ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ НАУЧНОЕ УЧРЕЖДЕНИЕ «ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИНСТИТУТ ЦИТОЛОГИИ И ГЕНЕТИКИ СИБИРСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК»

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

Специальность 1.5.8
Математическая биология, биоинформатика

Диссертация на соискание учёной степени
Кандидата биологических наук

Научный руководитель:
к.б.н. Миронова Виктория Владимировна

Новосибирск
2021

Список сокращений

ГО - Генная Онтология (GO - Gene Ontology)

ФГО - Фолд-специфичная ГО категория

ДЭГ - дифференциально экспрессирующиеся гены

РПЖ - рак предстательной железы

ИУК – индолилуксусная кислота

SEA - Singular Enrichment Analysis (Анализ уникального обогащения)

GSEA - Gene Set Enrichment Analysis (Анализ обогащения набора генов)

KEGG - Kyoto Encyclopedia of Genes and Genomes (Киотская энциклопедия генов и геномов)

RPKM - Reads Per Kilobase per Million mapped reads (количества прочтений для конкретного экзона на длину гена на миллион картированных прочтений)
FPKM - Fragments Per Kilobase per Million mapped reads (количества фрагментов для конкретного экзона на длину гена на миллион картированных прочтений)
RLE - Relative Log Expression (относительный логарифм экспрессии)
TMM - Trimmed Mean of M values (усеченное среднее значений M)
TPM - Transcripts Per Kilobase per Million mapped reads (количества транскриптов на длину гена на миллион картированных прочтений)
ANOVA - Analysis of variance (Дисперсионный анализ)
FDR - False Discovery Rate (уровень ложноположительных результатов)
FWER - Family Wise Error Rate (групповая вероятность ошибки)
RNA-Seq - секвенирование РНК
GAF - GO Annotation File (Файл аннотации Генной Онтологии)
OBO - Open Biomedical Ontologies (Открытые биомедицинские онтологии)
FSEA - Fold-change Specific Enrichment Analysis (Анализ фолд-специфичного обогащения)

GEO - Gene Expression Omnibus (Сборник данных по экспрессии генов)

GPCR - G Protein-Coupled Receptors (Рецепторы, сопряженные с G-белком)

WGCNA - Weighted Gene Co-expression Network Analysis (Анализ взвешенной сети коэкспрессии генов)

Введение

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 ответственных за формирование хозяйственно-ценных признаков (Ashburner et al., 2000), или генов, участвующих в исследуемых биологических процессах, например, развитии ишемической болезни сердца у человека или устойчивости

пшеницы к засоленности почвы (Balashanmugam et al., 2019; 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 в 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 анализа 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 методы 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 (100 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.

Публикации 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. - T.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. - T. 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

Тезисы 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
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, Novosibirsk, Russia, August 21-24, *2018*
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, Novosibirsk, Russia, August 21-24, *2018*
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
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

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 государственное регистрации базы данных №2018665628

1. Обзор литературы

1.1. Транскриптомные эксперименты в исследовании экспрессии генов

Совокупность всех мРНК и некодирующих РНК в клетке называется транскриптомом. В отличие от геномной ДНК транскриптом клетки сильно меняется под действием внешних и внутренних факторов, таких как неблагоприятные условия среды, обработка биологически-активными соединениями, заболевания, циркадные ритмы и прочее. Оценка количественного и качественного состава транскриптома позволяет изучать, как внешние и внутренние факторы влияют на экспрессию генов на молекулярном уровне. Как правило, такой анализ заключается в выявлении генов, статистически значимо изменивших свою экспрессию в ответ на экспериментальные условия (дифференциально экспрессирующиеся гены, ДЭГ). Также исследование транскриптома позволяет изучать процессы регуляции экспрессии генов и различия в экспрессии в разных тканях (Lowe et al., 2017).

На данный момент основными методами для исследования транскриптома являются экспрессионные ДНК-микрочипы (Nelson, 2001) и секвенирование РНК (RNA-seq) (Wang et al., 2009). Экспрессионный ДНК-микрочип представляет собой подложку (матрицу), на которой в определенном порядке расположены пробы-олигонуклеотиды (Zhang et al., 2009). Каждая “ячейка” матрицы содержит несколько пикомолей заякоренной в подложке ДНК с определенной последовательностью нуклеотидов, которая специфически комплементарна участку определённой мРНК. Затем данную матрицу используют для гибридизации с кДНК, полученной путем обратной транскрипции мРНК, меченной флуоресцентными красителями. Гибридизация регистрируется по активности флуоресценции в ячейках микрочипа. В итоге можно получить данные о

содержании в образце мРНК, для которых в микрочипе имеются специфичные пробы. Данный метод основывается на предположении, что интенсивность флуоресценции при гибридизации отражает уровень экспрессии определенного гена. Исследование экспрессии генов при помощи микрочипов является сравнительно недорогим методом с низкими трудозатратами (Heller, 2002), однако он имеет серьезное ограничение, а именно данный подход предполагает наличие предварительного знания о последовательностях генов и транскриптов. На данный момент существуют готовые микрочипы для модельных и множества экономически значимых организмов.

Секвенирование транскриптома (RNA-Seq) является комбинацией технологии высокопроизводительного секвенирования и вычислительных методов для количественной и качественной оценки состава транскриптома в исследуемом образце (Ozsolak и Milos, 2011). Преимуществами данного подхода по сравнению с использованием ДНК-микрочипов являются: 1) более точное определение уровней экспрессии генов, 2) возможность детектировать экспрессию у большего числа генов, 3) возможность определять уровни экспрессии у разных изоформ гена при изучении альтернативного сплайсинга (Agrawal et al., 2014). Так же, одной из важных особенностей секвенирования транскриптома является возможность анализа транскриптов организмов без референсного генома (*de novo* сборка транскриптома) (Wang et al., 2009). Стоит также отметить, что для секвенирования транскриптома требуется меньшее количество РНК (нанограммы) чем для исследования на ДНК-микрочипах (микрограммы), что позволяет изучать экспрессию генов в отдельных клетках (Hashimshony et al., 2012).

Рассмотрим более подробно процедуру биоинформатического анализа данных RNA-Seq, которую можно подразделить на несколько этапов: контроль качества, картирование, квантификация, нормировка.

1.1.1. Оценка качества прочтений

Основным форматом представления “сырых” данных секвенирования, является файл формата *fastq*. Он представляет собой текстовый документ, в котором каждому прочтению соответствует строки с идентификатором гена, последовательностью нуклеотидов и оценкой качества прочтения. Одним из наиболее популярных инструментов для контроля качества данных секвенирования представленных в формате *.fastq* является программа FastQC (Andrews, 2010), которая позволяет оценивать такие характеристики, как:

1) качество прочтений (в разрешении оснований и прочтений);
2) нуклеотидный состав последовательностей;
3) GC состав последовательностей;
4) количество позиций в прочтениях, которым не присвоен определенный нуклеотид;
5) распределений длин прочтений;
6) уровень дуплицированности в библиотеке прочтений;
7) перепредставленность последовательностей;
8) содержание адаптерных последовательностей;
9) состав *k*-меров.

Основной информацией, которую можно извлечь из данных характеристик, является общая оценка качества секвенирования и перепредставленность различных последовательностей, например, адаптеров и векторов.

1.1.2. Картирование

После анализа прочтений и удаления некачественных данных производится картирование прочтений на референсный геном или транскриптом. Задача картирования заключается в нахождении уникальной позиции в геноме для каждого прочтения. На первоначальном этапе используются быстрые эвристические алгоритмы поиска, использующие, как правило, хэш-таблицы (Chen et al., 2009) или преобразование Барроуза-

Уилера (Li и Durbin, 2009). Затем, когда для прочтений составлен набор кандидатных позиций, применяют более точные и ресурсоемкие алгоритмы выравнивания. Иногда одному прочтению могут соответствовать сразу несколько позиций в референсном геноме. В таком случае могут быть применены несколько подходов: исключение прочтений из анализа (Langmead et al., 2009), выбор позиции случайным образом (Li et al., 2008), выбор позиции по оценке локального покрытия (Cloonan et al., 2008). Кроме того, для избежания проблем, связанных с избыточностью прочтений, картируемых сразу на несколько позиций в референсном геноме, используют прочтения, секвенированные с двух концов одной молекулы мРНК, или двусторонние прочтения (paired-end reads). Важной особенностью картирования РНК на референсный геном является необходимость использовать информацию о сайтах сплайсинга, так как источником материала для секвенирования является РНК, и прочтения, пересекающие границы экзонов, не могут быть картированы надлежащим образом (Sultan et al., 2008).

1.1.3. Подсчет картированных прочтений (квантификация)

Как только для каждого прочтения найдена уникальная позиция в референсном геноме, для каждого гена производят подсчет прочтений картированных на их последовательности. Однако значительная часть прочтений может быть картирована вне границ известных экзонов. Такие прочтения могут соответствовать экзонам, специфичным для определенных типов клеток, или ранее неизвестным экзонам (Pickrell et al., 2010). Одним из решений данной проблемы является включение в анализ прочтений, соответствующих интронам. Однако такой подход зачастую не позволяет различать перекрывающиеся изоформы транскриптов.

1.1.4. Нормировка данных

Для сравнения экспрессии между генами внутри одного образца или между образцами необходимо проводить нормировку. Так как при сравнении экспрессии между образцами, как правило, сравнивают экспрессию одного и того же гена, необходимость в нормализации на длину гена отпадает. Однако разные образцы могут быть секвенированы с разным покрытием и разной глубиной (количество прочтений, уникально картированных на участок референсного генома), поэтому проводят нормализацию на размер библиотеки для каждого образца (Robinson и Smyth, 2007). Для сравнения экспрессии внутри одного образца проводят нормировку на длину гена, так как чем больше длина гена, тем больше прочтений будут картированы на него (Mortazavi et al., 2008). Стандартной процедурой нормировки внутри одного образца является расчет количества прочтений для односторонних прочтений (single-end) или фрагментов для двусторонних прочтений (paired-end) для конкретного гена на миллион картированных прочтений (RPKM, Reads Per Kilobase per Million mapped reads; FPKM, Fragments Per Kilobase per Million mapped reads). Значение RPKM вычисляется путем деления исходного количества прочтений для каждого гена на общее количество прочтений в образце в миллионах и дальнейшего деления полученного значения на длину данного гена в тысячах пар нуклеотидов (п.н.). Однако значения RPKM не позволяют сравнивать пропорции генов между образцами, так как суммарное значение RPKM для всех генов в разных образцах будет отличаться. Решить данную проблему позволяет расчет количества транскриптов на миллион картированных прочтений (TPM, Transcripts Per Kilobase per Million mapped reads). Расчет значения TPM отличается от RPKM порядком операций, сначала производится нормировка на длину гена с получением значения RPK (Reads Per Kilobase), затем значение RPK для каждого гена делится на суммарное значение RPK для каждого образца в миллионах. Обе величины RPKM и TPM учитывают длину гена и размер библиотеки, однако при их расчете не учитывается влияние генов с высоким уровнем изменения

экспрессии или генов, экспрессирующихся только в одной группе образцов. При приблизительно одинаковой глубине секвенирования, такие гены, при нормировании на размер библиотеки, будут размывать вклад остальных генов, с менее выраженной экспрессией в образце. На этом фоне, в остальных образцах может быть обнаружена ложно-положительная дифференциальная экспрессия данных генов. Для учета таких факторов используют нормировку с использованием коэффициентов масштабирования, наиболее популярными методами расчета которых являются RLE (Relative Log Expression) (Love et al., 2014) и TMM (Trimmed Mean of M values) (Robinson et al., 2010), реализованные в пакетах DESeq2 и EdgeR, соответственно. Коэффициент RLE представляет из себя медианное значение выборки, состоящей из количеств прочтений для каждого образца в отдельности, деленных на среднее геометрическое значение количеств прочтений для каждого гена и всех образцов, и рассчитывается по следующей формуле:

$s_j = median_j\left(\frac{k_{i,j}}{\prod_{v=1}^{m} k_{i,v}^{\frac{1}{m}}}\right)$, где i - индекс определенного гена, j - индекс определенного образца, k - количество прочтений, m - общее количество образцов. Для получения нормированных значений, количество прочтений для каждого гена делят на полученный коэффициент RLE. Коэффициент TMM рассчитывается для каждого образца относительно заранее выбранного референсного образца как взвешенное среднее после удаления низко и высоко экспрессируемых генов, которые определяются исходя из значений степени изменения экспрессии $log_2 \frac{Y_{gk}/N_k}{Y_{gk\prime}/N_{k\prime}}$ и абсолютной экспрессии $\frac{1}{2} log_2(Y_{gk}/N_k \cdot Y_{gk\prime}/N_{k\prime})$, где Y_{gk} и $Y_{gk\prime}$ - количества прочтений для гена g в образцах k и k', соответственно, а N_kи $N_{k'}$ - размер библиотеки прочтений для образцов k и k'. Логарифм взвешенного среднего для образца k по сравнению с образцом r после удаления части данных выглядит как $log_2 TMM_k^{(r)} = \frac{\sum_{g\in G} w_{gk}^r m_{gk}^r}{\sum_{g\in G} w_{gk}^r}$, где $m_{gk}^r = \frac{log_2(Y_{gk}/N_k)}{log_2(Y_{gr}/N_r)}$ и $w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}}$.

Нормировка данных при помощи коэффициента RLE позволяет проводить сравнения между образцами, а так же выявлять дифференциальную экспрессию (Love et al., 2014), в то время как TMM нормировка, помимо вышеперечисленного, позволяет также производить сравнение экспрессии генов в пределах одного образца (Robinson et al., 2010).

