с большим объемом оперативной памяти vkopq и vkop2q (запрашивался 1 терабайт оперативной памяти).

Методы

Картирование ридов на геном

Риды картировали на геном мыши версии mm9. Геном, включающий аутосомы и хромосомы X, Y и M (митохондриальная ДНК) был получен из базы данных UCSC (https://genome.ucsc.edu/). На основе полученного генома программой bowtie2-build был собран индекс mm9_bt2 (Langmead et al., 2012). Поскольку каждый рид представляет собой молекулу ДНК, соединяющую два фрагмента генома, картирование обоих концов рида велось независимо. Так как точное положение места лигирования двух фрагментов ДНК в риде неизвестно, для картирования использовался специальный алгоритм, предложенный группой Мирного (Imakaev et al., 2012). Для библиотек фибробластов и сперматозоидов, сначала, проводился анализ 25 п.о., расположенных на 5'-конце рида. Если местоположение в геноме не определялось однозначно, длина фрагмента увеличивалась на 5 п.о. и вновь проводилось картирование. Так проводилось постепенное увеличение длины фрагмента рида либо до тех пор, пока не удавалось однозначно картировать фрагмент на геном, либо пока длина анализируемого фрагмента не достигала 50 п.о. (половина всего анализируемого рида). Риды, для которых не удавалось установить однозначно локализацию в геноме при длине фрагмента 50 п.о., считались некартируемыми. Аналогично картировали 3'-конец рида, начиная с фрагмента в 25 п.о. и добавляя 5 пар оснований на каждом шаге. Для реализации основной части алгоритма использовался модуль mapping.py из библиотеки скриптов, разработанных группой Мирного (Mirnlib версии 0d30147f052f и Hi-C lib версии d28d8d985120, http://mirnylab.bitbucket.org/hiclib/).

Hi-C библиотеки ЭСК и кортекса были проанализированы аналогичным образом, с учетом отличающейся длины рида в этих библиотеках.

Непосредственно картирование (поиск анализируемых фрагментов в геноме) проводили при помощи программы bowtie2 версии 2.2.1 (Langmead et al., 2012) с

использованием индекса mm9_bt2 и опциями «-q - -5 0 -3 75 -p 8 --very-sensitive». Результаты картирования сохранялись в файлах, содержащие для каждого рида координаты двух участков генома: соответствующих 5’-концу рида и 3’-концу рида.

Фильтрация ридов

Полученные в результате картирования риды подвергали следующим процедурам фильтрации:

1. Удаление ридов, расположенных слишком близко к сайту рестрикции (здесь и далее под «сайтом рестрикции» имеется в виду сайт рестрикции фермента, использованного для приготовления Hi-C библиотеки). Такие риды не могут быть правильно картированы, поскольку фрагмент, используемый для картирования, слишком мал. Мы удаляли все риды, находящиеся ближе, чем 5 п.н. от сайта рестрикции, поскольку 5 п.н. является слишком маленьким участком, чтобы картировать рид в геноме.

2. Удаление ПЦР-дубликатов.

3. Удаление ридов из крайне маленьких (<100 п.о.) и крайне больших (>100 000 п.о.) рестрикционных фрагментов генома mm9. Рестрикционным фрагментом называется участок генома mm9, ограниченный двумя сайтами рестрикции. Если длина фрагмента рестрикции велика, его 3’-конец может взаимодействовать с участками генома, от которых «середина» фрагмента лежат на большом расстоянии и наоборот. Поскольку в методе Hi-C мы не можем различить взаимодействие каких-либо частей рестрикционного фрагмента, слишком большие фрагменты удалялись из анализа.

Слишком маленькие фрагменты (<100 п.о.) могут не иметь достаточной конформационной свободы для сближения с другим участком ДНК и лигирования, и поэтом также удалялись.

4. Удаление оверрепрезентированных (сверх часто представленных) фрагментов. Теоретически, все фрагменты генома должны быть одинаково представлены в Hi-C библиотеке (этот феномен будет более подробно обсуждаться ниже). Например, под действием фермента рестрикции HindIII, в геноме мыши mm9 образуется 823 370 фрагментов рестрикции, и на каждый из них должно приходиться

(при равномерном распределении) 0.000121 % ридов. В реальном эксперименте, в связи с особенностями первичной последовательности ДНК, доступностью сайтов рестрикции, неравномерной амплификацией и другими причинами, некоторые фрагменты могут оказаться оверрепрезентироваными. Поэтому, если на фрагмент рестрикции приходилось более, чем 0.5% всех ридов, все эти риды удалялись.

5. Удаление колец. В процессе лигирования, не исключена ситуация, когда концы одного рестрикционного фрагмента взаимодействую друг с другом. Более того, при неполной рестрикции, может быть, что такое кольцо включает несколько последовательных сайтов рестрикции. После секвенирования, такие фрагменты можно отличить по тому, выравниваются ли 5'- и 3'-концы рида на одну и ту же, или на разные цепи ДНК. Риды, относящиеся к кольцам, удалялись.

6. Удаление «свисающих концов» (в оригинале – "dangling ends"). В процессе приготовления Hi-C библиотеки, фактически на всех стадиях, может происходить спонтанная, несвязанная с действием фермента рестрикции, фрагментация ДНК. Такая фрагментированная ДНК находится в растворе, не связана с белками формальдегидом, и может спонтанно лигироваться по тупому концу, создавая при анализе данных секвенирования шум. После картирования 5'- и 3'-концов ридов возможно реконструировать фрагмент гибридной ДНК, давший начало этому риду. Для этого необходимо найти последовательность ДНК, лежащую между картированными участками и ближайшим к ним сайтам рестрикции, и объединить их. Если полученный фрагмент много больше размера фрагментов, подвергавшихся секвенированию, или если на участке генома между местами локализации 5'- и 3'-концов рида вообще нет сайта рестрикции, такой рид считается сформированным из случайных, не опосредованных хроматином лигирований (описанных выше) и поэтому удалялся.

Параметры фильтрации, а также мотивация для введения данных фильтров была обоснована работой (Imakaev et al., 2012). Для реализации основной части алгоритма использовался модуль fragmentHiC.py из библиотеки скриптов, разработанных группой Мирного (Mirnlib и Hi-C lib).

Построение матрицы пространственных контактов

Матрица пространственных контактов (A) строилась как таблица NxN, строки и столбцы которой представляли собой номера локусов в геноме mm9, а в ячейках которой (A_{ij}) находилось число контактов между соответствующими локусами (i и j). Локусы нумеровались, начиная с первого нуклеотида первой хромосомы, имели сквозную нумерацию и заканчивались последним локусом хромосомы X. Хромосомы Y и M не участвовали в дальнейшем анализе. Контактом между двумя локусами считался рид, концы которого были картированы внутри двух этих локусов. Если в локусе находилось больше одного рестрикционного фрагмента, их контакты суммировались. Размеры локусов, использовавшихся в данном анализе, составляли от 40 000 п.о. до 1 000 000 п.о., и назывались «разрешением» матрицы. Естественно, при увеличении размера локуса уменьшалось их число, но увеличивалось среднее число ридов (и контактов), приходящихся на один локус. Мотивация выбора разрешения приведена ниже (см. раздел Идентификация TAD доменов) и в главе «Результаты». Каждый локус ДНК, представленный одной строчкой матрицы контактов, называется бином.

Контакты диагональных элементов матрицы считали равными 0. Контакты соседних локусов (первых над- и поддиагональных элементов) также считали равным 0, чтобы не учитывать взаимодействие соседних фрагментов рестрикции.

Построение матрицы пространственных контактов было выполнено при помощи модуля binnedData.py из библиотеки скриптов, разработанных группой Мирного (Mirnlib и Hi-C lib).

Мы оценили ошибку для каждого взаимодействия, указанного в матрице контактов (SE), как

$$SE = \frac{K}{\sqrt{K}}$$

Где K – число ридов, поддерживающих данное взаимодействие. Ошибка вычислялась, учитывая только те взаимодействия, для которых K было отличным от 0.

Коррекция матрицы контактов

Для коррекции матрицы контактов использовался алгоритм iterative correction из работы (Imakaev et al., 2012). При этом каждый элемент исходной матрицы контактов (A_{ij}) представляется в виде

$$A_{ij} = B_i * B_j * T_{ij}\,, \quad (2)$$

Где B_i и B_j – некие коэффициенты, отражающие биофизические особенности локусов i и j, влияющие на их репрезентацию в Hi-C библиотеке (особенности первичной последовательности ДНК, влияющие на эффективность лигирования, доступности сайтов рестрикции и т.д.), а T_{ij} - частота контактов локуса i и j, которая наблюдалась бы без влияния особенностей B_i и B_j. Если бы особенностей B_i и B_j не было, все локусы были бы одинаково представлены в библиотеке Hi-C, что означает сумму по каждой строке или столбцу матрицы равной некой константе:

$$\sum_{j=1}^{N} A_{i,j} = Const \quad (3)$$

Поскольку контакты оцениваются друг относительно друга, константу можно выбрать любой, например, равной единице. (2) и (3) образуют систему уравнений, которые могут быть решены численно. При этом будут найдены T_{ij}, которые и составляют скорректированную матрицу контактов.

Для реализации основной части алгоритма коррекции использовался модуль BinnedData.py из библиотеки скриптов, разработанных группой Мирного (Mirnlib и Hi-C lib).

В дальнейшем, под матрицами контактов понимаются скорректированные матрицы, если не указано другое.

Поиск различий в профилях контактов локусов фибробластов и сперматозоидов

Для выявления районов, различающих матрицы контактов фибробластов и сперматозоидов, использовались три метода: вычисление Евклидова расстояния, коэффициентов корреляции Пирсона и Спирмена и значений E^1. Сравнение всегда проводилось для матриц одинакового разрешения.

Сравнение на основе Евклидова расстояния

Вычисление евклидова расстояния (E) проводилось по формуле

$$E = \sqrt{\sum_{j=c_{st}..c_{end}, j \neq i} \left(Sp_{ij} - Fib_{ij}\right)^2}$$

где E – Евклидово расстояние между бинами (локусами) i и j на хромосоме C, c_{st} – первый бин хромосомы C, c_{end} – последний бин хромосомы C, Sp_{ij} и Fib_{ij} – число контактов между бинами i и j для сперматозоидов и фибробластов соответственно. Чем большее значение E, тем больше считалось различие между бинами.

Сравнение на основе коэффициентов корреляции

Для вычисления коэффициентов корреляции Пирсона и Спирмена, использовали внутрихромосомные контакты каждого бина в виде одномерного массива, между которыми вычисляли коэффициенты корреляции функциями scipy.stats.spearmanr и scipy.stats.pearsonr (встроенные функции языка Python). Поскольку коэффициент корреляции Спирмена является ранговым, он более чувствительным к небольшим различиям между образцами, наблюдаемыми в областях с очень маленькими количествами контактов (имеются в виду взаимодействия с сильно удаленными от анализируемого бина областями, для которых число контактов находится практически на уровне шума, порядка 1-2 ридов на контакт). Поскольку такие небольшие частоты контактов (1-2 рида на контакт) могут быть связаны с неспецифическим, случайным лигированием фрагментов ДНК и, в этом случае, не имеют биологического смысла, мы использовали в дальнейшем коэффициент Пирсона.

Следует отметить, что уровень шума может влиять на результаты анализа, полученные с помощью коэффициента корреляции Пирсона или Спирмена. Для локусов, в контактах которых, по каким-либо причинам, соотношение сигнал\шум значительно ниже, чем в других, корреляция контактов будет ниже, чем для локусов с меньшим уровнем шума (будем называть в дальнейшем такие характеризующиеся высоким уровнем шума локусы «неинформативные»). Таким образом, при сравнении паттернов контактов на основе коэффициентов корреляции, низкая корреляция характерна для двух типов локусов: тех, для которых паттерн контактов значительно различается в рассматриваемых типах клеток и для неинформативных локусов. Для того, чтобы результат сравнения адекватно отображал реальные

физические различия пространственной организации, последние требуется исключить.

Мы предложили решение этой проблемы на основе анализа «референсных» выборок (фактически, метод статистического бутстрэппинга, bootstrapping (Efron, 1979)). Под референсными выборками имеются в виду две выборки, сгенерированные случайным образом из данных Hi-C ЭСК, количество данных в которых было сходным с количеством данных для фибробластов и сперматозоидов. Будучи случайными подвыборками одной генеральной совокупности, референсные выборки должны, в теории, быть одинаковыми и иметь одинаковый, стремящийся к 1 коэффициент корреляции для всех локусов. Однако, особенности локусов ДНК, создающие неравномерность в распределении шума, приводят к тому, что коэффициент корреляции Спирмена различается для разных локусов референсных выборок. Мы определили неинформативные локусы как такие, для которых выполняется

$$C_i < M - SD$$

где C_i – коэффициент корреляции Спирмена/Пирсона для бинов i в референсных выборках, M и SD – медиана и стандартное отклонение распределения коэффициентов корреляции Спирмена/Пирсона всех бинов референсных выборок, соответственно.

Мы подтвердили, что использование коэффициента корреляции Пирсона предпочтительно по сравнению с использованием коэффициента корреляции Спирмена, поскольку использование коэффициента корреляции Пирсона давало лучшие (т.е. в среднем более высокие) результаты при сравнении референсных выборок.

Мы провели анализ коэффициентов корреляции контактов для неинформативных локусов и их окружения (см. приложение 1). В соответствии с критерием выбора неинформативных локусов, для них наблюдался значительно более низкий коэффициент корреляции Пирсона (~0.79), чем средний по геному (~1.0). Более того, значительное снижение коэффициента корреляции (~0.84) наблюдалось и для локусов, расположенных на расстоянии +/- 1 бин от неинформативных локусов.

Чтобы учесть данные особенность локусов, мы исключили из дальнейшего рассмотрения все неинформативыне и смежные с ними бины, и рассчитали коэффициент корреляции для оставшихся локусов фибробластов и сперматозоидов. Более высокое значение коэффициента корреляции означало большее сходство локусов.

Сравнение на основе значений E^1

Значения первого собственного вектора (E^1) вычислялись исходя из уравнения (1), для k=1 (см. раздел «A- и B-домены хроматина» главы «Обзор литературы» для пояснения биологического смысла E^1). Вычисление E^1 проводилось, используя модуль binnedData.py из библиотеки Hi-C lib. Для сравнения полученных значений E^1 фибробластов и сперматозоидов, нами был предложен следующий алгоритм. Каждый бин был представлен в виде точки в двухмерном пространстве, X- и Y-координатами которой являлись значение E^1 фибробластов и сперматозоидов. Затем, для полученного множества точек, была подсчитана линия линейной регрессии, используя метод наименьших квадратов (подсчет осуществлялся при помощи функции scipy.stats.linregress, встроенная функции языка Python). Наконец, было подсчитано геометрическое расстояние от этой линии до каждой точки. Это расстояние считалось критерием схожести значения E^1: чем больше расстояние от точки до прямой, тем большим принимался уровень различий для бинов, соответствующих данным точкам.

После вычисления Евклидова расстояния, коэффициента корреляции Пирсона и различий E^1, мы присвоили каждому бину 3 характеристики, соответствующие полученным в этих трех методах сравнения значениям. Для каждой из характеристик, из общего количества в 2410 бинов (на разрешении матрицы контактов в 1 Mb), выбиралось некоторое количество (TopN) наиболее различающихся бинов. Таким образом, нами было получено три списка наиболее различающихся в фибробластах и сперматозоидах бинов.

Для каждой из использованных характеристик сходства – Евклидова расстояния, коэффициентов корреляции или анализа E^1, могут существовать свои локус-специфичные особенности, делающие ряд бинов неинформативными (пример

таких особенностей – соотношение сигнал\шум, описанной для коэффициентов корреляции). Чтобы учесть такие особенности, мы подсчитали Евклидово расстояние и сходство E^1 для референсных выборок, описанных выше, выбрали TopN наиболее различающихся бинов и исключили их из списков, сформированных для сперматозоидов и фибробластов (для коэффициентов корреляции такой анализ не выполнялся, поскольку неинформативные для данного метода анализа бины уже были исключены ранее).

Из-за специфических особенностей (например, чувствительности к соотношению сигнал\шум, описанному для коэффициентов корреляции), один и тот же бин может иметь разный уровень сходства между фибробластами и сперматозоидами, при определении этого уровня разными методами. Кроме того, разные математические методы могут выявлять различия, имеющие разную биофизическую природу. Чтобы учесть эти эффекты, мы считали наиболее сильно различающимися локусы, оказавшиеся в пересечении всех трех полученных выборок бинов. Мы оценили размер случайного пересечения выборок как

$$\frac{N_{Eucl} * N_P * N_{E1}}{{N_T}^2}$$

где N_{Eucl} – число бинов, выбранных на основе Евклидова расстояния (TopN за вычетом отфильтрованных на основе анализа референсных выборок бинов), N_p, N_{E1} – тоже для коэффициента корреляции Пирсона и анализа сходства E^1, N_T – общее число бинов (2410 для разрешения 1 Mb).

Определение уровня сходства значений E^1 фибробластов и сперматозоидов.

Для определения сходства E^1 фибробластов и сперматозоидов, использовалось два метода: вычисление коэффициента корреляции Спирмена и коэффициента взаимной информаци. Коэффициент Спирмена вычислялся при помощи функции scipy.stats.spearmanr, коэффициент взаимной информации вычислялся в соответствии с алгоритмом, описанным в (Reshef et al., 2011), используя модуль

MINE языка Python (http://minepy.sourceforge.net/) со стандартными параметрами (alpha=0.6, c=15). Следует отметить, что в отличие от предыдущего метода (раздел «выявление районов, различающих фибробласты и сперматозоиды»), в данном случае сравнивались не индивидуальные значения E^1 для каждого бина, а массивы E^1, характеризующие матрицу контактов целиком.

Выявление статистических различий в частотах индивидуальных контактов между сперматозоидами и фибробластами

Для описания вероятностей пространственных контактов, полученных в Hi-C эксперименте, мы использовали статистическую модель биномиального распределения, предложенную ранее в (Duan et al., 2010). Используя эту модель, Hi-C эксперимент можно представить себе следующим образом: для определенного контакта, каждый рид из библиотеки является независимым испытанием; испытание считается успешным, если рид поддерживает данный контакт, неудачным – если он поддерживает любой другой контакт. В таком случае, вероятность контактов между бинами i и j ($P^{i,j}$) составляет

$$P^{i,j} = \frac{m}{M}$$

где m – число контактов между локусами i и j, а M - общее число контактов в матрице (учитывая симметричность матрицы, M можно вычислить как сумму всей матрице разделенную на 2). При этом контакты, представленные только одним ридом (имеется в виду реальное число ридов в матрице до коррекции) не учитывались, поскольку они неотличимы от шума (мотивация и выбор порога для количества ридов равного 1 основывается на расчетах предложенных в (Duan et al., 2010); кроме того, такие контакты не удовлетворяют аппроксимации биномиального распределения нормальным, см. ниже).

При определенных условиях биномиальное распределение может быть представлено как нормальное. Мы использовали критерий

$$M*P^{i,j}*(1- P^{i,j}) > 9,$$

предложенный в (Schader et al., 1989), и исключили из рассмотрения все контакты, неудовлетворяющие критерию. Мы протестировали нулевую гипотезу

$$H_0: P_{Sp}^{i,j}=P_{Fib}^{i,j},$$

где $P_{Sp}^{i,j}, P_{Fib}^{i,j}$ - вероятности контактов между бинами i и j для сперматозоидов (P_{Sp}) и фибробластов (P_{Fib}).

