Секвенирование нового поколения и области его применения в онкогематологии

И.М. Бархатов[1], А.В. Предеус[2], А.Б. Чухловин[1]

[1]НИИ детской онкологии, гематологии и трансплантологии им. Р.М. Горбачевой ФГБОУ ВО «Первый Санкт-Петербургский государственный медицинский университет им. И.П. Павлова» Минздрава России; Россия, 197022 Санкт-Петербург, ул. Льва Толстого, 6–8;

[2]Институт биоинформатики ФГБУ «Санкт-Петербургский национальный исследовательский Академический университет Российской академии наук»; Россия, 194100 Санкт-Петербург, ул. Кантемировская, 2

***Контакты:** Ильдар Мунерович Бархатов i.barkhatov@gmail.com*

Обзорная статья касается основных принципов и технологий генного секвенирования нового поколения (next-generation sequencing, NGS), а также его приложений к оценке генных мутаций в лейкозных клетках. Обсуждается ряд современных работ, касающихся применения NGS в исследованиях генетической гетерогенности при миелопролиферативных заболеваниях, для выявления генов высокого клинического риска, в том числе мутаций, связанных с резистентностью к терапии, анализа эпигеномных нарушений при лейкозах, а также молекулярных аспектов эволюции злокачественных клонов. Отдельный раздел посвящен основным проблемам, связанным с биоинформатикой и корректным анализом больших компьютерных баз результатов NGS-исследований. Адекватный выбор программных продуктов очень важен для адекватной обработки и интерпретации данных NGS.

***Ключевые слова:** секвенирование нового поколения, технологии, лейкозы, генные мутации, биоинформатика*

DOI: 10.17650/1818-8346-2016-11-4-56-63

Next-generation gene sequencing and its applications in oncohematology

I.M. Barkhatov[1], A.V. Predeus[2], A.B. Chukhlovin[1]

[1]R.M. Gorbacheva Memorial Research Institute of Children Oncology, Hematology and Transplantation, I.P. Pavlov First Saint Petersburg State Medical University, Ministry of Health of Russia; 6–8 L'va Tolstogo St., Saint Petersburg 197022, Russia;

[2]Bioinformatics Institute, Saint Petersburg National Research Academic University of the Russian Academy of Sciences; 2 Kantemirovskaya St., Saint Petersburg 194100, Russia

The review bears on basic principles and technologies of next-generation sequencing (NGS), as well as its applications for detection of gene mutations in leukemic cells. We discuss some novel data concerning NGS approach to studies of genetic heterogeneity in myeloproliferative disorders, detection of high-risk genes, including drug resistance mutations, epigenomic changes associated with leukemias, as well as molecular aspects of clonal evolution. A special section concerns basic problems with bioinformatics and adequate analysis of large digital databases obtained with NGS approach. Optimal choice of appropriate software is of utmost importance for adequate retrieval and interpretation of the NGS data.

***Key words:** next generation sequencing, technologies, leukemia, gene mutations, bioinformatics*

Введение

Общеизвестно, что развитие злокачественных опухолевых заболеваний системы крови связано с возникновением ряда последовательных мутаций генов, ведущих к развитию злокачественного клеточного клона.

Известно также, что в процессе терапии первичная популяция злокачественных клеток может приобретать вторичные мутации, изменяясь под влиянием различных факторов отбора в организме больного (иммунологическая селекция, действие цитостатических препаратов и др.). Эти злокачественные субклоны с сочетаниями генных мутаций могут впоследствии приводить к рецидивам заболевания. Разнообразие генной патологии в организме больного существенно влияет на прогноз и исходы лечения больных с лейкозами и лимфомами. Для оценки клинического риска в каждом конкретном случае требуется детальная характеристика генетических аберраций, а для этого необходимо определение последовательности нуклеотидов целевого гена в злокачественных клетках.

Современная молекулярная диагностика при лейкозах основана на выявлении конкретных генных мутаций, характерных для того или иного вида лейкоза. Основным методом является молекулярно-генетическая диагностика на базе полимеразной цепной реакции (ПЦР) с анализом целевых участков генов, которая проводится в лабораториях крупных онкогематологических клиник. Кроме того, с начала 1990-х годов для задач дешифровки генов, в первую очередь смысловых последовательностей (сиквенсов) молекул матричной РНК (мРНК), применяли главным образом классическое секвенирование по Сэнгеру. Проект «Геном человека», основной задачей которого была сборка референсной последовательности человеческого

генома, потребовал участия 23 лабораторий, 13 лет работы и общих затрат в 3 млрд долларов. Однако этот проект послужил мощным стимулом для развития методов секвенирования. В ходе его выполнения были разработаны новые технологии расшифровки генома – секвенирование следующего поколения, или высокопроизводительное секвенирование (next generation sequencing, NGS).

Настоящий обзор имеет целью ознакомить гематологов с основными принципами методов NGS, показать его эффективность на ряде клинических примеров и обозначить важность адекватного применения современных методов биоинформатики при анализе данных NGS. Первичные результаты применения NGS в онкогематологии уже обсуждались ранее [1].

