Тема работы: Ранний и поздний транскрипционный ответ на низкие концентрации экзогенной салициловой кислоты в корне *Arabidopsis thaliana* L.

Состав коллектива:

Елгаева Елизавета Евгеньевна (владелец учетной записи, проводивший работу на кластере); elgaeva.liza@yandex.ru; место работы: Лаборатория компьютерной транскриптомики и эволюционной биоинформатики НГУ, лаборант-исследователь; место учебы: ФЕН НГУ, кафедра информационной биологии, 4 курс, группа 14412, обучение закончено 25.06.2018; с помощью ИВЦ НГУ выполнялась дипломная работа, которую планируется продолжить и довести до публикации во время обучения в магистратуре

Миронова Виктория Владимировна (руководитель); kviki@bionet.nsc.ru; место работы: ИЦиГ СО РАН, к. б. н., в. н. с. Сектора системной биологии морфогенеза растений, заведующая Лабораторией компьютерной транскриптомики и эволюционной биоинформатики НГУ

Землянская Елена Васильевна (соруководитель); ezemlyanskaya@bionet.nsc.ru; место работы: ИЦиГ СО РАН, к. б. н., с. н. с. Сектора системной биологии морфогенеза растений, заведующая Лабораторией компьютерной транскриптомики и эволюционной биоинформатики НГУ

Лашин Сергей Александрович (соруководитель); lashin@bionet.nsc.ru; место работы: ИЦиГ СО РАН, к. б. н., в. н. с. Сектора компьютерного анализа и моделирования биологических систем, Лаборатория транскриптомики растений, с. н. с Лаборатории компьютерной транскриптомики и эволюционной биоинформатики НГУ

Новикова Дарья Дмитриевна (экспериментатор); место работы: ИЦиГ СО РАН, Сектор системной биологии морфогенеза растений, инженер; место учебы: ИЦиГ СО РАН, окончила 3 курс аспирантуры

Мухин Алексей Максимович (программист-консультант); место работы: ИЦиГ СО РАН, научно-образовательный отдел, лаборант; место учебы: ФИТ НГУ, кафедра систем информатики, группа 17227, окончил 1 курс магистратуры

Научное содержание работы:

1. Постановка задачи.

Целью данной работы было выявление генов-мишеней в пути передачи сигнала низких концентраций салициловой кислоты в раннем и позднем ответе в корне *Arabidopsis thaliana* L.

Для достижения поставленной цели нами были выдвинуты следующие **задачи**:

1) провести биоинформатический анализ RNA-Seq эксперимента по действию экзогенной салициловой кислоты на корень *A. thaliana* (1 ч и 6 ч, 20 мкмоль/л);

2) выявить первичные и вторичные мишени в транскрипционном ответе на низкие концентрации СК;

3) провести функциональную аннотацию генов-мишеней в раннем и позднем ответе на низкие концентрации СК.

2. Современное состояние проблемы.

Салициловая кислота (СК) — это фенольное соединение, являющееся гормоном стресса растений, обеспечивающим реализацию защитных механизмов в ответ на абиотические и биотические стрессорные факторы. Кроме того, СК выполняет в организме растений роль регулятора морфогенеза – процессов роста и развития. Хронологически второй эффект СК был установлен позднее и по-прежнему недостаточно изучен.

Молекулярно-генетические механизмы действия СК традиционно изучались на наземных частях растений и с применением концентраций, вызывающих защитные реакции и СПУ (системную приобретенную устойчивость растений к патогенам) - более 100 мкмоль/л. Важно также, что в большинстве исследований изучался поздний транскрипционный ответ на высокие концентрации СК, данных по раннему ответу (до 1 часа) недостаточно для понимания механизмов, активирующих ответ на СК. По этой причине молекулярно-генетические механизмы, лежащие в основе различий в фенотипических эффектах от воздействия низких и высоких концентраций СК на развитие корня, остаются неизвестными.

Данная дипломная работа стала продолжением исследования действия различных концентраций СК в корнях растений, проводившихся в Секторе системной биологии морфогенеза ИЦиГ СО РАН. В ходе этой работы мои коллеги провели RNA-Seq эксперимент для изучения изменений транскриптома в клетках корней *A. thaliana*, индуцированных низкими концентрациями СК (20 мкмоль/л) при раннем (1 час) и позднем (6 часов) ответе. Полученные RNA-Seq данные были переданы мне для проведения биоинформатического анализа.

3. Подробное описание работы, включая используемые алгоритмы.

В рамках дипломной работы проводился биоинформатический анализ данных RNA-Seq.

