ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ №2(33), 2017, с. 45–68

УДК 57.05; 004.031.2

А. М. Спицина, А. О. Брагин, А. И. Дергилев, И. В. Чадаева,
Н. Н. Твердохлеб, Э. Р. Галиева, Л. Э. Табиханова,
Ю. Л. Орлов

# Компьютерные средства анализа транскриптомных данных: программный комплекс ExpGene

Аннотация. Технологии высокопроизводительного секвенирования ДНК позволяют получать данные экспрессии генов в масштабе генома, как на микрочипах, так и на основе транскриптомного профилирования. Необходимо развитие новых компьютерных методов анализа таких данных, опирающихся на суперкомпьютерные технологии. Рассмотрены задачи анализа транскриптом в контексте вычислительной сложности. Представлены примеры применения программного комплекса ExpGene для статистической обработки и визуализации транскриптомных и микрочиповых данных. Показаны приложения для анализа транскриптом отделов мозга лабораторных животных.

*Ключевые слова и фразы:* биоинформатика, программный комплекс, секвенирование ДНК, транскриптом, экспрессия генов, микрочипы, базы данных.

## Введение

Биоинформационный анализ молекулярных механизмов экспрессии генов помогает в решении задач естественных наук, биотехнологии и медицины. Анализ таких биоинформационных данных все больше входит в сферу суперкомпьютерных исследований [1, 2]. Методы секвенирования ДНК позволяют не только измерять уровни транскрипции генов (количество мРНК) в клетке, но и решать качественно новые научные проблемы исследования закономерностей взаимодействия генов, их согласованной работы в тканях организма, предоставляя все возрастающий объем транскриптомных данных [3, 4]. Продолжающееся развитие современных технологий секвенирования позволило собрать большой объем экспериментальных транскриптомных данных, для

Исследование поддержано бюджетным проектом ИЦиГ СО РАН 0324-2016-0008.

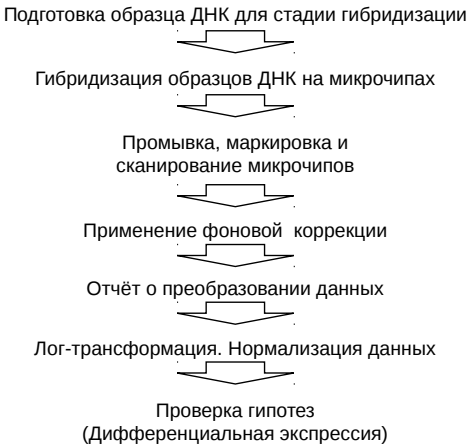


Рис. 1. Схема обработки данных микрочипов для анализа экспрессии генов

обработки которых необходимы новые компьютерные решения. Требуется применение компьютерных методов и программ, позволяющих обрабатывать данные быстро и с высокой точностью, предоставлять удобный интерфейс решения пользователю-специалисту в области естественных наук [4, 5].

Для измерения экспрессии генов используются микрочипы [6, 7] и высокопроизводительное секвенирование ДНК (транскриптомное профилирование) [8–10].

На рис. 1 представлена схема подготовки образцов ДНК и дальнейший анализ данных для определения уровней экспрессии генов на микрочипах.

Высокая экспрессия ряда генов может служить маркером для диагностики различных заболеваний, например, онкологических [10–13]. Выявление паттернов экспрессии позволяет находить списки генов (сигнатуры) для распознавания патологических состояний организма.

На рис. 2 представлена схема анализа транскриптомного профиля: от выравнивания и отображения считывания до оценки экспрессии дифференциального гена и анализа альтернативного сплайсинга [14].

Мы рассмотрим основные подходы и программные решения в данной области, представим собственное программное обеспечение, и примеры применения для анализа данных на лабораторных животных [15].

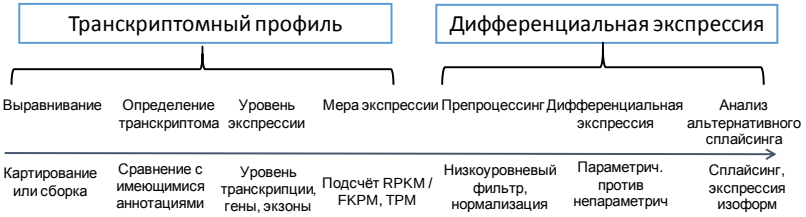


Рис. 2. Схема анализа транскриптомного профиля

Исследование регуляции экспрессии генов в масштабе генома требует развития программных средств интеграции данных, включая данные технологий RNA-Seq, ChIP-Seq, Hi-C, так же, как и данные, полученные с помощью микрочипов [4, 5]. В ИЦиГ СО РАН разработан ряд программных средств такой интеграции данных [3, 16–19]. Тестирование программ и анализ проводили на вычислительных ресурсах Сибирского Суперкомпьютерного Центра СО РАН.

## 1. Экспериментальные данные

Существуют международные базы, содержащие данные по экспрессии генов, полученные с помощью различных экспериментальных технологий (BioGPS [20], Gene Expression Omnibus (GEO) NCBI [21]).

Для получения данных экспрессии генов используются несколько технологических платформ микрочипов, одна из наиболее распространенных — GeneChip, разработанные компанией Affymetrix [22], которые используют технологию синтеза коротких олигонуклеотидных зондов на поверхности микрочипа.

В настоящее время стандартом определения экспрессии генов являются технологии, основанные на секвенировании — определениии первичной структуры нуклеотидных последовательностей. Современное полногеномное секвенирование транскриптом (RNA- seq) основано на прямом секвенировании комплементарной дезоксирибонуклеиновой кислоты (кДНК) [23, 24]. В результате секвенирования создается библиотека ридов (коротких прочтений фрагментов ДНК). Длина прочтения варьируется от 25 до 200 нуклеотидов. Затем риды картируются (выравниваются) на референсный геном. В настоящее время существует широкий спектр программ для количественного анализа экспрессии генов (Cufflinks [25], rSeq [26]), доступны репозитарии и базы данных RNA-Seq (RNA-Seq Atlas [27], GEO NCBI [21]).

В качестве приложения разработанных программ были проанализированы данные РНК-секвенирования нескольких отделов мозга лабораторных мышей и крыс [15]. Выборка мышей представляла собой три группы особей, две из которых подвергались повторному опыту агрессии посредством организации конфликтов с другими мышами (поведенческий эксперимент) [28].

Крысы были представлены двумя линиями, селектированными на агрессивное и ручное поведение по отношению к человеку в течение более 70 поколений [15]. В работе использовали образцы трех отделов головного мозга животных — гипоталамуса, покрышки среднего мозга и периакведуктума серого вещества. В случае мышей брали по три реплики для каждого отдела мозга каждой группы животных, в случае крыс — по 2 реплики [29].

Для анализа данных РНК-секвенирования использовали по три реплики для каждого отдела мозга каждой группы мышей и по две реплики для каждого отдела мозга каждой группы крыс.

РНК-секвенирование образцов осуществлялось на платформе Illumina. Использовался протокол NEBNext mRNA Library PrepReagent Set for Illumina (NEB, USA). Компьютерную обработку данных РНК-секвенирования проводили на суперкомпьютере ССКЦ СО РАН. Для картирования библиотек на геномы крысы RGSC Rnor_5.0\rn5 и мыши mm10 использовали программу TopHat2 [30]. Аннотацию генов в формате получили с сайта UCSC Genome Browser [31].

