КОЛИЧЕСТВЕННЫЙ И БЕЗОШИБОЧНЫЙ АНАЛИЗ ДАННЫХ МАССИРОВАННОГО СЕКВЕНИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ МОЛЕКУЛЯРНОГО БАРКОДИРОВАНИЯ

Е. С. Егоров[1,2], М. А. Израельсон[2], С. А. Касацкая[2], Д. М. Чудаков[1,2]✉, С. А. Лукьянов[1,2]

[1] Российский национальный исследовательский медицинский университет имени Н. И. Пирогова, Москва
[2] Институт биоорганической химии имени академиков М. М. Шемякина и Ю. А. Овчинникова РАН, Москва

За последние несколько лет технологии высокопроизводительного секвенирования (High Throughput Sequencing, HTS) прочно вошли в обиход современной биологии и медицины. Открылись принципиально новые возможности для секвенирования геномов и глубокого анализа различных ДНК- и РНК-библиотек. В то же время новые технологии содержат и новые подводные камни — ограничения в получении достоверной количественной и качественной информации при глубоком анализе сложных библиотек. Четкое понимание этих ограничений необходимо как для корректной интерпретации получаемой информации, так и для поиска технологических решений, позволяющих минимизировать их влияние. В обзоре мы показываем, как молекулярное баркодирование позволяет нормировать образцы и устранять ошибки полимеразной цепной реакции (ПЦР) и секвенирования при сохранении реального разнообразия библиотек в ходе HTS-анализа сложных библиотек.

Ключевые слова: молекулярное баркодирование, высокопроизводительное секвенирование, ДНК-библиотека, РНК-библиотека, анализ библиотек

✉ **Для корреспонденции**: Дмитрий Михайлович Чудаков
117997, Москва, ул. Островитянова, д. 1; chudakovdm@mail.ru
Статья поступила: 28.09.2015 **Статья принята к печати:** 22.10.2015 г

QUALITATIVE ERROR-FREE ANALYSIS OF MASS SEQUENCING DATA USING MOLECULAR BARCODING

Egorov ES[1,2], Israelson MA[2], Kasatskaya SA[2], Chudakov DM[1,2]✉, Lukyanov SA[1,2]

[1] Pirogov Russian National Research Medical University, Moscow
[2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow

Over the past few years the High Throughput Sequencing technologies have become well established in modern biology and medicine. Fundamentally new opportunities for genome sequencing and in-depth analysis of various DNA and RNA libraries have been discovered. However, new technologies come with new pitfalls such as limitations in obtaining valid qualitative and quantitative data in the in-depth analysis of complex libraries. A clear understanding of these limitations is necessary for the accurate interpretation of the collected data as well as for the search for technological solutions to minimize their impact. The following review shows how molecular barcoding helps normalize samples efficiently and eliminate PCR and sequencing errors while preserving the actual diversity of libraries in the course of HTS analysis of complex libraries.

Keywords: molecular barcoding, High Throughput Sequencing, DNA library, RNA library, library analysis

✉ **Correspondence should be addressed:** Dmitriy Chudakov
1, Ostrovityanova st., Moscow, 117997; chudakovdm@mail.ru
Received: 28.09.2015 **Accepted**: 22.10.2015

ОШИБКИ ПЦР И СЕКВЕНИРОВАНИЯ. МЕТОДИКИ ИХ КОРРЕКЦИИ

Возможность надежно идентифицировать редкие варианты генов в сложных образцах делает высокопроизводительное секвенирование (High Throughput Sequencing, HTS) эффективным подходом в онкодиагностике [1], пренатальной диагностике [2], анализе неоднородности и изменчивости опухолей [3], в исследованиях бактериальных [4] и вирусных [5] инфекций и микробиомов [6], а также в эволюционных исследованиях [7] и исследованиях репертуаров иммунных рецепторов [8–13]. Однако одной из ключевых проблем в анализе данных HTS является накопление ошибок полимеразной цепной реакции (ПЦР) и секвенирования. Некоторые направления практического применения технологии, например ранняя онкодиагностика, требуют надежной детекции редких мутаций, присутствующих в образце в концентрации менее 1 % и даже в концентрации порядка 0,1 % [13–18]. В таких случаях редкие подварианты последовательностей генов часто невозможно отличить от подвариантов, являющихся результатом ошибок секвенирования и предшествующей ПЦР-амплификации. Самый простой способ устранения ошибок основан на том, что редко встречающиеся подварианты рассматриваются как ошибочные производные одной и той же часто встречающейся последовательности и по этому признаку исключаются из последующего анализа (frequency-based filtration) [19–21]. Фильтрация минорных подвариантов может опираться на более или менее достоверную модель накопления ошибок ПЦР, но в целом произвольно регулируется выбором соотношения часто и редко встречающихся последовательностей в массиве данных. Такой способ фильтрации хотя и устраняет большую часть накопленных ошибок, влечет за собой также потерю значимой части реального разнообразия редко представленных подвариантов нуклеотидных последовательностей в образце. В результате наиболее часто используемые методы HTS позволяют достоверно обнаружить мутации, присутствующие