Предполагая нормальную аппроксимацию биномиального распределения, мы рассчитали критерий (p-value) для нулевой гипотезы:

$$pval^{i,j} = 2 * Norm\left(\frac{P_{Sp}^{i,j} - P_{Fib}^{i,j}}{\sqrt{\frac{P_{Sp}^{i,j} * \left(1 - P_{Sp}^{i,j}\right)}{M_{Sp}} + \frac{P_{Fib}^{i,j} * \left(1 - P_{Fib}^{i,j}\right)}{M_{Fib}}}}\right) - 1$$

где *Norm* - функция нормального распределения, $pval^{ij}$ – критерий достоверности различий между бинами i и j. Умножив каждое из значений pval на число проанализированных локусов, мы получили значения q (q-values) – статистический критерий достоверности, учитывающие поправку на множественное тестирование гипотез.

Моделирование «компрессии» генома

Чтобы смоделировать «компрессию» генома фибробластов, мы использовали следующий алгоритм. Мы, сначала, высчитывали значения K_j следующим образом:

$$K_j = \frac{Sp_j}{Fib_j}$$

где Sp_j и Fib_j представляют собой суммы элементов диагонали номер j в матрице контактов сперматозоидов и фибробластов. Фактически, K_j отражает среднее превышение частоты контактов для локусов, расположенных на расстоянии j в линейной молекуле ДНК, в сперматозоидах по сравнению с фибробластами. Мы оценили относительную погрешность при вычислении K_j следующим образом:

$$Err_j = \frac{1}{\sqrt{Sp_j}} + \frac{1}{\sqrt{Fib_j}}$$

Затем, мы генерировали «скомпрессированную» матрицу, умножая каждый элемент диагонали j матрицы контактов фибробластов на соответствующий коэффициент K_j. Мы проводили коррекцию только в том случае, если погрешность определения K_j составляла менее 5%. Наконец, мы проводили нормировку контактов, так чтобы сумма контактов исходной матрицы и «скомпрессированной»

оказались одинаковыми. Для этого «скомпрессированную» матрицу умножали на коэффициент нормировки

$$Knorm = \frac{\sum_{i,j} Fib}{\sum_{i,j} Fib\_compressed}$$

Где Fib – исходная матрица контактов, Fib_compressed – матрица контактов, полученная в ходе «компрессии».

Компрессия других матриц выполнялась аналогично.

Идентификация TAD доменов

TAD домены в геноме определяли согласно (Dixon et al., 2012). Матрицу контактов генерировали с разрешением 40 000 п.о., экспортировали в текстовый формат (для этого использовали скрипт на языке Python) и анализировали полученный файл программой DI_from_matrix.pl из пакета domaincall_software (пакет доступен по адресу http://bioinformatics-renlab.ucsd.edu/collaborations/sid/domaincall_software.zip). При этом параметр window size выставляли равным bin_size*50, в соответствии с рекомендацией (Dixon et al., 2012) (для матрицы с разрешением 40 000 п.о. bin_size составляет 2 Mb).

Программа DI_from_matrix.pl скринирует геном и, для каждого бина, позволяет получить значение DI (domain index), основанное на анализе контактов этого бина с локусами, расположенными на расстоянии не более чем window size от него. Переменная DI показывает разницу частот контактов слева и справа от данного бина. Поскольку топологический домен представляет собой участок ДНК, состоящий из локусов, контактирующих друг с другом больше, чем с окружающими локусами (см. главу «Обзор литературы», где более систематически обсуждаются определения доменов), граница топологического домена должна характеризоваться большим по модулю значением DI, тогда как участки внутри домена должны характеризоваться значениями DI близкими к нулю. Для определения границ доменов по полученным значениям DI, использовалась Скрытая Марковская Модель, как предложено (Dixon et al., 2012).

Определение границ проводилось программой HMM_calls, входящей в состав domaincall_software. Полученные карты доменов конвертировались в формат UCSC bigwig и анализировались при помощи скриптов на языке Python.

Моделирование пространственного расположения локусов в TAD доменах

Моделирование пространственного расположения локусов в TAD доменах проводилось при помощи алгоритма BACH, с использованием одноименного программного обеспечения (доступного по адресу http://www.people.fas.harvard.edu/~junliu/BACH/) (Hu et al., 2013). Сначала, генерировали входные данные для проведения анализа. Для этого, использовали функцию getAnnotatedRestrictionSites() из пакета HiTC (Servant et al., 2012) для генома mm9 (в качестве генома использовалась переменная BSgenome.Mmusculus.UCSC.mm9 из пакета Bioconductor). Данная функция позволяет получить параметр “mappability” – выраженную в числовом формате характеристику, показывающую насколько эффективно картирование ридов, происходящих из окрестности генома рядом с сайтом рестрикции. Другие входные параметры: GC-состав в районе сайтов рестрикции и число сайтов рестрикции в анализируемых локусах, получали при помощи встроенных функций библиотеки Mirnlib. Агрегацию входных данных и конвертацию в соответствующий формат выполняли при помощи скрипта на языке Python.

Входные данные были сгенерированы для всех TAD-доменов сперматозоидов и фибробластов, после чего программа BACH была запущена независимо для каждого из доменов. В результате, для каждого домена была получена пространственная модель, представлявшая собой физические координаты каждого локуса домена. Для анализа полученных данных, мы реализовали алгоритм, основанный на (Hu et al., 2013). Сначала, положение домена в пространстве изменялось параллельным переносом так, чтобы его геометрический центр находился в начале координат. Фактически, это преобразование соответствовало следующему:

$$x_{new} = x - \langle x \rangle$$

Где x_{new} – новая координата (по одной из осей), x – старая координата (по этой же оси), <x> - среднее значение координат всех участков домена по этой оси.

Затем, пространственную структуру топологического домена аппроксимировали цилиндром и разворачивали в пространстве так, чтобы его наиболее длинная ось совпадала с осью X. Это делали при помощи следующей

математической операции. Пусть M – матрица размером N*3, где N (число строк матрицы) соответствует числу локусов анализируемого TAD, и каждая строка содержит 3 значения, соответствующие координатам данного локуса по исходным осям X, Y, Z. Тогда матрица M_{new}, соответствующая координатам локусов в новых координатах (где ось X совпадает с главной осью цилиндра), вычисляется как

$$M_{new} = M \times V$$

где x – оператор умножения матриц, M – исходная матрица, V – унитарная матрица правых сингулярных векторов, вычисляемая сингулярным разложением матрицы $\frac{M}{\sqrt{N}}$ (более подрбно о геометрическом смысле такого разложения можно прочитать в (Shlens, 2005)). Сингулярное разложение проводилось при помощи функции numpy.linalg.svd (встроенная функция языка Python).

Длина полученного цилиндра (H) вычислялась как

$$H = P_{90}\text{-}P_{10}$$

Где P_{90} и P_{10} – соответствующие процентили (90% и 10%) разброса координат по оси X, а диаметр цилиндра (D) высчитывался как

$$D = \frac{\sum_N(\sqrt{y^2 + z^2})}{N}$$

Где z и y – значения координат по соответствующим осям для каждого из локусов, N – число локусов в TAD.

Для каждого полученного цилиндра высчитывался HD параметр, как отношение длины цилиндра к диаметру (см. главу «Результаты» для пояснения физического и биологического смысла величины HD). Статистическая значимость различий длин и HD параметров для TAD сперматозоидов и фибробластов проверялась тестом Манна-Уитни (Mann–Whitney), и при вероятности отвержения гипотезы (P-value) менее 0.001 различия принимались за значимые.

Анализ межхромосомных контактов

Анализ межхромосомных контактов проводили по аналогии с алгоритмами, описанными ранее (Lieberman-Aiden et al., 2009; Kalhor et al., 2012). Для этого, для каждой пары хромосом i и j, рассчитывали соотношение между ожидаемым и детектированным количеством межхромосомных взаимодействий следующим

образом:

$$\frac{S_{ij}}{(S_i * \frac{S_j}{T - S_i} + S_j * \frac{S_i}{T - S_i}) * 0.5}$$

где S_i и S_j - суммы межхромосомных контактов хромосомы i или j со всем остальным геномом, S_{ij} - сумма межхромосомных контактов i и j друг с другом, T - общая сумма всех межхромосомных контактов.

Анализ зависимости частоты взаимодействий от расстояния между локусами в линейной молекуле

Анализ зависимости частоты взаимодействий (P) от расстояния между локусами в линейной молекуле (s) выполнялся согласно (Lieberman-Aiden et al., 2009). Скрипт на языке Python для расчета P(s) был предоставлен Помазным М.Ю.

Результаты

Оценка количества и качества данных массового параллельного секвенирования.

Мы провели анализ данных массового параллельного секвенирования Hi-C библиотек фибробластов и сперматозоидов мыши. Как видно из рисунка 2, полученные нами данные (библиотека 1) по объему сильно уступали данным, полученным в предыдущих экспериментах Hi-C и описанным в литературе (Dixon et al., 2012; Battulin et al., 2014). В связи с этим, мы предприняли два дополнительных раунда секвенирования и получили объединенную библиотеку, сравнимую по объему опубликованными. Финальная библиотека включала около 150 и 400 миллионов ридов для фибробластов и сперматозоидов, соответственно (рис. 2).

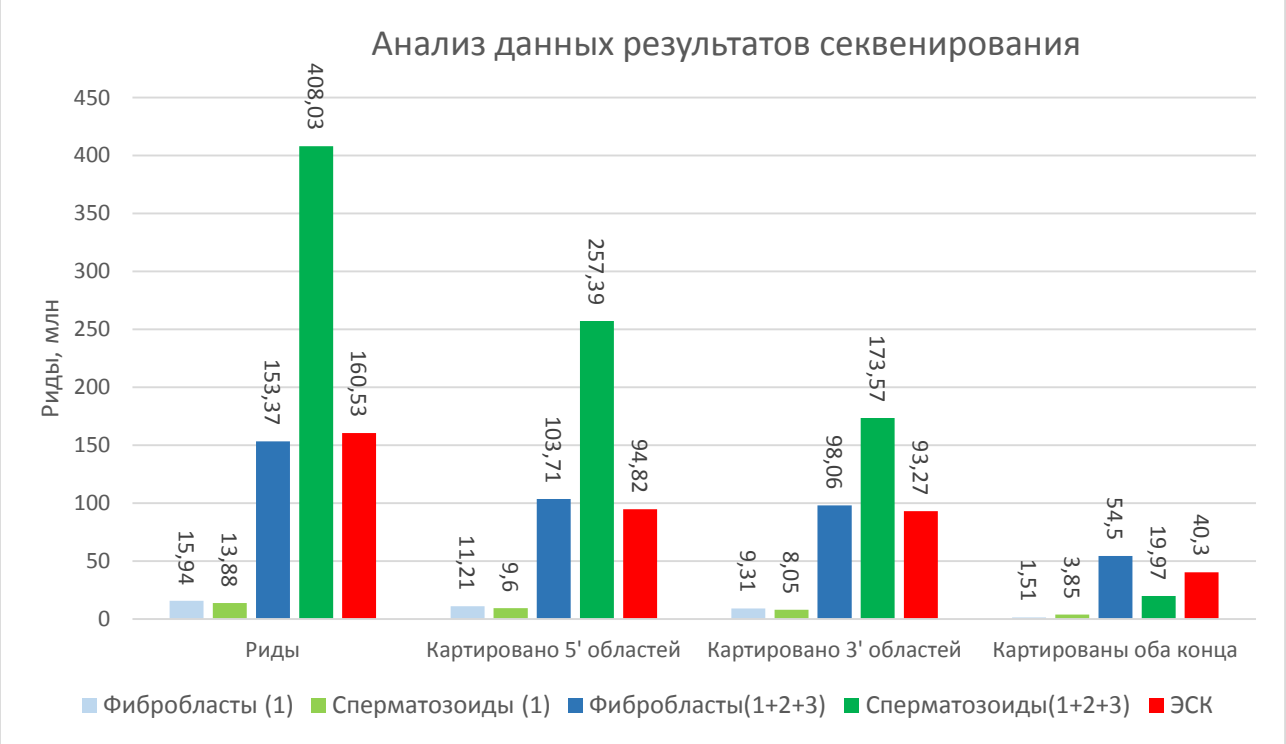


Рисунок 2. Анализ количества и качества данных массового секвенирования. Приведено количество ридов в библиотеке 1 и объединенных библиотек 1, 2 и 3, в миллионах. В качестве контроля использовался один технический повтор данных секвенирования Hi-C библиотеки ЭСК мыши, взятый из работы (Dixon et al., 2012). Риды – общее число ридов в библиотеке, картировано 5’(3’) областей – число ридов, у которых картирована 5’(3’) область, картированы оба конца – число ридов, у которых картированы и 5’- и 3'-область, после прохождения фильтров.

Мы провели картирование ридов финальной библиотеки на геном и осуществили фильтрацию, позволившую, исключить, по крайней мере частично, следующие неспецифические продукты: ПЦР дубликаты, продукты случайного (не опосредованного хроматином) взаимодействия фрагментов ДНК, продукты внутримолекулярного лигирования с образованием колец и т.д. (см. главу «Материалы и методы). Для 51% (76,98 млн) всех ридов библиотеки фибробластов и 38% (150,90 млн) ридов библиотеки сперматозоидов удалось картировать и 5’- и 3’-конец. После применения фильтров число ридов сократилось до 36% (54,5 млн) для фибробластов и 5% (19,97 млн) для сперматозоидов (рис. 3). Мы обнаружили, что наиболее наибольшее число ридов было удалено фильтром «свисающие концы», который удаляет фрагменты неспецифического лигирования, причем для сперматозоидов действие этого фильтра было особенно выраженным (удалено 85% всех картированных ридов, в то время как для фибробластов это значение составляло 21.5%).

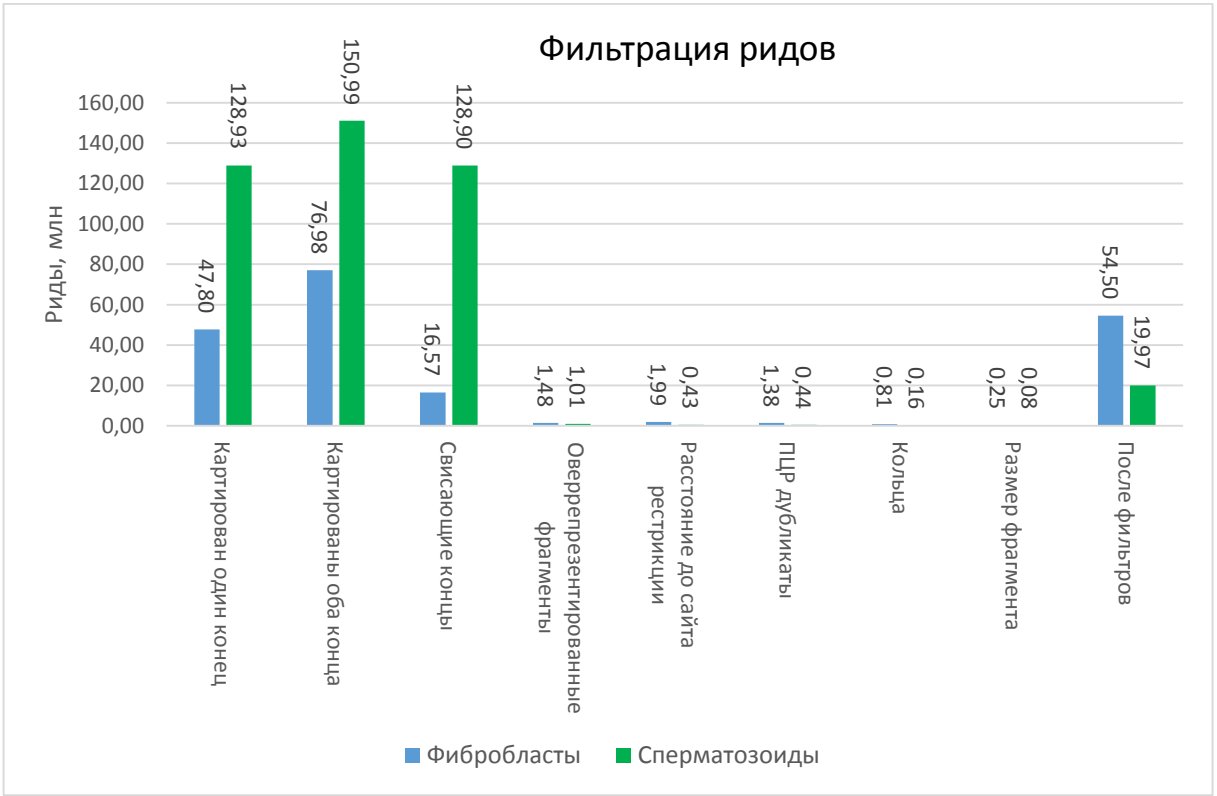


Рисунок 3. Фильтрация ридов. Показано число ридов, у которых картирован только один конец, оба конца, и число ридов, удаленных различными фильтрами (названия фильтров соответствуют тем, что приведены в главе «Материалы и Методы»), для фибробластов и сперматозоидов.

В геноме мыши содержится порядка 820 000 сайтов рестрикции HindIII. Теоретически, все они должны быть представлены в Hi-C библиотеке. Мы оценили, сколько сайтов рестрикции представлены в полученных нами данных, и обнаружили, что более чем 92,5% всех имеющихся в геноме сайтов встречаются в ридах каждой из трех проанализированных библиотек (сперматозоидов, фибробластов и ЭСК) после фильтрации (рис. 4). При этом число сайтов рестрикции в разных библиотеках варьировало незначительно (92,5 – 94,5% от всех имеющихся в геноме сайтов).

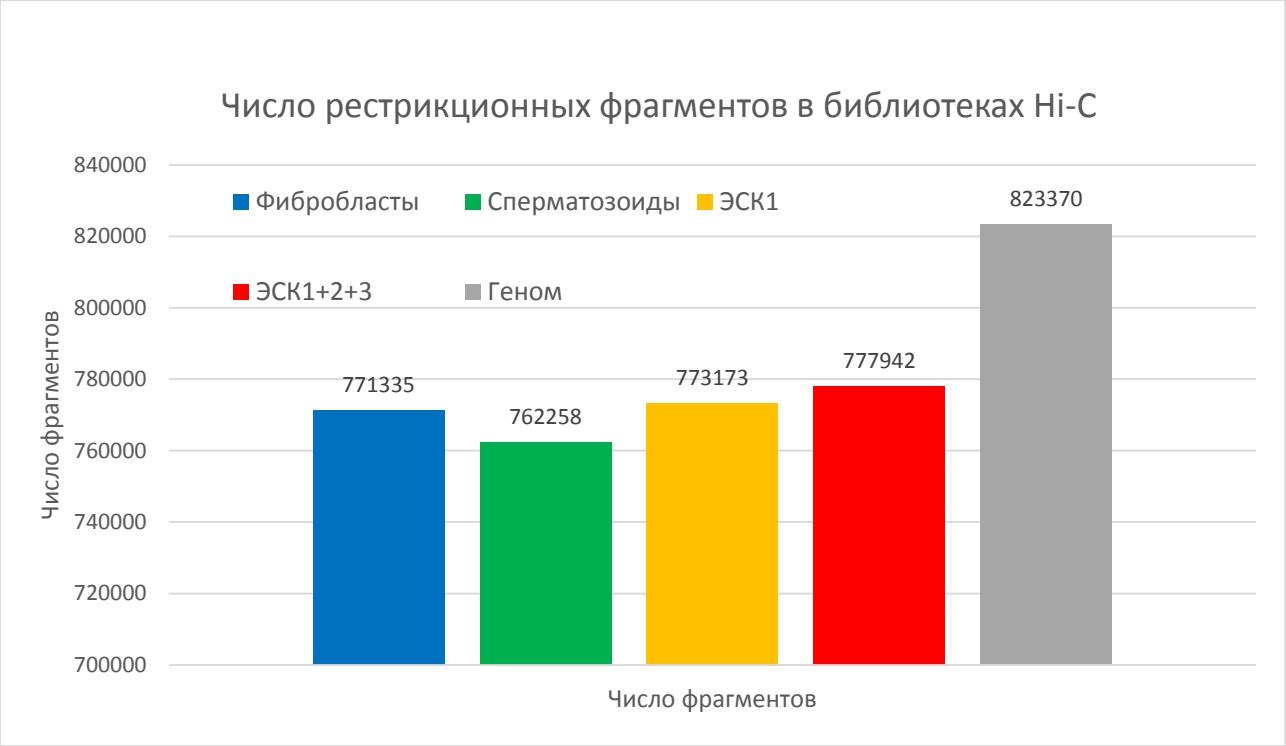


Рисунок 4. Число рестрикционных фрагментов в библиотеке Hi-C. ЭСК1 - один технический повтор данных секвенирования Hi-C библиотеки ЭСК мыши (данные по которому представлены на рис. 2), ЭСК1+2+3 – объединенная библиотека, содержащая 3 технических повтора секвенирования Hi-C библиотеки ЭСК мыши (один из которых - ЭСК1). ЭСК1+2+3 содержит приблизительно в 3 раза больше исходных данных, чем библиотека ЭСК1. Геном – общее число сайтов HindIII в геноме мыши.

