- Наименование работы – Исследование пространственной организации геномов фибробластов и сперматозоидов методом HiC
- Состав коллектива исполнителей, контактное лицо (ФИО полностью и адреса электронной почты всех членов коллектива) – Фишман Вениамин Семенович, minja-f@ya.ru
- Научное содержание работы:

1. Постановка задачи.

Несмотря на то, что основой геномов большинства организмов, живущих на земле, является молекула двухцепочечной ДНК, способы её хранения значительно различаются. В ходе эволюции, особенно при переходе от бактерий к многоклеточным эукариотам, произошло небольшое увеличение числа генов, и огромное – на 2-3 порядка – увеличение размера генома (от миллионов до биллионов нуклеотидов). При этом пространственная организация генома в ядре, помимо функции компактизации, также приобрела регуляторную функцию – так, например, 3D-фолдинг хромосом компартментализует геном и может обеспечить сближение регуляторных элементов. Поэтому изучение 3D-структуры содержимого ядра позволяет не только описать пространственную организацию генома, но и выяснить механизмы регуляции таких фундаментальных процессов, как транскрипция, репликация и т.д.

Попытки изучить пространственную организацию ядра проводятся учеными более ста лет. За это время наши знания в этой области значительно расширились. Во многом, этот прогресс связан с разработкой новых методологических подходов. Поскольку классические методы изучения пространственной организации геномов основаны на микроскопии, их фундаментальным ограничением является относительно низкая разрешающая способность. Для анализа укладки генома на более детальном уровне, в 2009 году был разработан метод Hi-C (Lieberman-Aiden et al., 2009).,позволяющий проводить подсчет числа пространственно-сближенных участков ДНК при помощи массового параллельного секвенирования. Этот метод позволил принципиально по-новому взглянуть на архитектуру ядра, выявив неизвестный ранее уровень организации генома – так называемые Hi-C домены, протяженные участки хромосомы, пространственно-сближенные и активно взаимодействующие друг с другом. Корреляция границ доменов с местами посадки регуляторных факторов, хромосомными территориями, зонами ранней и поздней репликации и другими функциональными единицами генома, а также их эволюционная консервативность показывают значимость этих структурных единиц. Тем не менее, не отвеченными

остаются фундаментальные вопросы о том, как формируются и поддерживаются Hi-C домены, и какую роль они играют в процессе функционирования генома.

Уникальным объектом для изучения пространственной организации генома являются сперматозоиды. Организация генетического материала в этих клетках сильно отличается от организации генома соматических клеток. Для уменьшения размера сперматозоида при его созревании происходит уплотнение ядра за счет специального механизма конденсации хроматина, при котором удаляются гистоны, а ДНК связывается с особыми белками - протаминами. Стоит подчеркнуть, что 3D-структура генома половых клеток (и в частности сперматозоидов) никогда не исследовалась молекулярными методами, поэтому неизвестно, сохраняются ли в геноме сперматозоида Hi-C домены. В связи с этим изучение взаимодействующих районов ДНК в ядре сперматозоида является важной фундаментальной научной задачей.

Конкретная фундаментальная задача в рамках проблемы, на решение которой направлен проект

В рамках проекта поставлена следующая фундаментальная задача:

1. Сравнить особенности пространственной организации геномов сперматозоидов и соматических клеток мыши.

2. Современное состояние проблемы.

После того, как стало известно, что молекулы ДНК, находящиеся в ядре, осуществляют роль хранения и передачи генетической информации, изучение механизмов её упаковки привлекло к себе внимание ученых и стало важным направлением молекулярной и клеточной биологии. Учитывая, что длина молекулы ДНК, кодирующей человеческий геном, составила бы в неупакованном состоянии около двух метров (что примерно в 200 000 раз больше, чем диаметр ядра), упаковка такой молекулы представляется крайней сложной задачей. Интересное само по себе, изучение принципов компактизации ДНК становится ещё более актуальным, если учесть, что упаковка ДНК осуществляет функцию регуляции различных процессов в геноме. Несмотря на то, что мы все ещё далеки от полного понимания того, как реализуется эта функция, современные технологии значительно расширили наши знания в этой области, позволив осуществить систематический и детальный анализ организации ядра.

