Аннотация

Мы выполнили ряд ключевых шагов для разработки нового метода диагностики наследственных заболеваний человека, основанном на комбинации методов захвата конформации хромосом и экзомного секвенирования.

Во-первых, мы существенно переработали существующий протокол приготовления Hi-C-библиотек при помощи ДНКазы, сделав его более воспроизводимым и эффективным, а также показав его преимущества перед классическим Hi-C для детекции однонуклеотидных вариантов в экзоме пациента. Мы совместили этот протокол с экзомным обогащением и, таким образом, разработали методику Exo-C.

Во-вторых, мы собрали достаточно большую выборку индивидуумов (более 40) с различными хромосомными перестройками. Для всех полученных образцов мы выполнили Exo-C-анализ, получив, таким образом, одну из самых больших коллекций унифицированных 3C-карт контактов для образцов человека.

В-третьих, благодаря анализу полученных 3C-данных, мы определили ранее описанных границы хромосомных перестроек и нашли новые, не доступные для детекции классическими методами, перестройки. В ряде случаев это позволило нам выдвинуть гипотезы о молекулярном механизме развития патологии. Проведение современных функциональных тестов, включая получение и дифференцировку индуцированных плюрипотентных клеток и полногеномный анализ экспрессии генов, дало возможность подтвердить часть из предложенных гипотез и, таким образом, установить молекулярную причину заболеваний.

В-четвертых, мы разработали две биоинформационные модели, позволяющие предсказывать трехмерные контакты перестроенных локусов. Модель 3DPredictor позволяет оценивать функциональное значение изменений архитектуры хроматина, вызванных хромосомной перестройкой. Эта модель полезна в тех случаях, когда хромосомная перестройка была найдена “классически” методом, а не с помощью Exo-C, но ее функциональные последствия могут реализовываться на уровне контактов хроматина. Алгоритм ReMOD предназначен для моделирования выбросов, отклонений и шумов в данных Exo-C и других Hi-C-данных. Этот алгоритм позволяет создавать распределения со статистиками, соответствующими наблюдаемым при хромосомных перестройках. Такие распределения необходимы для разработки и валидации биоинформационных алгоритмов для детекции перестроек.

- **Тема работы**

Разработка новых методов детекции хромосомных перестроек человека

- **Состав коллектива:**

Фишман Вениамин Семенович, НГУ, старший преподаватель

Полина Станиславовна Белокопытова, НГУ, аспирант

Евгений Александрович Можейко, ИЦиГ, аспирант

- **Информация о гранте: РФФИ 18-29-13021, 2018-2022**
- **Научное содержание работы:**
    1. **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.

2. 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 (Chromosome Conformation Capture, 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.

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.

3. Подробное описание работы, включая используемые алгоритмы.

Анализ трехмерной организации генома

Разработанные алгоритмы для поиска хромосомных перестроек подробно описаны в аспирантской работе Можейко Е.А., а также коротко приведены ниже.

1. Алгоритм поиска межхромосомных транслокаций:

Разобьем весь геном на равные непересекающиеся участки длиной в 10 КБ. Отсортируем такие участки по возрастанию геномной координаты, и присвоим каждому такому участку b порядковый номер k. Также присвоим каждой хромосоме порядковый номер: для человека это от 1 до 24. Обозначим за C_k^iколичество контактов участка b_kсо всеми участками хромосомы i . При транслокации участка b_kс хромосомы i на другую хромосому j, относительно глубины секвенирования снизится величина C_k^i и повысится C_k^j. Это произойдет потому, что частота межхромосомных контактов существенно ниже, чем частота внутрихромосомных. Пусть C_k^i и C_k^j- значения в контрольном образце без перестроек, а x и y - соответствующие значения в исследуемом образце. Также предположим, что случайные величины x и y независимы друг от друга. Тогда вероятность, что в исследуемом образце нет транслокации с хромосомы i на другую хромосому j участка b_k, можно вычислить по следующей формуле:

$$V = P^{i,j,k}(x,y) = P(x < C_k^i) \cdot P(y > C_k^j).$$

Порог V, который разделял транслоцированные и нормальные локусы мы подбирали эмпирически.

2. Алгоритм поиска cis-транслокаций.

Разобьем весь геном на равные непересекающиеся участки длиной в 10 КБ . Отсортируем такие участки по возрастанию геномной координаты, и присвоим каждому такому участку b порядковый номер k. Для каждого участка b_k разобьем область генома левее от участка b_k и область правее на более крупные участки (5 МБ) $b'_{\square}$, принадлежащие той же хромосоме, что и b_k. Также отсортируем такие участки по возрастанию геномной координаты, и присвоим каждому такому участку $b'_{\square}$ порядковый номер l. На основе разбиения b', для каждого участка b_kпостроим вектор контактов этого участка со своей хромосомой v_k. В отличие от алгоритма поиска межхромосомных транслокаций, в алгоритме поиска внутрихромосомных транслокаций мы все же строим одномерную матрицу контактов v_k для исследуемых участков. Пусть этот вектор v_kпосчитан для исследуемого образца. Обозначим за u_k соответствующий вектор в контрольном образце. Тогда дискретный интеграл

$$I = \int_{\square}^{\square} ¿ v_k - u_k \vee ¿ ¿$$

и есть искомое значение, на основе которого мы определям, транслоцирован ли участок b_k . Чем больше I, тем больше вероятность, что участок b_k, транслоцировался в другое место на своей хромосоме. Порог значения I также подбирался эмпирически.

