На правах рукописи

Вибе Даниил Станиславович

ВЫЯВЛЕНИЕ ВЗАИМОСВЯЗИ МЕЖДУ ВЕЛИЧИНАМИ ИЗМЕНЕНИЯ ЭКСПРЕССИИ И ФУНКЦИЯМИ ДИФФЕРЕНЦИАЛЬНО ЭКСПРЕССИРУЮЩИХСЯ ГЕНОВ НА ОСНОВЕ КОМПЬЮТЕРНОГО АНАЛИЗА ТРАНСКРИПТОМОВ АРАБИДОПСИСА И ЧЕЛОВЕКА

Математическая биология, биоинформатика
1.5.8

АВТОРЕФЕРАТ
Диссертации на соискание учёной степени
кандидата биологических наук

Новосибирск
2021

Работа выполнена в секторе системной биологии морфогенеза растений Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», г. Новосибирск.

Научный руководитель:	**Миронова Виктория Владимировна** к.б.н., в.н.с. сектора системной биологии морфогенеза растений ФГБНУ «ФИЦ Институт цитологии и генетики СО РАН», г. Новосибирск
Официальные оппоненты:	**Пенин Алексей Александрович** к.б.н., зав. лабораторией №19 ФГБУН «Институт проблем передачи информации им. А. А. Харкевича РАН», г. Москва
	Ратушняк Александр Савельевич д.б.н., в.н.с. отдела информационных технологий в медицине и биологии ФГБНУ «ФИЦ информационных и вычислительных технологий СО РАН», г. Новосибирск.
Ведущее учреждение:	Федеральное государственное бюджетное учреждение науки «Институт общей генетики им. Н.И. Вавилова РАН»

Защита диссертации состоится «___»_____ 2021г. на утреннем заседании диссертационного совета 24.1.239.01 (Д 003.011.01) на базе Федерального государственного бюджетного научного учреждения «Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук», в конференц-зале Института по адресу: пр. академика Лаврентьева 10, г. Новосибирск, 630090, тел.: (383) 363-49-06 (1321); факс: (383) 333-12-78; e-mail: dissov@bionet.nsc.ru

С диссертацией можно ознакомиться в библиотеке ИЦиГ СО РАН и на сайте института www.bionet.nsc.ru.
Автореферат разослан «___»_____ 2021г.

Учёный секретарь диссертационного совета,
доктор биологических наук Хлебодарова Т.М.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Биологические системы характеризуются сложными взаимодействиями между генами и биологически-активными веществами, активность которых может быть изучена через измерения уровней экспрессии генов. Технологии полногеномного анализа активности генов изменили методологию исследований биологических систем. В настоящее время процедура исследования полногеномной экспрессии генов методом RNA-Seq стала рутинной и может применяться для любого живого организма. Одним из стандартных подходов к анализу транскриптомных данных является поиск дифференциально экспрессирующихся генов (ДЭГ) с последующим анализом функционального обогащения. Методы функционального обогащения позволяют выявлять статистически значимое обогащение списка ДЭГ генами с одинаковой характеристикой, описанной в Генной Онтологии (ГО) (Ashburner et al., 2000) или в других словарях.

Выявление групп генов, связанных одной или несколькими характеристиками позволяет значительно сократить размерность данных полногеномных экспериментов и, следовательно, упрощает их дальнейший анализ. Анализ функционального обогащения генов суммирует информацию о процессах, которые усилились или ослабли при воздействии исследуемого стимула или состояния. Значимо обогащенные категории генов являются источником предсказаний генов-кандидатов и, например, используются для поиска предполагаемых онкогенов (Li et al., 2017), генов ассоциированных с развитием ишемической болезни сердца у человека (Balashanmugam et al., 2019), генов ответственных за формирование хозяйственно-ценных признаков (Ashburner et al., 2000), например, устойчивостью пшеницы к засоленности почвы (Xiong et al., 2017).