1.2. Анализ дифференциальной экспрессии генов

Обработанные данные транскриптомных экпериментов обычно представляют собой таблицы, в которых столбцы соответствуют исследуемым образцам и их репликам, а строки соответствуют пробам в случае ДНК микрочипов или идентификаторам генов из аннотации для исследуемого организма в случае RNA-seq данных. В каждой ячейке таблицы содержится значение, соответствующее количеству прочтений, выровненных на соответствующий ген в случае секвенирования транскриптома или соответствующее интенсивности гибридизации в случае использования ДНК-микрочипов. Наиболее часто данные об экспрессии генов из таких таблиц используются для поиска дифференциально экспрессирующихся генов при помощи различных статистических тестов (*t*-тест, ANOVA, метод Байеса, тест Манна-Уитни), которые учитывают вариацию между экспериментальными репликами и их количество, минимизируя ошибки первого и второго рода (Wei et al., 2004). Выбор статистического теста зависит от дизайна эксперимента и от особенностей используемых данных. Например, при сравнении экспрессии двух образцов, как правило, используют *t*-тест, при одновременном сравнении большего количества образцов применяют метод анализа вариации ANOVA. Однако вышеупомянутые статистические тесты применяются при предположении о нормальности распределения случайной величины, которое далеко не всегда выполняется для транскриптомных экспериментов. В таком случае

рекомендуется применять непараметрические тесты такие как, например, тест Манна-Уитни (Troyanskaya et al., 2002). Однако в то время как значения интенсивности гибридизации, полученные на ДНК-микрочипах, представлены непрерывной переменной, значения количества прочтений полученные из RNA-seq экспериментов представлены дискретной величиной, которую обычно аппроксимируют распределением Пуассона. Учитывая дискретный характер RNA-seq данных, самым простым подходом к выявлению ДЭГ является точный тест Фишера (Fisher Exact Test), при расчете которого сравниваются доли количества прочтений для каждого гена в двух образцах, таким образом обеспечивая нормирование на глубину секвенирования в каждом образце. Ввиду того, что точный тест Фишера не учитывает биологическую вариабельность экспрессии при анализе с использованием биологических реплик, которые необходимы для надежного выявления дифференциальной экспрессии, используют тест отношения правдоподобия (Likelihood Ratio test; LR) (Marioni et al., 2008). Одним из основных требований к использованию распределения Пуассона для оценки статистической значимости при выявлении ДЭГ является равенство математического ожидания и дисперсии, однако в данных некоторых RNA-Seq экспериментов было обнаружено значительное различие между дисперсией и математическим ожиданием, источником которого могла послужить, например, гетерогенность анализируемой клеточной популяции или сильная вариабельность среди биологических реплик (Auer и Doerge, 2011; Robinson и Smyth, 2007). При таких особенностях данных рекомендуется использовать для аппроксимации отрицательное биномиальное (Robinson и Smyth, 2008), бета-биномиальное и квази-биномиальное распределения (Auer и Doerge, 2011).

Наиболее популярными инструментами для анализа дифференциальной экспрессии являются пакеты, реализованные для языка программирования R. Пакет *limma* (Ritchie et al., 2015) используется для анализа ДНК-микрочип данных, а *DESeq2* (Love et al., 2014) и *edgeR*

(Robinson et al., 2010) - для анализа RNA-seq данных. Во всех перечисленных инструментах реализованы процедуры нормировки данных и различные статистические тесты, что позволяет применять один и тот же инструмент при разных дизайнах эксперимента (количество биологических реплик, количество исследуемых фенотипов и условий и.т.д.). Также существуют инструменты, которые могут применяться при специфических дизайнах экспериментов, таких как, например, отсутствие биологических или технических реплик. В качестве примера такого инструмента можно привести NOISeq (Tarazona et al., 2015), который позволяет симулировать технические реплики, однако стоит с осторожностью интерпретировать результаты такого анализа (Zaim et al., 2019).

В результате процедуры выявления ДЭГ исследователь может лишь проверить гипотезу об ассоциации изменения экспрессии генов и экспериментальных условий, что дает довольно поверхностную информацию об изучаемом биологическом процессе. На данный момент существует множество методов для дальнейшего анализа ДЭГ, которые позволяют выявлять группы генов, объединенных одной функцией, и различные взаимосвязи между генами. Рассмотрим наиболее распространенные подходы для анализа дифференциально экспрессирующихся генов.

1.2.1. Генная Онтология

Генная Онтология (ГО) представляет собой структурированный словарь биологических терминов, который подразделяется на три непересекающихся подсловаря для описания биологических процессов, молекулярных функций и клеточных компонент (Ashburner et al., 2000). Помимо общего словаря существуют специализированные словари для различных групп организмов. По структуре ГО представляет собой направленный ациклический граф, в котором вершинами являются категории ГО, а ребра представляют отношения между категориями, причем вершины высокого уровня представляют более общее описание процессов, например,

три “корневые” вершины графа ГО описывают все биологические процессы, молекулярные функции и клеточные компоненты, а вершины более низких уровней описывают более специфичные процессы и компоненты. В ГО выделяют четыре основных типа отношений: *is a*, *part of*, *has part* и *regulates*. Основным отношением определяющим структуру ГО является отношение *is a*, если оно определено между двумя категориями ГО, то можно сказать что одна категория описывает сущность являющуюся подтипом второй категории. Отношения *part of* и *has part* обозначают ситуацию в которой одна категория описывает сущность являющуюся частью второй категории. Например категория ГО описывающая митохондрию (GO:0005739) имеет отношение *part of* к категории ГО описывающей цитоплазму (GO:0005737) но между ними нет отношения *is a* так как митохондрии являются частью цитоплазмы, но не являются ее более специфичным подтипом. Отношение *regulates* описывает ситуацию, когда сущность описываемая одной категорией ГО напрямую влияет на проявление сущности описываемой другой категории. Также стоит отметить что отношения типа *part of, has part* и *regulates* могут существовать между категориями из разных подсловарей, например между категорией описывающей молекулярную функцию и биологический процесс, в то время как отношения типа *is a* могут существовать только между категориями внутри одного словаря (Gene Ontology Consortium, 2010). Такая структура словаря и отношений позволяет описать множество сложных взаимодействий происходящих в биологических системах в удобном для компьютерной обработки виде.

Для применения ГО к данным геномных экспериментов используются аннотации в которых описывается связь между генами и категориями ГО. Существует множество таких аннотаций для различных организмов, в которых для каждой аннотации гена к категории ГО указывается тип источника информации на основании которой данная аннотация была сделана. Всего в структуре ГО существует шесть основных типов источников информации (Gene Ontology Consortium, 2013):

1. Экспериментальное свидетельство - связь гена и сущности описываемой ГО категорией подтверждена в эксперименте;
2. Филогенетическое свидетельство - связь гена и сущности описываемой ГО категорией подтверждена на основании филогенетического анализа;
3. Свидетельство полученное на основе компьютерного анализа - связь гена и сущности описываемой ГО категорией подтверждена на основании *in silico* анализа ранее опубликованных данных;
4. Утверждение автора - аннотация гена к сущности описываемой ГО категорией сделана на основании опубликованного в статье утверждения;
5. Утверждение куратора - аннотация гена к сущности описываемой ГО категорией сделана на основании утверждения куратора, в том случае если данная аннотация не попадает в остальные типы источников;
6. “Электронное” свидетельство - связь гена и сущности описываемой ГО категорией подтверждена на основании *in silico* анализа без возможности отследить источник исходных экспериментальных данных.

Словари ГО хранятся и поддерживаются на базе ресурса Gene Ontology (http://geneontology.org) (Ashburner et al., 2000), который также предоставляет аннотации для более чем 20 организмов, включая модельные, такие как *Mus musculus*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Danio rerio*. Существует несколько распространенных вариантов использования Генной Онтологии. Самый простой вариант это использование аннотации и онтологии для исследования функций неизвестных или мало-изученных генов на основе схожести последовательностей с генами или белками, чья функция известна (Horan et al., 2008), или на основе данных о их взаимодействии, например белок-белковом (Kuppuswamy et al., 2014).

1.2.2. Анализ уникального обогащения (SEA)

Одним из первых подходов к анализу ДЭГ для выявления функциональных групп является анализ уникального обогащения списка ДЭГ категориями Генной Онтологии (ГО) (Ashburner et al., 2000).

На данный момент существует множество инструментов, представленных в виде компьютерных программ и веб-сервисов, позволяющих проводить такой анализ. В качестве примеров можно привести такие веб-сервисы, как DAVID (Huang et al., 2007), AgriGO (Tian et al., 2017), PANTHER (Mi et al., 2019), плагин BinGO (Maere et al., 2005) для программной платформы Cytoscape (Shannon, 2003), и множество пакетов для различных языков программирования: topGO (Alexa и Rahnenfuhrer, 2019) и GOstats (Falcon и Gentleman, 2007) для языка R, GOATOOLS (Klopfenstein et al., 2018) для Python и.т.д.

В основе всех вышеперечисленных программных продуктов лежит один и тот же алгоритм. Для каждой категории ГО рассчитывается обогащение в списке генов, представляющих интерес, например, ДЭГ, по сравнению со всеми генами, рассматриваемыми в исследовании. Обогащение рассчитывается одним из статистических методов по таблице сопряженности (Таблица 1), наиболее часто используется точный тест Фишера (Fisher's exact test), который позволяет оценить значимость взаимосвязи между двумя переменными. Вероятность наблюдать определенные значения в таблице сопряженности рассчитывается по формуле:

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$, где a, b, c, d, n - значения из Таблицы 1.

Таблица 1. Таблица сопряженности для расчета обогащения категориями ГО в исследуемом списке генов.

	Кол-во генов в исследуемом списке генов	**Кол-во генов вне исследуемого списка генов**	**Общее кол-во генов**
Кол-во генов, ассоциированных с категорией ГО	a	b	a+b
Кол-во генов, не ассоциированных с категорией ГО	c	d	c+d
Общее кол-во генов	a+c	b+d	n

Так как ГО имеет структуру ациклического графа, сначала оценивается обогащение для категорий, являющихся конечными узлами графа, то есть наиболее специфичных категорий ГО. Далее рассматриваются более общие категории, в аннотацию которых добавляются гены, аннотированные к более специфичным категориям, с которыми они имеют отношение предок - потомок (parent - child; True Path Rule) (Ashburner et al., 2000). ГО категории, прошедшие порог на множественное тестирование (False Discovery Rate (FDR), Family Wise Error Rate (FWER)), считаются обогащенными в исследуемом списке генов. Такой анализ позволяет выявить функциональные группы генов, ассоциированные с исследуемым процессом или признаком (Grossmann et al., 2007).

Часто, как результат, исследователь получает избыточный список ГО категорий, обогащенных в исследуемой группе генов. Анализ этого списка может быть трудоемким процессом. Для извлечения биологически значимого результата из анализа функционального обогащения разработаны различные методы снижения информационной избыточности, такие как

алгоритмы parent-child, elim, weight (Grossmann et al., 2007), и различные методы кластеризации. Рассмотрим более подробно метод кластеризации генов для разделения их на функциональные группы, реализованный в веб-сервисе DAVID (Huang et al., 2007). Данный метод основан на расчете каппа-статистики K_{mn} для генов m и n по формуле:

$K_{mn} = \frac{O_{mn} - A_{mn}}{1 - A_{mn}}$, где O_{mn} - наблюдаемая сопредставленность генов в аннотациях, взятых из различных баз данных (ГО, KEGG, и.т.д.), A_{mn} - сопредставленность, ожидаемая по случайным причинам, которая рассчитывается как вероятность принадлежности генов m и n одинаковому набору аннотаций. Значение K_{mn}, равное 1 означает максимальную сопредставленность генов в одной группе, а значение, равное 0 соответствует сопредставленности, которая может быть получена по случайным причинам. Данный подход авторы считают более подходящим для анализа функционального обогащения, чем классические, основанные на сходстве последовательностей или принадлежности к одному семейству белков, так как зачастую, продукты генов объединенных одной функцией не имеют таких сходств. Затем к рассчитанным значениям K_{mn} применяется алгоритм эвристического многоуровневого разбиения (heuristic fuzzy multiple-linkage partitioning). В качестве преимуществ по сравнению с другими распространенными методами кластеризации, такими как метод К-средних (Kanungo et al., 2004), авторы метода отмечают возможность включить ген сразу в несколько кластеров, динамическое определение количества кластеров, а также снижение “шума” за счет исключения из кластеров слабо ассоциированных генов. Полученные группы генов затем сортируются по значимости, которая рассчитывается на основе перепредставленности в данных группах генов, аннотированных к категориям ГО, для которых была выявлена статистически значимое обогащение в исследуемом списке генов по сравнению с геномом (Huang et al., 2007).

Перечисленные подходы позволяют выявить группы генов, объединенные одной биологической функцией. Однако информация об уровне экспрессии исследуемых генов может быть учтена только на этапе подготовки списка генов, которая, как правило, заключается в выявлении дифференциально экспрессирующихся генов.

1.2.3. Анализ обогащения набора генов (GSEA)

Альтернативным методом анализа функционального обогащения, в котором используются данные об экспрессии генов, является GSEA (Gene Set Enrichment Analysis) (Subramanian et al., 2005). Основная идея GSEA заключается в выявлении групп генов, объединенных одной характеристикой (функцией, сайтом связывания транскрипционного фактора и.т.д.) и демонстрирующих неравномерное распределение метрики используемой для ранжирования генов, которая, как правило, рассчитывается с использованием данных об экспрессии. Список генов, объединенных одной характеристикой, может быть составлен исследователем, взят из базы данных ГО, или из из других специализированных баз данных. В качестве альтернативы ГО, например, может быть использована база данных Molecular Signature Database (MSigDB) (Liberzon et al., 2011), которая содержит списки генов, ассоциированные с различными характеристиками, такими как связь с ГО терминами и различными заболеваниями, нахождением регуляторных элементов в промоторах, и другие. Важной особенностью метода GSEA является то, что он использует для анализа информацию по всем исследуемым в эксперименте генам, тогда как при анализе методом SEA предполагается предварительный отбор ДЭГ.

Рассмотрим метод GSEA более подробно. На первом этапе производится ранжирование исходного набора генов с использованием различных метрик, рассмотрим несколько из них.

Основной метрикой, которая используется для ранжирования генов и которую авторы метода предлагают использовать по умолчанию, является *Signal2Noise* (Subramanian et al., 2005). Она рассчитывается по формуле: $\frac{\mu_a-\mu_b}{\sigma_a+\sigma_b}$, где μ - среднее значение экспрессии среди реплик, σ - стандартное отклонение, a и b - исследуемые группы.

Еще один подход основан на расчете *t*-теста по формуле: $\frac{\mu_a-\mu_b}{\sqrt{\frac{\sigma_a^2}{n_a}+\frac{\sigma_b^2}{n_b}}}$, где μ - среднее значение экспрессии среди реплик, σ - стандартное отклонение, n - количество реплик, a и b - исследуемые группы.

Метод GSEA позволяет использовать сторонние подходы для расчета метрик. Например, многие исследователи в качестве метрики для ранжирования используют уже готовые значения, полученные при анализе дифференциальной экспрессии. В качестве примера можно привести следующую формулу расчета метрики: $\frac{log\,FC_n}{p_n}$, где $log\,FC_n$ - логарифм степени изменения экспрессии для n-ого гена, p_n - значение *p-value* для n-ого гена.

После расчета подходящей метрики, гены ранжируют с учетом полученных значений. Затем исследователь подбирает список генов с интересующими характеристиками. Часто такие списки составляют при помощи анализа уникального обогащения, а именно используют списки генов, которые соответствуют ГО терминам, значимо обогащенным в ДЭГ. Далее GSEA производит расчет оценки обогащения (enrichment score; ES) соответствующую взвешенной статистике Колмогорова-Смирнова (Massey, 1951), отображающей обогащение генов с интересующей характеристикой в верхней части ранжированного исследуемого списка генов. Затем при помощи пермутационного теста рассчитывается уровень значимости оценки обогащения. Обычно такой анализ проводят сразу для нескольких характеристик и, соответственно, для нескольких списков генов, поэтому на последнем этапе анализа производится расчет нормированной оценки

обогащения для каждого списка и коррекция на множественное тестирование. Как результат, GSEA позволяет выявить функциональные группы генов, значимо ассоциированные с исследуемым признаком.