Принципы осуществления высокопроизводительного секвенирования

NGS представляет собой процесс определения последовательности нуклеотидов в геномной ДНК или в совокупности информационных РНК (транскриптоме) путем амплификации множества коротких участков генов. Это разнообразие генных фрагментов в итоге покрывает всю совокупность целевых генов или, при необходимости, весь геном (или транскриптом как совокупность молекул РНК). На следующем этапе осуществляется одновременное многократное прочтение (секвенирование) этого множества участков генов. Далее проводится компьютерная обработка полученного массива данных и сравнение с «образцовыми» (референсными) последовательностями генов. Это позволяет установить присутствие сотен и тысяч генных мутаций или вариантов в исследуемых клетках за относительно короткий период времени (от нескольких часов до нескольких суток).

Использование подходящего метода NGS, в принципе, позволяет выявлять все известные типы соматических мутаций раковых клеток, в том числе замены нуклеотидов, делеции, инсерции оснований, изменения копийности генов, транслокации генетического материала и т. д. Это особенно важно для анализа мутаций в так называемых горячих точках генома, особо подверженных мутационным изменениям [2].

Особенности различных технологий высокопроизводительного секвенирования

Существует целый ряд технологических платформ для проведения NGS, которые различаются по способу осуществления секвенирования. Основные этапы NGS в целом сходны, а именно:

1) получение множества коротких фрагментов ДНК или молекул мРНК;

2) амплификация этих коротких последовательностей посредством большого числа специфических ДНК-зондов и с применением мультиплексной ПЦР;

3) получение так называемой библиотеки генов (т. е. набора фрагментов ДНК из данного образца) для последующего секвенирования;

4) высокопроизводительное прочтение нуклеотидных последовательностей в этом множестве генных фрагментов.

Прочтение сиквенсов, т. е. определение последовательности нуклеотидов в коротких фрагментах ДНК, может проводиться разными способами: с помощью пиросеквенирования (сейчас применяется редко), гибридизации на микрочастицах, микро-pH-метрии, масс-спектрометрии и т. д.

Ныне применяемые технологии секвенирования формально делят на NGS II и III поколения. Ко II поколению относятся секвенаторы, позволяющие получить большое количество коротких прочтений (25–800 пар оснований), в частности 454 Life Sciences, Illumina, Ion Torrent. К III поколению NGS относятся секвенаторы Pacific Biosciences и Oxford Nanopore, позволяющие прочитывать более длинные участки генов (2000–200 000 пар нуклеотидов). Сейчас наиболее широко используются секвенаторы Illumina. Так, по состоянию на 2014 г. эти модели занимали 70 % рынка секвенаторов и произвели более 90 % всех задокументированных данных. Сильными сторонами Illumina является высокая точность секвенирования (не более 0,1 % ошибок), высочайшая производительность, высокое соотношение производительность/цена, а также возможность получения парных прочтений с 2 концов фрагмента длиной до 750 пар оснований. Последнее позволяет увеличить точность выравнивания фрагментов на референсный геном, а также сильно улучшает точность в сборке геномов и транскриптомов *de novo*.

По мере развития NGS-технологий и конкуренции между разными методическими подходами стоимость NGS в расчете на 1 изучаемый образец неуклонно снижается. Так, в настоящее время силами одной лаборатории можно секвенировать весь геном человека за несколько дней и всего за ~ 1000 долларов, что сделало эту процедуру доступной для многих западных университетов и крупных медицинских учреждений [3].

Варианты применения NGS в медицинских исследованиях многообразны. Условно их можно разделить на следующие группы:

1) определение последовательности всей ДНК (полногеномное секвенирование – whole-genome sequencing, WGS);

2) определение последовательности белоккодирующих участков генома (полноэкзомное секвенирование – whole-exome sequencing, WES);

3) определение последовательности интересующих генов (сюда относятся многие коммерческие решения – от «клинических экзомов» CES, включающих около 5000 медицински значимых генов, до малых таргетных панелей, анализирующих 1–3 гена);

4) секвенирование транскриптома (РНК-секвенирование, RNA-seq), которое часто используется в онкологии для классификации опухолей, нахождения неоантигенов, поиска новых химерных генов и т. д.

Секвенирование экзомов и таргетных панелей уже заняло прочное место в диагностике редких наследственных заболеваний и постепенно набирает популярность и в онкологии. Следует отметить: несмотря на то, что секвенирование полного или клинического экзома значительно дешевле и проще в обработке, с его помощью практически невозможно находить крупные вставки и делеции (инделы), большие геномные перестройки и другие структурные варианты. Для этого используются стратегии полногеномного секвенирования, а также более старые методы, в частности сравнительная геномная гибридизация (array-CGH) и, конечно, изучение кариотипа данных клеток.

Именно знание референсной последовательности ДНК генома человека позволяет отбирать для NGS-анализа комплексы генов со значимыми мутациями и сравнивать их «образцовые» последовательности генома с результатами NGS. К таргетным методам относятся и такие распространенные подходы, как 16S-секвенирование, секвенирование генов, кодирующих молекулы иммунного ответа, в частности HLA-типирование, и многие другие приложения. Сверхглубокое таргетное секвенирование также позволяет анализировать замены и небольшие вставки/делеции в клеточных клонах (при частоте до 1 % в популяции клеток).