в образце только в значительной концентрации порядка 5 % [22–24] До недавнего времени не существовало метода, с помощью которого можно было бы проводить безошибочное глубокое секвенирование сложных библиотек генов при сохранении реального разнообразия гомологичных вариантов последовательностей. Прорыв в области количественного и безошибочного массированного секвенирования произошел с внедрением так называемого уникального молекулярного баркодирования [25, 26]. При этом подходе каждая исходная анализируемая молекула ДНК или кДНК маркируется уникальной нуклеотидной последовательностью, и эта информация затем используется в ходе программного анализа выходных данных массированного секвенирования (рисунок).

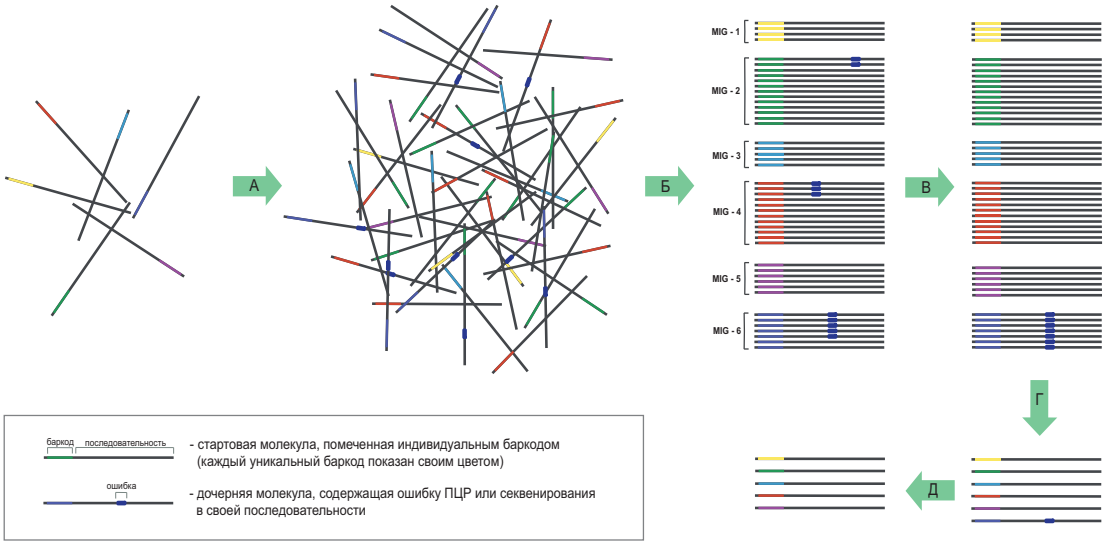
Опишем возможный вариант анализа молекулярно-баркодированных данных на примере разработанного нами подхода, получившего название Molecular Identifier Groups-based Error Correction (MIGEC) [10]. Он базируется на двухстадийном биоинформатическом анализе. Первая стадия относительно проста и основана на идее «безопасного секвенирования» (“Safe-Seqs”), предложенной в оригинальной работе проф. Vogelstein и соавторов [25]. Прочитанные последовательности, несущие один и тот же уникальный молекулярный идентификатор (баркод), объединяются в одну группу (кластер) — Molecular Identifier Group (MIG) (рисунок, Б). Наличие идентичного молекулярного идентификатора показывает, что данные прочитанные последовательности были наработаны с одной и той же стартовой молекулы ДНК или кДНК.

Соответственно, кластеризация данных секвенирования по уникальному идентификатору позволяет установить исходную нуклеотидную последовательность по доминирующей последовательности в каждой группе (рисунок, В).

Таким образом можно устранить многочисленные ошибки, накопленные в процессе амплификации, и исправить практически все ошибочно установленные в процессе секвенирования нуклеотиды. Однако на практике применения такого прямолинейного алгоритма оказывается недостаточно, для того чтобы добиться безошибочного глубокого анализа библиотеки интереса. Дело в том, что ошибки ПЦР, произошедшие на ранних стадиях амплификации, могут в определенной доле случаев образовывать доминирующую нуклеотидную последовательность в MIG вследствие стохастической природы ПЦР-амплификации, и результирующий ошибочный вариант последовательности интерпретируется как реально существующее разнообразие (рисунок, Г). Такие события составляют значимый процент анализируемой информации при глубоком HTS-анализе библиотек и не могут быть безопасно (для реального разнообразия образца) отсечены фильтрацией редких подвариантов.

