Разработка метода для одновременной детекции хромосомных перестроек и точковых мутаций в геноме человека

Краткая аннотация

Проект направлен на создание ногвого метода для поиска генетических вариантов разного масштаба в геноме человека. Мы разработали и апробировали биохимический протокол, включающий получение 3С-библиотек из клеток периферической крови пациентов, обогащение созданных библиотек экзомными последовательностями и высокопроизводительное секвенирование. Проведенный нами анализ данных секвенирования позволил предложить ряд технических модификаций протокола, позволяющих увеличить его эффективность. Кроме этого, нами были разработаны и протестированы биоинформационные инструменты для поиска хромосомных перестроек на основе данных секвенирования обогащенных 3С-библиотек и для предсказания трехмерной укладки хроматина в перестроенных районах. Первая группа методов, направленная на поиск перестроек, ещё нуждается в существенной доработке, в то время как алгоритм для предсказания трехмерной организации хроматина представляет собой законченный продукт, который можно адаптировать к использованию в клинической практике.

Состав коллектива:

Полина Станиславовна Белокопытова (аспирант НГУ)

Можейко Евгений Александрович (аспирант ИЦиГ СО РАН)

Валеев Эмиль Салаватович (студент МедФ НГУ)

Фишман Вениамин Семенович (внс ИЦиГ СО РАН, старший преподаватель ФЕН НГУ), почта minja-f@ya.ru

Работа выполняется в рамках гранта РФФИ 18-29-13021 (2018-2021)

Постановка задачи

Исследование путей реализации наследственной информации в признаках и свойствах организмов является одной из центральных задач генетики со времен её становления как самостоятельной науки. С развитием технологий высокопроизводительного секвенирования мы получили беспрецедентные возможности для сопоставления генотипа и фенотипа человека. Тем не менее очевидно, что даже обладая огромным массивом геномных данных мы все еще далеки от полного объяснения механизмов формирования нормальных и патологических фенотипов.

Можно выделить две фундаментальные проблемы, которые стоят на пути от расшифровки генома человека к пониманию фенотипических последствий отдельных вариаций. Во-первых, для эукариот характерна поразительно сложная система регуляции активности генов, которая реализуется на многих уровнях, начиная со специфического взаимодействия функциональных элементов генома в пространстве ядра. Поэтому, фенотип организма определяется не только вариациями кодирующих последовательностей, но и огромным количеством регуляторных элементов, полиморфизмы которых мы пока ещё не в полной мере умеем интерпретировать.

Во‑вторых, существующие методы секвенирования индивидуальных геномов недостаточно адаптированы для детекции структурных вариаций, в особенности сбалансированных хромосомных перестроек. Микрочиповые или полноэкзомные методы могут идентифицировать только лишь крупные несбалансированные делеции и дупликации или полиморфизмы в кодирующих регионах, а полногеномное секвенирование всё ещё остается слишком дорогим методом для рутинного

применения в клинической практике. Поэтому доступной информации о клинически значимых сбалансированных хромосомных перестройках субмикроскопического масштаба намного меньше, чем, например, об однонуклеотидных полиморфизмах в экзонах. В то же время, для большинства хромосомных синдромов характерны сложные плейотропные эффекты, часть из которых реализуется за счет нарушения плохо исследованных механизмов регуляции генов. Для их понимания необходимо создать и проанализировать большой массив данных, включающих подробную информацию о структурных вариациях генома.

Более того, большинство фенотипических признаков обусловлены комплексом разных вариаций, как структурных, так и точечных, реализующихся в одном геноме. Для поиска ассоциаций в столь сложных, комплексных системах, необходимо создание методов, которые позволили бы одновременно выявлять разные типы мутаций, и были бы достаточно доступны для рутинного применения. Разработка таких методов и подходов для интерпретации полученных данных позволит лучше понять принципы реализации генетической информации и механизмы формирования генетических патологий человека.

Целью Проекта является разработка нового подхода к молекулярной диагностике хромосомных болезней на основе комбинаций технологий захвата конформации хромосом, целевого обогащения и массового параллельного секвенирования.

Задачами Проекта являются:

1. Разработка метода для создания 3C-библиотек с обогащением экзонными и промоторными последовательностями (EPiC, **E**xome and **P**romomter 3C-**C**apture библиотек).

2. Создание биоинформационных подходов для поиска и интерпретации структурных и точковых мутаций на основе данных секвенирования EPiC библиотек.

3. Оценка эффективности метода EPiC в качестве инструмента генетической диагностики в сравнении с полноэкзомным и полногеномным секвенированием и методом CGH-гибридизации на ДНК-микрочипах (array-CGH, CMA).

