УДК 004.9:004.9:004.8:577.21

Современные алгоритмы обработки данных транскриптомов: обзор методов и результаты апробации

М. В. Спринджук, к. т. н., старший научный сотрудник лаборатории математической кибернетики

E-mail: stepanenkomatvei@yandex.ru

Объединенный институт проблем информатики НАН Беларуси, ул. Сурганова, д. 6, 220012, г. Минск, Республика Беларусь

Л. В. Можаровская, научный сотрудник лаборатории геномных исследований и биоинформатики

E-mail: milamozh@yandex.ru

Институт леса НАН Беларуси, ул. Пролетарская, д. 71, 246001, г. Гомель, Республика Беларусь

А. П. Кончиц, к. б. н., ведущий научный сотрудник лаборатории лесной селекции и семеноводства

E-mail: konchits@yandex.ru

Институт леса НАН Беларуси, ул. Пролетарская, д. 71, 246001, г. Гомель, Республика Беларусь

Л. П. Титов, д. м. н., профессор, член-корреспондент НАН Беларуси, заведующий лабораторией клинической и экспериментальной микробиологии РНПЦ эпидемиологии и микробиологии

РНПЦ эпидемиологии и микробиологии, ул. Филимонова, д. 23, 220114, г. Минск, Республика Беларусь

Аннотация. Анализ биоинформатических данных является актуальной проблемой современной вычислительной биологии и прикладной математики. С развитием биотехнологий, а также инструментальных средств получения и обработки информации о биологических объектах и системах, 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

Для цитирования: Спринджук, М. В. Современные алгоритмы обработки данных транскриптомов: обзор методов и результаты апробации / М. В. Спринджук, Л. В. Можаровская, А. П. Кончиц, Л. П. Титов // Цифровая трансформация. – 2021. – № 1 (14). – С. 53–64.

Modern Tanscriptome Data Rrocessing Algorithms: a Review of Methods and Results of Approbation

M. V. Sprindzuk, Candidate of Science (Technical), Senior Researcher, Laboratory of Mathematical Cybernetics

E-mail: stepanenkomatvei@yandex.ru

United Institute for Informatics Problems of the NAS of Belarus, 6 Surganova Str., 220012 Minsk, Republic of Belarus

L. V. Mozharovskaya, Researcher, Laboratory of Genomics Research and Bioinformatics

E-mail: milamozh@yandex.r

Forest Research Institute of the NAS of Belarus, 71 Proletarskaya Str., 246001 Gomel, Republic of Belarus

A. P. Konchits, Candidate of Science (Biological), Leading Researcher, Forest Tree Breeding and Seed Production Laboratory

E-mail: konchits@yandex.ru

Forest Research Institute of the NAS of Belarus, 71 Proletarskaya Str., 246001 Gomel, Republic of Belarus

L. P. Titov, Doctor of Sciences (Medical), Professor, Corresponding Member of the NAS of Belarus, Head of the Laboratory for Clinical and Experimental Microbiology

RRPC for Epidemiology and Microbiology, Republic of Belarus, 23 Filimonova Str., 220114 Minsk, Republic of Belarus

Abstract. Analysis of bioinformatics data is an actual problem in modern computational biology and applied mathematics. With the development of biotechnology, as well as tools for obtaining and processing information derived from biological objects and systems, unresolved issues of the development and application of new algorithms and software have emerged. The authors propose practical algorithms and methods for processing transcriptome data for effective results of annotation, visualization and interpretation of data.

Key words: transcriptome, genomics, bioinformatics, data analysis, software, algorithms

For citation: Sprindzuk M. V., Mozharovskaya L. V., Konchits A. P., Titov L. P. Modern Tanscriptome Data Rrocessing Algorithms: a Review of Methods and Results of Approbation. *Cifrovaja transformacija* [Digital transformation], 2021, 1 (14), pp. 53–64 (in Russian).

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 наилучшего качества и оптимального объема. На рисунках 1-6 схематически представлены авторские алгоритмы обработки транскриптомных данных, успешно апробированные на транскриптомах, полученных в лаборатории геномных исследований и биоинформатики Института леса НАН Беларуси (Гомель, Беларусь) [5 – 9]. Алгоритмы были разработаны с целью сохранения информации об исходных данных и для конечного результата лучшего аннотирования. Более подробная информация о конкретных элементах каждого из предлагаемых алгоритмов представлена в соответствующей документации программного обеспечения (https://github.com/).

