

УДК 574.4:631.46

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭКСПРЕССИИ ГЕНОВ СТРЕССОВОГО ОТВЕТА У РАСТЕНИЙ ПО ДАННЫМ ПУБЛИЧНЫХ ТРАНСКРИПТОМНЫХ БАЗ

Грекова М.В.

аспирант кафедры физиологии растений Московского государственного университета им. М.В. Ломоносова

e-mail: grekov.AMV@mail.ru

Аннотация

Представлена обзорно-методическая схема сравнительного анализа транскриптомных ответов растений на засуху, засоление и комбинированные абиотические воздействия. Рассмотрены три общедоступные серии RNA-seq *Arabidopsis thaliana* из базы Gene Expression Omnibus: GSE201379, GSE113138 и GSE146206. Определены критерии выбора сопоставимых образцов и последовательность обработки прочтений, оценки дифференциальной экспрессии и функционального обогащения. Систематизированы ожидаемые паттерны для генов АВА-сигнализации, осмотической защиты, ионного гомеостаза, антиоксидантной системы и фотосинтеза на основе опубликованной литературы. Обоснована необходимость анализа взаимодействия стрессовых факторов и контроля возможных эффектов времени, ткани и лаборатории. Статья не содержит вымышленных результатов самостоятельного пересчета RNA-seq.

Ключевые слова: *транскриптомика, RNA-seq, Arabidopsis thaliana, абиотический стресс, засуха, засоление, дифференциальная экспрессия, DESeq2, GEO.*

COMPARATIVE ANALYSIS OF PLANT STRESS-RESPONSE GENE EXPRESSION USING PUBLIC TRANSCRIPTOMIC DATABASES

Grekoval M.V.

Postgraduate student at the Department of Plant Physiology, Lomonosov Moscow State University

e-mail: grekov.AMV@mail.ru

Abstract

This review and methodological study outlines a comparative RNA-seq workflow for plant responses to drought, salinity and combined abiotic stresses. Three public *Arabidopsis thaliana* series from the Gene Expression Omnibus are examined: GSE201379, GSE113138 and GSE146206. Sample eligibility, read-quality assessment, gene quantification, differential-expression inference and functional enrichment are described. Literature-supported response patterns are considered for abscisic-acid signalling, osmoprotection, ion homeostasis, antioxidant defence and photosynthetic genes. The study emphasizes treatment-by-time interactions, tissue-specific effects and the limitations of cross-study integration. No newly computed RNA-seq differential-expression results are claimed.

Keywords: *transcriptomics, RNA sequencing, Arabidopsis thaliana, abiotic stress, drought, salinity, differential expression, DESeq2, GEO.*

Введение

Сохранение продуктивности растений в условиях изменения климата требует понимания механизмов ответа на водный дефицит, засоление и температурные воздействия. При этом одно и то же физиологическое проявление, например, замедление роста – может сопровождаться различными молекулярными событиями в зависимости от силы и длительности стресса. Современная транскриптомика предоставляет возможность оценивать активность тысяч генов одновременно, но сопоставление экспериментов остается непростой статистической задачей.

База Gene Expression Omnibus (GEO) хранит как необработанные данные секвенирования, так и обработанные матрицы экспрессии. Использование публичных наборов

снижает стоимость исследования и позволяет повторно проверить опубликованные гипотезы. Однако доступность данных не означает их автоматическую сопоставимость: различия генотипов, возраста растений, освещения, ткани, продолжительности воздействия, платформы и стратегии подготовки библиотек могут превышать собственно эффект стресса [1, 2].

Цель настоящей работы – предложить воспроизводимую методику сравнения групп генов стрессового ответа на основе проверяемых публичных RNA-seq серий. Для этого необходимо обосновать выбор наборов данных, унифицировать этапы вычислительной обработки, определить подход к статистическому выявлению дифференциально экспрессирующихся генов и сформулировать правила интерпретации функциональных категорий. В отличие от исследования с самостоятельно рассчитанной матрицей экспрессии, здесь выполняется критическое сопоставление протоколов, биологических результатов опубликованных работ и аналитических возможностей доступных массивов.

Методы и методики

В качестве модельного вида выбран *Arabidopsis thaliana*: для него доступны качественная референсная сборка, аннотация генов и многочисленные исследования регуляторных путей. Предварительные условия включения серии: публичный идентификатор GEO, эксперимент с высокопроизводительным секвенированием RNA, наличие биологических контролей, доступность метаданных об обработке и достаточное количество повторов для оценивания внутригрупповой вариации. Массивы микрочипов исключены из основной сравнительной схемы, поскольку интенсивности гибридизации нельзя непосредственно объединять с числами RNA-seq прочтений.

