DOI: 10.25557/2073-7998.2018.05.19-22

Исключение неверно картированных прочтений при таргетном высокопроизводительном секвенировании ДНК с использованием технологии Ion AmpliSeq

Карандашева К.О.[1], Аношкин К.И.[1,2], Володин И.В.[1], Кузнецова Е.Б.[1,3], Залетаев Д.В.[1,3], Стрельников В.В.[1,2], Танас А.С.[1,2]

[1] ФГБНУ «Медико-генетический научный центр»,
Москва, 115478, ул. Москворечье, д.1, e-mail: christinavader@gmail.com
[2] ФГБОУ ВО «Российский национальный исследовательский медицинский университет им. Пирогова» Министерства здравоохранения Российской Федерации,
Москва, 117997, ул. Островитянова, д. 1; e-mail: tanas80@gmail.com
[3] ФГАОУ ВО Первый Московский государственный медицинский университет имени И.М. Сеченова Министерства здравоохранения Российской Федерации (Сеченовский Университет),
Москва, 119991, ул. Трубецкая, д. 8, стр. 2; e-mail: zalnem@mail.ru

Актуальность. Использование высокопроизводительного параллельного секвенирования (ВПС) сопряжено с ошибками: не все генетические варианты, выявляемые с использованием ВПС, являются истинными и подтверждаются альтернативными методами. Частичный вклад в выявление ложноположительных вариантов вносят недостоверно картированные прочтения. Мы полагаем, что при проведении таргетного секвенирования с использованием технологии Ion AmpliSeq (таргетное обогащение мультиплексной ПЦР) задача исключения недостоверно картированных прочтений может быть решена алгоритмически с использованием информации о геномных координатах таргетных регионов и длине праймеров, использованных при амплификации фрагментов библиотек. Эта информация позволяет оценить правильность картирования каждого прочтения и исключить из дальнейшего анализа прочтения, не соответствующие дизайну эксперимента. **Цель.** Разработать алгоритм минимизации вклада ошибок картирования в спектр генетических вариантов, выявляемых при ВПС. **Материалы и методы**. С использованием информации о геномных координатах таргетных регионов и праймерах проанализировано 30 файлов формата BAM, полученных по результатам ВПС с использованием технологии таргетного секвенирования Ion AmpliSeq и коммерческих панелей праймеров Ion AmpliSeq Comprehensive Cancer Panel (15,992 таргетных региона) и Ion AmpliSeq Inherited Disease Panel (10,309 таргетных регионов). Алгоритм исключения неверно картированных прочтений реализовали средствами языка программирования Python. **Результаты.** Проведено сравнение исходного набора прочтений и набора, полученного после исключения неверно картированных прочтений. Определены три группы генетических вариантов: (1) выявляемые в обоих наборах прочтений — 6072 варианта; (2) выявляемые исключительно в исходном наборе — 127 (варианты, наблюдаемые в большинстве образцов, могут быть интерпретированы как результат систематических ошибок выравнивания); (3) выявляемые только в наборе прочтений, полученном удалением неверно картированных, — 63 истинных, ранее скрытых варианта. **Выводы.** Использование дополнительной информации об ожидаемом начале и окончании прочтения при таргетном исследовании позволяет (1) уменьшить количество выявляемых вариантов, обусловленных ошибочно картированными прочтениями, (2) детектировать ранее не обнаруживаемые по причине ложной малой аллельной частоты, (3) получить более достоверные значения аллельных частот выявляемых вариантов. Применение разработанного нами алгоритма исключения неверного картирования прочтений фрагментов ДНК повышает качество интерпретации результатов ВПС, что особенно важно при проведении ДНК-диагностики.

Ключевые слова: NGS, таргетное секвенирование, Ion AmpliSeq, ошибки картирования.

Авторы заявляют об отсутствии конфликта интересов.

