ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ДЕТЕКЦИИ ОДНОНУКЛЕОТИДНЫХ ВАРИАЦИЙ В ПУЛИРОВАННЫХ ОБРАЗЦАХ ДНК ПРИ АНАЛИЗЕ ДАННЫХ ШИРОКОМАСШТАБНОГО ПАРАЛЛЕЛЬНОГО СЕКВЕНИРОВАНИЯ

Гаврик О.А., Гусев Ф.Е., Решетов Д.А., Гольцов А.Ю., Андреева Т.В., Тяжелова Т.В., Рогаев Е.И.

ФГБОУ науки Институт общей генетики им. Н.И. Вавилова РАН, Москва; ФГБОУ Научный Центр Психического Здоровья РАМН, Москва

Появление и развитие новых подходов к определению последовательностей с применением методологии масштабного параллельного секвенирования на новейших технологических платформах позволяет сократить время и трудоемкость, которое характерно для стандартных методов поиска мутаций. Учитывая высокую стоимость эксперимента с использованием новейших технологических платформ для секвенирования, при решении ряда задач, является целесообразным предварительное пулирование образцов ДНК, что позволяет существенно снизить затраты на секвенирование. Ниже представлен обзор программного обеспечение для обработки данных параллельного секвенирования пулированных образцов ДНК.

При пулировании образцов ДНК частота уникального аллельного варианта, существующего в индивидуальных образцах, в общем пуле снижается вплоть до порогового значения ошибки секвенирования, поскольку вероятность выявления мутантного аллеля, присутствующего у одного индивида уменьшается с 1/n (где n – плоидность организма) в случае секвенирования одного образца, содержащего этот аллель, до 1/mn при секвенировании пула из m образцов. Кроме того, число прочтений, полученных в результате секвенирования содержащих вариацию, не всегда коррелирует с истинной частотой вариации в пуле [1]. В связи с этим важное значение приобретает выбор программного обеспечения для предсказания однонуклеотидных вариаций в случае анализа данных, полученных при секвенировании пулированных образцов.

Программы Syzygy (http://www.broadinstitute.org/software/syzygy/; [2]), CRISP (https://sites.google.com/site/vibansal/software/crisp; [3]) и SNVer (http://snver.sourceforge.net/cite.html; [4]) работают напрямую с *.bam файлами – продуктами выравнивания коротких прочтений на референсный геном. Эти программы разработаны специально для анализа пулированных данных. В свою очередь, программа VarScan (http://varscan.sourceforge.net/) универсальна и пригодна для поиска вариаций в индивидуальных геномах, так и в пулированных данных [5,6]. Входным файлом для программы VarScan служит не *.bam файл, а полученный из него с помощью программы SAMtools файл в *.pileup формате. Таким образом, анализ данных с применением программы VarScan содержит один дополнительный этап. Общим для применения всех перечисленных выше программ является то, что заранее включать в эксперимент контрольные образцы не требуется.

Для использования программы SPLINTER (http://www.ibridgenetwork.org/wustl/splinter; [7]), напротив, требуется заранее спланировать эксперимент таким образом, чтобы в состав библиотек для секвенирования входили два синтезированных контрольных образца ДНК: положительный контроль (смесь фрагментов ДНК, часть из которых содержит известные однонуклеотидные замены) и отрицательный контроль, не содержащий вариаций. Использование контролей увеличивает общую стоимость эксперимента, однако в этом случае производится калибровка программы под конкретный эксперимент, что позволяет существенно повысить эффективность анализа и вероятность детекции редкой вариации.

Тестирование программ Syzygy, VarScan и CRISP на собственных экспериментальных данных показало значительные различия в чувствительности и специфичности работы программ. Алгоритмы, используемые представленными выше программами, отличны друг от друга, что ведет к значительным расхождениям конечных результатов, полученных при анализе одного набора данных разными программами для детекции однонуклеотидных вариаций. Таким образом, при выборе программы для анализа данных широкомасштабного параллельного секвенирования перед проведением анализа необходимым этапом является проведение предварительного тестирования используемых программ для анализа последовательностей с заранее известными вариациями и их частотой встречаемости в пуле. Это позволит оценить точность получаемых результатов и эффективность предсказания вариаций.

Работа выполнена при финансовой поддержке Министерства науки и образования Российской Федерации (ГК 16.512.11.2083, ГК 02.740.11.0854).

Литература

1. Raineri E et al. 2012. BMC. Bioinformatics, 13:239.
2. Rivas M.A. et al. 2011. Nat Genet, 43: 1066-1073.
3. Bansal V. 2010. Bioinformatics, 26:i318-i324.
4. Wei Z. et al. 2011. Nucleic Acids Res, 39:e132.
5. Koboldt D.C. et al. 2009. Bioinformatics, 25:2283-2285.
6. Koboldt D.C. et al. 2012. Genome Res, 22:568-576.
7. Vallania F. et al. 2012. J Vis Exp, 23: 3943.