4. Применение метода EPiC для установления молекулярных механизмов развития патологий на выборке пациентов с врожденными наследственными заболеваниями.

Современное состояние проблемы (на момент начала работы)

Ежегодно в мире рождается несколько миллионов человек с генетически обусловленными врожденными пороками развития. Существует более 7 000 наследственных заболеваний (Mendelian Inheritance in Men database, MIM) и, хотя большинство таких патологий является редкими, в совокупности их частота в популяции превышает 7% (Baird, et al., 1988).

Из-за большого числа генетических синдромов, нередко с недостаточно полным описанием всего спектра фенотипических проявлений, а также из-за большой генетической гетерогенности многих наследственных патологий (например, нарушений интеллектуального развития), диагностика врожденного заболевания может представлять сложную задачу даже для опытного клинициста. Постановка первичного диагноза врачом-генетиком основывается на оценке фенотипа пациента и использовании методов цитогенетического кариотипирования, CGH-кариотипирования, биохимических анализов и секвенирования индивидуальных генов для уточнения молекулярных основ патологии (Kashevarova, et al., 2013; Kashevarova, et al., 2014; Shashi, et al., 2014). Однако даже при использовании самых современных клинических методов точный диагноз удается установить только в 46% случаев, причем суммарная стоимость обследований более чем в половине случаев превышает $25 000 (Shashi, et al., 2014).

Возможность установить причину заболевания на молекулярно-генетическом уровне прямо связана с используемыми методами диагностики. На сегодняшний день, самым применяемым в клинике подходом является микрочиповой анализ (CMA, chromosome hybridization array) (Miller, et al., 2010). Несмотря на то, что диагностическая точность этого метода превышает кариотипирование при помощи G-бендинга более чем в 2 раза, использование микрочипового анализа не позволяет поставить точный диагноз более чем в 80-85% случаев (Miller, et al., 2010). Поэтому, врачи

дополняют полученные результаты микрочипового анализа секвенированием индивидуальных генов-кандидатов или генных панелей. Такой подход, требующий от врача сформулированной *a priori* гипотезы о генетических нарушениях, вызвавших заболевание, часто оказывается неэффективным и не позволяет подтвердить диагноз.

Усилия последних лет, направленные на совершенствование и удешевление методов массового секвенирования, привели к появлению и распространению полноэкзомного секвенирования в диагностике врожденных патологий. Существующие оценки показывают, что анализ клинического экзома позволяет установить причину заболевания в ~25% случаев (Yang, et al., 2014; Lee, et al., 2014). Исследование клинического экзома пациента дает возможность не только идентифицировать вариации в кодирующих последовательностях, но и улавливать крупные делеции и дупликации (D'Aurizio, et al., 2016), однако уступает микрочиповому анализу в детекции относительно небольших вариантов (<100 КВ), особенно если перестройки расположены вне экзонов. В связи с этим, у пациентов с наследственной патологией приходится последовательно проводить как микрочиповой, так и полноэкзомный анализ.

Наконец, оба вышеперечисленных метода не позволяют идентифицировать сбалансированные перестройки (транслокации и инверсии) и сложные структурные вариации (хромосомные перестройки, в которых принимает участие одновременно несколько локусов), поэтому эти нарушения чаще всего выявляют цитогенетическим кариотипированием с использованием методов световой микроскопии. Именно классический цитогенетический анализ остается по-прежнему "золотым стандартом", поскольку это единственный метод, позволяющий визуализировать структуру хромосомной перестройки. Однако разрешение анализа, достигаемое при цитогенетическом кариотипировании, часто является недостаточным для постановки диагноза и требует применения молекулярно-цитогенетических технологий для уточнения тонкой структуры хромосомных аберраций, особенно в точках разрыва хромосом или вблизи них (Liehr, et al., 2018; Kashevarova, et al., 2018). Секвенирование геномов 230 пациентов со сбалансированными (по результатам цитогенетического анализа) транслокациями показало, что около 20% перестроек являются сложными, т.е. содержат в месте перестройки субмикроскопические участки более чем двух хромосом, причем часть перестроек (около 12 %) приводила к появлению крупных несбалансированных регионов (>100 КБ) (Redin, et al., 2017). Более того, около 30% исследованных сбалансированных перестроек сопровождаются разрывом гена, ассоциированного с наблюдаемым у пациента фенотипом (Redin, et al., 2017).