Выявление множественных нуклеотидных замен, делеций и иных аберраций в стволовых клетках у пациентов онкогематологического профиля становится все более распространенным. В последние 5–6 лет, с появлением технологических платформ и приборов для NGS, задачи одновременного анализа множества генов стали не только выполнимыми, но и достаточно экономичными.

Области использования секвенирования нового поколения в онкогематологии

Молекулярные варианты лейкозов

В настоящее время выявлено большое число генных и хромосомных мутаций, наблюдаемых при острых лейкозах. Так, на основании этих данных уже возможна детальная молекулярная классификация острого миелобластного лейкоза (ОМЛ) [4]. В ходе мультицентрового исследования авторы изучали последовательность 111 генов у 1540 пациентов после индукционной химиотерапии. С помощью цитогенетического анализа и глубокого секвенирования ДНК выявлены так называемые драйверные мутации, т. е. генные аномалии, связанные с развитием ОМЛ. К ним относили слитные гены (fusion genes), другие онкогенные мутации, а также нарушения плоидности хромосом. Всего обнаружены 5234 различные мутации в 76 генах. У 86 % пациентов найдены как минимум по 2 мутации. Наиболее часто выявлялись аберрации генов *FLT3*, *NPM1*, *DNMT3A*, *NRAS*. Мутации *NPM1* обычно возникали позже, что отражает их роль в дальнейшей эволюции лейкозной трансформации.

При этом сравнительно часто отмечались совместные мутации, например *NPM1* и *NRASG12/13*, а также комбинированные мутации гена *FLT3*. Выделены 11 подгрупп, или классов, ОМЛ по функциональным параметрам вовлеченных генов. Наиболее часто встречались случаи ОМЛ с мутациями *NPM1* (27 %). Вторая по частоте подгруппа включала клинические случаи с мутациями генов, кодирующих сплайсинг (структурные перестройки) молекул РНК, состояние хроматина или транскрипцию РНК. У 13 % пациентов выявлены мутации *TP53* со сложными хромосомными аберрациями, что позволило выделить их в особую патогенетическую группу.

Хотя возможности клонального анализа мутаций были ограниченными, наиболее рано обнаруживали мутации в генах, кодирующих белки эпигенеза (*DNTM3A*, *ASXL1*, *IDH1/2*, *TET2*). Все эти разнородные молекулярно-генетические данные сопоставляли с клинической картиной и исходами заболевания. Выяснилось, например, что неблагоприятный исход чаще наблюдался в группах с мутациями генов, регулирующих функции ядерного хроматина и транскрипцию РНК. Другой группой повышенного риска были пациенты с мутациями гена *TP53* и сложными нарушениями кариотипа. Эти молекулярные критерии, безусловно, подтвердят положения действующей классификации ОМЛ и позволят лучше прогнозировать индивидуальный риск для больных.

С учетом различий в спектре генных мутаций при различных лейкозах за последние годы предложено несколько «типовых» генно-диагностических панелей для индивидуальной оценки отдельных типов лейкозов. Так, одно из исследований касалось миелоидных неоплазий [5]. Наблюдали группу из 46 пациентов среднего возраста. Совокупность их мРНК (экзом) подвергали частичному секвенированию по 54 конкретным генам. За основу брали данные, полученные при анализе 568 ампликонов (всего 141 000 пар оснований). Для работы применялась методика компании Illumina, а секвенирование проводили с помощью системы MiSeq. Всего выявлено 77 мутаций в 24 исследуемых генах. В каждом позитивном образце выявляли в среднем 2 мутации. Кроме того, авторы показали совпадение между результатами, полученными с помощью NGS и стандартных молекулярно-биологических тестов.

Выявление генов риска неблагоприятного исхода

В ближайшие годы методики «глубинного» секвенирования генома, или NGS, обещают стать вполне конкурентными с ныне принятыми методическими подходами к генодиагностике мутаций. Одно из ранних исследований подобного рода было проведено V. Grossmann и соавт. [6] с применением пиросеквенирования. В работе использовали материал от 22 пациентов с острыми лейкозами и миелоидными новообразованиями. Серийное изучение ДНК методом

пиросеквенирования множественных участков сочетали с гибридизацией на биочипах для анализа длинных последовательностей генов. В целом этот методический подход позволил выявить разнообразные замены, делеции и вставки наряду со сбалансированными хромосомными аберрациями, включая транслокации и инверсии. В частности, авторы обнаружили случаи ОМЛ с аномалиями гена *RUNX1* и ранее неизвестными «слитными генами» с участием этого гена, который кодирует важный фактор транскрипции РНК. Интересно, что мутации *RUNX1* связаны с конкретными хромосомными поломками и имеют большое значение для течения ОМЛ и его исхода [7].

