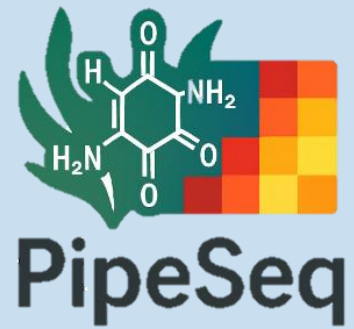


PipeSeq: программа автоматического анализа RNA-Seq

А.М. Нерезенко*, П.А. Виролаинен, Е.М. Чекунова
Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

*alexnerezenko@gmail.com

doi 10.18699/PlantGen2025-



Цель исследования - разработка и внедрение программного решения PipeSeq для полной автоматизации анализа данных RNA-seq, включая нормализацию, оценку дифференциальной экспрессии генов и визуализацию результатов на примере генов, кодирующих факторы транскрипции семейства GATA у *Chlamydomonas reinhardtii*.

РНК секвенирование (RNA-seq) — это метод, позволяющий получить количественную оценку экспрессии генов, выявить новые транскрипты и изучить общую структуру транскриптома. Он широко применяется в биологии и биотехнологии для анализа регуляторных сетей, стрессовых ответов и особенностей клеточного функционирования.

Современные тенденции в анализе RNA-seq данных направлены на повышение доступности и воспроизводимости вычислительных пайплайнов, снижение требований к навыкам программирования у исследователей, а также на улучшение визуального представления результатов. PipeSeq был разработан именно с учётом этих задач.

Объект исследования

Одноклеточная зеленая водоросль с гаплофазным жизненным циклом *C. reinhardtii* — модельный объект генетики фотосинтеза.

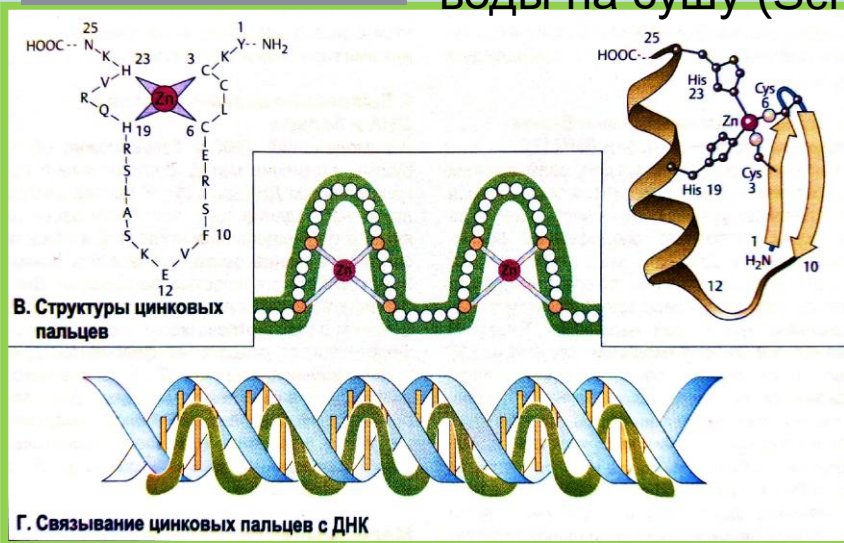
Поскольку зеленые водоросли и наземные растения (бриофиты и сосудистые растения) произошли от общего предка, изучение факторов транскрипции семейства GATA водорослей может пролить свет на происхождение кодирующих их генов, которые предположительно оказались важны для эволюции растений, способствовав их переходу из воды на сушу (Schwechheimer et al., 2022).

Предмет исследования

Транскрипционные факторы семейства GATA – специфические белки, содержащие цинк-пальцевые структуры. Они связываются с консенсусной последовательностью (A/T)GATA(A/G).

ДНК-связывающие GATA-факторы найдены только у эукариот, включая позвоночных, беспозвоночных, растения и грибы.