Анализ ONT данных. Данные ONT выравнивали инструментом minimap2, затем анализировали выравнивания в программе для визуализации данных IGV

4. Полученные результаты.

- **Разработка экспериментального протокола Exo-C, позволяющего совместить захват конформации хромосом с экзомным обогащением.**

Идея проекта базировалась на создании 3C-метода, не требующего использования эндонуклеаз рестрикции с фиксированным сайтом узнавания, чтобы обеспечить равномерное покрытие прочтениями экзонов генов. В рамках первого года работ мы показали, что при использовании ДНКазы I наблюдается гораздо более равномерное покрытие прочтениями, чем при использовании рестриктазы DpnII с фиксированным сайтом узнавания (рис. 1). Поэтому, мы сфокусировались на совмещении протокола ДНКазной-Hi-C (который к нашей удаче был опубликован группой Ma et al. в 2018 году, практически одновременно с положительным решением о предоставлении грантового финансирования), с этапом экзомного обогащения.

Первые наши эксперименты оказались неудачными во всех отношениях:

приготовленные по протоколу Ma et al. Hi-C-библиотеки показали низкое качество (большое количество случайных, не опосредованных хроматином, лигирований), а степень обогащения экзомными последовательностями существенно уступала показателям классического экзомного секвенирования (рис. 2 Pool1). Важным результатом нашей работы стала оптимизация отдельных этапов протокола Hi-C, которая позволила существенно повысить качество данных, а также уровень экзомного обогащения.

Несмотря на существенное улучшение качества библиотек, в версии протокола, опубликованной в статье Gridina et al., остался один параметр, который нам хотелось бы улучшить - количество нехимерных фрагментов ДНК (dangling ends). В статье Gridina et al., анализируя библиотеки приготовленные с использованием олигонуклиотидного адаптера, мы показали, что эти фрагменты могут образовываться за счет обратных лигирований, т.е. не обязательно, что их присутствие свидетельствует о неэффективности этапов рестрикции/лигирования. Тем не менее по нашим наблюдениям в библиотеках Exo-C стабильно наблюдалось в 2-3 раза больше dangling ends, чем в обычных Hi-C-библиотеках. В протоколе Exo-C проводится биотин-стрептавидиновая селекция химерных фрагментов, которая основана на том, что на концы молекул перед лигированием вводится биотинилированная буква, а после лигирования фрагменты ДНК обрабатывают T4 полинуклеотидкиназой. В результате, биотинилированные нуклеотиды оказываются только внутри фрагментов ДНК, образовавшихся в ходе лигирования.

Мы предположили, что никазная активность ДНКазы (этот фермент может вносить как одно-, так и двухцепочечные разрывы в ДНК) приводит к включению биотинилированных нуклеотидов на этапе мечения концов и внутри фрагментов. Биотинилированные таким образом молекулы будут оставаться в библиотеке после стрептавидиновой селекции независимо от того, вступали они в реакцию лигирования или нет. Из-за этого, снижается эффективность биотин-стрептавидинового обогащения продуктами лигирования и повышается доля нехимерных фрагментов в библиотеке.

Мы предположили, что гидролиз альтернативным ферментом (без никазной активности) может помочь решить эту проблему. Протестировав несколько нуклеаз, мы обнаружили, что S1-нуклеаза способна гидролизовать ДНК в условиях фиксированного хроматина. После подбора условий и оптимизации, нам удалось получить 3C-библиотеки с использованием этого фермента. Оказалось, что без потери в качестве данных, использование S1-нуклеазы позволяет снизить количество dangling ends приблизительно в 1.5 раза (по сравнению с ДНКазой I; рис. 2, пул 12). Следует подчеркнуть, что, насколько нам известно, это первый в мире опыт использования S1-нуклеазы в 3C-протоколах. В момент написания отчета ещё около десятка библиотек, приготовленных нами по протоколу с S1-нуклеазой, находятся на секвенировании.

