УДК: 004.9:004.9:004.8:577.21

М. В. СПРИНДЖУК[1], Л. П. ТИТОВ[2], А. П. КОНЧИЦ[3], Л. В. МОЖАРОВСКАЯ[3]

СОВРЕМЕННЫЕ АЛГОРИТМЫ ОБРАБОТКИ ДАННЫХ ТРАНСКРИПТОМОВ: ОБЗОР МЕТОДОВ И РЕЗУЛЬТАТЫ АПРОБАЦИИ

[1] Объединенный институт проблем информатики НАН Беларуси
[2] РНПЦ Микробиологии и эпидемиологии
[3] Институт леса НАН Беларуси

Анализ биоинформатических данных является актуальной проблемой современной вычислительной биологии и прикладной математики. С развитием биотехнологий и инструментальных средств получения и обработки данной информации появились нерешенные вопросы разработки и применения новых алгоритмов и программного обеспечения.

Авторы предлагают практические алгоритмы и методы обработки транскриптомных данных для эффективных результатов аннотирования, визуализации и интерпретации биоинформатических данных.

Ключевые слова: *транскриптом, геномика, биоинформатика, анализ данных, программное обеспечение, алгоритмы.*

Введение

С развитием технологий высокопроизводительного секвенирования геномов и транскриптомов, появилась актуальная проблема оптимизации обработки и анализа полученной информации. Известен ряд рекомендаций «лучшей практики» для обработки данных транскриптомов [1–4]. В реальных условиях практики необходимо адаптировать известные алгоритмы обработки данных, комбинировать и отбирать эффективные компоненты и параметры исполнения программных модулей с целью получения информации максимально возможного качества и оптимального объема. На рис. 1–6 схематически представлены разработанные алгоритмы обработки транскриптомных данных, успешно апорбированные на транскриптомах, полученных в лаборатории геномных исследований и биоинформатики Института леса НАН Беларуси (Гомель, Беларусь) [5–9], с последующим пояснением основных шагов разработанных алгоритмов для обработки биоинформатических данных транскриптомов. Более подробная информация о каждом элементе предлагаемого алгоритма представлена в соответствующей документации программного обеспечения (https://github.com/). Разработка эффективных алгоритмов анализа данных транскриптомов представляется междисциплинарной задачей, требующей знаний объектно-ориентировнного программирования, биоинформатики, биологической и технической терминологии.

Транскриптомный анализ

Транскриптом – совокупность всех РНК-транскриптов одной клетки или группы клеток. Тип и количество транскрибированных генов зависит от вида клеток и от изменений окружающей среды, влияющих на регуляцию транскрипции. Нарушение транскрипции часто приводит к патологическим процессам или заболеваниям [10].

За последние 20 лет накопился значительный опыт получения и анализа транскриптомных данных для бактерий, грибов растений и животных. Транскриптомные технологии особенно востребованы для эффективного выполнения новых задач в экологии, биотехнологии и молекулярной биологии, ветеренарии, судебной генетике и медицине.

Ряд зарубежных публикаций содержит подробные практические рекомендации по сборке и анализу транскриптомов для различных научных целей [11, 12]. В русскоязычных источниках также приводится методология обработки транскриптомных данных [13–15].

Операция по обработке данных транскриптомов, как правило, состоит из нескольких последовательных шагов, которые составляют общий алгоритм.

Контроль качества данных

Контроль качества (с англ. quality control, QC) исходных данных секвенирования основывается на подсчете числа прочтений и бальной оценке качества каждого прочтения в отдельности (показатель качества Phred) [16]. FastQC [17] и NGSQC [18] – наиболее распространенные программные инструменты для оценки качества первичных данных секвенирования. Также для оценки качества исходных данных транскриптомов применяют метрики результатов оценки картирования/выравнивания и аннотации на референсный транскриптом. В таком случае выполняется анализ вариантов с последующей даунстрим аннотацией, а для сборки транскриптома отбираются прочтения более высокого качества, которые распознаются в референсном транскриптоме или геноме. На этом этапе обработки данных выполняют вычисление оптимального размера k-меров для эффективной *de novo* сборки [19].

Картирование/выравнивание прочтений и *de novo* сборка

Прочтения транскриптома могут быть прокартированы/выравнены (с англ. alignment – выравнивание) на референсный геном или известный по структуре, проаннотированный транскриптом. Процедура картирования на референс выполняется, если целью эксперимента является идентификация генных изоформ. Для данного типа анализа используют как программные комплексы свободного доступа (Bowtie [20] и Bowtie2 [21], STAR [22], TopHat2 [23]), так и коммерческого: OmicsBox (стоимость лицензии на месяц стоит около ≈ 100 $), NextGene, Converge, CLC Genomics, JMP Genomics (стоимость лицензии ≈ 16 000 €). Когда отсутствует референсный транскриптом, в качестве эталона можно использовать родственный вид (как например *Arabidopsis thaliana* для растений, отмечен символом «*» на рис. 2) или выполнить *de novo* сборку. Для этой цели в биоинформатике применяют программное обеспечение rnaSPADes [24], Trinity [25–28], Oases [29,30], SOAPdenovo-trans, Abyss [31–33], NextGene Floton и другие. Более длинные прочтения или прочтения с парными концами (как при секвенировании обеих цепей ДНК) способствуют получению лучших результатов *de novo* сборки. Рекомендовано также комбинировать множество транскриптомов для получения единой комбинированной сборки с целью формирования множества эталонных контигов. Эти контиги затем можно использовать для картирования, подсчета экспрессии и сравнения между группами образцов транскриптомов.