Необходимо отметить, что сбалансированные хромосомные перестройки часто являются причиной развития патологий. Например, у пациентов с нарушениями интеллектуального развития сбалансированные хромосомные перестройки встречаются приблизительно в 5 раз чаще, чем в среднем в популяции (Nielsen, et al., 1991; Ravel, et al., 2006; Marshall, et al., 2008; Funderburk, et al., 1977; Jacobs, et al., 1974). Поэтому развитие высокочувствительных методов, способных детектировать сбалансированные перестройки, является актуальной задачей в клинической диагностике.

Наиболее эффективными методом для детекции всех видов генетических вариаций, включая сбалансированные перестройки, является полногеномное секвенирование. Этот метод имеет в четыре раза более высокую диагностическую эффективность чем микрочиповой анализ (Stavropoulos, et al., 2016). Однако высокая стоимость полногеномного секвенирования не позволяет применять его в рутинной клинической практике. Для уменьшения стоимости полногеномного секвенирования можно использовать нестандартные подходы к конструированию геномной библиотеки, которые позволяют более эффективно идентифицировать хромосомные перестройки при меньшей глубине секвенирования (Talkowski, et al., 2011).

Стандартная paired-end библиотека Illumina позволяет секвенировать концы фрагментов ДНК размером 200-400 п.о. Таким образом, только фрагменты генома, находящиеся на расстоянии не более 400 п.о. от перестройки, будут нести информацию об её границе. Для идентификации перестроек в этом случае необходимо приблизительно 30‑кратное покрытие генома (200-400 млн

ридов). Создание так называемых mate-pair и jumping-library библиотек позволяет секвенировать концы фрагментов размером 3-4 тысячи п.о. В этом случае гораздо бОльшая доля просеквенированых фрагментов будет нести информацию о перестройке, что, в свою очередь, позволяет снизить глубину секвенирования до ~20 млн ридов, а значит – уменьшить стоимость анализа в разы (Vergult, et al., 2014; Talkowski, et al., 2011). В перспективе, диагностика структурных перестроек может опираться и на альтернативные методы секвенирования, которые позволяют напрямую определять последовательность длинных фрагментов ДНК (до 100 КБ), например, PackBio или Oxford Nanopore (Weirather, et al., 2017). Однако на сегодняшний день эти методы слишком дороги и имеют целый ряд технических сложностей для внедрения в клиническую практику.

Альтернативой mate-pair библиотек является создание 3C-библиотек, в которой каждый фрагмент ДНК ковалентно связан с пространственно-близким регионом генома (Fishman, et al., 2018). Изначально, технология 3C (**C**hromosome **C**onformation **C**apture, захват конформации хромосом) разрабатывалась для исследования трехмерной организации генома (Dekker, et al., 2002). Однако полногеномный вариант метода под названием Hi-C (Lieberman-Aiden, et al., 2009; Rao, et al., 2014), который позволяет анализировать в одном эксперименте пространственную организацию в масштабе всего генома, показал, что вероятность контакта двух участков в пространстве ядра зависит, в первую очередь, от линейного расстояния между ними и описывается степенной функцией (Battulin, et al., 2015; Fishman, et al., 2018). В типичной Hi-C-библиотеке, 15% всех контактов участка приходится на локусы, лежащие на расстоянии менее 10 КБ от него; ещё 15% распределены по в десять раз более протяженному региону на расстоянии 10-100 КБ; 18% - по регионам, удаленным на 100 КБ – 1МБ, и так далее (Dudchenko, et al., 2017). За счет хромосомных перестроек изменяется расположение участков в геноме – и это существенно отражается на частоте их пространственных контактов. Хотя изменения частот контактов хроматина будут наиболее выраженными вблизи границы перестройки, хромосомная аномалия будет влиять и на частоты удаленных контактов. Таким образом, информацию о перестройке несет большое количество разных контактов (=ридов), что позволяет выявлять структурные вариации используя небольшую глубину секвенирования.

Первые работы, предлагающие использовать Hi-C-библиотеки для сборки геномов, идентификации структурных вариантов и гаплотипирования, появились в 2013 году (Korbel, et al., 2013; Burton, et al., 2013; Kaplan, et al., 2013). На тот момент идея идентификации структурных перестроек на основе Hi-C-данных упоминалась только как теоретически возможная, а основной акцент был сделан на использование Hi-C для сборки геномов. Недавно было опубликовано ещё две работы, непосредственно нацеленные на использование Hi-C для детекции хромосомных перестроек в раковых клетках (Harewood, et al., 2017; Chakraborty, et al., 2018). При этом эффективность такого подхода оказалась крайне высокой, что позволило снизить глубину секвенирования до уровня, сходного с mate-pair библиотеками.

