Вавиловский журнал генетики и селекции. 2016;20(2):247-254
DOI 10.18699/VJ16.148

Особенности экспериментального планирования при исследовании транскриптомов методами высокопроизводительного секвенирования

П.Н. Меньшанов[1, 2] ✉, Н.Н. Дыгало[1, 2]

[1] Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», Новосибирск, Россия
[2] Федеральное государственное автономное образовательное учреждение высшего образования «Новосибирский национальный исследовательский государственный университет», Новосибирск, Россия

В обзоре проанализированы отдельные проблемы планирования экспериментов с использованием методов RNA-Seq и Ribo-Seq, а также консолидированы ранее опубликованные рекомендации консорциума ENCODE (2011) и других авторов по вопросам планирования экспериментов при изучении транскриптомов как у млекопитающих, так и у других животных и растений. Существует предел увеличения глубины секвенирования для идентификации практически всех активно транскрибируемых в образце генов, который зависит от размера транскриптома у объекта исследования. Увеличение глубины прочтения транскриптома выше рекомендуемой не даст значительного повышения статистической мощности исследования. У млекопитающих для идентификации активно транскрибируемых генов оптимальная глубина секвенирования составляет ~2×10^9 п. н. на биологический образец. Для остальных видов глубина секвенирования на образец определяется с учетом данного значения, но должна быть пересчитана относительно протяженности транскриптома и удельного количества РНК на клетку в сравнении с транскриптомом млекопитающих. Выявление дифференциально экспрессируемых генов и стыков сайтов сплайсинга в мРНК можно улучшить, повышая число анализируемых биологических образцов в экспериментальных группах. Минимально допустимое число биологических повторов в группе должно быть равно двум. В то же время оптимальное число биологических повторов при соблюдении вышеозначенной глубины секвенирования составляет 5–8 образцов (как и при количественной оценке экспрессии отдельных генов методом qRT-PCR). При выполнении определения последовательности транскриптов рекомендуется использовать технологии секвенирования, точность определения буквы последовательности для которых ≥ 0,999. Учитывая удельную себестоимость секвенирования, для метода RNA-Seq целесообразно использовать технологии, дающие риды длиной ≥ 75 п. н. Удельные затраты на секвенирование в контрольных группах можно снизить за счет увеличения числа опытных экспериментальных групп путем компоновки нескольких сходных экспериментов или логического усложнения исходного эксперимента. Данные рекомендации могут быть использованы для планирования экспериментов по изучению транксриптомов в функциональной геномике.

Ключевые слова: высокопроизводительное секвенирование; транскриптом; RNA-Seq; Ribo-Seq; планирование эксперимента.

The design of experiments for the transcriptome studies by high-throughput sequencing methods

P.N. Menshanov[1, 2] ✉, N.N. Dygalo[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia

The common questions in the design of the high-throughput sequencing experiments using RNA-Seq or Ribo-Seq methods are reviewed. The ENCODE guidelines (2011) as well as the recently published advances in the design of the studies of mammalian, animal and plant transcriptomes are also summarized in this review. The optimal limit of the sequencing depth does exist for the identific tion of almost all actively transcribed genes. This limit depends on the transcriptome size in the biological object studied. Additional sequencing does not provide any substantial additional information about the transcriptome complexity. For mammals, the optimal limit of the sequencing depth for the identific tion of the actively transcribed genes is equal to ~2×10^9 bp per biological sample. For other species, the optimal limit of the sequencing depth per biological sample is determined similarly for mammals; however, the transcriptome size and the mean RNA content in the studied object should be taken into account, in comparison to the mammalian transcriptomes. The discovery of differentially expressed genes, as well as the identific tion of splicing sites in the mRNA could be enhanced by increasing the number of biological samples analyzed per each experimental group. The minimal number of biological replicates per experimental group is equal to 2. However, the optimal number of biological replicates per experimental group is equal to 5–8 (similar to the experiments quantifying the expression of single genes by qRT-PCR). For the transcriptome studies, it is recommended to use the sequencing technologies that have the accuracy of sequencing ≥ 0.999 per bp. For RNA-Seq, it is also recommended to use the technologies that are able to produce reads equal to or larger than 75 bp, to minimize the cost of the effective identific tion of the sequences. The relative cost for the sequencing of the control samples could be

REVIEW
Received 29.09.2015 г.
Accepted for publication 30.11.2015 г.
© AUTHORS, 2016

✉ e-mail: MenshanovPN@icg.sbras.ru

reduced by increasing the number of experimental groups in the experiment or by combining several independent experiments with similar control groups. The present notes could be utilized during the design step in the experimental studies devoted to the research of transcriptomes.

Key words: high-throughput sequencing; transcriptome; RNA-Seq; Ribo-Seq; design of the experiment.

КАК ЦИТИРОВАТЬ ЭТУ СТАТЬЮ:

Меньшанов П.Н., Дыгало Н.Н. Особенности экспериментального планирования при исследовании транскриптомов методами высокопроизводительного секвенирования. Вавиловский журнал генетики и селекции. 2016;20(2):247-254. DOI 10.18699/VJ16.148

HOW TO CITE THIS ARTICLE:

Menshanov P.N., Dygalo N.N. The design of experiments for the transcriptome studies by high-throughput sequencing methods. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2016;20(2):247-254. DOI 10.18699/VJ16.148

Две технологии высокопроизводительного секвенирования, используемые при изучении транскриптомов у животных и растений – метод RNA-Seq для определения последовательностей транскрибируемого генома и метод Ribo-Seq для определения последовательностей транслируемого генома – постепенно становятся базовыми инструментами современной функциональной генетики для исследования экспрессии генов в самых разнообразных моделях *in vivo* и *in vitro* (Wang et al., 2009; Sims et al., 2014; Ingolia, 2014). Использование данных методов уже позволило выявить существование значительного числа новых транскриптов с неизвестной функцией, считываемых в самых различных тканях организма (van Bakel et al., 2011; Aspden et al., 2014; Xie et al., 2014), обнаружить резкую «перестройку» транскриптома после стресса (Moskalev et al., 2015), подтвердить функциональный хаос регуляции транскрипции у гибридов (McManus et al., 2010). При изучении тканевых транскриптомов также были составлены подробные молекулярные атласы представленности различных функционально значимых транскриптов в тканях организма у человека, крысы, мыши и других модельных животных и растений (например, Hawrylycz et al., 2012; O'Rourke et al., 2014).