Любопытно, что такие ошибочные события, во всяком случае наиболее представленные из них, не устраняются также независимыми повторностями амплификации и секвенирования образца, так как частота определенных ошибок ПЦР в определенном контексте ДНК оказывается хорошо воспроизводима [10]. Для того чтобы идентифицировать и устранить подобные ошибочные варианты, мы ввели вторую стадию анализа данных (рисунок, Д). Она основывается на том, что высокочастотные (в каждом конкретном контексте ДНК) ошибки ПЦР носят повторяющийся характер, что позволяет отличить их от реального разнообразия. Такие ошибки «выдают» себя тем, что встречаются в качестве минорного подварианта в большом числе MIG и, соответственно, могут быть идентифицированы на основе относительной частоты встречаемости варианта последовательности в виде «мажора» или «минора» в MIG [10].

Двухстадийный алгоритм позволяет отфильтровать ошибочные варианты последовательностей с высокой точностью, сохраняя при этом естественное разнообразие библиотеки, и предоставляет возможность проводить глубокий безошибочный анализ сложных библиотек, как было показано нами на примере анализа данных секвенирования библиотек иммуноглобулинов и Т-клеточных рецепторов [10], а также на примере мультиплексного генетического анализа для онкодиагностики (наши неопубликованные данные). На примере анализа гомологичных контрольных последовательностей с различной представленностью мы показали, что анализ MIGEC устраняет практически всё искусственное разнообразие, которое содержится в данных секвенирования. Соотношение сигнал-шум, посчитанное как соотношение числа прочтений наиболее крупной контрольной последовательности к самому представленному ошибочному подварианту с одной и двумя нуклеотидными заменами, возросло с 1 000:1 и 20 000:1 для стандартно процессированных данных до 12 000:1 и 60 000:1 соответственно при обработке MIGEC. В то же время контрольные низкочастотные подварианты последовательности с заменами одного и двух нуклеотидов были сохранены.



Коррекция ошибок и нормировка данных с использованием молекулярного баркодирования. **(А)** В ходе приготовления библиотек происходит многократное и неравномерное увеличение копий стартовых молекул кДНК или ДНК. Некоторые из дочерних молекул неизбежно содержат ошибки. **(Б)** Первая стадия анализа: прочитанные молекулы, имеющие один и тот же баркод, объединяются в одну группу (MIG). **(В)** По доминирующей последовательности внутри каждой группы происходит установление исходной последовательности. **(Г)** Кластеризация MIG — переход на счет стартовых молекул. **(Д)** Вторая стадия анализа (используется для коррекции ошибок при глубоком секвенировании): высокочастотные ошибки детектируются и удаляются из дальнейшего анализа.

КОЛИЧЕСТВЕННЫЙ АНАЛИЗ ДАННЫХ HTS

Независимо от выбранной технологии создания библиотек — на основе геномной ДНК либо на основе кДНК — невозможно обеспечить равную эффективность реакций на всех этапах пробоподготовки и секвенирования. Любая из стадий подготовки библиотеки генов к HTS-анализу (забор образца клеток, выделение ДНК или РНК, ПЦР-амплификация), как и само секвенирование, может приводить и непременно приводит к потерям и искажению информации о реальном количестве молекул и соотношении их подвариантов в исходном образце. Использование молекулярного баркодирования при анализе данных высокопроизводительного секвенирования позволяет контролировать число реально анализируемых стартовых молекул, успешно прошедших все стадии амплификации и секвенирования. В дальнейшем сравнительном анализе образцов также становится возможным оперировать не числом прочтений секвенирования, а количеством стартовых молекул. Таким образом, этот подход принципиально важен для понимания качества проведенного анализа и нормированного сравнения многих библиотек, в том числе полученных в разное время и разными лабораториями.

Так, в эксперименте, где с исходного образца, содержащего 1 000 гомологичных молекул, получают 10^6 прочтений секвенирования, число уникальных вариантов последовательности может варьироваться от 1 до 1 000 в зависимости от разнообразия этого образца. Однако детекция, например, 30 уникальных вариантов последовательностей среди полученного миллиона прочтений не дает однозначной информации о реальном составе образца. Действительно, эти 30 вариантов могут составлять разнообразие всей 1 000 исходных молекул или только лишь 30 молекул, успешно вошедших в амплификацию. В то же время обнаружение, например, 100 уникальных вариантов молекулярных баркодов позволяет с уверенностью говорить, что были проанализированы последовательности именно 100 стартовых молекул. Очевидно, что разрешающая способность такого эксперимента не может быть повышена за счет увеличения покрытия секвенирования, так как с увеличением числа прочтений число реально анализируемых молекул все равно останется равным 100. С помощью молекулярного баркодирования можно не только корректировать ошибки секвенирования и ПЦР, но и контролировать реальную узость «горлышка бутылки» для конкретного эксперимента, а также эффективно нормировать образцы для сравнительного анализа [9].