Важно отметить, что секвенирование Hi-C библиотеки позволяет не только определить порядок расположения локусов в линейном геноме и, за счет этого, диагностировать хромосомные перестройки, но и дает возможность оценить расстояние между локусами в пространстве ядра. Известно, что в хроматин уложен в ядре неслучайным образом (Dixon, et al., 2012; Rao, et al., 2014; Battulin, et al., 2015). Исходя из анализа Hi-C данных, можно выделить компактно расположенные регионы генома – топологически ассоциированные домены (ТАДы), которые пространственно изолированы от окружения. Внутри этих регионов энхансеры сближены с теми промоторами, регуляцию которые они обеспечивают (Dixon, et al., 2012). Инсуляция ТАДов необходима для предотвращения неспецифического взаимодействия энхансеров с промоторами (Lupiáñez, et al., 2015). Показано, что хромосомные перестройки, затрагивающие границы ТАДов, могу реализовывать свой патологический эффект не за счет нарушений кодирующих последовательностей генов, а за счет изменения трехмерной организации локусов и появления “незаконных” взаимодействий промоторов и энхансеров (Lupiáñez, et al., 2015; Franke, et al., 2016;

Zepeda-Mendoza, et al., 2017; Ordulu, et al., 2016). По последним оценкам, до 7% патологических сбалансированных перестроек связаны с нарушением трехмерной организации ТАДов (Redin, et al., 2017). При этом Hi-C и другие 3C методы являются практически единственным эффективным способом исследования тонких нарушений трехмерной архитектуры ядра.

Подводя итог, следует отметить, что на сегодняшний день не существует технологий, которые могли бы применяться в рутинной клинической практике для детекции различных мутаций, таких как точковые модификации в экзонах, субмикроскопические делеции, дупликации и сбалансированные перестройки. Разработка новых технологий, а также подходов к интерпретации выявленных геномных вариантов - как с точки зрения нарушения известных функциональных элементов генома, так и с точки зрения трехмерной архитектуры ядра, является актуальным вопросом современной генетики.

Полученные результаты

1. Разработан экспериментальный протокол «Exo-C», являющийся модификацией метода ДНКазного Hi-C (Ma et al., 2019). Протокол Exo-C включает забор периферической крови пациентов, выделение фракции лимфоцитов, создание Hi-C библиотек и их последующее обогащение экзомными последовательностями, используя панель *Roch MedExome target capture* (см. протокол Exo-C). Используя разработанный метод, мы собрали и просеквенировали геномные библиотеки шести пациентов из группы, описанной в п. 1, а также библиотеку для линии клеток K562 (табл. 1). Клетки линии K562 несут ряд хорошо охарактеризованных хромосомных перестроек, в геноме этих клеток описаны точковые полиморфизмы и для K562 ранее получена Hi-C карта высокого разрешения. Таким образом, клетки линии K562 могут быть использованы для оценки точности разрабатываемых подходов для поиска генетических полиморфизмов на основе метода Exo-C.

Кроме того, для одного из пациентов мы собрали, помимо Exo-C, Hi-C-библиотеку, используя для фрагментации хроматина эндонуклеазу рестрикции DpnII, вместо ДНКазы I. Полученная библиотека, как и все остальные, была обогащена экзомными последовательностями и секвенирована.

2. Первичный анализ результатов секвенирования показал, что в полученных данных присутствует значительная доля неинформативных последовательностей, т.е. последовательностей, не несущих информацию о трехмерных контактах хроматина и/или о геномных полиморфизмах у пациентов.

--- Во-первых, значительное количество прочтений содержало последовательность адаптеров Illumina на конце (13-81%). Доля адаптеров Illumina коррелировала с характерным размером фрагментов в библиотеках, определенным по результатам анализа электрофоретической подвижности. Нами был сделан вывод, что в будущем необходимо увеличить размер продуктов гидролиза ДНКазой, условия озвучивания ДНК и оптимизировать этап селекции фрагментов по размерам (AMPure size-select) так, чтобы добиться большего размера фрагментов в библиотеках.

--- Во-вторых, мы обнаружили, что в библиотеках Exo-C содержится значительное количество биотинилированных адаптеров (последовательности адаптеров могут выявляться в 40-50% всех прочтений) и их конкатемеров, что уменьшает эффективность картирования прочтений на ~10% и, вероятно, приводит к увеличению количества нелигированных фрагментов ДНК (dangling ends) в библиотеках. Детальный анализ расположения и структуры конкатемеров адаптеров (см. раздел методы и подходы) позволил предположить, что изменение последовательностей использованных адаптеров и их концентрации, а также ферментативная рецессия концов ДНК при помощи T4-полимеразы перед обогащением биотином, позволит увеличить долю целевых последовательностей в протоколах ДНКазного Hi-C и Exo-C.

