

ПРИЛОЖЕНИЕ 3

Логика подхода и определение средней ошибки суммарного анализа распределения частот выбранных SNP

Определение средней ошибки суммарного анализа распределения частот выбранных SNP

Первоначально на основании результатов полноэкзомного секвенирования была проведена оценка частот встречаемости SNP для кодирующих и некодирующих областей генов. Основной анализ был проведен для кодирующих областей генов клонального гемопоэза (КГ), для чего были взяты данные распределения частот, анализируемых генов в образцах ДНК МНК в точках 0 (до терапии), 4, 8, 12, 27 мес. после проведенной реинфузии МНК, обогащенных реконструированными ГСК (Приложение 2: табл. 1 и 2). В финальном варианте сравнивались значения в нулевой и конечной точках наблюдения (Приложение 2: табл. 2, рис. 1). Дополнительно для выбора критерия для достоверной оценки снижения частоты SNP в конечной точке наблюдения по сравнению с нулевой точкой были отобраны несколько SNP из некодирующих областей генов (Приложение 2: табл. 3).

Характеристика табличных значений изменения частот встречаемости SNP кодирующей области генов КГНП

Анализ табличных значений частоты встречаемости SNP кодирующих и некодирующих областей генов свидетельствует о следующем. Общее количество детектируемых мутантных аллелей кодирующих областей составило 19, некодирующих – 109. Частоты детектируемых SNP в кодирующих областях генов и их доброкачественный статус свидетельствуют о том, что это герминальные мутации. Согласно распределению частот SNP, гены находятся как в гомо-, так и в гетерозиготном состоянии.

Для кодирующих вариантов:

- SNP, выявленные в 12 генах, определяются с частотой ~100 % на протяжении всего периода наблюдения, что свидетельствует о том, что это гомозиготные гены с аллелями герминального происхождения;
- 6 и 19 SNP кодирующих областей определяются с частотой ~50 % на протяжении всего периода наблюдения, что свидетельствует о том, что это гетерозиготные гены с аллелями герминального происхождения;
- к 27-му месяцу наблюдения для 6 и 19 SNP кодирующих областей (~30 %) частота встречаемости понизилась на 8–18 %.

Для некодирующих вариантов анализ не проводился ввиду сложно интерпретируемых результатов.

Препарат hDNA^{gr}

Частоты аллелей различных генов варьируют значительно, в диапазоне 17–100 %, что связано с происхождением препарата (ДНК получена от ~100 здоровых рожениц) (Приложение 2: рис. 1).

Генетические изменения, связанные с проведенной терапией, могли проявиться или как полное исчезновение SNP, или как снижение частоты встречаемости SNP в конечной точке анализа, которые были исходно (до обработки) выявлены в МНК. Если такое снижение будет обнаружено, то этот факт будет означать, что мутантные аллели были откорректированы в результате гомологического обмена с немутантными аллелями, пришедшими с экстраклеточными фрагментами (невозможность появления новых значений SNP, связанных с активацией проведенной обработкой ГСК, покоящихся с момента возникновения в эмбриогенезе, обсуждается в теле статьи).

При анализе SNP в кодирующих областях генов было обнаружено значимое снижение частоты встречаемости шести мутантных аллелей (*ASXL1*, *RAD21*, *DNMT1*, *SF3B1-1*, *SF3B1-2*, *SF3B1-3*) через 27 мес. наблюдения. Требовалось найти критерии достоверности этого снижения.

Объективные трудности в оценке достоверности снижения частоты встречаемости SNP в выбранных для анализа генах, находящихся в гетерозиготе

Для того чтобы провести сравнение и показать достоверное снижение частот встречаемости контролируемых SNP, требовалось разработать адекватный поставленной задаче подход.

Оказалось, что для 12 анализируемых генов экзонов частота встречаемости SNP составляет 100 %, что свидетельствует о том, что эти SNP находятся в гомозиготе. Частота встречаемости указанных SNP не изменилась на протяжении всего времени наблюдения. Это означает, что данные мутации могут находиться под давлением отбора. В этом случае, если в них «в моменте» произошли генетические изменения, связанные с

гомологической рекомбинацией с ДНК экстрахромосомальных фрагментов, то клоны, содержащие эти откорректированные мутантные аллели, были конкурентно вытеснены.

В анализ бралась ДНК, выделенная из МНК крови. Это означало, что анализировалось интегральное состояние гемопоэтической системы. Индивидуальный вклад костномозговых предшественников в частоту встречаемости того или иного SNP будет влиять в том случае, если часть аллеля находится в гомозиготном, а часть – в гетерозиготном состоянии.

Для шести анализируемых генов частота SNP колеблется между ~40–60 %. Эти значения могли означать, что: 1) это гетерозиготы с большим разбросом значений; 2) это гетерозиготы с включением в вариационный ряд также пролиферирующих клонов, содержащих гомозиготы по обнаруженному SNP; 3) это гетерозиготы с включением в вариационный ряд аллеля, пришедшего из препарата hDNA^{gr}, или исключением из вариационного ряда аллеля, в котором обнаруженное в нулевой точке SNP заместилось на «здоровый, немутантный аллель», пришедший из препарата hDNA^{gr}; 4) смешанные в различных комбинациях варианты.