Разработка эффективных алгоритмов анализа данных транскриптомов представляется междисциплинарной задачей, требующей знаний объектно-ориентированного 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 аннотирования (определения местоположения и функциональной принадлежности) на референсный транскриптом. В таком случае выполняется анализ вариантов с последующим аннотированием по направлению транскрипции, 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* (фрагментов подстрок геномного текста длины *k*) для эффективной *de novo* (от лат. *de novo* – заново, впервые) сборки – без картирования на референсный геном. Качество собранного генома оценивается по количеству контигов – строк из нуклеотидов, представляющих консенсусную последовательность ДНК; числу ошибок сборки и объему достоверного аннотрования генов – определения их местоположения и функциональной принадлежности [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 в том случае, когда 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 ≈ 100 $), NextGene, Converge, CLC Genomics, JMP Genomics (стоимость лицензии ≈ 16 000 €). В случае отсутствия референсного транскриптома, в качестве эталона можно использовать родственный вид (как, например, *Arabidopsis thaliana* для растений, отмечен символом «*» на рисунке 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], NextGeneFloton и другие. Более длинные прочтения или прочтения с парными концами (как при секвенировании обеих цепей ДНК) способствуют получению лучших результатов *de novo* сборки. С целью формирования множества эталонных консенсусных последовательностей рекомендуется комбинировать множество транскриптомов для получения единой комбинированной сборки. В дальнейшем эталонные консенсусные последовательности можно использовать для картирования, подсчета экспрессии и сравнения между группами образцов транскриптомов.

Аннотирование собранного транскриптома. Аннотирование (в литературе встречается также термин *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/). 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 белков. Аннотирование 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 в соответствии с доменным поиском. 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.

Таблица 1. Технические характеристики программного обеспечения, предназначенного для вычисления экспрессии генов транскриптома [36]. МП – максимальное правдоподобие, ВБ – вариационный метод Байеса
Table 1. Technical characteristics of the software tool designed to calculate transcriptome gene expression [36]. Maximum Likelihood, Variational Bayesian Method.

Название	Оперативная память (ГБ)	Время затрат	Алгоритм	Мультипоточность
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

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 консенсусной последовательностью ДНК, несущей 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 тысячу нуклеотидных оснований на миллион картированных прочтений, RPKM (с англ. reads per kilobase per million mapped reads); число фрагментов на тысячу нуклеотидных оснований на миллион картированных прочтений и число транскриптов на миллион картированных прочтений, FRKM и TPM соответственно (с англ., fragments или transcripts per million reads). Перечисленные вычислительные методы реализованы в алгоритмах бесплатного программного обеспечения CuffLinks, RSEM, eXpress, Kallisto и представлены в таблице 1.

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 прочтений транскриптома вычисляется как среднее от нормализованных данных.

Наиболее эффективными программами для количественной оценки полученных данных и качественного аннотирования являются веб-сервисы FastAnnotator [37], EggNog [38], TRAPID [39], InterProScan [42-45] с генерацией HTML отчета.

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.

Так, с целью сохранения информации об исходных данных и для конечного результата лучшего аннотирования, нами были разработаны и успешно апробированы [5-9] алгоритмы и методологические основы обработки данных транскриптомов растений, авторские алгоритмы схематически представлены на рисунках 1-6.

Как видно из рисунка 1, в разработанном общем алгоритме обработки данных транскриптомов растений представлены необходимые программные инструменты и основные шаги обработки данных: от *de novo* сборки полученных коротких непарных чтений до их картирования и аннотации.

Практический алгоритм обработки данных транскриптомов растений представлен на рисунке 2. Для реализации каждого этапа данного алгоритма рекомендуется использование следующих программных компонентов: (1) FastQC, Trimmomatic; (2) Kmergenie, NextGene, выбор k-меров 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;

Рис. 1. Разработанный общий алгоритм обработки данных транскриптомов растений
Fig. 1. Developed general algorithm for processing plant transcriptome data