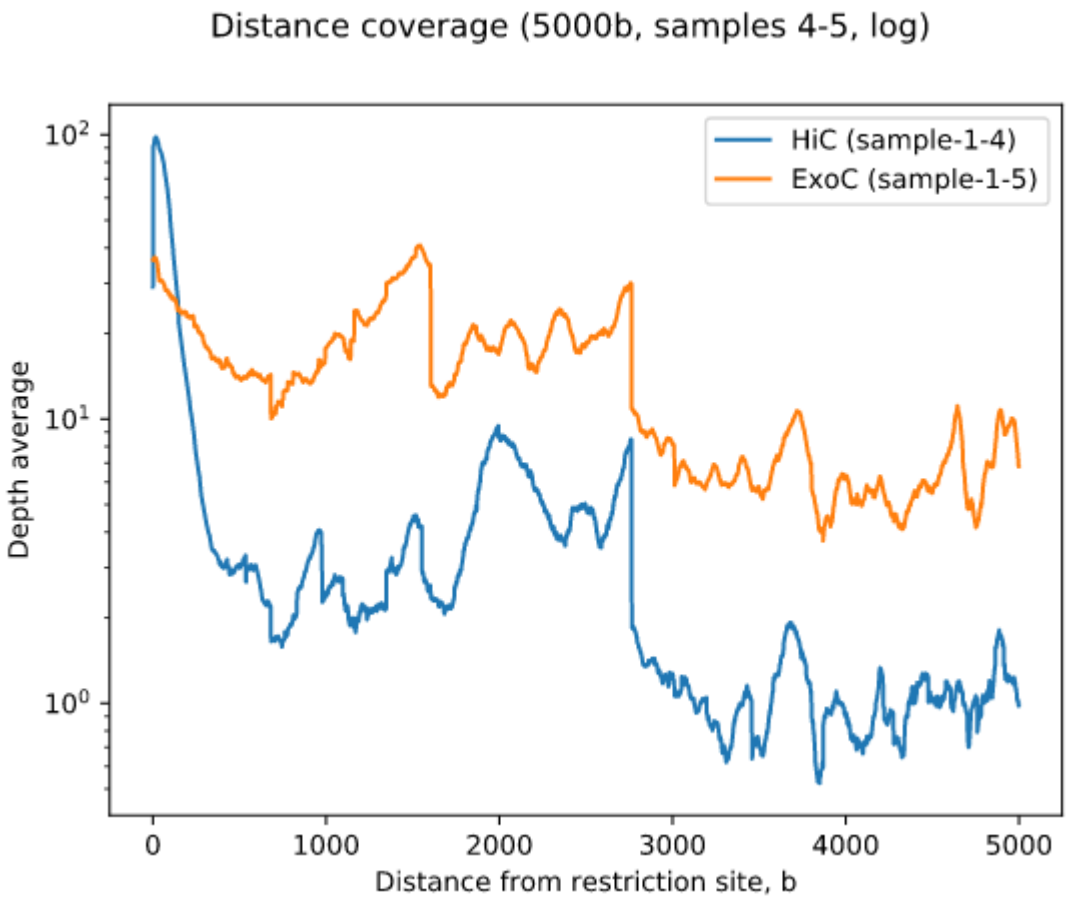

Рисунок 1. Зависимость геномного покрытия (ось Y, логарифмическая шкала) от расстояния до ближайшего сайта рестрикции DpnII в нуклеотидах (ось X) для образца, приготовленного с использованием ДНКазы I (Exo-C) или эндонуклеазы рестрикции DpnII (Hi-C).