Становится возможным проводить точное сравнение двух и более библиотек генов, даже если они были получены с разного количества клеток и секвенированы с разной глубиной покрытия. Для этого, например, можно в ходе анализа использовать из каждого образца равное число случайно выбранных последовательностей, помеченных разными уникальными молекулярными баркодами. Поскольку каждая прочтенная последовательность с уникальным баркодом соответствует отдельной молекуле ДНК либо кДНК, такая нормализация данных секвенирования позволяет кардинально снизить уровень искажения количественной информации, накопленной в ходе амплификации и секвенирования кДНК-библиотеки. В результате снижается разброс данных для независимо полученных реплик и повышается относительное детектируемое разнообразие, так как каждое новое прочтение соответствует новой молекуле ДНК или кДНК. Применение молекулярного баркодирования также позволяет снизить количественные искажения относительной представленности вариантов последовательностей внутри каждой конкретной библиотеки [12], так как в ходе анализа элиминируются искажения, связанные с неравной эффективностью работы праймеров при мультиплексной ПЦР, стохастической природой ПЦР, предпочтений ПЦР и секвенирования относительно различных матриц [9, 25–27].

ПОДВОДНЫЕ КАМНИ МОЛЕКУЛЯРНОГО БАРКОДИРОВАНИЯ

Несмотря на все преимущества применения молекулярного баркодирования, эта технология имеет ограничения, которые необходимо учитывать при проведении качественного глубокого анализа данных HTS. В частности, ПЦР-ошибки внутри самой последовательности уникального идентификатора могут вносить существенную погрешность в определение количества стартовых молекул кДНК или ДНК. По нашему опыту молекулярный баркод длиною в 12 случайных нуклеотидов после амплификации и секвенирования 10^4 раз обычно производит до 10–20 искусственных подвариантов баркода, и в совокупности они могут быть представлены 100–200 прочтениями секвенирования. В итоге после группировки прочтений по уникальным идентификаторам мы получаем от 11 до 21 стартовой молекулы кДНК, в то время как на самом деле она была только одна. Ещё большее число искусственных подвариантов молекулярных баркодов было обнаружено в модельной системе с последовательностями уникальных идентификаторов в 16 нуклеотидов [28]. Однако существует простой подход, позволяющий достаточно эффективно отфильтровывать такие искусственные подварианты молекулярных баркодов. Как правило, при биоинформатическом анализе заметный процент данных составляют баркоды, прочитанные в ходе секвенирования один или два раза и по последовательности отличающиеся от более представленных «родительских» вариантов всего на 1 нуклеотид. Такая ситуация типична и воспроизводилась в серии проанализированных нами данных для различных экспериментов, в которых стартовый образец прошел 27–35 раундов амплификации. Подавляющее большинство вариантов баркодов с низким покрытием секвенирования представляет собой искусственное разнообразие, возникающее из-за ошибок ПЦР на поздних раундах амплификации. Устранять такое искусственное разнообразие библиотек можно с помощью фильтрации баркодов по количеству полученных прочтений. То есть в дальнейшем анализе используются последовательности уникальных молекулярных баркодов, секвенированных не менее определенного числа раз. Величина оптимального порога, выраженного в минимальном количестве прочтений на каждый уникальный баркод, может варьироваться в зависимости от размера стартовой библиотеки и достигнутого покрытия секвенирования. Также возможно проводить frequency-based фильтрацию по принципу наличия более представленных «родительских» вариантов молекулярных баркодов.

Другим важным моментом является естественное повторение случайно синтезированных уникальных идентификаторов («коллизий») [12, 28–29]. Так, при использовании в качестве молекулярных баркодов последовательностей длиною в 12 случайных нуклеотидов их теоретическое разнообразие составляет порядка $1{,}7\times10^7$ уникальных вариантов. Стоит отметить, что реальное наблюдаемое разнообразие всегда несколько меньше теоретического: по нашим оценкам, порядка $1{,}4\times10^7$ для последовательности из 12 случайных нуклеотидов. Такое снижение происходит вследствие неравномерного синтеза праймеров и, предположительно, наличия некоторых нуклеотидных предпочтений в ходе амплификации. Вероятность того, что в образце со случайно синтезированными, например, 10 000 молекулярными баркодами, найдутся два идентичных или различающихся лишь на 1 нуклеотид варианта достаточно мала. Однако очевидно, что при глубоком секвенировании, когда стартовое число молекул достигает нескольких миллионов, коллизии 12-тинуклеотидных баркодов случаются гораздо чаще. В таком случае бывает сложно отличить естественные коллизии от искусственного разнообразия молекулярных баркодов, вызванного накоплением в них ошибок ПЦР и секвенирования. Например, для 1 миллиона анализируемых молекул можно