Если рассматривать вариант изменения частоты встречаемости SNP как результат гомологического замещения аллелей генома на привнесенный аллель, то обнаружение такого изменения через несколько месяцев после проведенной терапии может означать, что эти аллели не подвергаются давлению отбора и вновь возникшие генетические изменения легко закрепляются в геноме.

Возможные технические причины разброса частот встречаемости SNP в нулевой и конечной точках анализа

Как было сказано выше, при анализе частот встречаемости SNP в шести генах (*ASXL1*, *RAD21*, *DNMT1*, *SF3B1-1*, *SF3B1-2*, *SF3B1-3*) обнаружен разброс значений между ~40–60 %, что предполагало, что эти аллели находятся в гетерозиготе. Одна из возможных причин такого разброса значений при определении частоты гетерозигот связана с методом селективного отбора кодирующих областей генов. Отбор ведется методом гибридизации. На подложке (или в растворе) присутствуют гомологичные последовательности всех генов. При добавлении ДНК для гомозиготных аллелей не важно, находится внесенная ДНК в избытке к матрице или нет, – всегда будет один и тот же результат: 100 % частота встречаемости. Для гетерозигот при добавлении избыточного количества ДНК может возникнуть диспропорция между аллелями в процессе гибридизации. Если присутствует избыток ДНК по отношению к матрице с использованием платформы $\times 100$, то всегда будет вероятность несимметричного заполнения всех валентностей матрицы, что приведет к искажению и разбросу данных от эксперимента к эксперименту. В условиях проводимых исследований мы не могли повлиять на этот параметр. При этом, если использовать платформу с покрытием $\sim \times 1000$, то возможность несимметричного заполнения матрицы многократно снижается, что дает более точный результат. В этом случае все (или близко к тому) аллели найдут свое «посадочное место» и будут секвенированы.

Несколько возможных причин, объясняющих появление разброса частот SNP гетерозиготных генов в течение периода наблюдения с использованием нескольких мультиплексных платформ разных производителей, ставили задачу поиска подхода оценки ошибки измерений.

Пределы ошибки анализируемых значений частот SNP для гетерозиготных аллелей, обсуждаемые в литературе

При анализе изменения (снижения) частоты встречаемости SNP в начальной и конечной точках наблюдения мы основывались на результатах исследований, характеризующих предел отсечения значений при анализе генов, находящихся в гетерозиготе по анализируемому SNP. В случае нормальной гетерозиготы пределом отсечения достоверных значений (не артефактов любого происхождения) будет частота встречаемости SNP (VAF) между 0.33 и 0.63 (Sears et al., 2025). Следует отметить, что полученные значения были рассчитаны при использовании данных полногеномного секвенирования с покрытием $\sim \times 100$ – $\times 200$, что делает результаты анализа полностью приемлемыми для настоящего исследования. Также следует отметить, что в анализ были взяты данные, полученные в различных лабораториях, в разное время, с использованием различных NGS инструментов. Мы полагаем, что видимый широкий разброс значений (0.33–0.63) при определении чистых гетерозигот связан именно с разнообразием методических вариантов, используемых для получения данных полногеномного секвенирования, выполненного в разных экспериментальных условиях. Тем не менее для подкрепленного материалами публикации значения отсечения артефактов мы взяли нижний предел отсечения (0.33), выбранный на основании анализа результатов массивов данных, полученных в разных лабораториях

в разное время. Кроме цитируемой работы, существуют рекомендации ассоциации молекулярной патологии (Association for Molecular Pathology, AMP), где также отмечается, что ошибка анализа для гетерозигот может достигать $\pm 15\%$ (Association for Molecular Pathology Training and Education Committee, n. d.). Приведенные цифры означают, что, если значения изменений частот SNP в отобранных биоинформатическим анализом экзонах генов клонального гемопоэза будут находиться в обозначенных пределах, то они будут вне пределов категории (артефактов любого происхождения (Sears et al., 2025)) и будут приемлемы для последующего статистического анализа. В источниках отмечается особо, что приведенный подход будет легитимен только в случае отсутствия мозаицизма любой природы (герминальный мозаицизм, опухоль-ассоциированный мозаицизм). Это замечание является важным обстоятельством, которое необходимо принимать во внимание при расчетах. В случае пациента К. как раз и предполагается, что снижение частот SNP связано именно с возникшим мозаицизмом вследствие замены мутантного SNP немутантной аллелью, пришедшей из терапевтической дцДНК hDNA^{gr}. Это означает, что предел отсекаемых достоверных значений (не артефактов любого происхождения) VAF между 0.33 и 0.63 требует внутренней коррекции. Мы решили, что корректным будет использовать имеющиеся данные исключительно настоящего исследования, при этом находясь в поле достоверности отсекаемых артефактов (0.33–0.63). В качестве оценки разброса частот, характерных для гетерозигот выбранных генов, связанного с использованием четырех различных мультифакторных панелей, был выбран следующий подход.