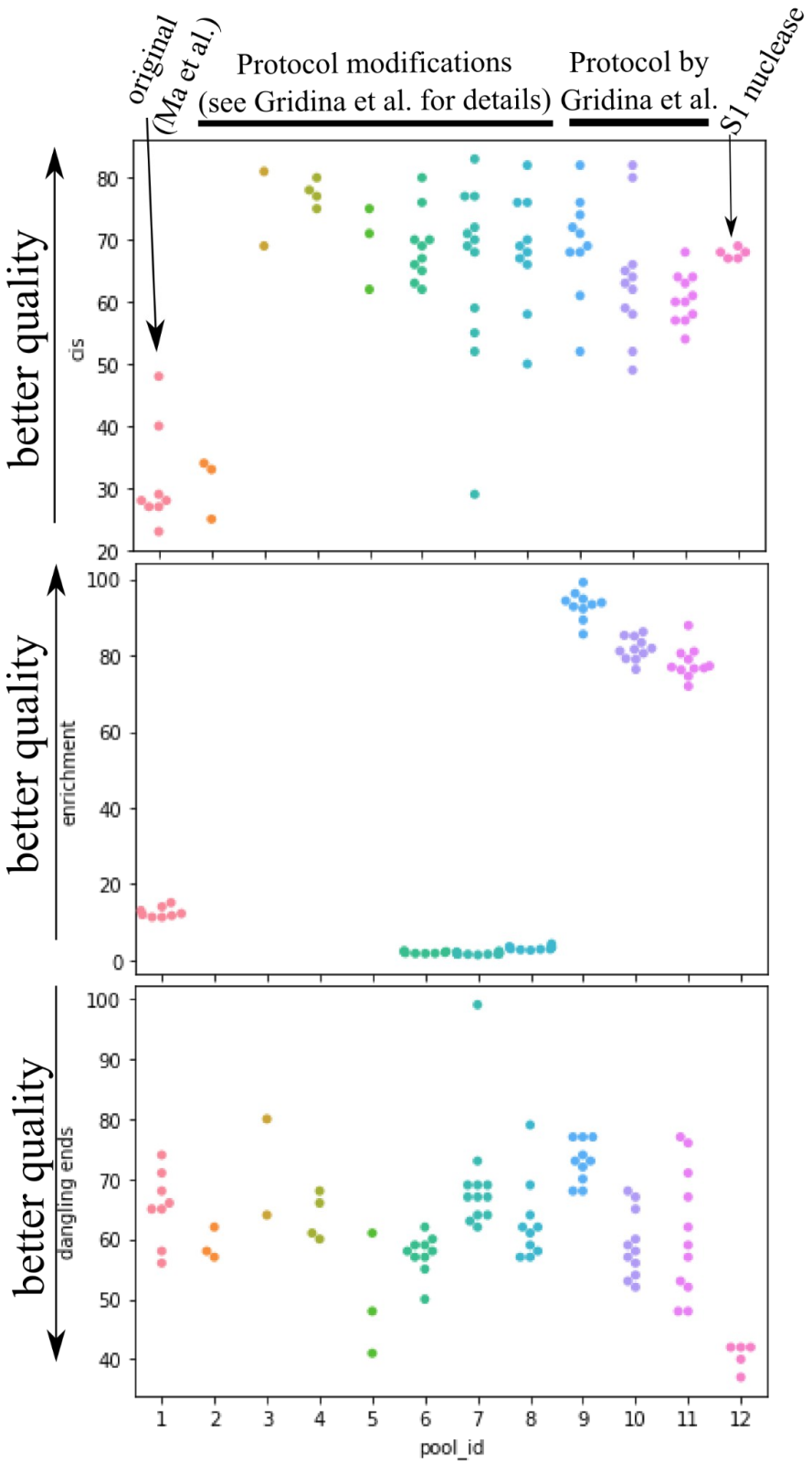

Рисунок 2. Метрики качества Exo-C-библиотек, приготовленных с использованием разных протоколов. Метрика cis (% внутрихромосомных контактов) отражает шум в 3C-данных (чем выше cis, тем меньше шум), метрика enrichment (отношение среднего

покрытия в экзоме к среднему покрытию в оставшейся части картируемого генома) отражает степень обогащения, метрика dangling ends (число нехимерных фрагментов ДНК, посчитанное на основе пропорции ридов в FF/RR/FR/RF ориентациях) отражает долю ридов, несущих информацию о трехмерных контактах (чем ниже метрика, тем больше информативных ридов). Следует отметить, что не все образцы доводили до стадии экзомного обогащения и глубокого секвенирования, поэтому для части пулов статистика по обогащению не представлена.

- **Получение карт трехмерных контактов хроматина для 45 индивидуумов, несущих различные хромосомные перестройки.**

Используя разработанный протокол Exo-C, мы получили карты трехмерных контактов для 45 индивидуумов (таблица 1). Мы показали возможность получения Exo-C библиотек из разного биоматериала: 34 библиотеки были получены из первичных лимфоцитов периферической крови без предварительной заморозки; 5 библиотек - из замороженных образцов мононуклеаров крови; 3 библиотеки - из культур перевивных клеточных линий (LNCaP, K542, A549); 3 библиотеки - из индуцированных плюрипотентных стволовых клеток, ранее полученных из первичных культур пациентов. Все 45 библиотек были секвенированы с глубиной от 30 до 90 млн прочтений. 31 библиотека была приготовлена с использованием наиболее продвинутой версии Exo-C протокола, описанной в статье Gridina et al., 2021, с высоким уровнем обогащения (покрытие в экзоме в 80-100 раз превышает покрытие остальной части генома). Совокупная карта 3C-контактов лимфоцитов, построенная по нашим данным, включает более чем 1.5 миллиарда прочтений.

Мы подбирали доноров материалов для сборки Exo-C-библиотек, основываясь на нескольких критериях. Во-первых, мы собрали выборку пациентов с описанными хромосомными перестройками: транслокациями (16 пациентов), инверсиями (6 пациентов; эта часть работы в 2022 году была выполнена при поддержке фонда РНФ, в связи с уменьшением запланированного финансирования гранта РФФИ в два раза), делециями и дупликациями (9 пациентов). Эту группу пациентов мы использовали для валидации разрабатываемых подходов к детекции геномных вариантов на основе Exo-C-данных. Во-вторых, мы взяли в анализ несколько пациентов, у которых был проведен тот или иной генетический тест первой линии (хромосомный микроматричный анализ и кариотипирование - в случае подозрения на хромосомную аномалию; экзомное секвенирование - в случае подозрения на моногенное заболевание), но при этом патогенный вариант найден не был.

Поскольку фокусом исследования была разработка метода детекции генетических вариантов, а не изучение какой-то определенной группы генетических заболеваний, мы не ограничивали набор участников исследования пациентами с какой-то одной нозологией. В нашей выборке оказались пациенты с различными фенотипами (таблица 1), однако большинство из них имело те или иные отставания в психоречевом и/или моторном

развитии.

В совокупности, полученные нами за время выполнения проекта данные представляют собой одну из самых больших коллекций унифицированных 3С-карт образцов периферической крови человека в мире.