Elimination of incorrectly mapped reads from the results of Ion AmpliSeq targeted NGS

Karandasheva K.O.[1], Tanas A.S.[1,2], Anoshkin K.I.[1,2], Volodin I.V.[1], Kuznetsova E.B.[1,3], Zaletayev D.V.[1,3], Strelnikov V.V.[1,2]

[1] Research Centre for Medical Genetics, Moscow, Russian Federation, 115478, Moskvorechie st.1, e-mail: christinavader@gmail.com
[2] Pirogov Russian National Research Medical University, Moscow, Russian Federation, 117997, Ostrovityanova St, 1, e-mail: tanas80@gmail.com
[3] I.M. Sechenov First Moscow State Medical University, Moscow, Russian Federation, 119991, Trubetskaya st. 8, e-mail: zalnem@mail.ru

Background. The use of high-throughput parallel sequencing (NGS) is fraught with errors: not all of the genetic variants that are detected by NGS are true and are confirmed by alternative methods. Incorrectly mapped reads contribute to the appearance of false-positive variants. We believe that for targeted sequencing using Ion AmpliSeq technology, the task of excluding erroneously mapped reads can be solved algorithmically. Additional information on the genomic coordinates of target regions and primers used in amplification allows us to evaluate the validity of the mapping of each reading and to exclude readings from the further analysis that do

not correspond to the design of the experiment. **Objective**. To develop an algorithm to minimize the contribution of mapping errors to the spectrum of genetic variants detected by NGS. **Material and methods.** Using the information on the genomic coordinates of target regions and primers, we have analyzed 30 BAM files obtained by Ion AmpliSeq (targeted multiplex PCR) NGS with commercial Ion AmpliSeq Comprehensive Cancer Panel (15,992 target regions) and Ion AmpliSeq Inherited Disease Panel (10,309 target regions). The algorithm for excluding incorrectly mapped reads was implemented using the Python programming language. **Result.** We have performed comparison of the initial set of reads and the set obtained after excluding incorrectly mapped reads. This comparison revealed three groups of genetic variants: (1) detectable in both sets of reads, 6072 variants; (2) detectable exclusively in the original set, 127 (the predominant part of these variants is present in most samples and can be interpreted as a result of systematic alignment errors); (3) detectable in the set generated by exclusion of erroneously mapped reads only, 63 (true positive, previously masked variants). **Conclusion.** The use of additional information on the expected start and end of the read in the targeted study allows to (1) reduce the number of false-positive genetic variants detected due to misleading reads, (2) detect new ones that were not previously detected due to a seemingly low allele frequency, (3) obtain more reliable values of allelic frequencies of the identified variants. The use of our algorithm to exclude the incorrect mapping of the of DNA fragment reads increases the quality of interpretation of the NGS results, which is especially important for DNA diagnostics.

Key words: NGS, target sequencing, Ion AmpliSeq, mapping errors.

Введение

Высокопроизводительное параллельное секвенирование (ВПС) — современный метод выбора при проведении ДНК-диагностики в случаях генетической гетерогенности фенотипа и/или значительной протяженности генов, мутации в которых ассоциированы с заболеванием [1].

В настоящее время в диагностике наследственной патологии методом ВПС применяют полногеномное (полноэкзомное) секвенирование и таргетную диагностику [2].

Полногеномное секвенирование ДНК позволяет выявлять предполагаемую причину заболевания даже в случаях высокой генетической гетерогенности или при неоднозначной клинической картине у пациента, не позволяющей врачу по клиническим критериям заранее установить гены, в которых необходимо вести поиск [1, 3]. Полногеномное исследование является оптимальным методом диагностики редких и неочевидных фенотипов [4], однако наиболее дорогостоящим методом, все еще недоступным в рамках рутинной диагностики.

Таргеное (целевое) секвенирование ДНК обеспечивает поиск мутаций в ограниченном наборе генов [5]. Такой подход широко используется при подозрении наличия у пациента конкретной наследственной патологии, для которой известны гены, изменения в которых являются молекулярной причиной заболевания. Концентрация на определенных участках генома значительно снижает временные и денежные затраты — процесс диагностики становится менее ресурсоемким, а анализ полученного результата занимает меньшее количество времени вследствие меньшего объема полученных данных, чем при полногеномном исследовании.