Аннотация собранного транскриптома

Аннотация транскриптома – наиболее важный этап в алгоритме анализа полученных данных, так как его результат–информация, имеющая научное значение в области биологии. Программное обеспечение TransDecoder идентифицирует локусы, кодирующие белки-кандидаты, на основе их нуклеотидного состава, длины открытой рамки считывания и наличия функциональных доменов в соответствии с базой данных семейств белковых доменов Pfam (https://pfam.xfam.org/). Данный ресурс [34] анализирует транскрипты, полученные с помощью *de novo* сборки транскриптома с использованием ряда компьютерных программ (Trinity, rnaSPADes, MIRA, Oases, Abyss, SOAPdenovo, NextGene и пр.) или сконструированные на основе выравнивания исследуемого транскриптома с референсом с использованием инструментов Tophat, Cufflinks и других аналогичных программ. Веб сервис FastAnnotator позволяет установить потенциальные функции исследуемых транскриптов на основе GO-аннотации (с англ. gene ontology – генная онтология, GO), идентифицируя в базах данных соответственные функциональные домены, кодируемых ими белковых. Аннотация в FastAnnotator состоит из четырех основных частей: поиск лучших совпадений в базе данных NCBInr, назначение идентификационных номеров согласно GO-классификации, EC (классификации ферментов; с англ. enzyme commission – комиссия по ферментам, EC) и присвоение номеров согласно доменного поиска. Онлайн сервис свободного доступа TRAPID (http: //bioinformatics.psb.ugent.be/webtools/trapid/) выполняет функциональный, сравнительный и филогенетический анализ транскриптомных данных на

основе использования 175 эталонных протеомов. GO-аннотация выполненяемая веб сервисом ShinyGO (http://ge-lab.org/go/) [35] характеризуется следующими функциями: (1) большой базой данных GO-аннотаций более чем для 200 видов растений и животных; (2) возможностью графической визуализации результатов обогащения и характеристик генов; (3) наличием интерфейса API (с англ. application programming interface, API – интерфейс прикладного программирования) для доступа к веб-ресурсам баз данных KEGG и STRING с целью поиска метаболических сетей и белок-белковых взаимодействий.

Количественный анализ экспрессии генов

Программные пакеты HTSeq и featuresCount вычисляют уровень экспрессии генов путем агрегации числа проаннотированных прочтений для каждого транскрипта. Результат сохраняется в файле формата GTF. При этом в программах заложены различные варианты определения пересечения фрагмента прочтения с той или иной последовательностью контига или скэффолда, несущих информацию о гене. Помимо способа агрегации исходного количества прочтений генов, широко применяются различные методы на основе нормализации данных транскриптомных образцов. В таком случае учитывают размеры библиотек прочтений и их длины. Метрики таких методов – количество прочтений на 1 тыс. н. о. на миллион картированных прочтений, RPKM (с англ. reads per kilobase per million mapped reads); число фрагментов на тысячу н. о. на миллион картированных прочтений и число транскриптов на миллион картированных прочтений, FRKM и TPM (с англ., fragments или transcripts per million reads). Перечисленные вычислительные методы реализованы в алгоритмах бесплатного доступа программ обеспечения CuffLinks, RSEM, eXpress, Kallisto (табл. 1).

Как правило, метрика FRKM используется для прочтений с парными концами, а RPKM для одноконцевых прочтений. TPM в отличие от RPKM не учитывает длину генов после нормализации показателя глубины секвенирования, что делает сумму показателей всех TPM во всех образцах одинаковыми и помогает в сравнении профиля экспрессии между различными транскриптомами. Избыточность

Таблица 1. **Технические характеристики программного обеспечения, предназначенного для вычисления экспрессии генов транскриптома [36]**

Название	Оперативная память (ГБ)	Время затрат	Алгоритм	Мульти-поточность
Cufflinks	3,5	117	МП	да
RSEM	5,6	154	МП	да
eXpress	0,55	30	МП	нет
TIGAR2	28,3	1045	ВБ	да
Kallisto	3,8	7	МП	да
Salmon	6,6	6	ВБ/МП	да
Salmon_aln	3	7	ВБ/МП	да
Sailfish	6,3	5	ВБ/МП	да

прочтений транскриптома вычисляется как среднее от нормализованных данных.

Объединение данных, оценка полученных результатов и формирование выводов

На последнем этапе, полученные данные можно объединить и структурировать для кластеризации, применить методы машинного обучения и построения онтологических сетей с формированием заключений о биологическом значении полученных результатов исследования.

Заключение

Разработанные алгоритмы и методологические основы обработки данных транскриптомов растений успешно апробированы на образцах данных сосны обыкновенной, лиственницы и березы [5–9].