Особый интерес представляют возможные прогностические генные мутации при хронических миелопролиферативных заболеваниях (МПЗ). Так, S. Delic и соавт. [8] с помощью диагностических панелей NGS проводили оценку мутационного спектра и валидацию метода диагностики по 28 генам, потенциально значимым для развития и эволюции МПЗ. Обследованы 100 пациентов с первичным миелофиброзом, истинной полицитемией и эссенциальной тромбоцитопенией. У 53 пациентов обнаружены 2 и более мутации. Показана относительная специфичность определенных мутаций (гены сплайсинга РНК *SF3B1*, *SRSF2* и *U2AF1*, гены *ASXL1* и *EZH2*, влияющие на хроматин) для случаев первичного миелофиброза.

Нарушения эпигенетической регуляции генома

При анализе первичных причин возникновения лейкозов некоторые авторы изучали роль эпигенетической регуляции, а именно метилирования, модификации гистонов и других факторов подавления экспрессии ключевых генов в патогенезе лейкозов, например при HTLV-I-ассоциированном Т-клеточном лейкозе. В работе [9], проведенной на 31 образце от больных с острым Т-лимфобластным лейкозом, осуществляли NGS-анализ для поиска мутаций в генах *SUZ12*, *DNMT1*, *DNMT3A*, *DNMT3B*, *TET1*, *TET2*, *IDH1*, *IDH2*, *MLL*, *MLL2*, *MLL3* и *MLL4*, контролирующих репрессию антионкогенов. Для секвенирования применяли систему Illumina Hiseq 2000. Мутации *MLL3* и *TET2* отмечались наиболее часто (32 % всех образцов), причем они выявлялись только в клетках этих пациентов. Авторы предполагают, что инактивация этих генов особо важна в патогенезе острого Т-лимфобластного лейкоза.

Поиск маркеров минимальной остаточной болезни

Оценка наличия и уровней минимальной остаточной болезни (МОБ) при лейкозах важна для определения тактики дальнейшего лечения. Однако при стандартном поиске онкогенов такие маркеры часто не выявляются. В этом плане представляет интерес статья H. Debarri и соавт. [10]. Авторы проводили мониторинг МОБ по мутациям *IDH1/2* и *DNMT3A* у пациентов с ОМЛ и ранее доказанной мутацией *NPM1*. Осуществлен детальный анализ мутаций гена *NPM1* и более редких генетических аномалий у больных с ОМЛ. Так, мутации *DNMT3A* оказались полезными для количественной оценки с помощью NGS на платформе Ion Torrent (2 млн прочтений на образец). В то же время мутации данного гена отражали наличие предлейкозных клонов в состоянии ремиссии, а не текущую стадию заболевания. Мутации *IDH1/2* и *NPM1* оказались более надежными маркерами МОБ, так как позволяли предсказывать рецидив ОМЛ или злокачественную трансформацию при миелодиспластическом синдроме. Таким образом, правильный выбор маркеров крайне важен для адекватного мониторинга МОБ при ОМЛ.

Анализ клональной эволюции злокачественных клеток

Большинство NGS-исследований клональной эволюции на фоне цитостатической терапии проводили при хроническом лимфолейкозе, что отражено в отечественном обзоре [11]. Данные о ряде оригинальных работ суммированы нами в таблице. В частности,

Применение NGS для анализа клональности при ХЛЛ

Цель работы	Гены, число маркеров	Система генотипирования	Размер выборки	Результат	Страна, ссылка
Выявление значимых мутаций при ХЛЛ	15 генов	MiSeq	136 случаев	Найдены 102 мутации в 8 генах (*TP53*, *SF3B1*, *NOTCH1*, *ATM*, *XPO1*, *MYD88*, *DDX3X*, *PTPN6*)	Германия [12]
Выявление малых клонов лимфоцитов с мутациями	*TP53*	Система 454	309 «свежих» случаев ХЛЛ	Худший прогноз у больных с мутациями *TP53* до терапии	Италия [13]
Анализ *de novo* мутаций после цитостатической терапии	*TP53*	MiSeq	60 случаев	Терапия ведет к отбору в пользу клональных мутаций *TP53*	Чехия [14]
Эффекты метилирования генов	*DAPK1*, 4 генных варианта	iPLEX Gold (Sequenom), MassARRAY	303 случая	Варианты *DAPK1* связаны с уровнями его метилирования	Германия [15]

Примечание. *ХЛЛ – хронический лимфолейкоз.*

показан менее благоприятный прогноз заболевания при выявлении мутаций *TP53*.