## 2. Программный комплекс ExpGene

В связи с постоянно растущим объемом данных экспрессии генов, как на микрочипах, так и RNA-seq, возникают следующие проблемы:

- невозможность обрабатывать данные вручную: требуется автоматизированная обработка, которая, в свою очередь, требует суперкомпьютерного подхода и алгоритмов Big Data, которые уменьшат время работы приложения;
- доступность уже разработанного программного обеспечения: (многие приложения не распространяются бесплатно, имеют закрытый исходный код, ограничены в формате данных);
- невозможность работы с данными без подключения к Интернету;
- удобство пользования интерфейсом (доступность для пользователя-биолога, не специалиста в информатике);
- необходимость сопоставления информации из различных баз данных для более полного анализа;
- визуализация результатов.

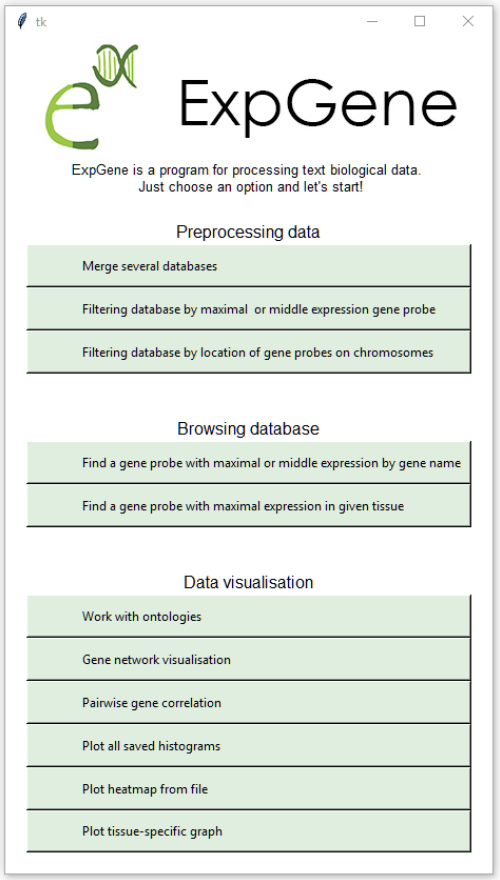


Рис. 3. Меню программы ExpGene

Разработан программный комплекс анализа данных на микрочипах, интерфейс которого представлен на рис. 3. Комплекс потребовал дополнительных опций для обработки экспрессионных данных в связи с развитием технологий и появлением новой информации, и данная работа посвящена улучшению и развитию комплекса с учетом этих требований.

Новый программный комплекс написан на высокоуровневом языке программирования Python, версия 3.5, и состоит из 15 модулей: главный модуль, модуль функций, модуль предварительной подго-

товки данных и 12 модулей опций. Программа имеет графический интерфейс для удобного взаимодействия, и работать с ней легко даже неопытному пользователю.

## 2.1. Описание модулей программы

### 2.1.1. *Главный модуль*

В главном модуле сосредоточены построение главного меню (рис. 3) и вызовы модулей опций.

### 2.1.2. *Модуль функций*

Модуль функций содержит общие функции, которые применяются в других модулях:

- FindMax — поиск усредненной пробы (пробы, которая является векторным средним арифметическим) по нескольким имеющимся пробам, относящимся к одному гену;
- FindMid — поиск максимальной пробы (пробы, которая является максимальной по длине ее вектора экспрессии) по нескольким имеющимся пробам, относящимся к одному гену;
- AddAlign — парсинг данных о расположении гена на хромосоме и вставка этой информации в базу данных, с которой в данный момент идет работа;
- PrintNetInHTML — вывод сети генов в файл *.html для дальнейшей отрисовки.

### 2.1.3. *Модуль предварительной подготовки данных*

Обработка и хранение данных производятся с помощью структур данных библиотеки pandas — Series и DataFrame. Series — это проиндексированный одномерный массив значений. DataFrame — это проиндексированный многомерный массив значений, каждый столбец DataFrame является структурой Series.

Таким образом, данные представляют собой таблицы (DataFrame), которые могут содержать как текстовые данные, так и численные. Так как программа предназначена для обработки различных данных, требуется учесть, что они не всегда имеют одинаковую структуру, что может привести к некорректной работе программы. Для этого была реализована функция выбора данных, позволяющая пользователю выбрать из базы данных только те значения (строки или столбцы), которые необходимы для применяемой опции. Такая организация рабочего процесса позволяет избежать ошибок и уменьшить использование памяти, что повышает производительность и скорость работы с большими данными.

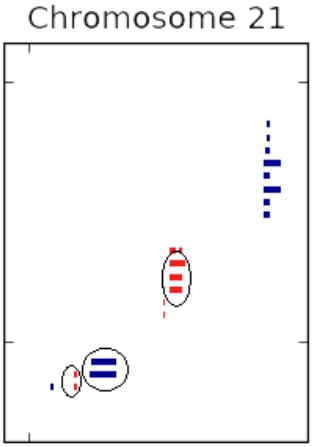


Рис. 4. Блоки проб-изоформ (выдача программы)

### 2.1.4. *Модули опций*

Каждая опция в программе представляет собой отдельный модуль, который может использовать только общие модули (модуль функций и модуль предварительной обработки данных), а также библиотеки языка Python. Такая структура делает программный комплекс гибким и позволяет встраивать в него неограниченное количество модулей, не нарушая общий принцип работы. Кроме того, отдельные модули могут быть интегрированы в другие программные комплексы без дополнительных изменений, что делает удобным совместные исследования. Модули опций делятся на три блока.

(*1*) Опции предобработки (для микрочиповых данных):
- объединение нескольких баз данных в одну с учетом возможных несоответствий;
- фильтрация базы данных по максимальной или усредненной пробе;
- фильтрация базы данных по расположению генов на хромосоме. На рис. 4 показано выделение изоформ гена на участке хромосомы 21. Красные линии на рисунке — гены, идущие в положительном направлении, синие линии — гены, идущие в отрицательном направлении; в кружках — изоформы одного и того же гена.

(*2*) Обзорные опции:

- нахождение пробы с максимальной или усредненной экспрессией по заданному имени гена;
- нахождение с максимальной экспрессией в данной ткани.

(3) Опции визуализации (для микрочиповых данных и данных RNA-Seq):

*Работа с генными онтологиями*

В разработанной программе были реализованы две опции для работы с онтологиями биологических процессов, клеточных компонентов и молекулярных функций на основе имеющейся в аннотационном файле информации (https://david.ncifcrf.gov/).

- Первая опция позволяет по введенному списку генов составить матрицу их контактов и собрать список процессов, отвечающий данной выборке. Кроме того, в файл выводится информация для каждого процесса: число генов, задействованных в нем, и список этих генов. По матрице затем может быть построена тепловая карта.
- Вторая опция позволяет по загруженной базе данных просмотреть все соответствующие ей процессы и количество генов, задействованных в них, и вывести список генов по каждому процессу в виде генной сети (полного графа).