Главным лимитирующим фактором, ограничивающим возможности экспериментаторов при планировании и проведении исследований транскриптомов, является ограниченность финансовых, трудовых и временны́х ресурсов, затраченных на подобные эксперименты (Sims et al., 2014). Удельная стоимость секвенирования 1×10^6 п. н. генома до 2010 г. составляла более одного доллара на всех имеющихся платформах (Mardis, 2008; Wetterstrand, 2015) (рис. 1). В связи с подобной высокой стоимостью секвенирования до 2010 г. эксперименты по изучению дифференциально экспрессируемых генов методами RNA-Seq и RiboSeq были по большей части невозможны в силу своей дороговизны, а в период 2011–2015 гг. эксперименты по сравнительному анализу транскриптомов чаще всего представляли собой прямое сравнение двух экспериментальных групп размером в один–три биологических повтора каждая (Sims et al., 2013). Вместе с тем активное развитие технологий секвенирования второго поколения привело в последние пять лет к значительному снижению удельной стоимости исследования генетической информации (рис. 1) и позволило ученым планировать эксперимент, принимая во внимание такие параметры, как глубина секвенирования, количество экспериментальных групп и образцов на группу.

В настоящее время существует значительное число обзоров литературы, детально рассказывающих о технических особенностях считывания ридов при секвенировании транскриптомов на различных платформах (например, Ansorge, 2009; Mutz et al., 2013), а также о программах и алгоритмах, которые могут быть использованы на второй стадии исследования – стадии сборки транскриптов (например, Martin, Wang, 2011; Florea, Salzberg, 2013; Ghosh, Chan, 2016). Вместе с тем в литературе мало внимания уделяется базовым правилам планирования экспериментов RNA-Seq и Ribo-Seq, а большинство информации, посвященной данному вопросу, встречается в разрозненном виде в обзорах литературы, описывающих технологию полногеномного секвенирования (например, Sim et al., 2014). Единственные правила проведения экспериментов (правила консорциума ENCODE – «Standards, Guidelines and Best Practices for RNA-Seq», редакция 1.0) были сформированы на основе данных, полученных при изучении транскриптомов у млекопитающих, и не обновлялись с 2011 г. (ENCODE 2011; Spies, Ciaudo, 2015). В этом обзоре проанализирован ряд важных вопросов планирования полногеномных экспериментов с использованием методов RNA-Seq и Ribo-Seq, а также консолидированы ранее опубликованные советы по планированию транскриптомных экспериментов, которые до настоящего времени были представлены в разрозненном виде в ключевых публикациях и обзорах (например, Corney, 2013; Hart et al., 2013a, b). Все эти рекомендации могут быть использованы для планирования экспериментов по изучению транскриптомов и у животных, и у растений в различных функциональных моделях *in vivo* и *in vitro*.

Глубина секвенирования транскриптома

Очевидно, что глубина секвенирования транскриптома определяется целью эксперимента, которая может сводиться как (1) к простому определению всех активно транскрибируемых экзонов, так и (2) к идентификации всех считываемых вариантов транскриптов, стыков сайтов сплайсинга и вариантов полиморфных сайтов в транскриптах образца (Veeneman et al., 2015). Отдельно следует

отметить, что правила консорциума ENCODE (2011) были сформулированы на основании данных, полученных при исследовании транскриптомов у ключевых модельных млекопитающих (мыши, крысы, люди) с размером генома ~ 10^9 п. н. По этой причине для остальных биологических видов необходимая глубина секвенирования на биологический образец также определяется с учетом этих правил, однако должна быть пересчитана относительно протяженности транскриптома (длины всех участков генома, кодирующих транскрипты) и удельного количества РНК на клетку в сравнении с транскриптомом млекопитающих.

Идентификация активно транскрибируемых мРНК у млекопитающих

Исходно в правилах ENCODE (2011) было зафиксировано (для млекопитающих), что для выявления активно транскрибируемых в образце мРНК достаточно определить последовательность у не менее чем 20–25 млн ридов, при этом длина рида должна быть ≥ 30 п. н. (эквивалентно ≥ $1{,}5 \times 10^9$ п. н. генетической информации). Этот тезис опирается на данные членов консорциума ENCODE (Mortazavi et al., 2008) и на более ранние оценки группы Galau (Galau et al., 1977), согласно которым в относительно гомогенной ткани у млекопитающих (например в печени) гены с активно работающими промоторами должны давать как минимум один транскрипт на клетку, и уровень таких мРНК должен быть не менее 0,3–1,0 FKPM. Как следствие, в большинстве экспериментов с применением RNA-Seq, опубликованных в начале 2010-х годов, глубина секвенирования на образец животного происхождения составляла порядка 1×10^9–3×10^9 п. н. (Sims et al., 2013).

Однако достаточно ли такой глубины прочтения транскриптома в современных экспериментах? Правило по глубине секвенирования от ENCODE неоднократно подвергалось критике другими членами сообщества (например, Hebenstreit et al., 2011). Более того, было установлено, что число идентифицируемых в образце транскриптов зависит от глубины секвенирования (Tarazona et al., 2011).