К сожалению, стандартные методы анализа функционального обогащения не используют значительную часть данных RNA-Seq. Например, количественные данные об уровнях экспрессии генов используются лишь косвенно, в качестве порога для выявления ДЭГ или для ранжирования генов. Как результат, существующие методы предсказывают лишь значимость изменения биологических процессов, но не силу, с которой происходит их изменение. Самыми значимыми могут оказаться, например процессы с большим количеством участников и небольшими, но значимыми степенями изменения экспрессии, но не процессы с меньшим количеством участников и сильными изменениями экспрессии некоторых из них. Таким образом, актуальной является задача разработки новых методов анализа функционального обогащения, способных выявлять статистически значимые взаимосвязи между функцией генов и степенью изменения их экспрессии, для получения новых знаний о молекулярно-генетической регуляции биологических процессов.

Цели и задачи работы

Целью данной работы является разработка метода функционального обогащения с учетом количественных данных о степени изменения экспрессии генов и его апробация в задачах анализа функционального обогащения в транскриптомных данных.

Для этого решаются следующие задачи:

1. Разработка алгоритмов и компьютерных приложений для анализа представленности функциональных групп генов в списке дифференциально-экспрессирующихся генов, с учетом степени изменения их экспрессии:

1.1. Разработка алгоритма для анализа обогащения групп генов категориями Генной Онтологии (ГО), с учетом разброса степеней изменения экспрессии.

1.2. Реализация разработанного алгоритма в виде пакета для языка программирования R с интегрированными средствами визуализации.
1.3. Оценка надежности разработанного метода и сравнение с существующими методами анализа представленности функциональных групп генов.
2. Апробация разработанного метода на данных десятков транскриптомных экспериментов.

Научная новизна

В настоящий момент стандартными методами анализа функционального обогащения являются SEA (анализ уникального обогащения, Singular Enrichment Analysis) (Huang et al., 2009) и GSEA (анализ обогащения набора генов, Gene Set Enrichment Analysis) (Subramanian et al., 2005). При использовании метода SEA информация о степени изменения экспрессии генов (fold-change) используется только на этапе отбора генов в список ДЭГ, а в GSEA значения степени изменения экспрессии могут быть использованы только при расчете метрики для ранжирования генов.

В данной работе был разработан новый метод анализа функционального обогащения FSEA (анализ фолд-специфичного обогащения, Fold-change Specific Enrichment Analysis), позволяющий выявлять статистически значимую взаимосвязь между функциональной характеристикой генов и степенью изменения их экспрессии в ответ на условия эксперимента. Применение FSEA на транскриптомных данных позволяет отранжировать категории ГО по силе транскрипционного ответа и более точно описать, какие изменения происходят в исследуемом образце и с какой силой. Тестирование FSEA на данных множества различных транскриптомных экспериментов показало существование множества ГО категорий, для которых характерна скоординированная фолд-специфическая экспрессия

вовлеченных генов. Для каждого эксперимента набор таких фолд-специфических ГО-категорий является уникальным.

Теоретическая и практическая значимость работы

Разработанный в данной работе метод FSEA дает исследователю дополнительную, ранее недоступную, информацию о силе транскрипционного ответа группы скоординированно-экспрессирующихся генов. С одной стороны это позволяет проранжировать процессы, которые происходят в анализируемой ткани по степени изменений (слабые, средние и сильные изменения). Например, в нашей работе по исследованию влияния экзогенного ауксина на корень растения, мы показали, что ГО категория “ответ на ауксин”, которая изучалась исследователями по всему миру как единственно важная, является лишь частным случаем транскрипционного ответа с сильной степенью изменения экспрессии генов. Есть и другие группы функционально-связанных и скоординированно-экспрессирующихся в ответ на ауксин генов, которые характеризуется меньшей степенью индукции/репрессии (Omelyanchuk et al., 2017).

Практическая значимость данной работы заключается в том, что FSEA позволяет лучше находить кандидатные гены для исследования причин масштабных изменений на молекулярно-генетическом уровне. Например, в исследовании данных по раку предстательной железы мы показали, что значительная часть дифференциально-экспрессирующихся генов, принадлежащих фолд-специфическим категориям, которые выявила FSEA, действительно описаны как онкосупрессоры (Wiebe et al., 2020).