Код пациента	**Контактное лицо, ответственное за передачу материала**	**Хромосомный пол**	**Хромосомные аномалии, найденные до проведения Exo-C-анализа**	**Краткое описание фенотипа участника исследования**
P1	**Назаренко Людмила Павловна**	**Женский**	**Не найдены**	**Предварительный диагноз муковисцидоз не подтвержден**
P2	**Назаренко Людмила Павловна**	**Мужской**	**Не найдены**	**Задержка психо-речевого развития. Отсутствие речи. Агрессивное поведение. Макроцефалия**
P3	**Назаренко Людмила Павловна**	**Женский**	**Не найдены**	**ДЦП. Открытое овальное окно. Симптоматическая фокальная эпилепсия. Микроцефалия**
P4	**Назаренко Людмила Павловна**	**Мужской**	**Не найдены**	**Задержка психо-речевого развития. Отсутствие речи. Агрессивное поведение**
P5	**Назаренко Людмила Павловна**	**Мужской**	**Не найдены**	**Задержка психо-речевого развития.**
P6	**Яблонская Мария Игоревна**	**Мужской**	**Не найдены**	**Нейрофиброматоз 1 типа**
P7	**Яблонская Мария Игоревна**	**Женский**	**Не найдены**	**Нейрофиброматоз 1 типа**

P8	**Кузьменко Наталья Борисовна**	**Женский**	**Не найдены**	**Неуточнённый иммунодефицит. Нейтропения. Тромбоцитопения.**
P9	**Кузьменко Наталья Борисовна**	**Женский**	**Не найдены**	**Микроцефалия. Органическое поражение ЦНС**
P10	**Шилова Надежда Владимировна**	**Женский**	**46 XX t(5;10)**	**задержка психомоторного развития**
K562			**Множественные (раковая линия)**	
A549			**Множественные (раковая линия)**	
LNCap			**Множественные (раковая линия)**	
P26	**Назаренко Людмила Павловна**	**Мужской**	**arr[hg19] 3q13.31(115890242_116022056)*1**	**Гидроцефалия**
P31	**Назаренко Людмила Павловна**	**Мужской**	**Не найдены**	**Задержка психо-речевого развития**
P32	**Назаренко Людмила Павловна**	**Женский**	**Не найдены**	**MIM: 611134**
P35	**Назаренко Людмила Павловна**	**Мужской**	**arr[hg19] 8p22(15362801_15842397)*1**	**Spina bifida S2**
TAF3	**Назаренко Людмила Павловна**	**Мужской**	**arr[hg19]3p26.3(1197623 -1492721)*1**	**Задержка писхоречевого развития.**

TAF4	**Назаренко Людмила Павловна**	**Мужской**	**dup[hg19]chr3:560685-1504666**	**Задержка писхоречевого развития.**
P37	**Шилова Надежда Владимировна**	**Женский**	**arr[hg19] 8q24.11(117921970_118869338)x3**	**Врожденная изолированная расщелина неба 3Б степени.**
P38	**Шилова Надежда Владимировна**	**Женский**	**arr[hg19] 8q24.11(117927697_118901851)x3**	**Норма, микроделеция 8q24.11 у дочери**
TAF9	**Назаренко Людмила Павловна**	**Женский**	**[hg19](chr2:32444-2667650)x1;[hg19] (chr2:2754133-24334429)x3**	**Задержка писхоречевого развития.**
P39	**Лукьянова Татьяна Витальевна**	**Мужской**	**arr[hg19] 8p22(16798008_17889195)x1, 8p21.3p21.1(19131256_28163113)x1**	**Множественные пороки развития**
P40	**Лукьянова Татьяна Витальевна**	**Женский**	**46, XX, inv(12)(q13.3;q24.1)**	**Фенотипические проявления отсутствуют. У плода кариотип 46,XY,inv(12)(q13.3;q24.1)mat.**
P42	**Шилова Надежда Владимировна**	**Мужской**	**46XY, t(4;12)(p14;q22)dn**	**Энцефалопатия. Задержка физического развития.**
P43	**Шилова Надежда Владимировна**	**Мужской**	**46XY, ins(2;4)**	**MIM: 618547, MIM: 156200**
P44	**Назаренко Людмила Павловна**	**Мужской**	**46XX, t(5;13)(p15;q22)**	**Норма, обращение в связи с проблемами репродукции**

P45	**Назаренко Людмила Павловна**	**Женский**	**45XX rob(13;14)**	**Рождение ребенка с ВПР (декстракардия, диафрагмальная грыжа), смерть на 3 сутки. Анэмбриония. У партнера хромосомных аномалий не выявлено.**
P46	**Назаренко Людмила Павловна**	**Мужской**	**46XY t(6;18;21) (p12q23, q23q22)**	**Норма, обращение в связи с проблемами репродукции**
P47	**Назаренко Людмила Павловна**	**Мужской**	**46XY, t(8;11;21) (q23;q22;q21)**	**MIM:133700**
P48	**Назаренко Людмила Павловна**	**Мужской**	**46XY inv(7)(p11q11.2)**	**Autism, ASD**
P49	**Назаренко Людмила Павловна**	**Женский**	**46 XX inv(7)(p11q11.2)**	**Хромосомная перестройка обнаружена у плода в ходе скрининга.**
P50	**Назаренко Людмила Павловна**	**Мужской**	**46XX, t(1;9) (q10.11;p10.11)**	**Норма, обращение в связи с проблемами репродукции**
P51	**Назаренко Людмила Павловна**	**Женский**	**46XX, t(1;9) (q10.11;p10.11)**	**Норма, обращение в связи с проблемами репродукции**
P52	**Назаренко Людмила Павловна**	**Женский**	**46XX, inv(3)**	**Норма, хромосомная перестройка обнаружена у плода в ходе скрининга.**
P53	**Назаренко Людмила Павловна**	**Женский**	**46XX**	**Миелодисплстический синдром. D46.2 Рефратерная анемия с избытком бластов**