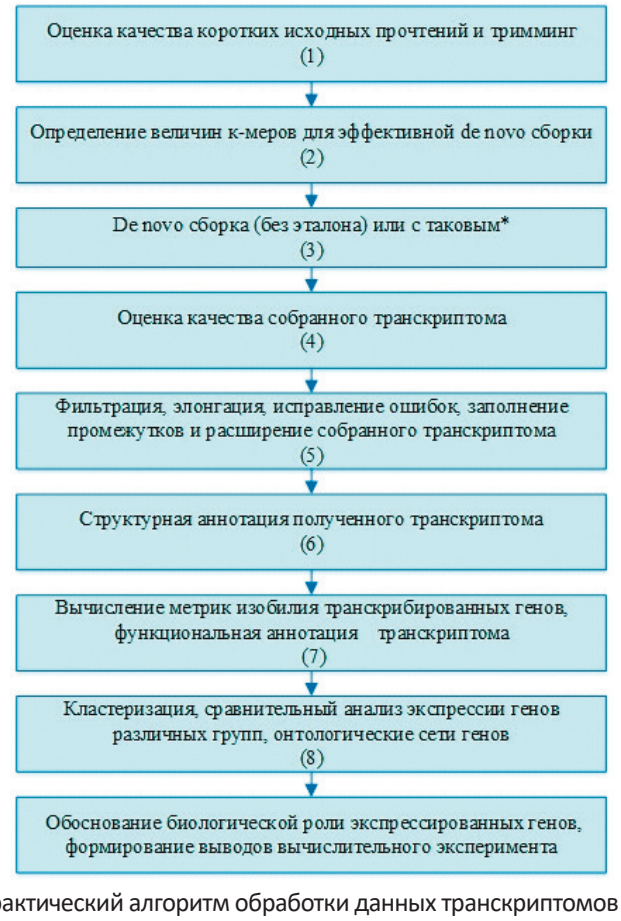

Рис. 2. Практический алгоритм обработки данных транскриптомов растений
Fig. 2. Practical algorithm for processing plant transcriptome data

InterProScan, EggNOG-mapper; (8) TRAPID, ShinyGO, Sailfish, Cufflinks, Kallisto.

Алгоритм сборки и постобработки трасннкриптомных данных (рисунок 3), включает следующие шаги: de novo сборку транскриптома; слияние полученных контигов – строк из нуклеотидов, представляющих консенсусную последовательность ДНК, с удалением повторов; оценку структуры и качества супертранскриптома; выделение кодирующей (участка последовательности ДНК несущего информацию о белке) и некодирующей части; кластеризацию супертранскриптома; постобработку, включающую закрытие промежутков последовательностей, консолидацию, элонгацию и удаление ошибок; оценку состава и качества; структурную, функциональную и онтологическую аннотацию супертранскриптома.

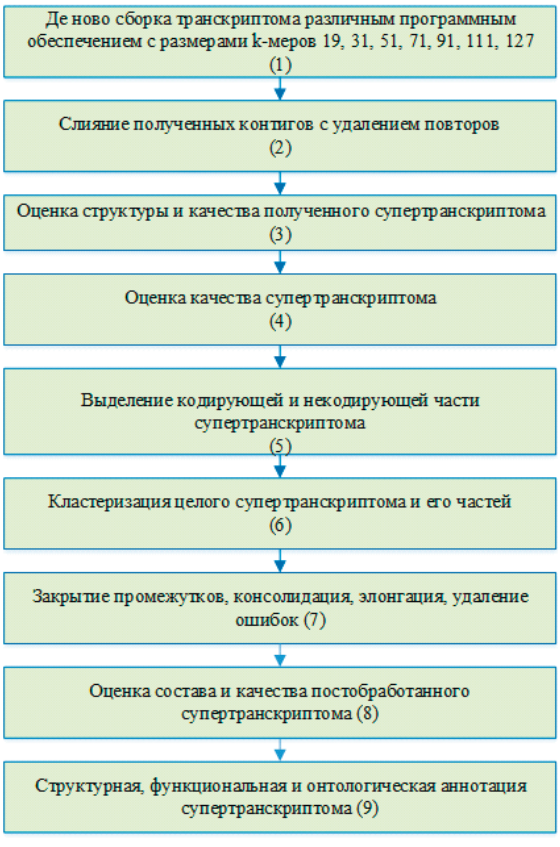

Рис. 3. Схема разработанного алгоритма сборки и постобработки траснкриптомных данных
Fig. 3. Scheme of the developed algorithm for collection and post-processing of transcriptomic data