Исследования, направленные на интенсификацию биоинфортматических методов метрики уровня экспрессии генов транскриптомов, высоко актуальны и нуждается в дальнейшем изучении в эксперименте *in silico*. Согласно литературным данным, уровень экспрессии генов лучше оценивать по отдельным компонентам, т. е. известные проаннотированные гены или их организованные кластеры, как например, с применением веб сервиса Genix [38]. Что позволяет отобрать гены с одинаковой аннотацей и проследить их метрики в отчетных таблицах программ с вычислением уровня экспресии генов транскриптома с последующим выполнением статистических расчетов по вычислению значимой разницы между группами образцов.

Наиболее эффективными программами для количественной оценки полученных данных и качественной аннотации являются

веб-сервисы FastAnnotator [39], EggNog [40], TRAPID [41], InterProScan [42–47] с генерацией HTML отчета.

Дальнейшими перспективами исследований методологии и алгоритмики обработки данных транскриптомов растений являются: освоение и применение новых программных инструментов для de novo сборки и постобработки [48], *in silico* выделение и изучение некодирующей РНК [49–68], улучшение и оптимизация автоматизации и организации обработки данных транскриптомов.

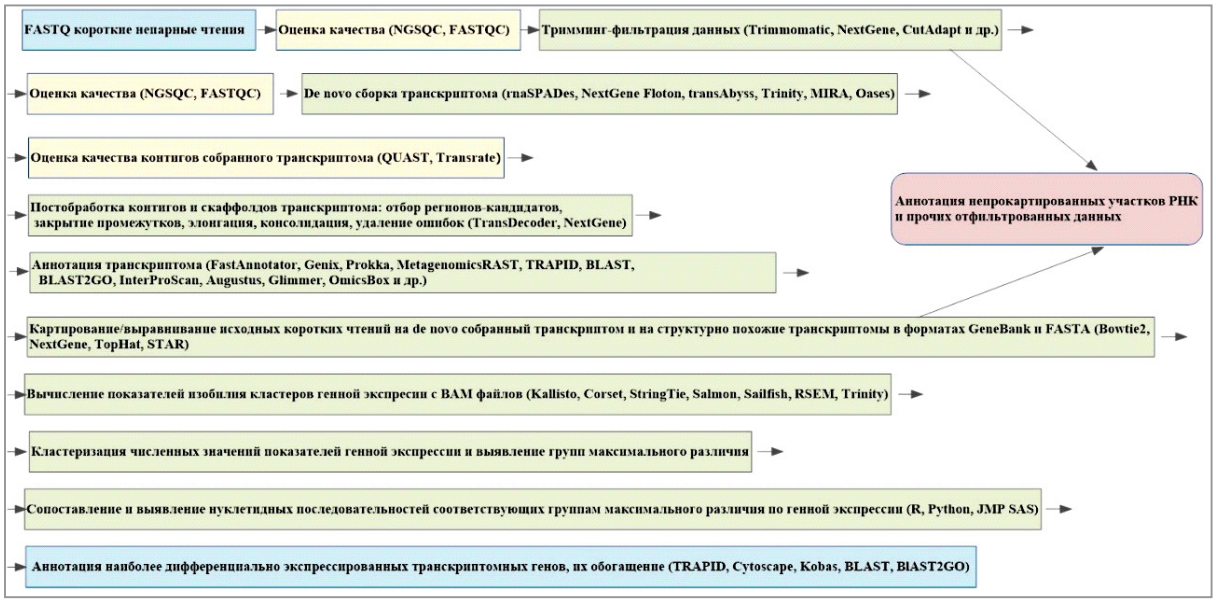

Рис. 1. Разработанный на основе анализа литературных данных и предыдущих исследованиях [5–9] общий алгоритм обработки данных транскриптомов растений. Представляет основные шаги обработки данных и необходимые программные инструменты.

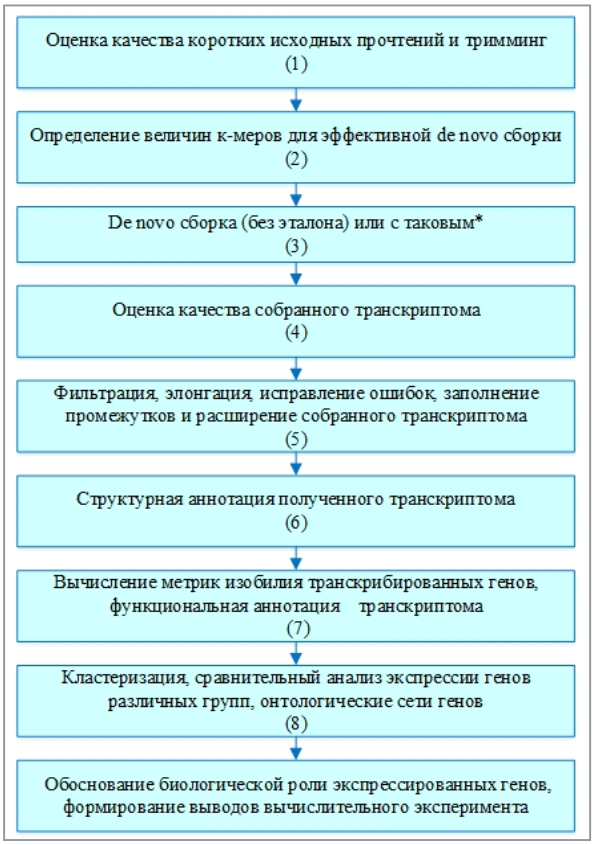

Рис. 2. Разработанный частный практический алгоритм обработки данных транскриптомов растений.