Есть множество вариаций метода GSEA для работы с различными данными, и характеристиками, а также с возможностью распараллеливания вычислений. Однако данный метод подвергается критике из-за низкой чувствительности (Tripathi et al., 2013). Важно также отметить, что при использовании разных метрик на одних и тех же данных результат анализа может сильно различаться (Zyla et al., 2017).

1.2.4. Анализ взвешенной сети коэкспрессии генов (WGCNA)

Еще одним подходом к изучению дифференциально экспрессирующихся генов является анализ взвешенных сетей коэкспрессии (WGCNA) (Zhang и Horvath, 2005). Данный подход позволяет выявлять модули генов, межмодульные хабы и узлы, основываясь на схожести профиля экспрессии. Для выявления генов, принадлежащих одному модулю сети, рассчитывается мера схожести:

$$s_{i,j}^{unsigned} = \left|cor(x_i, x_j)\right|,$$

где x_i и x_j- профили экспрессии генов i и j, однако такая мера может привести как к потере важной информации, так и к ложноположительным результатам, так как не учитывает направление изменения уровня экспрессии. Для учета активации и подавления экспрессии рассчитывается “подписанная” мера схожести:

$$s_{i,j}^{signed} = \frac{1 + cor(x_i, x_j)}{2}.$$

Такая мера принимает значения в диапазоне $[0, 1]$, При отрицательной корреляции она будет равна 0 и 1 при положительной корреляции. Далее из полученных значений мер схожести для каждой пары генов составляется матрица схожести $S = \left[S_{i,j}\right]$ и рассчитывается мера связности между двумя генами путем установления “мягкого” порога: $a_{i,j} = \left(s_{i,j}\right)^{\beta}$, где β- пороговый

параметр, который по умолчанию равен 12. Высокие значения параметра превращают высокие значения схожести в высокие значения связности и устремляют низкие значения к нулю. Таким образом, мы получаем матрицу связности $A = [a_{i,j}]$. Для каждой пары генов в полученной матрице рассчитывается мера топологического перекрывания (*TOM)*, учитывающая связность двух генов и силу связей этих генов с другими генами. Далее рассчитанные значения используются как входные данные для кластеризации. Как правило, следующим этапом после выявления модулей является анализ обогащения модулей ГО терминами. Исходя из предположения о том, что коэкспрессирующиеся гены являются функционально связанными, данный подход позволяет выявлять функциональные гены, для которых нет аннотации (Singer et al., 2005). Также данный подход широко используется для выявления ранее неизвестных генов связанных с различными заболеваниями (Ala et al., 2008). Одним из самых важных ограничений данного метода является нестабильность его работы при небольшом количестве биологических повторностей (менее 15-ти) для каждого из экспериментальных условий, а для большей надежности результатов необходимо не менее 100 биологических повторностей (Liesecke et al., 2019).

1.3 Исследование ответа на ауксин в транскриптомных экспериментах

Ауксины - семейство фитогормонов, которые являются регуляторами множества процессов роста и развития растений. Наиболее распространенной формой активного ауксина является индолилуксусная кислота (ИУК; IAA), низкомолекулярное химическое соединение, производное триптофана. Несмотря на то, что ауксин представляет собой соединение с относительно простой химической структурой, он контролирует такие разнообразные и комплексные процессы как рост, деление и дифференцировка клеток (Paque и

Weijers., 2016). Ауксин неравномерно распределен в тканях растений (Ljung et al., 2005), его концентрация очень высока в молодых листьях, откуда он затем транспортируется в другие ткани, в том числе в корень. В тканях концентрация ауксина значительно варьирует в зависимости от типа клеток, например в меристеме корня, которая является нишей стволовых клеток растения, наибольшая концентрация ауксина наблюдается в так называемом покоящемся центре, вокруг которого располагаются стволовые клетки корня (Petersson et al., 2009). Формирование градиента концентрации ауксина в тканях является крайне важным для правильного развития растения и поддержания ниши стволовых клеток. Актуальной задачей является исследование влияния ауксина на экспрессию генов в различных тканях растения и на различных стадиях развития и в зависимости от концентрации.

Подробный функциональный анализ ДНК-микрочип экспериментов по исследованию активности генов в тканях корня *A. thaliana* с различным временем обработки ауксином был опубликован в работе (Lewis et al., 2013). Авторы обнаружили что после 1 часа обработки ауксином (ИУК) запускается транскрипционная активность ауксин-чувствительных генов, экспрессия которых возвращается к контрольным значениям после 24 часов обработки фитогормоном. Затем авторы провели кластеризацию выявленных ауксин-чувствительных генов для идентификации генов со схожими временными профилями экспрессии. Функциональный анализ полученных кластеров показал, что гены показавшие ранний транскрипционный ответ были ассоциированы с аннотациями Генной Онтологии, относящимися к ответу на ауксин, а, например, среди генов показавших наибольшую транскрипционную активность при длительной обработке ауксином (8-12 часов), были обогащены процессы описывающие клеточное деление и формирование боковых корней, что совпадало по времени с наблюдаемой инициацией боковых корней. Кроме того, с транскрипционным ответом на ауксин оказались ассоциированы такие процессы, как: “репликация ДНК”, “процессинг РНК”, “организация клеточной стенки”, “ответ на осмотический

стресс” и другие. Список обогащенных ГО категорий достаточно большой, и хотя он дает представление о том, что происходит в корне в ответ на ауксин, важность и вклад разных процессов остается непонятен. Можно предположить, что дополнительный учет того, с какой силой активировались гены ответственные за определенные процессы, позволит лучше понять процессы лежащие в основе сложной системы регуляции процессов формирования органов растения. Например, такие процессы как взаимодействие ауксина с другими фитогормонами, такими как цитокинины, этилен, гиббереллины и др.

Применение Генной Онтологии позволяет изучать взаимодействие различных гормонов без проведения дополнительных экспериментов. Например, в работе (Paponov et al., 2008) авторы провели детальный анализ десятка транскриптомных экспериментов по обработке ауксином различных органов растения с разными концентрациями фитогормона. В результате они показали что ауксин скоординированно действует с основными фитогормонами, подавляя или активируя экспрессию ассоциированных с ними генов. При этом гены ассоциированные с разными фитогормонами изменяли свою экспрессию с разной силой. Подробное изучение связи функциональных характеристик генов с силой их проявления позволит исследовать сложные взаимодействия влияющие на формирование многоклеточных организмов.

1.4. Исследование транскрипционной активности генов в клетках рака предстательной железы

Одной из наиболее часто используемой моделью для изучения транскриптома рака предстательной железы (РПЖ) человека является клеточная линия LNCaP, полученная из клеток метастазы аденокарциномы предстательной железы человека (Horoszewicz et al., 1983). Аденокарцинома предстательной железы является наиболее распространенных некожных онкозаболеваний среди взрослых мужчин (Siegel et al., 2016). Главным

направлением исследования РПЖ является изучение перехода к форме заболевания не поддающейся стандартной терапии (кастрационно-резистентный РПЖ; CRPC) (Saad и Hotte, 2010). Известно что возникновение агрессивной формы РПЖ связано с аффинностью андрогенового рецептора, исследованию данного процесса посвящено множество работ (Cottard et al., 2017; Gelmann, 2002; Takayama и Inoue, 2013; Lai et al., 2012).

Основной причиной возникновения кастрационно-резистентного РПЖ считаются мутации гена андрогенового рецептора, такие как нонсенс-мутация AR-Q640X (Céraline et al., 2004) или экспрессия сплайс-варианта гена андрогенового рецептора AR-V7 (Watson et al., 2010). Наличие таких вариантов гена андрогенового рецептора вызывают пролиферацию клеток рака даже при проведении гормональной терапии (Guo et al., 2009). Таким образом важной и актуальной задачей является изучение путей регуляции пролиферации клеток рака посредством сигнального пути андрогенового рецептора и поиск новых мишеней для терапии кастрационно-резистентного РПЖ. В недавно опубликованных работах при помощи анализа функционального обогащения данных транскриптомных экспериментов по исследованию экспрессии генов в РПЖ были выявлены процессы обогащенные в генах значимо изменяющих свою экспрессию в клетках рака. Например, процессы затрагивающие передачу сигнала через сигнальный путь белка p53 являющегося онкосупрессором (Song et al., 2019), и сигнальный путь G-белков (Huang et al., 2019), которые являются потенциальной мишенью для терапии РПЖ (Weng et al., 2006). Тем не менее, методы эффективной терапии кастрационно-резистентного РПЖ до сих пор не найдены, и использование данных транскриптомных экспериментов для решения данной проблемы является перспективным направлением.

1.5. Заключение по обзору литературы

Данные полногеномных экспериментов по исследованию дифференциальной экспрессии генов хранят в себе важную информацию о молекулярно-генетических механизмах функционирования объекта исследования. Для извлечения такой информации разработано множество методов, которые нацелены на поиск ассоциаций между различными характеристиками генов, такими как функциональная аннотация, профиль экспрессии или наличие сайтов связывания в промоторных участках. Тем не менее из-за большого количества характеристик (список которых постоянно расширяется благодаря развитию экспериментальных технологий), а также из-за индивидуальных особенностей объектов исследования в данной области нет стандартного протокола обработки данных. Кроме того, методология исследований все еще находится на этапе развития, поэтому развитие существующих и разработка новых подходов для обработки данных транскриптомных экспериментов является актуальной задачей.

На данный момент информация о степени изменения экспрессии генов, как правило, используется только в трех следующих задачах: (1) как порог для отбора генов с самой сильной степенью изменения экспрессии, (2) как порог для выбора ДЭГ, (3) для кластеризации генов со схожим паттерном изменения экспрессии. В рамках данной диссертационной работы разрабатывается метод, позволяющий выявлять статистически значимую взаимосвязь между функцией генов и степенью изменения экспрессии.

2. Материалы и методы

2.1. Материалы

2.1.1. Транскриптомные данные

Для тестирования возможностей нового метода, разрабатываемого в данной диссертационной работе, использовались данные трех десятков транскриптомных экспериментов из базы данных GEO (Gene Expression Omnibus) (Edgar et al., 2002) (Таблица 2). Данные были подобраны так, чтобы в них содержалось достаточное число ДЭГ, были представлены разные технологии профилирования экспрессии и разные модельные организмы. Кроме того, детальный анализ был проведен на результатах четырех из них (Таблица 2). Во-первых, это были данные RNA-seq эксперимента по обработке корней *Arabidopsis thaliana* ауксином (1 мкмоль ИУК, 6 часов), полученные в ИЦиГ СО РАН на платформе SOLiD 5500 (GSE97258) (Omelyanchuk et al., 2017). Во-вторых, были использованы данные микрочип эксперимента по изучению влияния ауксина на экспрессию генов корня *Arabidopsis thaliana* с различным временем обработки (1 мкмоль ИУК, 30 мин-24 часа) (Lewis et al., 2013), полученные на чипе Affymetrix Arabidopsis ATH1 Genome Array. В третьих, были использованы данные RNA-Seq эксперимента по изучению экспрессии генов в клетках рака предстательной железы человека (клеточная линия LNCaP) по сравнению с нормальными клетками (клеточная линия HPrEC) (GSE70466), а также с индуцированной экспрессией сплайс-варианта гена андрогенового рецептора AR-V7, (GSE71334) (Cottard et al., 2017), полученные на платформе Illumina HiSeq 2500. Во всех экспериментах было по три биологических реплики. Списки ДЭГ из данных RNA-seq экспериментов были получены при помощи программы DESeq2 (Love et al., 2014) с порогом FDR равным 0.05 и порогом степени изменения экспрессии (fold-change) равным 0.

Таблица 2. Набор транскриптомных экспериментов, использованных для апробации метода FSEA.

№**	ID	Условие	Организм	PubMed ID	Авторы	Платформа	Год
-	GSE42007	IAA, 0.5 - 24 ч	*Arabidopsis thaliana*	24045021	Lewis et al	ДНК-микрочип	2013
1	GSE124643	*DMSO vs 2x*	*Homo sapiens*	30924641	Zucconi et al	RNA-seq	2019
2	GSE124643	*DMSO vs 228*	*Homo sapiens*	30924641	Zucconi et al	RNA-seq	2019
3	GSE142504	*UndiffCtl vs UndiffC9*	*Homo sapiens*	31843624	Chai et al	RNA-seq	2019
4	GSE130729	*CBF 0 vs 3*	*Arabidopsis thaliana*	27353960	Jia et al	RNA-seq	2019
5	GSE92705	*WT vs Rab11*	*Drosophila melanogaster*	31213502	Nie et al	RNA-seq	2019
6	GSE12404	*b vs t*	*Arabidopsis thaliana*	23319655	Belmonte et al	ДНК микрочип	2008
7	GSE141873	*tumor F1*	*Homo sapiens*	32024004	Roche et al	ДНК микрочип	2019
8	GSE12404	*h vs g*	*Arabidopsis thaliana*	23319655	Belmonte et al	ДНК микрочип	2008
9	GSE130729	*WT 0 vs 3*	*Arabidopsis thaliana*	27353960	Jia et al	RNA-seq	2019
10	GSE71334	*AR-WT vs AR-V7*	*Homo sapiens*	29069764	Cottard et al	RNA-seq	2016
11	GSE40216	*T0 vs 4h*	*Arabidopsis thaliana*	-/-	Blanvillain-Baufumé et al	RNA-seq	2017
12	GSE141873	*control F1*	*Homo sapiens*	32024004	Roche et al	ДНК микрочип	2019
13	GSE97258*	*Control vs IAA 6h*	*Arabidopsis thaliana*	28559568	Omelyanchuk et al	RNA-seq	2017
14	GSE142504	*DiffCtl_vs_DiffC9*	*Homo sapiens*	31843624	Chai et al	RNA-seq	2019
15	GSE67332	*SW warm vs SW 2 weeks cold*	*Arabidopsis thaliana*	26369909	Gehan et al	RNA-seq	2015

16	GSE12404	*m vs b*	*Arabidopsis thaliana*	23319655	Belmonte et al	ДНК микрочип	2008
17	GSE142504	*UndiffC9 vs DiffC9*	*Homo sapiens*	31843624	Chai et al	RNA-seq	2019
18	GSE67332	*IT warm vs IT 2 weeks cold*	*Arabidopsis thaliana*	26369909	Gehan et al	RNA-seq	2015
19	GSE40216	*T0 vs 8h*	*Arabidopsis thaliana*	-/-	Blanvillain-Baufumé et al	RNA-seq	2017
20	GSE67332	*SW warm vs SW 1 week cold*	*Arabidopsis thaliana*	26369909	Gehan et al	RNA-seq	2015
21	GSE12404	*g vs o*	*Arabidopsis thaliana*	23319655	Belmonte et al	ДНК микрочип	2008
22	GSE130729	*CBF 0 vs 24*	*Arabidopsis thaliana*	27353960	Jia et al	RNA-seq	2019
23	GSE67332	*IT warm vs IT 1 week cold*	*Arabidopsis thaliana*	26369909	Gehan et al	RNA-seq	2015
24	GSE63406	*WT control vs WT 1h*	*Arabidopsis thaliana*	26170331	Schlaen et al	RNA-seq	2015
25	GSE142504	*UndiffCtl vs DiffCtl*	*Homo sapiens*	31843624	Chai et al	RNA-seq	2019
26	GSE130729	*WT 0 vs 24*	*Arabidopsis thaliana*	27353960	Jia et al	RNA-seq	2019
27	GSE130729	*CBF 3 vs 24*	*Arabidopsis thaliana*	27353960	Jia et al	RNA-seq	2019
28	GSE63406	*WT control vs WT 24h*	*Arabidopsis thaliana*	26170331	Schlaen et al	RNA-seq	2015
29	GSE70466	*LNCaP vs PrEC*	*Homo sapiens*	-/-	French et al	RNA-seq	2018
30	GSE73784	*LNCaP vs PrEC*	*Homo sapiens*	27053337	Taberlay et al	RNA-seq	2016

* - секвенирование было осуществлено на SOLiD 5500 в ЦКП ФИЦ ИЦиГ СО РАН ** - порядковый номер эксперимента из Рисунка 11.