Определение ошибки измерений значений частот SNP гетерозиготных генов, полученных в результате использования четырех различных NGS панелей в конкретной лаборатории

На первом этапе мы оценили имеющийся разброс значений частот SNP шести гетерозиготных генов (*ASXL1*, *RAD21*, *DNMT1*, *SF3B1-1*, *SF3B1-2*, *SF3B1-3*) в нулевой и конечной точках согласно описанным выше рекомендациям. Оказалось, что все SNP полностью удовлетворяют требованиям предела отсекаемых артефактных значений и находятся выше значения 33 %.

Для достоверной оценки полученных результатов дополнительно требовалось оценить ошибку, выдаваемую используемыми мультифакторными панелями с покрытием $\sim \times 100$ и $\sim \times 1000$. Чтобы оценить ошибку измерений при определении гетерозигот четырьмя различными NGS платформами, был выбран следующий подход.

В нашем распоряжении, помимо анализа частот SNP в кодирующих областях генов клонального гемопоэза, находился анализ частот SNP в некодирующих областях, смежных с экзонами. Частоты SNP некодирующих областей нескольких генов оказались по своим значениям через 27 мес. наблюдения близки к значениям в нулевой точке и колебались вблизи отметки 50 % в промежуточных точках анализа (более низкие значения частот для одного случая, около 44 %, также были взяты в анализ, поскольку разброс находится в пределах одного значения во всех точках наблюдения) (Приложение 2: табл. 3). Сказанное означает, что выбранные аллели стабильно находятся в гетерозиготном состоянии, одинаково определяются как системой с покрытием $\times 100$, так и системой с покрытием $\times 1000$, на частоты их проявления не влияют другие факторы, как, например, гомозиготные клоны по анализируемой SNP, которые могут вытеснять клоны с гетерозиготными аллелями, и, значит, эти значения можно использовать для расчета ошибки определения гетерозигот выбранными мультиплексными панелями.

Оказалось, что все панели как занижают, так и завышают значения по сравнению с нулевой и некоторыми промежуточными точками. Эти изменения частот находятся в пределах нескольких процентов, и этот процент можно будет считать ошибкой использования нескольких мультиплексных панелей. Объединив разброс значений во всех точках наблюдения для выбранных аллелей, можно рассчитать средний % ошибки и сравнить нулевую и конечную точки (стандартное отклонение в %). Этот процент ошибки был определен и учтен при суммарном анализе изменений в частотах гетерозигот (Приложение 2: табл. 3).

Результаты сравнения частот встречаемости SNP в нулевой и конечной точках наблюдения для шести гетерозиготных генов, частота SNP которых в нулевой точке колебалась вблизи отметки 50 %

Из проведенных рассуждений следует, что определить изменения, связанные с замещением мутантного локуса в гетерозиготах, можно только в случае: 1) если распределение частот в нулевой точке находится вблизи 50 %, что подразумевает незначительный вклад гомозиготного варианта или его полное отсутствие; 2) если имеется достоверная зона отсекаемых ошибки определения частот гетерозиготы.

Оказалось, что достоверное снижение частоты встречаемости SNP выявляется у трех генов, *SF3B1-1*, *SF3B1-2*, *DNMT1*. Для трех других генов, *ASXL1*, *RAD21*, *SF3B1-3*, в рамках выбранного критерия достоверных отличий нет, но присутствует выраженная тенденция к снижению частоты встречаемости SNP. Сравнивались стандартные отклонения в нулевой и конечной точках наблюдения. Если эти значения не пересекались, то отличия считались достоверными в рамках выбранного критерия. Результаты сравнения приведены в теле статьи на рис. 1, А.

Было проведено конечное сравнение снижения частот встречаемости SNP генов *SF3B1-1*, *SF3B1-2*, *DNMT1*. Чтобы не учитывать возможную контаминацию минимальным количеством случайно появившихся гомозигот, для анализируемых генов мы выбрали интегральный подход оценки изменения частот встречаемых аллелей в трех отобранных гетерозиготных генах, нулевое распределение частот которых колеблется вокруг 50 % (см. рис. 1, Б). Сравнивались медианы частот в нулевой и конечной точках наблюдения.

Список литературы

- Association for Molecular Pathology Training and Education Committee Molecular In My Pocket ... ONCOLOGY : Interpretation of Genomic Assays n.d.
- Sears K., Hickey C., Vincent R., Stocks-Candelaria J., Tate J., Bumgardner C., Zhang S., Miller J.B. Establishing a variant allele frequency cutoff for manual curation of medical exome sequencing data. *J Mol Diagnostics*. 2025;27(1):36-41. doi 10.1016/J.JMOLDX.2024.09.006