Препроцессинг данных RNA-Seq

На первом этапе был проведен контроль качества транскриптомных данных при помощи программы FASTQC v0.11.5. Затем для удаления адаптеров и фрагментов их последовательностей была применена опция ILLUMINACLIP в программе Trimmomatic-0.36, с использованием встроенной PE-библиотеки адаптеров TruSeq3-PE. Для оценки качества данных, прошедших предобработку в программе Trimmomatic-0.36, был повторно проведен контроль качества ридов в FASTQC v0.11.5.

Картирование ридов на геном

Предобработанные данные были картированы в программе STAR-2.5.4b с привлечением файла аннотации из базы данных Ensemble. Референсный геном был собран из отдельных данных по каждой хромосоме. Для проверки статистики картирования использовалась программа SAMtools-1.7, а подсчет числа ридов в генах был осуществлен с применением R-пакета HTSeq-1.7.

Выявление дифференциально-экспрессирующихся генов

Анализ дифференциальной экспрессии был проведен в программах DESeq2 v1.18.1 и edgeR v3.20.9. Для оценки достоверности данных использовалась поправка на множественное сравнение Бенджамини-Хохберга. При обработке в DESeq2 v1.18.1 функция автоматической фильтрации ДЭГ с низким уровнем изменения экспрессии была отключена (independentFiltering = FALSE). ДЭГ были отфильтрованы в Excel по значению q-value<0.05 и разделены на активируемые и подавляемые в ответ на СК по знаку log2FoldChange.

Функциональная аннотация

Функциональная аннотация ДЭГ проводилась при помощи веб-сервиса AgriGO v2.0 с использованием инструмента SEA (Singular Enrichment Analysis). Для статистической обработки данных был применен тест Фишера и поправка на множественное сравнение Бенджамини-Хохберга. Другие параметры SEA не менялись.

Помимо AgriGO для функциональной аннотации использовался DAVID v6.8 - инструменты Functional Annotation для выявления функциональных кластеров и Gene Name Batch Viewer для получения списка названий ДЭГ.

4. Полученные результаты.

Первичный анализ СК-индуцированных транскриптомов *A. thaliana*

Первичный контроль качества данных, полученных из BGI Tech, показал следующие характеристики входных данных: высокая степень покрытия (21 - 22 М как для прямых, так и для обратных ридов), содержание нуклеотидов G и C - 45%, длина прочтений 100 п. н., качество нуклеотида в каждой позиции оценивается значением score>25 (для прямых прочтений >30), адаптерные последовательности не выявлены. В совокупности данные характеристики свидетельствуют о высоком качестве ридов. Однако было обнаружено обогащение k-мерами на 5'- и 3'-концах ридов. Анализ наиболее представленных k-мерных последовательностей показал, что некоторые из них являются фрагментами адаптеров. С целью их удаления была проведена предобработка ридов в программе Trimmomatic. Повторный контроль качества данных, прошедших препроцессинг, показал, что количество k-меров сократилось, удаление фрагментов адаптеров существенно не повлияло на длину ридов и не сказалось на качестве данных.

В ходе картирования парных ридов для каждой реплики картировалось примерно 42 М прочтений (96%), порядка 94% прочтений картировалось уникально.

В пользу высокого качества RNA-Seq также свидетельствуют графики (рис. 1-3), полученные при помощи программы DESeq2. Как для данных по одночасовой обработке корней экзогенной СК, так и для данных по шестичасовой обработке была выявлена высокая степень кластеризации образцов в соответствии с дизайном эксперимента (контроль-опыт) (рис. 1). Анализ данных методом главных компонент также показал четкое разделение образцов по признаку «контроль-опыт» (рис. 2, 3).

Поиск дифференциально-экспрессирующихся в ответ на обработку СК генов

Анализ дифференциальной экспрессии проводили двумя методами: DESeq2 и edgeR. Программы выдали сходные списки ДЭГ, однако сравнение результатов работы программ DESeq2 и edgeR показало, что edgeR обнаруживает большее количество ДЭГ. Несмотря на то, что программа DESeq2 обнаружила меньшее количество ДЭГ, чем edgeR, она выявляла больше генов со значительным (2<log2FoldChange) изменением уровня экспрессии. Небольшое расхождение результатов, предоставляемых программами, обусловлено различиями в алгоритмах их работы, в частности, нормализации.

На основании сравнительного анализа выходных данных DESeq2 и edgeR, нами было принято решение о проведении функциональной аннотации только ДЭГ, обнаруженных программой DESeq2, данные edgeR в дальнейший анализ не вовлечены.

Функциональная аннотация дифференциально-экспрессирующихся генов

В ходе функциональной аннотации ДЭГ в DAVID и AgriGO было выявлено подавление и активация различных физиологических процессов растений. Среди ДЭГ найдено большое количество различных транскрипционных факторов, белков-транспортеров и передатчиков сигнала. Результаты функциональной аннотации хорошо согласуются с литературными данными и подтверждают высокое качество данной работы.