Традиционно, изучение 3D-организации ядра проводилось методами микроскопии. Однако, около десяти лет назад, Деккером с соавторами была разработана принципиально новая технология, названная 3C (chromosome conformation capture) -

биохимическая методика, позволяющая подсчитать частоту контактов между двумя выбранными участками генома (Dekker *et al.*, 2002). В отличие от методик, основанных на микроскопии, метод 3C и другие методы, разработанные на его основе, не фиксируют конкретное событие взаимодействия в пределах одной клетки, а измеряют вероятность взаимодействия двух районов генома в большой (10^5–10^6) популяции клеток. Метод основан на принципе лигирования сближенных в пространстве молекул ДНК и может быть разбит на 4 этапа. Первый этап – фиксация клеток формальдегидом для сохранения нативной трехмерной структуры ядра. Формальдегид фиксирует белок-белковые, белок-ДНК и белок-РНК-взаимодействия за счет образования ковалентных связей между первичной аминогруппой белка и нуклеиновой кислотой (Orlando *et al.*, 1997; Jackson, 1999; Dekker *et al.*, 2002; Fujita, Wade, 2004).

После того как взаимодействующие молекулы ДНК оказываются сшитыми с белками, опосредующими это взаимодействие, ДНК фрагментируется с помощью рестриктаз. После этого проводят лигирование в условиях сильного разбавления. В таких условиях лигируются только концы молекул ДНК, сближенных в пространстве. Далее химерные молекулы ДНК выделяются и очищаются. Таким образом, создается библиотека попарно взаимодействующих молекул ДНК. Причем относительное обогащение в библиотеке специфических районов генома, лигированных друг с другом, отражает вероятность взаимодействия этих районов в трехмерном пространстве ядра в популяции клеток. Стоит подчеркнуть, что типичная 3C библиотека содержит огромное (до 10^{11}) количество уникальных пар взаимодействующих районов, и выбор дальнейшего метода анализа библиотеки зависит от задач исследования (Belton *et al.*, 2012). Особенно перспективным на сегодняшний день является метод Hi-C, который позволяет определять пространственную структуру хроматина в масштабе всего генома с очень высоким разрешением в большом количестве клеток (Lieberman-Aiden *et al.*, 2009). Метод Hi-C представляет собой объединение технологии 3C и технологий массового параллельного секвенирования. Полное секвенирование 3C-библиотеки теоретически позволяет установить все хроматиновые контакты, существующие в геноме. На практике количество установленных контактов сильно зависит от глубины секвенирования библиотеки (Shaw, 2010). Основной технической особенностью создания Hi-C библиотеки является этап обогащения библиотеки продуктами межмолекулярного лигирования. Такое обогащение достигается за счет заполнения липких концов, возникших при обработке рестриктазой фиксированного хроматина, биотинилированными нуклеотидами. На следующем этапе проводится лигирование по тупым концам. Таким образом, продукты межмолекулярного лигирования, собственно

молекулы, несущие информацию о контактах ДНК, оказываются меченными биотинилированными нуклеотидами. Использование на следующем этапе магнитных частиц, покрытых стрептавидином, позволяет сконцентрировать продукты лигирования и подготовить их для массового параллельного секвенирования (Lieberman-Aiden *et al.*, 2009).

При помощи применения Hi-C было независимым методом подтверждено наличие в ядре хромосомных территорий, существование которых ранее было показано методом FISH (Lieberman-Aiden et al., 2009). Предсказанные на основе данных Hi-C частоты межхромосомных контактов хорошо согласовывались с данными полученными на основе 3D FISH анализа, что еще раз подтвердило адекватность информации о пространственной организации генома полученных не в прямых экспериментах по микроскопированию, а с помощью молекулярно-биологических подходов (Kalhor et al., 2011).

Метод Hi-C позволяет реконструировать карту пространственных взаимодействий ДНК в ядре клетки, однако разрешение подобной карты сильно зависит от глубины секвенирования ДНК библиотеки. В первых работах Либермана с соавт. и Калхор с соавт. глубина секвенирования была относительно небольшой, и карта взаимодействующих районов имела разрешение порядка 1 миллиона пар нуклеотидов (м.п.н.). В проведенном недавно исследовании пространственной структуры генома эмбриональных стволовых клеток, фибробластов и нейронов коры головного мозга у мыши и человека методом Hi-C удалось достичь разрешения карты порядка 100 т.п.н. (Dixon et al., 2012). Десятикратное увеличение разрешения метода позволило авторам предложить модель фундаментальной организации генома соматических клеток – модель Hi-C доменов. Авторы показали, что «элементарной частицей» генома является Hi-C домен – протяженный участок хромосомы порядка 0,8-1 м.п.н. большая часть внутрихромосомных контактов которого приходится на самого себя. Таким образом, Hi-C домен представляет собой плотно упакованный автономный участок хромосомы. По длине хромосомы такие домены разделены районами «границ», для которых характерно малое количество внутрихромосомных контактов. Также было показано, что в районах границ часто локализуются сайты посадки инсуляторных белков, таких как CTCF, активно транскрибирующиеся гены домашнего хозяйства и SINE элементы (Dixon et al., 2012). Важным открытием стало то, что организация с помощью Hi-C доменов является эволюционно консервативной, то есть гомологичные последовательности генома человека и мыши организованы в топологические домены одинаково. Кроме того, структура Hi-C доменов устойчиво сохраняется в таких разных