2.1.2. Данные функциональной аннотации и онтологии генов

Для функциональной аннотации генов в работе использовались данные об ассоциации генов с определенной молекулярной функцией, биологическим процессом или клеточным компонентом (аннотация генов), представленные в файлах формата GAF (GO Annotation File), доступных в базе данных ресурса Gene Ontology (http://geneontology.org) (Ashburner et al., 2000), для *Arabidopsis thaliana* и полученные с помощью R пакета “org.Hs.eg.db” (Carlson, 2017), и данные о взаимоотношениях молекулярных функций, биологических процессов и клеточных компонент (Генная Онтология), представленные в файле формата OBO (Open Biomedical Ontologies). Файлы формата GAF и OBO были взяты из базы данных консорциума Генная Онтология (Gene Ontology Consortium) (Ashburner et al., 2000). Наборы генов, содержащие в своих промоторах регуляторные элементы (набор C3), и ассоциированные с процессами канцерогенеза (набор C6) были взяты из базы данных MSigDB (Liberzon et al., 2011).

2.2. Методы

В рамках данной работы был создан метод функциональной аннотации генов, позволяющий производить анализ ассоциации характеристик генов, например, категории ГО с определенным уровнем изменения экспрессии. Метод реализован в виде пакета FoldGO для языка программирования R версии 3.5.0 и выше. Для предоставления свободного доступа пакет был размещен в репозитории Bioconductor (http://bioconductor.org/packages/release/bioc/html/FoldGO.html). Исходный код пакета доступен через репозиторий GitHub (https://github.com/DanWiebe/FoldGO).

Использование пакета FoldGO предполагает базовые знания языка программирования R, поэтому, для того, чтобы метод был доступен более широкой аудитории пользователей, в сотрудничестве с к.б.н. Лашиным С.А. и Мухиным А.М. метод был реализован в виде веб-сервиса, который доступен по ссылке: http://webfsgor.sysbio.cytogen.ru/.

2.2.1. Метод анализа представленности функциональных групп генов с учетом степени изменения транскрипции (FSEA)

В качестве входных данных метода используется набор генов $G = G_1, \ldots\ldots, G_n$ и соответствующие каждому гену логарифмированные значения экспрессии $X = X_1, \ldots\ldots, X_n$, взятые по модулю. Исходный набор генов сортируется по значению экспрессии $G_{(1)}, \ldots\ldots, G_{(n)}$, таким образом, что $X_{(1)} < X_{(2)} < \ldots\ldots < X_{(n)} \mid X_{(1)} = min(X_1, \ldots\ldots, X_n), X_{(n)} = max(X_1, \ldots\ldots, X_n)$. Далее отсортированный по значению степени изменения экспрессии X набор генов G разбивается на k квантилей $Q_1, \ldots\ldots, Q_k$, таким образом, что для каждого $Q_i = G_{i,1}, \ldots\ldots, G_{i,m}$ выполняются следующие условия:

1) $X_{i,j} < f(\frac{i}{k})$;
2) $X_{i,j} \geq f(\frac{i-1}{k})$, при $i > 1$;
3) $X_{i,j} \geq min(X_1, \ldots\ldots, X_n)$, при $i = 1$,

где $X_{i,j}$ - значение степени изменения экспрессии для $G_{i,j}$, $i \in \{1, \ldots, k\}$,$j \in \{1, \ldots, m\}$, а f- функция, принимающая в качестве аргумента долю значений степени изменения экспрессии, лежащих ниже границы соответствующей квантили, и возвращающая значение степени изменения экспрессии, соответствующее границе данной квантили. Далее генерируется $\sum_{n=2}^{k} n$ вариантов объединений соседних квантилей $\cup_{n=i}^{j} Q_n$, где $i, j \in \{1, \ldots, k\}, i < j, i \neq 1 \wedge j \neq k$. Далее для каждой ГО категории из заранее подготовленного набора $GO = GO_1, \ldots\ldots, GO_s$, где $GO_i = \{G_{i,1}, \ldots\ldots, G_{i,t}\} | \forall G_{i,j} \in G, i \in \{1, \ldots\ldots, s\}, j \in \{1, \ldots\ldots, t\}$ (t - количество

генов, аннотированных к категории GO_i) и всех квантилей и их объединений производится оценка обогащения с использованием точного теста Фишера по таблице сопряженности:

Таблица 3. Таблица сопряженности для расчета ассоциации между аннотацией гена к категории ГО и принадлежности к определенному интервалу степени изменения экспрессии.

	Q_r+	Q_r-
GO_i+	A	B
GO_i-	C	D

, где A = $|GO_i \cap Q_r|$, B = $|GO_i \backslash Q_r|$, C = $|Q_r \backslash GO_i|$, D =$|G \backslash (Q_r \cup GO_i)|$, $i \in \{1, \ldots\ldots, s\}$, $r \in \{1, \ldots\ldots, k\}$.

2.2.2. Структура пакета программ FoldGO

Разработанный пакет FoldGO можно подразделить на три функционально независимых модуля:

1) Модуль обработки данных полногеномных экспериментов;
2) Модуль функциональной аннотации;
3) Модуль сопоставления данных и выявления фолд-специфичности.

2.2.2.1. Модуль обработки данных полногеномных экспериментов

Входные данные подаются в виде таблиц, содержащих следующие столбцы: идентификаторы генов, значения степени изменения экспрессии (фолд), оценки уровня значимости (*p-value*). Такие данные могут быть получены при помощи пакета программ *limma* (Ritchie et al., 2015) для микрочип-экспериментов или *edgeR* (Robinson et al., 2010) (Табл. 4) и *DESeq2* (Love et al., 2014) для RNA-seq экспериментов.

Таблица 4. Пример таблицы полученной при помощи программы *edgeR* (Robinson et al., 2010), содержащей данные по дифференциальной экспрессии

генов, которая может быть использована в качестве входных данных программы FoldGO. В первом столбце указаны идентификаторы генов, столбец logFC демонстрирует степень изменения экспрессии для конкретного гена, столбец logCPM - логарифм количества прочтений на миллион картированных прочтений, столбец PValue - уровень значимости этого изменения.

	logFC	logCPM	PValue
FBgn0000003	2.25662	-2.083	5.184e-7
FBgn0000008	-0.05492	3.036	8.24e-6
FBgn0000015	-3.78099	-1.793	2.51e-4
FBgn0000017	-0.24803	8.440	7.89e-3
FBgn0000018	-0.03655	4.940	0.043
...	...	...	...

На первом этапе обработки входных данных гены сортируются по степени изменения экспрессии. Далее отсортированный список генов разбивается на равные группы (квантили). При этом генерируются все возможные комбинации соседних групп генов без перестановок (Рис. 1). Более подробно процесс разбиения списка генов на квантили описан в подглаве 2.2.1.

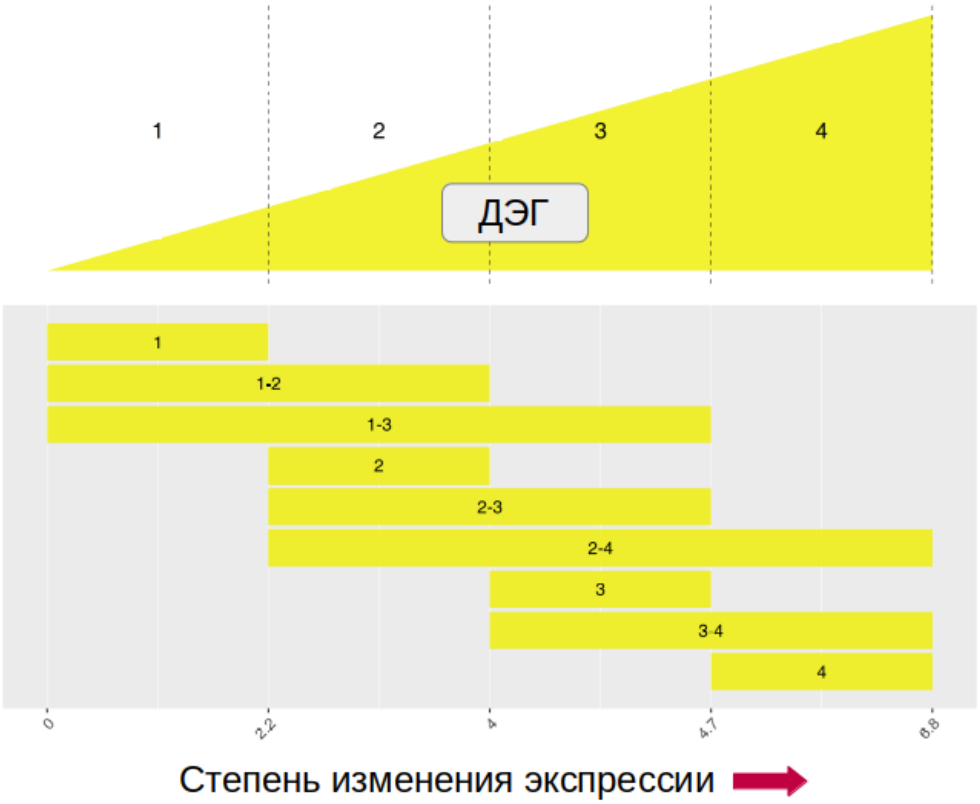

Рисунок 1. Схематическое изображение разбиения на четыре квантили списка дифференциально экспрессирующихся генов, отсортированного по возрастанию степени изменения экспрессии генов.

2.2.2.2. Модуль функциональной аннотации

Полученные списки генов (квантили и их объединения) используются в качестве входных данных для функциональной аннотации списков генов, которая производится при помощи стороннего R пакета topGO (Alexa и Rahnenfuhrer, 2019). Категории ГО, показавшие значимое обогащение в аннотации для полного интервала отбираются для дальнейшего анализа. Данный этап может быть пропущен, в таком случае анализ проводится для всех категорий ГО из используемой аннотации.

2.2.2.3. Модуль сопоставления данных и выявления фолд-специфичности

Для определения того, является ли конкретная категория ГО или мотив, обогащенными в определенном интервале степени изменения экспрессии

(фолд-специфичной), для каждой категории или мотива рассчитывается точный тест Фишера. Для расчета точного теста Фишера используется таблица сопряженности, содержащая значения количества генов, ассоциированных и неассоциированных с определенной категорией ГО или содержащих и не содержащих определенный мотив в регуляторных районах, значимо изменивших свою экспрессию в определенном интервале значений степени изменения экспрессии (Табл. 3). Затем для каждой категории или мотива отбирается интервал степени изменения экспрессии с минимальным значением *p-value*. Далее отбираются категории или мотивы, у которых значение *p-value* проходит порог, установленный с учетом коррекции на множественное сравнение. Для коррекции на множественное тестирование в программе используются следующие методы: коррекция Бонферрони (Bonferroni correction) (Bland и Altman, 1995), Бенджамини-Хохберга (Benjamini-Hochberg correction) (Benjamini и Hochberg, 1995), Бенджамини-Якутели (Benjamini-Yekutieli correction) (Benjamini и Yekutieli, 2001). Отобранные таким образом термины или мотивы будут считаться фолд-специфичными.

2.2.3. Расчет доли ложноположительных результатов

Для оценки доли ложноположительных результатов использовались данные по дифференциальной экспрессии генов клеточной линии LNCaP с конститутивно активным андрогенным рецептором (AR-V7 сплайс-вариант) (Cottard et al., 2017). Для нарушения связи между значениями степени изменения экспрессии проводилась процедура пермутации, в результате которой идентификаторы генов были случайным образом перемешаны. Далее производился отбор дифференциально экспрессирующихся генов и их разделение на гены повышающие и снижающие экспрессию. Таким образом было сгенерировано 2000 наборов ДЭГ. Полученные наборы данных использовались для выявления фолд-специфичных ГО категорий. Далее была

рассчитана доля наборов данных, в которых была выявлена хотя бы одна фолд-специфичная категория ГО с пороговым значением уровня ложноположительных результатов (False Discovery Rate, далее по тексту FDR), равным 0.05.

2.2.4. Оценка чувствительности метода

Для того чтобы оценить способность разработанного метода находить группы генов с высокой корреляцией по уровню экспрессии, мы создали набор симулированных данных с заданной корреляционной структурой. Для симуляции значений экспрессии использовалось многомерное нормальное распределение с ковариационной матрицей заданной таким образом, чтобы сформировать восемь групп генов с различными параметрами:

- 6 групп ДЭГ с сильной внутригрупповой корреляцией размером 5, 10, 20, 30, 40, 50 генов, $\mu = 1$, $\rho > 0.7$;
- группа ДЭГ без внутригрупповой корреляции размером 100 генов, $\mu = 1$, $\rho \sim 0$;
- группа не ДЭГ без внутригрупповой корреляции размером 700 генов, $\mu = 0$, $\rho \sim 0$.

Для проверки разработанного метода на симулированных данных было составлено 100 выборок из сгенерированного многомерного нормального распределения, а также составлена аннотация, состоящая из первых шести групп генов с сильной внутригрупповой корреляцией. Полученные данные использовались для оценки чувствительности метода.

3. Результаты

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

Логично предположить, что гены, объединенные одной функцией, должны иметь схожий профиль экспрессии, чтобы минимизировать затраты клетки на наработку всех регуляторов одного и того же процесса. Существующие подходы для анализа функционального обогащения (SEA, GSEA) не позволяют напрямую выявлять ассоциацию между функцией гена и степенью изменения его экспрессии. Поэтому мы разработали метод FSEA (Анализ фолд-специфичного обогащения, Fold-change Specific Enrichment Analysis), который позволяет выявлять ГО категории, 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 группы одинаковой размерности (квантили), генерации всех объединений соседних квантилей, и статистического теста для проверки гипотезы об обогащении определенных характеристик генов в интервалах степеней изменения экспрессии по сравнению со всеми дифференциально экспрессирующимися генами. Метод описан в главе 2.2.1.