Методология и методы диссертационного исследования

В данной работе разработан, протестирован и апробирован новый метод анализа функционального обогащения FSEA. В рамках анализа надежности разработанного метода, были оценены доля ложноположительных

результатов и чувствительность метода. Расчет доли ложноположительных результатов производился на пермутированных данных, полученных из реального транскриптомного эксперимента. Оценка чувствительности метода производилась на данных, сгенерированных из многомерного нормального распределения, содержащих заведомо известные группы генов с сильной внутригрупповой корреляцией по степени изменения экспрессии. Детальный анализ результатов апробации метода проведен на данных транскриптомных экспериментов по исследованию влияния фитогормона ауксина на экспрессию генов в корне *Arabidopsis thaliana* (Omelyanchuk et al., 2017) и изучению экспрессии генов в клеточной линии рака предстательной железы человека LNCaP (Wiebe et al., 2020).

Положения, выносимые на защиту:

1) Существует статистически достоверная взаимосвязь между функциональными характеристиками дифференциально экспрессирующихся генов и степенями изменения их экспрессии. Метод анализа фолд-специфичного обогащения выявляет эту взаимосвязь в транскриптомных экспериментах.
2) В клетках рака предстательной железы человека (LNCaP) активность генов, ассоциированных с важными для канцерогенеза процессами скоординирована не только по направлению изменения экспрессии (активация и ингибирование), но и по силе транскрипционного ответа.

Структура работы

Работа состоит из введения, списка публикаций по теме диссертации, обзора литературы, обзора использованных в работе материалов и методов, результатов, заключения, выводов, списка литературы (118 наименований) Материал изложен на 117 страницах, содержит 21 рисунок, 4 таблицы и 3 приложения.

Личный вклад автора

Основные результаты, изложенные в диссертации, получены автором самостоятельно. Автор участвовал в разработке алгоритма FSEA и самостоятельно реализовал его в программном пакете на языке R, тестирование пакета и апробация метода FSEA проводились автором лично.

Апробация результатов

Результаты работы вошли в отчет по гранту Российского Фонда Фундаментальных Исследований (№ 18-34-00871, руководитель Вибе Д.С.). Основные результаты были представлены на научных конференциях в виде устных и стендовых докладов: Всероссийская конференция с международным участием “Высокопроизводительное секвенирование в геномике” (HGS 2017, г. Новосибирск, Россия), международная конференция по биоинформатике регуляции и структуры геномов и системной биологии/симпозиум “Математическое моделирование и высокопроизводительные вычисления в биоинформатике, биомедицине и биотехнологии” (BGRS\SB-2018/MM&HPC-BBB-2018, г. Новосибирск, Россия), европейская конференция по вычислительной биологии (ECCB 2018, г. Афины, Греция), международная конференция по исследованию Арабидопсиса (ICAR 2019, г. Ухань, Китай), международная конференция “Математика. Компьютер. Образование” (МКО 2020, г. Дубна, Россия).

Метод FSEA, разработанный в рамках данной работы, опубликован в одном из крупнейших репозиториев биологического программного обеспечения Bioconductor (https://www.bioconductor.org) и имеет более ста скачиваний в месяц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Обзор литературы

В последнее десятилетие технология секвенирования транскриптома (RNA-seq) получила широкое применение в задачах исследования транскриптомов. В главе 1.1 рассматриваются основные этапы предобработки данных транскриптомных экспериментов, которые представлены процедурами оценки качества, картирования, квантификации и нормировки данных. В большинстве случаев, при анализе RNA-seq данных выявляют дифференциально-экспрессирующиеся гены и проводят анализ функционального обогащения. Основные методы такого анализа, такие как анализ уникального обогащения (SEA), анализ обогащения наборов генов (GSEA) и анализ взвешенной сети коэкспрессии генов (WGCNA) детально рассмотрены в главе 1.2. В главах 1.3 и 1.4 приведен краткий обзор экспериментов по исследованию транскрипционной активности генов в ответ на обработку ауксином в различных тканях арабидопсиса и в клетках рака предстательной железы человека.