Использование ВПС сопряжено с ошибками: не все выявляемые генетические варианты являются истинными и подтверждаются альтернативными методами [6].

Причинами наблюдения ложных генетических вариантов могут являться несоблюдение протокола исследования, контаминация исследуемого образца чужеродной ДНК, ошибки секвенирования (случайные или обусловленные особенностями используемой технологической платформы), а также ошибки картирования коротких прочтений на референсный геном.

В связи с этим при использовании методов ВПС в диагностических целях обязательным этапом является визуальный анализ данных, направленный на исключение артефактов секвенирования и ошибок картирования прочтений. На этом этапе перед исследователем возникает необходимость оценки достоверности обнаруживаемых генетических вариантов. Визуальный анализ результатов ВПС — трудоёмкая работа, чреватая ошибками вследствие субъективности оценок оператора.

Мы полагаем, что при проведении таргетного высокопроизводительного секвенирования с использованием технологии Ion AmpliSeq задача исключения недостоверно картированных прочтений может быть решена алгоритмически. Дополнительная информация о геномных координатах таргетных регионов и длине праймеров, использованных при амплификации, позволяет оценить валидность картирования каждого прочтения и исключить из дальнейшего анализа прочтения, не соответствующие дизайну проведенного эксперимента.

Целью настоящего исследования было разработать алгоритм минимизации вклада ошибок картирования в спектр генетических вариантов, выявляемых при ВПС.

Материалы и методы

Для разработки и валидации алгоритма использовали 30 файлов формата BAM, полученных по результатам ВПС на технологических платформах Ion PGM и Ion S5 (Life Technologies, США) с использованием технологии таргетного секвенирования Ion AmpliSeq (таргетное обогащение методом мультиплексной ПЦР) и коммерческих панелей праймеров Ion AmpliSeq Comprehensive Cancer Panel (15,992 таргетных региона) и Ion AmpliSeq Inherited Disease Panel (10,309 таргетных регионов).

Алгоритм исключения неверно картированных прочтений реализовали средствами языка программирова-

ния Python. В основу алгоритма легло утверждение, что после амплификации целевых районов генома все таргетные нуклеотидные последовательности, полученные в ходе секвенирования, должны иметь геномные координаты, предопределенные дизайном эксперимента. В качестве входных данных алгоритма использовали файл формата BAM, содержащий прочтения, картированные на референсный геном, и текстовый файл, содержащий информацию о координатах таргетных регионов и длине праймеров, использованных при амплификации.

Алгоритм основан на последовательном переборе всех прочтений из файла формата BAM. Если длина прочтения превышает 0,3 от длины целевого района (границу наименьшей допустимой доли покрытия определили по точке минимума второй производной функции на логарифмическом графике распределения частоты доли покрытия прочтением таргетного региона — рис. 1), а координаты концов корректны (соответствуют конечным координатам таргетных регионов с учетом возможности присутствия неотрезанного праймера, соответствующего данному таргетному региону), то такое прочтение является допустимым.

Результатом работы алгоритма является новый файл формата BAM, содержащий только прочтения ВПС, удовлетворяющие выдвинутым требованиям.

Для оценки эффективности алгоритма произвели поиск генетических вариантов в исходном и сгенерированном файлах и оценили их изменения, сравнив аллельные частоты выявленных вариантов до и после устранения неверно картированных прочтений.

Результаты

В исходном наборе прочтений, полученном в виде файла формата BAM, с использованием программного обеспечения Torrent Variant Caller (TVC) выявлено 6072 варианта. В наборе прочтений, сгенерированном нами с использованием разработанного алгоритма исключения неверного картирования, TVC выявил 6007 вариантов. При этом 5817 вариантов оказались общими для обоих наборов прочтений, 127 вариантов выявлены исключительно в исходном наборе (варианты, наблюдаемые в большинстве образцов, могут быть интерпретированы как результат систематических ошибок выравнивания), и 63 — исключительно в сгенерированном наборе прочтений (истинные, ранее скрытые генетические варианты).