*Визуализация генных сетей*

- Для наглядного представления взаимодействия генов на основе информации об их совместной экспрессии была реализована опция визуализации генных сетей. Для построения генной сети считается матрица корреляций, и затем, в зависимости от типа связи (отрицательная или положительная корреляция) и силы корреляции (от $-1$ до 1) между двумя вершинами-генами, задается дуга-связь, строится граф (рис. 5). Граф является интерактивным — можно разворачивать сеть и перемещать узлы; показан фрагмент, построенный по корреляциям экспрессии генов из базы BioGPS. Данный граф построен по списку тканеспецифичных генов.
- Можно отметить большое количество положительных корреляций, а также четко выделенные кластеры генов, каждый из которых соответствует определенной ткани или органу.

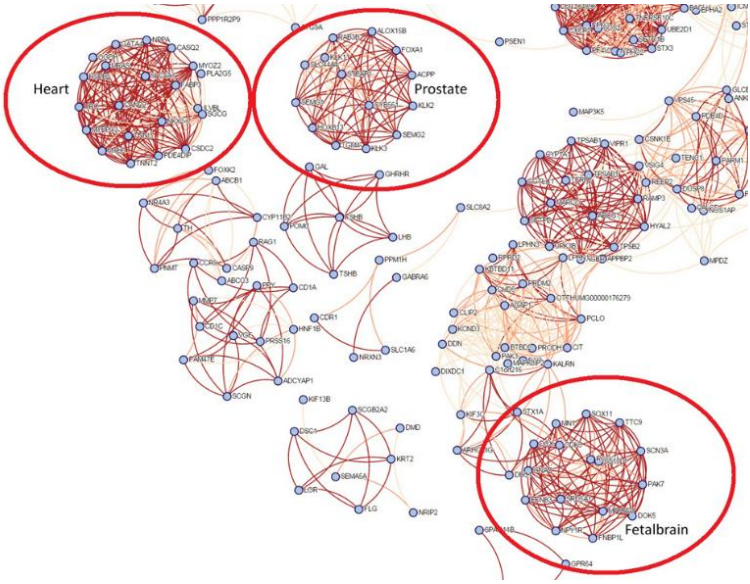


Рис. 5. Кластеры тканеспецифичных генов

*Попарная корреляция экспрессии генов*

- Попарная корреляция представляет собой симметричную матрицу, каждый элемент которой — коэффициент корреляции между двумя генами. Корреляция может быть рассчитана для трех типов коэффициентов: линейный коэффициент корреляции Пирсона, ранговый коэффициент корреляции Спирмена, коэффициент корреляции Кендалла.
- При расчете матрицы линейной или ранговой корреляции строится матрица значимости всех коэффициентов, и есть возможность строить генную сеть только для значимых коэффициентов (уровень значимости вводится пользователем).

*Построение гистограмм корреляций*

- После построения матрицы корреляций на экран выводится гистограмма, которая наглядно представляет связь в заданной выборке генов. Гистограмма может быть построена блочно (для маленьких выборок) или с помощью интерполирующей кривой (для больших выборок) (рис. 6).

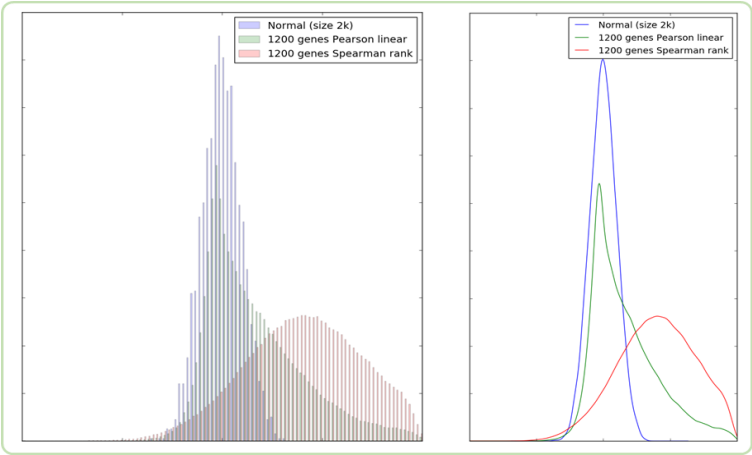


Рис. 6. Пример выдачи программы для случайной выборки размером 1200 генов (коэффициенты Пирсона и Спирмена) и сравнение с выборкой, подчиняющейся нормальному распределению

*Построение тепловых карт*

В программе реализовано несколько возможностей построения тепловых карт.

- Для построения тепловой карты по матрице корреляций необходимо выбрать соответствующую опцию, и тепловая карта будет построена в цветовой палитре “coolwarm”: отрицательные коэффициенты корреляции будут визуализированы в синей гамме, а положительные – в красной, что упростит визуальный анализ больших корреляционных матриц.
- При построении матрицы контактов используется цветовая палитра “Reds”: цвет варьируется от белого (нулевое количество контактов) до темно-красного (максимальное количество контактов).

Кроме того, оба типа тепловых карт строятся в двух вариантах:

- классическая тепловая карта (элементы матрицы не сортируются, а строятся непосредственно по матрице корреляций);

- тепловая карта с кластеризацией элементов матрицы корреляций по центроидному невзвешенному методу, который использует для пересчёта матрицы расстояний. В качестве расстояния

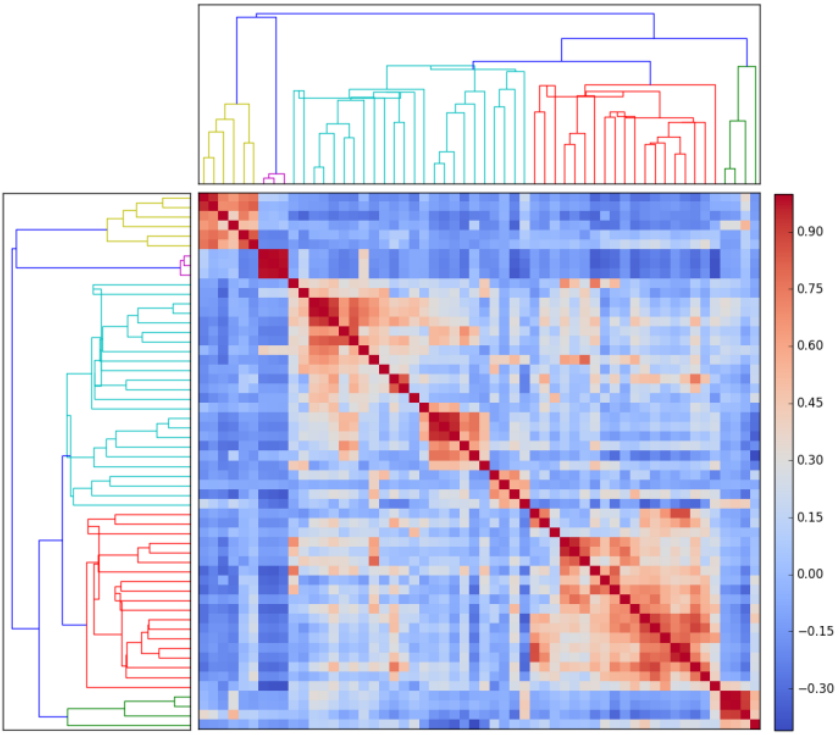


Рис. 7. Пример визуализации тепловой карты по матрице корреляций

между двумя кластерами в этом методе берётся расстояние между их центрами тяжести [32]. Такая тепловая карта наглядно показывает кластеры высокоэкспрессирующихся групп генов, и для нее на графике отражена древовидная кластеризация (рис. 7). Синие тона — отрицательные коэффициенты, красные тона — положительные коэффициенты. Также строится дендрограмма по полученным кластерам.