3.2. Разработка пакета программ FoldGO для функциональной аннотации транскриптомных данных методом FSEA

Данный метод реализован в виде пакета FoldGO для языка R (https://www.R-project.org). В качестве входных данных пакет программ использует данные транскриптомных экспериментов. Блочная структура пакета программ позволяет использовать сторонние программы для аннотации генов. Выходные данные пакета программ могут быть представлены в виде таблиц для дальнейшей работы с данными (Рис. 2), а

также могут быть представлены графически, в виде диаграмм (Рис. 3). Для организации публичного доступа пакет программ размещен в репозитории Bioconductor (http://bioconductor.org/packages/release/bioc/html/FoldGO.html). Подробное описание пакета программ представлено в главе Материалы и методы.

ids	namespace	name	padj	interval
GO:0003723	molecular_function	RNA binding	0.000043	1-3
GO:0003735	molecular_function	structural constituent of ribosome	5.433338e-17	1
GO:0005198	molecular_function	structural molecule activity	1.811485e-15	1
GO:0005488	molecular_function	binding	0.008135	1-5
GO:0005575	cellular_component	cellular_component	0.007051	1-3
GO:0005730	cellular_component	nucleolus	0.009593	2-5
GO:0005840	cellular_component	ribosome	1.539269e-12	1
GO:0006412	biological_process	translation	8.169046e-18	1
GO:0006518	biological_process	peptide metabolic process	2.84712e-17	1
GO:0006807	biological_process	nitrogen compound metabolic process	0.004839	1

Рисунок 2. Выходная таблица веб-сервиса FoldGO (https://webfsgor.sysbio.cytogen.ru), отображающая фолд-специфичные категории ГО. Колонки содержат идентификаторы (*ids*) и названия (*name*) ГО категорий, а также значения значения *p-value* с учетом коррекции на множественное тестирование (*padj*)и обозначения интервалов степени изменения экспрессии для которых была показана наилучшая статистическая значимость (*interval*).

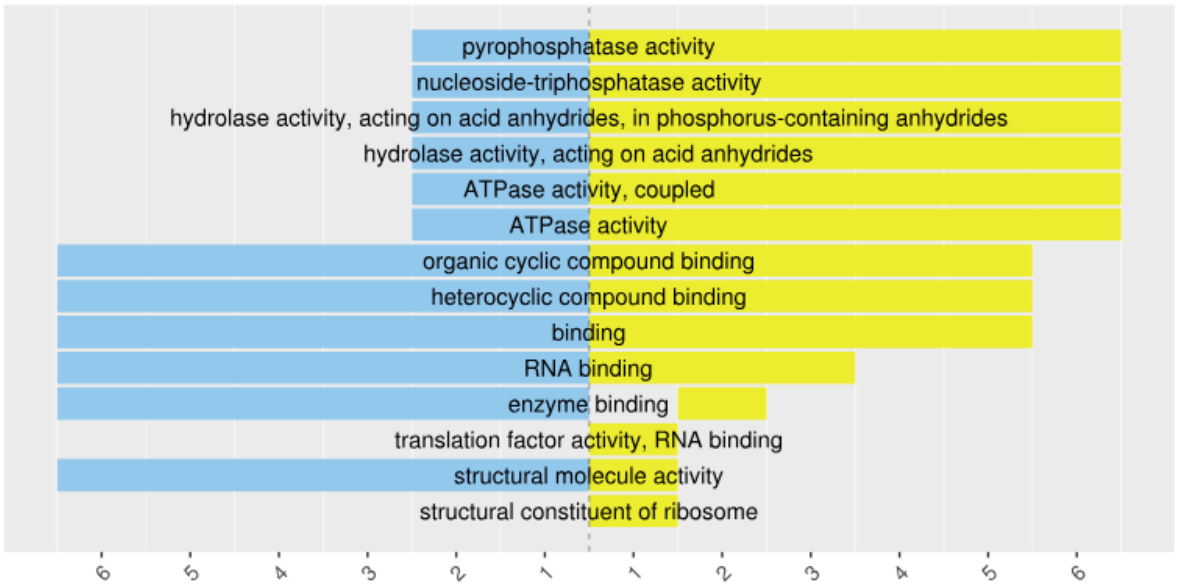

Рисунок 3. Выходная гистограмма веб-сервиса FoldGO

(https://webfsgor.sysbio.cytogen.ru), отображающая фолд-специфичные категории ГО. Для каждой категории ГО отображен интервал степени изменения экспрессии, в котором экспрессия генов регулируется фолд-специфично. Активация транскрипции отмечена желтым цветом, подавление транскрипции отмечено синим цветом.

3.3. Оценка применимости метода FSEA

3.3.1. Оценка доли ложноположительных результатов

Метод был протестирован для анализа нескольких десятков транскриптомных экспериментов из базы данных GEO (Таблица 2). Данные подбирались по принципу достаточности количества ДЭГ (более 200), так, чтобы в анализе участвовали данные полученные разными технологиями профилирования экспрессии и для разных модельных организмов. Оценка в 200 генов была получена опытным путем: с транскриптомами с меньшим количеством ДЭГ, метод FSEA может работать нестабильно из-за малого количества аннотаций генов к ГО категориям. Для всех проанализированных транскриптомов (Таблица 2), FSEA находил хотя бы несколько фолд-специфичных ГО категорий (подробно описано в главе 3.3.4). Для оценки адекватности статистической процедуры поиска фолд-специфичных ГО категорий был произведен расчет доли ложноположительных результатов (см. Материалы и методы, п. 2.2.3.). Для этого мы использовали данные эксперимента по исследованию экспрессии гена конститутивно активного андрогенового рецептора (AR-V7 сплайс-вариант) в клеточной линии LNCaP (Cottard et al., 2017). В этих данных была выявлена 381 фолд-специфичная ГО категория с долей ложноположительных результатов, равной 0.034 (Рис. 4).

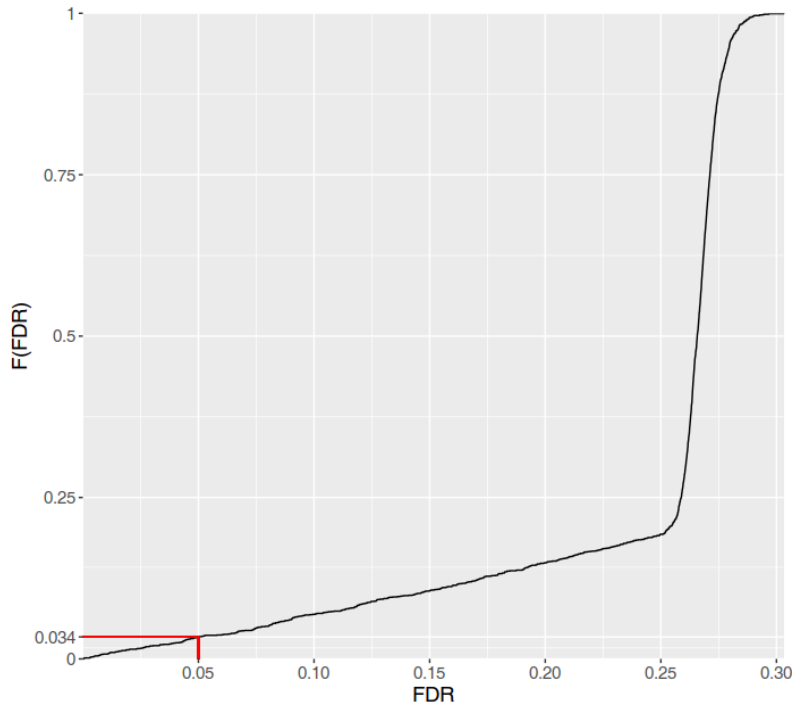

Рисунок 4. График эмпирической функции распределения минимальных значений *p-value,* рассчитанных FSEA после коррекции на множественное тестирование (FDR) на пермутированных данных по дифференциальной экспрессии генов в линии LNCaP (см. Материалы и Методы, п.2.2.3). Красными линиями отмечена точка, которой соответствует доля результатов с минимальным значением FDR < 0.05.

Далее мы проверили, как влияет варьирование параметров для анализа FSEA на долю ложноположительных результатов. Сначала мы проверили, как зависит доля ложноположительных результатов от количества квантилей, на которое производится разбиение исходного набора ДЭГ. Для этого мы провели анализ FSEA без коррекции на множественное тестирование на пермутированных данных по дифференциальной экспрессии генов в линии LNCaP (см. Материалы и Методы, п.2.2.3) с разбиением от 2 до 10 квантилей и обнаружили что при разбиении на более чем 3 квантили, доля ГО категорий, для которых метод FSEA показал значение *p-value* < 0.05 в среднем составляет 0.048 ± 0.007% и продолжает повышаться при увеличении количества квантилей, достигая 0.217 ± 0.013% при разбиении на

10 квантилей (Рис. 5 А). Далее, принимая за ложноположительный результат обнаружение хотя бы одной фолд-специфичной категории для каждой квантили и каждого количества ДЭГ, мы провели аналогичный анализ но с коррекцией на множественное тестирование учитывающей количество анализируемых ГО категорий. В результате метод FSEA показал что при разбиении на более чем 6 квантилей доля ложноположительных результатов превышает 5% и продолжает повышаться при увеличении количества квантилей, достигая ~12 % при разбиении на 10 квантилей (Рис. 5 Б).

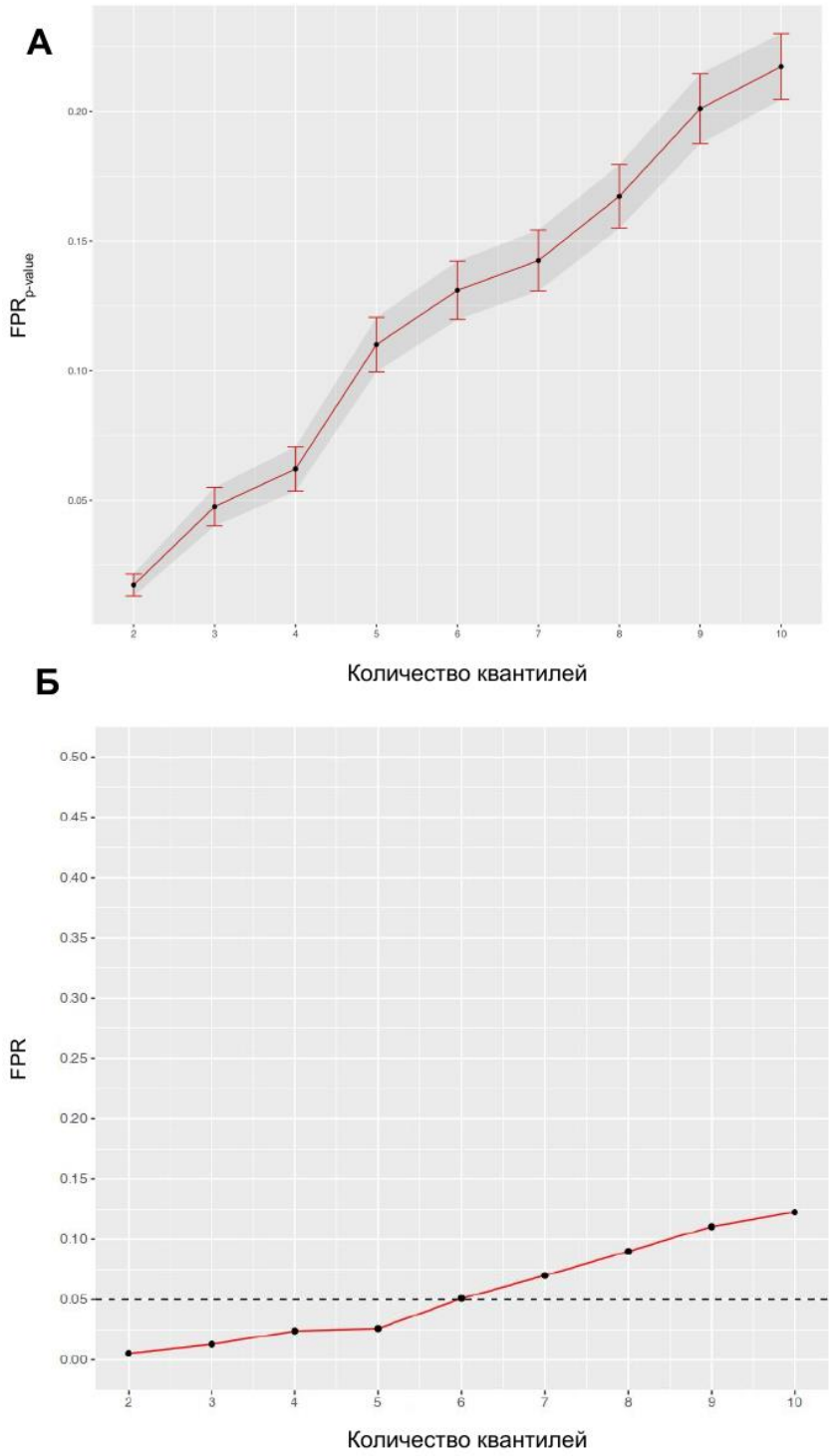

Рисунок 5. Графики зависимости доли ложноположительных результатов от количества квантилей , на которое производится разбиения набора ДЭГ при использовании FSEA. (**А**) Зависимость доли ГО категорий, для которых метод FSEA показал значение p-value < 0.05 без коррекции на множественное тестирование ($FPR_{p\text{-}value}$), от количества квантилей. Вертикальными линиями отмечены границы стандартного отклонения. (**Б**) Зависимость доли результатов FSEA в которых была выявлена одна или более фолд-

специфичная ГО категория (FPR) от количества квантилей. Горизонтальной пунктирной линией отмечена доля ложноположительных результатов, равная 5 %.

Для проверки влияния количества ДЭГ на долю ложноположительных результатов также использовались пермутированные данные, полученные на линии LNCaP (Cottard et al., 2017) (см. Материалы и Методы, п.2.2.3). После перемешивания идентификаторов генов отбор ДЭГ производился не при помощи установления порога значимости, а случайным образом по количеству генов от 200 до 2000 с шагом в 200 генов. Данный анализ также производился с варьированием количества квантилей, на которое производилось разбиение от 2 до 10 и коррекцией на множественное тестирование учитывающей количество анализируемых ГО категорий. В результате мы не наблюдали строгой зависимости между количеством генов, взятых в анализ и долей ложноположительных результатов при разбиении от 2 до 10 квантилей (Рис. 6). Таким образом можно постановить, что феномен фолд-специфинчного обогащения ГО категориями не является следствием случайной корреляции генов по степени изменения экспрессии.