5. Иллюстрации, визуализация результатов

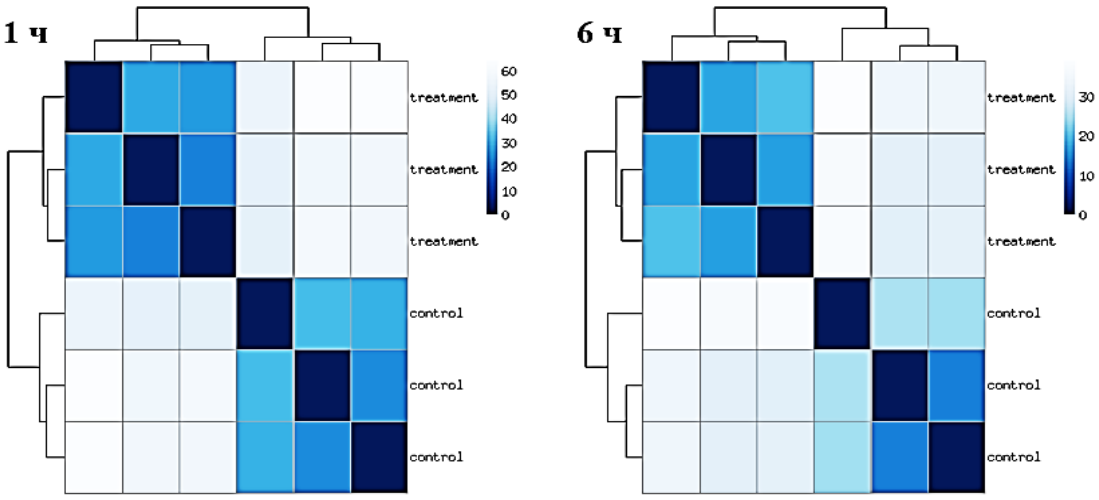

Рисунок 1 – Кластерный анализ RNA-Seq данных по 1ч- и 6ч-обработке СК, тепловые карты. Шкала отображает Евклидово расстояние между образцами

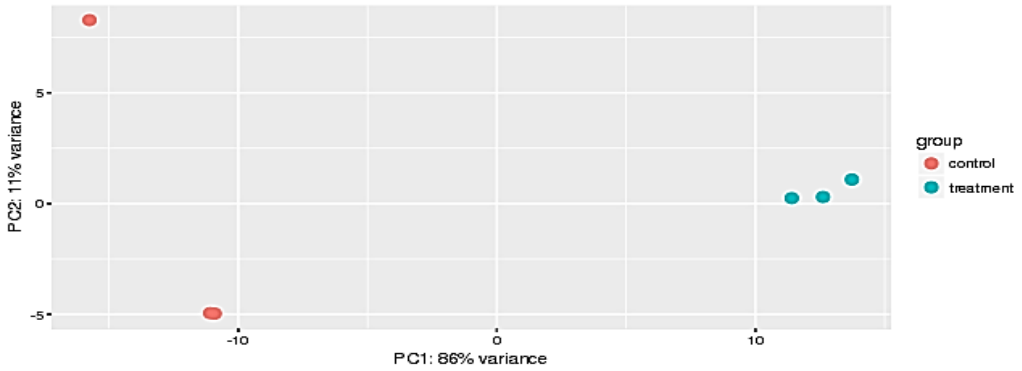

Рисунок 2 – Визуализация RNA-Seq данных по 1ч-обработке СК, метод главных компонент

Рисунок 3 – Визуализация RNA-Seq данных по 6ч-обработке СК, метод главных компонент

Эффект от использования кластера в достижении целей работы: все этапы биоинформатического анализа, за исключением функциональной аннотации дифференциально экспрессирующихся генов (осуществлялась с применением GUI),

проводились с использованием вычислительных мощностей ИВЦ НГУ, что позволило хранить и быстро обрабатывать большие массивы данных.

Перечень публикаций, содержащих результаты работы:

Материалы 56-й Международной научной студенческой конференции МНСК-2018: Биология/ Новосибирский Государственный Университет, Новосибирск, 2018, 202 с. (Доклад по данной теме занял первое место в секции «Биоинформатика».)

К публикации в сборнике материалов конференции BGRS\SB’2018, которая пройдет в августе 2018 года, также приняты тезисы по данной работе. Сборнику присваивается индекс ISBN, по окончании мероприятия он будет размещен в РИНЦ.

Ведется работа по написанию статьи в научный журнал на основе проделанного исследования.