Существующие методы анализа функционального обогащения не позволяют использовать информацию об активности генов в полной мере. А именно существующие подходы для анализа функционального обогащения (SEA, GSEA) не позволяют напрямую выявлять ассоциацию между функцией гена и степенью изменения его экспрессии. Поэтому целью данной работы является разработка метода анализа представленности функциональных групп генов с учетом степени изменения их экспрессии и его апробация в задачах анализа функционального обогащения в транскриптомных данных.

Разработка метода FSEA для анализа обогащения с учетом степени изменения экспрессии генов

В рамках настоящей работы нами был разработан метод анализа фолд-специфичного обогащения FSEA (Fold-change Specific Enrichment Analysis),

который позволяет выявлять категории Генной Онтологии (ГО), обогащенные в группах генов со схожей степенью изменения экспрессии (фолд-специфичная ГО категория, ФГО). Данный метод заключается в разбиении списка дифференциально экспрессирующихся генов по степени изменения экспрессии на квантили, генерации всех объединений соседних квантилей и статистического теста для проверки гипотезы об обогащении определенных характеристик генов в квантилях по сравнению со всеми дифференциально экспрессирующимися генами. Оценка надежности метода FSEA показала долю ложноположительных ниже пяти процентов. Для организации публичного доступа к разработанному методу, он был реализован в виде пакета программ FoldGO для языка программирования R и размещен в одном из крупнейших репозиториев биологического программного обеспечения Bioconductor (https://bioconductor.org/packages/release/bioc/html/FoldGO.html), а также реализован в виде веб-сервиса (https://webfsgor.sysbio.cytogen.ru/). Детальное описание метода FSEA и его реализаций представлено в главах 2.2, 3.1, 3.2, 3.3.

Апробация метода.

Метод FSEA был апробирован на данных 30 транскриптомных экспериментов находящихся в свободном доступе в базе данных GEO (Gene Expression Omnibus). В результате апробации фолд-специфичные категории ГО были выявлены во всех исследованных экспериментах, что указывает на универсальность фолд-специфичного транскрипционного ответа (Рис. 1).