Последующие исследования подтвердили тезис ENCODE по глубине секвенирования для выявления уникальных транскриптов. В частности, Hart с коллегами (2013b) на основании данных 127 экспериментов RNA-Seq установили, что в подавляющем большинстве случаев свыше 90 % транскриптов определяются на уровне > 0,1 FKPM даже в образцах из неоднородных тканей. Marinov с коллегами (2014) также установили, что для подавляющего большинства активных белок-кодирующих генов в одной клетке млекопитающих в каждый конкретный момент времени присутствует лишь от 50 000 до 1 транскрипта на клетку. В результате члены консорциума ENCODE подтвердили, что активные гены обычно имеют уровень экспрессии не ниже 0,5–5 FKPM (чем больше значение, тем меньше тотальной РНК было в исследуемой клетке), что в грубом приближении эквивалентно одному транскрипту на клетку (Kellis et al., 2014).

Еще одним доказательством данного правила ENCODE стало исследование Hart с коллегами (2013a), в котором была предложена модель, классифицирующая все гены на: (а) гены с неактивированными промоторами в гетерохроматине и (б) гены с активными промоторами,

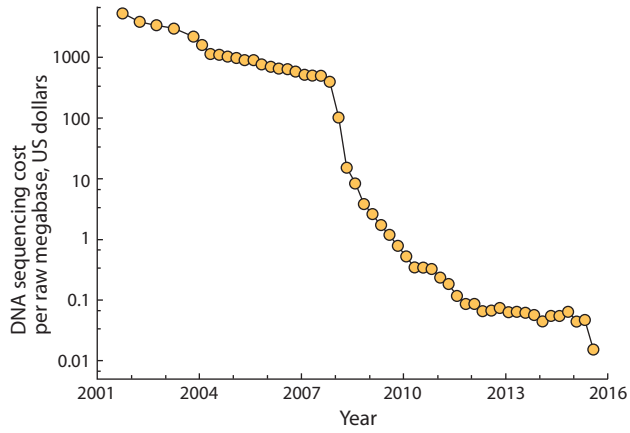

Fig. 1. The cost of DNA sequencing in US dollars per raw megabase in 2001–2015 (according to the NHGRI Genome Sequencing Program).

локализованные в эухроматине (Hart et al., 2013a). В этой работе было показано, что для простой идентификации подавляющего большинства генов с активным промотором у млекопитающих вполне достаточно 20–30 млн ридов (≥ 2×10^9 п. н. на образец), и дальнейшее увеличение глубины секвенирования не дает практически никакой дополнительной информации об отдельных транскриптах (Hart et al., 2013a).

Все эти исследования свидетельствуют о том, что не требуется бесконечное увеличение глубины секвенирования для идентификации практически всех активно транскрибируемых в образце экзонов – достаточно определить последовательность у более чем 20–30 млн ридов с длиной не менее чем 50 п. н. (порядка 2×10^9 п. н. генетической информации). Даже если транскрипт будет представлен не во всех клетках (лишь в ~20–50 % от общего числа клеток), имеющихся данных будет вполне достаточно для идентификации большинства таких мРНК у млекопитающих (Hart et al., 2013b).

Идентификация альтернативно транскрибируемых мРНК у млекопитающих

Для целей идентификации стыков сайтов сплайсинга, получившихся в результате альтернативного сплайсинга пре-мРНК, а также вариантов полиморфных сайтов в транскриптах подобной глубины прочтения в одном образце будет заведомо недостаточно. По мнению консорциума ENCODE (2011), уверенная детекция большинства стыков сайтов сплайсинга и полиморфных вариантов возможна лишь при прочтении не менее 100–200 млн парных ридов длиной не менее 75 п. н. (15×10^9–30×10^9 п. н. информации). Наши данные также указывают на то, что использование ридов с длинами 75 п. н. и более для детекции стыков сайтов сплайсинга является наиболее оптимальным с экономической точки зрения, если медианная длина экзона у транскриптов равна или больше таковой у млекопитающих (Menshanov, Dygalo, 2015).

Вместе с тем следует обратить внимание на существование зависимости между длиной рида (пары ридов)

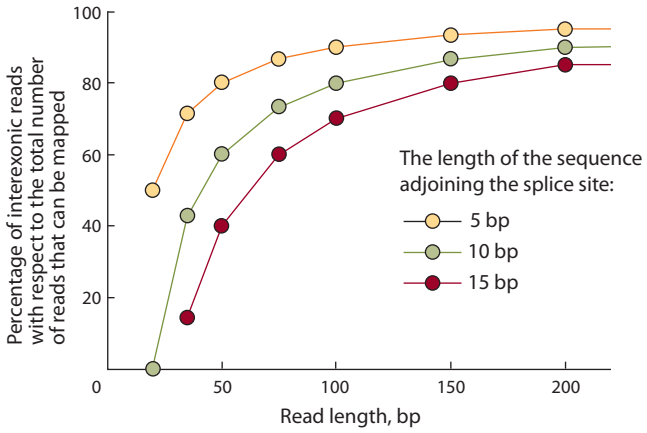

Fig. 2. The maximal portion of interexonic reads in the total number of reads of a specified si e (*X* axis: base pairs) that can be mapped with a preset length of anchors on each side of the junction used to detect a splice site (calculated for mammals, the median exon size being 140 bp).

и долей ридов, несущих стыки сайтов сплайсинга (рис. 2). В частности, это важно для метода Ribo-Seq, для которого эффективная длина рида составляет всего лишь ~30 п. н., именно такой участок транслируемой мРНК защищен рибосомой от деградации (Ingolia, 2014). В результате даже при использовании современных алгоритмов для сборки транскриптов не более чем 30–40 % ридов будут содержать необходимую для экспериментатора и однозначно восстанавливаемую информацию о местах стыка сайтов сплайсинга. По этой причине при использовании метода Ribo-Seq надежная детекция стыков сайтов сплайсинга для транскриптома млекопитающих будет достигнута лишь при считывании не менее чем 500 млн ридов.

Для метода RNA-Seq расчетные данные свидетельствуют о нецелесообразности использования сверхдлинных ридов (более 150–200 п. н.) и достижении высокой эффективности детекции стыков сайтов сплайсинга уже при использовании парных ридов длиной 100 п. н. (рис. 2) (Menshanov, Dygalo, 2015). Сверхдлинные риды целесообразно использовать лишь для специфических задач, определенных целями эксперимента.