*Построение графика тканеспецифичности*

- График тканеспецифичности представляет собой кривую, которая отражает распределение уровней экспрессии генов по тканям. Список генов выбирается с помощью модуля предварительной подготовки данных. Затем считается пороговое значение $M$. Оно задается как медиана экспрессии по всей базе данных (такое число выборки, что ровно половина из элементов выборки

Таблица 1. Фрагмент результата обработки данных РНК-секвенирования животных программой Cuffdiff

| Индекс гена | локализация | посчитанная экспрессия | | p-value | Значимость* |
|---|---|---|---|---|---|
| | | образец 1 | образец 2 | | |
| NM_001099457 | chr1:388172-401149 | 0 | 0 | 1 | no |
| NM_001099462 | chr1:688297-696950 | 0 | 0 | 1 | no |
| NM_001106217 | chr1:336312-3464946 | 54.9833 | 42.3015 | 0.12325 | no |
| NM_001128191 | chr1:348467-3478577 | 5.65749 | 7.64213 | 0.0879 | no |
| NM_001106227 | chr1:79664959-79667668 | 73.0068 | 45.1946 | 0.00045 | yes |

*Значимость с учетом поправки на множественное сравнение

больше него, а другая половина меньше него). После подсчетов программа выдает график общей тканеспецифичности, который показывает, у скольких генов экспрессия превышает порог $M$ в $j$ тканях, $j = 1, \ldots, K$, где $K$ — общее число тканей, а также значение тканеспецифичности по тканям — сколько генов экспрессируется в каждой ткани организма. В отдельный файл выводится список генов и для каждого гена — число тканей, в которых он экспрессируется.

## 3. Примеры применения программы

Результатом работы программы являются численные и текстовые данные и иллюстрации. При выборе какой-либо опции программа запрашивает у пользователя ввод названия файла для выходных данных, затем все необходимые данные выводятся в файл в виде таблиц, с которыми можно работать с помощью программы Microsoft Excel. Иллюстрации выводятся во всплывающем окне, которое имеет кнопки масштабирования и кнопку сохранения иллюстрации в высоком разрешении. Программа применялась к данным экспрессии генов у лабораторных животных [15].

Определение уровня экспрессии генов в единицах FPKM (fragments per kilobase of transcript per million mapped reads), и поиск дифференциально экспрессирующихся генов между библиотеками агрессивных и ручных крыс осуществляли программами Cufflinks [25]. Работы проведены сотрудниками Лаборатории нейроинформатики поведения ИЦиГ СО РАН. Пример полученных данных представлен в таблице 1.

Как видно из таблицы 1, программа Cufflinks (Cuffdiff) рассчитывает нормированное значение экспрессии гена с заданными координатами в анализируемых образцах с последующей оценкой значимости

различий в экспрессии. Расчет корреляции экспрессии генов в различных образцах по таким данным позволяет с помощью данной программы строить распределения и тепловые карты, показанные на рисунках выше.

При интерпретации данных RNA-seq следует учитывать, что в результате секвенирования происходят систематические ошибки, связанные с технологией и экспериментальной платформой (например Illumina), которые могут значительно влиять на оценку экспрессии [33]. Разработан ряд подходов процессинга и фильтрации входных данных, статистических оценок ошибок возникающих при секвенировании [34, 35]. Их использование требует дальнейшего расширения функционала программного комплекса.

## Заключение

Задачи биоинформатики являются важным направлением развития Суперкомпьютерных Центров коллективного пользования [1, 2, 36]. Прикладные задачи анализа экспрессии генов состояли в разработке быстродействующего программного комплекса и выявления с его помощью особенностей генов, активно экспрессирующихся в различных тканях и органах. Мы исследовали совместную экспрессию генов, функционирующих в составе известных генных сетей [16]. Анализировали особенности экспрессии пар транскриптов, ко-локализованных в геноме, в том числе цис-антисенс транскриптов [37], в геномах модельных организмов, таких как зебровая рыбка *D.rerio* [38], дрожжи [39].

В результате работы был реализован инструментарий для обработки данных микрочипов Affymetrix U133 и данных RNA-Seq на языке *Python*. Включены алгоритмы оценок коэффициентов корреляции и фильтрации проб по качеству для проб платформы Affimetrix [40, 41]. Включены опции визуализации данных (построение распределений коэффициентов корреляции, кластеризация и построение тепловых карт, визуализация сети). Для чистоты эксперимента анализ проводили на данных, уже отфильтрованных с помощью разработанной программы.

С помощью полногеномного секвенирования (RNA-Seq) ранее были выявлены дифференциально экспрессирующиеся гены в отделах мозга у агрессивных и ручных животных (крыс). Исследованы следующие отделы мозга: гипоталамус (hypothalamus), район покрышки среднего мозга (mesencephalic tegmentum) и периакведуктум серого вещества (periaqueductum grey matter)[15, 29]. С использованием

разработанного комплекса была реконструирована сеть генов, связанных с агрессивным поведением, проанализированы гистограммы распределения их корреляций.

С помощью программы были реконструированы генные сети регуляции холестерина и циркадного ритма, представленные ранее [5, 33], построены профили тканеспецифичности для этих сетей. Выделены тканеспецифичные гены и гены с повышенной экспрессией в отделах мозга. Были получены списки генов, расположенных относительно пар сайтов связывания CTCF, построены их профили тканеспецифичности и распределения коэффициентов корреляций совместной экспрессии генов в таком геномном окружении [42].

Развитие новых экспериментальных методов секвенирования привело к стремительному росту объемов данных и разработке новых компьютерных программ, в том числе использующих суперкомпьютерные технологии [1, 5, 42].

Системное исследование экспрессии генов в клетках мозга с помощью взаимодополняющих экспериментальных подходов является необходимой основой междисциплинарных нейробиологических исследований, которые могут быть продолжены на новых экспрессионных данных, в том числе для модельных животных, с помощью разработанного программного комплекса.