Таким образом, по результатам анализа данных секвенирования, выявлен ряд недостатков метода Exo-C и предложены способы их решения.

3. Показан нуклеосомо-опосредованный паттерн в профиле трехмерных контактов. Мы обнаружили повышенное количеством контактов между участками генома, находящихся на расстояниях 75, 190, 385, 575, 775 п.о. (и далее с интервалом ~200 п.о; рис. 1). Такой паттерн наблюдался только в случае гидролиза хроматина ДНКазой I, в то время как при использовании фермента DpnII для приготовления библиотек частоты контактов равномерно падают с расстоянием. Вероятно, найденный паттерн отражает преференцию ДНКазы I к гидролизу хроматина в межнуклеосомных районах. Выявленная закономерность позволяет предположить, что данные Exo-C можно использовать не только для исследования трехмерной организации генома, но и для изучения закономерностей распределения нуклеосом. Например, интересным кажется сравнение паттерна контактов в эу- и гетерохроматиновых районах, для которых известны различия в длине межнуклеосомного линкера. Также интересно оценить паттерн контактов вблизи ДНКаза-чувствительных районов и около +1-нуклеосомы транскрипционно-активных генов. Эти задачи запланированы на следующий год выполнения Проекта.

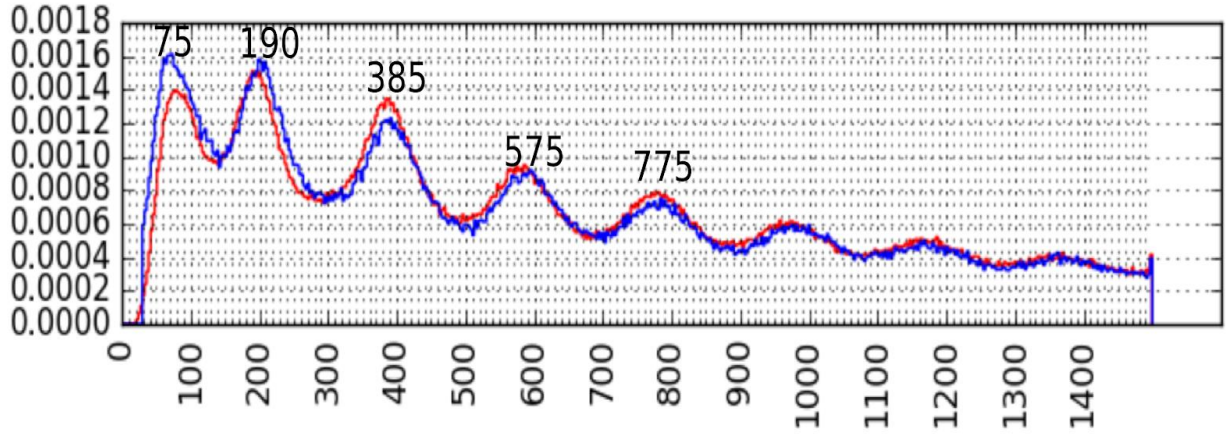

Рисунок 1. Зависимость частоты контактов от геномного расстояния для образцов Exo-C. По оси X – геномное расстояние в нуклеотидах (от 1 до 1500 п.о.), по оси Y – частота контактов. Красным показаны данные для прочтений с ориентациями FF и RR. Синим – для прочтений с ориентацией RF. См. также расшифровку обозначений ориентаций прочтений и их смысл в разделе «методы и походы».

4. Анализ глубины покрытия прочтениями показал, что целевые области (экзоны генов использованной панели *Roch MedExome target capture*) обогащены прочтениями по сравнению с остальным геномом приблизительно в 12 раз. Глубина прочтений варьирует между библиотеками и зависит, в первую очередь, от размера секвенированных фрагментов (при коротких фрагментах прочитанными оказываются адаптеры Illumina, что снижает количество прочитанных букв генома). При глубине в 60 млн прочтений в режиме PE2x150 (парные прочтения, 2*150 букв), среднее покрытие целевых районов составляет ~50X в библиотеках с оптимальным размером вставки. Таким образом, протокол Exo-C позволяет эффективно обогащать библиотеки экзомными последовательностями.