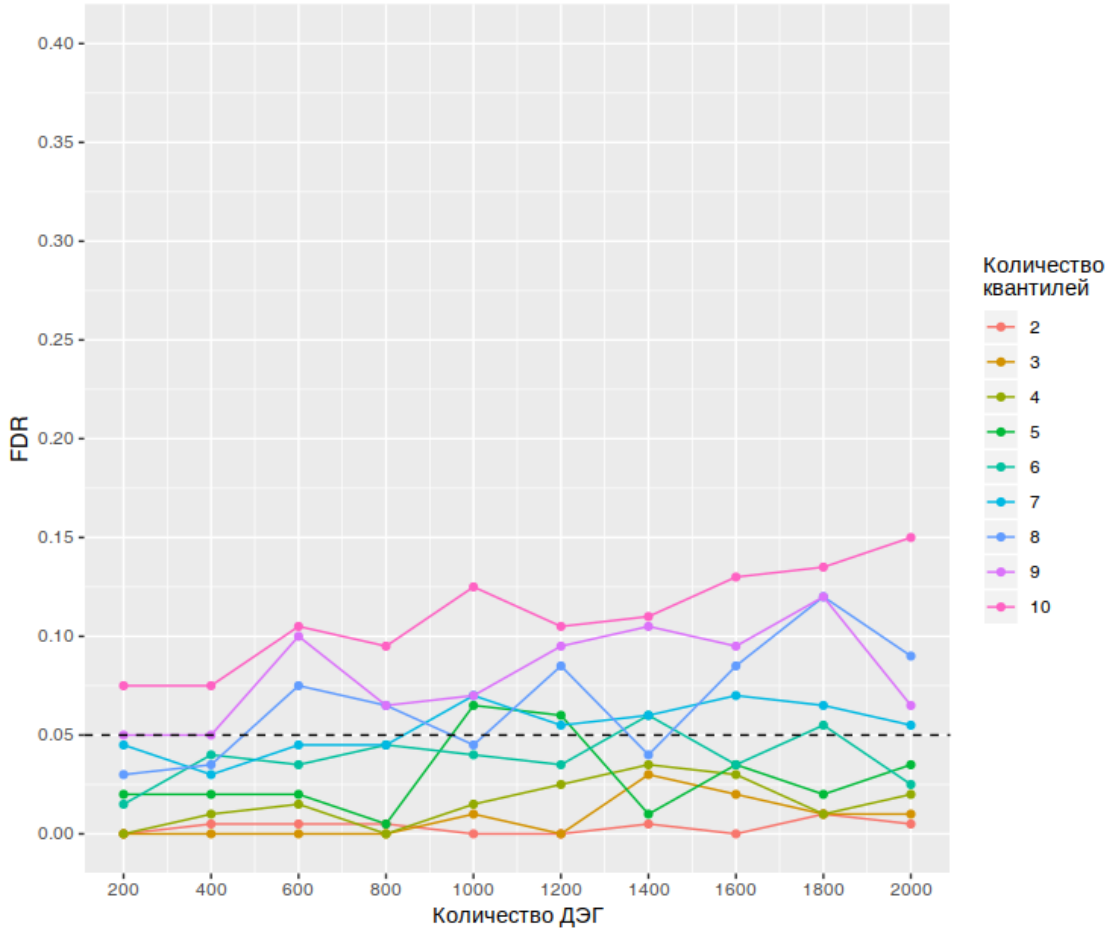

Рисунок 6. График зависимости количества ДЭГ и доли ложноположительных результатов, полученных при использовании FSEA с коррекцией на множественное тестирование учитывающей количество анализируемых ГО категорий. Горизонтальной пунктирной линией отмечена доля ложноположительных результатов равная 5 процентам. Цветами отмечено количество квантилей на которое производится разбиения набора ДЭГ.

Данный анализ показывает, что при разбиении исследуемого набора ДЭГ не более чем на 6 квантилей, можно ожидать долю ложноположительных результатов, незначительно превышающую 5 % независимо от количества ДЭГ.

3.3.2. Оценка чувствительности метода

Для того, чтобы оценить чувствительность метода FSEA был сгенерирован набор данных по дифференциальной экспрессии (см. Материалы и Методы, п. 2.2.4). При апробации метода на сгенерированных данных производилось варьирование количества квантилей, на которое будет производится разбиение исходного набора ДЭГ, с целью оценки влияния данного параметра на количество обнаруженных групп. В результате такого анализа было обнаружено, что при разбиении на 5 квантилей и более, количество обнаруженных групп размером более 5 генов со схожей степенью изменения экспрессии превосходит 78 процентов. Коэффициент корреляции для групп, различающихся по количеству генов, отбирался из равномерного распределения случайным образом с минимальным допустимым значением, равным 0.7.

Для сравнения результатов FSEA с существующими методами был произведен анализ функционального обогащения на симулированных данных с использованием метода GSEA (Subramanian et al., 2005). В результате было показано, что при разбиении на 6 квантилей, FSEA обнаруживает обогащенными примерно на 8 процентов больше групп, чем GSEA. Стоит отметить, что FSEA показывает заметно лучший результат при выявлении групп генов небольшого размера (до 20 генов включительно). В частности FSEA находит на 26% больше групп генов со схожей степенью изменения экспрессии, размером 10 генов (Рис. 7). Единственная группа, на которой GSEA показала результат лучше, чем FSEA, это группа, состоящая из 30 генов. При разбиении этой группы на 4 и более квантили FSEA обнаружила 78 % таких групп, в то время как GSEA обнаружила 99 %. Данная группа отлична от других высоким значением коэффициента корреляции ($\rho \sim 0.9935$).

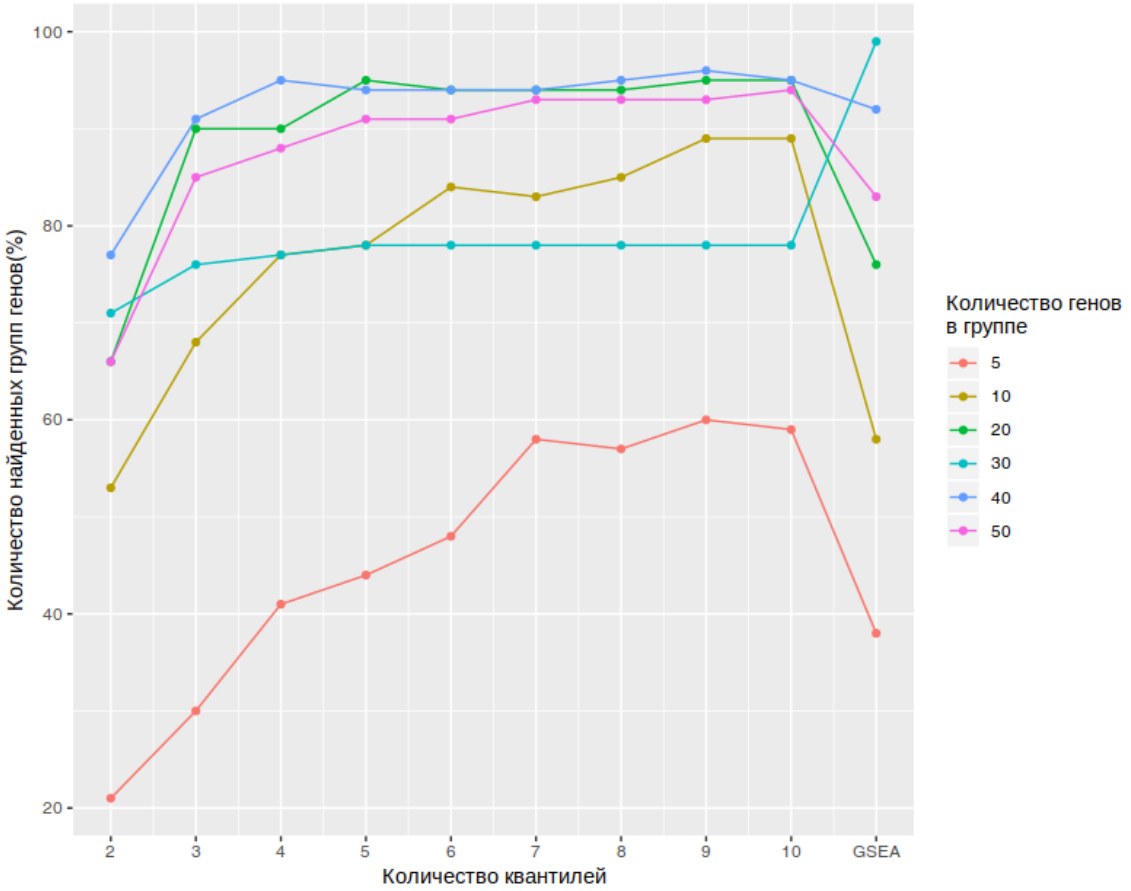

Рисунок 7. График зависимости доли обнаруженных с помощью метода FSEA групп генов со схожей степенью изменения экспрессии от количества квантилей, на которые разбивается исходный набор генов. По оси x отмечено количество квантилей, на которое производилось разбиение исходного списка генов, а также результат, полученный с использованием метода GSEA.

Данный анализ показал, что метод FSEA лучше, чем GSEA обнаруживает группы с невысокой, но значимой корреляцией (от 0.7 до 0.9) по степени изменения экспрессии. Однако GSEA показывает лучший результат для групп генов с высокой корреляцией ($\rho > 0.9$). Важно напомнить, что в отличие от GSEA и SEA, метод FSEA обладает уникальной способностью давать оценку силе транскрипционного ответа функционально-связанной группы генов.

3.3.3. Оптимизация метода для работы с произвольными значениями параметров

В связи с выявленной тенденцией к повышению доли ложноположительных результатов при увеличении количества квантилей (Рис. 5) возникла необходимость в коррекции метода с целью обеспечить долю ложноположительных результатов менее 5 % при разбиении набора ДЭГ на произвольное количество квантилей. Вероятной причиной роста доли ложноположительных результатов при увеличении количества квантилей является процедура коррекции на множественное тестирование, которая производится с учетом количества анализируемых категорий ГО, и применяется на этапе выявления фолд-специфичных категорий ГО (см. Материалы и методы, п. 2.2.2.3). Однако реальное количество тестов, производимое при выявлении фолд-специфичных категорий ГО, равно количеству исследуемых интервалов степени изменения экспрессии (квантили и их объединения), умноженному на количество анализируемых категорий ГО (1).

(1) $n_t = n_{go} \times \sum_{q=2}^{k} q$, где n_t - количество статистических тестов, n_{go} - количество анализируемых категорий ГО, k - количество квантилей.

Поэтому мы модифицировали метод FSEA с учетом коррекции на полное количество статистических тестов. В результате оценки доли ложноположительных результатов для модифицированного метода FSEA мы наблюдали FDR < 0.05 при количестве ДЭГ от 100 до 1000 и при разбиении на количество квантилей от 2 до 10 без выраженной тенденции к росту при увеличении значения параметров (Рис. 8).

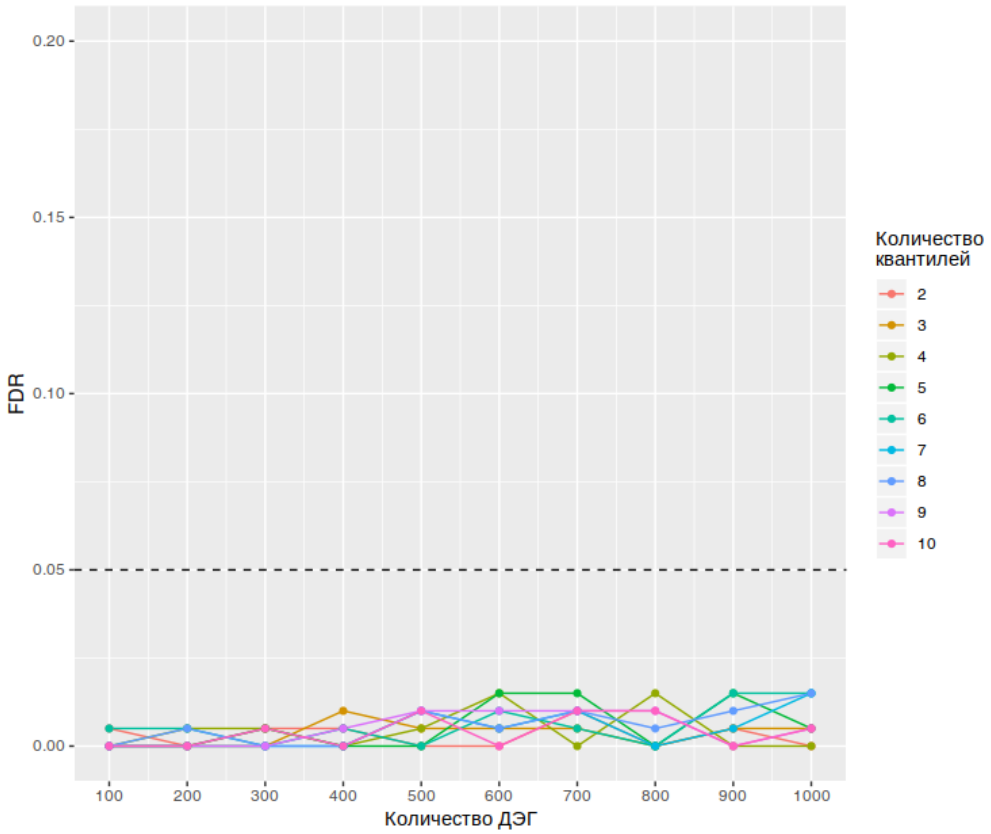

Рисунок 8. График зависимости количества ДЭГ и доли ложноположительных результатов при использовании FSEA с коррекцией на полное количество статистических тестов, рассчитанное по формуле (1). Горизонтальной пунктирной линией отмечена доля ложноположительных результатов равная 5 процентам. Цветами отмечено количество квантилей на которое производится разбиения набора ДЭГ.

Чтобы проверить, как изменилась чувствительность метода FSEA с коррекцией на множественное тестирование, учитывающей количество интервалов степени изменения экспрессии, мы повторили анализ на симулированных данных аналогично п. 3.3.2. Повторный анализ показал, что метод обнаруживает более 70 % групп генов, состоящих из 20 и более генов со схожей степенью изменения экспрессии. Для сравнения работы FSEA с коррекцией на множественное тестирование, учитывающей только количество ГО категорий, метод обнаруживал более 90 % групп размером 20, 40 и 50 генов при разбиении на 5 и более квантилей. Наилучший результат

(~90 %) был показан для групп, состоящих из 40 генов при разбиении на 9 квантилей. При разбиении исходного набора генов на любое количество квантилей, FSEA находил порядка 30 % и 45 % групп, состоящих из 5 и 10 генов, соответственно, что на ~ 20 % и 40 %, соответственно, меньше, чем было обнаружено при коррекции только на количество исследуемых категорий ГО (Рис. 9).