ожидать более 30 000 пар случайно синтезированных 12-тинуклеотидных вариантов молекулярных баркодов с идентичной последовательностью. Таким образом, при анализе данных глубокого секвенирования с использованием молекулярного баркодирования необходимо учитывать и теоретическое, и реальное разнообразие используемых уникальных идентификаторов. Предпочтительно использовать большее разнообразие вариантов (то есть большую длину случайно синтезируемой последовательности в составе используемого адаптера) при глубоком анализе больших библиотек.

МАССИРОВАНОЕ СЕКВЕНИРОВАНИЕ В ИССЛЕДОВАНИЯХ АДАПТИВНОГО ИММУНИТЕТА

В иммунологии молекулярное баркодирование открывает новые возможности для анализа сложных репертуаров вариабельных фрагментов антител и Т-клеточных рецепторов, разнообразие которых в организме каждого человека может составлять сотни миллионов вариантов.

Разнообразие Т-клеточных рецепторов и антител внутри одного образца также может быть чрезвычайно велико, а высокогомологичные варианты могут присутствовать в различных пропорциях и быть практически неотличимыми от накопленных ошибок при HTS-анализе. Молекулярное баркодирование делает возможным точный анализ и сопоставление информации о репертуарах антител и Т-клеточных рецепторов для различных субпопуляций лимфоцитов, органов и тканей здоровых и больных индивидуумов и позволяет надежно отслеживать изменения в разнообразии иммунных репертуаров и судьбу клональных популяций лимфоцитов с течением времени или после проведенной терапии. С использованием метода молекулярного баркодирования нами был проведен нормированный сравнительный анализ разнообразия репертуаров бета-цепей Т-клеточных рецепторов периферической крови здоровых людей различного возраста [9].

Было показано, что наблюдаемое разнообразие Т-клеточных рецепторов практически линейно убывает в течение жизни. Одновременно с возрастом происходит активное заполнение гомеостатического пространства высокопредставленными клонами Т-клеток при значительном падении процентного содержания единично представленных клонотипов. Учитывая относительно стабильное общее количество Т-лимфоцитов (оно лишь незначительно снижается с возрастом) разрастание субпопуляций эффекторных Т-лимфоцитов и Т-лимфоцитов памяти неизбежно приводит к снижению относительного количества наивных Т-лимфоцитов. Это приводит к снижению наблюдаемого и экстраполируемого разнообразия вариантов Т-клеточных рецепторов и вероятности инициации эффективного иммунного ответа против новых патогенов и онкологических заболеваний. Молекулярное баркодирование также повышает надежность анализа репертуаров рецепторов иммунных клеток для малых чисел исследуемых лимфоцитов. Это касается ситуаций, когда работа проводится с малыми популяциями сортированных или культивируемых лимфоцитов либо с образцами ткани, содержащими небольшое количество лимфоцитов.

При глубоком секвенировании библиотек с малым числом стартовых молекул искусственное разнообразие накопленных ошибок амплификации начинает существенно превалировать над реальным, а численные значения представленности вариантов последовательностей искажаются вследствие стохастической природы ПЦР. Использование молекулярного баркодирования эффективно устраняет как ложные подварианты последовательностей, так и накопленные количественные искажения, при этом полностью сохраняя нативную информацию о разнообразии репертуаров иммунных клеток в образце [12].

Исследование индивидуальных репертуаров антител и Т-клеточных рецепторов методом HTS в последнее время все чаще применяется в медицине: при количественной оценке минимальной остаточной болезни при терапии лимфопролиферативных заболеваний [30–34]; при отслеживании результатов аутологичной трансплантации гемопоэтических клеток крови [35–37]; при количественном определении наличия и клональности проникающих в солидную опухоль лимфоцитов (Tumor infiltrating lymphocytes, TILs) [38]; при отслеживании изменений периферического Т-клеточного репертуара, вызванных иммунотерапией рака [39]; при поиске противораковых Т-клеточных рецепторов среди проникающих в солидную опухоль лимфоцитов [40–42] и др.

ЗАКЛЮЧЕНИЕ

Технология молекулярного баркодирования представляет собой мощный инструмент для нормированного безошибочного HTS-анализа.