Программные компоненты этапов, которые можно использовать для реализации данного алгоритма: (1) FastQC, Trimmomatic; (2) Kmergenie, NextGene, эмпирически или автоматически; (3) rnaSPADes, MIRA, Nextgene Floton, DeBrujin; (4) Quast; (5) TransDecoder, CD-HIT-EST, NextGene; (6) Genix, Augustus, tRNAScan, Glimmer, BLAST; (7) FastAnnotator, TRAPID; InterProScan, EggNOG-mapper (8) TRAPID, ShinyGO, Sailfish, Cufflinks, Kallisto

Рис. 3. Схема разработанного алгоритма сборки и постобработки траснкриптомных данных. Алгоритм был разработан с целью сохранения информации об исходных данных и для конечного результата лучшей аннотации. На примере метатранскриптома показал эффективность, отображенную в метриках качества сборок и полноты аннотации

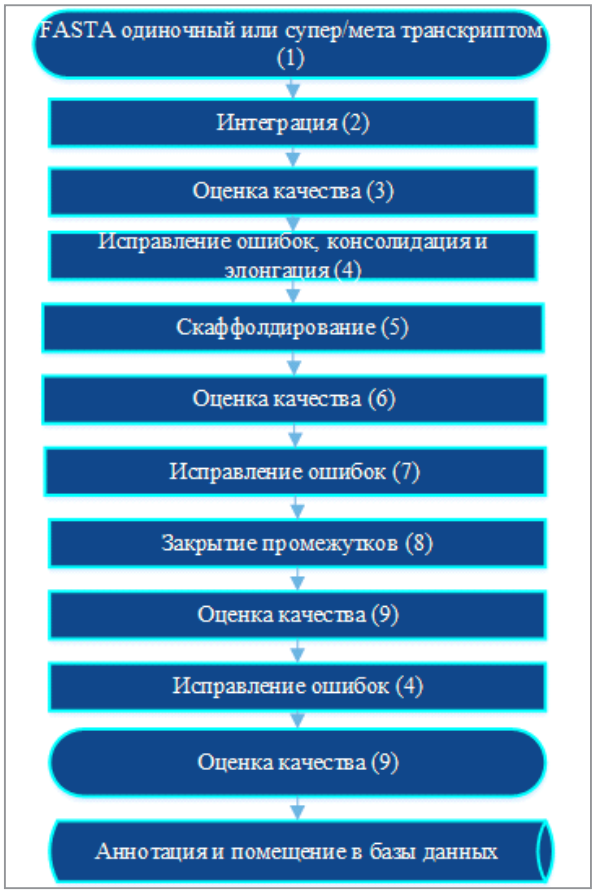

Рис. 4. Концептуальный алгоритм для улучшения качества собранных *de novo* транскриптомов. Алгоритм актуален и для геномов различного происхождения

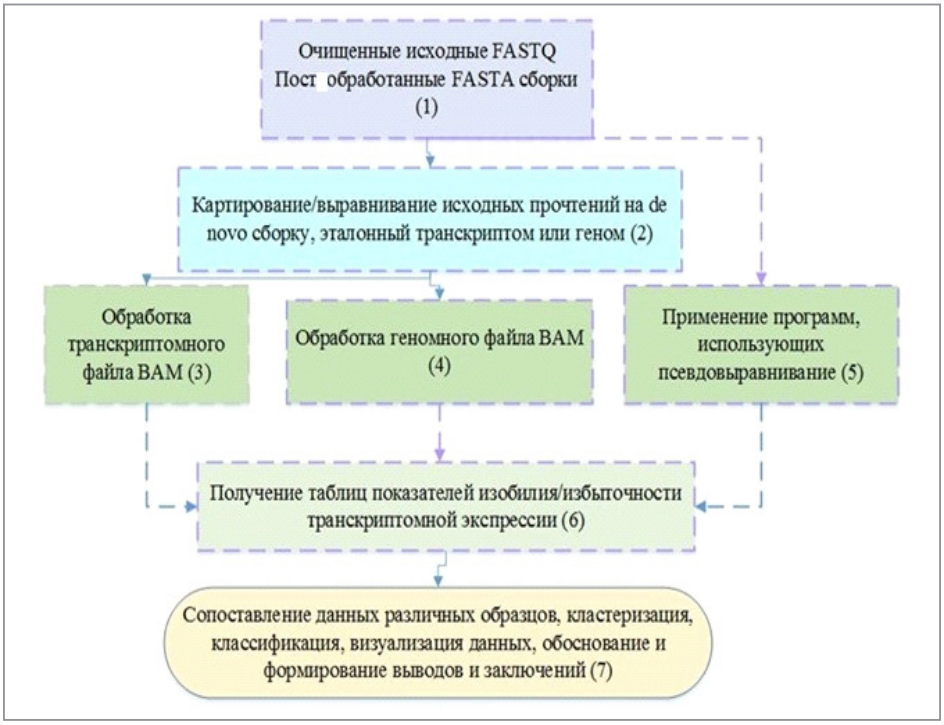

Рис. 5. Концептуальный алгоритм обработки данных, предназначенный для извлечения и анализа информации экспрессии генов исследуемого транскриптома