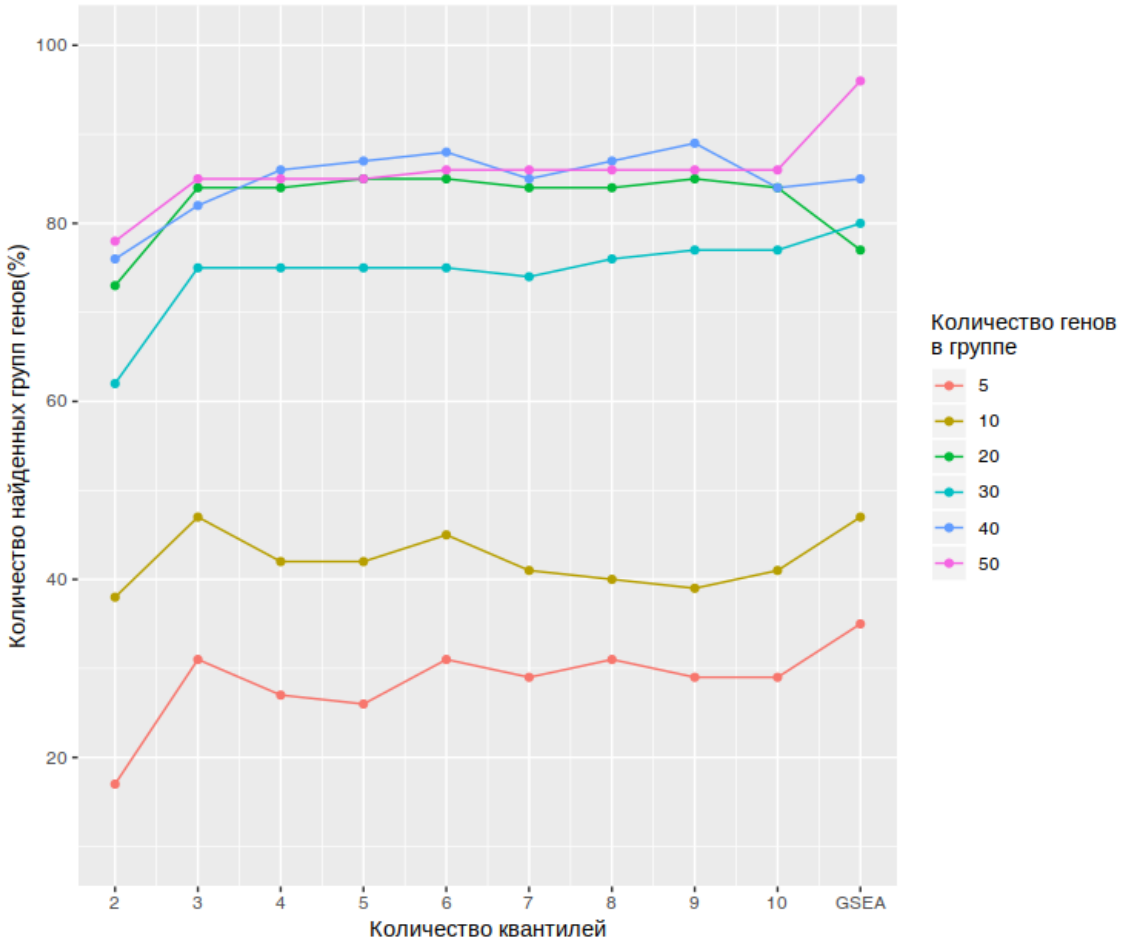

Рисунок 9. График зависимости количества квантилей, на которые разбивается исходный набор генов, и доли обнаруженных групп генов со схожей степенью изменения экспрессии при использовании FSEA с коррекцией на полное количество статистических тестов, рассчитанное по формуле (1). По оси x отмечено количество квантилей, на которое производилось разбиение исходного списка генов, а также результат, полученный с использованием метода GSEA.

Исходя из полученных результатов можно заключить, что метод FSEA показывает сопоставимые результаты как при коррекции на количество исследуемых категорий ГО (при разбиении не более чем 6 квантилей), так и при коррекции на полное количество статистических тестов. При разбиении более чем на 6 квантилей, для надежного контроля за долей ложноположительных результатов, рекомендуется использовать коррекцию на множественное тестирование с учетом полного количества статистических тестов.

3.3.4. Валидация метода на выборке транскриптомных экспериментов из базы данных GEO

Для масштабного тестирования метода FSEA на реальных данных, были использованы данные транскриптомных экспериментов, проведенных на человеке, дрозофиле и арабидопсисе из базы данных GEO (Edgar et al., 2002) (Таблица 2). Анализ производился с разбиением на 5 квантилей. По результатам анализа транскриптомных данных с помощью метода FSEA были найдены фолд-специфичные категории ГО во всех исследованных экспериментах, что указывает на универсальность фолд-специфичного транскрипционного ответа (Рис. 10 А, Б). Детальный анализ результатов FSEA показал, что каждый эксперимент характеризуется собственным, уникальным по составу, набором фолд-специфичных ГО категорий. Однако для всех транскриптомов была найдена общая закономерность: наибольшее количество фолд-специфичных ГО категорий было представлено биологическими процессами (GO BP; Biological Process), в то время как ГО категории, представляющие молекулярные функции (GO MF; Molecular Function), относительно редко выявлялись как фолд-специфичные.

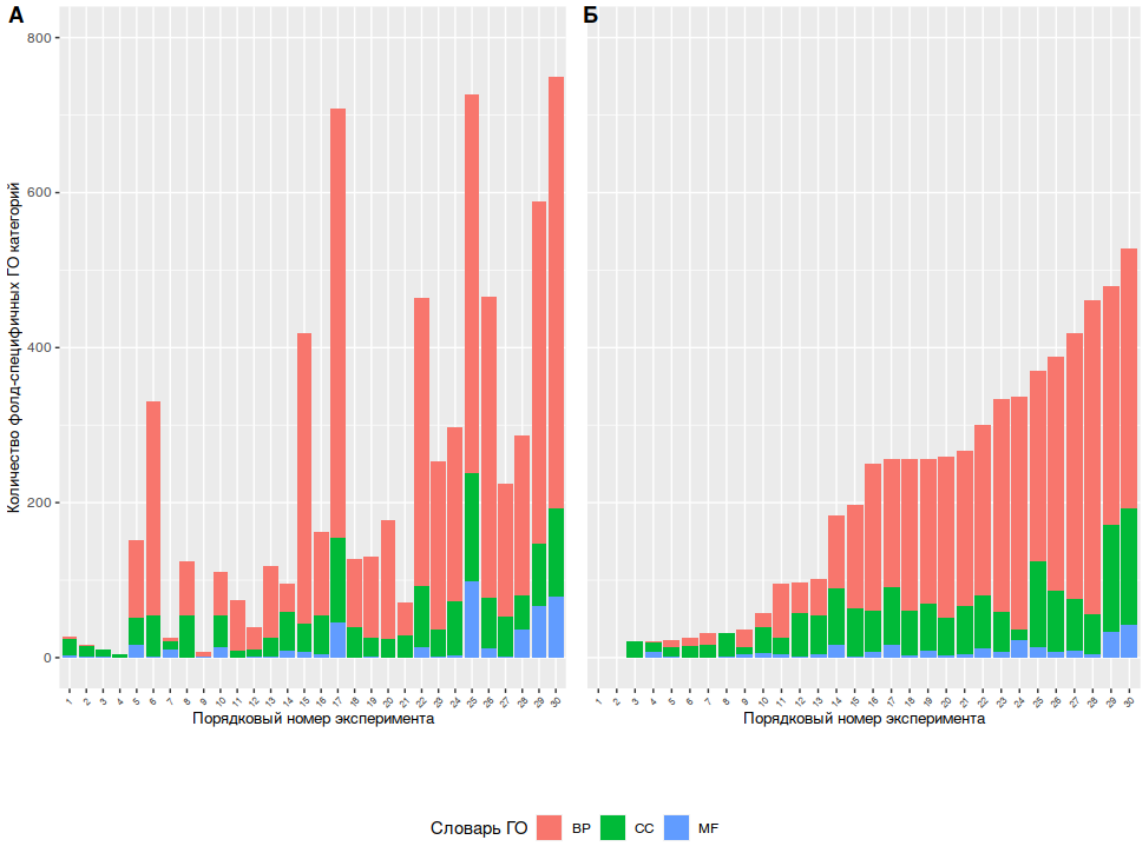

Рисунок 10. Количество фолд-специфичных терминов, обнаруженных при анализе экспериментов, перечисленных в таблице 2 для ДЭГ, экспрессия которых подавляется (**А**) и активируется (**Б**). По оси x отмечено количество фолд-специфичных ГО категорий. По оси y отмечены порядковые номера экспериментов, соответствующие номерам из таблицы 2. Цветами отмечены словари ГО: красным - биологические процессы (BP; Biological Process), зеленым - клеточные компоненты (CC; Cellular Components), синим - молекулярные функции (MF; Molecular Function).

Еще одной интересной особенностью было то, что для категорий ГО, представляющих клеточные компоненты (GO CC; Cellular Components), была показана наибольшая статистическая значимость по фолд-специфичности по сравнению с двумя другими словарями ГО (Таблица П3). Такие особенности можно объяснить тем, что регуляторные сети, обеспечивающие функционирование клеточных компонент и биологических процессов,

характеризуются высокой сложностью. Это, вероятно, и приводит к тому, что FSEA находит больше фолд-специфичных ГО категорий среди биологических процессов по сравнению с молекулярными функциями.

На этих же данных мы проанализировали каким интервалам степени изменения экспрессии чаще соответствует фолд-специфичный транскрипционный ответ (Рис. 11 А, Б). Суммирование данных по 30 экспериментам показало, что фолд-специфичный ответ, как правило, характерен для групп ДЭГ с низкой или очень высокой степенью изменения экспрессии. Это правило наблюдалось как для активируемых, так и для подавляемых ДЭГ. Этот результат является принципиально важным, так как в большинстве случаев исследователи уделяют внимание генам с наибольшей степенью изменения экспрессии и процессам, в которых они участвуют, таким образом игнорируя наличие другого типа транскрипционного ответа. Как будет показано в следующим главах, транскрипционный ответ с небольшой степенью изменения экспрессии характерен для большого количества важных молекулярно-генетических процессов, и он может быть ассоциирован с определенным типом цис-регуляторных элементов, роль которых была плохо изучена ранее. Важность таких слабо-активируемых процессов была предсказана ранее (St. Laurent et al., 2013), но на них долгие годы не обращали внимание. Метод FSEA является удобным инструментом для выявления таких процессов и предоставляет всю необходимую информацию для их дальнейшего исследования.

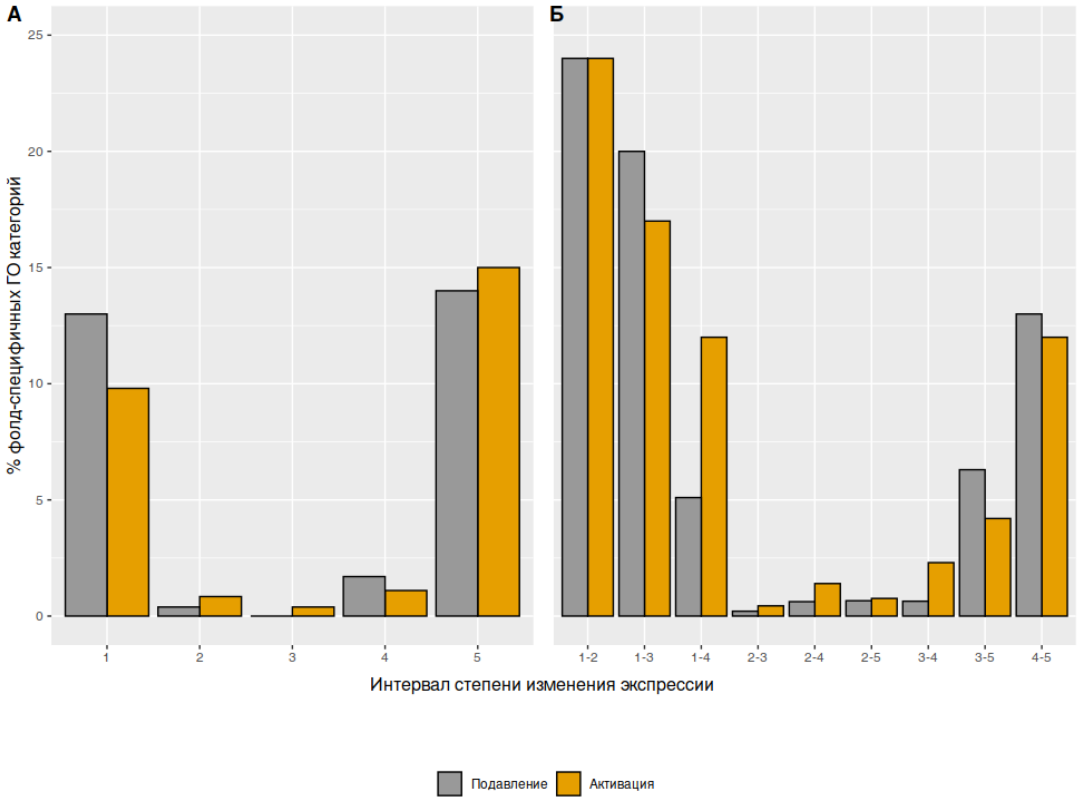

Рисунок 11. Процент фолд-специфичных категорий ГО в одиночных (**А**) и объединенных (**Б**) интервалах степени изменения экспрессии среди всех фолд-специфичных категорий ГО, обнаруженных в экспериментах, перечисленных в таблице 2 для ДЭГ, экспрессия которых подавляется (серые столбцы) и активируется (оранжевые столбцы).

3.4. Апробация метода. Анализ данных по 6-часовой обработке ауксином корней арабидопсиса

Развитие метода FSEA началось с анализа транскриптома, полученного на корнях модельного растения *Arabidopsis thaliana*, обработанного фитогормоном ауксином. Транскриптом был получен моими коллегами в ИЦиГ СО РАН для исследования морфогенетического эффекта фитогормона ауксина. В работе нам хотелось оценить, существуют ли отличия в силе транскрипционного ответа для разных функционально-связанных групп генов. Так как метода для такого анализа не существовало, мы разработали прототип FSEA, который описан ниже. Впоследствии он был доработан до готового продукта, описанного в главе 2.2.1.

Анализ дифференциальной экспрессии ауксин-индуцируемых генов выявил 789 ДЭГ, значимо ассоциированных с повышением экспрессии, и 659 ДЭГ, значимо ассоциированных с понижением экспрессии в ответ на ауксин (Benjamini-Hochberg FDR < 0.05) с разбросом степени изменения экспрессии от 1.5 до 88 и от 1.4 до 143, соответственно. Дифференциально-экспрессирующиеся гены были разделены на 6 равных групп (квантилей), а также был сгенерирован набор всех возможных комбинаций объединений соседних квантилей, для этого был создан модуль предобработки траснкиптомных данных (см. Материалы и методы, п. 2.2.2.1.). В результате был сформирован набор из 21 группы генов, для каждой группы из которого была проведена функциональная аннотация с использованием ресурса AgriGO (Tian et al., 2017) по точному тесту Фишера с коррекцией Бонферрони (Bland и Altman, 1995). Точные границы интервалов степени изменения экспрессии генов отображены на рисунке 13 (Рис. 12 А).