При достаточном уровне покрытия секвенирования использование уникальных молекулярных идентификаторов позволяет применять методы эффективной коррекции ошибок ПЦР и секвенирования, сохраняя при этом естественное разнообразие исследуемого образца. Последнее исключительно важно для анализа библиотек, включающих высоко гомологичные варианты последовательностей. Применение молекулярного баркодирования также делает возможным точный контроль глубины анализа данных высокопроизводительного секвенирования — в единицах анализируемых молекул исходного образца, а также проведение корректного сравнения данных по секвенированию сложных библиотек генов в условиях различного количества клеток/молекул на старте, различного качества ДНК, РНК или кДНК и различной глубины секвенирования.

В биологии молекулярное баркодирование активно используется для решения разнообразных задач, таких как анализ геномов, транскриптомов [25, 26, 43] и микробиомов [6], исследование точности работы полимераз [25], оценка уровня ошибок в ходе транскрипции [44], синтез праймеров [25] или собственно секвенирование [29], а также анализ разнообразия репертуаров рецепторов иммунных клеток [8–11].

В медицине распространение молекулярного бакродирования в HTS-анализе позволит достичь принципиально нового уровня надежности и чувствительности, что откроет новые возможности для ранней онкодиагностики и пренатальной диагностики по ДНК плазмы крови, для анализа гетерогенности и изменчивости опухолей, инфекционных агентов и микробиомов, достоверного анализа репертуаров рецепторов иммунных клеток, который все чаще находит применение в медицинской практике.

Работа поддержана Российским научным фондом, грант №14-35-00105.

Литература

1. Diehl F, Schmidt K, Durkee KH, Moore KJ, Goodman SN, Shuber AP, et al. Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. Gastroenterology. 2008; 135: 489–98.
2. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA. 2008; 105: 16266–71.
3. Burrell RA, Swanton C. The evolution of the unstable cancer genome. Curr Opin Genet Dev 2014; 24: 6–7.
4. Colman RE, Schupp JM, Hicks ND, Smith DE, Buchhagen JL, Valafar F, et al. Detection of Low-Level Mixed-Population Drug Resistance in Mycobacterium tuberculosis Using High Fidelity Amplicon Sequencing. PLoS One. 2015; 10: e0126626.
5. Van Laethem K, Theys K, Vandamme AM. HIV-1 genotypic drug resistance testing: digging deep, reaching wide? Curr Opin Virol. 2015; 14: 16–23.
6. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science. 2013; 341: 1237439.
7. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013; 14: 827–39.
8. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci USA. 2013; 110: 13463–8.
9. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J Immunol. 2014; 192: 2689–98.
10. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014; 11: 653–5.
11. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. Sci Rep. 2014; 4: 6778.
12. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova OV, Sharonov GV, Staroverov DB, et al. Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. J Immunol. 2015; 194: 6155–63.
13. Tie J, Kinde I, Wang Y, Wong HL, Roebert J, Christie M, et al. Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. Ann Oncol. 2015; 26: 1715–22.
14. Diehl F, Li M, Dressman D, He Y, Shen D, Szabo S, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci USA. 2005; 102: 16368–73.
15. Fleischhacker M, Schmidt B. Circulating nucleic acids (CNAs) and cancer--a survey. Biochim Biophys Acta. 2007; 1775: 181–232.
16. Chen Z, Feng J, Buzin CH, Liu Q, Weiss L, Kernstine K, et al. Analysis of cancer mutation signatures in blood by a novel ultra-sensitive assay: monitoring of therapy or recurrence in non-metastatic breast cancer. PLoS One. 2009; 4: e7220.
17. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014; 6 (224): 224ra224.
18. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014; 20: 548–54.
19. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics. 2011; 12: 106.
20. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. Eur J Immunol. 2012; 42: 3073–83.
21. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, et al. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. PLoS One. 2013; 8: e70388.
22. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22: 568–76.
23. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013; 31: 213–19.
24. Harismendy O, Schwab RB, Bao L, Olson J, Rozenzhak S, Kotsopoulus SK, et al. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol. 2011; 12: R124.
25. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA. 2011; 108: 9530–35.
26. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012; 9: 72–4.
27. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013; 14: R51.
28. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 2011; 39: e81.
29. Deakin CT, Deakin JJ, Ginn SL, Young P, Humphreys D, Suter CM, et al. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. Nucleic Acids Res. 2014; 42: e129.
30. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. Blood. 2012; 120: 5173–80.
31. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med. 2012; 4: 134ra163.
32. Martinez-Lopez J, Lahuerta JJ, Pepin F, Gonzalez M, Barrio S, Ayala R, et al. Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma. Blood. 2014; 123: 3073–79.
33. Ladetto M, Bruggemann M, Monitillo L, Ferrero S, Pepin F, Drandi D, et al. Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. Leukemia. 2014; 28: 1299–1307.
34. Logan AC, Vashi N, Faham M, Carlton V, Kong K, Buno I, et al. Immunoglobulin and T Cell Receptor Gene High-Throughput Sequencing Quantifies Minimal Residual Disease in Acute Lymphoblastic Leukemia and Predicts Post-Transplantation Relapse and Survival. Biol Blood Marrow Transplant. 2014; 20 (9): 1307–13.
35. Mamedov IZ, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, Zvyagin IV, et al. Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. EMBO Mol Med. 2011; 3 (4): 201–7.
36. Britanova OV, Bochkova AG, Staroverov DB, Fedorenko DA, Bolotin DA, Mamedov IZ, et al. First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy. Bone Marrow Transplant. 2012; 47: 1479–81.
37. Muraro PA, Robins H, Malhotra S, Howell M, Phippard D, Desmarais C, et al. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. J Clin Invest. 2014; 124: 1168–72.
38. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, et al. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. J Pathol. 2013; 231: 433–40.
39. Cha E, Klinger M, Hou Y, Cummings C, Ribas A, Faham M, et al. Improved Survival with T Cell Clonotype Stability After Anti-CTLA-4 Treatment in Cancer Patients. Sci Transl Med. 2014; 6: 238ra270.
40. Dudley ME, Wunderlich JR, Robbins PF, Yang JC, Hwu P, Schwartzentruber DJ, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. Science. 2002; 298: 850–4.
41. Linnemann C, Mezzadra R, Schumacher TN. TCR repertoires of intratumoral T-cell subsets. Immunol Rev. 2014; 257: 72-82.
42. Gros A, Robbins PF, Yao X, Li YF, Turcotte S, Tran E, et al. PD-1 identifies the patient-specific CD8(+) tumor-reactive repertoire infiltrating human tumors. J Clin Invest. 2014; 124: 2246–59.
43. Grun D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014; 11: 637–40.
44. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci USA. 2013; 110: 18584–9.