Построение матрицы пространственных контактов

Основываясь на ридах, прошедших фильтры, мы построили матрицы пространственных контактов для сперматозоидов, фибробластов и ЭСК (рис. 5 и 6). Матрица контактов представляет собой таблицу, в ячейках которых указаны частоты контактов между соответствующими локусами (см. главу «Материалы и Методы»

для более детального описания). Визуальный анализ полученных матриц показал, что индивидуальные хромосомы выявляются в них компактными, контакт-обогащенными кластерами с четкими границами. Рисунок контактов внутри индивидуальных хромосом имеет явно выраженный «клетчатый» паттерн, что подразумевает наличие внутри хромосом протяженных участков, разделенных крупными областями в линейной молекуле, но взаимодействующих друг с другом в пространстве ядра.

Мы провели коррекцию полученных контактов при помощи алгоритма iterative correction (Imakaev et al., 2012). Идея этого алгоритма основывается на том, что каждый фрагмент в Hi-C библиотеке, в теоретическом идеальном эксперименте, должен быть представлен одинаковым количеством ридов. Однако, в реальности, это не так. Например, сайты рестрикции могут различаться по доступности ферменту (некоторые из них могут быть закрыты белками хроматина), или фрагмент может содержать повторяющиеся мотивы ДНК, что снижает эффективность картирования ридов в области этого фрагмента. Алгоритм iterative correction, основываясь на предположении о равной представленности всех сайтов в «идеальном» эксперименте, позволяет скорректировать вышеперечисленные особенности локусов ДНК. Удобство этого алгоритма заключается в том, что для коррекции не нужно знать биофизические причины таких особенностей, их количество и распределение среди локусов. Математическое объяснение алгоритма приведено в главе «Материалы и методы». После коррекции, «клетчатый» паттерн контактов, как и обособленность индивидуальных хромосом, стала ещё более выраженной (рис. 5 и 6).

В полученных матрицах пространственных контактов весь геном поделен на дискретные локусы (бины) фиксированной длины. Все контакты фрагментов рестрикции внутри одного бина суммируются. Мы проанализировали размер погрешности определения частоты пространственных контактов, и обнаружили явно выраженный рост этого параметра при увеличении разрешения матрицы (т.е. при уменьшении размера бина). Так, при разрешении 1 Mb, средняя ошибка для внутрихромосомных контактов составляла ~24 %, а при увеличении разрешения на порядок (до 100 kb) средняя ошибка вырастала до ~88 %. Поэтому, в большинстве

дальнейших вычислений использовалась матрица, составленная с разрешением 1 Mb.

При визуальном анализе мы обнаружили высокое сходство карт пространственных контактов фибробластов и сперматозоидов (рис. 6). Чтобы подтвердить это сходство, мы провели анализ значений E^1 для этих матриц.

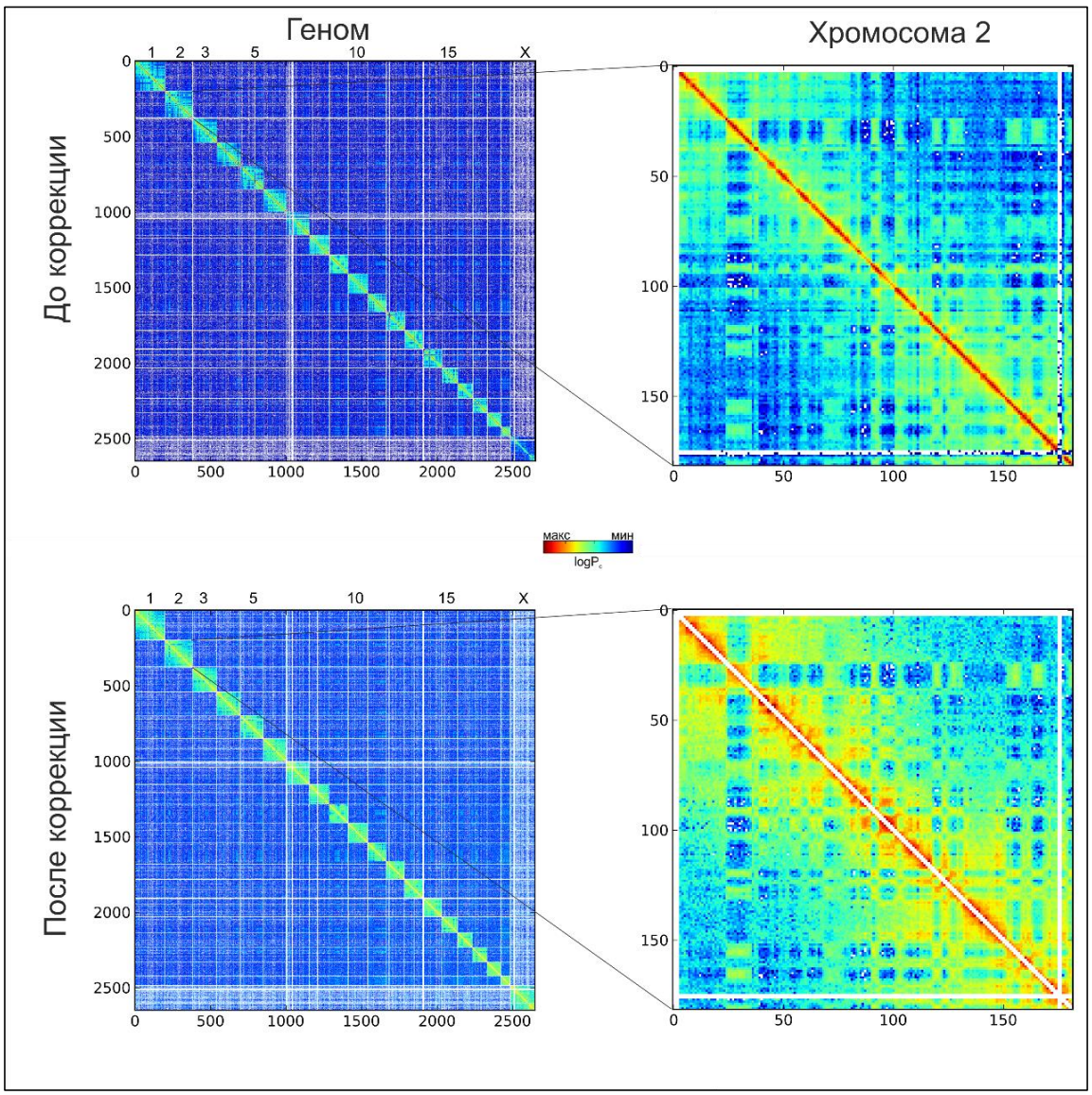


Рисунок 5. Влияние корректировки на матрицу пространственных контактов. На рисунке (слева) изображена матрица пространственных контактов сперматозоидов до коррекции (сверху) и после коррекции (снизу). Справа показано увеличенное изображение внутрихромосомных контактов в хромосоме 2. Сбоку и снизу каждой матрицы контактов указаны номера бинов (локусов ДНК), а в ячейках матрицы отображен логарифм относительных частот контактов между парами локусов ($\log P_c$), в цветовой шкале: красный – максимальный (наибольшая вероятность контактов), синий – минимальный. Белые полосы соответствуют некартируемым участкам. Сверху каждой матрицы указаны номера соответствующих хромосом. Для хромосомы 2 нумерация бинов начинается заново. Матрицы приведены с разрешением 1 Mb.

A/B-домены в геномах сперматозоидов и фибробластов

Ранее было показано, что значения E^1 характеризуют пространственную организацию локусов в клетках млекопитающих (Imakaev et al., 2012; Lieberman-Aiden et al., 2009) (см. главу «Обзор литературы»). Мы провели вычисление значений собственных векторов для полученных матриц пространственных контактов фибробластов и сперматозоидов, а также матриц ЭСК и клеток кортекса, описанных в (Dixon et al., 2012) (рис. 6,7; таблицы 1,2). Сравнение полученных значений E^1 - представлено на рисунке 7.

Полученные данных хорошо согласуются с визуальным сходством матриц пространственных контактов: коэффициент корреляции Спирмена для значений E^1 сперматозоидов и фибробластов составил 0.878. Следует отметить, что значения E^1 сперматозоидов и других типов клеток для других типов клеток также показывали высокий уровень сходства: коэффициент корреляции для значений E^1 сперматозоидов и ЭСК составил 0.878, сперматозоидов и кортекса - 0.901 (таблица 1). Эти результаты были подтверждены при использовании альтернативной меры сходства двух наборов данных: коэффициента максимальной информации (таблица 2) (Reshef et al., 2011). Коэффициент максимальной информации может быть использован для выявления различных типов взаимосвязей между двумя наборами данных (в нашем случае - значениями E^1 в разных типах клеток), включающих не только линейную зависимость (которая выявляется такими мерами сходства, как коэффициент Спирмена или Пирсона), но и логарифмическую и другие нелинейные зависимости. Полученные высокие коэффициенты максимальной информации указывают на сходство значений E^1 для разных типов клеток.

В работе (Lieberman-Aiden et al., 2009) было показано, что геном может быть поделен на различные по своим характеристикам A- и B- домены, основываясь на значениях E^1. Чередование областей, характеризующихся положительными и отрицательными значениями E^1 указывает на присутствие этих доменов в геномах сперматозоидов и фибробластов. А сходство значений E^1 сперматозоидов и фибробластов свидетельствует о сходстве этих доменов в разных данных типах клеток.

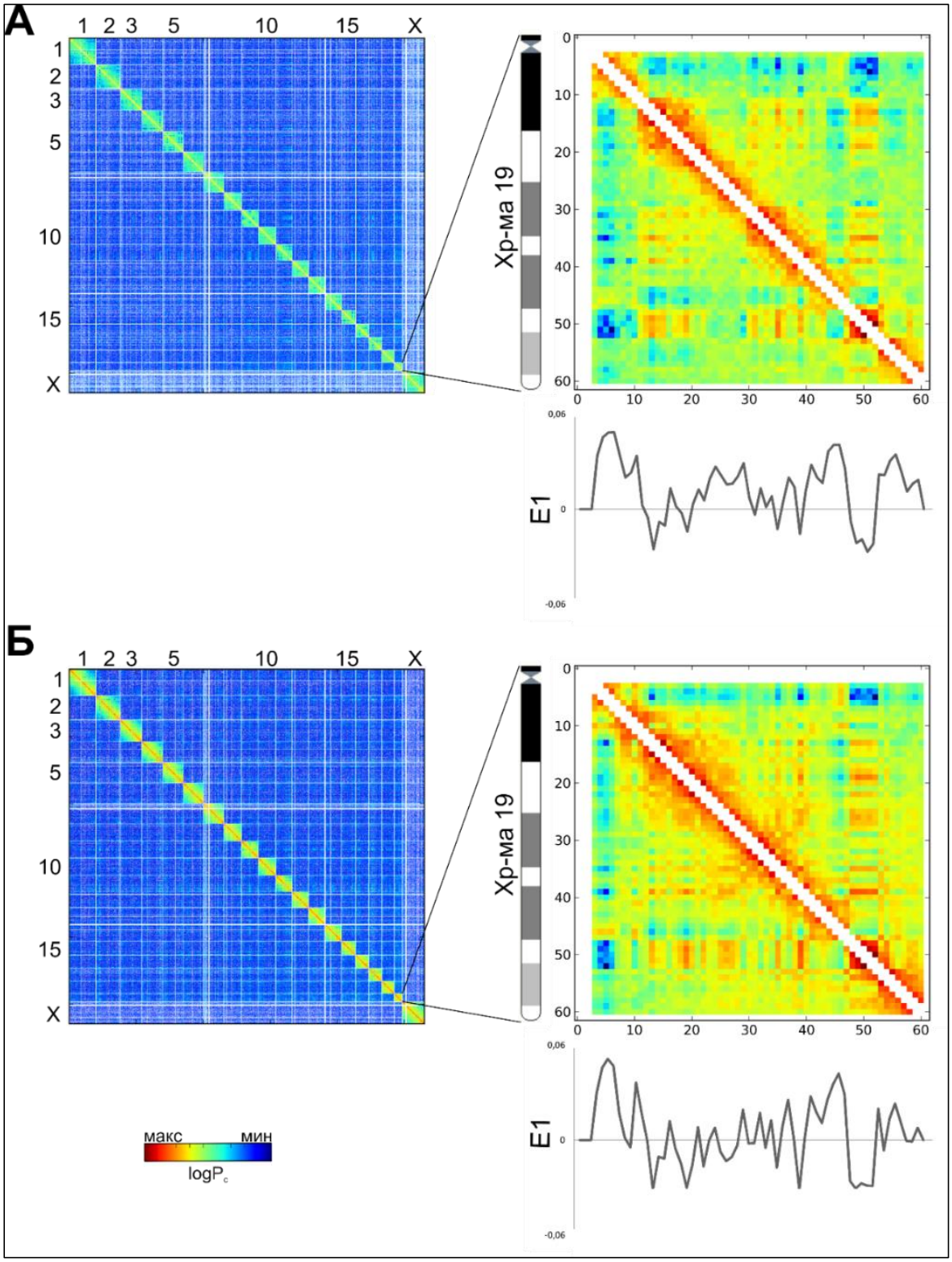


Рисунок 6. Матрицы пространственных контактов фибробластов и сперматозоидов. Показаны полногеномные (слева) матрицы пространственных контактов фибробластов (А) и сперматозоидов (Б) и внутрихромосомные контакты хромосомы 19 (справа) для этих типов клеток. Обозначения как на рис. 5. Под матрицей контактов хромосомы 19 показаны значения E^1 соответствующих локусов. Матрицы приведены с разрешением 1 Mb, после коррекции.

Таблица 1. Коэффициенты корреляции Спирмена для значений E^1 матриц пространственных контактов фибробластов, сперматозоидов, ЭСК и клеток кортекса

	Сперматозоиды	Фибробласты	Кортекс	ЭСК
Сперматозоиды	1,000	0,899	0,901	0,878
Фибробласты	0,899	1,000	0,823	0,821
Кортекс	0,901	0,823	1,000	0,893
ЭСК	0,878	0,821	0,893	1,000

Таблица 2. Значение коэффициента максимальной информации (maximal information coefficient) (Reshef et al., 2011) для значений E^1 матриц пространственных контактов фибробластов, сперматозоидов, ЭСК и клеток кортекса

	Сперматозоиды	Фибробласты	Кортекс	ЭСК
Сперматозоиды	1,000	0,673	0,693	0,651
Фибробласты	0,673	1,000	0,567	0,575
Кортекс	0,693	0,567	1,000	0,683
ЭСК	0,651	0,575	0,683	1,000

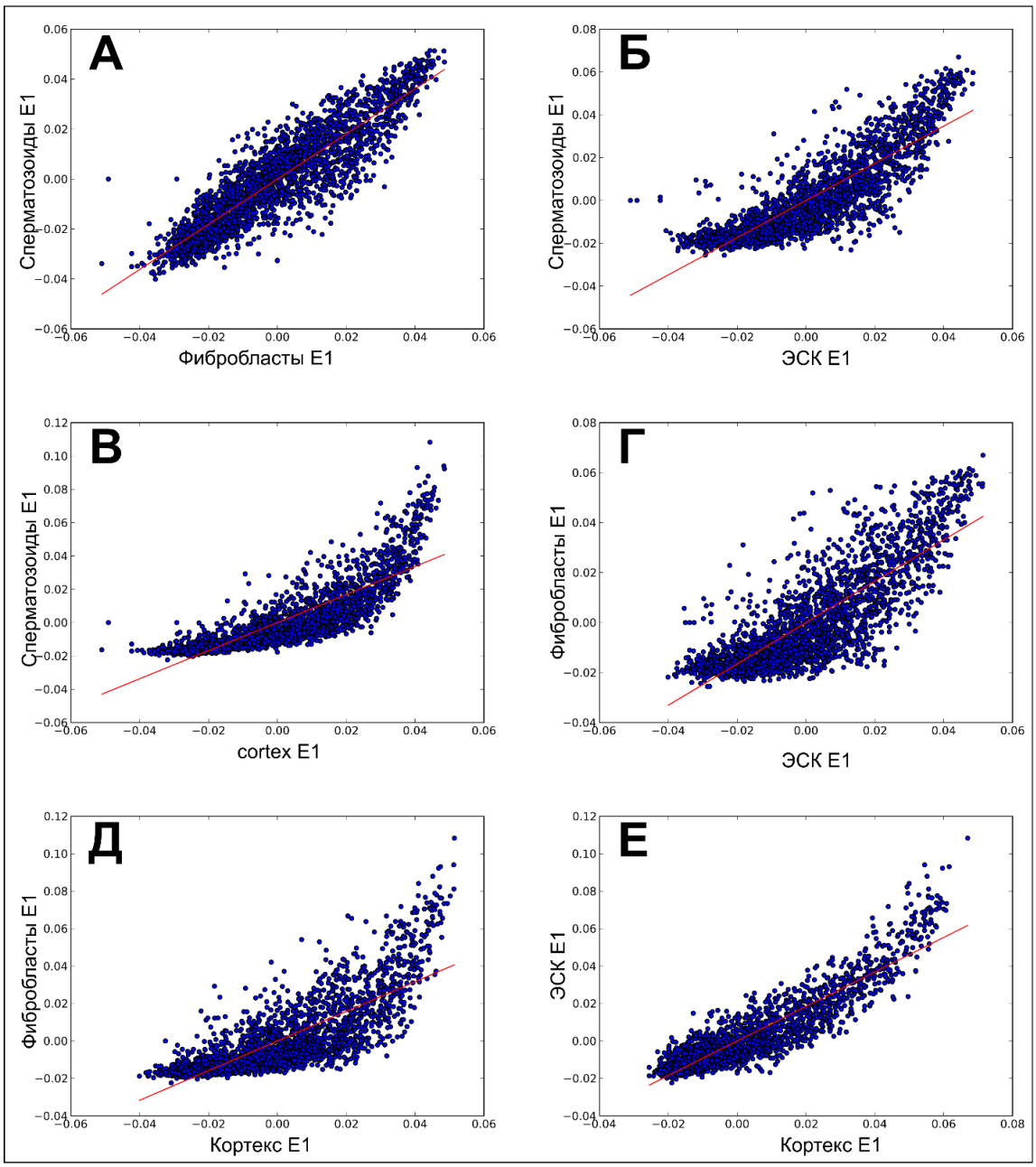


Рисунок 7. Фибробласты, сперматозоиды, ЭСК и клетки кортекса демонстрируют сходство значений E^1. Показаны попарные сравнения значений E^1 для четырех типов клеток. Каждая точка графиков (А-Е) отображает значение E^1 для одного из бинов матрицы пространственных контактов. Абсцисса каждой точки соответствует значению E^1 данного бина в типе клеток, обозначенного под осью абсцисс, ордината - значению E^1 в типе клеток, обозначенного сбоку от оси ординат. На каждом графике показана линия линейной регрессии.