Особенности мутационного процесса при хронических МПЗ изучали P. Lundberg и соавт. [16]. Авторы использовали целевое NGS 104 генов для выявления соматических мутаций в группе из 197 больных с хроническими МПЗ. ДНК для анализа получали из гранулоцитов. Полученные библиотеки генных фрагментов исследовали с помощью систем Illumina и подтверждали мутации на платформе Ion Torrent PGM. Для выявления максимального числа мутаций исследовали ДНК из наиболее поздних образцов материала. При этом оценивали индивидуальные мутационные профили для последующего анализа клональной эволюции в серийно взятых пробах. Для оценки копийности генов проводили РНК-секвенирование, а мутации гена *CALR* определяли с помощью ПЦР. В результате у 90 % больных наблюдались определенные соматические мутации, причем в 37 % случаев отмечены более редкие мутации, нежели JAK2V617F и *CALR*. Наличие более 2 мутаций коррелировало со снижением общей выживаемости пациентов и повышенным риском развития ОМЛ, особенно *TP53* с утратой гетерозиготности. Наиболее интересные данные связаны с опытами по исследованию генных мутаций в единичных колониях длительных культур лейкозных клеток. Опорным маркером служила частая мутация JAK2V617F. Было обнаружено, что мутации *TET2* и *DNMT3A* появлялись главным образом до появления JAK2V617F или выявлялись в отдельных клонах, что говорило о биклональном состоянии. Мутации генов *ASXL1* и *EZH2* возникали как до JAK2V617F, так и после нее, а у 3 больных отмечено появление мутации *IDH1* после возникновения мутации JAK2V617F. Авторы выделяют 2 возможные модели развития клональных мутаций: линейный характер приобретения мутаций внутри клона и возможная биклональная структура возникающих злокачественных популяций. В то же время число выявленных мутаций было сходным в ранние и поздние сроки заболевания, что говорит о низкой частоте мутаций у больных с хроническими МПЗ.

Поиск мутаций, связанных с резистентностью к терапии

На протяжении ряда лет при хроническом миелолейкозе (ХМЛ) и некоторых других формах лейкозов применяется терапия ингибиторами тирозинкиназы (ИТК). На фоне такого лечения часто происходит отбор мутированной (резистентной) формы *BCR/ABL*, что определяет необходимость секвенирования этого гена. Даже небольшие клоны со свежими мутациями *BCR/ABL* при ХМЛ могут быть обнаружены с помощью методик NGS. Так, S. Soverini и соавт. [17] попытались выявить тех больных, которые сохраняли исходные мутации резистентности после терапии ИТК II поколения (ИТК2), и выяснить, могут ли минорные клоны с такими мутациями размножиться в ходе данного лечения. Авторы провели оценку иматиниб-резистентных пациентов (ХМЛ, Ph^+). Мутационные профили и их изменения сравнивали между больными, ответившими и не ответившими на терапию ИТК2. Результаты NGS указывают на предсуществование невысоких уровней клинически актуальных мутаций (повышающих риск рецидива) у 43 % пациентов, не ответивших на лечение. Напротив, у больных, успешно леченных ИТК2, не было исходно найдено мутаций, ведущих к резистентности. Таким образом, NGS-мониторинг иматиниб-резистентных больных может эффективно выявлять ряд ранних прогностически важных мутаций.

Проблемы биоинформатики при анализе больших баз данных секвенирования нового поколения

В настоящее время силами одной лаборатории можно секвенировать весь геном человека за несколько дней и за несколько тысяч долларов, что делает эту процедуру доступной для многих университетов и крупных медицинских учреждений. В обзоре R.R. Gullapalli и соавт. [3] описан этот круг методик, а также обсуждаются ограничения в их практическом использовании. На текущий момент наиболее важной задачей является создание структурированных баз данных о геноме исследованных клеток, адекватных способов их компьютерной обработки и выдачи результатов, касающихся изменений в геноме данного индивида. Следует обратить внимание на то, что биоинформационная обработка данных NGS часто не только представляет собой практическую сложность, но и может радикально повлиять на диагноз.

Эволюция программных продуктов для анализа результатов геномного секвенирования детально изложена в статье G.A. Van der Auwera и соавт. [18]. Типичная последовательность обработки данных зависит от природы эксперимента (WGS, WES, target panel, RNA-seq). Важнейшей первой стадией обработки любого NGS-эксперимента является контроль качества «сырых» прочтений, который осуществляется с помощью программы FastQC. Первичный анализ позволяет оценить проблемы, которые имели место в подготовке библиотеки или при выделении материала, и скорректировать их. Геномные эксперименты подразумевают «выравнивание» данных на референсный геном. Наиболее популярной программой для этой цели является BWA. Так, для экспериментов RNA-seq необходимо выравнивание с учетом сплайсинга («перекройки» молекул РНК после их считывания в клетке). На текущий момент самой быстрой и точной программой для такого выравнивания является STAR.

После выравнивания геномного эксперимента (WGS, WES, таргетная панель) порядок действий следующий:

1) определение вариантов (variant calling) – стадия, в которой программа определяет варианты, отличающиеся от референсной последовательности;

2) аннотация вариантов (variant annotation) – стадия, в которой для хорошо охарактеризованных замен описывается их патогенность (на основании баз Clin-Var, OMIM и др.), а для малоизвестных замен предсказывается эффект с помощью одной из программ – предикторов эффекта (PROVEAN, SIFT, Polyphen, MutPred и др.);

3) ранжирование вариантов – стадия, в которой найденные и аннотированные варианты сортируются по ряду критериев, таких как частота аллели в публичных базах данных, патогенность, гомо-/гетерозиготность замены и др.