References

1. Diehl F, Schmidt K, Durkee KH, Moore KJ, Goodman SN, Shuber AP, et al. Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. Gastroenterology. 2008; 135: 489–98.
2. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA. 2008; 105: 16266–71.
3. Burrell RA, Swanton C. The evolution of the unstable cancer genome. Curr Opin Genet Dev 2014; 24: 61–7.
4. Colman RE, Schupp JM, Hicks ND, Smith DE, Buchhagen JL, Valafar F, et al. Detection of Low-Level Mixed-Population Drug Resistance in Mycobacterium tuberculosis Using High Fidelity Amplicon Sequencing. PLoS One. 2015; 10: e0126626.
5. Van Laethem K, Theys K, Vandamme AM. HIV-1 genotypic drug resistance testing: digging deep, reaching wide? Curr Opin Virol. 2015; 14: 16–23.
6. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science. 2013; 341: 1237439.
7. Barrick JE, Lenski RE. Genome dynamics during experimental evolution. Nat Rev Genet. 2013; 14: 827–39.
8. Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. Proc Natl Acad Sci USA. 2013; 110: 13463–8.
9. Britanova OV, Putintseva EV, Shugay M, Merzlyak EM, Turchaninova MA, Staroverov DB, et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. J Immunol. 2014; 192: 2689–98.
10. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014; 11: 653–5.
11. He L, Sok D, Azadnia P, Hsueh J, Landais E, Simek M, et al. Toward a more accurate view of human B-cell repertoire by next-generation sequencing, unbiased repertoire capture and single-molecule barcoding. Sci Rep. 2014; 4: 6778.
12. Egorov ES, Merzlyak EM, Shelenkov AA, Britanova OV, Sharonov GV, Staroverov DB, et al. Quantitative Profiling of Immune Repertoires for Minor Lymphocyte Counts Using Unique Molecular Identifiers. J Immunol. 2015; 194: 6155–63.
13. Tie J, Kinde I, Wang Y, Wong HL, Roebert J, Christie M, et al. Circulating tumor DNA as an early marker of therapeutic response in patients with metastatic colorectal cancer. Ann Oncol. 2015; 26: 1715–22.
14. Diehl F, Li M, Dressman D, He Y, Shen D, Szabo S, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. Proc Natl Acad Sci USA. 2005; 102: 16368–73.
15. Fleischhacker M, Schmidt B. Circulating nucleic acids (CNAs) and cancer--a survey. Biochim Biophys Acta. 2007; 1775: 181–232.
16. Chen Z, Feng J, Buzin CH, Liu Q, Weiss L, Kernstine K, et al. Analysis of cancer mutation signatures in blood by a novel ultra-sensitive assay: monitoring of therapy or recurrence in non-metastatic breast cancer. PLoS One. 2009; 4: e7220.
17. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014; 6 (224): 224ra224.
18. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, et al. An ultrasensitive method for quantitating circulating tumor DNA with broad patient coverage. Nat Med. 2014; 20: 548–54.
19. Nguyen P, Ma J, Pei D, Obert C, Cheng C, Geiger TL. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. BMC Genomics. 2011; 12: 106.
20. Bolotin DA, Mamedov IZ, Britanova OV, Zvyagin IV, Shagin D, Ustyugova SV, et al. Next generation sequencing for TCR repertoire profiling: platform-specific features and correction algorithms. Eur J Immunol. 2012; 42: 3073–83.
21. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, Andersson B, et al. PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. PLoS One. 2013; 8: e70388.
22. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012; 22: 568–76.
23. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013; 31: 213–19.