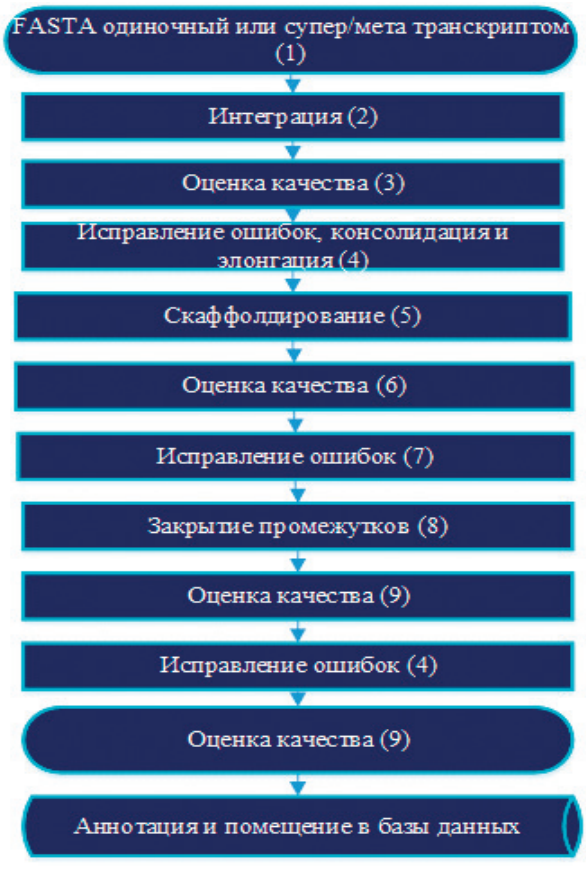

Рис. 4. Концептуальный алгоритм для улучшения качества собранных de novo транскриптомов
Fig. 4. Conceptual algorithm for improving the quality of assembled de novo transcriptomes

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

Fig. 5. Conceptual data processing algorithm designed to extract and analyze gene expression information of the transcriptome under study

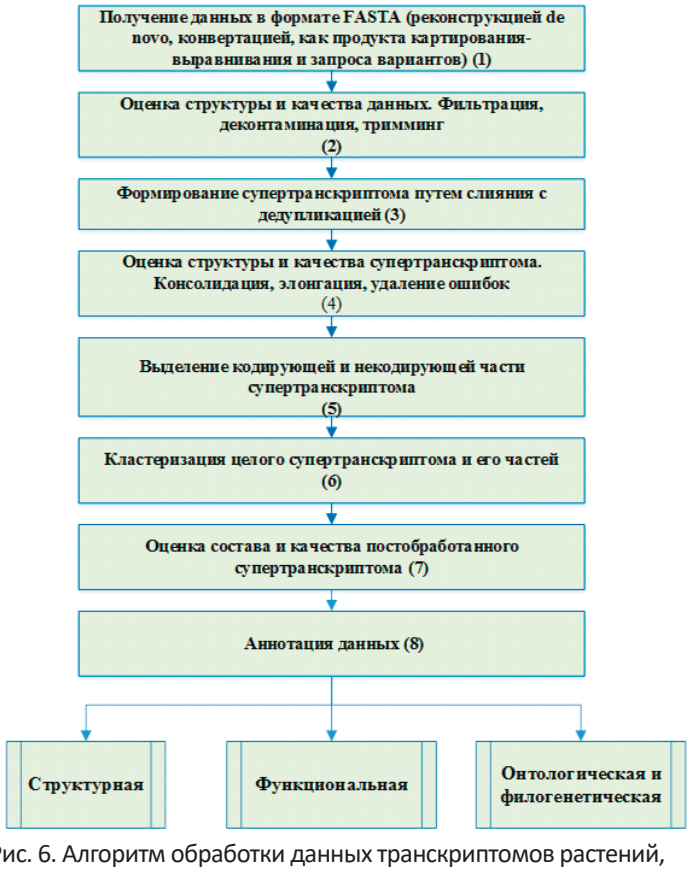

Рис. 6. Алгоритм обработки данных транскриптомов растений, позволяющий максимизировать эффективность аннотации и улучшить объем и качество биологически интерпретируемой информации

Fig. 6. Algorithm for processing plant transcriptome data that maximizes annotation efficiency and improves the amount and quality of biologically interpreted information

Алгоритмы для улучшения качества собранных de novo транскриптомов и обработки данных для извлечения и анализа информации экспрессии генов исследуемого транскриптома представлены на рисунках 4 и 5.

Разработанный алгоритм обработки данных транскриптомов расте-ний (рисунок 6), в отличие от аналогов, позволяет максимизировать эффективность аннотирования и улучшить объем и качество биологически интерпретируемой ин-

