ОБЗОРЫ

ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ СЕКВЕНАТОРОВ ТРЕТЬЕГО ПОКОЛЕНИЯ ДЛЯ КОЛИЧЕСТВЕННОГО ПРОФИЛИРОВАНИЯ ТРАНСКРИПТОМА

С.П. Радько[1]*, Л.К. Курбатов[1], К.Г. Птицын[2], Я.Ю. Киселёва[3], Е.А. Пономаренко[1], А.В. Лисица[1], А.И. Арчаков[1]

[1]Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича, 119121, Москва, ул. Погодинская 10;*эл. почта : radkos@yandex.ru
[2]Научно-исследовательский институт Аджиномото-Генетика, 117545, Москва, 1-й Дорожный проезд 1
[3]Российский научный центр рентгенорадиологии, 117997, Москва, ул. Профсоюзная 86

Транскриптомное профилирование широко используется для анализа динамики транскриптома при исследовании различных биологических процессов на клеточном и тканевом уровне. В отличие от секвенаторов второго поколения, которые секвенируют относительно короткие фрагменты нуклеиновых кислот, ДНК/РНК секвенаторы третьего поколения, разработанные биотехнологическими компаниями “PacBio” и “Oxford Nanopore Technologies”, позволяют секвенировать транскрипты как единичные молекулы и могут рассматриваться как потенциальные молекулярные счётчики, способные измерять количество копий каждого транскрипта с высокой производительностью, чувствительностью и специфичностью. В данном обзоре рассмотрены особенности технологий одномолекулярного секвенирования, предлагаемых компаниями “PacBio” и “Oxford Nanopore Technologies”, и применение технологий одномолекулярного секвенирования для целей транскриптомного анализа, включая анализ изоформ транкриптов. Также обсуждаются перспективы и ограничения их использования для количественного профилирования транскриптома.

Ключевые слова: секвенирование третьего поколения; транскриптом; количественное профилирование

DOI: 10.18097/BMCRM00086

ВВЕДЕНИЕ

Инициирование и успешное выполнение проекта “Геном человека” дало начало новым, так называемым “омиксным” подходам к исследованию фундаментальных и прикладных аспектов функционирования живых существ. Характерной чертой этих подходов является получение и биоинформатический анализ огромных массивов данных [1]. Дополнительно к геномике, сегодня окончательно сложились и активно развиваются такие “омиксные” направления, как протеомика, транскриптомика, метаболомика и интерактомика, каждый из которых опирается на свои технологические платформы [1]. Транскриптомика, которая в первую очередь фокусируется на идентификации транскриптов во всем разнообразии их изоформ и на количественной оценке их представленности в разнообразных типах клеток (количественное профилирование транскриптома), опиралась в начале своего развития на технологию микрочипов, первоначально разработанную для геномного анализа [2]. В настоящее время использование технологий секвенирования, известных как “секвенирование нового поколения” (Next Generation Sequencing, NGS), стало доминирующим методологическим подходом к анализу транскриптома, включая его количественное профилирование [3]. Сегодня NGS подразделяют на секвенирование второго и третьего поколений (SGS и TGS – Second and Third Generation Sequencing) [4], технологии которых существенно различаются.

Одним из характерных различий SGS и TGS является размер секвенируемых молекул. В случае SGS секвенируются относительно небольшие (от 25 до 500 пар оснований (п.о.)) фрагменты нуклеиновых кислот с перекрывающимися последовательностями. По аналогии с принятым в англоязычной литературе термином “reads”, последовательности ДНК таких фрагментов в русскоязычной литературе называют “чтениями”. После завершения этапа секвенирования, “чтения” собираются в протяженные участки генома с помощью специальных математических алгоритмов (aligners), позволяющих проводить их выравнивание [5]. При транскриптомном анализе проводят секвенирование фрагментов кДНК, которые потом собирают в последовательности транскриптов путём их выравнивания с использованием референсных геномов или транскриптомов (картирование), или без их использования (секвенирование транскриптома *de novo*). При количественном профилировании транскриптомов методами SGS в качестве метрики используют нормированное специальным образом количество “чтений”, картирующихся на определённый ген (транскрипт): RPKM (Reads Per Kilobase Per Million mapped reads) или FPKM (Fragments Per Kilobase Per Million mapped reads), что позволяет определять относительную представленность транскриптов в исследуемом образце [6].

В случае TGS длина “чтения” может достигать десятков тысяч п.о. [7]. Применительно к транскриптому такая длина “чтения” делает возможным проведение так называемого “одномолекулярного секвенирования” (single molecule sequencing), когда практически любой транскрипт секвенируется как единичная молекула, что устраняет необходимость последующей “сборки”. Таким образом, секвенаторы третьего поколения можно рассматривать как потенциальные молекулярные счётчики, позволяющие определять представленность каждого транскрипта в транскриптоме с высокой производительностью, чувствительностью и специфичностью. В данном обзоре рассмотрены основные технологические особенности одномолекулярного секвенирования с использованием технологических платформ, разработанных компаниями “Pacific Biosciences” (США) и “Oxford Nanopore Technologies” (Великобритания), а также их применение в транскриптомном анализе, включая количественное профилирования транскриптома.

1. ТЕХНОЛОГИИ ОДНОМОЛЕКУЛЯРНОГО СЕКВЕНИРОВАНИЯ

Термин “одномолекулярное секвенирование” первоначально появился как часть названия технологии секвенирования ДНК, разработанной компанией “Pacific Biosciences” и получившей название “одномолекулярное секвенирование в реальном времени” (Single Molecule Real Time sequencing – SMRT) [8]. В дальнейшем он был применён для обозначения технологии

нанопорового секвенирования, предложенной компанией "Oxford Nanopore Technologies", также характеризующегося длиной "чтения", достигающей десятки тысяч п.о. [9, 10].

1.1. SMRT-секвенирование

Технология SMRT-секвенирования основана на наблюдении в реальном времени за синтезом ДНК-полимеразой новой цепи на ДНК-матрице, которая и представляет секвенируемую молекулу ДНК [11, 12]. Иллюстрация концепции SMRT-секвенирования представлена на рисунке 1. Технология включает два принципиальных компонента: 1) плоский световод, который позволяет направлять световую энергию в ячейку, размер которой значительно меньше длины волны видимого света (zero-mode waveguide, ZMW); 2) дезоксинуклеозидтрифосфаты (dNTP), у которых к концевой фосфатной группе присоединён флуорофор.

ZMW-ячейка представляет подложку из плавленого кварца, покрытую металлической плёнкой (алюминиевой или золотой) толщиной ~100 нм, в которой сделано отверстие диаметром ~100 нм. Особенность распространения света в отверстии с такой апертурой состоит в том, что его интенсивность быстро затухает, в результате чего освещённой оказывается лишь малая часть ячейки (объёмом около 20 зептолитров), прилегающая к поверхности подложки [4]. Время диффузии флуоресцентно-меченных dNTP (далее фм-dNTP) через такой малый объём составляет несколько микросекунд. Дно ячейки дериватизировано полиэтиленгликолем (ПЭГ), конъюгированным с биотином. В качестве ДНК-полимеразы используется мутантный вариант рекомбинантной ДНК-полимеразы бактериофага φ29 с пониженной 3′-5′-экзонуклеазной активностью, биотинилированный *in vivo* на N-конце [13]. После формирования комплексов ДНК-полимеразы с ДНК-матрицей, к ним добавляется стрептавидин, что позволяет иммобилизовать комплекс полимераза-матрица на дне ячейки через формирование мультимерного комплекса ПЭГ-стрептавидин-полимераза-матрица [12].

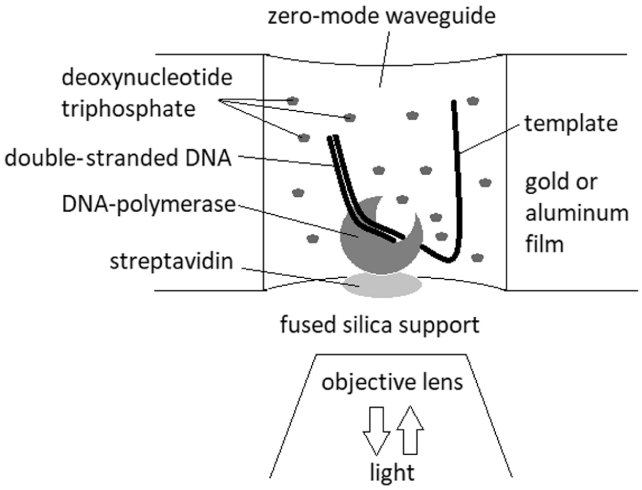

Рисунок 1. Схематическая иллюстрация одномолекулярного секвенирования ДНК в реальном времени (SMRT-секвенирования). ДНК-полимераза в комплексе с ДНК-матрицей иммобилизована на дне ZMW-ячейки. Секвенирование матрицы происходит путём наблюдения в реальном времени за процессингом ДНК-полимеразой дезоксинуклеотидтрифосфатов, меченных флуорофорами через концевую фосфатную группу, при включении нуклеотидов в растущую цепь ДНК.

Присоединение флуорофора к концевой фосфатной группе dNTP (в отличие от более распространённого подхода к его конъюгации с dNTP через присоединение к основанию) приводит к отщеплению флуорофора полимеразой при включении нуклеотида в синтезируемую цепь ДНК. В результате нуклеотиды в синтезируемой цепи не мечены флуорофорами и не дают вклада во флуоресцентный сигнал. Характерное время между захватом фм-dNTP активным центром ДНК-полимеразы и отщеплением флуорофора при включении нуклеотида в растущую цепь ДНК составляет несколько мс, что позволяет дискриминировать такое событие от прохождения фм-dNTP через освещённый объём ячейки в силу броуновского движения, основываясь на длительности сигнала. Каждый тип dNTP мечен флуорофором, характеризующимся своим "цветом" (испускающим свет определенной длины волны), что позволяет наблюдать процесс синтеза в реальном времени как последовательность флуоресцентных сигналов разного "цвета" [4, 12]. Регистрация сигнала во времени происходит с помощью опто-электронной системы, сопрягающей возможности конфокальной микроскопии, фото-электронного усиления сигнала и CCD-камеры (CCD – charge-coupled device) и позволяющей детектировать единичные молекулярные события, такие как включение флуоресцентно-меченного dNTP в растущую цепь ДНК [11, 12, 14].