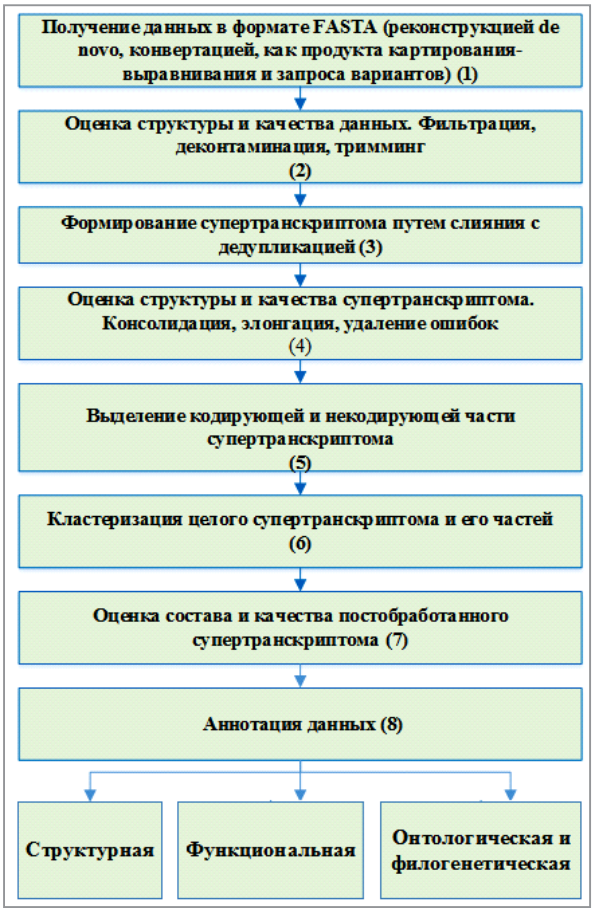

Рис. 6. Алгоритм обработки данных транскриптомов растений, позволяющий, в отличие от аналогов, максимизировать эффективность аннотации и улучшить объем и качество биологически интерпретируемой информации

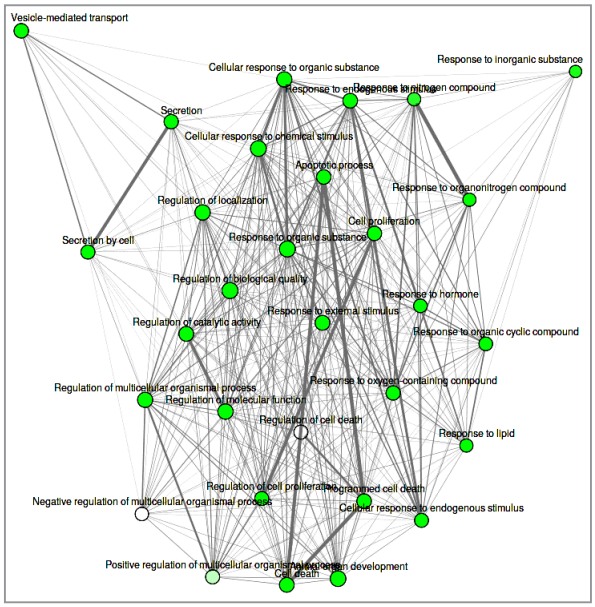

Рис. 7. Метаболическая сеть на основе генной онтологии, с использованием инструментов ShinyGo, для образца патосистемы сосна-фитоплазма (на основе анализа данных полученных в [37])

ЛИТЕРАТУРА

1. **Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A. et al.** A survey of best practices for RNA-seq data analysis // Genome biology.– 2016.– V. 17, № 1.– P. 13.
2. **Eldem, V., Zararsiz, G., Taşçi, T., Duru, I.P., Bakir, Y. et al.** Transcriptome Analysis for Non-Model Organism: Current Status and Best-Practices // Applications of RNA-Seq and Omics Strategies-From Microorganisms to Human Health.– 2017.– V. 1, № 2.– P. 1–19.
3. **Liu, X., Li, N., Liu, S., Wang, J., Zhang, N. et al.** Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review // Front Bioeng Biotechnol.– 2019.– V. 7,– P. 358.
4. **Mutz, K. O., Heilkenbrinker, A., Lönne, M., Walter, J.-G., Stahl, F.** Transcriptome analysis using next-generation sequencing // Current opinion in biotechnology.– 2013.– V. 24, № 1.– P. 22–30.
5. **Можаровская, Л. В.** Идентификация и функциональная аннотация патоген-индуцированных генов проростков сосны обыкновенной / Л. В. Можаровская, С. В. Пантелеев, О. Ю. Баранов, В. Е. Падутов // Молекулярная и прикладная генетика: сб. науч. тр. / Институт генетики и цитологии НАН Беларуси; редкол.: А. В. Кильчевский (гл. ред.) [и др.].– Минск: Институт генетики и цитологии НАН Беларуси, 2019.– Т. 26.– С. 69–78.
6. **Можаровская, Л. В.** Сравнительный анализ транскрипционных профилей проростков сосны обыкновенной (Pinus sylvestris L.) различающихся температурными условиями выращивания / Л. В. Можаровская // Проблемы лесоведения и лесоводства: Сб. науч. Трудов ИЛ НАН Беларуси.– Вып. 78.– Гомель: ИЛ НАН Беларуси, 2018.– С. 70–78.
7. **Можаровская Л. В., Пантелеев С. В., Разумова О. А., Баранов О. Ю.** Выявление сайтов редактирования мРНК в хлоропластном геноме сосны обыкновенной (Pinus sylvestris L.) Сборник научных трудов [Институт леса Национальной академии наук Беларуси]/ Национальная академия наук Беларуси, Институт леса.– Гомель, 2019.– Вып. 79: Проблемы лесоведения и лесоводства.– С. 54–61
8. **Кирьянов П. С., Баранов О. Ю., Падутов В. Е.** Выявление генетических особенностей среди форм березы повислой, различающихся по признаку узорчатости древесины // Лесное хозяйство: материалы 84-й науч.-техн. кон-