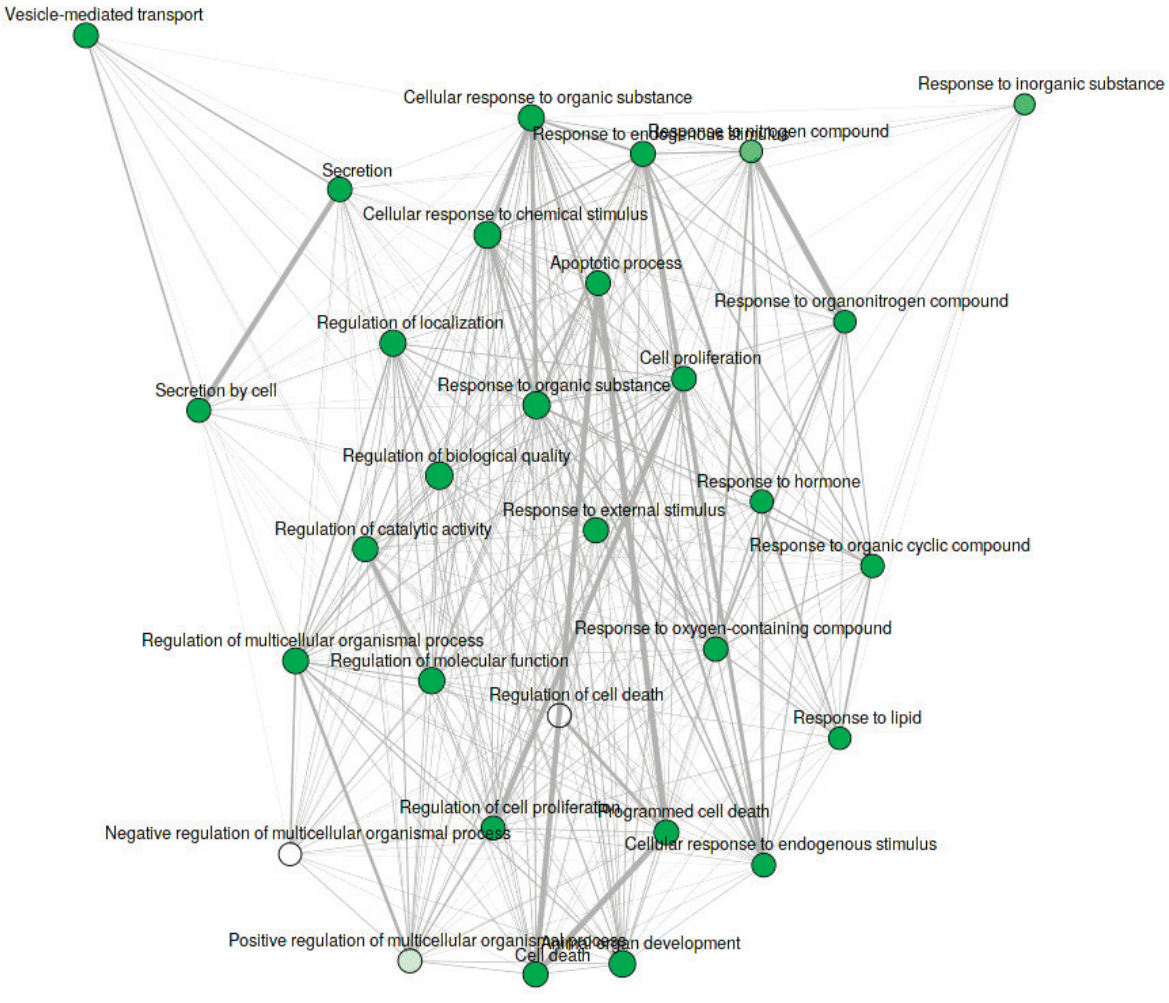

Рис. 7 . Метаболическая сеть на основе генной онтологии с использованием инструментов ShinyGo.
Патосистема сосна – фитоплазма
Fig. 7. Gene ontology-based metabolic network using ShinyGo tools.
Pathosystem - pine-phytoplasma

формации. На этапе оценки структуры и качества данных, включает фильтрацию, деконтаминацию и тримминг – удаление прочтений с низким качеством и малой длины.

На основе проведенного ранее исследования патосистемы сосна обыкновенная – фитоплазма [46], с использованием разработанных алгоритмов (рисунок 1-6) был проведен транскриптомный анализ, с построением метаболической сети на основе генной онтологии и с применением инструментов ShinyGo. Визуализация полученных данных представлена на рисунке 7.

Заключение. Исследования, направленные на развитие биоинформатических методов оценки уровня экспрессии генов транскриптомов, актуальны и нуждаются в дальнейшем изучении. Уровень экспрессии генов при изучении транскриптомов лучше оценивать по отдельным компонентам. При этом необходимо отбирать прочтения генов с одинаковым аннотированием и проследить их метрики в отчетных таблицах программ, с последующим выполнением статистических расчетов по вычислению значимой разницы между группами образцов.