24. Harismendy O, Schwab RB, Bao L, Olson J, Rozenzhak S, Kotsopoulus SK, et al. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing. Genome Biol. 2011; 12: R124.
25. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci USA. 2011; 108: 9530–35.
26. Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods. 2012; 9: 72–4.
27. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013; 14: R51.
28. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP. A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res. 2011; 39: e81.
29. Deakin CT, Deakin JJ, Ginn SL, Young P, Humphreys D, Suter CM, et al. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. Nucleic Acids Res. 2014; 42: e129.
30. Faham M, Zheng J, Moorhead M, Carlton VE, Stow P, Coustan-Smith E, et al. Deep-sequencing approach for minimal residual disease detection in acute lymphoblastic leukemia. Blood. 2012; 120: 5173–80.
31. Wu D, Sherwood A, Fromm JR, Winter SS, Dunsmore KP, Loh ML, et al. High-throughput sequencing detects minimal residual disease in acute T lymphoblastic leukemia. Sci Transl Med. 2012; 4: 134ra163.
32. Martinez-Lopez J, Lahuerta JJ, Pepin F, Gonzalez M, Barrio S, Ayala R, et al. Prognostic value of deep sequencing method for minimal residual disease detection in multiple myeloma. Blood. 2014; 123: 3073–79.
33. Ladetto M, Bruggemann M, Monitillo L, Ferrero S, Pepin F, Drandi D, et al. Next-generation sequencing and real-time quantitative PCR for minimal residual disease detection in B-cell disorders. Leukemia. 2014; 28: 1299–1307.
34. Logan AC, Vashi N, Faham M, Carlton V, Kong K, Buno I, et al. Immunoglobulin and T Cell Receptor Gene High-Throughput Sequencing Quantifies Minimal Residual Disease in Acute Lymphoblastic Leukemia and Predicts Post-Transplantation Relapse and Survival. Biol Blood Marrow Transplant. 2014; 20 (9): 1307–13.
35. Mamedov IZ, Britanova OV, Bolotin DA, Chkalina AV, Staroverov DB, Zvyagin IV, et al. Quantitative tracking of T cell clones after haematopoietic stem cell transplantation. EMBO Mol Med. 2011; 3 (4): 201–7.
36. Britanova OV, Bochkova AG, Staroverov DB, Fedorenko DA, Bolotin DA, Mamedov IZ, et al. First autologous hematopoietic SCT for ankylosing spondylitis: a case report and clues to understanding the therapy. Bone Marrow Transplant. 2012; 47: 1479–81.
37. Muraro PA, Robins H, Malhotra S, Howell M, Phippard D, Desmarais C, et al. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. J Clin Invest. 2014; 124: 1168–72.
38. Emerson RO, Sherwood AM, Rieder MJ, Guenthoer J, Williamson DW, Carlson CS, et al. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. J Pathol. 2013; 231: 433–40.
39. Cha E, Klinger M, Hou Y, Cummings C, Ribas A, Faham M, et al. Improved Survival with T Cell Clonotype Stability After Anti-CTLA-4 Treatment in Cancer Patients. Sci Transl Med. 2014; 6: 238ra270.
40. Dudley ME, Wunderlich JR, Robbins PF, Yang JC, Hwu P, Schwartzentruber DJ, et al. Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. Science. 2002; 298: 850–4.
41. Linnemann C, Mezzadra R, Schumacher TN. TCR repertoires of intratumoral T-cell subsets. Immunol Rev. 2014; 257: 72–82.
42. Gros A, Robbins PF, Yao X, Li YF, Turcotte S, Tran E, et al. PD-1 identifies the patient-specific CD8(+) tumor-reactive repertoire infiltrating human tumors. J Clin Invest. 2014; 124: 2246–59.
43. Grun D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014; 11: 637–40.
44. Gout JF, Thomas WK, Smith Z, Okamoto K, Lynch M. Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci USA. 2013; 110: 18584–9.