Анализ TAD-доменов в геномах сперматозоидов и фибробластов.

Нами был проведен поиск TAD-доменов в геномах фибробластов и сперматозоидов. Мы идентифицировали 2 590 TAD-доменов в геноме фибробластов, средний размер которых составлял 928 kb (медиана распределения размеров доменов составила 680 kb). Число TAD-доменов фибробластов и TAD-доменов, описанных ранее для ЭСК (2 200 доменов с медианой 880 kb (Dixon et al., 2012)) было сходным. Интересно, что число доменов, идентифицированных в сперматозоидах, оказалось несколько меньшим: 1 856 доменов со средним размером 1 226 kb и медианой 1 000 kb (рис. 8).

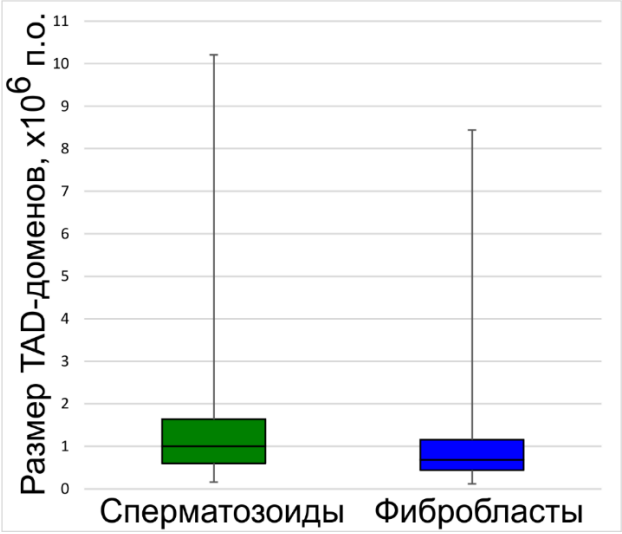


Рисунок 8. Размеры TAD-доменов фибробластов и сперматозоидов. Диаграмма размаха показывает размеры TAD-доменов (указаны по оси Y, в миллионах п.о.) для фибробластов и сперматозоидов.

Проанализировав значение переменной DI (domain index, переменная, значение которой вычисляется для каждого участка генома и косвенно характеризует его принадлежность к домену; описана более подробно в главе «Материалы и методы») в данных фибробластов и сперматозоидов, мы обнаружили высокий уровень сходства (коэффициент корреляции Спирмена 0,759). Это свидетельствует о высоком сходстве TAD-доменов фибробластов и сперматозоидов. Сходство также подтверждается при визуальном анализе TAD-доменов этих клеток (рис. 9).

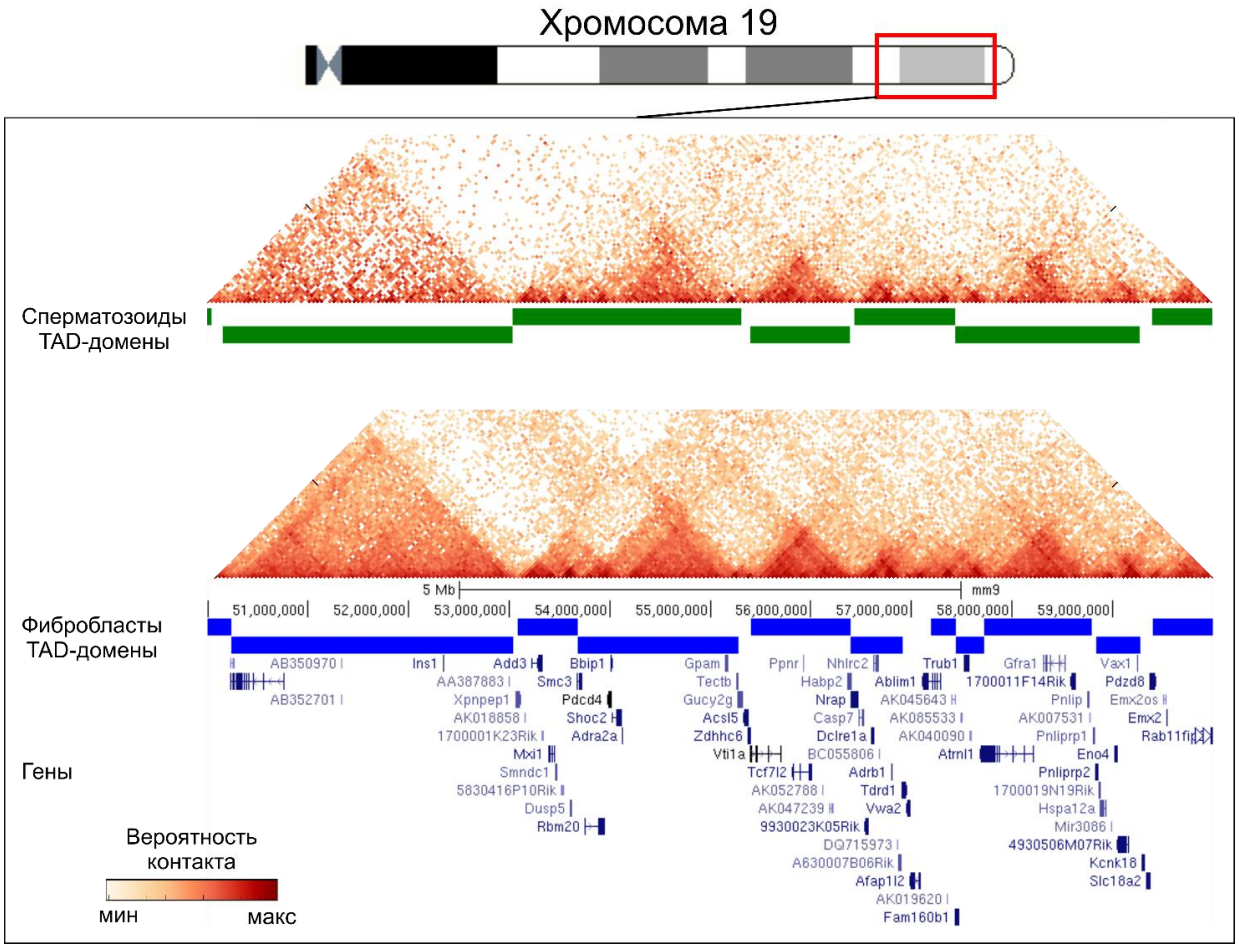


Рисунок 9. TAD-домены в фибробластах и сперматозоидах. Репрезентативный участок генома (район хромосомы 19), представлен матрицей пространственных контактов с разрешением 40 000 п.о. На данном рисунке, показан участок матрицы с одной стороны от главной диагонали, развернутый на 90 градусов. Вероятность контакта отражена цветом в соответствии со шкалой в нижней части рисунка. Границы доменов показаны под матрицей в виде прямоугольников: синего цвета для фибробластов, зеленого для сперматозоидов. Внизу также отмечены гены, расположенные в данном участке генома, в соответствии с информацией геномного браузера USCS.

Для того, чтобы объяснить различия в размерах TAD-доменов фибробластов и сперматозоидов, мы предприняли более детальное сравнение их локализации в геномах этих клеток. Мы обнаружили, что в ряде случаев расположенные последовательно в геноме фибробластов домены «сливаются» друг с другом в геноме сперматозоидов (рис 9). Этот эффект может частично объяснить увеличение средних размеров TAD генома сперматозоидов по сравнению с фибробластами.

Таким образом, нами были выявлены TAD-домены в геномах фибробластов и сперматозоидов. Домены демонстрировали сходную локализацию в этих клетках,

однако домены сперматозоидов были несколько большими, чем домены фибробластов. Кроме того, в некоторых случаях TAD-домены фибробластов, расположенные друг рядом с другом, сливались в один домен в сперматозоидах.

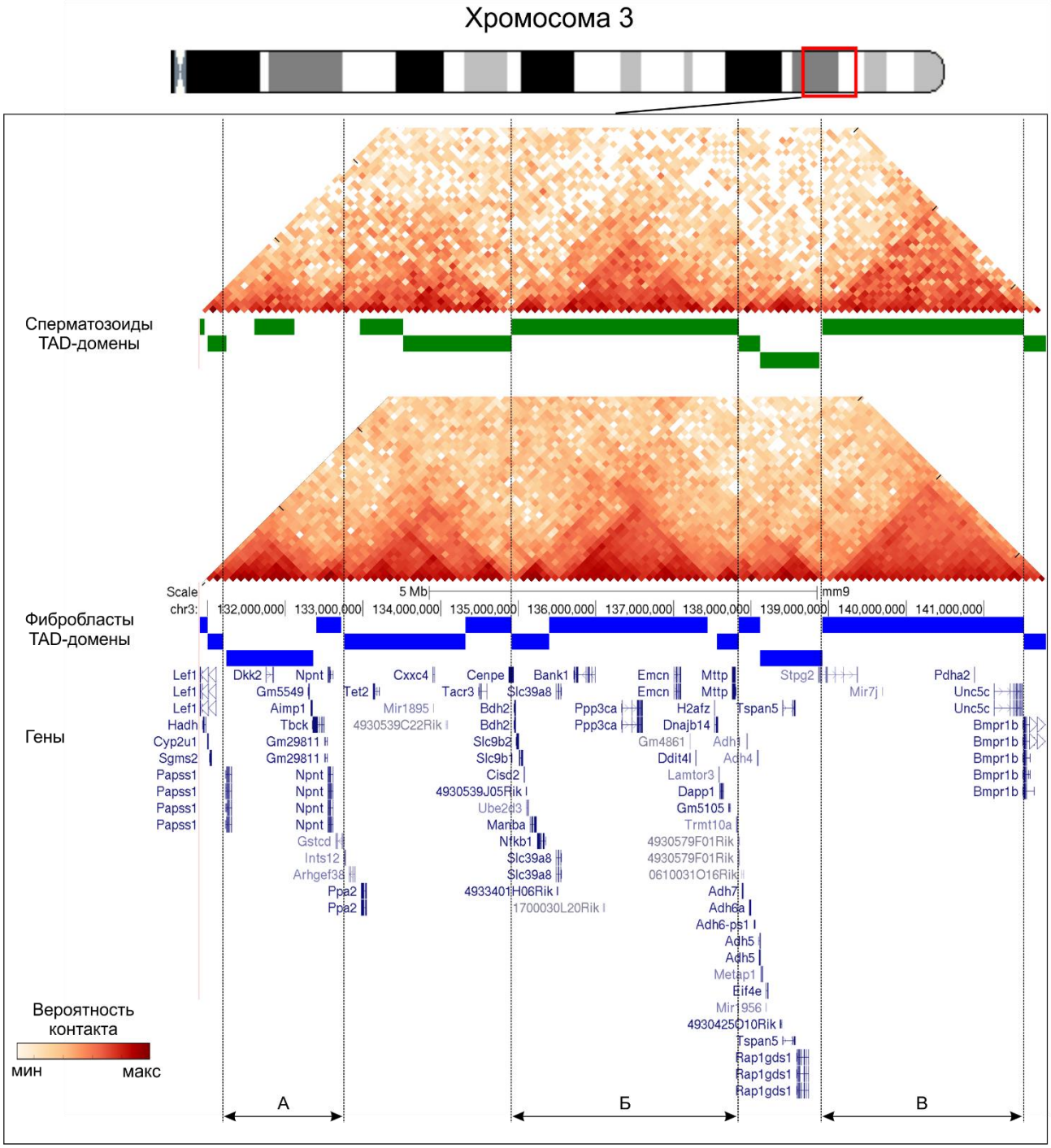


Рисунок 10. Различия TAD-доменов фибробластов и сперматозоидов. Репрезентативный участок генома (район хромосомы 3), представлен как на рис. 9. На рисунке можно видеть три различных региона: на одном из них TAD-домены фибробластов и сперматозоидов различаются (А), на другом – домены фибробластов сливаются в сперматозоидах (Б) и на третьем домены совпадают (В).

Поиск различий в пространственной укладке геномов сперматозоидов и фибробластов

Несмотря на вышеописанное общее сходство матриц пространственных контактов фибробластов и сперматозоидов, даже при их визуальном анализе можно выделить небольшое количество районов, контакты которых отличаются. О присутствии таких контактов свидетельствует также наличие несовпадающих или сливающихся TAD-доменов и отличный от единицы коэффициент корреляции Спирмена значений E^1.

Различия в укладке определенных локусов

Мы использовали ряд методов для поиска отдельных локусов, контакты которых различаются в фибробластах и сперматозоидах: коэффициенты корреляции Пирсона и Спирмена, Евклидово расстояние и близость значений E^1. Каждый из этих методов имеет свои преимущества и лимитирующие факторы, и, кроме того, разные методы могут выявлять различия, связанные с разными биофизическими особенностями. Мы получили вышеупомянутые характеристики (коэффициенты корреляции, Евклидово расстояние и значения E^1) для индивидуальных бинов (рис. 11, А). Общее сходство пространственной укладки сперматозоидов и фибробластов регистрируется всеми тремя методами (рис. 11, А): оно выражается в высоких коэффициентах корреляции, небольшом Евклидовом расстоянии и корреляции значений E^1. Можно отметить небольшое уменьшение коэффициентов корреляции для участков, расположенных в центре хромосомы, по сравнению с участками на её концах, связанное, вероятнее всего, с особенностями распределения сигналов и шумов в структуре пространственных контактов.

На разрешении 1 Mb, геном мыши представлен в полученных нами данных приблизительно 2400 бинами. Для того, чтобы выявить наиболее сильно различающиеся по пространственной укладке участки, мы выбирали определенное число (TopN) наиболее различающихся бинов, используя каждую из трех вышеперечисленных методик сравнения независимо. Затем, мы определяли интересующие нас бины (3xTopN) как попавшие в TopN для каждой из трех методик.

Для того, чтобы учесть индивидуальную чувствительность каждого из методов

сравнения к природе и особенностям распределения шумов в полученных Hi-C данных, мы исключили из 3xTopN бинов те, которые попадали в TopN различающихся при сравнении двух случайных подвыборок одной генеральной совокупности данных («референсных» выборок; детали корректировки подробно описаны в главе «Материалы и методы»). Таким образом, мы получили скорректированный набор наиболее различающихся участков генома 3xTopN_corr. Набор бинов 3xTopN_corr содержит участки, наиболее сильно различающиеся в фибробластах и сперматозоидах, выявленные тремя разными математическими методами, с учетом индивидуальной чувствительности каждого из этих методов к распределению шума в данных Hi-C.

Очевидно, что выбор числа TopN определяет количество бинов, попадающих в список 3xTopN_corr (см. приложение 2). Для больших TopN (более половины всех бинов), список 3xTopN будет большим, однако его размер будет резко уменьшаться после коррекции (например, для TopN равного числу бинов в геноме все участки ДНК будут отфильтрованы из 3xTopN после коррекции), и результирующий список 3xTopN_corr будет маленьким. Больший интерес представляют небольшие значения TopN и соответствующие им списки 3xTopN_corr, содержащие наиболее сильно различающиеся по пространственной укладке ДНК бины. Для значений TopN до 20 нами не было получено ни одного бина в списках 3xTopN_corr (приложение 2, Б). При увеличении TopN вплоть до 120 (что соответствует выбору ~5% наиболее различающихся бинов в геноме), мы идентифицировали 7 локусов, попавших в список 3xTopN_corr (таблица 3). Идентифицированные регионы располагались на хромосомах 5, 12, 13 и 19, причем три из семи участков оказались расположенными на хромосоме 19 (что, при случайном распределении участков, должно быть маловероятным, поскольку хромосома 19 является самой короткой в мышином геноме). Важно отметить, что для всех проанализированных значений TopN (начиная от 30) полученное число локусов в списке 3xTopN_corr было значимо больше, чем ожидаемое (см. приложение 2).

Таким образом, мы показали значительное сходство пространственной укладки ДНК фибробластов и сперматозоидов, выявленное всеми тремя использованными подходами. Кроме того, нам удалось определить ряд участков в геномах

исследованных клеток, имеющих наиболее существенные различия в пространственной организации.

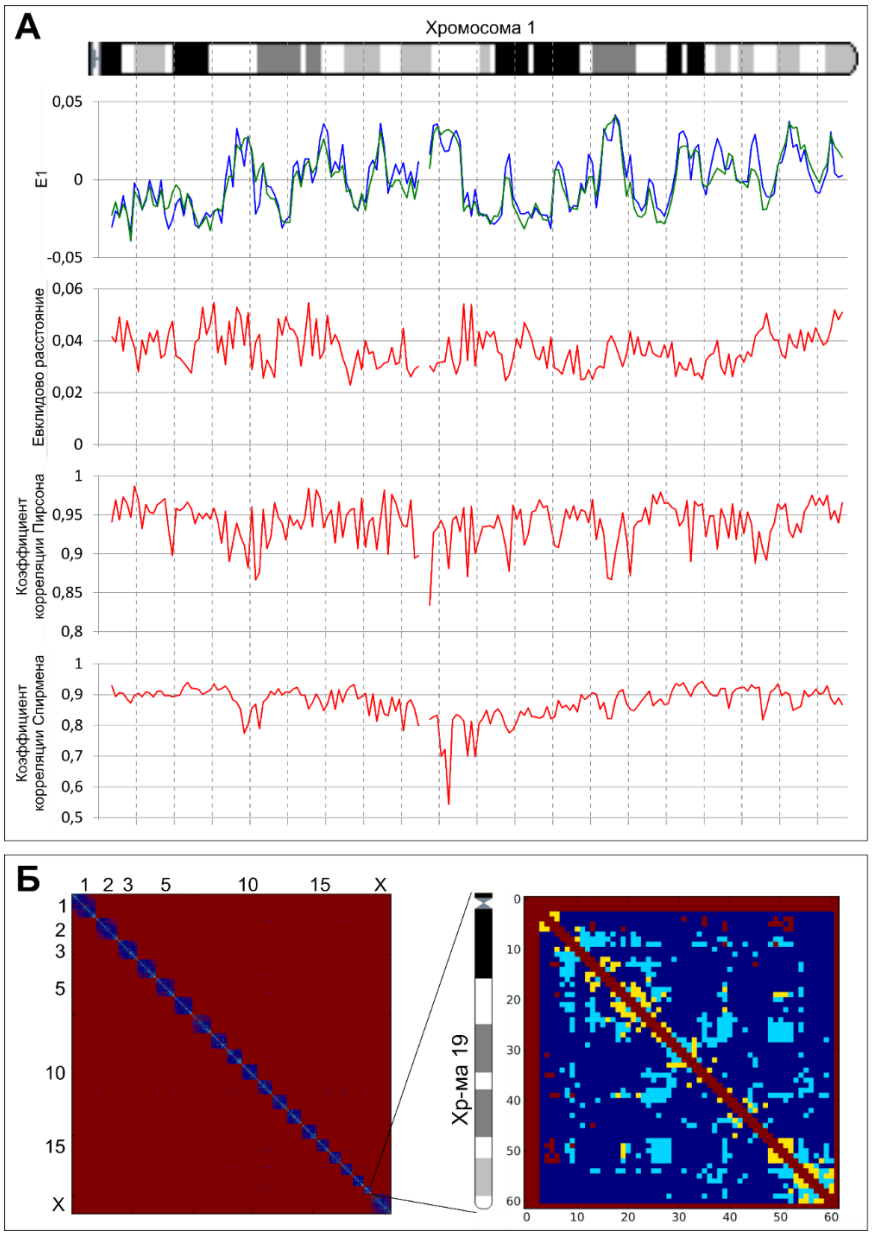


Рисунок 11. Сравнение пространственных контактов в геномах сперматозоидов и фибробластов. А. Сравнение индивидуальными методами локусов в геноме: приведены (сверху вниз) значения E^1, Евклидово расстояние и коэффициенты корреляции Пирсона и Спирмена для всех бинов 1-ой хромосомы сперматозоидов и фибробластов. Б. Показаны результаты анализа различий в частотах индивидуальных контактов для всего генома (слева) и хромосомы 19 (справа). Результаты представлены в виде матрицы, в столбцах и строках которой находятся бины, а в ячейках – результаты сравнения контактов данных бинов в цветовой шкале. Синий цвет означает неинформативный контакт, розовый – контакт, для которого показаны статистически значимые различия у фибробластов и сперматозоидов, желтый – статистически значимые, более чем двукратные различия.