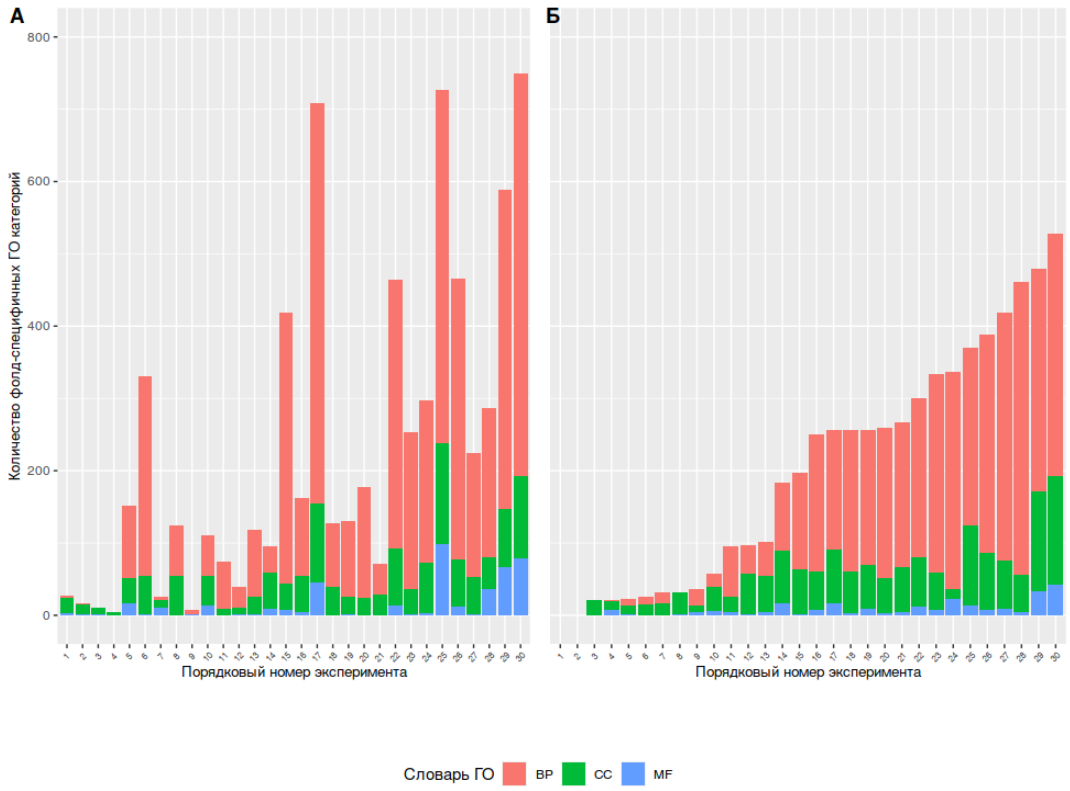

Рисунок 1. Количество фолд-специфичных терминов, обнаруженных при анализе экспериментов, перечисленных в таблице 2 для ДЭГ, экспрессия которых подавляется (**А**) и активируется (**Б**). По оси y отмечено количество фолд-специфичных ГО категорий. По оси x отмечены порядковые номера экспериментов, соответствующие номерам из таблицы 2 диссертации. Цветами отмечены словари ГО: красным - биологические процессы (BP; Biological Process), зеленым - клеточные компоненты (CC; Cellular Components), синим - молекулярные функции (MF; Molecular Function).

Также мы обнаружили, что фолд-специфичный ответ, как правило, характерен для групп ДЭГ с низкой и очень высокой степенью изменения экспрессии (Рис. 2).

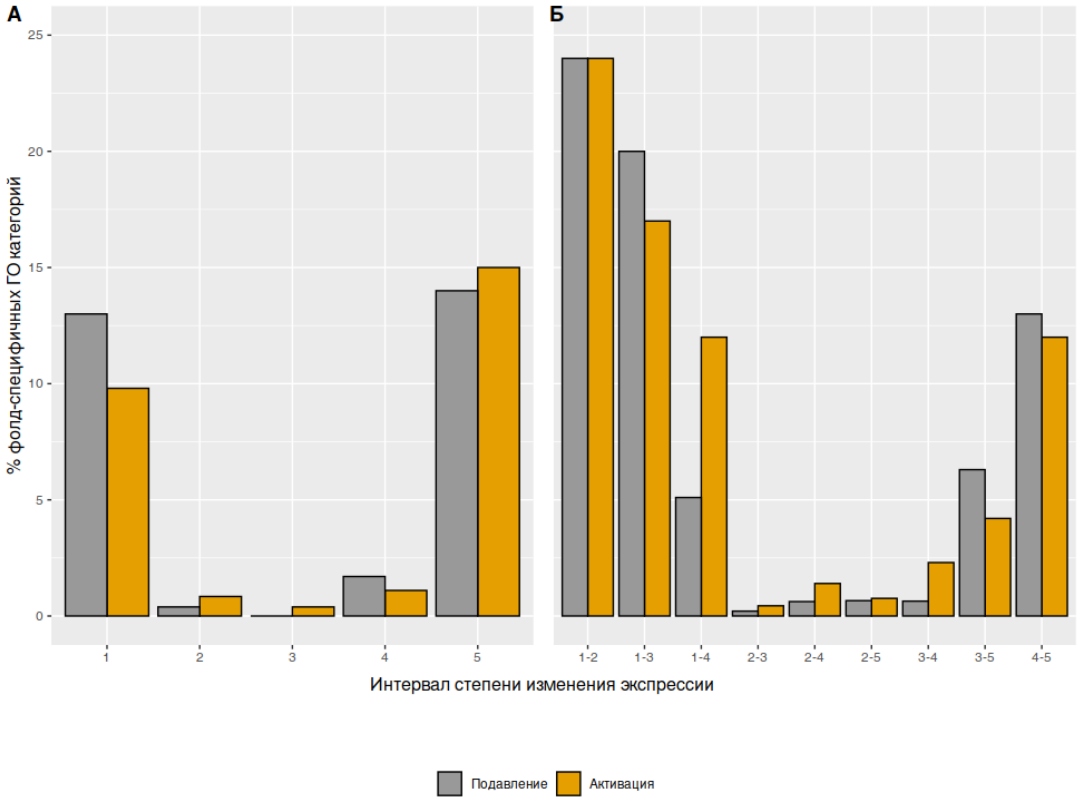

Рисунок 2. Процент фолд-специфичных категорий ГО в одиночных (**А**) и объединенных (**Б**) интервалах степени изменения экспрессии среди всех фолд-специфичных категорий ГО, обнаруженных в экспериментах, перечисленных в таблице 2 для ДЭГ, экспрессия которых подавляется (серые столбцы) и активируется (оранжевые столбцы).