Глубина секвенирования транскриптома, протяженность транскриптома и удельное количество РНК в клетках

Необходимо помнить, что все вышеперечисленные рекомендации, в том числе и рекомендации ENCODE, опираются на данные, полученные в экспериментах RNA-Seq у млекопитающих, протяженность транскриптома у которых не превышает 100 млн п. н. Однако следует отметить, что протяженность транскриптома слабо коррелирует с размером генома у различных животных и растений, что затрудняет предсказание данного показателя для транскриптома у видов, для которых известен лишь размер генома (Corney, 2013; Kagale et al., 2014; Coate, Doyle, 2015). По этим причинам планирование глубины секвенирования транскриптома в каждом конкретном эксперименте должно учитывать не только расчетную, но и по возможности ранее определенную протяженность транскриптома у исследуемого вида. Например, протяженность транскриптома дрозофилы ненамного меньше транскриптома у млекопитающих (Nfonsam et al., 2012). В то же время у злаковых средняя протяженность транскриптома в два–три раза больше, чем у мыши (Krasileva et al., 2013), поэтому для этих видов растений целесообразно установить глубину секвенирования не менее чем в $4–5 \times 10^9$ п. н.

Активно используемое в западной литературе понятие «размер транскриптома» (transcriptome size) может быть определено и как количество транскриптов и РНК в клетке (Coate, Doyle, 2010, 2015), которое также лишь частично коррелирует с размером генома (Marguerat, Bähler, 2012). Удельное количество РНК и транскриптов в клетке может варьировать не только у различных видов эукариот и прокариот, но и у разнополых особей отдельных видов, у одного организма в различных тканях, в одной и той же ткани на разных этапах развития и даже в одной и той же ткани до и после отдельных экспериментальных воздействий (Coate, Doyle, 2010, 2015). Например, удельное количество РНК в листьях мангровых деревьев в четыре раза выше, чем в камбиальной ткани ствола (Reef et al., 2010). Удельное содержание РНК в гепатоцитах более чем в 20 раз превышает аналогичный показатель в тимоцитах (Schmidt, Schibler, 1995). Также следует отметить, что воздействие глюкокортикоидами приводит к увеличению удельного количества РНК в печени крыс (Thompson et al., 1976; Flusser et al., 1989). В то же время в мозге, тимусе, а также ряде других тканей у млекопитающих под влиянием глюкокортикоидов происходит не увеличение, а уменьшение удельного количества РНК (Zimmerman et al., 1970). Соответственно, при планировании глубины секвенирования необходимо учитывать не только протяженность транскриптома, но и удельное количество РНК к ДНК и целевых РНК в исследуемых клетках.

Число экспериментальных групп

В связи с высокой себестоимостью полногеномного анализа одним из важнейших вопросов планирования экспериментов RNA-Seq и Ribo-Seq является определение минимального и оптимального числа экспериментальных групп для разных категорий экспериментов (сравнение экспрессии в различных тканях, временна́я динамика транскриптома после воздействия, диурнальная и возрастная динамика транскриптома). Поскольку существует бесконечное множество различных дизайнов экспериментов, правила ENCODE (2011) никак не регламентируют число экспериментальных групп, оставляя данный параметр на усмотрение экспериментатора.

Pembroke с коллегами (2015) установил, что в ходе диурнального цикла значительное число транскриптов имеет не простую синусоидальную динамику изменений, как считалось ранее, а более сложную – с двумя пиками. Индукция (или снижение) уровня ряда транскриптов после воздействия экспериментального фактора также может быть неодновременной, что проявляется в появлении нескольких «волн» изменений у так называемых «ранних» и «поздних» дифференциально экспрессируемых генов (например, Arner et al., 2015). Подобные неодновремен-

ные изменения экспрессии регуляторных генов (например транскрипционных факторов) могут приводить к появлению и более сложных паттернов экспрессии отдельных мРНК после воздействия (например, Shishkina et al., 2015).

Учитывая все вышесказанное, (1) для экспериментов, предполагающих лишь банальное сравнение базального уровня транскриптов в нескольких тканях (в одной ткани у разных видов и т. д.), целесообразно планировать по одной экспериментальной группе на каждый объект исследования; (2) для исследований, предполагающих изучение динамики уровня транскриптов в онтогенезе, также следует готовить по одной экспериментальной группе на каждый планируемый к изучению возраст (например, Kozhevnikova et al., 2013); (3) для экспериментов, предполагающих изучение диуринальных ритмов изменения транскрипции, минимально должны быть изучены две временные точки. Однако в свете данных Pembroke с коллегами (2015), более корректным будет исследование уровня транскриптов в не менее чем четырех временных точках (днем, ночью, утром и вечером, как у Hughes с коллегами (2012), а в идеале – через каждые 3–4 ч; (4) для исследований, предполагающих изучение динамики транскриптома после воздействия экспериментального фактора, минимально допустимым является определение либо «ранних» (через 30–120 мин после воздействия), либо совокупности «ранних» и «поздних» дифференциально экспрессируемых генов (через 2–12 ч после воздействия). Вместе с тем в таких случаях более разумным представляется определение динамики изменений как минимум в двух опытных точках – для установления и «ранних», и «поздних» генов, а также генов с уникальным паттерном экспрессии (помимо определения базального уровня транскриптов).

Следует отметить, что многие важные гены за счет существующих сетей регуляции транскрипции имеют сложный, неволноподобный паттерн экспрессии, который не укладывается в простую модель ранней и поздней «волны» изменений (Arner et al., 2015). Одним из наиболее характерных примеров таких генов является ген *bcl-X*, уровень длинного транскрипта которого может быть индуцирован глюкокортикоидами в стволе мозга через 6 ч после начала воздействия, однако уже через 2 ч после этого уровень мРНК Bcl-XL восстанавливается до базального (Shishkina et al., 2015). Очевидно, что для идентификации таких генов потребуется более чем две экспериментальные точки для оценки уровня целевых транскриптов после исходного воздействия.