Представленные алгоритмы и методологические основы обработки данных транскриптомов растений позволяют максимизировать эффективность аннотации и улучшить объем и качество биологически интерпретируемой информации.

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

Список литературы

1. Conesa, A. A survey of best practices for RNA-seq data analysis / A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera et al. // Genome biology.– 2016.– V. 17, № 1.– P. 13.
2. Eldem, V. Transcriptome Analysis for Non-Model Organism: Current Status and Best-Practices / V. Eldem, G. Zararsiz, T. Taşçi, I. P. Duru, Y. Bakir, et al.// Applications of RNA-Seq and Omics Strategies-From Microorganisms to Human Health.– 2017.– V. 1, № 2.– Pp. 1-19.
3. Liu, X. Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review / X. Liu, N. Li, S. Liu, J. Wang, N. Zhang et al. // Front BioengBiotechnol.– 2019.– V. 7. – P. 358.
4. Mutz, K.-O. Transcriptome analysis using next-generation sequencing / K.-O. Mutz, A. Heilkenbrinker, M. Lönne, J.-G. Walter, F. Stahl // Current opinion in biotechnology.– 2013.– V. 24, № 1.– P. 22-30.
5. Можаровская, Л. В. Идентификация и функциональная аннотация патоген-индуцированных генов проростковсосны обыкновенной. Л. В. Можаровская, С. В. Пантеле-ев, О.Ю. Баранов, В.Е. Падутов// Молекулярная и прикладная генетика: сб.науч.тр./ Ин-ститут 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 (Pinussylvestris 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. Выявление сайтов редактирования мРНК в хлоропластном геноме сосны обыкновенной (Pinussylvestris L.)/ Л. В. Можаровская, С. В. Пантелеев, О. А. Разумова, О. Ю. Баранов, // 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. Выявление генетических особенностей среди форм березы повислой, различающихся по признаку узорчатости древесины/ П. С. Кирьянов, О. Ю. Баранов, В. Е. Падутов // Лесное хозяйство : материалы 84-й науч.-техн. конференции профессорско-преподавательского состава, научных сотрудников и аспирантов (с международным уча-стием), Минск, 03–14 февраля 2020 г. / отв. за издание И. В. Войтов; УО БГТУ. – Минск: БГТУ, 2020. – С. 106-107.
9. Падутов, В. Е. Сравнительный анализ транскрипционных профилей каллусных культур лиственницы сибирской с различным эмбриогенным потенциалом / В. Е. Падутов, И. Н.Третьякова, Л. В. Можаровская, А. В. Константинов, Д. В. Кулагин, М. П. Кусенкова // 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. RNA-Seq: a revolutionary tool for transcriptomics / Z. Wang, M. Gerstein, M. Snyder // Nature reviews genetics. – 2009. – V. 10. – №. 1. – Pp. 57-63.
11. Haas, B.J. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis / B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood et al. // Nat Protoc.– 2013.– V. 8, № 8.– Pp. 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.Новейшие технологии высокопроиз-водительного секвенированиятранскриптома отдельных клеток / Е. А. Водясова, Э. С. Челебиева, О. Н. Кулешова//Вавиловский журнал генетики и селекции. – 2019. – Т. 23. – №. 5. – С. 508-518.
16. Ewing, B. Base-calling of automated sequencer traces using phred. II. Error proba-bilities / B. Ewing, P. Green, //Genome research. – 1998. – V. 8. – №. 3. – P. 186-194
17. Brown, J. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool / J. Brown, M. Pirrung, L. A. McCue// Bioinformatics.– 2017.– V. 1, № 1.– P. 1-9.
18. Dai, M. NGSQC: cross-platform quality analysis pipeline for deep sequencing data / M. Dai, Thompson, R. C. Maher, R. Contreras-Galindo, M. H. Kaplan et al. // 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 bio-informatics software without a gold standard: a case study of BWA and Bowtie / E. Giannoulatou, S.H. Park, D. T. Humphreys, J. W. Ho // BMC Bioin-formatics.– 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. Characterization of bHLH/HLH genes that are involved in brassinosteroid (BR) signaling in fiber development of cotton (Gossypiumhirsutum) / R. Lu, J. Zhang, D. Liu, Y. L. Wei, Y. Wang, et al. // BMC Plant Biol.– 2018.– V. 18, № 1.– P. 304.
23. Kim, D. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions /Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. et al. // Genome Biol.– 2013.– V. 14, № 4.– P. R36.