В библиотеке, приготовленной с использованием фермента рестрикции DpnII (а не ДНКазы), покрытие целевых последовательностей также оказалось высоким (~50X при глубине секвенирования в 60 млн прочтений). Однако, распределение прочтений по целевым последовательностям оказалось намного менее равномерным, по сравнению с Exo-C библиотеками. Так, целевые последовательности, непосредственно прилегающие к сайту рестрикции (<20 пар оснований от сайта DpnII), были прочитаны в 10 и более раз чаще, чем нуклеотиды на расстоянии >300 пар оснований. Поэтому, процент непокрытых прочтениями последовательностей для DpnII-библиотеки оказался ~6.5 раз выше, чем для Exo-C библиотек со сравнимым средним покрытием (5.9% целевых последовательностей не покрыты ни одним прочтением в DpnII-библиотеке, 0.87% - в Exo-C-библиотеке со сравнимым размером фрагментов, из расчета на 60 млн. прочтений; при этом среднее покрытие ~50X для обеих библиотек). Аналогично, медианное покрытие оказалось ниже

для DpnII-библиотеки почти в два раза (17 против 33). Таким образом, целесообразно использовать фермент ДНКазуI, или аналогичные ферменты, не имеющие фиксированного сайта узнавания, для создания обогащенных библиотек.

В результате анализа данных секвенирования для каждого из пациентов найден ряд геномных полиморфизмов. Полученные данные будут проанализированы биоинформатиками и врачами-генетиками для поиска вариантов, которые могут являться причиной патологического фенотипа пациентов в рамках следующего года выполнения Проекта.

5. Показана возможность поиска хромосомных транслокаций на основе Exo-C данных на примере линии клеток K562. Мы провели анализ обогащенных 3C-контактов в клетках K562 и H1, полученных в работе Ma. et al. 2019. В этом анализе мы использовали простейший, «наивный» метод поиска транслокаций, основанный на том, что частота межхромосомных контактов заметно ниже частоты контактов двух соседних локусов хромосомы. Используя информацию об известных (Dixon et al. 2018) межхромосомных транслокациях в клетках K562, мы валидировали этот метод – нам удалось основе 3C-данных выявить все границы транслокаций с разрешением 1 MB. Важно, что данные Ma et al., 2019, были получены при помощи обогащения 3C-библиотек последовательностями, суммарная длина которых в разы меньше, чем в использованном нами наборе *Roch MedExome target capture.* Тем не менее, этих данных уже было достаточно для поиска ряда перестроек. Это означает, что анализ перестроек пациентов, анализируемых в рамках Проекта, можно смело проводить с использованием *Roch MedExome target capture* и других, сопоставимых по длине целевой (обогащаемой) области генома, наборов.

Следует также отметить, что полученные нами результаты были опубликованы в журнале *Russian Journal of Genetics* (WOS).

6. Построены карты пространственных контактов хроматина для пяти пациентов. Репрезентативный пример полученной карты приведен на рис. 2. В дальнейшем, мы проведем анализ полученных профилей контактов для поиска структурных вариантов в геномах.