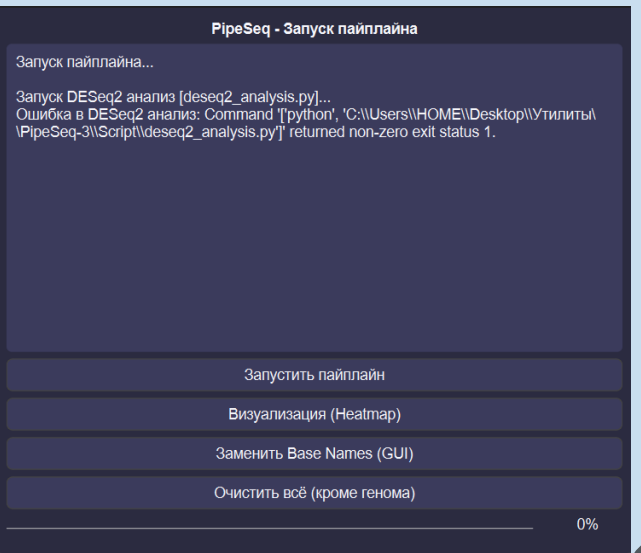
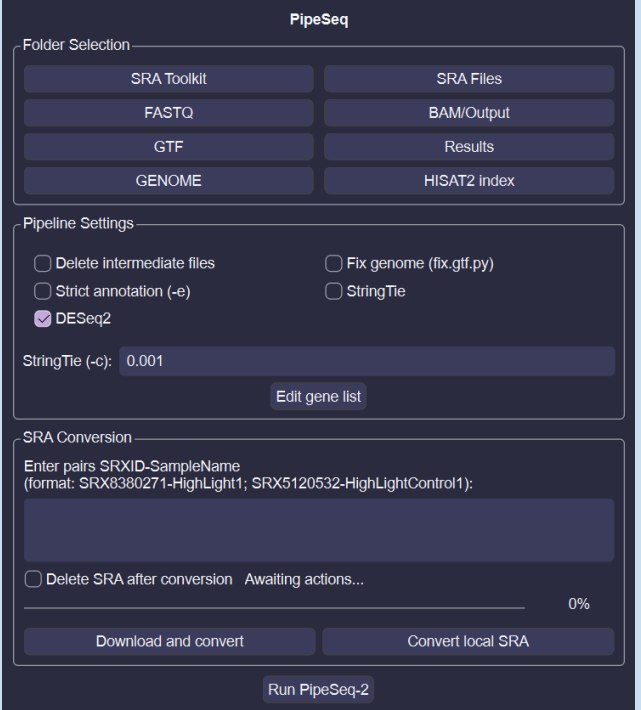
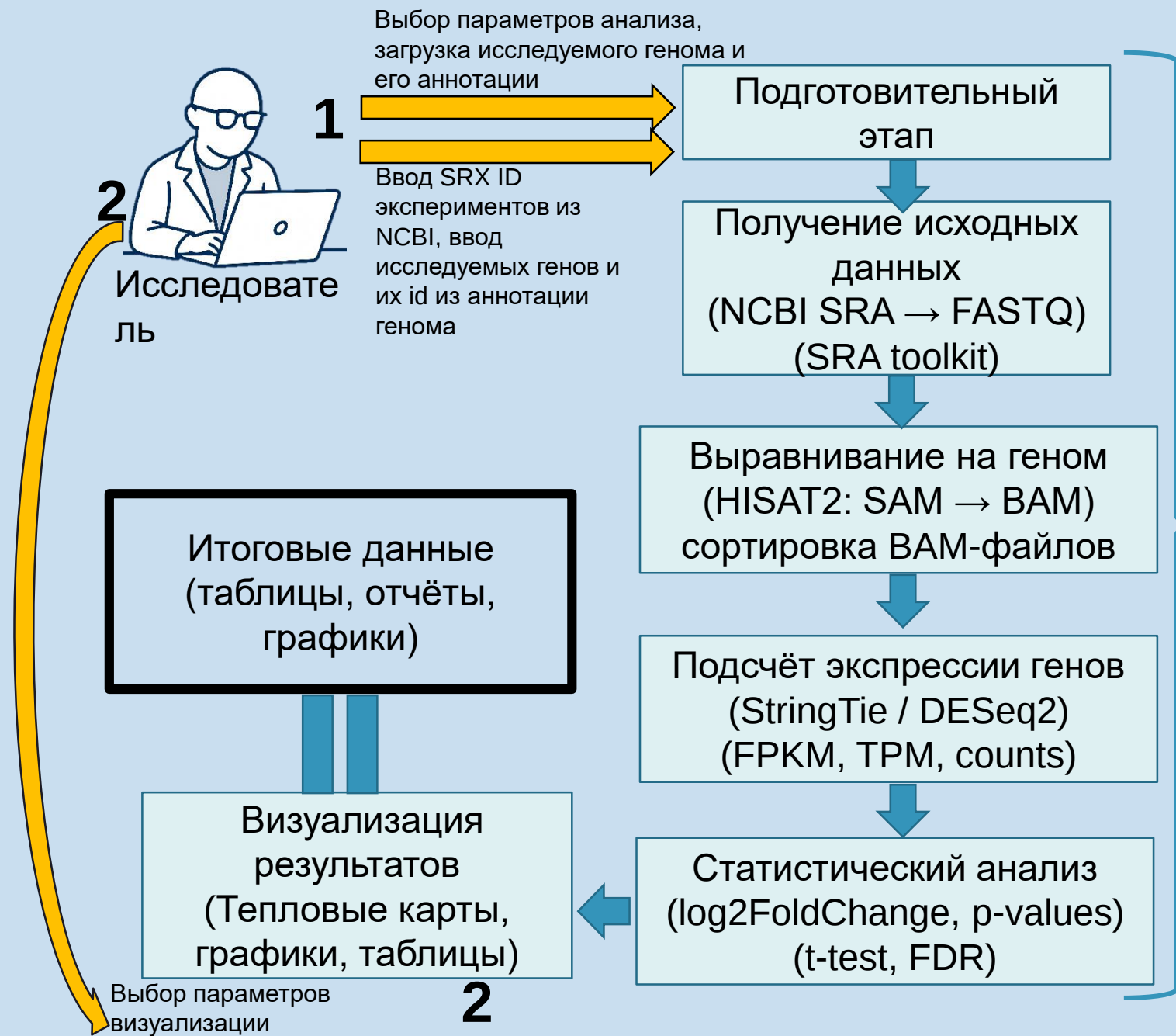
- У высших растений показана их роль в контроле роста и развития и в ответе на различные стрессы, а также в метаболизме азота и хлорофилла;
- У грибов GATA-факторы регулируют азотный и углеродный метаболизм.