24. Bankevich, A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing / Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M. et al.// J Comput Biol.– 2012.– V. 19, № 5.– P. 455-77.
25. Bankar, K.G., Ameliorated de novo transcriptome assembly using Illumina paired end sequence data with Trinity Assembler / Todur, V.N., Shukla, R.N., Vasudevan, M.// Genom Data.– 2015.– V. 5. – P. 352-9.
26. Cabau, C. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies / Cabau, C., Escudie, F., Djari, A., Guiguen, Y., Bobe, J. et al.// PeerJ.– 2017.– V. 5. – P. e2988.
27. Haas, B. J. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis / Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D. et al.// Nat Protoc.– 2013.– V. 8, № 8.– P. 1494-512.
28. Kim, C.S. K-mer clustering algorithm using a MapReduce framework: application to the parallelization of the Inchworm module of Trinity / Kim, C.S., Winn, M.D., Sachdeva, V., Jordan, K.E.// BMC Bioinformatics.– 2017.– V. 18, № 1.– P. 467.
29. Cabau, C. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies / Cabau, C., Escudie, F., Djari, A., Guiguen, Y., Bobe, J. et al.// PeerJ.– 2017.– V. 5, – P. e2988.
30. Schulz, M. H. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels / Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E./ Bioinformatics.– 2012.– V. 28, № 8.– P. 1086-92.
31. Birol, I. De novo transcriptome assembly with ABySS / Birol, I., Jackman, S.D., Nielsen, C. B., Qian, J. Q., Varhol, R. et al.// Bioinformatics.– 2009.– V. 25, № 21.– P. 2872-7.
32. Jackman, S.D. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter / Jackman, S.D., Vandervalk, B.P., Mohamadi, H., Chu, J., Yeo, S. et al.// Genome Res.– 2017.– V. 27, № 5.– P. 768-777.
33. Simpson, J.T. ABySS: a parallel assembler for short read sequence data / Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.// Genome Res.– 2009.– V. 19, № 6.– P. 1117-23.
34. Boerner, S. Computational Analysis of LncRNA from cDNA Sequences /Boerner, S., McGinnis, K.M. // 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 quan-tification //BMC genomics. – 2017. – V. 18. – №. 1. – P. 583.
37. Chen, 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.
38. Huerta-Cepas, J. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences /Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D. et al. / Nucleic Acids Research.– 2016.– V. 44, № D1.– P. D286-D293.
39. Van Bel, M. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes /Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y. et al. // Genome Biol.– 2013.– V. 14, № 12.– P. R134.
40. Jones, P. InterProScan 5: genome-scale protein function classification / Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W. et al.// Bioinformatics.– 2014.– V. 30, № 9.– P. 1236-40.
41. Kelly, R.J. IPRStats: visualization of the functional potential of an InterProScan run /Kelly, R.J., Vincent, D.E., Friedberg, I. // BMC Bioinformatics.– 2010.– V. 11 Suppl 12. – P. S13.
42. Mulder, N. InterPro and InterProScan: tools for protein sequence classification and comparison / Mulder, N., Apweiler, R.// Methods Mol Biol.– 2007.– V. 396, – P. 59-70.
43. Quevillon, E. InterProScan: protein domains identifier /Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N. et al. / Nucleic Acids Research.– 2005.– V. 33. № Web Server issue.– P. W116-20.
44. Syed, A. Java GUI for InterProScan (JIPS): a tool to help process multiple InterProScans and perform ortholog analysis / Syed, A., Upton, C.// BMC Bioinformatics.– 2006.– V. 7, – P. 462.
45. Zdobnov, E.M. InterProScan--an integration platform for the signature-recognition methods in InterPro /Zdobnov, E.M., Apweiler, R. // Bioinformatics.– 2001.– V. 17, № 9.– P. 847-8.
46. Пантелеев, С. В. Молекулярно-генетическая диагностика инфекционных агентов по-бегов сосны обыкновенной с признаками «ведьминых метел» / С. В. Пантелеев, О. Ю. Ба-ранов, И. Э. Рубель // Сб. науч. тр. / НАН Беларуси, Институт леса. – Гомель, 2016. – Вып. 76 : Проблемы лесоведения и лесоводства. – С. 242–249.