типах клеток, таких как эмбриональные стволовые клетки, фибробласты и клетки коры головного мозга.

Организация хроматина в сперматозоидах.

В отличие от всех других клеточных типов большинство локусов ДНК в ядре сперматозоидов связана не с гистонами, а с другими белками – протаминами. Протамины представляют собой, относительно небольшие сильно основные белки. Специфическую упаковку генома сперматозоиды приобретают по мере созревания. Реорганизация и компактизация хроматина, по-видимому, происходит сходным образом у всех млекопитающих. По мере созревания сперматозоида хроматин переходит из «открытого» активного состояния к очень компактному, электронно плотному полностью неактивному состоянию. 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. 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 ценре (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 и общей конденсации хроматина. FISH с использованием хромосомоспецифических зондов показала наличие хромомсомных территорий в ядре зрелого сперматозоида человека (Zalenskaya et al., 2004). Белковый состав ядерного матрикса меняется по мере диффренцировки 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., 2001, 2004). 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 переукладке 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. Поэтому не известно, сохраняется ли типичная для соматических и стволовых клеток организация генома Hi-C доменами (Dixon et al., 2012) в сперматозоидах.

3. Подробное описание работы, включая используемые алгоритмы.

Пространственная организация геномов сперматозоидов и соматических клеток (фибробластов) мыши была изучена при помощи метода Hi-C, являющегося наиболее продвинутой молекулярно-биологической технологией исследования 3D-структуры генома. При помощи массового параллельного секвенирования, были получены базы данных, содержащие меж- и внутрихромосомные контакты для фибробластов и сперматозоидов. Эти данные были обработаны при помощи биоинформационных методов, в частности, будут использованы разработанные в 2012 году алгоритмы iteractive correction (Imakaev et al, 2012) (для нормализации марицы контактов) и domain search (Dixon et al., 2012) (для поиска доменов). Кроме того, при анализе исходных данных был использованы разработанный Мирным и коллегами набор фильтров, позволяющий исключить из рассмотрения возникающие во время подготовке Hi-C библиотеки артефакты. Следует отметить, что описанная методика обработки данных является наиболее продвинутой и позволяет значительно улучшить качество выходных данных.

4. Полученные результаты.

Секвенирование библиотек проводили на платформе Illumina GAII по методике парного секвенированиия (paired-end sequencing), с длиной прочтения 50 п.н. В результате было получено 119,35 млн. ридов для Hi-C библиотеки сперматозоидов и 153,37 млн. ридов для Hi-C библиотеки эмбриональных фибробластов мыши (рис 2). Для того чтобы оценить качество полученных данных, все дальнейшие манипуляции с данными секвенирования выполнялись параллельно с данными по секвенированию Hi-C библиотеки эмбриональных стволовых клеток (ЭСК), опубликованными ранее

(Dixon *et al.*, 2012). Далее проводили картирование ридов на геном мыши. Левый и правый концы каждого рида (5’ и 3’ соответственно) картировали отдельно по методике, описанной в работе Имакаева с соавторами (Imakaev *et al*., 2012). Количество достоверно картированных ридов для каждой из библиотек представлено на рис. 2. Далее проводили фильтрование картированных ридов от артефактных последовательностей, которые могли возникнуть на этапе приготовления ДНК библиотек, и которые не несут информации о пространственной конфигурации молекул ДНК. В результате, из дальнейшего анализа были исключены риды, картированные только с одной из сторон, риды, возникшие в результате замыкания в кольцо отдельного фрагмента ДНК, и риды, картированные далеко от сайта узнавания HindIII в геноме. Фильтрование проводили по методике, описанной в работе Имакаева с соавторами (Imakaev et al., 2012). Количество ридов, картированных с двух сторон и прошедших фильтрование, представлено на рис. 2. Именно данные, прошедшие фильтрование, несут информацию о пространственной организации исследуемого генома. Можно отметить, что доля качественных данных в разных библиотеках неодинакова. Так, для Hi-C библиотеки фибробластов доля ридов, картированных с двух сторон и прошедших фильтрование, составляет ~29%, для Hi-C библиотеки сперматозоидов ~9%, а для Hi-C библиотеки ЭСК ~32%. По-видимому, более плотная укладка генома сперматозоидов приводит к возникновению большего количества артефактных последовательностей в процессе приготовления Hi-C библиотеки, что уменьшает долю информативных ридов в исходных данных. Однако можно сделать вывод, что плотная укладка генома сперматозоида, тем не менее, не препятствует получению качественных данных с помощью метода Hi-C.