P54	**Беляева Елена Олеговна**	**Мужской**	**t(2;18)**	**Задержка писхоречевого развития**
P56	**Шилова Надежда Владимировна**	**Женский**	**46,XX,t(2;20)(p24.2;p13)dn**	**Задержка писхоречевого развития.**
P55	**Шилова Надежда Владимировна**	**Мужской**	**46,XY,t(7;16)(p13;q23)dn**	**Задержка писхомоторного развития.**
P58	**Шилова Надежда Владимировна**	**Мужской**	**46,XY,inv(7)(p13q21.2)pat**	**Гидроцефалия**
P57	**Шилова Надежда Владимировна**	**Мужской**	**46,XY,t(3;10)(p21;q11.2)**	**Шизотипическое расстройство**
P62	**Шилова Надежда Владимировна**	**Мужской**	**46,XY,t(1;4)(p32.2;q33)dn**	**Задержка психоречевого развития**
P69	**Шилова Надежда Владимировна**	**Мужской**	**46,XY,t(2;6)(p13;q13),t(7;11)(q31.2;p15.3)dn**	**Задержка психоречевого развития**
P63	**Мусатова Елизавета Валерьевна**	**Мужской**	**46,XY,inv(2)(p21q23);t(3;7)(p13;q11.2)**	**Задержка психоречевого развития, аутизм**
P71	**Беляева Елена Олеговна**	**Мужской**	**arr[hg19]5q34(163585144_167946316)*1,9p24.2p23(2267812_12275177)*1**	**Не предоставлено**

Таблица 1. Список участников исследования (совокупные данные за 3 года). Цветом обозначены три пациента, не взятые в валидационную выборку при поиске

транслокаций (см. подпункт 4 ниже)

- **Детекция клинически-значимых вариантов у пациентов с генетическими заболеваниями.**

Среди пациентов с неустановленной молекулярной причиной патологии по результатам Exo-C-эксперимента удалось найти патогенные или вероятно патогенные однонуклеотидные варианты в экзомах трех человек. Кроме того, мы смогли подтвердить наличие патогенных хромосомных перестроек у подавляющего большинства пациентов, для которых патогенный вариант уже был аннотирован до начала анализа с использованием классических цитогенетических методов диагностики (см. также информацию о подтвержденных вариантах в подпункте 4 этого раздела).

Среди найденных нами патогенных вариантов, некоторые не были описаны ранее в литературе и представляют отдельный интерес для изучения. Мы также обнаружили некоторые варианты, которые могут быть отнесены к вариантам неопределенного клинического значения согласно текущим стандартам диагностики, и провели ряд подтверждающих исследований для уточнения клинического значения этих вариантов.

- **Разработка алгоритмов для автоматического поиска хромосомных перестроек на основе Exo-C-данных и их экспериментальная валидация.**

Крупные, цитологически-видимые хромосомные перестройки достаточно легко обнаружить при беглом визуальном анализе Exo-C-карт. Однако мелкие хромосомные перестройки размером несколько КБ или десятки КБ визуально обнаружить гораздо сложнее. Для поиска таких хромосомных перестроек Е.М. Можейко, в рамках выполнения аспирантской работы под руководством В.С. Фишмана, разработал ряд биоинформационных алгоритмов.

Суть алгоритма для поиска сбалансированных транслокаций (инсерций) заключается в следующем: мы рассматриваем вероятность наблюдать изменение соотношения частот cis/trans-контактов фиксированного локуса относительно контрольных значений cis/trans-контактов в этом локусе. Таким образом, при транслокации, за счет того, что частота cis-контактов существенно больше частоты trans-контактов, мы увидим отклонение отношения cis/trans от контрольного в данном локусе. Теоретическое обоснование данного алгоритма приведено нами в статье Mozheiko and Fishman, 2019,. Предложенный нами подход имеет сразу несколько преимуществ по сравнению с существующими:

- В отличие от существующих методов, мы учитываем, что при межхромосомной транслокации повышается не только частота межхромосомных контактов, но и, при этом, также снижается частота внутрихромосомных контактов. Таким образом, дополнительная информация, о снижении частоты внутрихромосомных контактов, позволяет повысить точность и специфичность метода.

- В нашем методе, мы не строим двумерную матрицу контактов, а рассматриваем вероятности наблюдать изменение соотношения частот cis/trans контактов каждого локуса в отдельности. Это позволяет на порядок уменьшить вычислительные затраты, и при этом кратно уменьшить минимальный размер детектируемых транслокаций. Эта главная особенность нашего метода, которая, как будет показано ниже, позволяет детектировать небольшие (меньше 10 КБ) транслокации, в то время как аналогичные методы работают на минимальном разрешении 40 КБ и гораздо большей глубине секвенирования.
- И, в заключении, наш алгоритм поиска межхромосомных транслокаций адаптирован для обогащенных последовательностями экзонов Hi-C данных, так как не привязан к взаимодействиям отдельных локусов между собой, а рассматривает сумму всех контактов локуса с хромосомой. Тем самым мы учитываем повышенную неравномерность покрытия, и полное отсутствие покрытия в некоторых участках генома.