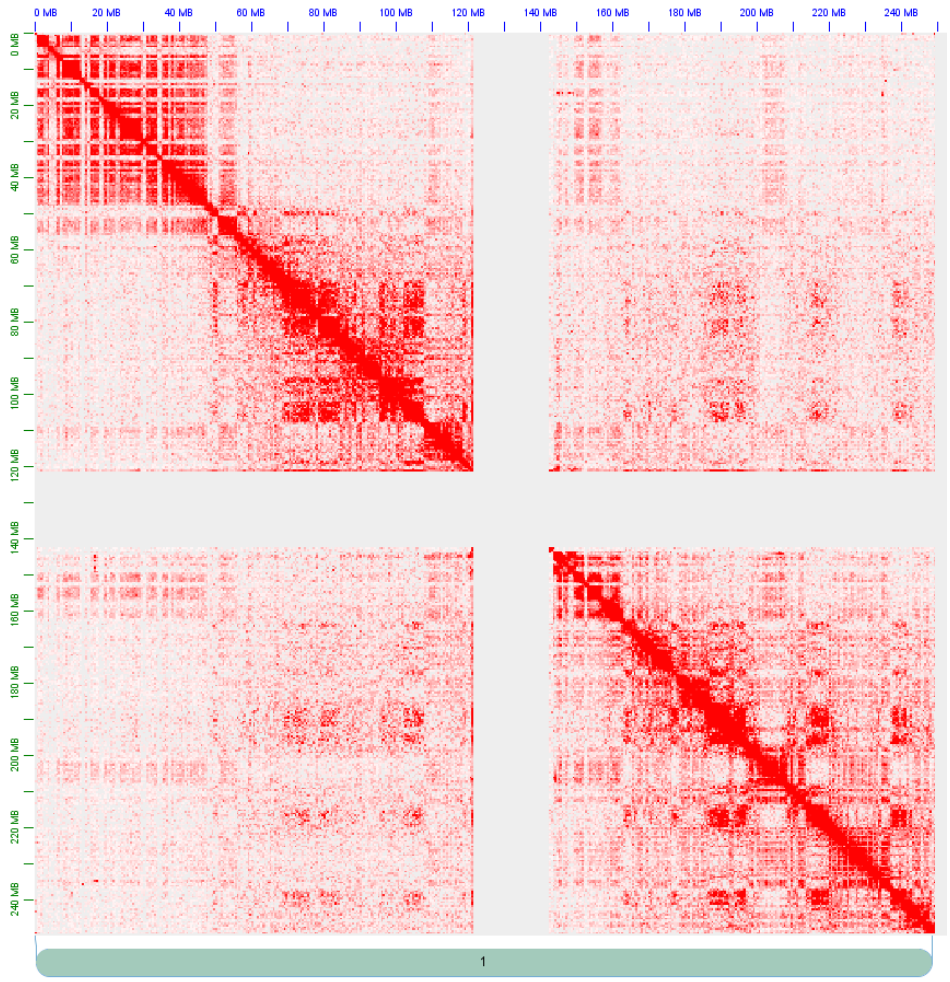

Рисунок 2. Карта пространственных контактов хромосомы 1 пациента 5. Карта получена на основе метода Exo-C.

7. Создан алгоритм для предсказания трехмерной организации перестроенных геномов. Мы разработали алгоритм *3DPredictor* для предсказания трехмерной организации хроматина в нормальных (неперестроенных) геномах с разрешением 25 тысяч пар оснований в рамках предыдущего проекта (РНФ 17-74-10143). В рамках текущего Проекта, мы расширили возможности *3DPredicor*, добавив специализированные модули для предсказания 3D-организации генома в случае делеций, дупликаций и инверсий, а также повысили разрешение анализа до 5 тысяч пар оснований. Кроме того, мы показали, что предсказания архитектуры хроматина могут быть сделаны с высокой точностью на основе лишь небольшого количества предикторов, среди которых наиболее информативными являются сайты связывания белка CTCF и их ориентация, геномное расстояние и данные об уровне транскрипции. Более того, информация о транскрипционной активности может быть заменена на сведения об участках открытого хроматина, обогащенных меткой H3K27Ac. Публикация, включающая результаты проделанной работы, опубликована в журнале *Genome Research*.

Подробное описание работы, включая используемые алгоритмы

Анализ прочтений потребовал тесной интеграции между биоинформатиками и молекулярными биологами. Благодаря их совместной работе, удалось не просто выявить последовательности

конкатемеризованных адаптеров, но и объяснить, они образуются в ходе Exo-C эксперимента, и как можно избежать их формирования в будущем.

Картирование прочтений

В соответствии с протоком Exo-C, секвенированные фрагменты ДНК представляют из себя две геномные последовательности, которые соединены одинаковым для всех адаптером (bridge-bridge). Этот адаптер может располагаться в случайном месте фрагмента, который секвенируется с двух концов, что в результате дает пару прочтений R1 и R2. Предполагается, что при лигировании двух последовательностей в ходе Exo-C-протокола, в R1 попадает часть одной геномной их них, а в R2 - часть второй. Но, если длина одной из лигированных последовательностей меньше длины прочтения, то в соответствующее прочтение попадет соединяющий два фрагмента адаптер и часть второй геномной последовательности, что помешает выровнять прочтение на геном (такие прочтения мы называем «химерными». Таким образом, прежде чем выравнивать прочтения необходимо удалить с 3’ конца последовательность адаптера вместе с частью второй последовательности. Для этой задачи был успешно использован cutadapt. При этом для анализа полиморфизмов мы «разрезали» химерные последовательности, сохраняя 5’- и 3’-части, объединяли все полученные последовательности (независимо от их принадлежности к R1 или R2) в один файл и проводили выравнивание утилитой bowtie2. Для анализа 3C-контактов мы удаляли 3’-части химерных последовательностей, сохраняя информацию о соответствии между прочтениями R1 и R2. После этого выравнивание и первичный анализ выровненных данных проводили утилитой Hi-C-Pro.

Анализ глубины покрытия в целевом геноме

Глубина покрытия вычислялась с помощью утилиты bedtools coverage (с параметром -hist — построение гистограммы, отображающей количество букв референса с конкретной глубиной покрытия).