Стоит отметить, что в большинстве случаев исследователи уделяют больше внимания генам с наибольшей степенью изменения экспрессии и процессам, в которых они участвуют. Однако ранее была показана важность процессов ассоциированных с низкой степенью изменения экспрессии (St. Laurent et al., 2013), но их регуляция на уровне транскрипции не была детально изучена. Метод FSEA является надежным инструментом для выявления таких процессов и предоставляет всю необходимую информацию для их дальнейшего исследования. Подробно результаты апробации метода FSEA рассмотрены в главе 3.3.4.

Анализ данных по 6-часовой обработке ауксином корней арабидопсиса

В главе 3.4 представлен детальный анализ результатов апробации метода FSEA на данных эксперимента по обработке ауксином корней *Arabidopsis thaliana.* В результате которой было выявлено более 100 категорий ГО ассоциированных с фолд-специфичным изменением экспрессии. Детальный анализ выявленных ГО категорий показал, что механизмы действия ауксина на транскриптом очень сбалансированы. Можно сделать вывод, что обработка ауксином позволяет сохранять жизненно важные ресурсы в период активного роста и развития тканей корня. Результаты апробации метода FSEA на данных RNA-seq эксперимента по обработке ауксином корней *Arabidopsis thaliana* опубликованы в журнале Nature Scientific Reports (Omelyanchuk et al., 2017).

Апробация метода на данных эксперимента по экспрессии генов в клеточной линии рака предстательной железы (LNCaP).

В главе 3.5 представлен детальный анализ апробации метода FSEA на данных экспериментов по исследованию экспрессии генов в клеточной линии аденокарциномы предстательной железы (LNCaP) с экспрессией гена андрогенового рецептора AR-V7 (GSE71334) (Cottard et al., 2017) и по сравнению с нормальными клетками эпителия предстательной железы (клеточная линия HPrEC) (GSE70466). Для обоих экспериментов было обнаружено большое количество фолд-специфичных категорий ГО связанных с процессами канцерогенеза, многие из которых не были обнаружены классическими методами функционального обогащения, такими как SEA (Huang et al., 2009) и GSEA (Subramanian et al., 2005) (Рис. 3).