Характерной особенностью SMRT-технологии является использование в качестве матрицы топологически замкнутой (кольцевой) ДНК (SMRTbell template [15]), которую получают лигированием ДНК-адаптеров, имеющих структуру "шпильки", к фрагменту двунитевой ДНК (рис. 2).

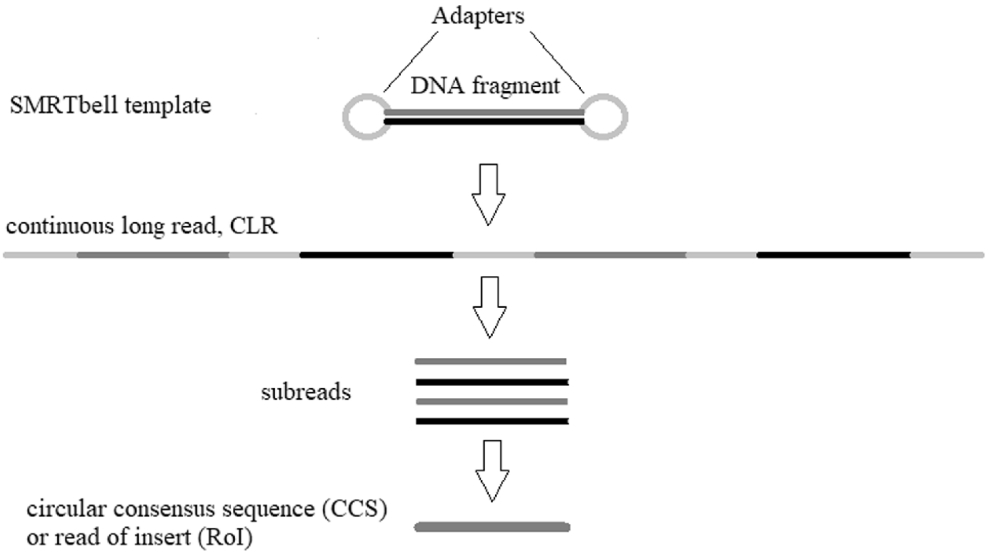

Рисунок 2. Схематическое изображение SMRTbell-матрицы и последовательности обработки различных типов "чтений" в процессе SMRT-секвенирования.

ДНК-полимераза фага φ29 является "вытесняющей ДНК-полимеразой" (strand displacing DNA polymerase), то есть способна вытеснять одну из нитей ДНК при синтезе, в результате чего "чтения" включают как "смысловые", так и "антисмысловые" последовательности секвенируемого фрагмента ДНК, разделённые последовательностями адаптеров и повторяющиеся много раз [15, 16]. Результатом является так называемое "непрерывное длинное чтение" (continuous long read, CLR), которое при обработке данных путём узнавания и устранения последовательностей адаптеров может быть преобразовано в набор "субчтений" (subreads). Консенсусная последовательность "субчтений" единичного CLR носит название "циркулярная консенсусная последовательность" (circular consensus sequence, CCS) или "чтение вставки" (read of insert, RoI). Обработка CLR и получение набора RoI происходят в автоматическом режиме с использованием программного обеспечения, предоставляемого "Pacific Biosciences". Биоинформатический инструментарий для анализа результатов SMRT-секвенирования также доступен на GitHub [17].

Высокая производительность секвенирования достигается тем, что синтез ДНК проводиться параллельно в тысячах ZMW-ячейках, расположенных с высокой плотностью на плоской кварцевой подложке (ZMW-чип) [12]. Так, в первом секвенаторе третьего поколения PacBio RS, который был выведен компанией "Pacific Biosciences" на рынок в 2011 г., использованы чипы, содержащие 3 тыс. ZMW-ячеек. В 2013 г. компания начала продажи секвенатора PacBio RS II, где параллельное секвенирование проводится в 150 тыс. ZMW-ячейках. Тем не менее, производительность SMRT-секвенирования заметно ниже, чем производительность многих секвенаторов второго поколения – как правило, пропускная способность одного чипа, содержащего 150 тыс. ZMW-ячеек, составляет 0.5-1 млрд. нуклеотидов (нт) [16]. Следует отметить также, что фактически только около 35-75 тыс. ZMW-ячеек из 150 тыс. содержат одну иммобилизованную молекулу ДНК-полимеразы на ячейку и могут быть использованы для секвенирования (остальные либо не содержат ДНК-полимеразы, либо содержат более одной молекулы полимеразы из-за пуассоновского распределения числа ДНК-полимераз по ZMW-ячейкам при иммобилизации) [12, 16]. В настоящее время "Pacific Biosciences" предлагает систему одномолекулярного секвенирования в реальном времени Sequel System (с 2016 г.), использующую чипы, содержащие 1 млн. ZMW-ячеек, что повышает пропускную способность до 3.5-7 млрд. нт на ZMW-чип [16]. Одновременно с увеличением пропускной способности происходила оптимизация и совершенствование протоколов секвенирования и используемых реагентов (включая рекомбинантную ДНК-полимеразу), что привело к возрастанию средней длины чтения: в секвенаторах PacBio RS медианное значение длины чтения составляло 4 тыс. нт, в PacBio RS II – около 10 тыс. нт, а в Sequel System – более 20 тыс. нт. Максимальная длина чтения превышает 60 тыс. нт, а в некоторых случаях может даже достигать более 90 тыс. нт [18].

Одним из существенных недостатков SMRT-секвенирования является высокая частота ошибки, которая достигает 15% [4, 19, 20]. Среди причин столь высокого уровня ошибок – захват dNTP активным центром ДНК-полимеразы без последующего включения основания в растущую цепь [19]. При этом время нахождения "захваченного" фм-dNTP в освещенном объёме ZMW-ячейки может также составлять мс и восприниматься системой как включение нуклеотида в растущую цепь. Поскольку ошибки распределены случайно по длине "чтения", они могут быть идентифицированы и устранены, если последовательность ДНК-фрагмента повторяется в "чтении" достаточное количество раз (что достигается использованием SMRTbell-матрицы). Так, при 15-кратном повторе вероятность ошибки становится менее 1% [20], а при 30-кратном точность секвенирования – более 99.999% [18]. Однако по мере увеличения длины секвенируемого ДНК-фрагмента количество повторов в "чтении" в среднем уменьшается, что приводит к возрастанию уровня ошибок [20].

Существенным недостатком SMRT-технологии является то, что иммобилизация комплексов ДНК-полимераза/SMRTbell-матрица в ZMW-ячейке зависит от размера матрицы: комплексы с матрицами меньшего размера иммобилизуются со значительно более высокой эффективностью, что приводит к превалированию коротких ДНК-фрагментов [16]. Применительно к транскриптомному анализу это означает, что короткие транскрипты будут секвенироваться со значительно большей частотой, чем они встречаются в секвенируемой популяции молекул РНК. Для решения данной проблемы "Pacific Biosciences" предлагает проводить предварительное разделение кДНК по размеру на фракции: 1-2 тыс. нт, 2-3 тыс. нт, 3-6 тыс. нт и 5-10 тыс. нт. Фракционирование кДНК проводится с помощью либо традиционного агарозного электрофореза, либо, что предпочтительно, гель-электрофореза в пульсирующем поле (pulsed field gel electrophoresis, PFGE) [16]. В последнем случае производитель рекомендует использовать PFGE-системы BluePippin или SageELF компании "Sage Science" (США). Так как процедура фракционирования кДНК по размеру приводит к значительному уменьшению количества материала, что может привести к потере низкокопийных транскриптов, рекомендуется проводить ПЦР-амплификацию кДНК перед фракционированием и каждой фракции после фракционирования перед созданием SMRTbell-библиотеки. Метод транскриптомного анализа, разработанный компанией Pacific Biosciences, включающий получение кДНК (1), пре-амплификацию кДНК (2), фракционирование кДНК по размеру (3), пост-амплификацию кДНК (4), создание SMRTbell-библиотеки для каждой фракции (5) и её последующее SMRT-секвенирование (6), получил название Iso-Seq [16].

Следует отметить, что технология SMRT может потенциально быть использована для прямого секвенирования молекул РНК, если ДНК-полимеразу заменить на обратную транскриптазу [21]. Используя обратную транскриптазу вируса иммунодефицита человека, авторы [21] показали возможность мониторинга синтеза кДНК на РНК-матрице в ZMW-ячейке в реальном времени. Метод позволил идентифицировать модификации оснований РНК и наблюдать перестройки вторичной структуры молекул РНК. Однако дальнейшего развития применение SMRT-технологии для прямого секвенирования РНК за прошедшие после публикации 5 лет так и не получило.