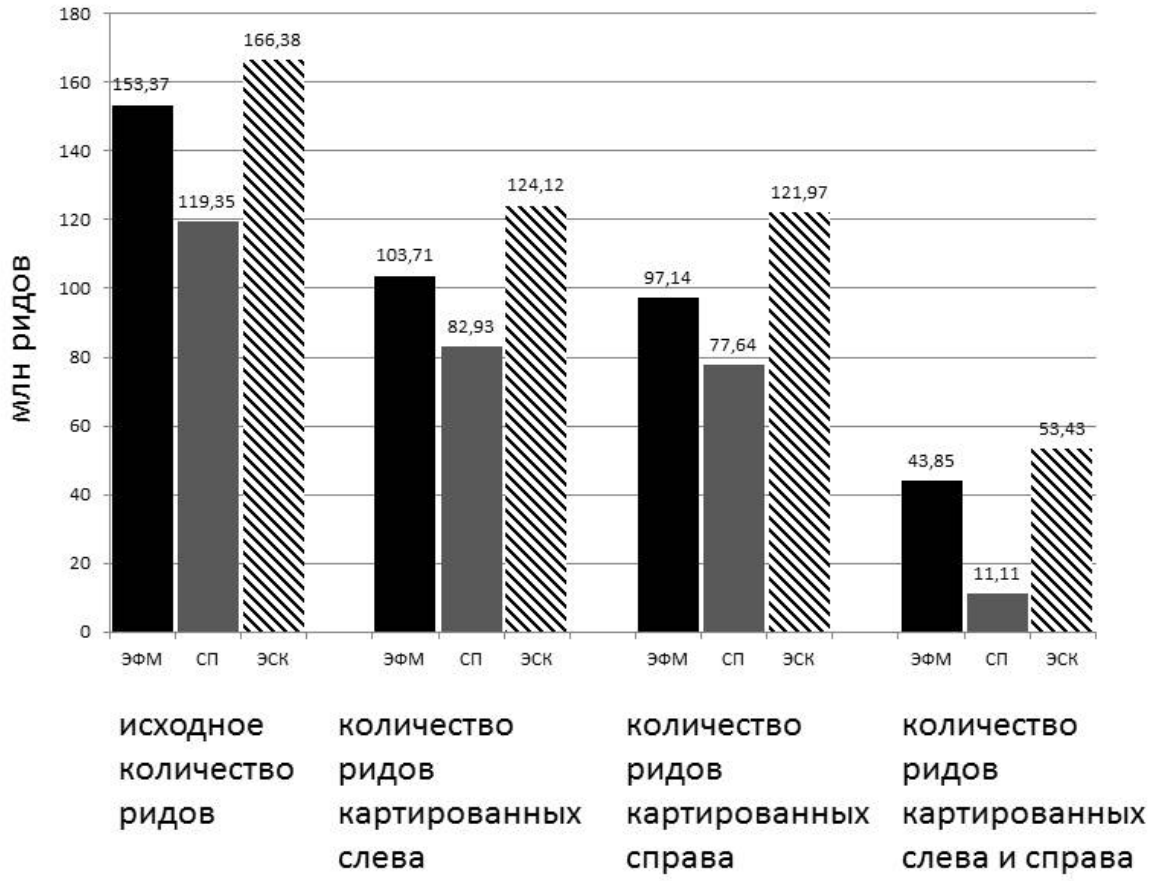

Рис. 2. Изменение количества ридов библиотек Hi-C по мере прохождения этапов обработки данных. ЭФМ – данные для Hi-C библиотеки фибробластов, СП – сперматозоидов, ЭСК - эмбриональных стволовых клеток.