References

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. 13 p.
2. Elden 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. pp. 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 BioengBiotechnol, 2019, V. 7. 358 p.
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. pp. 22-30.
5. Mozharovskaya L.V., Panteleev S.V., Baranov O.Yu., Padutov V.E. Identification and Functional Annotation of Pathogen-

Induced Genes of the Scots Pine Seedlings. Molecular and Applied Genetics, Minsk, 2019, V. 26. pp.69-78. (in Russian).
6. Mozharovskaya, L.V. Comparative Analysis of the Transcription Profiles from Pine Seedlings (Pinus Sylvestris L.) Grown Under Various Temperature Conditions. Problemy lesovedeniya i lesovodstva, Gomel, V. 78. pp. 70-78. (in Russian).
7. Mozharovskaya L.V., Panteleev S. V., Razumova O.A., Baranov O. Yu. Identification of mRNA Editing Sites in the Chloroplast Genome Of Pine (Pinus Sylvestris L.). Problemy lesovedeniya i lesovodstva, Gomel, 2019, V. 79. pp. 54-61. (in Russian).
8. Kiryanov P.S., Baranov O. Yu., Padutov V.E., Identification of Genetic Features Among the Forms of Silver Birch, Differing by the Characteristic of Wood Patterning // Forestry: materials of the 84th scientific-technical. conferences of faculty, researchers and graduate students (with in-ternational participation), Minsk: BSTU, 2020. pp. 106-107. (in Russian).
9. Padutov V.E., Tretyakova I.N., Mozharovskaya L.V. Konstantinov A.V., Kulagin D.V., Kus-enkova M.P. Comparative Analysis of Transcriptional Profiles of Callus Cultures of Siberian Larch with Different Embryogenic Potential // Forestry: materials of the 84th scientific-technical. conferences of faculty, research staff and graduate students (with international participation), Minsk: BSTU, 2020. p. 131.
10. Wang Z., Gerstein M., Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews genetics, 2009, V. 10., №. 1. pp. 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. pp. 1494-512.
12. Wang, Y., Sun, M.-a. Transcriptome Data Analysis: Methods and Protocols. Springer, 2018.
13. Available at: http://bioinformaticsinstitute.ru/sites/default/files/07-28-04-kasyanov.pdf. (accessed: 04.09.2020) (in Russian).
14. Kasyanov A. S. New methods of data processing obtained using modern sequencing technologies for solving problems of gene expression analysis: author. diss. Cand. physical-mat. sciences, 2012. (in Russian).
15. Vodyasova E.A., Chelebieva E.S., Kuleshova O.N. The latest technologies for high-performance sequencing of the transcriptome of individual cells. Vavilovskiy Zhurnal Genetics and Breeding, 2019, V. 23, №5. - pp. 508-518.
16. Ewing B., Green P. Base-calling of automated sequencer traces using phred. II. Error proba-bilities. Genome research, 1998, V. 8, №. 3. pp. 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.– pp. 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. p. S7.
19. Romanenkov K.V. Method for assessing the quality of genome assembly based on frequen-cies of k-mers. Preprints M.V. Keldysh. 2017. No. 11. 24 p. doi: 10.20948 / prepr-2017-11
20. Giannoulatou E., Park S.H., Humphreys D.T., Ho J.W. Verification and validation of bioin-formatics software without a gold standard: a case study of BWA and Bowtie. BMC Bioinfor-matics, 2014,V. 15 Suppl 16. pp. 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. pp. 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 (Gossypiumhirsutum). BMC Plant Biol., 2018, V. 18, № 1. pp. 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. pp. 455-477.
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. pp. 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. pp. 1494-1512.
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. pp. 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. pp. 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. pp. 1086-1092.
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. pp. 2872-2877.
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. pp. 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. pp. 1117-1123.
34. Boerner S., McGinnis K.M. Computational Analysis of LncRNA from cDNA Sequences. Methods In Molecular Biology (Clifton, N.J.)., 2016, V. 1402. pp. 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 quan-tification. BMC genomics, 2017, V. 18, №. 1. pp. 583.
37. Chen, 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. pp. S9.
38. Huerta-Cepas J., Szklarczyk D., Forslund K., Cook H., Heller D. et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research, 2016, V. 44, № D1. pp. D286-D293.
39. 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. pp. R134.
40. 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.pp. 1236-40.
41. 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. pp. S13.
42. Mulder N., Apweiler R. InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol., 2007, V. 396. P. 59-70.
43. Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N. et al. InterProScan: protein domains identifier. Nucleic Acids Research, 2005, V. 33. № pp. W116-20.
44. 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.
45. Zdobnov E.M., Apweiler R. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics, 2001,V. 17, № 9. pp. 847-8.
46. Panteleev S. V., Baranov O. Yu., Rubel I. E. Molecular-genetic diagnostics of infectious agents of Scots pine shoots with signs of "witch`s brooms". Problemy lesovedeniya i lesovodstva, Gomel, 2016, V. 76. pp. 242–249. (in Russian).

Received: 30.09.2020 *Поступила: 30.09.2020*