Различия в индивидуальных контактах

Пространственная организация генома может различаться не только на уровне укладки целых локусов, но и за счет индивидуальных различий в попарных контактах отдельных локусов. Мы разработали статистический метод, позволяющий оценить различия в частотах индивидуальных контактов двух типов клеток (описан в главе «Материалы и методы»). При помощи этого метода, мы протестировали в сперматозоидах и фибробластах все контакты, представленные более чем 1 ридом («информативные» контакты). Оказалось, что как для фибробластов, так и для сперматозоидов, число информативных контактов составляет около 153 000 (~4,31 %) из потенциальных 3.5×10^7 контактов при разрешении 1 Mb. Из 153 363 информативных для этих типов клеток контактов, 8 947 (5,85%) имели статистически значимые различия в частотах взаимодействия (уровень значимости q-value <0.05). Более того, из этих 8 947 взаимодействий, для 6 586 контактов частоты различались более чем в 2 раза (рис. 11, Б). Следует отметить, что вышеупомянутые участки хромосомы 19, различающиеся по укладке в сперматозоидах и фибробластах, имели в этих клетках большое количество статистически значимо различающихся контактов с различными локусами генома (рис. 11, Б).

Таблица 3. Участки, различающиеся по пространственной организации в сперматозоидах и фибробластах

Хромосома	Первый нуклеотид локуса (в соответствии с картой мышиного генома mm9)	Последний нуклеотид локуса (в соответствии с картой мышиного генома mm9)
Хромосома 5	119000000	120000000
Хромосома 5	142000000	143000000
Хромосома 12	35000000	36000000
Хромосома 13	57000000	58000000
Хромосома 19	54000000	55000000
Хромосома 19	8000000	9000000
Хромосома 19	9000000	10000000

Анализ зависимости частоты контактов локусов от расстояния в линейной молекуле

Анализ зависимости частоты контактов локусов (P) от расстояния в линейной молекуле (s), P(s), позволяет определить наиболее вероятный тип укладки ДНК в ядре (Mirny, 2011; Lieberman-Aiden et al., 2009; Naumova et al., 2013). Мы рассчитали зависимость P(s) для сперматозоидов и фибробластов и обнаружили, что и для тех, и для других, наблюдалось сильно выраженное уменьшение числа контактов с ростом расстояния между локусами в линейной молекуле. Для сперматозоидов зависимость выражалась как $P(s) \sim s^{-1,07}$, для фибробластов - $P(s) \sim s^{-1,27}$ (рис. 12, А). Мы оценили стандартную ошибку показателей степени (-1,07 и -1,27) как не более, чем 0,01 и показали, что значения показателей степеней статистически значимо различаются между собой. Полученные значения также значимо отличались от значения -1 ($P(s) \sim s^{-1}$), характерного для гипотетической идеальной фрактальной глобулы ДНК. Однако, тип упаковки сперматозоидов была более близким к фрактальной, чем упаковки фибробластов.

Интересно, что для фибробластов частоты контактов локусов, удаленных менее чем на 10 Mb., были выше, чем для сперматозоидов. Это различие компенсировалось увеличением частот контактов локусов в сперматозоидах, расположенных на расстоянии 10^7-10^8 п.о. Эти данные означают, что сперматозоиды имеют больше контактов между удаленными локусами, чем фибробласты. Более детальный анализ, результаты которого представлены на рисунке 12, Б, показал, что частоты контактов в фибробластах были выше для регионов, распложенных ближе, чем 40 Mb. Для регионов, разделенных 50-150 Mb, сперматозоиды показывали более чем двукратное увеличение частот контактов, по сравнению с фибробластами.

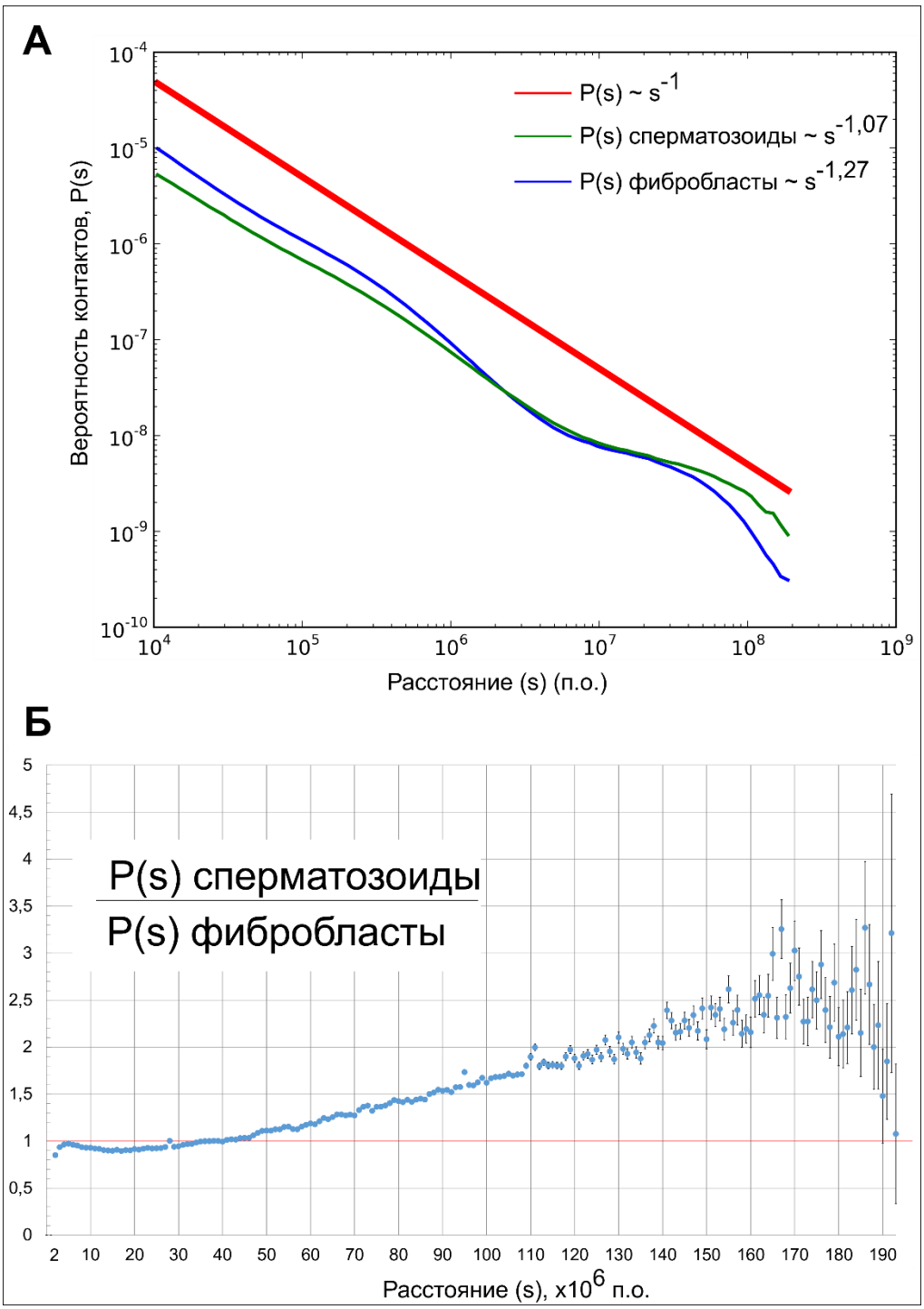


Рисунок 12. Геном сперматозоидов обогащен взаимодействиями удаленных участков. А. Приведен график зависимости P(s) для фибробластов, сперматозоидов и теоретический график, ожидаемый для идеальной фрактальной глобулы. Б. Показано соотношение количества контактов в фибробластах и сперматозоидах, в зависимости от расстояния между локусами в линейной молекуле ДНК. По оси абсцисс отложено расстояние между локусами (в миллионах п.о.), а по оси ординат – усредненное по всем локусам, расположенным на данном расстоянии, отношение частот контактов сперматозоидов и фибробластов. Для каждой точки указан размер ошибки. Горизонтальная линия на уровне 1 соответствует равной частоте контактов сперматозоидов и фибробластов.

Чтобы выявить влияние различий P(s) на характеристики пространственной организации, мы, используя алгоритм BACH (Hu et al., 2013), смоделировали пространственную структуру TAD-доменов сперматозоидов и фибробластов. После моделирования, мы представили каждый TAD-домен как цилиндр, и измерили соотношение длины такого цилиндра и его радиуса (HD-ratio). У более вытянутых цилиндров такое соотношение будет больше, чем у более компактных, (минимальное HD-ratio цилиндра совпадает с HD-ratio шара и равно 1). Если представлять себе компактизацию генома, при которой происходит линейное сжатие вдоль какого-либо направления, то в ходе этого процесса HD-ratio будет уменьшаться. Мы обнаружили, что TAD-домены сперматозоидов являются более компактными (имеют меньшее HD-ratio), чем TAD-домены фибробластов (последние являются более «вытянутыми»). Различия в значениях HD-ratio для TAD-доменов фибробластов и сперматозоидов, показанные в приложении 3, являются статистически значимыми.

Таким образом, нами было показано степенное падение частоты контактов с ростом линейного расстояния между участками ДНК для сперматозоидов и фибробластов. При этом сперматозоиды демонстрировали большее количество контактов между удаленными на значительное (более 40 Mb) расстояние бинами. Кроме этого, мы показали, что TAD-домены сперматозоидов являются более компактными, сжатыми, по сравнению с TAD-доменами фибробластов.

Анализ межхромосомных контактов в геномах фибробластов и сперматозоидов

В ряде работ было показано, что в геномах клеток млекопитающих количество межхромосомных контактов много меньше, чем внутрихроимосомных (Lieberman-Aiden et al., 2009; Kalhor et al., 2012; Rao et al., 2014). Мы также наблюдали эту тенденцию в наших картах пространственных контактов: более чем 90% всех контактов в рассмотренных клетках приходились на внутрихромосомные (рис. 5).

Выявление статистически достоверных различий в частотах отдельных межхромосомных контактов не представляется возможным из-за их небольшого количества. Однако возможно провести статистический анализ, суммируя определенные категории межхромосомных контактов: например, все

межхромосомные контакты одной хромосомы.

Мы рассчитали соотношение внутри- и межхромосомных контактов для каждой хромосомы фибробластов и сперматозоидов (рис. 13, А). Мы обнаружили, что для всех хромосом наблюдаются две одинаковые тенденции. Во-первых, число межхромосомных контактов в 10-40 раз меньше, чем внутрихромосомных. Во-вторых, в сперматозоидах это соотношение контактов (внутри- к межхромосомным) ниже, чем в фибробластах: для сперматозоидов оно составляет 10-20 раз, тогда как для фибробластов – 20-40 раз. Это означает, что в сперматозоидах наблюдается много больше межхромосомных контактов, чем в фибробластах.

Мы также оценили частоты контактов индивидуальных хромосом друг с другом (рис. 13, Б и В). Оказалось, что длинные хромосомы (1-5 и X) имеют тенденцию взаимодействовать друг с другом чаще, чем с короткими (хромосомами 10-19). Эта тенденция может быть визуализирована как обогащенный красным сигналом квадрат в левом верхнем углу матриц, представленных на рисунках 13, Б и В. Наблюдалась ещё одна аналогичная тенденция – увеличение частоты контактов длинных хромосом друг с другом, по сравнению с частотами контактов длинных хромосом с короткими. Однако, это тенденция была менее выраженной. Мы подтвердили полученные результаты о преференциях межхромосомных контактов, зависящих от длин хромосом, проведя анализ распределения частот межхромосомных контактов от соотношения их длин. Нами была обнаружена обратная корреляция этих двух параметров, характеризующаяся коэффициентом Пирсона -0,44 (рис. 13, Г).

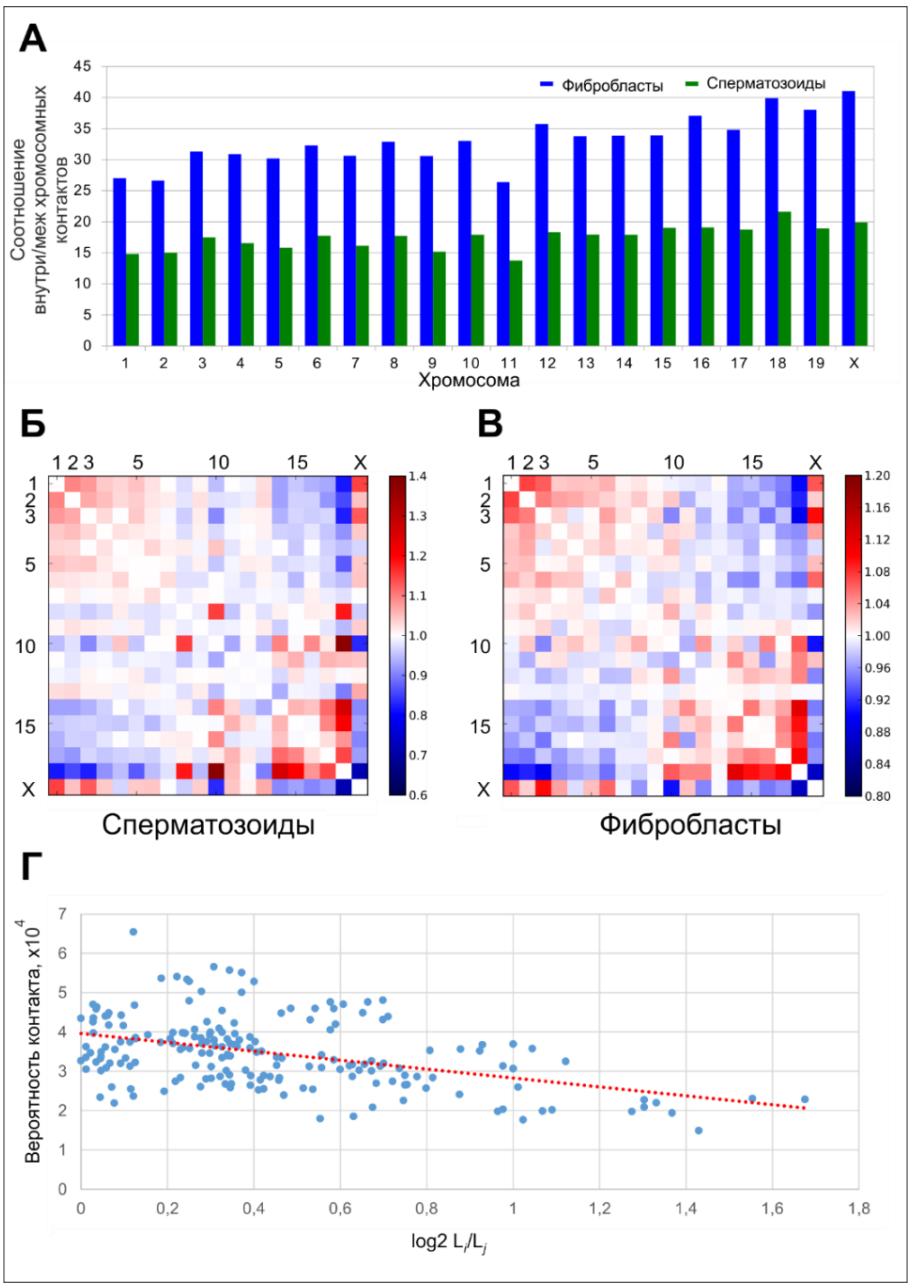


Рисунок 13. Анализ межхромосомных контактов в геномах фибробластов и сперматозоидов. А. Соотношение внутри- и межхромосомных контактов для каждой из хромосом. Б и В – матрица пространственных контактов индивидуальных хромосом для сперматозоидов (Б) и фибробластов (В). Каждый элемент матрицы относится к контактам хромосом, указанных в соответствующей строке и столбце. Элементы матрицы отражают соотношение полученной в эксперименте частоты контактов к ожидаемому для случайного (равномерного) распределения, в соответствии с цветовой шкалой приведенной сбоку. Красный цвет соответствует превышению числа контактов над ожидаемым, синий – уменьшение по сравнению с ожидаемым. Г. Зависимость числа контактов между хромосомами “i” и “j” от двоичного логарифма соотношения их длин, $\log_2(L_i \backslash L_j)$. Для хромосом одинаковой длины $\log_2(L_i \backslash L_j)$ стремится к 0, для хромосом различной – к бесконечности. Таким образом, чем дальше от начала координат находится точка на оси абсцисс, тем больше отношение длин хромосом контакты которых она отражает. График приведен для сперматозоидов, для фибробластов зависимость выглядит сходно.

Влияние компактизации генома сперматозоидов на особенности пространственной организации этих клеток

Различия в пространственной организации сперматозоидов и фибробластов, описанные выше, могут иметь как минимум две потенциальные причины. Во-первых, характерная для сперматозоидов более плотная упаковка генома, связанная с уменьшением размера ядра и более плотной упаковкой ДНК при помощи протаминов, приводит к локус-неспецифическим изменениям контактов в масштабе всего генома и может, в конечном счете, быть причиной различий в частотах индивидуальных контактов. Во-вторых, различия в частотах контактов могут объясняться локальными, локус-специфическими изменениями в структуре пространственных взаимодействий, имеющие значение для функционирования данных локусов.

Чтобы оценить роль этих причин (неспецифической полногеномной компактизации и локус-специфичного изменения паттерна контактов), мы разработали метод «компрессии» генома соматических клеток (фибробластов) в соответствии с параметрами сперматозоидов. Проводя «компрессию» генома соматических клеток, мы хотели бы получить матрицу пространственных контактов ДНК «теоретической клетки», геном которой компактизован также, как и в сперматозоидах, но содержит локус-специфические контакты, характерные для исходных соматических клеток. Идея метода состоит в том, чтобы, не меняя распределение частот контактов для локусов, расположенных в линейной молекуле на одинаковом расстоянии, сблизить все локусы друг с другом. Таким образом, «компрессия» фибробластов в соответствии с параметрами сперматозоидов приводит к тому, что функция P(s) для этих клеток становится одинаковой, однако при этом в фибробластах сохраняется соотношение контактов для локусов, расположенных на одинаковом расстоянии друг от друга. Мы обозначили пространственную структуру, полученную после компрессии, как C_{sp}-фибробласты («компрессированные» по типу сперматозоидов фибробласты). Мы обнаружили, что число значимо различающихся контактов между C_{sp}-фибробластами и сперматозоидами после компрессии уменьшилось приблизительно на 25% (рис. 14),

и составило 6 962 (до компрессии – 8 974). Кроме того, число контактов, для которых различия в частоте составили более чем 2 раза, уменьшилось с 6 586 (до компрессии) до 5 009.