Серия GSE201379 посвящена развитию засухи у *Arabidopsis* после предварительной обработки этанолом либо без нее. Отбор надземной части проводился в несколько сроков от начала прекращения полива: 0, 1, 3, 5 и 7 суток [3]. Для исследования эффекта засухи необходимы сравнения внутри одинаковых уровней фактора «этанол» и в пределах одного времени, иначе взаимодействие предварительной обработки с дефицитом воды будет ошибочно интерпретировано как общий стрессовый ответ.

Серия GSE113138 описывает различие между фазами ответа на соль при постепенном введении NaCl и условиях солевого шока. Для каждой из трех групп – контроль, Phase 0 и Phase 1 – указано по три повтора [4]. Это позволяет поставить корректный вопрос о том, какие транскрипционные признаки относятся к немедленной реакции, а какие – к более длительной адаптации, однако нельзя без дополнительных допущений отождествлять этапы с временными точками засухи.

В GSE146206 исследованы отдельные и комбинированные воздействия соли, осмотического и теплового стресса [5]. В опубликованном описании показано, что сочетания с нагреванием существенно перестраивают глобальный профиль экспрессии и метаболитов. Именно поэтому серия полезна не для создания искусственной «средней реакции на стресс», а для проверки гипотезы о неаддитивности совместных воздействий. Перед анализом требуется уточнять схему повторов для каждого конкретного контраста.

Результаты

Наиболее изученное направление – передача сигналов абсцизовой кислоты (ABA). Изменение водного режима может запускать каскады, связанные с рецепторами PYR/PYL, фосфатазами PP2C, киназами SnRK2 и транскрипционными факторами ABF/AREB. Гены этих семейств представляют интерес как кандидаты для оценки раннего сигнального ответа; однако величина экспрессии отдельного компонента не тождественна активности всего пути, регулируемого также на уровне фосфорилирования [6].

Осмотическая защита связана с накоплением совместимых осмолитов и защитных белков. К этой группе относят, в частности, гены биосинтеза пролина, дегидринов и LEA-белков. Их индукция может быть характерна для дегидратации и засоления, но временная динамика зависит от физиологической стадии и степени повреждения. Для обоснования роли гена желательно сопоставлять его экспрессию с измеряемыми признаками водного потенциала, содержанием пролина и восстановлением роста после снятия воздействия.

Для солевого стресса особое значение имеют транспорт ионов и регуляция соотношения Na^+/K^+ , включая системы SOS и вакуолярные транспортеры. При длительном засолении изменения экспрессии транспортных компонентов могут отражать механизм поддержания ионного гомеостаза. Вместе с тем однократный транскриптомный срез не определяет реального потока ионов или устойчивости растения: необходима проверка физиологических показателей [6].

Антиоксидантная защита охватывает супероксиддисмутазы, каталазы, аскорбатпероксидазы и связанные с ними редокс-системы. Их транскрипционная регуляция может быть общей частью ответа на разные типы стресса, но знак изменения зависит от интенсивности окислительного воздействия. Гены фотосинтетического аппарата, напротив, в условиях тяжелых комбинированных воздействий нередко демонстрируют снижение активности; в GSE146206 выраженное влияние тепла на транскрипцию фотосинтетических белков отдельно отмечено авторами [5].

Сначала формируется паспорт каждого образца: идентификатор GEO/SRA, генотип, ткань, возраст, вариант воздействия, доза, время, биологический повтор, прибор секвенирования, парность прочтений и номер лабораторной партии. При отсутствии нужного поля образец не следует автоматически относить к контрольной группе. Таблица метаданных должна сохраняться вместе с кодом анализа; именно она определяет корректность дизайна.

Затем FASTQ-файлы проверяют на качество оснований, содержание адаптерных последовательностей, избыток определенных фрагментов, долю низкокачественных прочтений и число дубликатов. Сведения о правилах контроля и необходимости единого референса систематизированы в руководстве A. Conesa и соавторов [1]. Для сравнения серий целесообразна одна версия аннотации *Arabidopsis* и одинаковая процедура подсчета прочтений на уровне генов.

После удаления технических адаптеров прочтения сопоставляют с референсным геномом либо осуществляют псевдовыравнивание по транскриптому с последующим агрегированием оценок к генам. Для DESeq2 необходимы исходные или корректно восстановленные целочисленные счета; подстановка TPM либо FPKM в качестве независимых исходных счетов искажает модель дисперсии. Отдельно проверяют общую долю сопоставленных прочтений и перекосы библиотеки, а образцы с аномальными характеристиками документированно исключают только по заранее заданным критериям.