На клеточной линии K562, у которой подробно описаны в литературе 19 крупных (больше 1 МБ) транслокаций (Dixon et. al 2018), мы оценили точность и специфичность разработанного метода для крупных (больше 1 МБ) транслокаций, и сравнили полученные результаты с другим, самым современным, методом детекции межхромосомных транслокаций HINT (Wang et. al 2020) (рис. 6). Несмотря на то, что на Hi-C данных без обогащения и с гораздо большей глубиной секвенирования, HINT показывает высокую эффективность [Wang et. al 2020], на рисунке 6 видно, что на данных Exo-C, он значительно уступает нашему методу.

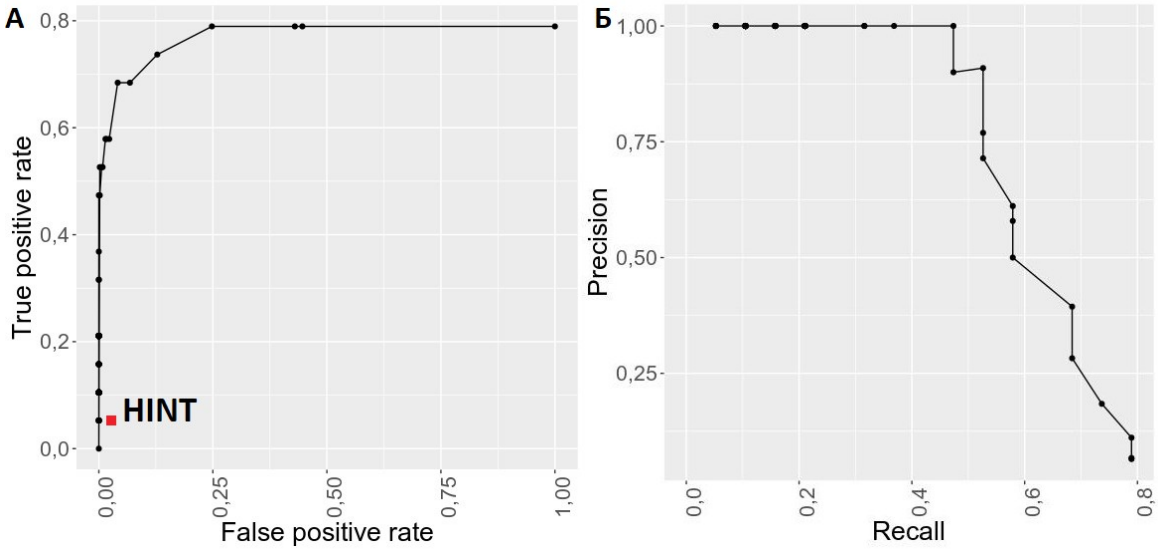

Рисунок 6. Оценка эффективности метода для поиска межхромосомных транслокаций на тестовой Exo-C-библиотеке клеточной линии K562. A. True positive rate - доля истинно позитивных транслокаций. False positive rate - доля ложно позитивных

транслокаций. За 100% было взято число возможных пар хромосом (23*22 пар). Красной точкой обозначено соответствующее значение для алгоритма HINT (Wang et. al 2020) Б. График Precision \ Recall (чувствительности \ специфичности) для тех же данных.

Убедившись в работоспособности алгоритма, мы применили его для анализа Exo-C-данных пациентов, исключив из анализа раковые линии клеток (в выборку попали 33 пациента из первых трех пулов; оставшаяся часть данных ещё не была секвенирована к моменту проведения анализа). У этих пациентов алгоритм аннотировал 634 локуса, для которых на Exo-C-карте наблюдается паттерн инсерции. Число 634 не следует интерпретировать как число найденных у пациентов уникальных инсерций. Во-первых, алгоритм тестирует возможность инсерции для каждого 10-КБ участка независимо, таким образом инсерции больших районов или обмены плечами могут приводить к появлению нескольких сигналов, а не одного. Во-вторых, мы обратили внимание на то, что большая часть из обнаруженных инсерций была найдена у двух или более индивидуумов. Таким образом, число уникальных обнаруженных нами инсерций существенно меньше 634.

С другой стороны, следует отметить, что распространенные в популяции хромосомные перестройки зачастую не выявляются в нашем анализе, поскольку детекция хромосомных перестроек основана на сравнение исследуемого образца с контролем. Следовательно, если у исследуемого образца и контроля присутствует одинаковая перестройка, она не будет обнаружена.

Мы провели визуальный анализ всех обнаруженных перестроек для создания валидационной выборки. Во-первых, мы обнаружили, что все 11 цитологически-видимых крупных транслокаций, взятые в анализ (таблица 1), были обнаружены алгоритмом (мы исключили из выборки транслокации родственных пациентов P50 и P51, а для пациента P56 данные ещё не были получены к моменту анализа; все остальные пациенты были взяты в это исследование). Интересно, что была обнаружена робертсоновская транслокация у пациента P45, которую при визуальном анализе Exo-C-карт мы детектировать уверенно не смогли.

Во-вторых, мы выбрали восемь примеров небольших, субмикроскопических инсерций для независимой валидации методом ПЦР. Размеры выбранных инсерций, уточненные по результатам визуального анализа Hi-C-карт, варьировали от 1 до 51 КБ, т.е. все они не могли быть найдены при цитологическом кариотипировании .