ференции профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным участием), Минск, 03–14 февраля 2020 г. / отв. за издание И. В. Войтов; УО БГТУ.– Минск: БГТУ, 2020.– С. 106–107.
9. **Падутов В. Е., Третьякова И. Н., Можаровская Л. В. Константинов А. В., Кулагин Д. В., Кусенкова М. П.** Сравнительный анализ транскрипционных профилей каллусных культур лиственницы сибирской с различным эмбриогенным потенциалом // Лесное хозяйство: материалы 84-й науч.-техн. конференции профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным участием), Минск, 03–14 февраля 2020 г. / отв. за издание И. В. Войтов; УО БГТУ.– Минск: БГТУ, 2020.– С. 131.
10. **Wang Z., Gerstein M., Snyder M.** RNA-Seq: a revolutionary tool for transcriptomics // Nature reviews genetics.– 2009.– V. 10.– № . 1.– P. 57–63.
11. **Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D. et al.** De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis // Nat Protoc.– 2013.– V. 8, № 8.– P. 1494–512.
12. **Wang, Y., Sun, M.-a.** Transcriptome Data Analysis: Methods and Protocols. Springer, 2018.
13. [Электронный ресурс] – Режим доступа: http://bioinformaticsinstitute.ru/sites/default/files/07–28–04-kasyanov.pdf.– Дата доступа: 04.09.2020.
14. **Касьянов А. С.** Новые методы обработки данных, полученных с помощью современных технологий секвенирования, для решения задач анализа экспрессии генов: автореф. дисс. канд. физ.-мат. наук.– 2012.
15. **Водясова Е. А., Челебиева Э. С., Кулешова О. Н.** Новейшие технологии высокопроизводительного секвенирования транскриптома отдельных клеток //Вавиловский журнал генетики и селекции.– 2019.– Т. 23.– № . 5.– С. 508–518.; Акберова Н. И. Анализ данных секвенирования транскриптома и метаболома: учебно-методическое пособие.– 2014.– 26 с.
16. **Ewing B., Green P.** Base-calling of automated sequencer traces using phred. II. Error probabilities // Genome research.– 1998.– V. 8.– № . 3.– P. 186–194.
17. **Brown, J., Pirrung, M., McCue, L. A.** FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool // Bioinformatics.– 2017.– V. 1, № 1.– P. 1–9.
18. **Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M.H. et al.** NGSQC: cross-platform quality analysis pipeline for deep sequencing data // BMC Genomics.– 2010.– V. 11 Suppl 4,– P. S7.
19. **Романенков К. В.** Метод оценки качества сборки генома на основе частот k-меров // Препринты ИПМ им. М. В. Келдыша. 2017. № 11. 24 с. doi:10.20948/prepr-2017-11.
20. **Giannoulatou, E., Park, S. H., Humphreys, D. T., Ho, J. W.** Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie // BMC Bioinformatics.– 2014.– V. 15 Suppl 16,– P. S15.
21. **Langdon, W. B.** Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks // BioData Min.– 2015.– V. 8, № 1.– P. 1.
22. **Lu, R., Zhang, J., Liu, D., Wei, Y. L., Wang, Y. et al.** Characterization of bHLH/HLH genes that are involved in brassinosteroid (BR) signaling in fiber development of cotton (Gossypium hirsutum) // BMC Plant Biol.– 2018.– V. 18, № 1.– P. 304.
23. **Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. et al.** TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions // Genome Biol.– 2013.– V. 14, № 4.– P. R36.
24. **Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M. et al.** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // J Comput Biol.– 2012.– V. 19, № 5.– P. 455–77.
25. **Bankar, K. G., Todur, V. N., Shukla, R. N., Vasudevan, M.** Ameliorated de novo transcriptome assembly using Illumina paired end sequence data with Trinity Assembler // Genom Data.– 2015.– V. 5,– P. 352–9.
26. **Cabau, C., Escudie, F., Djari, A., Guiguen, Y., Bobe, J. et al.** Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies // PeerJ.– 2017.– V. 5,– P. e2988.
27. **Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D. et al.** De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis // Nat Protoc.– 2013.– V. 8, № 8.– P. 1494–512.
28. **Kim, C. S., Winn, M. D., Sachdeva, V., Jordan, K. E.** K-mer clustering algorithm using a MapReduce framework: application to the parallelization of the Inchworm module of Trinity // BMC Bioinformatics.– 2017.– V. 18, № 1.– P. 467.
29. **Cabau, C., Escudie, F., Djari, A., Guiguen, Y., Bobe, J. et al.** Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies // PeerJ.– 2017.– V. 5,– P. e2988.
30. **Schulz, M. H., Zerbino, D. R., Vingron, M., Birney, E.** Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels // Bioinformatics.– 2012.– V. 28, № 8.– P. 1086–92.
31. **Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R. et al.** De novo transcriptome assembly with ABySS // Bioinformatics.– 2009.– V. 25, № 21.– P. 2872–7.
32. **Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S. et al.** ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter // Genome Res.– 2017.– V. 27, № 5.– P. 768–777.
33. **Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. et al.** ABySS: a parallel assembler for short read sequence data // Genome Res.– 2009.– V. 19, № 6.– P. 1117–23.
34. **Boerner, S., McGinnis, K. M.** Computational Analysis of LncRNA from cDNA Sequences // Methods In Molecular Biology (Clifton, N.J.).– 2016.– V. 1402,– P. 255–269.
35. **Ge, S., Jung, D.** ShinyGO: a graphical enrichment tool for animals and plants. 2018.
36. **Zhang C. et al.** Evaluation and comparison of computational tools for RNA-seq isoform quantification //BMC genomics.– 2017.– V. 18.– № . 1.– P. 583.