Для того чтобы оценить, насколько полно полученные данные отражают реальную картину пространственной конфигурации генома, важно понять, насколько полно полученные данные покрывают геном. Всего в геноме мыши около 840000 рестрикционных фрагментов HindIII. Мы оценили количество рестрикционных фрагментов HindIII, для которых была получена информация об их пространственных взаимодействиях. Для Hi-C библиотеки фибробластов таких фрагментов HindIII составило около 760000, для Hi-C библиотеки сперматозоидов – 730000, для контрольных данных по ЭСК – 770000. Поскольку не все фрагменты можно картировать, например, фрагменты в районе повторов ДНК не поддаются однозначному картированию, а некоторые фрагменты плохо секвенируются из–за особенностей первичной структуры ДНК, то естественно, что в полученных данных мы не можем увидеть все фрагменты HindIII. Однако тот факт, что в полученных данных и

для фибробластов, и для сперматозоидов имеется информация о пространственных взаимодействиях значительной части фрагментов HindIII, позволяет заключить, что полученные данные достаточно полно покрывают весь геном.

Еще одна важная характеристика данных Hi-C – это количество взаимодействий между различными фрагментами генома. Если предположить, что каждый рестрикционный фрагмент может взаимодействовать с любым другим рестрикционным фрагментом, то получится более 10^11 различных вариантов таких взаимодействий. Мы оценили количество различных взаимодействий в полученных данных. Для Hi-C библиотеки фибробластов число уникальных взаимодействий составило около 36,3 млн., для Hi-C библиотеки сперматозоидов – 9,12 млн, для контрольных данных по ЭСК – 49,54 млн. Видно, что реальное число взаимодействий, наблюдаемое в эксперименте, значительно ниже теоретически возможного, поскольку, вероятно, не все они существуют in vivo. Кроме того, для того, чтобы наблюдать в данных 10^11 возможных комбинаций взаимодействий, необходимо секвенировать Hi-C библиотеку с глубиной не мене чем 10^11 ридов, что невозможно на современном уровне развития технологии секвенирования. Однако можно сделать вывод, что количество уникальных взаимодействий фрагментов генома в полученных данных для фибробластов и сперматозоидов по порядку величины не значительно отличается от количества уникальных взаимодействий фрагментов генома для опубликованных данных по ЭСК (Dixon *et al.*, 2012). А значит, полученных данных достаточно для исследования общих принципов пространственной организации генома половых и соматических клеток.

Таким образом, мы впервые применили метод Hi-C для получения ДНК библиотеки взаимодействующих районов генома сперматозоида. Было показано, что применение методики приготовления ДНК библиотеки взаимодействующих районов генома методом Hi-C позволяет получать качественные данные для исследования пространственной организации генома сперматозоидов.

5. Иллюстрации, визуализация результатов.

- Эффект от использования кластера в достижении целей работы.

Обработка данных пространственной структуры генома в сперматозоидах на вычислительной машине, доступной в лаборатории, заняло 5 дней. На вычислительном кластере НГУ – менее одного дня.

- Перечень публикаций, содержащих результаты работы.

1. Исследование пространственной организации генома сперматозоидов и фибробластов мыши методом Hi-C

Н.Р. Баттулин[1,2], В.С. Фишман[1,2], А.А. Хабарова[1], М.Ю. Помазной[1], Т.А. Шнайдер[1,2], Д.А. Афонников[1,2], О.Л. Серов[1,2]

2. Battulin N, **Fishman V.S.**, Mazur A.M., Pomaznoy M., Khabarova A.A., Afonnikov D.A., Prokhortchouk E.B., Serov O.L. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. // Genome Biology, 2015, V 16, № 77

- Функционирование кластера крайне удобно. До начала работ на кластере НГУ, вычисления проводились на кластере ВЦ. Стоит отметить, что на кластере НГУ загруженность вычислительных узлов была ниже (т.е. время ожидание в очереди – меньше), размер предоставляемого дискового пространства (что критически важно для экспериментов, проводимых нами, поскольку одна база данных может занимать 50-100 ГБ) – много выше. Особенно приятным оказалось наличие вычислительных узлов, оснащенных большим количеством оперативной памяти (500-1000 ГБ), поскольку это было использовано на определённых этапах анализа данных. Также следует отметить высокую скорость, с которой системный администратор контактировал с пользователем. В частности, в ходе работы были оперативно установлены на сервер определенные библиотеки языка python, не входившие в базовый комплект. Кроме того, оперативно осуществлялась поддержка пользователей, не обладающих большим опытом работ с Linux-системами, по общим вопросам.