Структура цинковых пальцев (В) и их связывание с большой бороздкой ДНК (Г)

(Wirth, 2017 в пер. Тихомировой, 2020)

Этапы анализа: схема пайплайна



1

2

3

Выбор параметров визуализации

Настройки тепловой карты

Заголовок: Тепловая карта 111DeSeq2 Log2 Fold Change

Подпись X (вверху): Названия генов

Подпись Y (слева): Условия

Цвет min (нижний): blue

Цвет 0 (середина): white

Цвет max (верхний): green

☒ Показывать p-value

Текущие настройки:
Заголовок: Тепловая карта 111DeSeq2 Log2 Fold Change
Подпись X: Названия генов
Подпись Y: Условия
Цвет min: blue
Цвет 0: white
Цвет max: green
Показывать p-value: True
Сортировать условия по max экспрессии: True
Сортировать гены по max экспрессии (Control первым): True
Убрать цвет у p-value > 0.05 и уменьшить шрифт: False
(Порядок условий и генов будет сохранён при сохранении)

☒ Сортировать условия по максимальной экспрессии (log2)

☒ Сортировать гены по максимальной экспрессии (Control всегда первым)

☐ Убрать цвет у ячеек с p-value > 0.05 и уменьшить шрифт в 2 раза

Порядок условий:

- 30 минут акклиматиза
- 30 минут в темноте, 6
- 6 часов в темноте, бе
- Акклиматизация к тем
- Выращивание в кисл
- Выращивание в щелс

Порядок генов:

- GATA-1
- GATA-2
- GATA-3
- GATA-4
- GATA-5
- GATA-6

Сохранить и продолжить

Интерфейс настройки тепловой карты

Замена Base Name

Выбран файл: C:/Users/HOME/Desktop/Утилиты/PipeSeq-3/Results/Deseq2.txt

Открыть файл

30 минут акклиматизации к свету → Новое имя (оставь пустым, чтобы не ...)

30 минут в темноте, без кислорода → Новое имя (оставь пустым, чтобы не ...)

6 часов в темноте, без кислорода → Новое имя (оставь пустым, чтобы не ...)

Акклиматизации к темноте → Новое имя (оставь пустым, чтобы не ...)

Выращивание в кислой среде 1 день → Новое имя (оставь пустым, чтобы не ...)

Выращивание в щелочной среде 1 день → Новое имя (оставь пустым, чтобы не ...)

Дефицит железа → Новое имя (оставь пустым, чтобы не ...)

Дефицит меди → Новое имя (оставь пустым, чтобы не ...)

Дефицит нитратов → Новое имя (оставь пустым, чтобы не ...)

Избыток железа → Новое имя (оставь пустым, чтобы не ...)