1.2. Нанопоровое секвенирование

Нанопоровое секвенирование нуклеиновых кислот основано на пропускании молекул однонитевой ДНК (онДНК) или РНК через поры диаметром 1-2 нм. Появлению нанопоровых секвенаторов предшествовала череда академических исследований, посвящённых анализу прохождения ДНК и РНК через белковые нанопоры, формируемые порinами – специализированными трансмембранными белками бактерий – в липидном бислое. Первое исследование такого рода относится к 1996 г., когда Kasianowicz и соавт. [22] показали, что гомоолигонуклеотиды ДНК под действием приложенного напряжения могут проходить через поры, образуемые α-гемолизином (α-ГЛ) при его встраивании в искусственную липидную мембрану (диаметр такой поры составляет 1.4 нм). При этом происходит

временная блокировка поры, которая сопровождается уменьшением количества ионов электролита, проходящих через пору под действием приложенного напряжения и, соответственно, снижением силы тока, протекающего через мембрану. Авторы также показали, что величина снижения силы тока зависела от типа нуклеотидов, образующих гомополимер ДНК, и что двунитевая ДНК (днДНК), представленная дуплексами гомоолигонуклеотидов, была не способна пройти через пору [22]. Позднее они же показали, что прохождение через α-ГЛ-нанопору синтетической РНК, состоящей из двух сегментов, образованных разными нуклеотидами (включая нуклеотиды с метилированными основаниями), может быть детектировано как ступенчатое падение силы тока [23]. Эти работы продемонстрировали принципиальную возможность секвенировать нуклеиновые кислоты, основываясь на детекции последовательности изменений величины силы тока, протекающего через мембрану, при транслокации молекулы онДНК или РНК через встроенную в мембрану белковую пору. Однако практическая реализация этой возможности требовала решения по меньшей мере двух задач [24]: (1) замедления транслокации молекул онДНК и РНК с 100-1000 нт в миллисекунду до ~1 нт в миллисекунду, чтобы позволить временное разрешение их прохождения через нанопору на однонуклеотидном уровне, и (2) распознавания последовательностей нуклеотидов, формирующих сигнал (длина поры в α-ГЛ такова, что сигнал генерируется участком последовательности, включающим 12 нт).

Возможные решения этих задач были показаны в работах [25-27]. Как оказалось, формирование комплекса нуклеиновой кислоты с моторным белком, в качестве которого была использована ДНК-полимераза фага φ29, может замедлить прохождение онДНК через α-ГЛ-нанопору до 2.5-40 нт/с, что позволяет регистрировать её транслокацию с временным разрешением на однонуклеотидном уровне [25, 26]. Авторы работ продемонстрировали, что различные активности φ29 ДНК-полимеразы (которые проявляются в зависимости от доступности для полимеразы ионов магния и/или дезоксинуклеотидтрифосфатов и включают геликазную, 3′-5′-экзонуклеазную и полимеразную активности) могут быть использованы для контроля скорости транслокации онДНК через пору (рис. 3). Решением второй задачи стала замена α-ГЛ на генетически модифицированный рекомбинантный вариант белка MspA (Mycobacterium smegmatis protein A – трансмембранный белок *Mycobacterium smegmatis*). В отличие от α-ГЛ, длина поры MspA такова, что сигнал генерируется участком онДНК, состоящим из 4 нт, при этом замена даже одного нуклеотида на другой в таком “квадромере” приводит к детектируемому изменению амплитуды сигнала [27]. Интегрирование возможностей, которые давало одновременное использование φ29 ДНК-полимеразы и “мутантной” MspA-нанопоры, продемонстрировало, что транслокация различных последовательностей ДНК длиной 42-53 нт через нанопору сопровождается своим (то есть характерным для каждой тестированой последовательности) паттерном изменений силы тока во времени [28]. Экспериментальное исследование характеристических паттернов изменений силы тока для всех 256 возможных четырёхбуквенных комбинаций нуклеотидов dA, dT, dC и dG позволило составить “карту квадромеров”, с помощью которой оказалось возможным конвертировать последовательности изменений в силе тока при транслокации длинных (до 4.5 тыс. нт) фрагментов ДНК фага φX174 в нуклеотидные последовательности, которые однозначно картировались на геном фага [29].

Очевидно, эти же принципы лежат в основе метода нанопорового секвенирования нуклеиновых кислот, разработанного и коммерциализированного компанией “Oxford Nanopore Technologies” (“ONT”), хотя она не раскрывает многих деталей устройства и функционирования предлагаемых ею секвенаторов. Тем не менее, в патентном портфеле “ONT” находятся патенты, защищающие её права на генетически модифицированные рекомбинатные α-ГЛ [30], MspA [31] и лизинин (порин, найденный у дождевого червя *Eisenia fetida*) [32] и на использование геликаз Hel308 [33] и XPD [34] в качестве моторных белков для контролируемой транслокации нуклеиновых кислот через белковую нанопору. Также компанией запатентован метод секвенирования днДНК, который состоит в лигировании ДНК-адаптера, имеющего шпилечную структуру, к одному из концов фрагмента днДНК (который может представлять фрагмент геномной ДНК или дуплексную кДНК), что обеспечивает последовательное “чтение” обоих нитей ДНК при использовании в качестве моторного белка ДНК-полимеразы фага φ29, которая, как оказалось, при определённых условиях способна контролировать транслокацию онДНК через нанопору даже взаимодействуя только с единичной ДНК-цепью [35]. В отличие от нанопорового секвенирования только одной цепи днДНК (так называемое 1D-секвенирование, в этом случае получают так называемое “чтение матрицы” – “template read”), в случае использования такого адаптера происходит “чтение” обоих цепей (2D-секвенирование, включающее “чтение матрицы” и “чтение комплемента” – “complement read”). Это позволяет при дальнейшем анализе данных провести коррекцию ошибок распознавания нуклеотидов, которые распределены случайно в обоих “чтениях”, и, таким образом, повысить точность

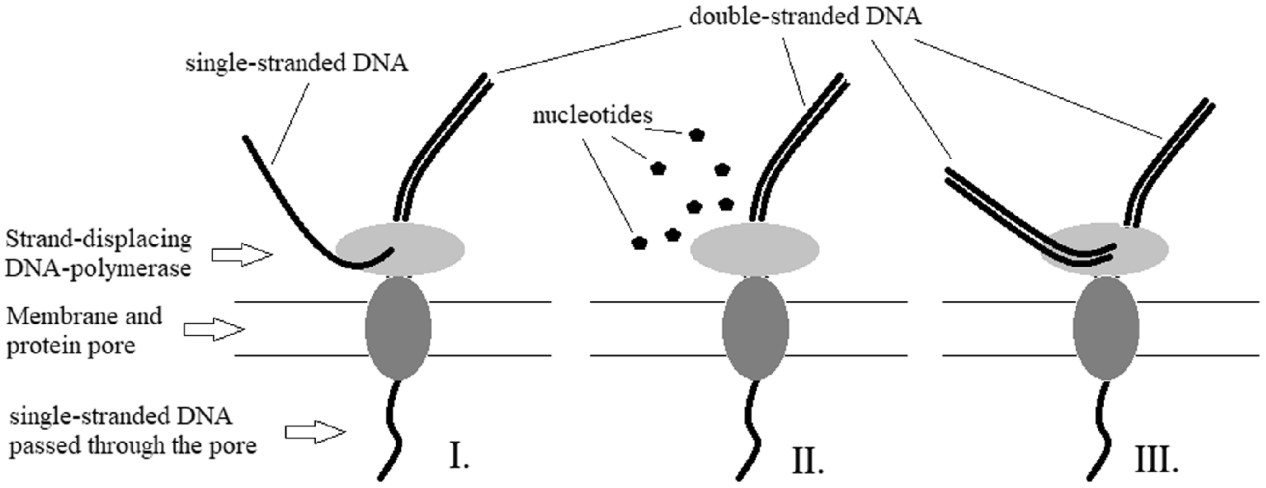

Рисунок 3. Иллюстрация различных активностей ДНК-полимеразы фага φ29 при использовании её в качестве моторного белка при нанопоровом секвенировании ДНК. I. Геликазная активность. II. 3′-5′-экзонуклеазная активность. III. Полимеразная активность.

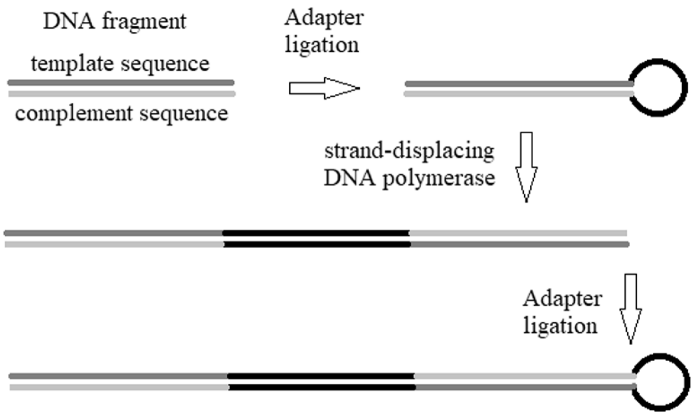

Рисунок 4. Схема конструирования двунитевой ДНК, содержащей чередующиеся последовательности "матрицы" и "комплемента".

секвенирования. Интересно, что используя вытесняющую ДНК-полимеразу, можно превратить фрагмент днДНК с адаптером в днДНК, в которой последовательности "матрицы" и "комплемента" будут повторяться дважды (рис. 4). Последующее лигирование адаптера к такой днДНК позволит "прочитать" дважды как "матрицу", так и "комплемент" исходного фрагмента днДНК, ещё более повышая точность нанопорового секвенирования [35]. Компанией "ONT" также запатентованы математический алгоритм распознавания нуклеотидов (base-calling), позволяющий конвертировать последовательность изменений силы тока во времени при трансклокации онДНК через нанопору в последовательность нуклеотидов, который основан на использовании скрытой марковской модели и алгоритма Витерби [36], а так же метод и аппарат для создания нанопоровых чипов [37] и инструментарий, необходимый для их работы (проточная микросистема, микроэлектроника и т.п.) [38, 39]. Нанопоровые чипы, разработанные "ONT", содержат от 500 до 3000 белковых пор, встроенных в искусственные амфифильные мембраны, и позволяют проводить независимое измерение силы тока, протекающего через каждую нанопору. Нанопоровый чип интегрирован в одноразовую проточную ячейку (flow cell), которая является заменяемой частью нанопорового секвенатора.

Первый коммерческий нанопоровый секвенатор MinION был предложен "ONT" в 2015 г. В настоящее время компания предлагает несколько нанопоровых секвенаторов разной производительности: MinION (производительность 10–30 млрд. нт на одну проточную ячейку в зависимости от используемого типа чипов), GridION (имеет до 5 проточных ячеек, производительность – до 150 млрд. нт за 48 ч) и PromethION (до 48 проточных ячеек, производительность – до 15 трлн. нт за 48 ч). MinION представляет портативный секвенатор (размер 10.5×2.3×3.3 см), который подключается к компьютеру через USB-разъём и рассчитан на одну проточную ячейку. Компания также объявила о разработке ещё более портативной версии нанопорового секвенатора – SmidgION, который рассчитан на работу с мобильными устройствами (в том числе таким, как смартфон).