Наиболее чувствительной и неоднозначной является 1-я стадия (определение вариантов), для которой существует ряд хорошо разработанных и поддерживаемых решений. Самым популярным и точным является комплекс программ Genome Analysis Toolkit (GATK). Особенности этого пакета – запрет на использование коммерческими структурами и необходимость наличия значительной когорты (30 и более образцов) для надежного определения вариантов. Хорошо разработанными альтернативами также могут служить пайплайны на основе samtools/VCFtools и freebayes.

Большой интерес вызывают результаты сравнения эффективности различных систем обработки сигналов NGS и точности выявления соответствующих генных вариантов [19]. Совокупность данных о генотипах, полученных на базе Illumina Infinium Human Exome v1.1 Beadchip, обрабатывали с помощью 4 программных систем: SAMtools, GATK, glftools и Atlas2, изучая варианты анализа одного или нескольких образцов ДНК. Для выравнивания прочтений использовали общую программу (BWA), и далее на этой основе были созданы 7 потоков информации, которые позволили проанализировать данные экзомного секвенирования 20 человек. В качестве «золотого стандарта» были взяты данные классического секвенирования по Сэнгеру, с которым и сравнивались результаты NGS-генотипирования. В целом анализ отдельно взятых образцов продемонстрировал высокое сходство по перекрыванию целевых генов (порядка 0,9). При этом пакет программ GATK показал наибольшую специфичность (> 0,9999). Введение параллельных образцов существенно повышало чувствительность. По результатам симуляционных экспериментов GATK превосходила SAMtools и glfSingle по чувствительности, особенно при невысоких уровнях перекрывания фрагментов.

Таким образом, помимо правильного выбора технологической платформы для проведения NGS важным является также подбор оптимального программного обеспечения для обработки большого объема полученной генетической информации и ее сопоставления со стандартными (референсными) последовательностями генома человека. За последние годы разработано множество специализированных пакетов программ для обработки данных NGS. В частности, S. Pabinger и соавт. [20] изучили возможности 205 программ для анализа материала при полногеномном и полноэкзомном секвенировании. Авторы выделяют 5 аспектов NGS-исследования, каждый из которых подлежит отдельной оценке: контроль качества, сопоставление с известными генными последовательностями, идентификация генных вариантов, их аннотация и визуальное представление. Были выбраны 32 программы для точного определения генных вариантов. Особое внимание было уделено функциональным свойствам и специальным условиям применения отдельных программных продуктов, что особенно интересно для специалистов по клинической генетике, использующих различные платформы для выявления специфических мутаций.

Проблемы, связанные с интерпретацией данных секвенирования, также рассматривались в работе E.L. Crowgey и соавт. [21]. Авторы предложили интегрированный подход к анализу этого потока данных от начала до конца процесса – от выявления генных вариантов до их функциональной оценки на основе принципа фильтрования биоинформации. Результаты считывания множества генных фрагментов переводили из bam-файлов в формат fastq с помощью bam2fastq. Параметры и число покрытий на 1 экзон (vertical) и среднее покрытие экзона (horizontal) вычисляли на каждый образец с помощью GATK DiagnoseTarget (v. 2.5–2). Специализированные потоки сведений создавали для аннотирования и фильтрации однонуклеотидных полиморфизмов (ОНП) и вставок/делеций. Соответствующие схемы обработки данных секвенирования, полученных в системе Illumina, были успешно применены в плане сопоставлений генных последовательностей, детекции ОНП, мелких делеций/вставок и изменений числа копий генных фрагментов. При этом идет аннотирование вариантов с учетом типа заболевания и вкладом данного генного варианта. В ходе работы с ранее валидированными объектами разработаны специальные алгоритмы для фильтрации вариантов, основанные на качестве варианта, его типе наследования и влиянии на функцию кодируемого белка. Соответствующие клинические ассоциации привязаны к информационному ресурсу iProXpress (данные о редких заболеваниях). Таким образом, эта система позволяет проводить следующие операции: аннотирование; ранжирование (приоритизацию); фильтрацию по характеру наследования, функциональной оценке и т. д., что дает возможность клинической оценки геномных изменений при редких генетических заболеваниях.

Применение различных методик биоинформатики при анализе результатов экзомного секвенирования отдельно рассматриваются в статье M. D'Antonio и соавт. [22]. Целью такого анализа является поиск информативных аллельных вариантов и их связь с фенотипом обследуемого человека. Авторы предлагают свой веб-ресурс WEP (Whole-Exome sequencing Pipeline), который включает несколько этапов обработки потока геномной информации:

ОНКОГЕМАТОЛОГИЯ 4'2016 ТОМ 11 | ONCOHEMATOLOGY 4'2016 VOL. 11

1) проверка целостности и качества данных на входе, фильтрация информации;

2) правильное сопоставление фрагментов генов;

3) конверсия в системе BAM, сортировка и индексация;

4) дублирование удаленных данных;

5) оптимизация сопоставлений вокруг точек вставок/делеций;

6) повторная калибровка по качеству;

7) детекция генных вариантов – ОНП и вставок/делеций;

8) аннотация вариантных генов;

9) сохранение результатов в специализированных базах данных для обеспечения возможности последующих сравнений, статистической обработки и др.

Авторы отмечают, однако, что для управления этим процессом нужен IT-персонал с соответствующими навыками.