*Авторы благодарны Н. Л. Подколодному, М. Чен, П. Третьякову, а также ССКЦ СО РАН за поддержку работы.*

## Список литературы

[1] Б. М. Глинский, Н. В. Кучин, И. Г. Черных, Ю. Л. Орлов, Н. Л. Подколодный, В. А. Лихошвай, Н. А. Колчанов. «Суперкомпьютерные технологии в решении задач биоинформатики», *Программные системы: теория и приложения*, **6**:4 (2015), с. 99–112, URL: http://psta.psiras.ru/read/psta2015_4_99-112.pdf ↑ 45,57,58

[2] Г. Э. Норман, Н. Д. Орехов, В. В. Писарев, Г. С. Смирнов, С. В. Стариков, В. В. Стегайлов, А. В. Янилкин. «Зачем и какие суперкомпьютеры экзафлопсного класса нужны в естественных науках», *Программные системы: теория и приложения*, **6**:4 (2015), с. 243–311, URL: http://psta.psiras.ru/read/psta2015_4_243-311.pdf ↑ 45,57

[3] Ю. Л. Орлов, А. О. Брагин, И. В. Медведева и др. «ICGenomics: программный комплекс анализа символьных последовательностей геномики», *Вавиловский журнал генетики и селекции*, **16**:4/1 (2012), с. 732–731, URL: http://vavilov.elpub.ru/index.php/jour/article/view/70 ↑ 45,47

[4] Ю. Л. Орлов. «Компьютерное исследование регуляции транскрипции генов эукариот с помощью данных экспериментов секвенирования и иммунопреципитации хроматина», *Вавиловский журнал генетики и селекции*, **18**:1 (2014), с. 193–206, URL: http://vavilov.elpub.ru/index.php/jour/article/view/240 ↑ 45,46,47

[5] А. М. Спицина, Ю. Л. Орлов, Н. Н. Подколодная, А. В. Свичкарев, А. И. Дергилев, М. Чен, Н. В. Кучин, И. Г. Черных, Б. М. Глинский. «Суперкомпьютерный анализ геномных и транскриптомных данных, полученных с помощью технологий высокопроизводительного секвенирования ДНК», *Программные системы: теория и приложения*, **6**:1 (2015), с. 157–174, URL: http://psta.psiras.ru/read/psta2015_1_157-174.pdf ↑ 46,47,58

[6] K. M. Bhawe, M. K. Aghi. "Microarray Analysis in Glioblastomas", *Methods Mol. Biol.*, **1375** (2016), pp. 195–206, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5056625/ ↑ 46

[7] P. H. Guzzi, M. Cannataro. "Micro-Analyzer: automatic preprocessing of Affymetrix microarray data", *Comput Methods Programs Biomed.*, **111**:2 (2013), pp. 402–409. ↑ 46

[8] H. C. Huang, Y. Niu, L. X. Qin. "Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software", *Cancer Inform.*, **14**, Suppl. 1 (2015), pp. 57–67, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678998/ ↑ 46

[9] X. Gu. "Statistical detection of differentially expressed genes based on RNA-seq: from biological to phylogenetic replicates", *Brief Bioinform.*, **17**:2 (2016), pp. 243–248. ↑ 46

[10] A. Poplawski, F. Marini, M. Hess, T. Zeller, J. Mazur, H. Binder. "Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective", *Brief Bioinform.*, **17**:2 (2016), pp. 21323. ↑ 46

[11] A. Perez-Diez, A. Morgun, N. Shulzhenko. "Microarrays for cancer diagnosis and classification", *Adv. Exp. Med. Biol.*, **593** (2013), pp. 74–85, URL: http://www.ncbi.nlm.nih.gov/books/NBK6624/ ↑ 46

[12] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, A. Mortazavi. "A survey of best practices for RNA-seq data analysis", *Genome Biol.*, **17** (2016), pp. 13, URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8 ↑ 46

[13] A. Dobin, T. R. Gingeras. "Mapping RNA-seq Reads with STAR", *Current protocols in bioinformatics*, **51** (2015), pp. 11.14.1–11.14.19, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631051/ ↑ 46

[14] C. R. Williams, A. Baccarella, J. Z. Parrish, C. C. Kim. "Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq", *BMC Bioinformatics*, **18**:1

(2017), pp. 38, URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1457-z ↑ 46

[15] V. N. Babenko, A. O. Bragin, A. M. Spitsina, I. V. Chadaeva, E. R. Galieva, G. V. Orlova, I. V. Medvedeva, Y. L. Orlov. "Analysis of differential gene expression by RNA-seq data in brain areas of laboratory animals", *Journal of Integrative bioinformatics*, **13**:4 (2016), 292, 15 p., URL: http://biecoll.ub.uni-bielefeld.de/volltexte/2017/5436/ ↑ 46,48,56,57

[16] Y. Orlov, H. Xu, D. Afonnikov et al. "Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell", *Journal of Integrative bioinformatics*, **9**:2 (2012), pp. 211, URL: http://www.ncbi.nlm.nih.gov/pubmed/22987856 ↑ 47,57

[17] О. С. Кожевникова, М. К. Мартыщенко, М. А. Генаев и др. «RatDNA: база данных микрочиповых исследований на крысах для генов, ассоциированных с заболеваниями старения», *Вавиловский журнал генетики и селекции*, **16**:4/1 (2012), с. 756–765, URL: http://vavilov.elpub.ru/index.php/jour/article/view/72 ↑ 47

[18] Д. А. Полунин, И. А. Штайгер, В. М. Ефимов. «Разработка программного комплекса JACOBI 4 для многомерного анализа микрочиповых данных», *Вестник НГУ. Серия: Информационные технологии*, **12**:2 (2014), с. 90–98, URL: http://www.nsu.ru/xmlui/bitstream/handle/nsu/4125/2014_V12_No2_11.pdf ↑ 47

[19] И. В. Медведева, О. В. Вишневский, Н. С. Сафронова и др. «Компьютерный анализ данных экспрессии генов в клетках мозга, полученных с помощью микрочипов и высокопроизводительного секвенирования», *Вавиловский журнал генетики и селекции*, **17**:4/1 (2013), с. 629–638, URL: http://vavilov.elpub.ru/index.php/jour/article/view/187 ↑ 47

[20] BioGPS, URL: http://biogps.org/ ↑ 47

[21] *Gene Expression Omnibus (GEO) NCBI*, URL: http://www.ncbi.nlm.nih.gov/geo/ ↑ 47

[22] *Affymetrix*, URL: http://www.affymetrix.com/ ↑ 47

[23] R. Hrdlickova, M. Toloue, B. Tian. "RNA-Seq methods for transcriptome analysis", *Wiley Interdiscip Rev RNA*, **8**:1, URL: https://www.ncbi.nlm.nih.gov/pubmed/27198714 ↑ 47

[24] B. J. Haas, M. C. Zody. "Advancing RNA-Seq analysis", *Nat Biotechnol.*, **28**:5 (2010), pp. 421–423, URL: https://www.ncbi.nlm.nih.gov/pubmed/20458303 ↑ 47

[25] Cufflinks, URL: http://cole-trapnell-lab.github.io/cufflinks/ ↑ 47,56

[26] rSeq: RNA-Seq Analyzer, URL: http://www-personal.umich.edu/~jianghui/rseq/ ↑ 47

[27] RNA Seq Atlas, URL: http://medicalgenomics.org/rna_seq_atlas ↑ 47

[28] И. Л. Коваленко, Д. А. Смагин, А. Г. Галямина, Ю. Л. Орлов, Н. Н. Кудрявцева. «Изменение экспрессии дофаминергических генов в структурах мозга самцов мышей под влиянием хронического социального стресса: данные RNA-seq», *Молекулярная биология*, **50**:1 (2016), с. 184–187. ↑ 48