Число внутригрупповых повторов

Ранее было установлено, что при достаточной глубине прочтения каждого биологического образца ($\geq 2 \times 10^9$ п. н. на образец) самым разумным методом повышения мощности исследования является увеличение числа биологических повторов, а не глубины секвенирования каждого образца (Ching et al., 2014; Liu et al., 2014). Однако сколько биологических повторов необходимо и достаточно делать при исследовании транскриптома?

Правила ENCODE (2011) устанавливают, что минимальное число биологических повторов в группе должно быть не менее двух. Подобный дизайн позволяет проверить внутригрупповую дисперсию и определить условное «качество» детекции целевых транскриптов – коэффициент детерминации R^2. В соответствии с правилами ENCODE, данный параметр должен быть не менее 0,90. В противном случае исследователь должен объяснить причину возникновения низкого значения коэффициента детерминации, либо воспроизвести результаты эксперимента.

Для выявления целевых дифференциально экспрессируемых генов (как и при количественной оценке экспрессии отдельных генов методом qRT-PCR) целесообразно оценивать необходимое число образцов в каждой группе, используя классический анализ статистической мощности и учитывая ожидаемую вариабельность уровня исследуемых транскриптов (Karlen et al., 2007; Spies, Ciaudo, 2015). Для идентификации практически всех мРНК, уровень которых изменился в $\geq 1{,}5$ раза и средняя копийность которых выше, чем один транскрипт на клетку, оптимальным размером группы будет пять–восемь образцов на экспериментальную группу (Karlen et al., 2007).

Выбор платформы для высокопроизводительного секвенирования

Каждая платформа для выполнения полногеномных исследований имеет ряд технических характеристик, которые предопределяют качество секвенирования. При планировании эксперимента следует обращать внимание на такие параметры, как длина и парность рида, которые дает платформа, а также на вероятность совершения ошибки при определении одного нуклеотида.

В табл. 1 приведены данные по доле ридов определенной длины с различным количеством технических ошибок, в зависимости от используемой платформы (Corney, 2013; Fox et al., 2014). Поскольку для метода Ribo-Seq эффективная длина рида составляет всего лишь ~30 п. н. (Ingolia, 2014), то для экспериментов с использованием данного метода платформами выбора будут платформы SOLID и Illumina с короткими непарными ридами длиной не более 50 п. н., дающие значительное число ридов без ошибок (табл. 1).

Табл. 2 содержит информацию об удельной цене секвенирования 1×10^9 п. н. на различных платформах для метода RNA-Seq. Учитывая данные Menshanov и Dygalo (2015), для RNA-Seq целесообразным будет использование платформы Illumina 4000 в варианте с парными ридами длиной ≥ 100 п. н., для которой стоимость эффективного секвенирования 1×10^9 п. н. будет ниже 50 долл. США (рис. 2, табл. 2).

Удельная цена секвенирования

Отдельно стоит оговорить возможности по экономии денежных ресурсов при выполнении экспериментов RNA-Seq и Ribo-Seq. Существует ряд возможностей оптимизации затрат на проведение полногеномных исследований путем корректного планирования проведения эксперимента. Например, очевидно, что число контрольных и опытных групп определяется целями эксперимента (см. выше). Удельную стоимость выполнения эксперимента можно снизить путем увеличения числа опытных (неконтрольных) экспериментальных групп, что обычно достигается за счет компоновки в одном исследовании

Table 1. Percentages of reads with specified numbers of basic sequencing e rors with different sequencing platforms

Platform	Probability of an improper determination of a single nucleotide	Percentages of 30-bp reads with certain numbers of sequence errors (Ribo-Seq)			Percentages of 100bp reads with certain numbers of sequence errors (RNA-Seq)		
		no errors	1 error	≥ 2 errors	no errors	1 error	≥ 2 errors
454 GS Junior PacBio RS Ion Torrent/Proton	~ 0.01	73.97 %	22.42 %	≤ 3.62 %	36.60 %	36.97 %	26.42 %
Solid	~ 0.001	97.04 %	2.91 %	≤ 0.05 %	–	–	–
Illumina	~ 0.001	97.04 %	2.91 %	≤ 0.05 %	90.48 %	9.06 %	≤ 0.47 %

Table 2. The cost of DNA sequencing in US dollars per raw gigabase (RNA-seq method) depends on the length of a single read and on the type of the sequencing platform

Sequencing platform	Single read length, bp	Cost of sequencing per raw gigabase, US dollars	Cost of obtaining equivalent numbers of interexonic reads, US dollars
SOLID	50	70	~ 120
SOLID	75	55.6	~ 65
Illumina 2000	50	110	~ 190
Illumina 2000	75	90	~ 105
Illumina 2000	100	75	~ 75
Illumina 4000	100	45	~ 45

According to Menshanov, Dygalo (2015), with additions.

Table 3. The proportion of the cost of data analysis in control experimental groups in the total cost of the sequence determination in all experimental groups (with the assumption of equal numbers of raw sequencing information obtained in each experimental group)

Number of control groups	Total number of experimental groups								
	2	3	4	5	6	7	8	9	10
1	50	33.3	25	20	16.7	14.3	12.5	11.1	10
2	–	66.7	50	40	33.3	28.6	25	22.2	20
3	–	–	75	60	50	42.9	37.5	33.3	30

нескольких сходных экспериментов или усложнения исходного эксперимента. В результате, если экспериментатор будет в одном исследовании сопоставлять несколько воздействий, то удельные затраты на секвенирование контрольных групп будут снижаться (табл. 3).