В качестве контроля, мы выполнили компрессию фибробластов по типу ЭСК, получив C_{ESC}-фибробласты. Общее число статистически различающихся контактов в сперматозоидах и C_{ESC}-фибробластах составило 10 848, что приблизительно на 20 % больше, чем число различий между сперматозоидами и фибробластами до компрессии. Число контактов, для которых разница частот была более чем двукратной, также увеличилось до 8 776 (на 33%). Следует отметить, что при этом C_{ESC}-фибробласты были более сходны с ЭСК, чем исходные фибробласты (рис. 14). Это свидетельствует о том, что компрессия уменьшает число различий в пространственной организации между разными типами клеток только тогда, когда она проводиться в соответствии со специфическими для данных типов клеток параметрами.

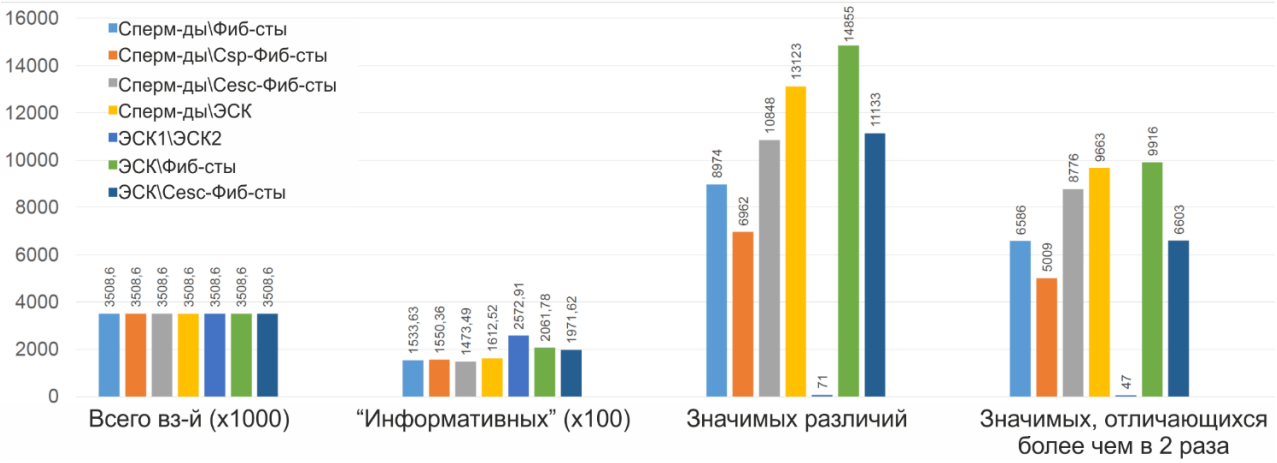


Рисунок 14. Компрессия генома объясняет часть различий между сперматозоидами и фибробластами. На гистограмме приведены (слева направо): общее количество потенциальных взаимодействий для матрицы контактов с разрешением 1 Mb, число «информативных» взаимодействий, число статистически значимо различающихся контактов, и число значимо различающихся контактов, для которых различие как минимум двукратно. Первые две категории отображены в 1000 и 100 кратном масштабе. Данные приведены для результатов попарного сравнения сперматозоидов, фибробластов, C_{sp}-фибробластов, C_{ESC}-фибробластов, а двух «референсных» выборок (ЭСК1 и ЭСК2), описанных в главе «Материалы и методы».

Таким образом, наша модель «компрессии» предсказывает, что около 25%

различий в частотах индивидуальных контактов между сперматозоидами и фибробластами может объясняться различиями в плотности упаковки ДНК в этих типах клеток, в то время как большинство различий связаны с другими причинами, среди которых можно выделить локус-специфические особенности пространственной организации.

Обсуждение

Построение матрицы пространственных контактов фибробластов и сперматозоидов

В данной работе нами впервые были получены данные о пространственной организации сперматозоидов и фибробластов мыши. Технология приготовления Hi-C библиотек соматических клеток была описана ранее (Lieberman-Aiden et al., 2009), однако получение таких библиотек для сперматозоидов, описанное нами, расширяет область применения метода Hi-C. В ряде работ было показано, что методы приготовления геномных библиотек, включающие этап ферментативного гидролиза ДНК без удаления белков хроматина, мало эффективны в случае их применения для сперматозоидов, поскольку плотная упаковка геномов этих клеток затрудняет доступ фермента гидролиза (Carone et al., 2014). Ряд параметров, выявленных при анализе первичных данных секвенирования Hi-C библиотеки сперматозоидов, указывает на её более низкое качество, по сравнению с библиотекой фибробластов. Это особенно заметно при сравнении доли ридов, не прошедших фильтры, для сперматозоидов и фибробластов: для последних она намного меньше (рис. 3). Однако, остается неясным является ли данный эффект результатом особенностей гидролиза более плотно упакованной ДНК сперматозоидов, или техническими артефактом приготовления библиотеки.

Нами были построены матрицы пространственных контактов для четырех типов клеток: фибробластов, сперматозоидов, ЭСК и клеток кортекса (последние два типа клеток по литературным данным). Как было отмечено в результатах, даже матрицы с низким разрешением имели достаточно большую среднюю погрешность в частотах контактов (порядка 20%). Однако, нужно отметить, что сходный результат был получен и в других экспериментах Hi-C (Lieberman-Aiden et al., 2009). Кроме того, нужно понимать, что размер ошибки для каждого контакта вычисляется исходя из его частоты. Чем выше частота контакта, тем больше ридов его поддерживает, и тем меньше будет ошибка. Анализируя график зависимости P(s), можно увидеть, что частоты пространственных контактов крайне быстро убывают с ростом расстояния между локусами в линейной молекуле. Поэтому, ошибки в частотах

пространственных контактов распределены крайне неравномерно. Для взаимодействий, удаленных на небольшое расстояние (порядка нескольких десятков бинов) фрагментов, ошибка в определении частоты много меньше, чем заявленная средняя, в то время как для разнесенных на сотни миллионов п.о. участков частоты контактов практически недостоверны. Этот эффект легко заметить, рассматривая ошибки точек на графике рисунка 12 Б (эти ошибки пропорциональны ошибкам индивидуальных контактов участков на заданном расстоянии). Такую неравномерность в распределении контактов отмечают и другие исследователи (Ay et al., 2014a; Duan et al., 2010). Однако следует отметить, что практически все виды анализа, приведенные в статье, либо оценивали индивидуальные ошибки частот взаимодействий и оперировали с контактами, для которых была показана высокая достоверность (как, например, статистический анализ различий частот индивидуальных контактов), либо проводились для близлежащих участков (как, например, построение TAD-доменов).

Сходство пространственной укладки геномов фибробластов и сперматозоидов

Несмотря на целый ряд принципиальных аспектов, различающих сперматозоиды и фибробласты, пространственная укладка генома в этих клетках демонстрирует поразительное сходство, по крайней мере, на масштабе миллионов п.о. Это проявляется как в визуальном сходстве матриц пространственных контактов, так и в наличие в обоих типах клеток базовых пространственных структур: A/B- и TAD-доменов. Кроме того, более чем 90% индивидуальных взаимодействий имели одинаковые частоты в геномах фибробластов и сперматозоидов.

Следует отметить, что при анализе значений E^1 мы выявили не только сходство фибробластов и сперматозоидов, но и их близость к другим типам клеток: ЭСК и клеткам кортекса. При этом сперматозоиды оказались по этому параметру даже несколько ближе к клеткам кортекса, чем к фибробластам. Этот нюанс важен, поскольку данные о пространственной организации генома в клетках кортекса и ЭСК получены оригинальным методом Hi-C (Dixon et al., 2012), тогда как библиотеки фибробластов и сперматозоидов получены модифицированным методом TCC

(Kalhor et al., 2012). Поэтому близость данных E^1 сперматозоидов и клеток кортекса, полученных с применением разных методов, показывает, что выявленные сходства отражают общность в закономерностях пространственной укладки, а не технические артефакты, специфические для методов Hi-C и TCC. Такая интерпретация хорошо согласуется отмеченным в других работах сходством результатов ($R>0.95$, показано в (Kalhor et al., 2012; Rao et al., 2014)), полученных разными модификациями методов Hi-C, таких как «классический» метод Hi-C (Lieberman-Aiden et al., 2009), TCC (Kalhor et al., 2012) и недавно предложенным *in situ* Hi-C (Rao et al., 2014).

Идентификация TAD-доменов в сперматозоидах

На сегодняшний день неизвестно, поддерживаются ли TAD-домены в клетках за счет каких-либо активных механизмов, или их формирование является побочным результатом процессов транскрипции и упаковки ДНК на нуклеосомном уровне (подробней этот вопрос обсуждается в главе «Обзор литературы»). В сперматозоидах, упаковка ДНК отличается от упаковки соматических клеток на базовом уровне: сохранено не более 10% нуклеосом, большая часть гистонов заменена протаминами (Mudrak et al., 2011; Hammoud et al., 2009; Carone et al., 2014). Более того, в сперматозоидах не происходит транскрипция, которая также играет большую роль в формировании пространственных петель и, следовательно, в поддержании структуры топологических доменов (De Laat et al., 2003). Присутствие TAD-доменов в сперматозоидах показывает, что вышеперечисленные элементы не являются необходимыми для их поддержания.

В ряде работ было показано, что белок CTCF присутствует в зрелых сперматозоидах (Carone et al., 2014; Tang et al., 2006). Учитывая важную роль этого фактора в процессе формирования TAD-доменов в соматических клетках (подробно обсуждается в главе «Обзор литературы»), можно предположить, что именно он обеспечивает присутствие этих пространственных структур в геноме сперматозоидов.

Различия пространственных контактов сперматозоидов и фибробластов

Несмотря на общее сходство организации геномов соматических и половых клеток, мы выявили также ряд различий между ними. Мы использовали для поиска

таких различий три различных математических метода (сравнение значений E^1, коэффициенты корреляции и Евклидово расстояние). Два метода (сравнение значений E^1 и использование коэффициентов корреляции) были предложены ранее (Kalhor et al., 2012; Lieberman-Aiden et al., 2009; Imakaev et al., 2012), но модифицированы в данной работе. Мы впервые использовали Евклидово расстояния для оценки сходства матриц контактов.

Несмотря на то, что пересечение результатов, полученных тремя различными методами, оказалось много больше ожидаемого для случайной выборки различающихся бинов, оно (пересечение) оказалось далеко от 100%. Одним из объяснений этого феномена может быть то, что разные математические методы при сравнении пространственной укладки локусов отражают разные биологические особенности этих локусов. Следует подчеркнуть, что до недавнего времени разные группы исследователей зачастую использовали какой-либо один из доступных методов сравнения матриц пространственных контактов (Lieberman-Aiden et al., 2009; Hou et al., 2012; Vietri Rudan et al., 2015). Кажется резонным проведение в будущем целенаправленных исследований, которые сравнивали бы различные математические методы анализа Hi-C данных систематически и предложили, в итоге, оптимальный метод.

Сравнивая укладку ДНК фибробластов и сперматозоидов в масштабе всего генома, мы обнаружили целый ряд свидетельств более плотной упаковки последней. Во-первых, сперматозоиды имеют больше контактов между отдаленными (в линейной молекуле) участками. Это хорошо согласуется с представлением о том, что их геном является более компактным, сжатым, так что удаленные участки оказываются ближе друг к другу, чем в геноме фибробластов. Во-вторых, TAD-домены фибробластов оказались более вытянутыми, а домены сперматозоидов – более компактными. Это также можно объяснить линейным сжатием генома. В-третьих, в сперматозоидах наблюдается большее количество межхромосомных контактов, что может быть логично объяснено сближением отдельных хромосом друг с другом в компактном ядре сперматозоидов.

Оценивая частоты индивидуальных контактов, мы обнаружили, что приблизительно 5% из них различают фибробласты и сперматозоиды. Для того,

чтобы оценить роль компактизации ДНК в формировании этих различий, мы разработали метод нормализации, учитывающий «компрессию» генома. В отличие от алгоритмов предложенных другими авторами (Hu et al., 2013; Mirny, 2011), наш метод не предполагает физическое моделирование структуры биополимера, а проводит математические операции с уже полученными частотами контактов.

Используя такой метод виртуальной «компрессии», мы показали, что около четверти всех различий в пространственной организации половых и соматических клеток можно объяснить равномерным, не связанным с особенностями тех или иных локусов, сжатием генома. Открытыми остаются вопросы о природе и биологической роли остальных различий пространственной организации. Можно предположить целый ряд гипотез, отвечающих на эти вопросы. Во-первых, различия могут происходить из контактов, имевших функциональное значение для регуляции транскрипции на ранних стадий сперматогенеза, и пассивно сохранившихся в транскрипционно-неактивном ядре зрелого сперматозоида. Во-вторых, особенности пространственной организации контактов сперматозоидов могут быть важны для регуляции генной экспрессии на ранних стадиях эмбриогенеза. В-третьих, различия могут происходить из-за неравномерностей локализации нуклеосом, поскольку пространственная укладка ДНК при помощи гистонов и протаминов имеет разные физические параметры (Carone et al., 2014; Erkek et al., 2013; Allen et al., 1997; Fuentes-Mascorro et al., 2000).

К сожалению, для детального изучения этих вопросов, необходимо исследование пространственной организации генома сперматозоидов с более высоким разрешением. Например, различия в укладке ДНК с помощью гистонов и протаминов должны быть наиболее выражены при исследовании пространственной организации участков с характерным размером тороидов: несколько десятков тысяч нуклеотидов (Allen et al., 1997; Fuentes-Mascorro et al., 2000). Анализ индивидуальных контактов между промоторами генов раннего эмбриогенеза и их регуляторными регионами также требует разрешения в несколько тысяч п.о.

Наконец, стоит отметить, что различия в пространственных контактах сперматозоидов и фибробластов могут быть связаны с особенностями укладки фибробластов, а не сперматозоидов. Как и любой другой тип клеток, фибробласты

имеют специфический профиль генной экспрессии, поддержание которого, вероятно, осуществляется, в том числе, за счет специфических пространственных контактов между регуляторными последовательностями.

Исследуя TAD-домены фибробластов и сперматозоидов, мы обнаружили значимые различия в их числе и размере. Слияние топологических доменов, наблюдаемое в фибробластах, а также их больший размер по сравнению с доменами сперматозоидов, логично согласовывается с данными, указывающими на большую компактизацию генома этих клеток. Однако следует отметить, что определение границ TAD-доменов может сильно зависеть от метода, которым этот поиск выполняется (обсуждается в главе «обзоре литературы» более подробно). В данной работе был использован наиболее распространенный алгоритм выявления TAD-доменов со стандартными параметрами запуска, чтобы сделать полученные данные сопоставимыми с результатами других работ (Symmons et al., 2014; Tark-Dame et al., 2014; Trimarchi et al., 2014). Учитывая вышесказанное, нужно признать, что показанные различия в TAD-доменах, могут означать не функциональное в биологическом смысле слияние доменов (или уменьшение их числа), а влияние эффекта компактизации на работу математического алгоритма, выявляющего домены. Таким образом, мы считаем, что в сперматозоидах могут присутствовать практически все TAD-домены фибробластов, но, из-за особенностей пространственной укладки, эти домены «невидимы» для математического алгоритма их поиска.

С другой стороны, нельзя исключать, что различия в структуре TAD-доменов фибробластов и сперматозоидов имеют биологическую роль. Такие, специфические для того или иного типа клеток, TAD-домены описаны в литературе (Dixon et al., 2012). Например, в геноме клеток кортекса содержится 1 519 доменов со средним размером 1,54 Mb (медиана 1,32 Mb) (Dixon et al., 2012), что отличается от клеток ЭСК и фибробластов и, более того, ближе всего к параметрам TAD-доменов сперматозоидов.

Модели укладки ДНК сперматозоидов как фрактальной и равновесной глобулы

В данной работе нам удалось получить распределение P(s) для фибробластов и сперматозоидов. Это распределение имеет значительное сходство с полученными ранее для других соматических клеток распределениями (Lieberman-Aiden et al., 2009; Naumova et al., 2013) - явно выраженное уменьшение частоты контактов с ростом расстояния в линейной молекуле. Распределение P(s) сперматозоидов по своим параметрам ($P(s) \sim s^{-1,07}$) лучше соответствовало фрактальной модели укладки ДНК ($P(s) \sim s^{-1}$), тогда как распределение фибробластов ($P(s) \sim s^{-1,27}$) – немного лучше описывалось моделью равновесной глобулы ($P(s) \sim s^{-1,5}$), чем фрактальной. Интересно, что укладка митотических хромосом значительно отличалась по функции распределения P(s) от укладки ДНК сперматозоидов, несмотря на то, что уровень компактизации митотических хромосом и генома зрелых половых клеток вполне сопоставимы (Naumova et al., 2013). Более того, в митотических хромосомах не были идентифицированы TAD-домены, а структура пространственных контактов была практически однородной (Naumova et al., 2013). Для сперматозоидов эти особенности не характерны. Из этого можно заключить, что компактизация генома сама по себе не является фактором достаточным, чтобы принципиально изменить пространственную архитектуру генома.

Пространственная организация ДНК передается в ряду поколений через геном сперматозоидов

Подводя итог, хотелось бы отметить, что сперматозоиды отвечают за передачу различных типов генетической информации в ряду поколений. Помимо передачи информации о первичной структуре ДНК, уже известно, что сперматозоиды передают и эпигенетическую информацию, в частности, через гистоновые модификации (Brykczynska et al., 2010; Erkek et al., 2013) и механизм геномного импринтинга (Bartolomei et al., 1993). Сходство организации пространственной структуры половых и соматических клеток позволяет предположить, что сперматозоиды, помимо прочего, выполняют функцию передачи информации о пространственной архитектуре генома в ряду поколений.

Выводы

1. Метод высокоэффективного конформационного захвата хромосом, в сочетании с биоинформационными алгоритмами фильтрации и нормализации данных, позволил впервые получить полногеномные карты пространственных контактов для фибробластов и сперматозоидов мыши с разрешением не менее одного миллиона п.о.

2. Анализ карт пространственных контактов фибробластов и сперматозоидов мыши позволил идентифицировать А- и В-компартменты и топологические домены, характерный размер которых составляет 680 и 1000 тысяч нуклеотидов, соответственно, в геномах исследованных клеток.

3. Сравнение пространственной укладки геномов фибробластов и сперматозоидов, проведенное тремя независимыми методами, показало высокий уровень сходства их организации. Статистический анализ различий частот индивидуальных пространственных контактов обнаружил 5% взаимодействий, различающихся в фибробластах и сперматозоидах. Выявлено семь локусов на хромосомах 5, 12, 13 и 19, трехмерная организация которых наиболее значительно различается в фибробластах и сперматозоидах.

4. Зависимость частоты контактов локусов от расстояния между ними в линейной молекуле для фибробластов и сперматозоидов характеризуется степенной функцией с коэффициентами степени -1,27 и -1,07. Укладка ДНК сперматозоидов лучше соответствует фрактальной модели, чем организация ДНК фибробластов. В ДНК сперматозоидов выявлено увеличение частоты контактов между локусами, располагающимися на расстоянии более 40 миллионов п.о., в сравнении с геномом фибробластов, что свидетельствует о большей степени компактизации генома половых клеток.