37. **Пантелеев, С. В.** Молекулярно-генетическая диагностика инфекционных агентов побегов сосны обыкновенной с признаками «ведьминых метел» / С. В. Пантелеев, О. Ю. Баранов, И. Э. Рубель // Сб. науч. тр. / НАН Беларуси, Институт леса.– Гомель, 2016.– Вып. 76: Проблемы лесоведения и лесоводства.– С. 242–249.
38. **Kremer, F. S., Eslabao, M. R., Dellagostin, O. A., Pinto, L. D.** Genix: a new online automated pipeline for bacterial genome annotation // FEMS Microbiol Lett.– 2016.– V. 363, № 23.
39. **T. W., Gan, R. C., Wu, T. H., Huang, P. J., Lee, C. Y. et al.** FastAnnotator – an efficient transcript annotation web tool // BMC Genomics.– 2012.– V. 13 Suppl 7,– P. S9.
40. **Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D. et al. e**ggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences // Nucleic Acids Research.– 2016.– V. 44, № D1.– P. D286-D293.
41. **Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y. et al.** TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes // Genome Biol.– 2013.– V. 14, № 12.– P. R134.
42. **Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W. et al.** InterProScan 5: genome-scale protein function classification // Bioinformatics.– 2014.– V. 30, № 9.– P. 1236–40.
43. **Kelly, R. J., Vincent, D. E., Friedberg, I.** IPRStats: visualization of the functional potential of an InterProScan run // BMC Bioinformatics.– 2010.– V. 11 Suppl 12.– P. S13.
44. **Mulder, N., Apweiler, R.** InterPro and InterProScan: tools for protein sequence classification and comparison // Methods Mol Biol.– 2007.– V. 396,– P. 59–70.
45. **Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N. et al.** InterProScan: protein domains identifier // Nucleic Acids Research.– 2005.– V. 33, № Web Server issue.– P. W116–20.
46. **Syed, A., Upton, C.** Java GUI for InterProScan (JIPS): a tool to help process multiple InterProScans and perform ortholog analysis // BMC Bioinformatics.– 2006.– V. 7,– P. 462.
47. **Zdobnov, E. M., Apweiler, R.** InterProScan – an integration platform for the signature-recognition methods in InterPro // Bioinformatics.– 2001.– V. 17, № 9.– P. 847–8.
48. **Kremer, F. S., McBride, A. J. A., Pinto, L. d. S.** Approaches for in silico finishing of microbial genome sequences // Genetics and molecular biology.– 2017.– V. 40, № 3.– P. 553–576.
49. **Abbas, Q., Raza, S. M., Biyabani, A. A., Jaffar, M. A.** A review of computational methods for finding non-coding RNA genes // Genes.– 2016.– V. 7, № 12.– P. 113.
50. **Abernathy, J., Overturf, K.** Expression of Antisense Long Noncoding RNAs as Potential Regulators in Rainbow Trout with Different Tolerance to Plant-Based Diets // Animal Biotechnology.– 2017.– V. 2, № 1.– P. 1–8.
51. **Andreia, S. R., Inês, C., Bruno Vasques, C., Yao-Cheng, L., Susana, L. et al.** Small RNA profiling in Pinus pinaster reveals the transcriptome of developing seeds and highlights differences between zygotic and somatic embryos // Scientific Reports.– 2019.– № 1.– P. 1.
52. **Babarinde, I. A., Li, Y., Hutchins, A. P.** Computational methods for mapping, assembly and quantification for coding and non-coding transcripts // Computational and structural biotechnology journal.– 2019.– V. 1, № 1.– P. 2–14.
53. **Bai, Y., Dai, X., Harrison, A. P., Chen, M.** RNA regulatory networks in animals and plants: A long noncoding RNA perspective // Briefings In Functional Genomics.– 2015.– V. 14, № 2.– P. 91–101.
54. **Boerner, S., McGinnis, K. M.** Computational Analysis of LncRNA from cDNA Sequences // Methods In Molecular Biology (Clifton, N.J.).– 2016.– V. 1402,– P. 255–269.
55. **Chaturvedi, S., Rao, A. L. N.** Riboproteomics: A versatile approach for the identification of host protein interaction network in plant pathogenic noncoding RNAs // PLoS ONE.– 2017.– V. 12, № 10.
56. **Chaves, I., Costa, B. V., Rodrigues, A. S., Bohn, A., Miguel, C. M.** miRPursuit-a pipeline for automated analyses of small RNAs in model and nonmodel plants // FEBS Letters.– 2017.– V. 591, № 15.– P. 2261–2268.
57. **Collemare, J., O'Connell, R., Lebrun, M. H.** Nonproteinaceous effectors: the terra incognita of plant–fungal interactions // New Phytologist.– 2019.– V. 223, № 2.– P. 590–596.
58. **Dhiman, H., Kapoor, S., Sivadas, A., Sivasubbu, S., Scaria, V.** zflncRNApedia: A Comprehensive Online Resource for Zebrafish Long Non-Coding RNAs // PLoS ONE.– 2015.– V. 10, № 6.– P. e0129997-e0129997.
59. **Fan, B., Wu, X. Q., Li, L., Chao, Y., Förstner, K. et al.** DRNA-seq reveals genomewide TSSs and noncoding RNAs of plant beneficial rhizobacterium Bacillus amyloliquefaciens FZB42 // PLoS ONE.– 2015.– V. 10, № 11.
60. **Hao, Z., Fan, C., Cheng, T., Su, Y., Wei, Q. et al.** Genome-Wide Identification, Characterization and Evolutionary Analysis of Long Intergenic Noncoding RNAs in Cucumber. 2015.
61. **Heera, R., Sivachandran, P., Chinni, S. V., Mason, J., Croft, L. et al.** Efficient extraction of small and large RNAs in bacteria for excellent total RNA sequencing and comprehensive transcriptome analysis // BMC Research Notes.– 2015.– V. 8,– P. 1–11.
62. **Hu, L., Xu, Z., Hu, B., Lu, Z. J.** COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features // Nucleic Acids Research.– 2017.– V. 45, № 1.– P. e2-e2.
63. **Matsui, A., Nakaminami, K., Seki, M. m. s. r. j.** Biological Function of Changes in RNA Metabolism in Plant Adaptation to Abiotic Stress // Plant & Cell Physiology.– 2019.– V. 60, № 9.– P. 1897–1905.
64. **Mingyang, Q., Jinhui, C., Deqiang, Z.** Exploring the Secrets of Long Noncoding RNAs // International Journal of Molecular Sciences.– 2015.– V. 16, № 3.– P. 5467–5496.
65. **Negri, T. D. C., Bugatti, P. H., Saito, P. T. M., Domingues, D. S., Paschoal, A. R. et al.** Pattern recognition analysis on long noncoding RNAs: A tool for prediction in plants // Briefings in Bioinformatics.– 2019.– V. 20, № 2.– P. 682–689.