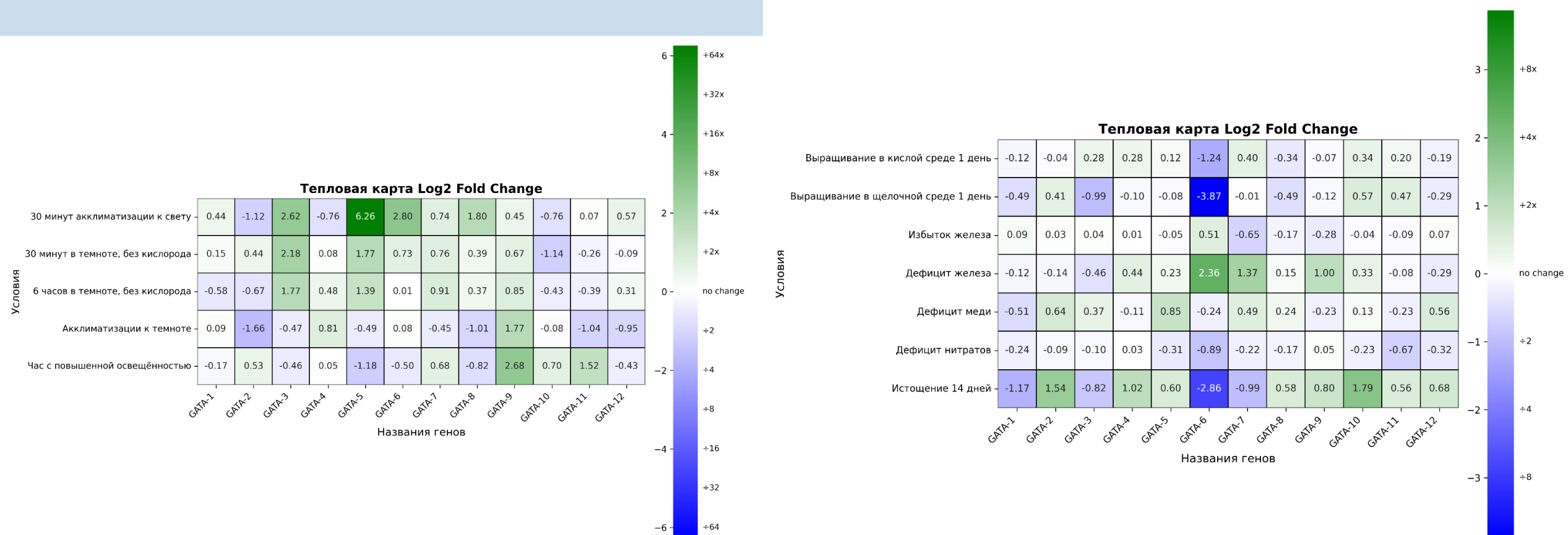
Истощение 14 дней → Новое имя (оставь пустым, чтобы не ...)

Час с повышенной освещённостью → Новое имя (оставь пустым, чтобы не ...)

Заменить и перезаписать файл

Интерфейс скрипта быстрой замены имён в файле результатов

Варианты полученных тепловых карт как результат работы программы



Заключение

Программа PipeSeq представляет собой настраиваемый и удобный инструмент для пошаговой или полной автоматической обработки данных RNA-seq. Она поддерживает работу с paired-end и одиночными FASTQ-файлами, автоматизирует этапы от выравнивания (HISAT2) и подсчёта экспрессии (StringTie) до статистического анализа (log2FoldChange, p-value) и визуализации уровня экспрессии анализируемых генов (тепловые карты).

Среди преимуществ PipeSeq:

1. минимальные требования к ручной настройке;
2. удобный GUI-интерфейс на PyQt6;
3. поддержка параллельной обработки и контроль качества данных на каждом этапе;
4. возможность гибкой работы как с публичными SRA-данными, так и с локальными экспериментами;
5. модульная архитектура и прозрачный лог-файл выполнения для отслеживания всех этапов.

PipeSeq успешно был нами применён для анализа экспрессии генов семейства GATA у *Chlamydomonas reinhardtii*, что позволило выявить изменения уровня транскрипции этих генов при различных условиях культивирования. Полученные данные служат основой для дальнейших функциональных исследований *in situ*. В перспективе PipeSeq может быть адаптирован для анализа RNA-seq данных других видов, а также использоваться в учебных курсах по биоинформатике как пример простого и гибкого инструмента анализа транскриптомов.

Спасибо за внимание!

Работа выполняется при поддержке СПбГУ, шифр проекта 124032000041-1