Уже в ранних работах молекулы РНК, наряду с молекулами ДНК, использовались для изучения прохождения нуклеиновых кислот через белковые нанопоры [23], что указывало на потенциальную возможность нанопорового секвенирования не только ДНК, но и РНК. Впервые возможность нанопорового секвенирования непосредственно РНК, без необходимости получения кДНК с помощью обратной транскрипции, была продемонстрирована компанией "ONT" в 2018 г. на примере polyA+РНК, выделенной из дрожжей [40]. Практически 80% "чтений", полученных в результате прямого секвенирования РНК, картировалось на геном дрожжей, а распределение длин "чтений" было одинаково с их распределением при секвенировании кДНК. В отличие от прямого секвенирования РНК с использованием SMRT-технологии, которое так и не получило развития, прямое нанопоровое секвенирование РНК уже применяется для секвенирования геномов РНК-вирусов [41] и транскриптомного анализа [42-44].

Нанопоровое секвенирование характеризуется достаточно длинными "чтениями". Так, при использовании одного из вариантов улучшенной технологии нанопорового секвенирования (известного как R9.0), медианное значение длины "чтения" составляет более 13 тыс. нт, максимальное – превышает 150 тыс. нт [45]. Основным недостатком нанопорового секвенирования является его невысокая точность. Несмотря на быстрый прогресс (в 2015 г. точность секвенирования составляла около 60% [10], а в 2018 – уже более 90% [42]), связанный с постоянным улучшением технологии нанопорового секвенирования и метода анализа первичных данных (например, использование бактериального мембранного белка CsgG для формирования нанопор и алгоритма Albacore для base-calling, основанного на использовании рекуррентной нейронной сети [45]), она существенно уступает точности SGS. Тем не менее, за счёт своей длины, "чтения" могут быть однозначно картированы на референсный геном с точностью, превышающей 97%, что позволяет уже сегодня успешно использовать нанопоровое секвенирование для детекции патогенных микроорганизмов и профилирования микробиомов [46-49].

2. ИЗУЧЕНИЕ ТРАНСКРИПТОМА С ПОМОЩЬЮ ТЕХНОЛОГИЙ ОДНОМОЛЕКУЛЯРНОГО СЕКВЕНИРОВАНИЯ

2.1. Применение SMRT-секвенирования в транскриптомном анализе

Исследование "ширины" транскриптома (многообразия транскриптов и их изоформ) с помощью SGS-технологий сталкивается со значительными трудностями, ассоциированными с проблемой точной реконструкции полноразмерных транскриптов из коротких "чтений" [16, 50]. Так, сравнение 14 математических алгоритмов компьютеризованной "сборки" полноразмерных транскриптов из данных SGS, полученных для образцов личинок *C. elegans* и *D. melanogaster* и клеток линии HepG2, показало, что в первых двух случаях максимум 73% всех транскриптов могут быть реконструировано как полноразмерные (в среднем – 50-55%), а в последнем – максимум 61% (в среднем – 41%) [51]. Эти результаты также указывают на то, что количественные профили анализируемого транскриптома могут заметно различаться в зависимости от использованного алгоритма обработки первоначальных данных. Значительная длина "чтений" при SMRT-секвенировании (в среднем 10-20 тыс. нт) позволяет получать полноразмерные последовательности для подавляющего большинства транскриптов (например, у человека медианное значение длины транскрипта составляет около 2.5 тыс. нт [50], что делает возможным прямую детекцию их изоформ, образующихся в результате альтернативного сплайсинга, и обнаружение новых генов. Однако точность идентификации изоформ транскриптов и новых генов существенно ограничивалась высоким уровнем ошибки, характерным для SMRT-секвенирования. Для преодоления этого ограничения, Au и соавт. разработали математический алгоритм, позволяющий корректировать ошибки в длинных "чтениях", получаемых при SMRT-секвенировании транскриптома, с помощью

коротких “чтений”, получаемых при его секвенировании методами SGS [52]. Алгоритм, названный LSC, включает пять шагов: гомополимерную компрессию (замену последовательностей одного нуклеотида на единичный нуклеотид) длинных и коротких “чтений”, контроль качества коротких “чтений”, картирование коротких “чтений” на длинные, корректировку ошибок и декомпрессию последовательностей. Этот алгоритм был успешно реализован при исследовании изоформ транскриптов эмбриональных стволовых клеток человека: комбинирование и совместный анализ данных, полученных с использованием SMRT- и SGS-технологий позволили обнаружить 2103 изоформы, неаннотированных на тот момент в базе данных RefSeq или других базах данных, а также идентифицировать 273 новых гена [50]. Позднее был предложен более быстродействующий вариант LSC, названный LSCplus [53]. Для корректировки ошибок в длинных “чтениях” с помощью коротких “чтений” также были предложены алгоритмы: LoRDEC [54] (корректировка длинных “чтений” с помощью коротких, основанная на использовании графов де Брёйна); PacBioToCA, который являлся частью программного пакета Celera Assembler (в настоящее время больше не поддерживается); proovread (также основанный на картировании коротких “чтений” на длинные) [55]. Хотя точность SMRT-секвенирования существенно улучшилась за последние годы (в том числе из-за внедрения в практику секвенирования протокола Iso-Seq), комбинированный или, как его часто называют, гибридный подход, включающий совместное использование SMRT- и SGS-технологий, и сегодня широко используется в транскриптомном анализе (например, [56-62]). Его несомненным достоинством является отсутствие необходимости в референсном геноме (или в его полноте), недостатком – высокая стоимость из-за использования двух методов секвенирования.

Протокол Iso-Seq, разработанный “Pacific Biosciences” в 2014 г., получил широкое распространение с момента его первого использования в 2015 г. [63] и является сегодня доминирующем подходом к анализу транскриптомов различного происхождения с использованием технологии SMRT-секвенирования [16]. Среди экспериментальных исследований, выполненных в 2018 г. с применением этого протокола, преобладают работы, посвящённые альтернативному сплайсингу [56-60, 64, 65] и аннотации геномов [58, 61, 66-69]. Наряду с этим, SMRT-секвенирование по протоколу Iso-Seq также используется для изучения альтернативного полиаденилирования [70-72] и детекции химерных (или “слитых”) генов (fusion genes) [73]. Использование SMRTbell-библиотек в протоколе Iso-Seq и последующая обработка получаемых CLR, результатом которой является набор RoI, представляющих консенсусные последовательности, существенно повышает конечную точность секвенирования. Так, SMRT-анализ по протоколу Iso-Seq транскриптома проростков сорго показал при сравнении с референсным геномом, что вероятность ошибки в расчёте на один нуклеотид составляет 2.34% и распределена следующим образом: однонуклеотидные замены – 0.64%, вставки – 1.07%, делеции – 0.63% [71]. Это позволяет при наличии референсного генома провести анализ “ширины” транскриптома SMRT-секвенированием без привлечения методов SGS. Алгоритм анализа, получивший название TAPIS (Transcriptome Analysis Pipeline for Isoform Sequencing), представляет собой итеративный процесс, который чередует картирование “чтений” на референсный геном и коррекцию ошибок [71]. Для картирования авторы использовали программу GMAP (Genome Mapping and Alignment Program) [74]. Анализ данных SMRT-секвенирования проростков сорго с использованием TAPIS позволил обнаружить более 11 тыс. новых сплайс-изоформ, сайты альтернативного полиаденилирования в приблизительно 11 тыс. экспрессирующихся генах и более 2100 новых генов [71]. Более того, существенное снижение ошибки SMRT-секвенирования при использовании протокола Iso-Seq позволило предложить алгоритм, получивший название ToFU (Transcript isOforms: Full-length and Unassembled), для анализа транскриптома *de novo* (то есть без использования референсного генома) без необходимости в коротких “чтениях” [75]. Алгоритм основан на анализе RoI и кластеризации их таким образом, чтобы получить набор консенсусных последовательностей, который и представляет набор транскриптов. Использование алгоритма ToFU для исследования транскриптома грибов вида *Plicaturopsis crisp*а позволило идентифицировать почти 23 тыс. различных изоформ, представляющих около 9 тыс. транскрибируемых локусов. Сопоставление полученных последовательностей транскриптов с ранее аннотированным геномом *Plicaturopsis crispa* с помощью программы GMAP показало, что точность секвенирования составляет 0.22% (0.06% – замены, 0.04% – вставки, 0.12% – делеции) [75]. Следует однако отметить, что значительное число изоформ транскриптов, предсказываемых ToFU, может являться артефактами секвенирования, как показала их верификация методом количественной ПЦР при машинном обучении классификатора SQANTI (Structural and Quality Annotation of Novel Transcript Isoforms), предназначенного для отсеивания “артефактных транскриптов” при анализе данных SMRT-секвенирования [76].