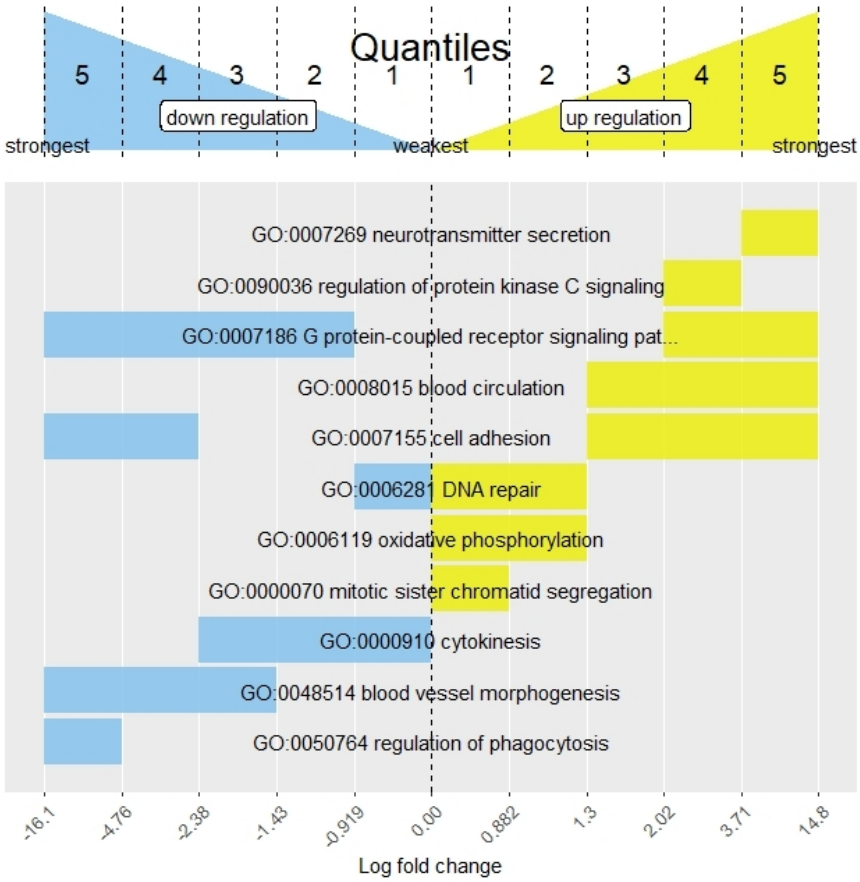

Рисунок 3. Выходная гистограмма веб-сервиса FoldGO (https://webfsgor.sysbio.cytogen.ru), отображающая категории ГО, ассоциированные с процессами канцерогенеза. Данные категории ГО, выявленные только при помощи метода FSEA, показали наибольшую статистическую значимость обогащения в определенных интервалах степени изменения экспрессии.

Применение метода FSEA к различным характеристикам генов

В главе 3.6 представлены результаты применения метода FSEA для анализа фолд-специфичного обогащения ДНК-мотивами и наборами онкогенов в данных по исследованию экспрессии сплайс-варианта *AR-V7* в клеточной линии LNCaP (GSE71334) (Cottard et al., 2017). В результате метод FSEA обнаружил ассоциацию с фолд-специфичной экспрессией для нескольких сайтов связывания транскрипционных факторов связанных с развитием рака предстательной железы по андроген-независимому пути и трансактивацией андрогенового рецептора.

ВЫВОДЫ

1) Разработан биоинформатический метод FSEA для анализа представленности категорий Генной Онтологии (ГО) в наборах генов со схожей степенью изменения экспрессии, выявленной в транскриптомных экспериментах. Метод FSEA позволяет оценить силу транскрипционного ответа в группах генов объединенных общей функцией.
2) Метод FSEA реализован в виде пакета программ FoldGO на языке R и размещен в открытом доступе в репозитории Bioconductor и в виде веб-сервиса.
3) Применение метода FSEA на транскриптомах животных и растений, находящихся в открытом доступе, показало наличие большого числа фолд-специфичных категорий генов в каждом из тестируемых транскриптомов, при условии значительного числа (>200) дифференциально-экспрессирующихся генов.
4) Массовый анализ функционального обогащения транскриптомов методом FSEA выявил три типа транскрипционного ответа у функционально-связанных групп генов: (1) не скоординированный по степени изменения экспрессии; (2) скоординированные слабые изменения; (3) скоординированные сильные изменения экспрессии.
5) Помимо хорошо изученного транскрипционного ответа на ауксин с высокой степенью изменения экспрессии генов у *Arabidopsis thaliana*, метод FSEA выявил ряд функциональных характеристик генов, которые на полногеномном уровне регулируются ауксином с низкой или промежуточной степенями изменения экспрессии.
6) Анализ транскриптомов клеточных линий рака предстательной железы человека с помощью FSEA выявил фолд-специфичные категории ГО, связанные с процессом онкогенеза, которые не могут выявить методы GSEA и SEA.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

По теме диссертации было опубликовано 11 научных работ, из них три статьи в зарубежных журналах из списка ВАК, восемь тезисов конференций, на разработанный пакет программ FoldGO получено авторское свидетельство.