66. **Ortogero, N., Hennig, G.W., Langille, C., Ro, S., Yan, W. et al.** Computer-assisted annotation of murine sertoli cell small RNA transcriptome // Biology of Reproduction.– 2013.– V. 88, № 1.
67. **Paschoal, A.R., Lozada-Chávez, I., Domingues, D.S., Stadler, P.F.** ceRNAs in plants: computational approaches and associated challenges for target mimic research // Briefings in Bioinformatics.– 2018.– V. 19, № 6.– P. 1273–1289.
68. **Zongbo, Q., Xiaojuan, L., Yuanyuan, Z., Manman, Z., Yinglang, W. et al.** Genome-wide analysis reveals dynamic changes in expression of microRNAs during vascular cambium development in Chinese fir, Cunninghamia lanceolata // Journal of Experimental Botany.– 2015.– V. 66, № 11.– P. 3041–3054.

Поступила	*После доработки*	*Принята к печати*
25.09.2020	*01.05.2021*	*01.06.2021*

SPRINDZUK M. V.[1], TITOV L. P.[2], KONCHITS A. P.[3], MOZHAROVSKAYA L. V.[3]

MODERN TRANSCRIPTOME DATA PROCESSING ALGORITHMS: A REVIEW OF METHODS AND RESULTS OF APPROBATION

[1] United Institute for Informatics Problems of the National Academy of Sciences of Belarus
[2] RSPC Microbiology and Epidemiology
[3] Institute of Forest research of the National Academy of Sciences of Belarus

Analysis of bioinformatics data is an actual problem in modern computational biology and applied mathematics. With the development of biotechnology and tools for obtaining and processing such information, unresolved issues of the development and application of new algorithms and software have emerged.

Authors propose practical algorithms and methods for processing transcriptomic data for efficient results of annotation, visualization and interpretation of bioinformatics data.

Keywords: *transcriptome, genomics, bioinformatics, data analysis, software, algorithms.*