Еще одной важной проблемой является неразвитость центров коллективного пользования на территории Российской Федерации и, по сути, региональная монополизация данной высокотехнологичной отрасли несколькими игроками, обладающими доступом к приборной базе. На мировом рынке секвенирования существует значительное число крупных и мелких игроков, предоставляющих свои услуги по полногеномному секвенированию за вполне приемлемую цену. Из крупных игроков наибольшего внимания заслуживают BGI Technologies (www.bgitech.com) и Macrogene (www.macrogenlab.com), которые занимают более 50 % всего рынка коммерческого секвенирования. В связи с этим сопоставление цен секвенирования в различных отечественных и зарубежных центрах, где может проводиться коммерческое определение последовательности транскриптомов методами RNA-Seq и Ribo-Seq, а также вмененных затрат на транспортировку образцов до таких центров, становится важным элементом планирования эксперимента.

Выбор между методами RNA-Seq и qRT-PCR

Группа Nonis с коллегами (2014) провела сравнительный анализ себестоимости количественной оценки уровней у различного числа транскриптов в предопределенном числе образцов при использовании высокопроизводительного секвенирования методом RNA-Seq и классического метода qRT-PCR. Было показано, что использование современных платформ высокопроизводительного секвенирования, дающих риды длиной не менее 75–100 п. н., для оценки уровней у более чем 200–250 транскриптов методом RNA-Seq всегда экономически выгоднее аналогичного исследования, выполненного методом qRT-PCR, если брать цены 2014–2015 гг. (подробнее Nonis et al.,

2014) (рис. 3). Так как в биологических объектах присутствуют не сотни, а тысячи различных транскриптов, исследование Nonis с коллегами (2014) подтвердило экономическое преимущество применения метода RNA-Seq над qRT-PCR для количественной оценки уровней транскриптов. Использование метода RNA-Seq позволяет исследовать не только уровни транскриптов, но и их последовательности. Как следствие, преимущество применения высокопроизводительного секвенирования в функциональной геномике является неоспоримым.

Представленные в настоящем обзоре рекомендации не являются исчерпывающими. Отдельно следует отметить, что для нужд клинической генетики (детекция соматических и генеративных полиморфных вариантов генома) уже разработан либо продолжает разрабатываться целый набор правил (Bennett, Farah, 2014; Gargis et al., 2015), который также может быть использован и в экспериментальных целях. Дальнейшее развитие и удешевление полногеномного секвенирования позволят повысить эффективность исследований в области функциональной геномики методами RNA-Seq и Ribo-Seq. По этой причине планирование эксперимента будет неотъемлемой компонентой, позволяющей избежать потерь ресурсов в ходе выполнения научно-исследовательской деятельности в условиях бюджетных и иных ограничений.

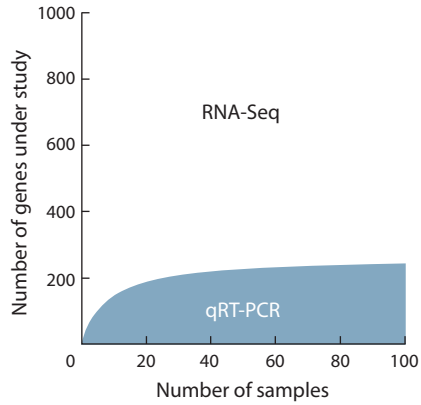

Fig. 3. The economic choice between RNA-Seq and qRT-PCR in the study of expression levels in a specified number of genes in a specifie number of experimental samples (according to Nonis et al., 2014).

The reduction of the RNA-Seq cost shifts the border of equal economic efficie y of RNA-Seq/qRT-PCR use downward to the *X*-axis. The reduction of the qRT-PCR cost will shift this border upward from the *X*-axis.

Acknowledgments

This study was supported by State Budgeted Project 0324-2015-0004, and the Russian Foundation for Basic Research, projects 14-04-00219, 15-34-20574_mol_a_ved, and 13-04-40014-N KOMFI.

Conflicts of interest

The authors declare no conflicts of interest.

References

Ansorge W.J. Next-generation DNA sequencing techniques. Nat. Biotechnol. 2009;25(4):195-203. DOI 10.1016/j.nbt.2008.12.009

Arner E., Beckhouse A., Briggs J., Ovchinnikov D., Wolvetang E., Wells C. and FANTOM Consortium. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science. 2015;347(6225):1010-1014. DOI 10.1126/science.1259418

Aspden J.L., Eyre-Walker Y.C., Phillips R.J., Amin U., Mumtaz M.A., Brocard M., Couso J.P. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. Elife. 2014;3:e03528. DOI 10.7554/eLife.03528

Bennett N.C., Farah C.S. Next-generation sequencing in clinical oncology: Next Steps Towards Clinical Validation. Cancers (Basel). 2014;6(4):2296-2312. DOI 10.3390/cancers6042296

Ching T., Huang S., Garmire L.X. Power analysis and sample size estimation for RNA-Seq differential expression. RNA. 2014;20(11):1684-1696. DOI 10.1261/rna.046011.114

Coate J.E., Doyle J.J. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol. Evol. 2010;2:534-546. DOI 10.1093/gbe/evq038

Coate J.E., Doyle J.J. Variation in transcriptome size: are we getting the message? Chromosoma. 2015;124(1):27-43. DOI 10.1007/s00412-014-0496-3

Corney D.C. RNA-seq using next generation sequencing. Mater. Methods. 2013;3:203. DOI 10.13070/mm.en.3.203

Florea L.D., Salzberg S.L. Genome-guided transcriptome assembly in the age of next-generation sequencing. IEEE/ACM Trans. Comput. Biol. Bioinform. 2013;10(5):1234-1240.