5. Разработана оригинальная модель «компрессии» генома фибробластов, позволившая установить, что около четверти выявленных различий в частотах пространственных контактов могут быть связаны с большим уровнем компактизации генома сперматозоидов, по сравнению с фибробластами.

6. Выявлены преференции в распределении межхромосомных контактов в

геномах фибробластов и сперматозоидов, заключающиеся в преимущественных контактах либо длинных хромосом 1-6 и хромосомы X, либо коротких хромосом 15-19 между собой. В целом геном сперматозоидов демонстрирует большее количество межхромосомных контактов, чем геном фибробластов, что может быть следствием его большей компактизации.

7. Несмотря на поразительные морфологические и функциональные различия, геномы фибробластов и сперматозоидов имеют сходные параметры пространственной архитектуры, что указывает на консерватизм принципов трехмерной организации ДНК в интерфазном ядре.

Благодарности

Автор выражает благодарность коллективу лаборатории генетики развития. Отдельно хочется отметить благодарностью О.Л. Серова и Н.Р. Баттулина, которые не только помогали автору ценными советами в решении конкретных задач в ходе выполнения диссертационной работы, но и внесли неоценимый вклад в формирование научного образа мышления, который, как надеется автор, у него сложился.

Данная работа не могла бы быть выполнена в полной мере без помощи Д.А. Афонникова и М.Ю. Помазного, которые объяснили автору многие технические аспекты биоинформационных методов и на всем протяжении выполнения диссертационной работы принимали участие в обсуждении полученных данных, делясь плодотворными идеями о направлениях дальнейших работ.

Автор признателен коллективу лаборатории пептидных гормонов центра им. Макса-Дельбрука, в частности, Наталье Алениной и Мишелю Бадеру, за ценные советы и за то, что они своей неиссякаемой энергией и трудолюбием подавали постоянный пример для подражания.

Неоценимый вклад в выполнение данной работы внесли родные и близкие автора. Благодарность за неустанную поддержку автор выражает, в первую очередь, своим родителям и жене. Автор благодарит свою тётю за то, что она открыла ему основы биологии и привила незаменимые в научной деятельности любознательность и трудолюбие. Автор также хочет выразить благодарность своим друзьям, в частности, А. Евдокимову, которые наполняли жизнь автора несвязанными с наукой интересами и ограничивали научное трудолюбие автора до разумных пределов.

Список литературы

Баттулин Н.Р., Фишман В.С., Орлов Ю.Л., Мензоров А.Г., Афонников Д.А., Серов О.Л. 3С-методы в исследованиях пространственной организации генома // Вавиловский журнал генетики и селекции - 2012. - Т. 16. - Н. 4/2. - С. 872–878

Allen M.J., Bradbury E.M., Balhorn R. AFM analysis of DNA-protamine complexes bound to mica // Nucleic Acids Research - 1997. - V. 25. - N 11. - P. 2221–2226

Andrulis E.D., Neiman A.M., Zappulla D.C., Sternglanz R. Perinuclear localization of chromatin facilitates transcriptional silencing. // Nature - 1998. - V. 394. - N 6693. - P. 592–595

Anton T., Bultmann S., Leonhardt H., Markaki Y. Visualization of specific DNA sequences in living mouse embryonic stem cells with a programmable fluorescent CRISPR/Cas system. // Nucleus (Austin, Tex.) - 2014. - V. 5. - N 2. - P. 163–72

Ay F., Bailey T.L., Noble W.S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. // Genome research - 2014a. -P. 1–23

Ay F., Bunnik E.M., Varoquaux N., Bol S.M., Prudhomme J., Vert J.P., Noble W.S., Le Roch K.G. Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression // Genome Research - 2014b. - V. 24. - P. 974–988

Balhorn R., Cosman M., Thornton K., Krishnan V. V, Corzett M., Bench G., Kramer C., Lee J.I., Hud N. V, Allen M., Prieto M., Meyer-Ilse W., Brown J.T., Kirz J., Zhang X., Bradbury E.M., Maki G., Braun R.E., Breed W. Protamine mediated condensation of DNA in mammalian sperm // Male Gamete - 1999. -P. 55–70

Balhorn R., Gledhill B.L., Wyrobek A.J. Mouse sperm chromatin proteins: quantitative isolation and partial characterization. // Biochemistry - 1977. - V. 16. - N 18. - P. 4074–4080

Bartolomei M.S., Webber A.L., Brunkow M.E., Tilghman S.M. Epigenetic mechanisms underlying the imprinting of the mouse H19 gene // Genes and Development - 1993. - V. 7. - N 9. - P. 1663–1673

Battulin N.R., Fishman V.S., Khabarova A.A., Pomaznoy M.Y., Shnaider T.A., Afonnikov D.A., Serov O.L. Investigation of the spatial genome organization of mouse sperm and fibroblasts by the Hi-C method // Russian Journal of Genetics: Applied Research - 2014. - V. 4. - N 6. - P. 556–560

Boveri T. Die Blastomerenkerne von Ascaris megalocephala und die Theorie der Chromosomenindividualität // Archiv für Zellforschung - 1909. - V. 3. - P. 181–268

Brianna Caddle L., Grant J.L., Szatkiewicz J., van Hase J., Shirley B.-J., Bewersdorf J., Cremer C., Arneodo A., Khalil A., Mills K.D. Chromosome neighborhood composition determines translocation outcomes after exposure to high-dose radiation in primary cells. // Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology - 2007. - V. 15. - N 8. - P. 1061–1073

Brinkers S., Dietrich H.R.C., de Groote F.H., Young I.T., Rieger B. The persistence length of double stranded DNA determined using dark field tethered particle motion. // The Journal of chemical physics - 2009. - V. 130. - N 21. - P. 215105

Brykczynska U., Hisano M., Erkek S., Ramos L., Oakeley E.J., Roloff T.C., Beisel C., Schübeler D., Stadler M.B., Peters A.H.F.M. Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. // Nature structural & molecular biology - 2010. - V. 17. - N 6. - P. 679–687

Carone B.R., Hung J.H., Hainer S.J., Chou M. Te, Carone D.M., Weng Z., Fazzio T.G., Rando O.J. High-resolution mapping of chromatin packaging in mouse embryonic stem cells and sperm // Developmental Cell - 2014. - V. 30. - N 1. - P. 11–22

Carrell D.T., Emery B.R., Hammoud S. Altered protamine expression and diminished spermatogenesis: What is the link? // Human Reproduction Update - 2007. - V. 13. - N 3. - P. 313–327

Carter D., Chakalova L., Osborne C.S., Dai Y., Fraser P. Long-range chromatin regulatory interactions in vivo. // Nature genetics - 2002. - V. 32. - N 4. - P. 623–626

Chen B., Gilbert L.A., Cimini B.A., Schnitzbauer J., Zhang W., Li G.-W., Park J., Blackburn E.H., Weissman J.S., Qi L.S., Huang B. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System // Cell - 2013. - V. 155. - N 7. - P. 1479–1491

Chen J.L., Guo S.H., Gao F.H. Nuclear matrix in developing rat spermatogenic cells. // Molecular reproduction and development - 2001. - V. 59. - N 3. - P. 314–321

Cope N.F., Fraser P., Eskiw C.H. The yin and yang of chromatin spatial organization. // Genome biology - 2010. - V. 11. - N 3. - P. 204

Cremer C., Cremer T. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. // Nature reviews. Genetics - 2001a. - V. 2. - P. 292–301

Cremer M., von Hase J., Volm T., Brero A., Kreth G., Walter J., Fischer C., Solovei I., Cremer C., Cremer T. Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. // Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology - 2001b. - V. 9. - N 7. - P. 541–567

Cremer M., Kupper K., Wagler B., Wizelman L., von Hase J., Weiland Y., Kreja L., Diebold J., Speicher M.R., Cremer T. Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei. // The Journal of cell biology - 2003. - V. 162. - N 5. - P. 809–820

Cremer T., Cremer C., Baumann H., Luedtke E.K., Sperling K., Teuber V., Zorn C. Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by

premature chromosome condensation and laser-UV-microbeam experiments. // Human genetics - 1982a. - V. 60. - N 1. - P. 46–56

Cremer T., Cremer C., Schneider T., Baumann H., Hens L., Kirsch-Volders M. Analysis of chromosome positions in the interphase nucleus of Chinese hamster cells by laser-UV-microirradiation experiments. // Human genetics - 1982b. - V. 62. - N 3. - P. 201–209

Croft J.A., Bridger J.M., Boyle S., Perry P., Teague P., Bickmore W.A. Differences in the localization and morphology of chromosomes in the human nucleus. // The Journal of cell biology - 1999. - V. 145. - N 6. - P. 1119–1131

Dekker J., Rippe K., Dekker M., Kleckner N. Capturing chromosome conformation. // Science (New York, N.Y.) - 2002. - V. 295. - P. 1306–1311

Le Dily F., Baù D., Pohl A., Vicent G.P., Serra F., Soronellas D., Castellano G., Wright R.H.G., Ballare C., Filion G., Marti-Renom M. a, Beato M. Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. // Genes & development - 2014. - V. 28. - N 19. - P. 2151–62

Dixon J.R., Selvaraj S., Yue F., Kim A., Li Y., Shen Y., Hu M., Liu J.S., Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions // Nature - 2012. - V. 485. - P. 376–380

Duan Z., Andronescu M., Schutz K., McIlwain S., Kim Y.J., Lee C., Shendure J., Fields S., Blau C.A., Noble W.S. A three-dimensional model of the yeast genome. // Nature - 2010. - V. 465. - P. 363–367

Duggal G., Wang H., Kingsford C. Higher-order chromatin domains link eQTLs with the expression of far-away genes // Nucleic Acids Research - 2014. - V. 42. - N 1. - P. 87–96

Efron B. Bootstrap Methods: Another Look at the Jackknife // The Annals of Statistics - 1979. - V. 7. - N 1. - P. 1–26

Erkek S., Hisano M., Liang C.-Y., Gill M., Murr R., Dieker J., Schübeler D., van der Vlag J., Stadler M.B., Peters A.H.F.M. Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. // Nature structural & molecular biology - 2013. - V. 20. - N 7. - P. 868–875

Federico C., Scavo C., Cantarella C.D., Motta S., Saccone S., Bernardi G. Gene-rich and gene-poor chromosomal regions have different locations in the interphase nuclei of cold-blooded vertebrates. // Chromosoma - 2006. - V. 115. - N 2. - P. 123–128

Feng S., Cokus S.J., Schubert V., Zhai J., Pellegrini M., Jacobsen S.E. Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis // Molecular Cell - 2014. - V. 55. - N 5. - P. 694–707

Filippova D., Patro R., Duggal G., Kingsford C. Identification of alternative topological domains in chromatin. // Algorithms for molecular biology : AMB - 2014. - V. 9. - N 1. - P. 14

Fuentes-Mascorro G., Serrano H., Rosado a Sperm chromatin. // Archives of andrology - 2000. - V. 45. - N 3. - P. 215–225

Fujita N., Wade P.A. Use of bifunctional cross-linking reagents in mapping genomic distribution of chromatin remodeling complexes // Methods - 2004. - V. 33. - N 1. - P. 81–85

Fullwood M.J., Liu M.H., Pan Y.F., Liu J., Xu H., Mohamed Y. Bin, Orlov Y.L., Velkov S., Ho A., Mei P.H., Chew E.G.Y., Huang P.Y.H., Welboren W.-J., Han Y., Ooi H.S., Ariyaratne P.N., Vega V.B., Luo Y., Tan P.Y., Choy P.Y., Wansa K.D.S.A., Zhao B., Lim K.S., Leow S.C., Yow J.S., Joseph R., Li H., Desai K. V, Thomsen J.S., Lee Y.K., Karuturi R.K.M., Herve T., Bourque G., Stunnenberg H.G., Ruan X., Cacheux-Rataboul V., Sung W.-K., Liu E.T., Wei C.-L., Cheung E., Ruan Y. An oestrogen-

receptor-alpha-bound human chromatin interactome. // Nature - 2009. - V. 462. - P. 58–64

Gao F., Wei Z., Lu W., Wang K. Comparative analysis of 4C-Seq data generated from enzyme-based and sonication-based methods. // BMC genomics - 2013. - V. 14. - N 1. - P. 345

Gavrilov A. a, Gushchanskaya E.S., Strelkova O., Zhironkina O., Kireev I.I., Iarovaia O. V, Razin S. V Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. // Nucleic acids research - 2013. - V. 41. - N 6. - P. 3563–3575

Gilbert N., Boyle S., Fiegler H., Woodfine K., Carter N.P., Bickmore W.A. Chromatin architecture of the human genome: Gene-rich domains are enriched in open chromatin fibers // Cell - 2004. - V. 118. - N 5. - P. 555–566

Goetze S., Mateos-Langerak J., Gierman H.J., de Leeuw W., Giromus O., Indemans M.H.G., Koster J., Ondrej V., Versteeg R., van Driel R. The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. // Molecular and cellular biology - 2007. - V. 27. - N 12. - P. 4475–4487

Grasser F., Neusser M., Fiegler H., Thormeyer T., Cremer M., Carter N.P., Cremer T., Muller S. Replication-timing-correlated spatial chromatin arrangements in cancer and in primate interphase nuclei. // Journal of cell science - 2008. - V. 121. - N 11. - P. 1876–1886

Grob S., Schmid M.W., Grossniklaus U. Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila // Molecular Cell - 2014. - V. 55. - N 5. - P. 678–693

Habermann F.A., Cremer M., Walter J., Kreth G., von Hase J., Bauer K., Wienberg J., Cremer C., Cremer T., Solovei I. Arrangements of macro- and microchromosomes in chicken cells. // Chromosome research : an international journal on the molecular,

supramolecular and evolutionary aspects of chromosome biology - 2001. - V. 9. - N 7. - P. 569–584

Hagège H., Klous P., Braem C., Splinter E., Dekker J., Cathala G., de Laat W., Forné T. Quantitative analysis of chromosome conformation capture assays (3C-qPCR). // Nature protocols - 2007. - V. 2. - P. 1722–1733

Hahn M.A., Wu X., Li A.X., Hahn T., Pfeifer G.P. Relationship between gene body DNA methylation and intragenic H3K9ME3 and H3K36ME3 chromatin marks // PLoS ONE - 2011. - V. 6. - N 4. - P. e18844

Hammoud S.S., Nix D. a, Zhang H., Purwar J., Carrell D.T., Cairns B.R. Distinctive chromatin in human sperm packages genes for embryo development. // Nature - 2009. - V. 460. - N 7254. - P. 473–478

Hattori M. Finishing the euchromatic sequence of the human genome // Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme - 2005. - V. 50. - N 2. - P. 162–168

Heng H.H., Krawetz S.A., Lu W., Bremer S., Liu G., Ye C.J. Re-defining the chromatin loop domain. // Cytogenetics and cell genetics - 2001. - V. 93. - N 3-4. - P. 155–161

Heng H.H.Q., Goetze S., Ye C.J., Liu G., Stevens J.B., Bremer S.W., Wykes S.M., Bode J., Krawetz S.A. Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. // Journal of cell science - 2004. - V. 117. - N Pt 7. - P. 999–1008

Hepperger C., Mannes A., Merz J., Peters J., Dietzel S. Three-dimensional positioning of genes in mouse cell nuclei. // Chromosoma - 2008. - V. 117. - N 6. - P. 535–551

Horowitz R.A., Agard D.A., Sedat J.W., Woodcock C.L. The three-dimensional architecture of chromatin in situ: Electron tomography reveals fibers composed of a continuously variable zig-zag nucleosomal ribbon // Journal of Cell Biology - 1994. - V. 125. - N 1. - P. 1–10

Hou C., Li L., Qin Z.S., Corces V.G. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains // Molecular Cell - 2012. - V. 48. - P. 471–484

Hu M., Deng K., Qin Z., Dixon J., Selvaraj S., Fang J., Ren B., Liu J.S. Bayesian Inference of Spatial Organizations of Chromosomes // PLoS Computational Biology - 2013. - V. 9. - N 1. - P. e1002893

Imakaev M., Fudenberg G., McCord R.P., Naumova N., Goloborodko A., Lajoie B.R., Dekker J., Mirny L.A. Iterative correction of Hi-C data reveals hallmarks of chromosome organization // Nature Methods - 2012. - V. 9. - P. 999–1003

Jackson V. Formaldehyde cross-linking for studying nucleosomal dynamics. // Methods (San Diego, Calif.)x - 1999. - V. 17. - N 2. - P. 125–139

Johannes S., Holwerda B., Laat W. De CTCF : the protein , the binding partners , the binding sites and their chromatin loops // Philosophical Transactions of the Royal Society B: Biological Sciences - 2013. - V. 368. - N 1620. - P. 20120369

Kalhor R., Tjong H., Jayathilaka N., Alber F., Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. // Nature biotechnology - 2012. - V. 30. - N 1. - P. 90–98

Khalil A., Grant J.L., Caddle L.B., Atzema E., Mills K.D., Arneodo A. Chromosome territories have a highly nonspherical morphology and nonrandom positioning. // Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology - 2007. - V. 15. - N 7. - P. 899–916

Koehler D., Zakhartchenko V., Froenicke L., Stone G., Stanyon R., Wolf E., Cremer T., Brero A. Changes of higher order chromatin arrangements during major genome activation in bovine preimplantation embryos. // Experimental cell research - 2009. - V. 315. - N 12. - P. 2053–2063

Kosak S.T., Skok J.A., Medina K.L., Riblet R., Le Beau M.M., Fisher A.G., Singh H. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. // Science (New York, N.Y.) - 2002. - V. 296. - N 5565. - P. 158–162

Kupper K., Kolbl A., Biener D., Dittrich S., von Hase J., Thormeyer T., Fiegler H., Carter N.P., Speicher M.R., Cremer T., Cremer M. Radial chromatin positioning is shaped by local gene density, not by gene expression. // Chromosoma - 2007. - V. 116. - N 3. - P. 285–306

De Laat W., Grosveld F. Spatial organization of gene expression: The active chromatin hub // Chromosome Research - 2003. - V. 11. - N 5. - P. 447–459

Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2 // Nature Methods - 2012. - V. 9. - N 4. - P. 357–359

Lee J.D., Allen M.J., Balhorn R. Atomic force microscope analysis of chromatin volumes in human sperm with head-shape abnormalities. // Biology of reproduction - 1997. - V. 56. - N 1. - P. 42–49

Lieberman-Aiden E., van Berkum N.L., Williams L., Imakaev M., Ragoczy T., Telling A., Amit I., Lajoie B.R., Sabo P.J., Dorschner M.O., Sandstrom R., Bernstein B., Bender M.A., Groudine M., Gnirke A., Stamatoyannopoulos J., Mirny L.A., Lander E.S., Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. // Science (New York, N.Y.) - 2009. - V. 326. - P. 289–293

Lin Y.C., Benner C., Mansson R., Heinz S., Miyazaki K., Miyazaki M., Chandra V., Bossen C., Glass C.K., Murre C. Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. // Nature immunology - 2012. - V. 13. - N 12. - P. 1196–204

Lupiáñez D.G., Kraft K., Heinrich V., Krawitz P., Brancati F., Klopocki E., Horn D., Kayserili H., Opitz J.M., Laxova R., Santos-Simarro F., Gilbert-Dussardier B., Wittler

L., Borschiwer M., Haas S.A., Osterwalder M., Franke M., Timmermann B., Hecht J., Spielmann M., Visel A., Mundlos S. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions // Cell - 2015. -P. 1–14

Ma H., Naseri A., Reyes-Gutierrez P., Wolfe S. a., Zhang S., Pederson T. Multicolor CRISPR labeling of chromosomal loci in human cells // Proceedings of the National Academy of Sciences - 2015. - V. 112. - N 10. - P. 3002–3007