Контроль матрицы включает распределения нормализованных счетов, главные компоненты и расстояния между повторностями. Если образцы группируются преимущественно по лаборатории или серии GEO, общий анализ без учета эффекта партии ненадежен. Предпочтительный вариант для трех рассмотренных исследований – отдельная оценка контрастов в каждом наборе с последующим функциональным сопоставлением. Прямое объединение сырых счетов возможно лишь при доказанной достаточной совместимости дизайнов и явном включении партии в модель.

В DESeq2 числа прочтений моделируются отрицательно-биномиальным распределением с размерным фактором библиотеки, средним уровнем и дисперсией [2]. Для каждой серии дизайн должен учитывать ее реальные экспериментальные факторы: в GSE201379 – время, предварительную обработку этанолом и их взаимодействие; в GSE113138 – категориальный фактор фазы засоления; в GSE146206 – главные эффекты воздействий и проверяемые взаимодействия между ними. Не следует интерпретировать контраст, если хотя бы в одной из его групп отсутствуют независимые биологические повторы.

Для гена рассчитывают логарифм отношения экспрессии и скорректированное значение p при множественном тестировании. Практический критерий можно задать как FDR менее 0,05 и абсолютное значение $\log_2\text{FC}$ не менее 1, но выбор порога необходимо обосновать исходной задачей и провести анализ чувствительности. Для слабовыраженных генов более корректно использовать усадку оценок эффекта и избегать ранжирования по крайне нестабильным логарифмам отношений [2].

В качестве меры сравнимости между сериями более надежны согласованность знака эффекта, процент дифференциально экспрессирующихся генов внутри заранее определенного набора и обогащение функциональных категорий. Само количество значимых генов сильно зависит от глубины секвенирования, дисперсии и числа повторов. Поэтому превосходство одного воздействия по числу DEG нельзя напрямую интерпретировать как более сильную биологическую реакцию.

Список генов должен сопровождаться версией аннотации, способом исключения низких счетов, названием контраста, числом повторов и методом корректировки значимости. Для проверки устойчивости целесообразны альтернативные фильтры низкоэкспрессированных генов, оценка влияния единичных выбросов и сопоставление результатов с функционально ориентированным анализом рангов всех протестированных генов.

$$K_{ij} = NB(\mu_{ij}, \alpha_i), \quad (1)$$

где K – число прочтений гена i в образце j ;

μ – ожидаемое значение с учетом размера библиотеки;

α – параметр дисперсии.

$$\log_2 FC_i = \log_2 \frac{\mu_{i, stress}}{\mu_{i, control}}, \quad (2)$$

Отношение интерпретируется для строго определенного контраста, а не для несопоставимых выборок из разных исследований.

$$FDR = BH(p_1, p_2, \dots, p_m), \quad (3)$$

Процедура Бенджамини-Хохберга учитывает множественные проверки; скорректированное значение p само по себе не задает биологическую значимость величины эффекта.

Обсуждение

Доступные описания трех серий демонстрируют разные научные вопросы. В GSE201379 акцент сделан на прогрессировании водного дефицита при наличии дополнительного фактора предварительной обработки. В GSE113138 изучаются фазы солевого стресса, позволяющие различать ранние и последующие изменения. В GSE146206 оцениваются не только отдельные воздействия, но и комбинации, которые не обязаны подчиняться сумме отдельных ответов [3-5].

Для корректного сравнительного анализа предлагается разделять универсальный стрессовый ответ и контекстно-зависимые механизмы. Общими кандидатными категориями являются АВА-сигнализация, осмопротекторы и редокс-регуляция. Отдельно рассматривают транспорт ионов при солевом воздействии и снижение транскрипции фотосинтетических компонентов в комбинациях с теплом. Такой синтез согласуется с современным обзором молекулярных механизмов абиотического стресса [6], но не заменяет реальный подсчет пересечений DEG на уровне идентификаторов генов.

Рассматриваемые материалы не дают оснований сообщать точное число «общих стрессовых генов», средний $\log_2 FC$ по группам либо собственные коэффициенты корреляции между тремя сериями: унифицированный пересчет исходных счетов в данной работе не проводился. При последующем воспроизводимом анализе целесообразно представлять диаграммы распределения эффектов, матрицу согласования функциональных категорий и коэффициенты устойчивости при повторной подвыборке, а не ограничиваться диаграммой пересечения списков.

При объединении опубликованных списков важно различать устойчивость к стрессу и сам факт активации ответа. Сильная экспрессия защитных генов может характеризовать поврежденное растение, которому пришлось компенсировать воздействие, тогда как толерантный генотип способен ограничить повреждение без столь высокой транскрипционной перестройки. Соотносить молекулярные маркеры с устойчивостью следует лишь при наличии фенотипических данных и сопоставимых критериев.