Происходит дальнейшее совершенствование анализа массивов данных по результатам NGS, в частности для идентификации ОНП обычно используют различные версии аннотаций данных NGS. Проблема, однако, состоит в должной интеграции и сопоставлении массивов данных, полученных разными методами секвенирования, а также с историческими контролями. Иногда бывает трудно полностью совместить отдельные массивы информации. Так, при анализе отдельных генетических вариантов в исследовании [23] использовали стандартную номенклатуру клинического секвенирования (CSN). Однако некоторые проблемные генные варианты, например инсерции/делеции нуклеотидов (инделы), с трудом выявлялись в этом массиве. Поэтому авторы предложили новый программный продукт CAVA (Clinical Annotation of Variants), соответствующий современным принципам анализа генного полиморфизма и приему результатов NGS. Это относительно простая система для введения и анализа потоков данных NGS, позволяющая уточнить свойства транскрипта, местоположение гена в цепи ДНК и провести сопоставление со сходными генными вариантами. Ускорение этого процесса позволяет облегчить клиническую оценку обнаруженного генного варианта в сравнении с другими базами данных. Все это дает возможность обеспечить стандартное, клинически значимое аннотирование данных NGS в соответствии с действующими клиническими стандартами.

Возможности секвенирования нового поколения в подборе пар донор–реципиент для трансплантации

Отдельную проблему составляет обработка результатов NGS в практике трансплантации гемопоэтических стволовых клеток, что требует тщательного подбора доноров гемопоэтических клеток по антигенам системы HLA, которые кодируются на хромосоме 6. Классические методы типирования HLA, основанные на ПЦР-диагностике, отработаны в совершенстве, однако не оценивают всего разнообразия клинически актуальных антигенов совместимости. Ранее упомянутые биоинформационные подходы, основанные на сравнении с референс-генами, тут неприменимы ввиду выраженной гетерогенности (полиморфизма) генов системы HLA.

Поэтому сейчас глубинное HLA-типирование и анализ сочетаний различных антигенных вариантов являются одним из ведущих направлений NGS – это направление иногда называется иммуноинформатикой. В частности, на его основе можно предсказывать взаимодействия B- и T-клеточных эпитопов (иммуноактивных участков белков) в организме больного после трансплантации. Для этих целей разработан целый ряд программных средств, которые совершенствуются с учетом внедрения методов NGS [24]. В данной работе содержится обзор большого числа средств иммуноинформатики с акцентом на предсказание эпитопов B- и T-клеток. Стало возможным оценивать с высокой точностью аллотипы (при возможности выбора оптимального донора) и вероятные влияния генных вариантов HLA на риск развития иммунных конфликтов между клетками пациента и трансплантата, а также развивать методы иммунотерапии для персонализированного лечения раковых заболеваний.

Заключение

Революционные технологии NGS за последние годы привели к повышению надежности и снижению стоимости процедур глубинного секвенирования геномных последовательностей, ускорению генодиагностики и расширению приложений геномного секвенирования. В дальнейшем можно ожидать и применения методов NGS для анализа минимальных образцов ДНК, например в единичных клетках или внеклеточной ДНК в плазме крови.

Однако широкое применение методов NGS в условиях клинической практики требует совершенствования компьютерных технологий для полноценной обработки больших баз данных, получаемых в разных лабораториях, разработки стандартизированных и воспроизводимых методов биоинформатики, пригодных для дифференциальной диагностики лейкозов и типирования антигенов гистосовместимости.