Manuelidis L. Individual interphase chromosome domains revealed by in situ hybridization. // Human genetics - 1985. - V. 71. - N 4. - P. 288–293

Mayer R., Brero A., von Hase J., Schroeder T., Cremer T., Dietzel S. Common themes and cell type specific variations of higher order chromatin arrangements in the mouse. // BMC cell biology - 2005. - V. 6. - P. 44

Mirny L. a The fractal globule as a model of chromatin architecture in the cell. // Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology - 2011. - V. 19. - N 1. - P. 37–51

Mizuguchi T., Fudenberg G., Mehta S., Belton J.-M., Taneja N., Folco H.D., FitzGerald P., Dekker J., Mirny L., Barrowman J., Grewal S.I.S. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe // Nature - 2014. - V. 516. - N 7531. - P. 432–435

Morey C., Da Silva N.R., Perry P., Bickmore W.A. Nuclear reorganisation and chromatin decondensation are conserved, but distinct, mechanisms linked to Hox gene activation. // Development (Cambridge, England) - 2007. - V. 134. - N 5. - P. 909–919

Mudrak O., Zalenskaya I., Zalensky A. Organization of Chromosomes During Spermatogenesis and in Mature Sperm // Springer Berlin Heidelberg - 2011. - P.261–277

Nagano T., Lubling Y., Stevens T.J., Schoenfelder S., Yaffe E., Dean W., Laue E.D., Tanay A., Fraser P. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. // Nature - 2013. - V. 502. - N 7469. - P. 59–64

Naumova N., Imakaev M., Fudenberg G., Zhan Y., Lajoie B.R., Mirny L. a, Dekker J. Organization of the mitotic chromosome. // Science (New York, N.Y.) - 2013. - V. 342. - N 6161. - P. 948–953

Neusser M., Schubel V., Koch A., Cremer T., Muller S. Evolutionarily conserved, cell type and species-specific higher order chromatin arrangements in interphase nuclei of primates. // Chromosoma - 2007. - V. 116. - N 3. - P. 307–320

Orlando V., Strutt H., Paro R. Analysis of Chromatin Structure byin VivoFormaldehyde Cross-Linking // Methods - 1997. - V. 11. - N 2. - P. 205–214

Osborne C.S., Chakalova L., Brown K.E., Carter D., Horton A., Debrand E., Goyenechea B., Mitchell J.A., Lopes S., Reik W., Fraser P. Active genes dynamically colocalize to shared sites of ongoing transcription. // Nature genetics - 2004. - V. 36. - N 10. - P. 1065–1071

Phillips-Cremins J.E., Sauria M.E.G., Sanyal A., Gerasimova T.I., Lajoie B.R., Bell J.S.K., Ong C.T., Hookway T.A., Guo C., Sun Y., Bland M.J., Wagstaff W., Dalton S., McDevitt T.C., Sen R., Dekker J., Taylor J., Corces V.G. Architectural protein subclasses shape 3D organization of genomes during lineage commitment // Cell - 2013. - V. 153. - P. 1281–1295

Pope B.D., Ryba T., Dileep V., Yue F., Wu W., Denas O., Vera D.L., Wang Y., Hansen R.S., Canfield T.K., Thurman R.E., Cheng Y., Gülsoy G., Dennis J.H., Snyder M.P., Stamatoyannopoulos J. a., Taylor J., Hardison R.C., Kahveci T., Ren B., Gilbert D.M. Topologically associating domains are stable units of replication-timing regulation // Nature - 2014. - V. 515. - N 7527. - P. 402–405

Rabl C. Über Zelltheilung // Morph. Jb - 1885. - V. 10. - P. 214–330

Rao S.S.P., Huntley M.H., Durand N.C., Stamenova E.K., Bochkov I.D., Robinson J.T., Sanborn A.L., Machol I., Omer A.D., Lander E.S., Aiden E.L. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping // Cell - 2014. - V. 159. - N 7. - P. 1665–1680

Reshef D.N., Reshef Y.A., Finucane H.K., Grossman S.R., McVean G., Turnbaugh P.J., Lander E.S., Mitzenmacher M., Sabeti P.C. Detecting Novel Associations in Large Data Sets // Science - 2011. - V. 334. - P. 1518–1524

Schader M., Schmid F. Two Rules of Thumb for the Approximation of the Binomial Distribution by the Normal Distribution // The American Statistician - 1989. - V. 43. - N 1. - P. 23–24

Schardin M., Cremer T., Hager H.D., Lang M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. // Human genetics - 1985. - V. 71. - N 4. - P. 281–287

Servant N., Lajoie B.R., Nora E.P., Giorgetti L., Chen C.J., Heard E., Dekker J., Barillot E. HiTC: Exploration of high-throughput “C” experiments // Bioinformatics - 2012. - V. 28. - P. 2843–2844

Sexton T., Yaffe E., Kenigsberg E., Bantignies F., Leblanc B., Hoichman M., Parrinello H., Tanay A., Cavalli G. Three-dimensional folding and functional organization principles of the Drosophila genome // Cell - 2012. - V. 148. - P. 458–472

Shaman J.A., Yamauchi Y., Ward W.S. The sperm nuclear matrix is required for paternal DNA replication // Journal of Cellular Biochemistry - 2007. - V. 102. - N 3. - P. 680–688

Shlens J. A Tutorial on Principal Component Analysis // Measurement - 2005. - V. 51. - P. 52

Simonis M., Kooren J., de Laat W. An evaluation of 3C-based methods to capture DNA interactions. // Nature methods - 2007. - V. 4. - N 11. - P. 895–901

Smith Z.D., Meissner A. DNA methylation: roles in mammalian development. // Nature reviews. Genetics - 2013. - V. 14. - N 3. - P. 204–20

Sneppen K., Dodd I.B. A Simple Histone Code Opens Many Paths to Epigenetics // PLoS Computational Biology - 2012. - V. 8. - N 8. - P. e1002643

Sofueva S., Yaffe E., Chan W.-C., Georgopoulou D., Vietri Rudan M., Mira-Bontenbal H., Pollard S.M., Schroth G.P., Tanay A., Hadjur S. Cohesin-mediated interactions organize chromosomal domain architecture. // The EMBO journal - 2013. - V. 32. - N 24. - P. 3119–3129

Solovei I., Kreysing M., Lanctôt C., Kösem S., Peichl L., Cremer T., Guck J., Joffe B. Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution // Cell - 2009. - V. 137. - N 2. - P. 356–368

Stewart M.D., Li J., Wong J. Relationship between histone H3 lysine 9 methylation, transcription repression, and heterochromatin protein 1 recruitment. // Molecular and cellular biology - 2005. - V. 25. - N 7. - P. 2525–2538

Symmons O., Uslu V.V., Tsujimura T., Ruf S., Nassari S., Schwarzer W., Ettwiller L., Spitz F. Functional and topological characteristics of mammalian regulatory domains // Genome Research - 2014. - V. 24. - P. 390–400

Tanabe H., Muller S., Neusser M., von Hase J., Calcagno E., Cremer M., Solovei I., Cremer C., Cremer T. Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. // Proceedings of the National Academy of Sciences of the United States of America - 2002. - V. 99. - N 7. - P. 4424–4429

Tang J.B., Chen Y.H. Identification of a tyrosine-phosphorylated CCCTC-binding nuclear factor in capacitated mouse spermatozoa // Proteomics - 2006. - V. 6. - N 17. - P. 4800–4807

Tark-Dame M., Jerabek H., Manders E.M.M., Heermann D.W., van Driel R. Depletion of the chromatin looping proteins CTCF and cohesin causes chromatin compaction:

insight into chromatin folding by polymer modelling. // PLoS computational biology - 2014. - V. 10. - N 10. - P. e1003877

Therizols P., Illingworth R.S., Courilleau C., Boyle S., Wood A.J., Bickmore W. a Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. // Science (New York, N.Y.) - 2014. - V. 346. - N 6214. - P. 1238–42

Tolhuis B., Palstra R.J., Splinter E., Grosveld F., De Laat W. Looping and interaction between hypersensitive sites in the active b-globin locus // Molecular Cell - 2002. - V. 10. - N 6. - P. 1453–1465

Trimarchi T., Bilal E., Ntziachristos P., Fabbri G., Dalla-Favera R., Tsirigos A., Aifantis I. Genome-wide mapping and characterization of notch-regulated long noncoding RNAs in acute leukemia // Cell - 2014. - V. 158. - P. 593–606

Vietri Rudan M., Barrington C., Henderson S., Ernst C., Odom D.T., Tanay A., Hadjur S. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture // Cell Reports - 2015. - V. 10. - N 8. - P. 1297–1309

De Wit E., de Laat W. A decade of 3C technologies: Insights into nuclear organization // Genes and Development - 2012. - V. 26. - N 1. - P. 11–24

Zalenskaya I.A., Zalensky A.O. Non-random positioning of chromosomes in human sperm nuclei // Chromosome Research - 2004. - V. 12. - N 2. - P. 163–173

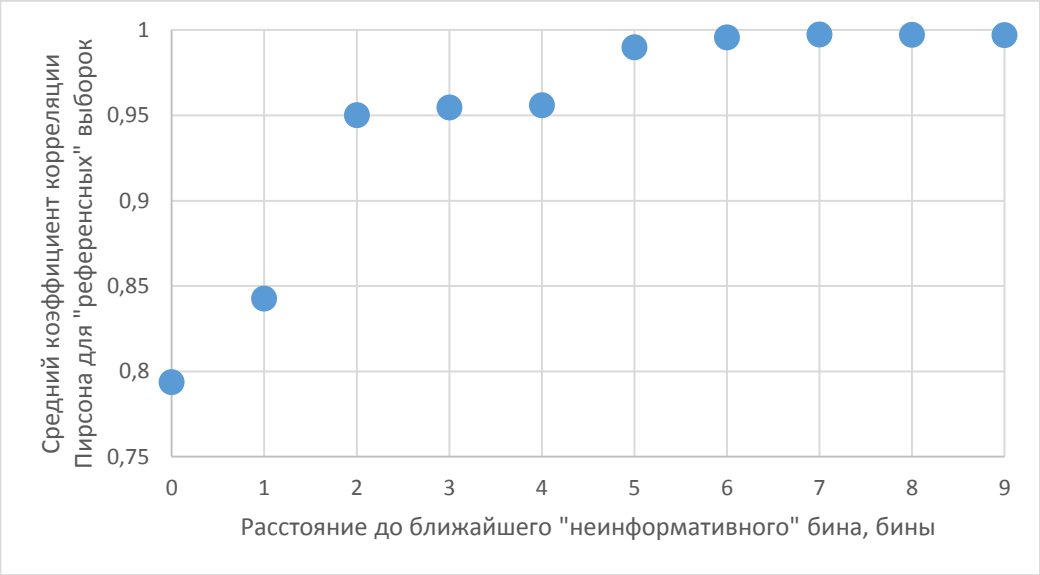
Zeitz M.J., Mukherjee L., Bhattacharya S., Xu J., Berezney R. A probabilistic model for the arrangement of a subset of human chromosome territories in WI38 human fibroblasts. // Journal of cellular physiology - 2009. - V. 221. - N 1. - P. 120–129

Zorn C., Cremer C., Cremer T., Zimmer J. Unscheduled DNA synthesis after partial UV irradiation of the cell nucleus. Distribution in interphase and metaphase. // Experimental cell research - 1979. - V. 124. - N 1. - P. 111–119

Zorn C., Cremer T., Cremer C., Zimmer J. Laser UV microirradiation of interphase nuclei and post-treatment with caffeine. A new approach to establish the arrangement of interphase chromosomes. // Human genetics - 1976. - V. 35. - N 1. - P. 83–89

Zuin J., Dixon J.R., van der Reijden M.I.J. a, Ye Z., Kolovos P., Brouwer R.W.W., van de Corput M.P.C., van de Werken H.J.G., Knoch T. a, van IJcken W.F.J., Grosveld F.G., Ren B., Wendt K.S. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. // Proceedings of the National Academy of Sciences of the United States of America - 2014. - V. 111. - N 3. - P. 996–1001

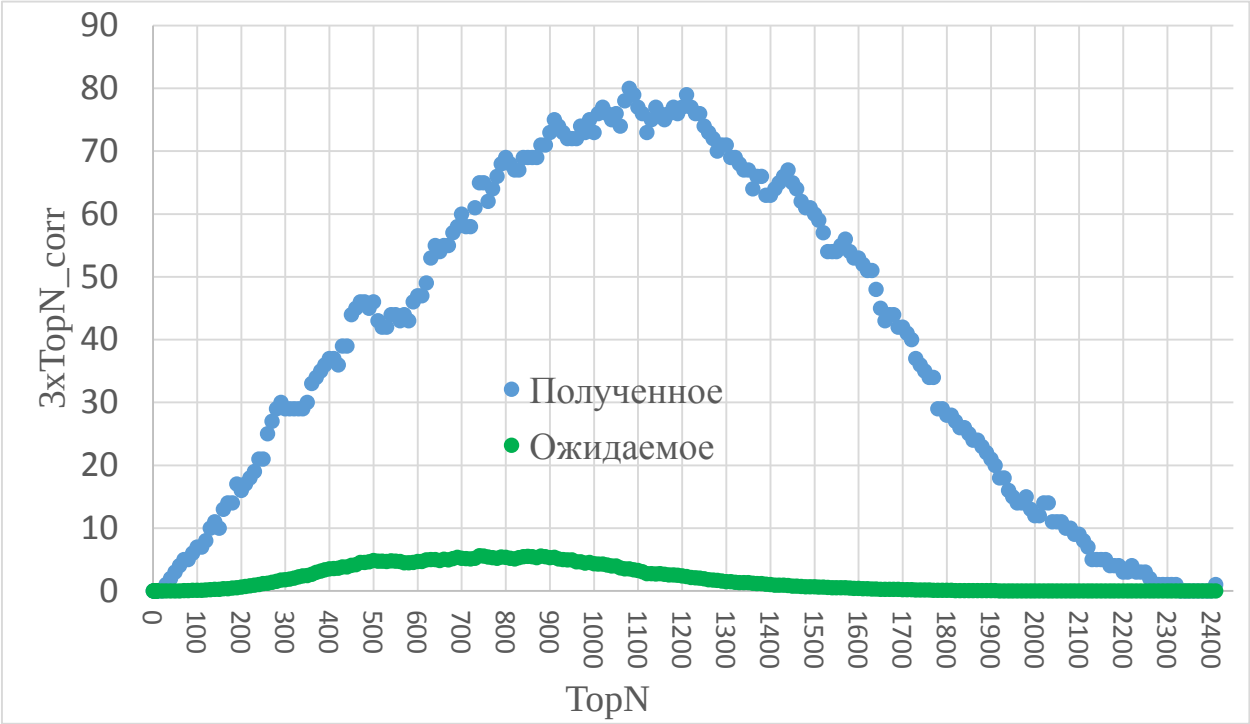
Приложение 1



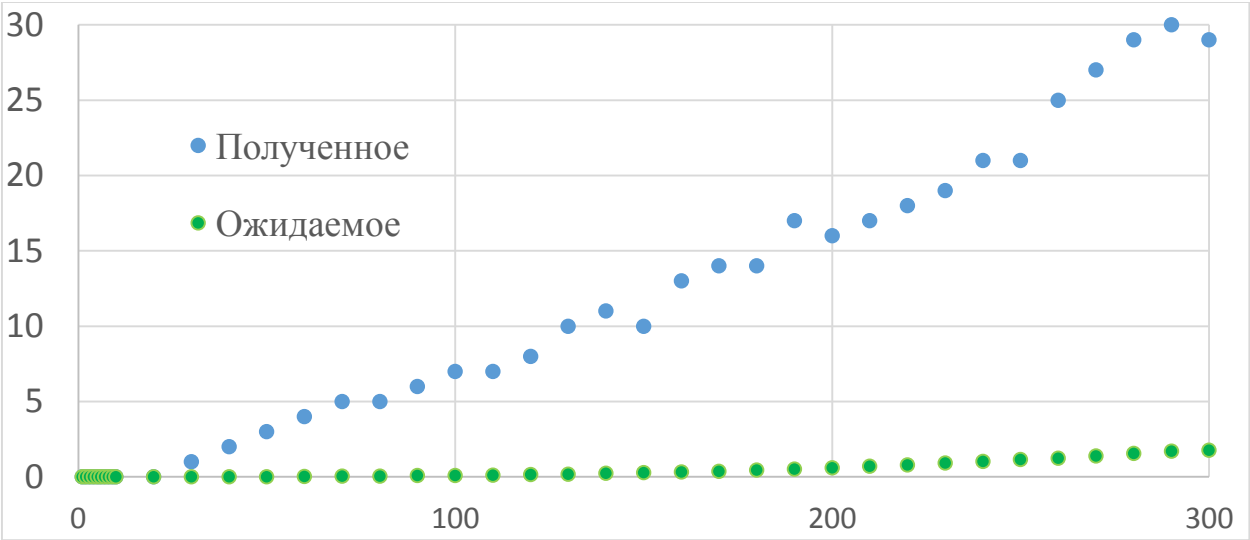
Зависимость коэффициента корреляции Спирмена для контактов локусов «референсных» выборок вблизи «неинформативных» бинов. По оси Y отложен коэффициент корреляции, по оси X – расстояние от данного локуса до ближайшего «неинформативного» (расстояние указано в бинах). Данные усреднены по всему геному.

Приложение 2

А

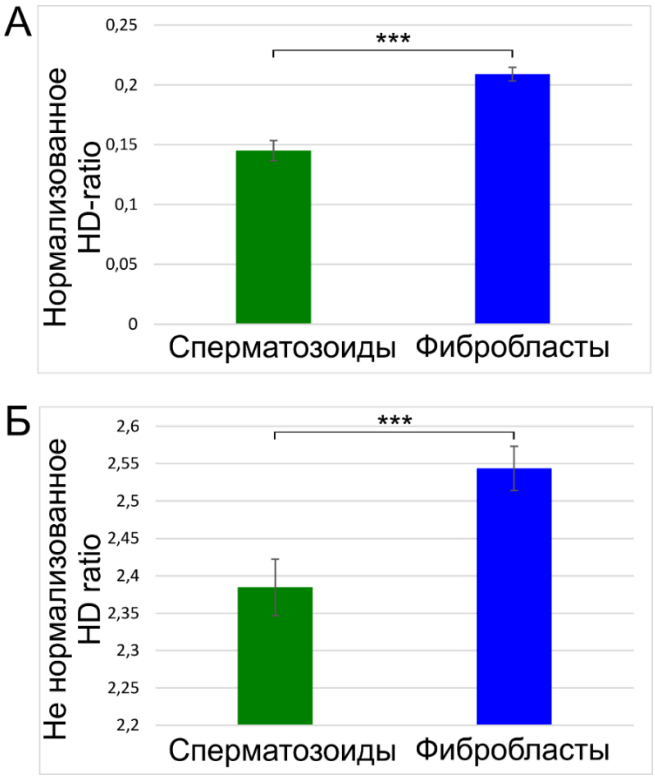


Б



Зависимость числа бинов в списке 3xTopN_corr от выбора числа TopN. Синим цветом показана зависимость, полученная на основе экспериментальных данных, зеленым – ожидаемое число бинов для каждого числа TopN, рассчитанное как указано в главе «Материалы и методы». А. Весь диапазон значений TopN. Б. Диапазон значений TopN от 0 до 300

Приложение 3



TAD-домены фибробластов более «вытянутые», чем домены сперматозоидов. На рисунке представлены значения HD-ratio для фибробластов и сперматозоидов. На гистограмме А значения нормализованы на длину доменов (выраженную в бинах), на рисунке Б показаны ненормализованные данные