Flusser G., Ginzburg V., Meyuhas O. Glucocorticoids induce transcription of ribosomal protein genes in rat liver. Mol. Cell. Endocrinol. 1989;64(2):213-222. DOI 10.1016/0303-7207(89)90148-2

Fox E.J., Reid-Bayliss K.S., Emond M.J., Loeb L.A. Accuracy of next generation sequencing platforms. Next Gener. Seq. Appl. 2014;1:1000106. DOI 10.4172/jngsa.1000106

Galau G.A., Klein W.H., Britten R.J., Davidson E.H. Significance of rare mRNA sequences in liver. Arch. Biochem. Biophys. 1977;179(2):584-599. DOI 10.1016/0003-9861(77)90147-3

Gargis A.S., Kalman L., Bick D.P., da Silva C., Dimmock D.P., Funke B.H., Gowrisankar S., Hegde M.R., Kulkarni S., Mason C.E., Nagarajan R., Voelkerding K.V., Worthey E.A., Aziz N., Barnes J., Bennett S.F., Bisht H., Church D.M., Dimitrova Z., Gargis S.R., Hafez N., Hambuch T., Hyland F.C., Luna R.A., MacCannell D., Mann T., McCluskey M.R., McDaniel T.K., Ganova-Raeva L.M., Rehm H.L., Reid J., Campo D.S., Resnick R.B., Ridge P.G., Salit M.L., Skums P., Wong L.J., Zehnbauer B.A., Zook J.M., Lubin I.M. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat. Biotechnol. 2015;33(7):689-693. DOI 10.1038/nbt.3237

Ghosh S., Chan C.K. Analysis of RNA-Seq Data using TopHat and Cufflinks. Method. Mol. Biol. 2016;1374:339-361. DOI 10.1007/978-1-4939-3167-5_18

Hart T., Komori H.K., LaMere S., Podshivalova K., Salomon D.R. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics. 2013;14:778. DOI 10.1186/1471-2164-14-778

Hart S.N., Therneau T.M., Zhang Y., Poland G.A., Kocher J.P. Calculating sample size estimates for RNA sequencing data. J. Comput. Biol. 2013;20(12):970-978. DOI 10.1089/cmb.2012.0283

Hawrylycz M.J., Lein E.S., Guillozet-Bongaarts A.L., Shen E.H., Ng L., Miller J.A., van de Lagemaat L.N., Smith K.A., Ebbert A., Riley Z.L., Abajian C., Beckmann C.F., Bernard A., Bertagnolli D., Boe A.F., Cartagena P.M., Chakravarty M.M., Chapin M., Chong J., Dalley R.A., Daly B.D., Dang C., Datta S., Dee N., Dolbeare T.A., Faber V., Feng D., Fowler D.R., Goldy J., Gregor B.W., Haradon Z., Haynor D.R., Hohmann J.G., Horvath S., Howard R.E., Jeromin A., Jochim J.M., Kinnunen M., Lau C., Lazarz E.T., Lee C., Lemon T.A., Li L., Li Y., Morris J.A., Overly C.C., Parker P.D., Parry S.E., Reding M., Royall J.J., Schulkin J., Sequeira P.A., Slaughterbeck C.R., Smith S.C., Sodt A.J., Sunkin S.M., Swanson B.E., Vawter M.P., Williams D., Wohnoutka P., Zielke H.R., Geschwind D.H., Hof P.R., Smith S.M., Koch C., Grant S.G., Jones A.R. An anato-

mically comprehensive atlas of the adult human brain transcriptome. Nature. 2012;489(7416):391-399. DOI 10.1038/nature11405

Hebenstreit D., Fang M., Gu M., Charoensawan V., van Oudenaarden A., Teichmann S.A. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol. Syst. Biol. 2011;7: 497. DOI 10.1038/msb.2011.28

Hughes M.E., Grant G.R., Paquin C., Qian J., Nitabach M.N. Deep sequencing the circadian and diurnal transcriptome of Drosophila brain. Genome Res. 2012;22(7):1266-1281. DOI 10.1101/gr.128876.111

Ingolia N.T. Ribosome profiling: new views of translation, from single codons to genome scale. Nat. Rev. Genet. 2014;15(3):205-213. DOI 10.1038/nrg3645

Kagale S., Koh C., Nixon J., Bollina V., Clarke W.E., Tuteja R., Spillane C., Robinson S.J., Links M.G., Clarke C., Higgins E.E., Huebert T., Sharpe A.G., Parkin I.A. The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. Nat. Commun. 2014;5:3706. DOI 10.1038/ncomms4706

Karlen Y., McNair A., Perseguers S., Mazza C., Mermod N. Statistical significance of quantitative PCR. BMC Bioinformatics. 2007;8:131. DOI 10.1186/1471-2105-8-131

Kellis M., Wold B., Snyder M.P., Bernstein B.E., Kundaje A., Marinov G.K., Ward L.D., Birney E., Crawford G.E., Dekker J., Dunham I., Elnitski L.L., Farnham P.J., Feingold E.A., Gerstein M., Giddings M.C., Gilbert D.M., Gingeras T.R., Green E.D., Guigo R., Hubbard T., Kent J., Lieb J.D., Myers R.M., Pazin M.J., Ren B., Stamatoyannopoulos J.A., Weng Z., White K.P., Hardison R.C. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. USA. 2014;111(17):6131-6138. DOI 10.1073/pnas.1318948111

Kozhevnikova O.S., Korbolina E.E., Ershov N.I., Kolosova N.G. Rat retinal transcriptome: effects of aging and AMD-like retinopathy. Cell Cycle. 2013;12(11):1745-1761. DOI 10.4161/cc.24825

Krasileva K.V., Buffalo V., Bailey P., Pearce S., Ayling S., Tabbita F., Soria M., Wang S., IWGS Consortium, Akhunov E., Uauy C., Dubcovsky J. Separating homeologs by phasing in the tetraploid wheat transcriptome. Genome Biol. 2013;14(6):R66. DOI 10.1186/gb-2013-14-6-r66

Liu Y., Zhou J., White K.P. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30(3): 301-304. DOI 10.1093/bioinformatics/btt688

Mardis E.R. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24(3):133-141. DOI 10.1016/j.tig.2007.12.007

Marguerat S., Bähler J. Coordinating genome expression with cell size. Trends Genet. 2012;28(11):560-565. DOI 10.1016/j.tig.2012.07.003

Marinov G.K., Williams B.A., McCue K., Schroth G.P., Gertz J., Myers R.M., Wold B.J. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014;24(3):496-510. DOI 10.1101/gr.161034.113

Martin J.A., Wang Z. Next-generation transcriptome assembly. Nat. Rev. Genet. 2011;12(10):671-682. DOI 10.1038/nrg3068

McManus C.J., Coolon J.D., Duff M.O., Eipper-Mains J., Graveley B.R., Wittkopp P.J. Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Res. 2010;20(6):816-825. DOI 10.1101gr.102491.109

Menshanov P.N., Dygalo N.N. Methodological aspects of read mapping and assembly of transcriptomes derived from the brain tissue samples of Rattus norvegicus. Rus. J. Genet: Appl. Res. 2015;5(4):401-406. DOI 10.1134/S2079059715040097

Mortazavi A., Williams B.A., McCue K., Schaeffer L., Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008;5(7):621-628. DOI 10.1038/nmeth.1226

Moskalev A., Zhikrivetskaya S., Krasnov G., Shaposhnikov M., Proshkina E., Borisoglebsky D., Danilov A., Peregudova D., Sharapova I., Dobrovolskaya E., Solovev I., Zemskaya N., Shilova L., Snezhkina A., Kudryavtseva A. A comparison of the transcriptome of Drosophila melanogaster in response to entomopathogenic fungus, ionizing radiation, starvation and cold shock. BMC Genomics. 2015;16(Suppl. 13):S8. DOI 10.1186/1471-2164-16-S13-S8

Mutz K.O., Heilkenbrinker A., Lönne M., Walter J.G., Stahl F. Transcriptome analysis using next-generation sequencing. Curr. Opin. Biotechnol. 2013;24(1):22-30. DOI 10.1016/j.copbio.2012.09.004

Nfonsam L.E., Cano C., Mudge J., Schilkey F.D., Curtiss J. Analysis of the transcriptomes downstream of Eyeless and the Hedgehog, Decapentaplegic and Notch signaling pathways in Drosophila melanogaster. PLoS One. 2012;7(8):e44583. DOI 10.1371/journal.pone.0044583

Nonis A., De Nardi B., Nonis A. Choosing between RT-qPCR and RNA-seq: a back-of-the-envelope estimate towards the definition of the break-even-point. Anal. Bioanal. Chem. 2014;406(15):3533-3536. DOI 10.1007/s00216-014-7687-x

O'Rourke J.A., Iniguez L.P., Fu F., Bucciarelli B., Miller S.S., Jackson S.A., McClean P.E., Li J., Dai X., Zhao P.X., Hernandez G., Vance C.P. An RNA-Seq based gene expression atlas of the common bean. BMC Genomics. 2014;15:866. DOI 10.1186/1471-2164-15-866

Pembroke W.G., Babbs A., Davies K., Ponting C.P., Oliver P.L. Temporal transcriptomics suggest that twin-peaking genes reset the clock. Elife. 2015;4.pii:e10518. DOI 10.7554/eLife.10518

Reef R., Ball M.C., Feller I.C., Lovelock C.E. Relationships among RNA:DNA ratio, growth and elemental stoichiometry in mangrove trees. Funct. Ecol. 2010;24(5):1064-1072. DOI 10.1111/j.1365-2435.2010.01722.x

Schmidt E.E., Schibler U. Cell size regulation, a mechanism that controls cellular RNA accumulation: consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y and the liver-enriched transcription factor DBP. J. Cell Biol. 1995;128(4):467-483.

Shishkina G.T., Kalinina T.S., Bulygina V.V., Lanshakov D.A., Babluk E.V., Dygalo N.N. Anti-apoptotic protein Bcl-xL expression in the midbrain raphe region Is sensitive to stress and glucocorticoids. PLoS One. 2015;10(12):e0143978. DOI 10.1371/journal.pone.0143978

Sims D., Sudbery I., Ilott N.E., Heger A., Ponting C.P. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 2014;15(2):121-132. DOI 10.1038/nrg3642

Spies D., Ciaudo C. Dynamics in transcriptomics: advancements in RNA-seq time course and downstream analysis. Comput. Struct. Biotechnol. J. 2015;13:469-477. DOI 10.1016/j.csbj.2015.08.004

Tarazona S., García-Alcalde F., Dopazo J., Ferrer A., Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011;21(12):2213-2223. DOI 10.1101/gr.124321.111

The ENCODE Consortium. Standards, Guidelines and Best Practices for RNA-Seq V1.0. 1.0. 6-1-2011.

Thompson W.L., Abeles F.B., Beall F.A., Dinterman R.E., Wannemacher R.W. Jr. Influence of the adrenal glucocorticoids on the stimulation of synthesis of hepatic ribonucleic acid and plasma acute-phase globulins by leucocytic endogenous mediator. Biochem. J. 1976; 156(1):25-32.

van Bakel H., Nislow C., Blencowe B.J., Hughes T.R. Response to "The reality of pervasive transcription". PLoS Biol. 2011;9(7):e1001102. DOI 10.1371/journal.pbio.1001102

Veeneman B.A., Shukla S., Dhanasekaran S.M., Chinnaiyan A.M., Nesvizhskii A.I. Two-pass alignment improves novel splice junction quantification. Bioinformatics. 2015;32:43-49. DOI 10.1093/bioinformatics/btv642

Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 2009;10(1):57-63. DOI 10.1038/nrg2484

Wetterstrand K. DNA sequencing costs: data from the NHGRI large-scale genome sequencing program. (2015). Available at http://www.genome.gov/sequencingcosts/

Xie C., Yuan J., Li H., Li M., Zhao G., Bu D., Zhu W., Wu W., Chen R., Zhao Y. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res. 2014;42(Database issue):D98-D103. DOI 10.1093/nar/gkt1222

Zimmerman E.F., Andrew F., Kalter H. Glucocorticoid inhibition of RNA synthesis responsible for cleft palate in mice: a model. Proc. Natl. Acad. Sci. USA. 1970;67(2):779-785.