Статьи в журналах

1) **Wiebe, D.S.**, Omelyanchuk, N.A., Mukhin, A.M., Grosse, I., Lashin, S.A., Zemlyanskaya, E.V., Mironova, V.V. Fold-Change-Specific Enrichment Analysis (FSEA): Quantification of Transcriptional Response Magnitude for Functional Gene Groups // Genes - 2020 г. - Т.11 - N 4 - C434. doi: 10.3390/genes11040434
2) Omelyanchuk N.A. #, **Wiebe D.S.** #, Novikova D.D., Levitsky V.G., Klimova N., Gorelova V., Weinholdt C., Vasiliev GV., Zemlyanskaya EV., Kolchanov N.A., Kochetov A.V., Grosse I., Mironova V.V. Auxin regulates functional gene groups in a fold-specific manner in Arabidopsis root // Nat Sci Rep – 2017 г. - Т. 7 - N 1 - C.2489. doi:10.1038/s41598-017-02476-8, # - equal contribution
3) Zemlyanskaya, E.V. #, **Wiebe, D.S.** #, Omelyanchuk, N.A., Levitsky, V.G., Mironova, V.V. Meta-analysis of transcriptome data identified TGTCNN motif variants associated with the response to plant hormone auxin in Arabidopsis thaliana L. // J Bioinform Comput Biol - 2016 г. - N 14(2). doi: 10.1142/S0219720016410092, # - equal contribution

Авторские свидетельства

1) **Вибе Д.С.**, Омельянчук Н.А., Миронова В.В. Функциональная аннотация дифференциально экспрессирующихся генов с учетом степени изменения экспрессии (FoldGO). Свидетельство о государственной регистрации программы для ЭВМ №2018665628

Тезисы материалов конференций

1) **Вибе Д.С.**, Мухин А.М., Омельянчук Н.А., Миронова В.В. FoldGO - программный комплекс для выявления фолд-специфичных ГО категорий в данных транскриптомных экспериментов. “Симпозиум Биофизика сложных систем. Вычислительная и системная биология. Молекулярное моделирование” 27 января – 1 февраля, 2020, Дубна, Россия
2) Nadya Omelyanchuk, **Daniil Wiebe**, Victoria Mironova. FoldGO: a web server to identify functional gene groups responding to a factor within specific ranges of fold changes. 30th International Conference on Arabidopsis Research (ICAR2019). June 16-21, 2019, Wuhan, China
3) **Вибе Д.С.**, Омельянчук Н.А., Мухин А.М., Лашин С.А., Миронова В.В. FoldGO - новый метод анализа функционального обогащения с учетом степени изменения транскрипционной активности. Сборник тезисов, 7ой съезд ВОГиС, 18 - 22 июня, 2019 Санкт-Петербург, Россия
4) **D.S. Wiebe**, N.A. Omelyanchuk, V.V. Mironova. FoldGO - the new method for functional enrichment analysis of transcriptome data to identify fold-change-specific GO categories. 17th european conference on computational biology (ECCB 2018), 8 – 12 September, 2018, Athens, Greece
5) **D.S. Wiebe**, A.M. Mukhin, N.A. Omelyanchuk, V.V. Mironova. FoldGO for functional annotation of transcriptome data to identify fold-change-specific GO categories. Mathematical Modeling and High Performance Computing in Bioinformatics, Biomedicine and Biotechnology, August 21-24, 2018, Novosibirsk, Russia
6) A.M. Mukhin, **D.S. Wiebe**, I. Grosse, S.A. Lashin, V.V. Mironova Developing FoldGO, the tools for multifactorial functional enrichment analysis. Mathematical Modeling and High Performance Computing in Bioinformatics, Biomedicine and Biotechnology, August 21-24 2018, Novosibirsk, Russia

7) Омельянчук Н.А., **Вибе Д.С.**, Миронова В.В. Auxin induced expression changes differ among functional gene groups. Сборник тезисов, Высокопроизводительное секвенирование в геномике, 18.07.2017 - 23.07.2017, Новосибирск, Россия
8) Миронова В. В., **Вибе Д. С**., Омельянчук Н.А. Auxin coordinates transcriptional fold changes for the genes belonging to particular functional groups Сборник тезисов, Конгресс биотехнология: состояние и перспективы развития, 20.02.2017 - 22.02.2017, Москва, Россия