[29] В. Н. Бабенко, А. О. Брагин, И. В. Медведева, И. В. Чадаева, А. И. Дергилев, А. М. Спицина, Н. Н. Кудрявцева, А. Л. Маркель, Ю. Л. Орлов. «Анализ транскриптомных данных экспрессии генов в отделах мозга крыс, селектированных по агрессивному поведению», *XVIII Всероссийская научно-техническая конференция «Нейроинформатика-2016»*, Сборник научных трудов. Т. 2, НИЯУ МИФИ, М., 2016, с. 82–92. ↑ 48,57

[30] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg. “TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions”, *Genome Biol.*, **14**:4 (2013), pp. R36, URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r36 ↑ 48

[31] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, W. J. Kent. “The UCSC Table Browser data retrieval tool”, *Nucleic Acids Res.*, **32**, Database issue (2004), pp. D493–496, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh103 ↑ 48

[32] P. H. Sneath, R. R. Sokal. *Numerical taxonomy. The principles and practices of numerical classification*, WH Freeman, San Francisco, 1973. ↑ 55

[33] Е. В. Кулакова, А. М. Спицина, Н. Г. Орлова, А. И. Дергилев, А. В. Свичкарев, Н. С. Сафронова, И. Г. Черных, Ю. Л. Орлов. «Программы анализа геномных данных секвенирования, полученных на основе технологий ChIP-seq, ChIA-PET и Hi-C», *Программные системы: теория и приложения*, **6**:2 (2015), с. 129–148, URL: http://psta.psiras.ru/read/psta2015_2_129-148.pdf ↑ 57,58

[34] R. te Boekhorst, F. M. Naumenko, N. G. Orlova, E. R. Galieva, A. M. Spitsina, I. V. Chadaeva, Y. L. Orlov, I. I. Abnizova. «Computational problems of analysis of short next generation sequencing reads», *Вавиловский журнал генетики и селекции*, **20**:6 (2016), с. 746–755 (in English), URL: http://vavilov.elpub.ru/jour/article/viewFile/845/846 ↑ 57

[35] I. Abnizova, R. te Boekhorst, Y. Orlov. “Computational Errors and Biases of Short Read Next Generation Sequencing”, *Journal of Proteomics & Bioinformatics*, **10** (2017), pp. 1–17, URL: https://www.omicsonline.org/open-access/computational-errors-and-biases-in-short-read-next-generationsequencing-jpb-1000420.php?aid=85469 ↑ 57

[36] Д. И. Харитонов, Г. В. Тарасов, Д. В. Леонтьев, Р. В. Парахин, В. В. Грибова. «Текущее состояние и перспективы развития центра коллективного пользования «Дальневосточный Вычислительный Ресурс»», *Программные системы: теория и приложения*, **7**:4

(2016), с. 197–208, URL: http://psta.psiras.ru/read/psta2016_4_197-208.pdf ↑ 57

[37] O. V. Grinchuk, P. Jenjaroenpun, Y. L. Orlov, J. Zhou, V. A. Kuznetsov. “Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns”, *Nucleic Acids Res.*, **38**:2 (2010), pp. 534–547, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp954 ↑ 57

[38] C. L. Winata, I. Kondrychyn, V. Kumar, K. G. Srinivasan, Y. Orlov, A. Ravishankar, S. Prabhakar, L. W. Stanton, V. Korzh, S. Mathavan. “Genome wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish”, *PLoS Genet.*, **9**:10 (2013), e1003852, URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003852 ↑ 57

[39] Ю. Г. Матушкин, В. Г. Левицкий, В. С. Соколов, В. А. Лихошвай, Ю. Л. Орлов. «Эффективность элонгации генов дрожжей коррелирует с плотностью нуклеосомной упаковки в 5’нетранслируемом районе», *Математическая биология и биоинформатика*, **8**:1 (2013), с. 248–257, URL: http://www.matbio.org/2013/Matushkin_8_248.pdf ↑ 57

[40] Ю. Л. Орлов, В. М. Ефимов, Н. Г. Орлова. «Статистические оценки экспрессии мобильных элементов в геноме человека на основе клинических данных экспрессионных микрочипов», *Вавиловский журнал генетики и селекции*, **15**:2 (2011), с. 327–339, URL: http://www.bionet.nsc.ru/vogis/pict_pdf/2011/15_2/12.pdf ↑ 57

[41] Y. L. Orlov, J. Zhou, L. Lipovich, A. Shahab, V. A. Kuznetsov. “Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis”, *In Silico Biol.*, **7**:3 (2007), pp. 241–260, URL: http://www.bioinfo.de/isb/2007/07/0041/ ↑ 57

[42] Е. В. Кулакова, А. М. Спицина, А. Г. Богомолов, Н. Г. Орлова, А. И. Дергилев, И. В. Чадаева, В. Н. Бабенко, Ю. Л. Орлов. «Программа анализа геномного распределения хромосомных контактов в ядре клетки по данным, полученным по технологиям ChIA-PET и Hi-C», *Программные системы: теория и приложения*, **8**:1 (2017), с. 219–242, URL: http://psta.psiras.ru/read/psta2017_1_219-242.pdf ↑ 58

Рекомендовал к публикации *Программный комитет*
Пятого национального суперкомпьютерного форума ***НСКФ-2016***

*Пример ссылки на эту публикацию:*

А. М. Спицина, А.О. Брагин, А. И. Дергилев и др. «Компьютерные средства анализа транскриптомных данных: программный комплекс ExpGene», *Программные системы: теория и приложения*, 2017, **8**:2(33), с. 45–68.

URL: http://psta.psiras.ru/read/psta2017_2_45-68.pdf

*Об авторах:*

**Анастасия Михайловна Спицина**

Аспирант ФИТ НГУ. Область научных интересов: биоинформатика, суперкомпьютерные вычисления

*e-mail:* anastasia.spitsina@gmail.com

**Анатолий Олегович Брагин**

к.б.н., н.с. ИЦиГ СО РАН. Область научных интересов: генетика поведения, биоинформатика, компьютерная геномика

*e-mail:* ibragim@bionet.nsc.ru

**Артур Игоревич Дергилев**

Магистр НГУ. Область научных интересов: биоинформатика, суперкомпьютерные вычисления

*e-mail:* arturd1993@yandex.ru

**Ирина Витальевна Чадаева**

м.н.с. ИЦиГ СО РАН. Область научных интересов: генетика поведения, биоинформатика, компьютерная геномика

*e-mail:* ichadaeva@bionet.nsc.ru

**Наталья Николаевна Твердохлеб**

Окончила Новосибирский Государственный Университет, младший научный сотрудник ИЦиГ СО РАН. Область научных интересов: генные сети, суперкомпьютерные вычисления

*e-mail:* nata@bionet.nsc.ru

**Эльвира Расимовна Галиева**

к.б.н., научный сотрудник ИЦиГ СО РАН и НГУ. Область научных интересов: цитогенетика, транскрипция, молекулярная биология

*e-mail:* galieva@bionet.nsc.ru

**Людмила Эдмундовна Табиханова**

Окончила Новосибирский Государственный Университет, научный сотрудник ИЦиГ СО РАН. Область научных интересов: популяционная генетика, молекулярная биология, генетика человека

*e-mail:* tabikhan@bionet.nsc.ru

**Юрий Львович Орлов**

д.б.н., проф. РАН, зав. лабораторией компьютерной геномики ФЕН НГУ, зав.лаб. нейроинформатики поведения ИЦиГ СО РАН. Область научных интересов: биоинформатика, компьютерная геномика

*e-mail:* orlov@bionet.nsc.ru

Anastasia Spitsina, Anatoliy Bragin, Artur Dergilev, Irina Chadaeva, Natal'ya Tverdokhleb, El'vira Galiyeva, Ludmila Tabikhanova, Yuriy Orlov. *Computer tools for analysis of transcriptomics data: program complex ExpGene.*

Abstract. High-performance DNA sequencing technologies allow to obtain gene expression data in the genome scale from microchips and transcriptome profiling.