ЛИТЕРАТУРА / REFERENCES

1. Смирнихина С.А., Лавров А.В., Адильгереева Э.П. и др. Клиническое значение полноэкзомных исследований миелоидных опухолей методом секвенирования следующего поколения. Клиническая онкогематология 2013;6(1):11–9. [Smirnikhina S.A., Lavrov A.V., Adil'gireeva E.P. et al. Clinical significance of whole-exome studies using next generation sequencing in myeloid neoplasia. Klinicheskaya onkogematologiya = Clinical Oncohematology 2013;6(1):11–9 (In Russ.)].
2. Ross J.S., Cronin M. Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol 2011;136(4):527–39. DOI: 10.1309/AJCPR1SVT1VHUGXW. PMID: 21917674.
3. Gullapalli R.R., Desai K.V., Santana-Santos L. et al. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. J Pathol Inform 2012;3:40. DOI: 10.4103/2153-3539.103013. PMID: 23248761.
4. Papaemmanuil E., Gerstung M., Bullinger L. et al. Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med 2016;374(23):2209–21. DOI: 10.1056/NEJMoa1516192. PMID: 27276561.
5. Au C.H., Wa A., Ho D.N. et al. Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. Diagn Pathol 2016;11:11. DOI: 10.1186/s13000-016-0456-8. PMID: 26796102.
6. Grossmann V., Kohlmann A., Klein H.U. et al. Targeted next-generation sequencing detects point mutations, insertions, deletions and balanced chromosomal rearrangements as well as identifies novel leukemia-specific fusion genes in a single procedure. Leukemia 2011;25(4):671–80. DOI: 10.1038/leu.2010.309. PMID: 21252984.
7. Gaidzik V.I., Bullinger L., Schlenk R.F. et al. RUNX1 mutations in acute myeloid leukemia: results from a comprehensive genetic and clinical analysis from the AML study group. J Clin Oncol 2011;29(10):1364–72. DOI: 10.1200/JCO.2010.30.7926. PMID: 21343560.
8. Delic S., Rose D., Kern W. et al. Application of an NGS-based 28-gene panel in myeloproliferative neoplasms reveals distinct mutation patterns in essential thrombocythaemia, primary myelofibrosis and polycythaemia vera. Br J Haematol 2016;175(3):419–26. DOI: 10.1111/bjh.14269. PMID: 27447873.
9. Yeh C.H., Bai X.T., Moles R. et al. Mutation of epigenetic regulators TET2 and MLL3 in patients with HTLV-I-induced acute adult T-cell leukemia. Mol Cancer 2016;15(1):15. DOI: 10.1186/s12943-016-0500-z. PMID: 26880370.
10. Debarri H., Lebon D., Roumier C. et al. IDH1/2 but not DNMT3A mutations are suitable targets for minimal residual disease monitoring in acute myeloid leukemia patients: a study by the Acute Leukemia French Association. Oncotarget 2015;6(39):42345–53. DOI: 10.18632/oncotarget.5645. PMID: 26486081.
11. Северина Н.А., Бидерман Б.В., Никитин Е.А., Судариков А.Б. Мутации генов при хроническом лимфолейкозе: новые аспекты патогенеза, открытые с помощью технологий полногеномного секвенирования. Гематология и трансфузиология 2014;59(3):41–8. [Severina N.A., Biderman B.V., Nikitin E.A., Sudarikov A.B. Gene mutations in chronic lymphocytic leukemia: new aspects of pathogenesis discovered by next generation sequencing. Gematologiya i transfuziologiya = Hematology and Transfusiology 2014;59(3):41–8. (In Russ.)].
12. Vollbrecht C., Mairinger F.D., Koitzsch U. et al. Comprehensive analysis of disease-related genes in chronic lymphocytic leukemia by multiplex PCR-based next generation sequencing. PLoS One 2015;10(6):e0129544. DOI: 10.1371/journal.pone.0129544. PMID: 26053404.
13. Rossi D., Khiabanian H., Spina V. et al. Clinical impact of small TP53 mutated subclones in chronic lymphocytic leukemia. Blood 2014;123(14):2139–47. DOI: 10.1182/blood-2013-11-539726. PMID: 24501221.
14. Malcikova J., Stano-Kozubik K., Tichy B. et al. Detailed analysis of therapy-driven clonal evolution of TP53 mutations in chronic lymphocytic leukemia. Leukemia 2015; 29(4):877–85. DOI: 10.1038/leu.2014.297. PMID: 25287991.
15. Landau D.A., Tausch E., Taylor-Weiner A.N. et al. Mutations driving CLL and their evolution in progression and relapse. Nature 2015;526(7574):525–30. DOI: 10.1038/nature15395. PMID: 26466571.
16. Lundberg P., Karow A., Nienhold R. et al. Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. Blood 2014;123(14):2220–8. DOI: 10.1182/blood-2013-11-537167. PMID: 24478400.
17. Soverini S., De Benedittis C., Polakova K.M. et al. Next-generation sequencing for sensitive detection of BCR-ABL1 mutations relevant to tyrosine kinase inhibitor choice in imatinib-resistant patients. Oncotarget 2016;7(16):21982–90. DOI: 10.18632/oncotarget.8010. PMID: 26980736.
18. Van der Auwera G.A., Carneiro M.O., Hartl C. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43:11.10.1–33. DOI: 10.1002/0471250953.bi1110s43. PMID: 25431634.
19. Liu X., Han S., Wang Z. et al. Variant callers for next-generation sequencing data: a comparison study. PLoS One 2013;8(9):e75619. DOI: 10.1371/journal.pone.0075619. PMID: 24086590.
20. Pabinger S., Dander A., Fischer M. et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 2014;15(2):256–78. DOI: 10.1093/bib/bbs086. PMID: 23341494.
21. Crowgey E.L., Stabley D.L., Chen C. et al. An integrated approach for analyzing clinical genomic variant data from next-generation sequencing. J Biomol Tech 2015;26(1):19–28. DOI: 10.7171/jbt.15-2601-002. PMID: 25649353.
22. D'Antonio M., D'Onorio De Meo P., Paoletti D et al. WEP: a high-performance analysis pipeline for whole-exome data. BMC Bioinformatics 2013;14 Suppl 7:S11. DOI: 10.1186/1471-2105-14-S7-S11. PMID: 23815231.
23. Münz M., Ruark E., Renwick A. et al. CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med 2015;7:76. DOI: 10.1186/s13073-015-0195-6. PMID: 26315209.
24. Backert L., Kohlbacher O. Immuno-informatics and epitope prediction in the age of genomic medicine. Genome Med 2015;7:119. DOI: 10.1186/s13073-015-0245-0. PMID: 26589500.