Нами проведено сравнение аллельных частот вариантов до и после исключения неверно картированных прочтений. Каждый генетический вариант был охарактеризован с использованием двух параметров: аллельная частота в исходном файле и аллельная частота в сгенерированном файле (рис. 2).

Среди наблюдаемых в обоих наборах прочтений, обнаружены генетические варианты, аллельная частота которых значительно изменилась. Например, на рис. 3 представлен вариант chr1:g.33,132,314 G>A, выявляемый в исходном наборе прочтений как гетерозиготный с аллельной частотой, равной 0,59. Исключение недостоверно картированных прочтений изменило его наблюдаемую аллельную частоту до 0,96, и следовательно, данный вариант является истинно гомозиготным.

Искажение недостоверно картированными прочтениями истинной аллельной частоты выявляемого генетического варианта имеет значение при проведении ДНК-диагностики, особенно при подозрении на наличие у пациента рецессивного заболевания.

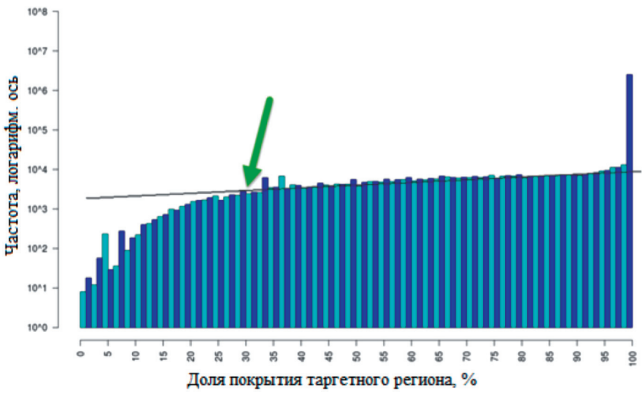

Рис. 1. Распределение доли покрытия прочтением таргетных регионов (выборка из 10 образцов).

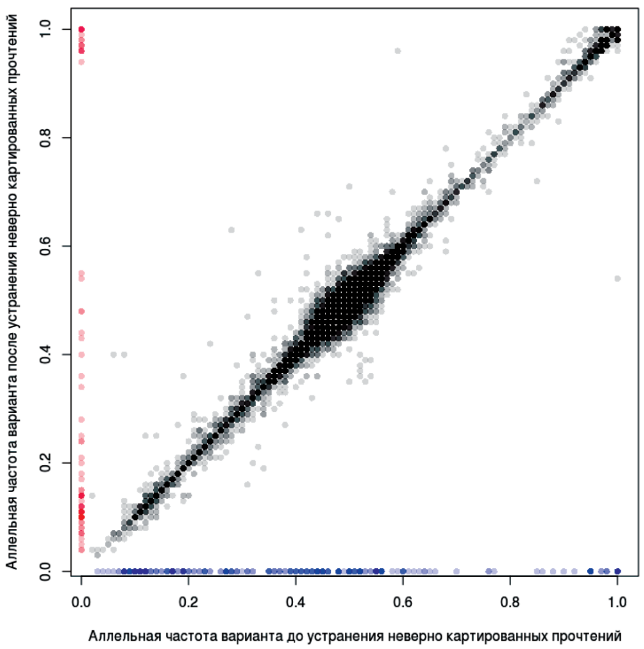

Рис. 2. Аллельные частоты выявляемых коротких вариантов нуклеотидных последовательностей до и после устранения неверно картированных прочтений. Красным отмечены генетические варианты, выявление которых стало возможным в результате применения алгоритма исключения неверно картированных прочтений. Синим отмечены варианты, выявляемые только в исходном наборе прочтений.