It is necessary to develop new computational methods, based on supercomputer technologies, to analyze such data. The problems of gene expression analysis in the context of computational complexity are considered. We present application examples of the multifunctional software package ExpGene for analysis and visualization of transcriptomics and microarray data. We consider examples of analysis of laboratory animal brain areas gene expression data. (*In Russian*).

*Key words and phrases:* bioinformatics, program complex, DNA sequencing, transcriptome, gene expression, microarrays, databases.

*References*


[1] B. M. Glinskiy, N. V. Kuchin, I. G. Chernykh, Yu. L. Orlov, N. L. Podkolodnyy, V. A. Likhoshvay, N. A. Kolchanov. "Bioinformatics and High Performance Computing", *Programmnyye sistemy: teoriya i prilozheniya*, **6**:4 (2015), pp. 99–112 (in Russian), URL: http://psta.psiras.ru/read/psta2015_4_99-112.pdf

[2] G. E. Norman, N. D. Orekhov, V. V. Pisarev, G. S. Smirnov, S. V. Starikov, V. V. Stegaylov, A. V. Yanilkin. "What for and which Exaflops Supercomputers are Necessary in Natural Sciences", *Programmnyye sistemy: teoriya i prilozheniya*, **6**:4 (2015), pp. 243–311 (in Russian), URL: http://psta.psiras.ru/read/psta2015_4_243-311.pdf

[3] Yu. L. Orlov, A. O. Bragin, I. V. Medvedeva i dr. "ICGenomics: a Program Complex for Analysis of Symbol Sequences in Genomics", *Vavilovskiy zhurnal genetiki i selektsii*, **16**:4/1 (2012), pp. 732–731 (in Russian), URL: http://vavilov.elpub.ru/index.php/jour/article/view/70

[4] Yu. L. Orlov. "Computer-Assisted Study of the Regulation of Eukaryotic Gene Transcription on the Base of Data on Chromatin Sequencing and Precipitation", *Vavilovskiy zhurnal genetiki i selektsii*, **18**:1 (2014), pp. 193–206 (in Russian), URL: http://vavilov.elpub.ru/index.php/jour/article/view/240

[5] A. M. Spitsina, Yu. L. Orlov, N. N. Podkolodnaya, A. V. Svichkarev, A. I. Dergilev, M. Chen, N. V. Kuchin, I. G. Chernykh, B. M. Glinskiy. "Supercomputer Analysis of Genomics and Transcriptomics Data Revealed by High-Throughput DNA Sequencing", *Programmnyye sistemy: teoriya i prilozheniya*, **6**:1 (2015), pp. 157–174 (in Russian), URL: http://psta.psiras.ru/read/psta2015_1_157-174.pdf

[6] K. M. Bhawe, M. K. Aghi. "Microarray Analysis in Glioblastomas", *Methods Mol. Biol.*, **1375** (2016), pp. 195–206, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5056625/

[7] P. H. Guzzi, M. Cannataro. "Micro-Analyzer: automatic preprocessing of Affymetrix microarray data", *Comput Methods Programs Biomed.*, **111**:2 (2013), pp. 402–409.

[8] H. C. Huang, Y. Niu, L. X. Qin. "Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software", *Cancer Inform.*, **14**, Suppl. 1 (2015), pp. 57–67, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4678998/

[9] X. Gu. "Statistical detection of differentially expressed genes based on RNA-seq: from biological to phylogenetic replicates", *Brief Bioinform.*, **17**:2 (2016), pp. 243–248.

[10] A. Poplawski, F. Marini, M. Hess, T. Zeller, J. Mazur, H. Binder. "Systematically evaluating interfaces for RNA-seq analysis from a life scientist perspective", *Brief Bioinform.*, **17**:2 (2016), pp. 21323.

[11] A. Perez-Diez, A. Morgun, N. Shulzhenko. "Microarrays for cancer diagnosis and classification", *Adv. Exp. Med. Biol.*, **593** (2013), pp. 74–85, URL: http://www.ncbi.nlm.nih.gov/books/NBK6624/

[12] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang, A. Mortazavi. "A survey of best practices for RNA-seq data analysis", *Genome Biol.*, **17** (2016), pp. 13,

URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

[13] A. Dobin, T. R. Gingeras. “Mapping RNA-seq Reads with STAR”, *Current protocols in bioinformatics*, **51** (2015), pp. 11.14.1–11.14.19, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4631051/

[14] C. R. Williams, A. Baccarella, J. Z. Parrish, C. C. Kim. “Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq”, *BMC Bioinformatics*, **18**:1 (2017), pp. 38, URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1457-z

[15] V. N. Babenko, A. O. Bragin, A. M. Spitsina, I. V. Chadaeva, E. R. Galieva, G. V. Orlova, I. V. Medvedeva, Y. L. Orlov. “Analysis of differential gene expression by RNA-seq data in brain areas of laboratory animals”, *Journal of Integrative bioinformatics*, **13**:4 (2016), 292, 15 p., URL: http://biecoll.ub.uni-bielefeld.de/volltexte/2017/5436/

[16] Y. Orlov, H. Xu, D. Afonnikov et al. “Computer and Statistical Analysis of Transcription Factor Binding and Chromatin Modifications by ChIP-seq data in Embryonic Stem Cell”, *Journal of Integrative bioinformatics*, **9**:2 (2012), pp. 211, URL: http://www.ncbi.nlm.nih.gov/pubmed/22987856

[17] O. S. Kozhevnikova, M. K. Martyshchenko, M. A. Genayev i dr. “RatDNA: Database on Microarray Studies of Rats Bearing Genes Associated with Age-Related Diseases”, *Vavilovskiy zhurnal genetiki i selektsii*, **16**:4/1 (2012), pp. 756–765 (in Russian), URL: http://vavilov.elpub.ru/index.php/jour/article/view/72

[18] D. A. Polunin, I. A. Shtayger, V. M. Yefimov. “JACOBI 4 Software for Multivariate Analysis of Microarray Data”, *Vestnik NGU. Seriya: Informatsionnyye tekhnologii*, **12**:2 (2014), pp. 90–98 (in Russian), URL: http://www.nsu.ru/xmlui/bitstream/handle/nsu/4125/2014_V12_No2_11.pdf

[19] I. V. Medvedeva, O. V. Vishnevskiy, N. S. Safronova i dr. “Computer Analysis of Data on Gene Expression in Brain Cells Obtained by Microarray Tests and High-Throughput Sequencing”, *Russian Journal of Genetics: Applied Research*, **4**:4 (2014), pp. 259–266.