На этапе анализа функционального обогащения (см. Материалы и методы, п. 2.2.2.2.) для ДЭГ, повышающих и понижающих свою экспрессию в ответ на ауксин, мы отобрали 225 и 307 ГО категорий, соответственно, которые были значимо обогащены (*p-adjusted* < 0.001, точный тест Фишера, коррекция Бонферрони (Bland и Altman, 1995)) хотя бы в одном из интервалов степени изменения экспрессии. Приблизительно 15% из отобранных категорий оказались обогащенными в группах генов с очень схожими степенями изменения экспрессии, что может говорить о довольно точной регуляции определенных биологических процессов ауксином. Более того, мы обнаружили, что для некоторых ГО категорий значение *p-value* отличалось на несколько порядков от значений *p-value,* полученных при анализе функционального обогащения данными категориями для всех ДЭГ. Например, категория “трансляция” оказалась обогащена в группе дифференциально-экспрессирующихся генов с *p-value* = 7.3×10^{-19} , а для группы генов со слабым и очень слабым ответом на ауксин *p-value* для данной категории оказалось гораздо меньше (*p-value* = 6.6×10^{-54}). Для того

чтобы оценить значимость таких различий, мы провели анализ на фолд-специфичность (см. Материалы и методы, п. 2.2.2.3.). В результате которого мы обнаружили, что значительная часть терминов, а именно 82 (36%) и 36 (12%) были ассоциированы с фолд-специфичным повышением и понижением экспрессии генов, соответственно (Рис. 12 Б). Оставшиеся термины (143 для повышения и 271 для снижения экспрессии) оказались не фолд-специфичными.

Для проверки влияния количества генов, попадающих в одиночные интервалы, мы провели аналогичный анализ, но с разбиением списков дифференциально-экспрессирующихся генов на 3, 4 и 8 подсписков. Несмотря на то, что в результате данной проверки мы получили разные количества фолд-специфичных ГО категорий, качественный состав остался прежним.

Таким образом был создан прототип FSEA и обнаружено, что многие группы функционально-схожих генов изменяют свою экспрессию скоординированно, с определенной силой транскрипционного ответа. Далее мы будем называть этот феномен фолд-специфичностью транскрипционного ответа.

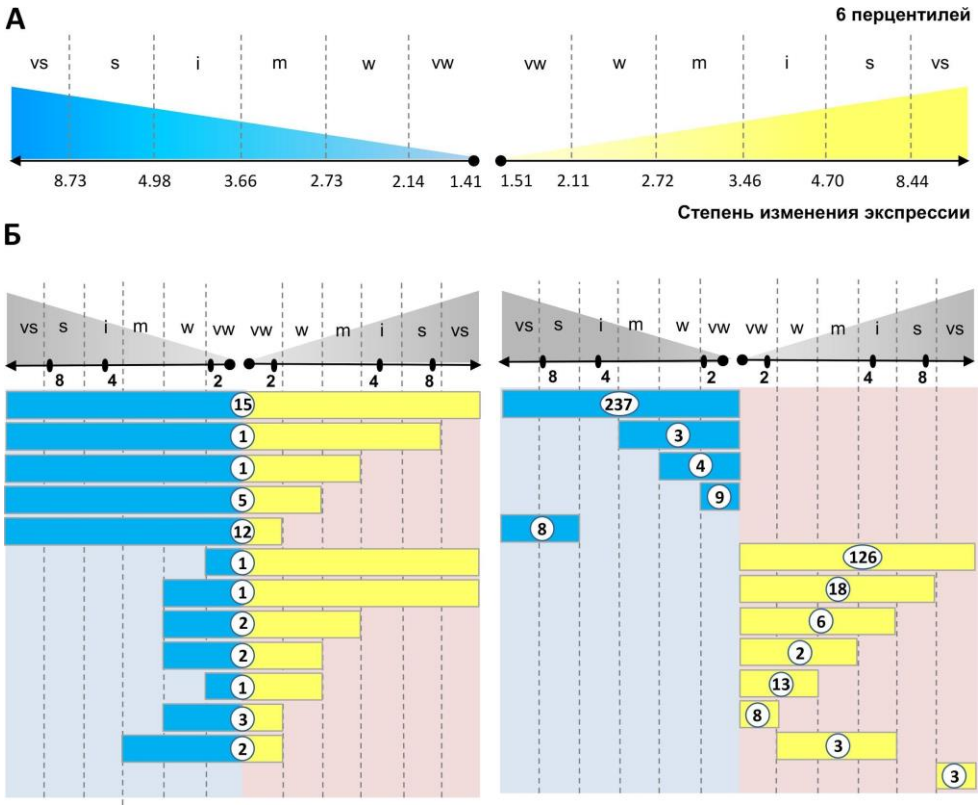

Рисунок 12. А. Интервалы степени изменения экспрессии, полученные разделением наборов дифференциально экспрессирующихся генов на 6 перцентилей (см. Материалы и методы, п. 2.2.2.1). Б. Количество категорий генной онтологии, обогащенных в подгруппах дифференциально экспрессирующихся генов, изменяющих свою экспрессию в определенных интервалах степени изменения экспрессии. Отрицательная регуляция ауксином отмечена синим цветом, положительная регуляция ауксином отмечена желтым цветом. Буквами отмечена степень ответа на ауксин: vs – очень сильный, s – сильный, i – средний, m – умеренный, w – слабый, vw – очень слабый.

3.4.1. Фолд-специфичная регуляция генов, ассоциированных с клеточными компонентами и молекулярными функциями

В результате качественного анализа фолд-специфичных ГО категорий в ответе на ауксин, мы обнаружили, что лишь некоторые из них относятся к словарю молекулярных функций. Например, гены, аннотированные к ГО категориям относящимся к активности факторов трансляции и структуре рибосом, были ассоциированы со слабым повышением экспрессии в ответ на ауксин. ГО категории, относящиеся к связыванию с РНК, оказались обогащенными среди генов с широким диапазоном повышения экспрессии - от очень слабого до сильного. Только гены, кодирующие ферменты с гидролазной активностью, показали снижение уровня экспрессии от очень слабого до среднего. Ауксин влияет на экспрессию генов, связанных с другими молекулярными функциями, но не фолд-специфично.

Среди генов, аннотированных к ГО категориям, относящимся к клеточным компонентам, большая часть показала фолд-специфичное изменение экспрессии в ответ на ауксин (Рис. 13). Не фолд-специфичный ответ показали гены, ассоциированные с аппаратом Гольджи, эндоплазматическим ретикулумом и митохондриями. Мы обнаружили, что ауксин не только однонаправленно изменяет экспрессию генов, ассоциированных с определенными клеточными компонентами, но еще и с примерно одинаковой степенью. Наиболее схожий фолд-специфичный ответ на ауксин показали гены, кодирующие рибосомальные РНК и белки.

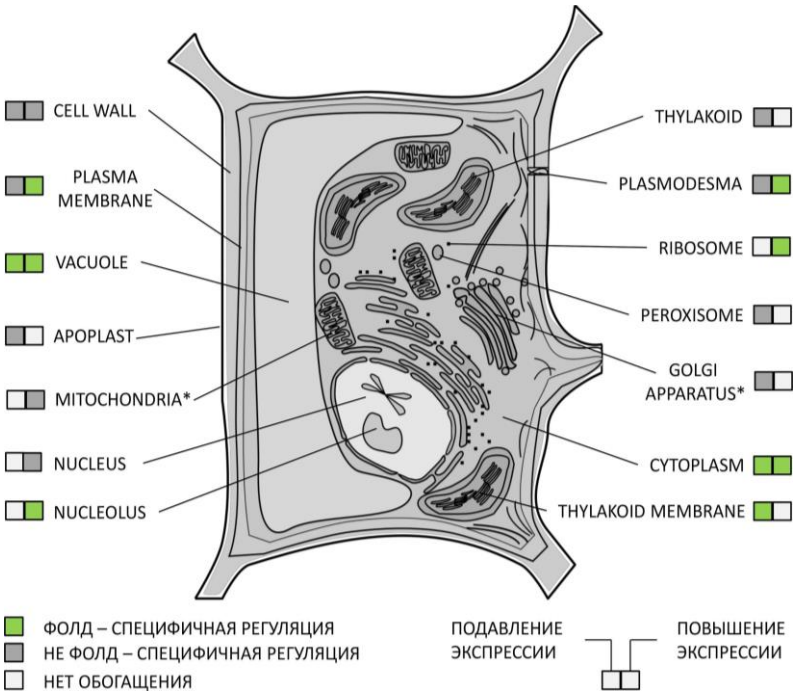

Рисунок 13. Клеточные компоненты, фолд-специфично регулируемые ауксином. Аппарат Гольджи, митохондрии и рибосомы включены в рисунок, так как их биогенез регулируется ауксином**.**

3.4.2. Фолд-специфичная регуляция генов, ассоциированных с биологическими процессами

Мы разделили процессы, регулируемые ауксином, на семь основных групп: ответ на гормональный стимул, процессы клеточного развития, процессы метаболизма в клетке, организация органелл, транспорт, клеточное деление и экспрессия генов (Рис. 14). Ауксин особым образом регулирует только два из этих процессов: не специфично - клеточное деление и очень слабо - экспрессию генов. Такие энергозатратные процессы, как клеточное деление, клеточный цикл, цитокинез, репликация ДНК, организация цитоскелета, хроматина и хромосом были ассоциированы с повышением экспрессии в ответ на ауксин. Для остальных процессов оказалось, что ауксин реализует хорошо сбалансированную и экономичную стратегию, которая, вероятно, позволяет клетке сохранять ресурсы во время деления. А именно: ауксин умеренно подавляет экспрессию генов, ассоциированных с

катаболическими процессами, и не фолд-специфично подавляет вторичный метаболизм, процессы транспорта и множество других. Однако его влияние на первичный метаболизм более сложное и фолд-специфичное. В основном ауксин ингибирует все основные метаболические процессы, за исключением метаболизма нуклеиновых кислот, белков и небольших молекул. Ауксин активирует экспрессию ДЭГ во всех трех процессах таким образом, что большая часть генов, ассоциированных с ними, активируется примерно в одинаковой степени. Такая координация также наблюдалась и для процессов, напрямую связанных с экспрессией генов, так как было выявлено три этапа, которые активировались ауксином фолд-специфично: модификация гистонов (от слабого до среднего уровня), трансляция (до слабого уровня) и метилирование РНК (до сильного уровня). По результатам анализа FSEA для клеточных компонент и молекулярных функций трансляция оказалась наиболее специфично регулируемым процессом среди остальных.

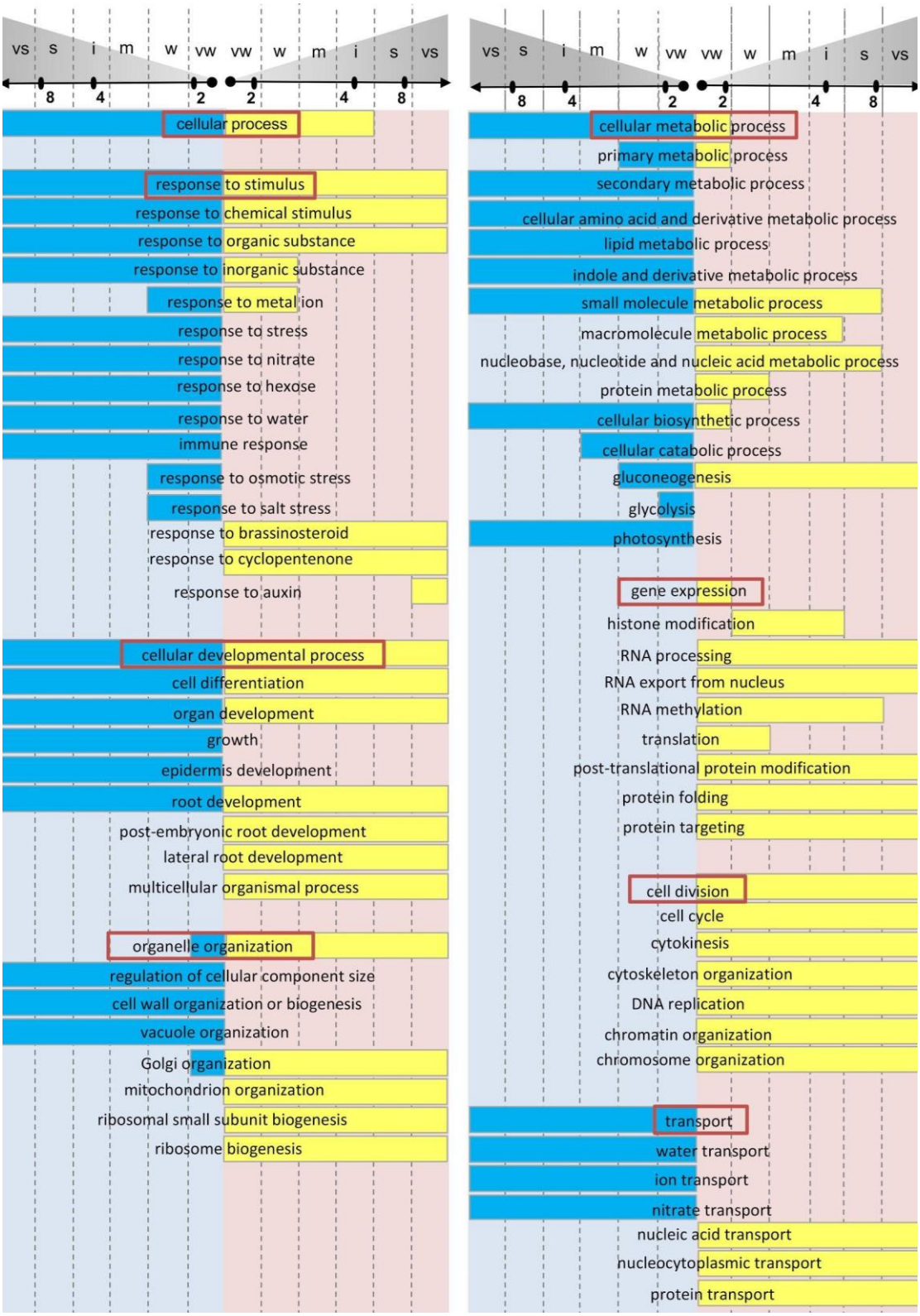

Рисунок 14. Биологические процессы, фолд-специфично регулируемые ауксином**.** Для каждой категории ГО отображен интервал степени изменения экспрессии, в котором экспрессия значительной части генов регулируется ауксином. Наиболее длинные интервалы (от очень слабого до очень сильного) являются не фолд-специфичными. Активация транскрипции

отмечена желтым цветом, подавление транскрипции отмечено синим цветом. В красные рамки помещены общие ГО категории. Буквами отмечена степень ответа на ауксин: vs – очень сильный, s – сильный, i – средний, m – умеренный, w – слабый, vw – очень слабый.

3.4.3. