

ИЗУЧЕНИЕ СТРУКТУРЫ ТРАНСКРИПТОМА РАСТЕНИЙ МЕТОДОМ ПРЯМОГО СЕКВЕНИРОВАНИЯ РНК НА ПЛАТФОРМЕ OXFORD NANOPORE

А.И. Глушкевич, А.Н. Князев, И.А. Фесенко

Институт биоорганической химии им. академиков М.М. Шемякина и Ю.А. Овчинникова РАН,
Москва, Россия

Современные технологии Oxford Nanopore позволяют проводить прямое секвенирование молекул нуклеиновых кислот. Такие методы нанопорового секвенирования основаны на измерении электрического сигнала в белковой поре при протягивании через них нативных молекул. Генерируемые сигналы переводят в нуклеотидные последовательности, причем в случае РНК секвенирования каждое прочтение соответствует одному полноразмерному транскрипту. Полученные результаты позволяют устанавливать экзон-интронную структуру каждой молекулы мРНК без дополнительной биоинформатической сборки и дает возможность определять длину 3'-полиА хвоста.

Недавнее использование прямого секвенирования РНК с помощью нанопор на модельном растении *Arabidopsis thaliana* позволило выявить 38 тысяч новых изоформ, слитые транскрипты, альтернативные сайты полиаденилирования и начала транскрипции [12]. Также было установлены и эпitransкриптомные изменения в виде метилирования аденина, которые происходят обычно в 3' некодирующей области и регулируют циркадные ритмы у *Arabidopsis* [8]. Однако, исследований транскриптомных изменений в процессе роста и развития растений до недавнего времени проведено не было.

Для изучения структуры транскриптома растений методом прямого секвенирования РНК на платформе Oxford Nanopore нами был выбран мох *Physcomitrium patens* (Hedw.) Mitt., который широко используется в молекулярно-биологических исследованиях. Геном этого растения секвенирован и аннотирован, что позволяет исключить сборку транскриптома *de novo*. Кроме того, простой жизненный цикл мха позволяет провести сравнительную характеристику структуры транскриптомов на разных этапах развития. Нами были получены растения на двух стадиях развития – протонемы и гаметофор, которые использованы в качестве вариантов исследований. Выделение тотальной РНК проводили с использованием TRIzol™ Reagent, чистую полиА фракцию РНК получали с помощью poly(A)Purist kit, ThermoFisher. Для секвенирования использовался прибор MinION (Oxford Nanopore) с рекомендуемыми ячейками FLO-MIN106 и наборами для приготовления библиотеки SQK-RNA002, позволяющими секвенировать непосредственно молекулы мРНК в направлении 3'-5'. В процессе пробоподготовки к молекулам матричной РНК на полиадеилинированный конец лигируются адаптеры, синтезируется кДНК цепь SuperScript III обратной транскриптазой. Данный комплекс позволяет стабилизировать молекулу мРНК. На заключительном этапе к 3' концу РНК присоединяют адаптеры с моторными белками, которые обеспечивают направленное движение молекулы РНК через пору. Каждый эксперимент был выполнен в трех биологических повторях.

Полученные прочтения в виде fasta5 файлов были переведены в нуклеотидные последовательности программой Guppy 3.6.0, Oxford Nanopore Technologies. Выравнивания на геном осуществляли с использованием minimap2 [5], а обработку полученных результатов – samtools 1.9 [4] и bedtools [10]. Для анализа дифференциальной экспрессии применяли featureCounts [6] и пакет DESeq2 [7]. Экспрессию изоформ определяли с помощью FLAIR [11]. Для идентификации длины полиА хвоста использовался Nanopolish, а для визуализации результатов пакет NanoTail (<https://github.com/smaegol/nanotail>). Поиск модификаций оснований осуществляли с помощью Nanocompare [3]. Использованные скрипты и пайплайн обработки представлены на https://github.com/Liverworks/moss_nanopore.

Для каждой стадии развития было получено более 4 миллионов ридов, каждый из которых представлял отдельную единицу транскриптома. Среднее качество всех прочтений (q-score) составляло 8.7. В дальнейшем анализе использовались риды со средним качеством нуклеотида больше 7, которые составляли более 90 % всех полученных. Последующее выравнивание на геном *Physcomitrella patens*, сборка V3 [2] позволило локализовать 92 % ридов. Причем 13 % выровненных прочтений принадлежат неаннотированным областям генома. Было обнаружено 2200 новых локусов, что составляет 9 % от всех экспрессирующихся локусов.

Сравнение экспрессии аннотированных генов на стадии протонем и гаметофоров показывает глобальные различия между этими двумя жизненными формами на транскриптомном уровне (рис. 1). Анализ в терминах Gene Ontology показал, что в протонеме активнее гены, связанные с транскрипцией и трансляцией, рибосомальные белки. В гаметофорах выявлено увеличение экспрессии большого количества транскрипционных факторов.

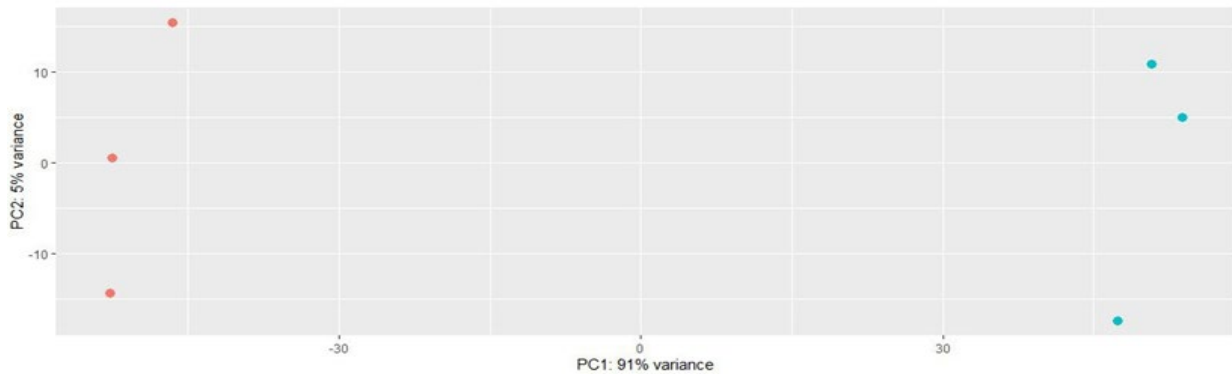


Рис. 1. PCA протонем и гаметофоров по экспрессии генов. Красными точками обозначены образцы гаметофоров, зелеными – протонем. В анализ включены все аннотированные гены

Существенные различия между двумя стадиями развития были установлены при идентификации конкретных изоформ и оценке их экспрессии. Именно благодаря уникальной технологии нанопрового секвенирования возможно определять всю последовательность мРНК одним прочтением и устанавливать соответствующую экзонную структуру (варианты сплайсинга). Всего было идентифицировано 37 000 различных транскриптов. Менее половины (14000) всех установленных транскриптов совпадают с существующими изоформами в аннотации. При этом из транскриптов, представленных в аннотации, более 90 % представлены как в гаметофорах, так и в протонемах. Многие из них экспрессируются с разной активностью и, как и при анализе экспрессии отдельных генов, показывают значительную разницу между двумя состояниями. В наших экспериментах удалось выявить транскрипты, которые содержали большое количество экзонов (например, 60 экзонов – локус Pr3c22_8910). Около половины изоформ из представленных в аннотации и около 35 % из всех транскриптов имеют более одного экзона. На рис. 2 представлены изоформы локуса Pr3c7_3990 (Chr07:2,476,701–2,482,165), в котором идентифицированные нами варианты сплайсинга не были предсказаны ранее.



Рис. 2. Локус Pr3c7_3990 в геномном браузере IGV [9]. а) транскрипты, полученные в результате нанопрового секвенирования б) аннотация локуса с предсказанными изоформами

Несомненным преимуществом нанопорового секвенирования является возможность устанавливать длину 3' хвоста полиаденилированных молекул РНК. Нами установлена средняя длина полиА хвоста мРНК на стадиях гаметофор и протонемы в размере 70 нуклеотидов. При этом показана слабая отрицательная корреляция между средней длиной полиА хвоста транскриптов гена и количеством этих транскриптов (коэффициент корреляции Спирмена $\rho = -0.205$, $p = 0.0$). Подобную закономерность указывают и для *Arabidopsis* [8]. Сравнение результатов по каждому гену в отдельности позволило обнаружить достоверную разницу по длине полиА хвостов для 110 генов на разных стадиях развития. На рис. 3 представлен пример транскрипта, для которого длина полиА хвоста идентифицируется по двум пикам частоты и составляет 40 и 60 нуклеотидов. Причем представленность коротких 3'-концов значительно выше на стадии протонемы. Также интересно, что белковый продукт этого транскрипта имеет селективный сайт связывания с РНК [1].

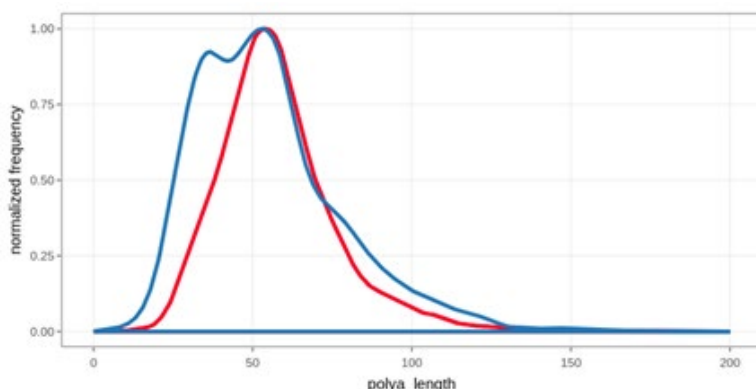


Рис. 3. Распределение длин полиА хвостов транскриптов гена Pr3c7_6560. Красным обозначены гаметофоры, синим – протонемы.

Нанопоровое секвенирование РНК у растений позволяет расширить возможности исследований транскриптома растений. В настоящем исследовании дана подробная характеристика транскриптома модельного растения *Physcomitrium patens* на двух стадиях жизненного цикла. Нами выявлены транскрибируемые участки генома, в том числе неаннотированные, и определены уровни их транскрипционной активности.

Найдены неизвестные транскрибируемые изоформы генов, которые образуются в результате альтернативного сплайсинга. Установлены различия длин полиаденилированных 3' концов для ряда генов в зависимости от стадии развития растения.

Работа выполнена при поддержке гранта РФФИ 20-04-00938 А.

ЛИТЕРАТУРА

1. Goodstein D.M. [и др.]. Phytozome: a comparative platform for green plant genomics // *Nucleic Acids Research*. 2012. № D1 (40). С. D1178-D1186.
2. Lang D. [и др.]. The *Physcomitrella patens* chromosome-scale assembly reveals moss genome structure and evolution // *The Plant Journal*. 2018. № 3 (93). С. 515–533.
3. Leger A. [и др.]. Genomics. RNA modifications detection by comparative Nanopore direct RNA sequencing. 2019.
4. Li H. [и др.]. The Sequence Alignment/Map format and SAMtools // *Bioinformatics*. 2009. № 16 (25). С. 2078–2079.
5. Li H. Minimap2: pairwise alignment for nucleotide sequences // *Bioinformatics*. 2018. № 18 (34). С. 3094–3100.
6. Liao Y., Smyth G.K., Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features // *Bioinformatics*. 2014. № 7 (30). С. 923–930.
7. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // *Genome Biology*. 2014. № 12 (15). С. 550.
8. Parker M.T. [и др.]. Nanopore direct RNA sequencing maps the complexity of *Arabidopsis* mRNA processing and m6A modification // *eLife*. 2020. (9). С. e49658.
9. Robinson J.T. [и др.]. Integrative genomics viewer // *Nature Biotechnology*. 2011. № 1 (29). С. 24–26.
10. Quinlan A.R., Hall I.M. BEDTools: a flexible suite of utilities for comparing genomic features // *Bioinformatics*. 2010. № 6 (26). С. 841–842.
11. Tang A.D. [и др.]. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns // *Nature Communications*. 2020. № 1 (11). С. 1438.
12. Zhang S. [и др.]. New insights into *Arabidopsis* transcriptome complexity revealed by direct sequencing of native RNAs // *Nucleic Acids Research*. 2020. С. gkaa588.