[20] BioGPS, URL: http://biogps.org/

[21] *Gene Expression Omnibus (GEO) NCBI*, URL: http://www.ncbi.nlm.nih.gov/geo/

[22] *Affymetrix*, URL: http://www.affymetrix.com/

[23] R. Hrdlickova, M. Toloue, B. Tian. “RNA-Seq methods for transcriptome analysis”, *Wiley Interdiscip Rev RNA*, **8**:1, URL: https://www.ncbi.nlm.nih.gov/pubmed/27198714

[24] B. J. Haas, M. C. Zody. “Advancing RNA-Seq analysis”, *Nat Biotechnol.*, **28**:5 (2010), pp. 421–423, URL: https://www.ncbi.nlm.nih.gov/pubmed/20458303

[25] Cufflinks, URL: http://cole-trapnell-lab.github.io/cufflinks/

[26] rSeq: RNA-Seq Analyzer, URL: http://www-personal.umich.edu/~jianghui/rseq/

[27] RNA Seq Atlas, URL: http://medicalgenomics.org/rna_seq_atlas

[28] I. L. Kovalenko, D. A. Smagin, A. G. Galyamina, Yu. L. Orlov, N. N. Kudryavtseva. “Changes in the Expression of Dopaminergic Genes in Brain Structures of

Male Mice Exposed to Chronic Social Defeat Stress: An RNA-Seq Study", *Molekulyarnaya biologiya*, **50**:1 (2016), pp. 184–187 (in Russian).

[29] V. N. Babenko, A. O. Bragin, I. V. Medvedeva, I. V. Chadayeva, A. I. Dergilev, A. M. Spitsina, N. N. Kudryavtseva, A. L. Markel', Yu. L. Orlov. "Analysis of transcriptome data of gene expression in brain areas of rats selected by aggressive behavior", *XVIII Vserossiyskaya nauchno-tekhnicheskaya konferentsiya "Neyroinformatika-2016"*, Sbornik nauchnykh trudov. V. 2, NIYaU MIFI, M., 2016, pp. 82–92 (in Russian).

[30] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions", *Genome Biol.*, **14**:4 (2013), pp. R36, URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r36

[31] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, W. J. Kent. "The UCSC Table Browser data retrieval tool", *Nucleic Acids Res.*, **32**, Database issue (2004), pp. D493–496, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh103

[32] P. H. Sneath, R. R. Sokal. *Numerical taxonomy. The principles and practices of numerical classification*, WH Freeman, San Francisco, 1973.

[33] Ye. V. Kulakova, A. M. Spitsina, N. G. Orlova, A. I. Dergilev, A. V. Svichkarev, N. S. Safronova, I. G. Chernykh, Yu. L. Orlov. "Supercomputer Analysis of Genomics and Transcriptomics Data Revealed by High-Throughput DNA Sequencing", *Programmnyye sistemy: teoriya i prilozheniya*, **6**:2 (2015), pp. 129–148 (in Russian), URL: http://psta.psiras.ru/read/psta2015_2_129-148.pdf

[34] R. te Boekhorst, F. M. Naumenko, N. G. Orlova, E. R. Galieva, A. M. Spitsina, I. V. Chadaeva, Y. L. Orlov, I. I. Abnizova. "Computational problems of analysis of short next generation sequencing reads", *Vavilovskiy zhurnal genetiki i selektsii*, **20**:6 (2016), pp. 746–755 (in English), URL: http://vavilov.elpub.ru/jour/article/viewFile/845/846

[35] I. Abnizova, R. te Boekhorst, Y. Orlov. "Computational Errors and Biases of Short Read Next Generation Sequencing", *Journal of Proteomics & Bioinformatics*, **10** (2017), pp. 1–17, URL: https://www.omicsonline.org/open-access/computational-errors-and-biases-in-short-read-next-generationsequencing-jpb-1000420.php?aid=85469

[36] D. I. Kharitonov, G. V. Tarasov, D. V. Leont'yev, R. V. Parakhin, V. V. Gribova. "State of the Art and Development Prospects of the Shared Resource Center "Far Eastern Computing Resource" IACP FEB RAS", *Programmnyye sistemy: teoriya i prilozheniya*, **7**:4 (2016), pp. 197–208 (in Russian), URL: http://psta.psiras.ru/read/psta2016_4_197-208.pdf

[37] O. V. Grinchuk, P. Jenjaroenpun, Y. L. Orlov, J. Zhou, V. A. Kuznetsov. "Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns", *Nucleic Acids Res.*, **38**:2 (2010), pp. 534–547, URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp954

[38] C. L. Winata, I. Kondrychyn, V. Kumar, K. G. Srinivasan, Y. Orlov, A. Ravishankar, S. Prabhakar, L. W. Stanton, V. Korzh, S. Mathavan. "Genome wide analysis reveals Zic3 interaction with distal regulatory elements of stage specific developmental genes in zebrafish", *PLoS Genet.*, **9**:10 (2013), e1003852,

URL: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003852

[39] Yu. G. Matushkin, V. G. Levitskiy, V. S. Sokolov, V. A. Likhoshvay, Yu. L. Orlov. "Yeast Gene Elongation Efficiency Correlates with Nucleosome Formation in 5'-Untranslated Region", *Matematicheskaya biologiya i bioinformatika*, **8**:1 (2013), pp. 248–257 (in Russian), URL: http://www.matbio.org/2013/Matushkin_8_248.pdf

[40] Yu. L. Orlov, V. M. Yefimov, N. G. Orlova. "Statistical Estimates of Transposable Element Expression in the Human Genome Based on Clinical Microarray Data on Expression", *Vavilovskiy zhurnal genetiki i selektsii*, **15**:2 (2011), pp. 327–339 (in Russian), URL: http://www.bionet.nsc.ru/vogis/pict_pdf/2011/15_2/12.pdf

[41] Y. L. Orlov, J. Zhou, L. Lipovich, A. Shahab, V. A. Kuznetsov. "Quality assessment of the Affymetrix U133A&B probesets by target sequence mapping and expression data analysis", *In Silico Biol.*, **7**:3 (2007), pp. 241–260, URL: http://www.bioinfo.de/isb/2007/07/0041/

[42] Ye. V. Kulakova, A. M. Spitsina, A. G. Bogomolov, N. G. Orlova, A. I. Dergilev, I. V. Chadayeva, V. N. Babenko, Yu. L. Orlov. "Program for Analysis of Genome Distribution of Chromosome Contacts in Cell Nucleus by the Data Obtained using ChIA-PET and Hi-C Technologies", *Programmnyye sistemy: teoriya i prilozheniya*, **8**:1 (2017), pp. 219–242 (in Russian), URL: http://psta.psiras.ru/read/psta2017_1_219-242.pdf

*Sample citation of this publication:*

Anastasia Spitsina, Anatoliy Bragin, Artur Dergilev, Irina Chadaeva, Natal'ya Tverdokhleb, El'vira Galiyeva, Ludmila Tabikhanova, Yuriy Orlov. "Computer tools for analysis of transcriptomics data: program complex ExpGene", *Program systems: Theory and applications*, 2017, **8**:2(33), pp. 45–68. (*In Russian*). URL: http://psta.psiras.ru/read/psta2017_2_45-68.pdf