В то время как SMRT-секвенирование *per se* или в сочетании с SGS широко используется для исследования альтернативного сплайсинга, альтернативного полиаденилирования, аннотации геномов и секвенирования транскриптомов *de novo*, оно практически не использовалось для количественного профилирования транскриптома. В работах, где транскриптомный анализ проводился с использованием гибридного секвенирования, сочетающего технологии SMRT и SGS, для количественного профилирования транскриптома использовались короткие “чтения”, генерируемые SGS [62, 77-81]. Одна из причин того, что SMRT-секвенирование не применяется для количественного профилирования, состоит в использовании протокола Iso-Seq, который предполагает фракционирование кДНК по размеру и последующую раздельную амплификацию фракций, что может существенно исказить оценку количества транскриптов по количеству RoI. Однако отказ от использования этого шага в протоколе Iso-Seq приведёт к предпочтительному секвенированию более коротких фрагментов [16]. Тем не менее, оценка дифференциальной экспрессии генов с помощью SMRT-секвенирования без использования шага фракционирования в протоколе Iso-Seq, по-видимому, в принципе возможна. При анализе транскриптомов вируса герпеса животных на разных стадиях литического цикла [82] и эндогенных вирусов человека при ВИЧ-инфекции [83], дифференциальную экспрессию генов вирусов оценивали используя количество RoI, полученных после обработки первоначальных данных SMRT-секвенирования. При этом фракционирование кДНК в протоколе Iso-Seq не проводилось [82, 83]. Хотя это может завышать количество более коротких транскриптов, авторы, очевидно, исходили из предположения, что оно завышено в одинаковой пропорции в секвенируемых образцах. Насколько правомочен такой подход к дифференциальному профилированию транскриптомов, чей размер существенно превосходит размер транскриптома вирусов, в настоящее время не известно. Возможно, добавление в секвенируемые образцы “калибрантов” – молекул РНК с уникальной последовательностью и известной концентрацией (подобных

тем, что разрабатываются в рамках проекта ERCC, "External RNA Controls Consortium" [84]) позволит проводить определение как относительного, так и абсолютного количество транскриптов по количеству RoI. При наличии линейки таких калибрантов, различающихся длиной молекул РНК, количественное профилирование транскриптомов с помощью SMRT-секвенирования может стать возможным как в абсолютных, так и относительных величинах даже в рамках протокола Iso-Seq. В любом случае, использование SMRT-секвенирования для количественного профилирования транскриптомов эукариот потребует достижения им производительности, не уступающей производительности SGS, чтобы обеспечить надёжную количественную детекцию низкокопийных транскриптов.

2.2. Анализ транскриптома с использованием нанопорового секвенирования

Как и в случае с SMRT-секвенированием, использование нанопорового секвенирования в транскриптомном анализе было в первую очередь сфокусировано на идентификации изоформ транскриптов. Первое такое исследование было посвящено характеризации экспрессии генов со сложной экзон-интронной организацией, а именно генов *Rdl*, *MRP*, *Mhc* и *Dscam1* дрозофилы [85], которые могут кодировать сотни и тысячи изоформ благодаря альтернативному сплайсингу. Авторы получали библиотеки полноразмерных кДНК транскриптов этих генов с помощью оптимизированной процедуры обратной транскрипции с последующей ПЦР со сменой матрицы (template-switching polymerase chain reaction, twPCR), которая обеспечивает избирательную амплификацию молекул кДНК, соответствующих полноразмерным транскриптам (full-length cDNA, FL-cDNA). Секвенирование FL-cDNA-библиотек на MinION и картирование полученных "чтений" на референсные последовательности с использованием алгоритма для выравнивания LAST [86] позволило обнаружить 7874 различных изоформ транскрипта гена *Dscam1* (*Dscam1* содержит 115 экзонов, 95 из которых могут участвовать в альтернативном сплайсинге и потенциально генерировать 30016 изоформ), а для генов *Rdl*, *MRP* и *Mhc* – 301, 337 и 112 изоформ, соответственно [85]. Аналогичный подход был также использован для анализа альтернативного сплайсинга гена *BRCA1* человека с использованием мРНК из клеток лимфобластоидной линии [87]. FL-cDNA-библиотека была секвенирована на MinION и полученные "чтения" картированы на последовательность гена с помощью программы GMAP. В результате было обнаружено 32 изоформы транскрипта гена *BRCA1* человека, из которых 20 изоформ не были описаны ранее. Нанопоровое секвенирование было также использовано для изучения транскриптома бакуловируса *Autographa californica multiple nucleopolyhedrovirus*, (AcMNPV) с использованием комбинации секвенирования кДНК-библиотеки и прямого секвенирования РНК [44]. Это позволило выявить 5 новых сплайс-изоформ известных транскриптов и 132 ранее не описанных транскрипта. Интересно, что секвенирование кДНК-библиотеки дало 324677 "чтений" (средняя длина 1053 нт), из которых 103133 были картированы на геном AcMNPV (используя программу GMAP; остальные "чтения" представляли транскрипты клетки-хозяина), в то время как прямое секвенирование РНК дало только 6482 "чтения" (средняя длина 614 нт; из них 2430 были картированы на вирусный геном) [44]. В ноябре 2018 г. была опубликована работа Seki и соавт. [42], в которой был впервые выполнен анализ транскриптома клеток человека (клеточная линия аденокарциномы лёгких LC2/ad) с использованием нанопорового секвенирования и идентификации изоформ транскриптов. В результате секвенирования FL-cDNA-библиотеки было получено 532956 "чтений", которые были картированы на транскрипты в базе данных RefSeq с использованием двух алгоритмов для выравнивания – LAST и BWA [88]. Как отметили авторы работы, использование LAST более предпочтительно, так как он позволил картировать большую долю "чтений" на референсный транскриптом (38.3% vs. 33.9%) с покрытием транскриптов "чтениями" 0.8 и более. В результате было идентифицировано 6018 изоформ транскриптов, из которых 158 – ранее не аннотированных. Кроме того, проведённый анализ позволил выявить 151 химерный ген [42].

Аналогично SMRT-секвенированию, нанопоровое секвенирование используется в гибридном секвенировании как технология, "комплементарная" SGS [89-92]. Применительно к анализу транскриптома, корректировка длинных "чтений" с помощью коротких (с использованием программы proovread) позволила снизить ошибку нанопорового секвенирования с 12% до 0-2% [93]. Нанопоровое секвенирование также было применено для секвенирования транскриптома *de novo* [94], для чего авторами был разработан алгоритм анализа длинных "чтений", основанный на их кластеризации на основе схожести без обращения к референсному геному или транскриптому, получивший название CARNAC-LR (Clustering coefficient-based Acquisition of RNA Communities in Long Reads). Кластеризация длинных "чтений", полученных в результате нанопорового секвенирования РНК из мозга мышей (всего 1 256 967 'чтений") и сравнение полученных результатов с результатами картирования этих "чтений" на референсный геном (с помощью программы BLAT [95]) показало приблизительно 80% совпадение, что указывает на сложность сборки больших транскриптомов *de novo* из данных нанопорового секвенирования без корректировки ошибок секвенирования с помощью референсного генома или транскриптома. Вероятно, это связано с существенной долей однонуклеотидных замен, вставок и делеций в "чтениях", которые, в отличие от SMRT-секвенирования, не могут быть корректированы при получении консенсусной последовательности из многих "субчтений". В случае нанопорового секвенирования, максимально возможное количество "субчтений" в настоящее время не превышает двух (2D- или $1D^2$-секвенирование).

В то время как SMRT-технология характеризуется избирательностью, заключающейся в предпочтительном секвенировании более коротких последовательностей ДНК (и, соответственно, необходимостью использования протокола Iso-Seq для того, чтобы уменьшить эту избирательность), нанопоровое секвенирование, по-видимому, подобной избирательности не имеет. В недавней работе Oikonomopoulos и соавт. [96], нанопоровое секвенирование FL-cDNA-библиотек, полученных для смесей 92 полиаденилированных транскриптов (отобранных в рамках проекта ERCC [84]) с заданной концентрацией, показало, что количество "чтений" хорошо согласуется с ожидаемой концентрацией соответствующего транскрипта (коэффициент корреляции Пирсона r_p равнялся 0.98) и не зависит от его длины или GC-состава молекул кДНК. Интересно, что в этом случае использование алгоритма для выравнивания LAST также давало наилучшее соответствие между количеством "чтений" и ожидаемым количеством транскрипта в сравнении с такими алгоритмами, как Margin-Align [97], BWA, BLASR [98], BLAST [99] и Smith-Waterman [100].

Отсутствие избирательности нанопорового секвенирования по длине и GC-составу транскриптов делает принципиально возможным оценку их относительного и/или абсолютного количества. В качестве меры количества каждого транскрипта используют число

“чтений”, картируемых на его последовательность [42, 43], что требует, соответственно, наличия референсного транскриптома или генома. Подобный подход был реализован при исследовании дифференциальной экспрессии генов дрожжей *Saccharomyces cerevisiae* [43]. Используя прямое нанопоровое секвенирование РНК, было получено ~530 и ~620 тыс. “чтений” с медианной длиной 1150 и 1263 нт, соответственно, для полиаденилированной РНК дрожжевых клеток, культивируемых в различных условиях (при избытке глюкозы или в присутствии этанола). “Чтения” были картированы на геном *S. cerevisiae* с помощью алгоритма выравнивания GraphMAP, что позволило идентифицировать 5433 различных транскриптов (91% от их ожидаемого числа). Алгоритм GraphMAP [101] был специально разработан для картирования на референсный геном длинных “чтений”, получаемых при нанопоровом секвенировании и характеризующихся относительно высокой ошибкой секвенирования (в данном случае она составляла 12% [43]). Отношение количества “чтений”, приходящихся на каждый транскрипт при различных условиях культивирования, было использовано для анализа дифференциальной экспрессии генов [43]. В другой публикации [42] нанопоровое секвенирование FL-cDNA-библиотеки, полученной для полиаденилированной РНК культивируемых клеток аденокарциномы лёгкого человека, было использовано для количественного профилирования их транскриптома. В качестве метрики авторы использовали RPM (Reads Per Million; доля “чтений”, картированных на данный транскрипт, от их общего количества, умноженная на миллион). Количественный профиль транскриптома, полученный с использованием нанопорового секвенирования, показал хорошее соответствие с профилем, полученным секвенированием с помощью SGS (r_p=0.91) [42]. При этом для SGS в качестве метрики использовали TPM (Transcripts Per Million mapped reads). TPM, наряду с RPKM и FPKM, часто используется для количественной оценки результатов SGS и рассматривается как метрика, устраняющая сдвиг в оценке количества транскриптов, возможный при использовании RPKM- и FPKM-метрик [102]. Следует отметить, что валидация количественного профилирования транскриптома культивируемых клеток аденокарциномы человека методом ОТ-ПЦР (обратная транскрипция и ПЦР в реальном времени) на выборке из 44 транскриптов показала хорошее соответствие с результатами нанопорового секвенирования (r_p=0.82) [42]. Таким образом, нанопоровое секвенирование позволяет проводить количественное профилирование транскриптомов эукариот, включая клетки человека. Очевидно, что добавление в секвенируемые образцы “РНК-калибрантов” сделает возможным наряду с оценкой относительного количества транскриптов также и определение их абсолютного количества.