Анализ 3C-контактов

Анализ 3C-контактов проводили при помощи утилиты Hi-C-Pro. Визуализацию карт контактов проводили при помощи JuiceBox. Определение доли нелигированных фрагментов ДНК (dangling ends) в библиотеках проводили на основе анализа распределения ориентаций прочтений.

У выровненных прочтений есть две возможные ориентации – прямая (Forward, F) или обратная (Reverse, R). Если прочтение совпадает с прямой ориентацией референсного генома, то имеет ориентацию F, если с обратной - R. Каждая пара прочтений может иметь любую пару ориентаций: FF, RR, FR, RF, где первая буква обозначает ориентацию более «левого» (лежащего на 5’-конце фрагмента генома) прочтения, а вторая – более «правого» (на 3’-конце). Ориентации пар прочтений в 3-C-контактах может быть любой из четырех, т.е. все варианты ориентации теоретически должны иметь соотношение 1:1:1:1. Но, кроме пар прочтений, обуславливающих трехмерный контакт, в 3-C-библиотеках содержатся нелигированные фрагменты ДНК (dangling ends), которые всегда имеют ориентацию FR. Кроме того, участки генома, лигированные сами с собой с образованием кольца, всегда имеют ориентацию RF. При этом прочтения, имеющие ориентации FF и RR, точно являются 3-C контактами. Таким образом, если количество FR и RF существенно выше количества прочтений в ориентациях FF и RR, можно приблизительно оценить процентное содержание нелигированных фрагментов и колец по формуле 100%*(RF+FR–RR-FF)/(RF+FR+RR+FF), где RF, FR, RR, FF – число прочтений в соответствующей ориентации.

После получения уникально картированных прочтений было обнаружено существенное превышение прочтений с ориентациями FR и RF во всех образцах (табл. 2).

Образец	Нелигированные фрагменты + кольца, % от уникальных выравненных пар
1	60,71272

2	48,6653
3	51,85185
4	38,04546
5	56,53974
6	58,31695
7	43,66197
8	58,94711
9	49,32432
DNAseI Ma et. al	35,53212

Таблица 2. Доля нецелевых фрагментов в полученных библиотеках.

При этом анализ распределения FR и RF прочтений показал, что нелигированные фрагменты составляют существенно большую долю прочтений, чем кольца.

Присутствие нелигированных фрагментов может быть связано с неэффективным обогащением биотином, маркирующим адаптер-опосредованные сайты лигирования (см. п. 6 раздела *Захват конформации хромосом*). Второй возможно причиной появления нелигированных фрагментов является «забивка» концов фрагментов ДНК конкатемерами адаптеров, которые имеют «тупой» конец и не могут вступать в дальнейшее лигирование с фрагментами ДНК, имеющими на конце нормальный «липкий» адаптер. Дискриминировать эти две причины появления нелигированных молекул можно по анализу последовательностей: первые не должны содержать последовательностей биотинилированных адаптеров, вторые, наоборот, должны содержать такие последовательности, причем преимущественно на концах. Мы написали программное обеспечение, которое выполняет такой анализ. Оказалось, что оба типа молекул (с адаптерами на концах и без адаптеров) имеют приблизительно равную частоту встречаемости среди нелигированных фрагментов. Таким образом, целесообразно в будущем, во-первых, увеличить количество отмывок на этапе обогащения биотинилированными последовательностями и, во-вторых, изменить последовательности адаптеров для предотвращения лигирования несвязанных с геномом адаптеров к концам молекул ДНК. Кроме того, кажется перспективным применение альтернативного метода сборки библиотек, опубликованный в августе этого года (BAT Hi-C; doi.org/10.1016/j.ymeth.2019.08.004), в котором предлагается лигирование фрагментов хроматина с участием одного «липкого» адаптера вместо двух, используемых в Exo-C в настоящий момент.

Идентификация хромосомных транслокаций на основе данных по секвенированию обогащенных 3C-библиотек

Работы по поиску перестроек в обогащенных 3C-данных подробно описаны в статье, опубликованной в журнале «Генетика». Статья приложена к данному отчету.

Моделирование архитектуры хроматина при помощи алгоритма 3DPredictor

Модель 3DPredictor подробно описана в тексте статьи (Belokopytova et al., 2020).

Эффект от использования кластера в достижении целей работы.

Вычислительные ресурсы кластера были необходимы для картирования последовательностей и других вычислительно-сложных операций.

Перечень публикаций, содержащих результаты работы

Belokopytova et al., 2020, Genome research, IF=11

https://genome.cshlp.org/content/early/2019/12/19/gr.249367.119