Три инсерции были найдены у нескольких пациентов. Для всех этих трех перестроек мы смогли найти в литературе данные о том, что данная перестройка распространена в популяциях человека (см. ссылки в графе “validation” в таблице 2).

Ещё пять хромосомных перестроек были уникальными, т.е. обнаружены только у одного пациента. Для двух случаев, мы смогли амплифицировать границу перестройки используя ПЦР (таблица 2). Кроме того, для одного из этих случаев граница перестройки была также подтверждена результатами, полученными при полногеномном секвенировании, выполненном для данного пациента ранее.

#	chr from	chr to	size	Found in multiple patients?	Validation
1	5	2	<1 KB	Yes	https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00762-1
2	2	6	<1 KB	No	PCR – not validated
3	13	11	23000	No	PCR – validated; WGA - validated
4	19	10	43000	No	PCR – validated
5	4	13	51000	No	PCR – validated
6	12	2	2000	Yes	https://academic.oup.com/gbe/article/11/6/1679/5498151
7	12	9	36000	No	PCR – not validated
8	15	13	20000	Yes	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5435966/#B21

Таблица 2. Список инсерций, обнаруженных при помощи Exo-C-анализа, для которых была проведена валидации.

Поиск внутрихромосомных транслокаций и инверсий

Поиск внутрихромосомных транслокаций более сложная задача, чем поиск межхромосомных. В существующих методах для поиска транслокаций на основе Hi-C данных не предложены подходы для поиска внутрихромосомных транслокаций. Мы предлагаем следующий метод: если для фиксированного локуса построить распределение частот контактов со всеми остальными локусами на той же хромосоме, то мы увидим характерную картину зависимости частот контактов от расстояния в линейной молекуле ДНК. Если произошла транслокация, мы увидим смещение частот контактов влево или вправо, в зависимости от геномного расстояния, на которое переместился локус при транслокации, поскольку транслоцированный локус окажется ближе к участкам слева или справа от исходного места локализации. Идея метода заключается в том, чтобы посчитать интеграл модуля разности кривых зависимости частоты контактов от расстояния, наблюдаемых в контрольном и экспериментальном образцах.

Применение разработанного алгоритма к имеющейся у нас группе пациентов позволило детектировать 58 потенциальных внутрихромосомных транслокаций. Из семи

образцов, для которых были описаны цитологически-видимые инверсии, алгоритм корректно установил границы четырех. Для двух из трех образцов, для которых инверсии не были детектированы алгоритмом, границы перестройки проходят в центромерной области и инверсионный паттерн не наблюдается при визуальном осмотре Exo-C-карт. Таким образом, предложенный метод позволяет детектировать инверсии в ряде случаев, однако детальная оценка его чувствительности и специфичности требует существенного расширения выборки внутрихромосомных транслокациях. Частично эту проблему может решить моделирование хромосомных перестроек, описанное ниже.

Поиск делеций и дупликаций

Для поиска делеций и дупликаций мы использовали стандартные инструменты GATK. В прошлых отчетах мы подробно описывали нашу оригинальную идею по нормализации геномного покрытия. Однако детальное сравнение с последними версиями нормализации покрытия алгоритма GATK показало, что наш подход не имеет решающих преимуществ по сравнению с методом нормализации GATK model segments. Результаты применения метода GATK model segments позволили обнаружить часть делеций и дупликаций, которые были найдены у пациентов методом хромосомного микроматричного анализа (см. таблицу 1). Однако поскольку чувствительность и специфичность для метода GATK хорошо описана в литературе, а в нашем случае применение этого метода ничем не отличается от стандартного экзомного анализа, мы не проводили детальный анализ и валидацию всех найденных вариантов.

Поиск однонуклеотидных вариантов

Для поиска и однонуклеотидных вариантов в экзомах пациентов мы также использовали стандартные подходы на основе GATK. Важным результатом в этой части работ являлось сравнение списка вариантов, детектируемых по результатам Exo-C-анализа и классического экзомного секвенирования. Для этого мы использовали опубликованные экзомные и геномные данные клеток K562. Мы показали, что более 95% вариантов в экзоме клеток K562, найденных в результате классического экзомного секвенирования, детектируется и в Exo-C-данных при сходной глубине секвенирования. При использовании более строгих фильтров (исключение из анализа вариантов, найденных в одном исследовании и не подтвержденных в других опубликованных работах), уровень пересечения между публичными данными и данными Exo-C возрастает до 99%. Таким образом, метод Exo-C позволяет находить однонуклеотидные варианты с точностью, сравнимой с классическим экзомным секвенированием.

При интерпретации найденных вариантов мы строго следовали опубликованным в РФ рекомендациям для интерпретации данных, полученных методом NGS. Это позволило нам обнаружить патогенные варианты для трех пациентов. Все эти варианты были проверены секвенированием по Сэнгеру, во всех случаях результаты Exo-C и секвенирования по Сэнгеру совпали. Эти результаты показывают, что метод Exo-C клиническую эффективность как диагностический метод для выявления точковых

вариантов в экзомах пациентов.

- **Эффект от использования кластера в достижении целей работы.**

Выполнение большой части поставленных задач было бы невозможно без использования ресурсов вычислительного кластера НГУ