Следовательно, интерпретация ограничена устойчивыми содержательными различиями дизайна и литературными выводами; точный общий список DEG должен быть результатом отдельного воспроизводимого расчета.

Основное ограничение вторичного анализа заключается в гетерогенности протоколов. Разные способы определения засухи (уменьшение влажности субстрата, прекращение полива, осмотическое воздействие маннитола) нельзя считать эквивалентными. Солевая нагрузка зависит от концентрации, скорости ее введения и продолжительности. Дополнительную неоднородность вносят фотопериод, циркадная фаза и орган растения, в котором измеряется экспрессия.

Статистический риск связан с псевдорепликацией: несколько технических библиотек из одного биологического образца не создают новые независимые наблюдения. Другой риск – селективная публикация только наиболее значимых генов, из-за чего списки из статей могут отличаться от результатов одинакового полного пайплайна. Недостаточно просто указать программу: требуются версии библиотек, команды, параметры, проверка контрольных сумм и фиксированный файл метаданных.

Для дальнейшего развития работы обоснован репозиторий с манифестом образцов, скриптами загрузки, средой исполнения и тремя отдельными отчетами дифференциальной экспрессии. Затем формируется согласованная таблица направлений изменения групп генов. В случае расширения на другие виды понадобится строгая схема ортологов, поскольку идентичные названия функциональных семейств не гарантируют взаимно-однозначного соответствия генов.

Практическая ценность подхода состоит в том, что даже без проведения нового опыта можно получить методологически проверяемое исследование и сформировать гипотезы для последующих RT-qPCR подтверждений. При этом публикация расчетных результатов допустима только после фактического выполнения анализа данных, документирования критериев контроля и оценки неопределенности.

Заключение

Сопоставление публичных RNA-seq исследований представляет собой обоснованный способ изучения стрессовых ответов растений при ограниченных лабораторных ресурсах. Три отобранные серии GEO позволяют рассматривать временную динамику засухи, фазы солевого воздействия и неаддитивность комбинированных стрессов у *Arabidopsis thaliana*. Их назначение различается, что исключает механическое слияние данных без учета экспериментального дизайна.

Наиболее информативным решением является единая процедура контроля качества, подсчета прочтений и статистической обработки при раздельной оценке контрастов. Для биологической интерпретации рекомендуются группы генов АВА-сигналикации, осмотической защиты, транспорта ионов, антиоксидантной системы и фотосинтеза. Направление и величина ответа требуют проверки отдельно для времени, ткани и типа воздействия.

Представленная схема определяет воспроизводимый порядок будущего сравнительного анализа и критерии его доказательности. Для заявления новых количественных результатов необходим повторный расчет исходных данных GEO; в настоящей статье сделаны обзорно-методические выводы, подтвержденные описаниями публичных серий и научными публикациями.

Литература

1. Conesa A., Madrigal P., Tarazona S. et al. A survey of best practices for RNA-seq data analysis // *Genome Biology*. 2016. Vol. 17. Art. 13.
2. Love M.I., Huber W., Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // *Genome Biology*. 2014. Vol. 15. Art. 550.
3. Gene Expression Omnibus. GSE201379: Determination of transcriptomic changes associated with progressive drought stress in ethanol treated vs non-treated control *Arabidopsis*

plants. 2022. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE201379> (дата обращения: 21.10.2026).

4. Gene Expression Omnibus. GSE113138: Transcript profile analysis suggests marker genes to distinguish between Phase 0 and Phase 1 responses of Arabidopsis to salt stress. 2021. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE113138> (дата обращения: 21.10.2026).

5. Gene Expression Omnibus. GSE146206: Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes. 2020. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE146206> (дата обращения: 21.10.2026).

6. Zhang H., Zhu J., Gong Z., Zhu J.-K. Abiotic stress responses in plants // Nature Reviews Genetics. 2022. Vol. 23. PP. 104-119.

7. Meta-analysis of public RNA sequencing data of abscisic acid-related abiotic stresses in Arabidopsis thaliana // Frontiers in Plant Science. 2024.

8. Gene Expression Omnibus. GSE199590: G-protein mutants RNA-seq analysis of salt response in Arabidopsis and Marchantia. 2023. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE199590> (дата обращения: 21.10.2026).

9. Gene Expression Omnibus. GSE108852: Control- and salt-treated Arabidopsis RNA-seq measurements (associated with Kramer et al.) URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108852> (дата обращения: 21.10.2026).