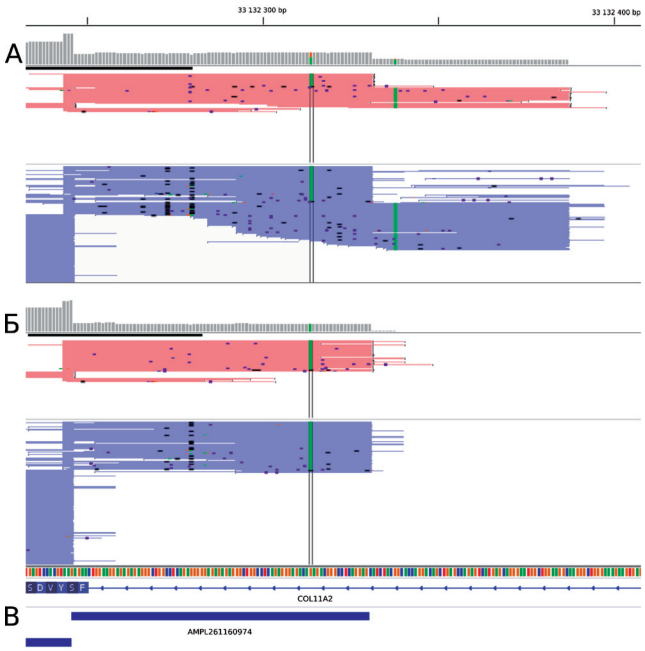

Рис. 3. Наблюдаемые аллельные частоты варианта chr1:g.33,132,314 G>A.
А — до устранения ошибочно картированных прочтений; Б — после устранения ошибочно картированных прочтений; В — разметка ампликонов согласно дизайну эксперимента; целевым является ампликон AMPL261160974.

Варианты, выявляемые только в исходном наборе прочтений, были расценены как следствие неправильного картирования части прочтений на референсный геном. Данные генетические варианты характеризуются:

1) систематичностью в выборке образцов;

2) выявляемостью на прочтениях преимущественно одного направления (strand bias).

Преобладающая часть таких вариантов встречается в большинстве образцов и может быть интерпретирована как результат систематических ошибок выравнивания. Исключение неверно картированных прочтений из дальнейшего анализа приводит к значительному сокращению числа обнаруживаемых генетических вариантов и выявлению скрытых истинных вариантов, ранее не определяемых по причине ложной низкой аллельной частоты.

Заключение

Среди ошибок, возникающих при ВПС, устранению на этапе обработки результатов подлежат ошибки картирования получаемых коротких последовательностей на референсный геном. Использование дополнительной информации об ожидаемом начале и окончании прочтения при таргетном исследовании позволяет:

1) уменьшить количество выявляемых вариантов, обусловленных ошибочно картированными прочтениями;

2) детектировать новые, ранее не обнаруживаемые по причине ложной малой аллельной частоты;

3) получить более достоверные значения аллельных частот выявляемых генетических вариантов.

Применение разработанного нами алгоритма исключения неверного картирования прочтений фрагментов ДНК повышает качество интерпретации результатов ВПС, что особенно важно при проведении ДНК-диагностики.

Список литературы

1. Г.В. Байдакова, Е.Ю. Захарова, И.В. Канивец, Ф.А. Коновалов, В.В. Стрельников, С.И. Куцев. Диагностика врожденных и наследственных болезней у детей: достижения и перспективы развития. Вестник Росздравнадзора. 2016; 3: 27-33.

2. Ребриков Д.В., Коростин Д.О., Шубина Е.С., Ильинский В.В. NGS: высокопроизводительное секвенирование, 2-е издание. М.: БИНОМ. 2015; С. 209.

3. Worthey E.A. Analysis and Annotation of Whole-Genome or Whole-Exome Sequencing-Derived Variants for Clinical Diagnosis. Human Genetics. 2013; 24: 1-24.

4. Yang Y., Muzny D.M., Reid J.G. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. The New England Journal of Medicine. 2013; 69: 501-512.

5. Sikkema-Raddatz B., Johansson L.F, de Boer E.N., Almomani R., Boven L.G. Targeted Next-Generation Sequencing can replace Sanger Sequencing in clinical diagnostics. Human Mutation. 2013; 10: 83-92.

6. Mi-Hyun Park, Hwanseok Rhee, Jung Hoon Park, Soo Kyung Koo. Comprehensive Analysis to Improve the Validation Rate for Single Nucleotide Variants Detected by Next-Generation Sequencing. PLOS ONE. 2014; 9(1): e86664.