ЗАКЛЮЧЕНИЕ

Одномолекулярное секвенирование с использованием технологий SMRT- и нанопорового секвенирования существенно расширяет возможности исследования многообразия изоформ транскриптов, образующихся в результате альтернативного сплайсинга и/или полиаденилирования, секвенирования транскриптомов *de novo*, аннотации геномов и идентификации химерных генов, особенно при сочетании с SGS. Применительно к количественному профилированию транскриптома, SMRT-технология имеет очевидные ограничения, связанные с предпочтительным секвенированием коротких фрагментов кДНК и с неопределённостями в оценке количества транскриптов, вносимыми необходимостью фракционирования кДНК по размеру в рамках протокола Iso-Seq. Напротив, нанопоровое секвенирование не имеет таких ограничений и может быть использовано для количественного профилирования транкриптома, особенно в формате прямого секвенирования РНК, в котором нанопоровый секвенатор может выступать в качестве молекулярного счётчика, позволяющего идентифицировать транскрипт с высокой селективностью и оценить его относительную представленность в транскриптоме. При использовании РНК-калибрантов – молекул РНК с уникальными последовательностями и известной концентрацией – нанопоровое секвенирование позволит определять абсолютное количество транскрипта каждого вида в анализируемых образцах.

БЛАГОДАРНОСТИ

Работа выполнена в рамках Программы фундаментальных научных исследований государственных академий наук на 2013–2020 годы.

ЛИТЕРАТУРА

1. Green, E.D., Watson, J.D., & Collins, F.S. (2015) Human Genome Project: Twenty-five years of big biology. Nature, 526(7571), 29-31. DOI: 10.1038/526029a

2. Malone, J.H., & Oliver, B. (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology, 9, 34. DOI: 10.1186/1741-7007-9-34

3. Costa-Silva, J., Domingues, D., & Lopes, F.M. (2017) RNA-Seq differential expression analysis: An extended review and a software tool. PloS One, 12(12), e0190152. DOI: 10.1371/journal.pone.0190152

4. Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016) High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J. Microbiol., 56(4), 394-404. DOI: 10.1007/s12088-016-0606-4

5. Morey, M., Fernandez-Marmiesse, A., Castineiras, D., Fraga, J.M., Couce, M.L., & Cocho, J.A. (2013) A glimpse into past, present, and future DNA sequencing. Mol. Genet. Metab., 110(1-2), 3-24. DOI: 10.1016/j.ymgme.2013.04.024

6. Choi, S.C. (2016) On the study of microbial transcriptomes using second- and third-generation sequencing technologies. J. Microbiol., 54(8), 527-536. DOI: 10.1007/s12275-016-6233-2

7. Pillai, S., Gopalan, V., & Lam, A. K. (2017) Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. Crit. Revs. Oncology/Hematology, 116, 58-67. DOI: 10.1016/j.critrevonc.2017.05.005

8. Schadt, E.E., Turner, S., & Kasarskis, A. (2010) A window into third-generation sequencing. Human Molecular Genetics, 19(R2), R227-R240. DOI: 10.1093/hmg/ddq416

9. Ashton, P.M., Nair, S., Dallman, T., Rubino, S., Rabsch, W., Mwaigwisya, S., Wain, J., & O'Grady, J. (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nature Biotechnol., 33(3), 296-300. DOI: 10.1038/nbt.3103

10. Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection Quantification, 3, 1-8. DOI: 10.1016/j.bdq.2015.02.001

11. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., & Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), 133-138. DOI: 10.1126/science.1162986

12. Korlach, J., Bjornson, K.P., Chaudhuri, B.P., Cicero, R.L., Flusberg, B.A., Gray, J.J., Holden, D., Saxena, R., Wegener, J., & Turner, S.W. (2010) Real-time DNA sequencing from single

polymerase molecules. Met. Enzymol., 472, 431-455. DOI: 10.1016/S0076-6879(10)72001-2

13. Beckett, D., Kovaleva, E., & Schatz, P.J. (1999) A minimal peptide substrate in biotin holoenzyme synthetase-catalyzed biotinylation. Protein Sci., 8(4), 921-929. DOI: 10.1110/ps.8.4.921

14. Lundquist, P.M., Zhong, C.F., Zhao, P., Tomaney, A.B., Peluso, P.S., Dixon, J., Bettman, B., Lacroix, Y., Kwo, D.P., McCullough, E., Maxham, M., Hester, K., McNitt, P., Grey, D. M., Henriquez, C., Foquet, M., Turner, S.W., & Zaccarin, D. (2008) Parallel confocal detection of single molecules in real time. Optics Letts., 33(9), 1026-1028

15. Travers, K.J., Chin, C.S., Rank, D.R., Eid, J.S., & Turner, S.W. (2010) A flexible and efficient template format for circular consensus sequencing and SNP detection. Nucl. Acids Res., 38(15), e159. DOI: 10.1093/nar/gkq543

16. An, D., Cao, H.X., Li, C., Humbeck, K., & Wang, W. (2018) Isoform Sequencing and State-of-Art Applications for Unravelling Complexity of Plant Transcriptomes. Genes, 9(1). DOI: 10.3390/genes9010043

17. https://github.com/PacificBiosciences/IsoSeq_SA3nUP, last accessed on December 1, 2018.

18. Nakano, K., Shiroma, A., Shimoji, M., Tamotsu, H., Ashimine, N., Ohki, S., Shinzato, M., Minami, M., Nakanishi, T., Teruya, K., Satou, K., & Hirano, T. (2017) Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Human Cell, 30(3), 149-161. DOI: 10.1007/s13577-017-0168-8

19. Mardis, E.R. (2013) Next-generation sequencing platforms. Ann. Rev. Anal. Chem., 6, 287-303. DOI: 10.1146/annurev-anchem-062012-092628

20. Rhoads, A., & Au, K.F. (2015) PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics, 13(5), 278-289. DOI: 10.1016/j.gpb.2015.08.002

21. Vilfan, I.D., Tsai, Y.C., Clark, T.A., Wegener, J., Dai, Q., Yi, C., Pan, T., Turner, S. W., & Korlach, J. (2013) Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. J. Nanobiotechnol., 11, 8. DOI: 10.1186/1477-3155-11-8

22. Kasianowicz, J.J., Brandin, E., Branton, D., & Deamer, D.W. (1996) Characterization of individual polynucleotide molecules using a membrane channel. Proc. Natl. Acad. Sci. USA, 93(24), 13770-13773.

23. Akeson, M., Branton, D., Kasianowicz, J.J., Brandin, E., & Deamer, D.W. (1999). Microsecond time-scale discrimination among polycytidylic acid, polyadenylic acid, and polyuridylic acid as homopolymers or as segments within single RNA molecules. Biophys. J., 77(6), 3227-3233. DOI: 10.1016/S0006-3495(99)77153-5

24. Bayley, H. (2015) Nanopore sequencing: from imagination to reality. Clin. Chem., 61(1), 25-31. DOI: 10.1373/clinchem.2014.223016

25. Lieberman, K.R., Cherf, G.M., Doody, M.J., Olasagasti, F., Kolodji, Y., & Akeson, M. (2010) Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. J. Am. Chem. Soc., 132(50), 17961-17972. DOI: 10.1021/ja1087612

26. Cherf, G.M., Lieberman, K.R., Rashid, H., Lam, C.E., Karplus, K., & Akeson, M. (2012) Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. Nat. Biotechnol., 30(4), 344-348. DOI: 10.1038/nbt.2147

27. Manrao, E.A., Derrington, I.M., Pavlenok, M., Niederweis, M., & Gundlach, J.H. (2011) Nucleotide discrimination with DNA immobilized in the MspA nanopore. PloS One, 6(10), e25723. DOI: 10.1371/journal.pone.0025723

28. Manrao, E.A., Derrington, I.M., Laszlo, A.H., Langford, K.W., Hopper, M.K., Gillgren, N., Pavlenok, M., Niederweis, M., & Gundlach, J.H. (2012) Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. Nat. Biotechnol., 30(4), 349-353. DOI: 10.1038/nbt.2171

29. Laszlo, A.H., Derrington, I.M., Ross, B.C., Brinkerhoff, H., Adey, A., Nova, I.C., Craig, J.M., Langford, K.W., Samson, J.M., Daza, R., Doering, K., Shendure, J., & Gundlach, J.H. (2014) Decoding long nanopore sequencing reads of natural DNA. Nat. Biotechnol., 32(8), 829-833. DOI: 10.1038/nbt.2950

30. US Patent N 9447152 B2.

31. US Patent N 9751915 B2.

32. US Patent Application N 2015/0068904 A1.

33. US Patent N 9758823 B2.

34. US Patent Application N 2015/0065354 A1.

35. US Patent Application N 2015/0152492 A1.

36. US Patent Application N 2016/0162634 A1.

37. US Patent Application N 2015/0014160 A1.

38. US Patent N 10036065 B2.

39. US Patent N 9651519 B2.

40. Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E.J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Brocklebank, D., Juul, S., Clarke, J., Heron, A.J., & Turner, D.J. (2018) Highly parallel direct RNA sequencing on an array of nanopores. Nature Methods, 15(3), 201-206. DOI: 10.1038/nmeth.4577

41. Keller, M.W., Rambo-Martin, B.L., Wilson, M.M., Ridenour, C.A., Shepard, S.S., Stark, T.J., Neuhaus, E.B., Dugan, V.G., Wentworth, D.E., & Barnes, J.R. (2018) Direct RNA Sequencing of the Coding Complete Influenza A Virus Genome. Sci. Rep., 8(1), 14408. DOI: 10.1038/s41598-018-32615-8

42. Seki, M., Katsumata, E., Suzuki, A., Sereewattanawoot, S., Sakamoto, Y., Mizushima-Sugano, J., Sugano, S., Kohno, T., Frith, M.C., Tsuchihara, K., & Suzuki, Y. (2018) Evaluation and application of RNA-Seq by MinION. DNA Research, DOI: 10.1093/dnares/dsy038

43. Jenjaroenpun, P., Wongsurawat, T., Pereira, R., Patumcharoenpol, P., Ussery, D.W., Nielsen, J., & Nookaew, I. (2018) Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of *Saccharomyces cerevisiae* CEN.PK113-7D. Nucl. Acids Res., 46(7), e38. DOI: 10.1093/nar/gky014

44. Moldovan, N., Tombacz, D., Szucs, A., Csabai, Z., Balazs, Z., Kis, E., Molnar, J., & Boldogkoi, Z. (2018) Third-generation Sequencing Reveals Extensive Polycistronism and Transcriptional Overlapping in a Baculovirus. Sci. Rep., 8(1), 8604. DOI: 10.1038/s41598-018-26955-8

45. Jain, M., Tyson, J.R., Loose, M., Ip, C.L.C., Eccles, D.A., O'Grady, J., Malla, S., Leggett, R.M., Wallerman, O., Jansen, H.J., Zalunin, V., Birney, E., Brown, B.L., Snutch, T.P., Olsen, H. E., Min, I.O.N.A., & Reference, C. (2017) MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9.0 chemistry. F1000Research, 6, 760. DOI: 10.12688/f1000research.11354.1

46. Butt, S.L., Taylor, T.L., Volkening, J.D., Dimitrov, K.M., Williams-Coplin, D., Lahmers, K.K., Miller, P.J., Rana, A.M., Suarez, D.L., Afonso, C.L., & Stanton, J.B. (2018) Rapid virulence prediction and identification of Newcastle disease virus genotypes using third-generation sequencing. Virology J., 15(1), 179. DOI: 10.1186/s12985-018-1077-5

47. Rames, E., & Macdonald, J. (2018) Evaluation of MinION nanopore sequencing for rapid enterovirus genotyping. Virus Res., 252, 8-12. DOI: 10.1016/j.virusres.2018.05.010

48. Li, C., Chng, K.R., Boey, E.J., Ng, A.H., Wilm, A., & Nagarajan, N. (2016) INC-Seq: accurate single molecule reads using nanopore sequencing. GigaScience, 5(1), 34. DOI: 10.1186/s13742-016-0140-7

49. Batovska, J., Lynch, S.E., Rodoni, B.C., Sawbridge, T.I., & Cogan, N.O. (2017). Metagenomic arbovirus detection using MinION nanopore sequencing. J. Virol. Methods, 249, 79-84. DOI: 10.1016/j.jviromet.2017.08.019

50. Au, K.F., Sebastiano, V., Afshar, P.T., Durruthy, J.D., Lee, L., Williams, B.A., van Bakel, H., Schadt, E.E., Reijo-Pera, R.A., Underwood, J.G., & Wong, W.H. (2013) Characterization of the human ESC transcriptome by hybrid sequencing. Proc. Natl. Acad. Sci. USA, 110(50), E4821-4830. DOI: 10.1073/pnas.1320101110

51. Steijger, T., Abril, J.F., Engstrom, P.G., Kokocinski, F., RGASP Consortium, Hubbard, T.J., Guigo, R., Harrow, J., & Bertone, P. (2013) Assessment of transcript reconstruction methods for RNA-seq. Nature Methods, 10(12), 1177-1184. DOI: 10.1038/nmeth.2714

52. Au, K.F., Underwood, J.G., Lee, L., & Wong, W.H. (2012) Improving PacBio long read accuracy by short read alignment. PloS One, 7(10), e46679. DOI: 10.1371/journal.pone.0046679

53. Hu, R., Sun, G., & Sun, X. (2016). LSCplus: a fast solution for improving long read accuracy by short read alignment. BMC Bioinformatics, 17(1), 451. DOI: 10.1186/s12859-016-1316-y.

54. Salmela, L., & Rivals, E. (2014) LoRDEC: accurate and efficient long read error correction. Bioinformatics, 30(24), 3506-3514. DOI: 10.1093/bioinformatics/btu538

55. Hackl, T., Hedrich, R., Schultz, J., & Forster, F. (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics, 30(21), 3004-3011. DOI: 10.1093/bioinformatics/btu392

56. Chao, Q., Gao, Z.F., Zhang, D., Zhao, B.G., Dong, F.Q., Fu, C.X., Liu, L.J., & Wang, B.C. (2018) The developmental dynamics of the *Populus* stem transcriptome. Plant Biotechnol. J., DOI: 10.1111/pbi.12958

57. Filichkin, S.A., Hamilton, M., Dharmawardhana, P.D., Singh, S.K., Sullivan, C., Ben-Hur, A., Reddy, A.S.N., & Jaiswal, P. (2018) Abiotic Stresses Modulate Landscape of Poplar Transcriptome via Alternative Splicing, Differential Intron Retention, and Isoform Ratio Switching. Frontiers in Plant Science, 9, 5. DOI: 10.3389/fpls.2018.00005

58. Piriyapongsa, J., Kaewprommal, P., Vaiwsri, S., Anuntakarun, S., Wirojsirasak, W., Punpee, P., Klomsa-Ard, P., Shaw, P.J., Pootakham, W., Yoocha, T., Sangsrakru, D., Tangphatsornruang, S., Tongsima, S., & Tragoonrung, S. (2018) Uncovering full-length transcript isoforms of sugarcane cultivar Khon Kaen 3 using single-molecule long-read sequencing. PeerJ, 6, e5818. DOI: 10.7717/peerj.5818

59. Zhang, G., Sun, M., Wang, J., Lei, M., Li, C., Zhao, D., Huang, J., Li, W., Li, S., Li, J., Yang, J., Luo, Y., Hu, S., & Zhang, B. (2018) PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J., DOI: 10.1111/tpj.14120

60. Zhu, J., Wang, X., Xu, Q., Zhao, S., Tai, Y., & Wei, C. (2018) Global dissection of alternative splicing uncovers transcriptional diversity in tissues and associates with the flavonoid pathway in tea plant (*Camellia sinensis*). BMC Plant Biology, 18(1), 266. DOI: 10.1186/s12870-018-1497-9

61. Kim, J.Y., Lim, H.Y., Shin, S.E., Cha, H.K., Seo, J.H., Kim, S.K., Park, S.H., & Son, G.H. (2018) Comprehensive transcriptome analysis of *Sarcophaga peregrina*, a forensically important fly species. Scientific Data, 5, 180220. DOI: 10.1038/sdata.2018.220

62. Zhu, C., Li, X., & Zheng, J. (2018) Transcriptome profiling using Illumina- and SMRT-based RNA-seq of hot pepper for in-depth understanding of genes involved in CMV infection. Gene, 666, 123-133. DOI: 10.1016/j.gene.2018.05.004

63. Singh, N., Sahu, D.K., Chowdhry, R., Mishra, A., Goel, M.M., Faheem, M., Srivastava, C., Ojha, B.K., Gupta, D K., & Kant, R. (2016) IsoSeq analysis and functional annotation of the infratentorial ependymoma tumor tissue on PacBio RSII platform. Meta Gene, 7, 70-75. DOI: 10.1016/j.mgene.2015.11.004

64. Sahlin, K., Tomaszkiewicz, M., Makova, K.D., & Medvedev, P. (2018) Deciphering highly similar multigene family transcripts from Iso-Seq data with IsoCon. Nature Commun., 9(1), 4601. DOI: 10.1038/s41467-018-06910-x

65. Balazs, Z., Tombacz, D., Szucs, A., Snyder, M., & Boldogkoi, Z. (2018) Dual Platform Long-Read RNA-Sequencing Dataset of the Human Cytomegalovirus Lytic Transcriptome. Frontiers Genetics, 9, 432. DOI: 10.3389/fgene.2018.00432

66. Workman, R.E., Myrka, A.M., Wong, G.W., Tseng, E., Welch, K.C., Jr., & Timp, W. (2018) Single-molecule, full-length transcript sequencing provides insight into the extreme metabolism of the ruby-throated hummingbird *Archilochus colubris*. GigaScience, 7(3), 1-12. DOI: 10.1093/gigascience/giy009

67. Yi, S., Zhou, X., Li, J., Zhang, M., & Luo, S. (2018) Full-length transcriptome of *Misgurnus anguillicaudatus* provides insights into evolution of genus *Misgurnus*. Scientific Reports, 8(1), 11699. DOI: 10.1038/s41598-018-29991-6

68. Nudelman, G., Frasca, A., Kent, B., Sadler, K.C., Sealfon, S.C., Walsh, M.J., & Zaslavsky, E. (2018) High resolution annotation of zebrafish transcriptome using long-read sequencing. Genome Res., 28(9), 1415-1425. DOI: 10.1101/gr.223586.117

69. Dong, L., Liu, H., Zhang, J., Yang, S., Kong, G., Chu, J.S., Chen, N., & Wang, D. (2015) Single-molecule real-time transcript sequencing facilitates common wheat genome annotation and grain transcriptome research. BMC Genomics, 16, 1039. DOI: 10.1186/s12864-015-2257-y

70. Wang, T., Wang, H., Cai, D., Gao, Y., Zhang, H., Wang, Y., Lin, C., Ma, L., & Gu, L. (2017) Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (*Phyllostachys edulis*). Plant J. , 91(4), 684-699. DOI: 10.1111/tpj.13597

71. Abdel-Ghany, S.E., Hamilton, M., Jacobi, J.L., Ngam, P., Devitt, N., Schilkey, F., Ben-Hur, A., & Reddy, A.S. (2016) A survey of the sorghum transcriptome using single-molecule long reads. Nature Commun., 7, 11706. DOI: 10.1038/ncomms11706

72. Zhang, S. J., Wang, C., Yan, S., Fu, A., Luan, X., Li, Y., Sunny Shen, Q., Zhong, X., Chen, J. Y., Wang, X., Chin-Ming Tan, B., He, A., & Li, C.Y. (2017) Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. Mol. Biol. Evol., 34(10), 2453-2468. DOI: 10.1093/molbev/msx212

73. Liu, X., Mei, W., Soltis, P.S., Soltis, D.E., & Barbazuk, W.B. (2017). Detecting alternatively spliced transcript isoforms from single-molecule long-read sequences without a reference genome. Molecular Ecology Resources, 17(6), 1243-1256. DOI: 10.1111/1755-0998.12670

74. Wu, T.D., & Watanabe, C.K. (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics, 21(9), 1859-1875. DOI: 10.1093/bioinformatics/bti310

75. Gordon, S.P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I.V., Figueroa, M., Schilling, J.S., Chen, F., & Wang, Z. (2015) Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PloS One, 10(7), e0132628. DOI: 10.1371/journal.pone.0132628

76. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., Edelmann, M., Ezkurdia, I., Vazquez, J., Tress, M., Mortazavi, A., Martens, L., Rodriguez-Navarro, S., Moreno-Manzano, V., & Conesa, A. (2018) SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res., DOI: 10.1101/gr.222976.117

77. Deng, Y., Zheng, H., Yan, Z., Liao, D., Li, C., Zhou, J., & Liao, H. (2018) Full-Length Transcriptome Survey and Expression Analysis of *Cassia obtusifolia* to Discover Putative Genes Related to Aurantio-Obtusin Biosynthesis, Seed Formation and Development, and Stress Response. Int. J. Mol. Sci., 19(9), DOI: 10.3390/ijms19092476

78. Ren, P., Meng, Y., Li, B., Ma, X., Si, E., Lai, Y., Wang, J., Yao, L., Yang, K., Shang, X., & Wang, H. (2018) Molecular Mechanisms of Acclimatization to Phosphorus Starvation and Recovery Underlying Full-Length Transcriptome Profiling in Barley (*Hordeum vulgare* L.). Frontiers Plant Sci., 9, 500. DOI: 10.3389/fpls.2018.00500

79. Xu, Z., Peters, R. J., Weirather, J., Luo, H., Liao, B., Zhang, X., Zhu, Y., Ji, A., Zhang, B., Hu, S., Au, K. F., Song, J., & Chen, S. (2015) Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis. Plant J., 82(6), 951-961. DOI: 10.1111/tpj.12865

80. Cao, H., Lai, Y., Bougouffa, S., Xu, Z., & Yan, A. (2017) Comparative genome and transcriptome analysis reveals distinctive surface characteristics and unique physiological potentials of *Pseudomonas aeruginosa* ATCC 27853. BMC Genomics, 18(1), 459. DOI: 10.1186/s12864-017-3842-z

81. Ning, G., Cheng, X., Luo, P., Liang, F., Wang, Z., Yu, G., Li, X., Wang, D., & Bao, M. (2017) Hybrid sequencing and map finding (HySeMaFi): optional strategies for extensively deciphering gene splicing and expression in organisms without reference genome. Sci. Rep., 7, 43793. DOI: 10.1038/srep43793

82. Tombacz, D., Balazs, Z., Csabai, Z., Moldovan, N., Szucs, A., Sharon, D., Snyder, M., & Boldogkoi, Z. (2017) Characterization of the Dynamic Transcriptome of a Herpesvirus with Long-read Single Molecule Real-Time Sequencing. Sci. Rep., 7, 43751. DOI: 10.1038/srep43751

83. Young, G.R., Terry, S.N., Manganaro, L., Cuesta-Dominguez, A., Deikus, G., Bernal-Rubio, D., Campisi, L., Fernandez-Sesma, A., Sebra, R., Simon, V., & Mulder, L.C.F. (2018) HIV-1 Infection of Primary CD4(+) T Cells Regulates the Expression of Specific Human Endogenous Retrovirus HERV-K (HML-2) Elements. J. Virol., 92(1). DOI: 10.1128/JVI.01507-17

84. Lee, H., Pine, P. S., McDaniel, J., Salit, M., & Oliver, B. (2016) External RNA Controls Consortium Beta Version Update. J. Genomics, 4, 19-22. DOI: 10.7150/jgen.16082

85. Bolisetty, M.T., Rajadinakaran, G., & Graveley, B.R. (2015) Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biol., 16, 204. DOI: 10.1186/s13059-015-0777-z.

86. Frith, M.C., Hamada, M., & Horton, P. (2010) Parameters for accurate genome alignment. BMC Bioinformatics, 11, 80. DOI: 10.1186/1471-2105-11-80

87. de Jong, L.C., Cree, S., Lattimore, V., Wiggins, G.A.R., Spurdle, A.B., kConFab Investigators, Miller, A., Kennedy, M.A., & Walker, L.C. (2017) Nanopore sequencing of full-length BRCA1 mRNA transcripts reveals co-occurrence of known exon skipping events. Breast Cancer Research, 19(1), 127. DOI: 10.1186/s13058-017-0919-1

88. Li, H., & Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754-1760. DOI: 10.1093/bioinformatics/btp324

89. Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.J., Buck, D., & Au, K.F. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000Research, 6, 100. DOI: 10.12688/f1000research.10571.2

90. Fu, S., Ma, Y., Yao, H., Xu, Z., Chen, S., Song, J., & Au, K.F. (2018) IDP-denovo: *de nov*o transcriptome assembly and isoform annotation by hybrid sequencing. Bioinformatics, 34(13), 2168-2176. DOI: 10.1093/bioinformatics/bty098

91. Moldovan, N., Szucs, A., Tombacz, D., Balazs, Z., Csabai, Z., Snyder, M., & Boldogkoi, Z. (2018) Multiplatform next-generation sequencing identifies novel RNA molecules and transcript isoforms of the endogenous retrovirus isolated from cultured cells. FEMS Microbiol. Letts., 365(5). DOI: 10.1093/femsle/fny013

92. Moldovan, N., Tombacz, D., Szucs, A., Csabai, Z., Snyder, M., & Boldogkoi, Z. (2017) Multi-Platform Sequencing Approach Reveals a Novel Transcriptome Profile in Pseudorabies Virus. Frontiers Microbiol., 8, 2708. DOI: 10.3389/fmicb.2017.02708

93. Hargreaves, A.D., & Mulley, J.F. (2015) Assessing the utility of the Oxford Nanopore MinION for snake venom gland cDNA sequencing. PeerJ, 3, e1441. DOI: 10.7717/peerj.1441.

94. Marchet, C., Lecompte, L., Silva, C.D., Cruaud, C., Aury, J.M., Nicolas, J., & Peterlongo, P. (2018) *De novo* clustering of long reads by gene from transcriptomics data. Nucl. Acids Res. DOI: 10.1093/nar/gky834

95. Kent, W. J. (2002) BLAT - the BLAST-like alignment tool. Genome Res., 12(4), 656-664. DOI: 10.1101/gr.229202

96. Oikonomopoulos, S., Wang, Y. C., Djambazian, H., Badescu, D., & Ragoussis, J. (2016) Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci. Rep., 6, 31602. DOI: 10.1038/srep31602.

97. Jain, M., Fiddes, I.T., Miga, K.H., Olsen, H.E., Paten, B., & Akeson, M. (2015) Improved data analysis for the MinION nanopore sequencer. Nature Methods, 12(4), 351-356. DOI: 10.1038/nmeth.3290

98. Chaisson, M.J., & Tesler, G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics, 13, 238. DOI: 10.1186/1471-2105-13-238

99. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990) Basic local alignment search tool. J. Mol. Biol., 215(3), 403-410. DOI: 10.1016/S0022-2836(05)80360-2

100. Smith, T.F., & Waterman, M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol., 147(1), 195-197

101. Sovic, I., Sikic, M., Wilm, A., Fenlon, S.N., Chen, S., & Nagarajan, N. (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nature Commun., 7, 11307. DOI: 10.1038/ncomms11307

102. Wagner, G.P., Kin, K., & Lynch, V.J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosciences, 131(4), 281-285. DOI: 10.1007/s12064-012-0162-3

Поступила: 03. 12. 2018.
Принята к публикации: 13. 12. 2018.

PROSPECTS FOR THE USE OF THIRD GENERATION SEQUENCERS FOR QUANTITATIVE PROFILING OF TRANSCRIPTOME

S.P. Radko[1]*, L.K. Kurbatov[1], K.G. Ptitsyn[2], Y.Y. Kiseleva[3], E.A. Ponomarenko[1], A.V. Lisitsa[1], A.I. Archakov[1]

[1]Institute of Biomedical Chemistry, 10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: radkos@yandex.ru
[2]Adzhinomoto-Genetika,1 Dorozhny 1-st Road, Moscow, 117545 Russia
[3]Russian Scientific Center of Roentgenoradiology, 86 Profsoyuznaya str., Moscow, 117997 Russia

Transcriptome profiling is widely employed to analyze transcriptome dynamics when studying various biological processes at the cell and tissue levels. Unlike the second generation sequencers, which sequence relatively short fragments of nucleic acids, the third generation DNA/RNA sequencers developed by biotechnology companies “PacBio” and “Oxford Nanopore Technologies” allow one to sequence transcripts as single molecules and may be considered as potential molecular counters capable to measure the number of copies of each transcript with high throughput, sensitivity, and specificity. In the present review, the features of single molecule sequencing technologies offered by “PacBio” and “Oxford Nanopore Technologies” are considered alongside with their utility for transcriptome analysis, including the analysis of transcript isoforms. The prospects and limitations of the single molecule sequencing technology in application to quantitative transcriptome profiling are also discussed.

Key words: third generation sequencing; transcriptome; quantitative profiling

ACKNOWLEDGMENTS

This work has been performed within the framework of the Fundamental Scientific Research Program of the State Academies of Sciences for 2013-2020.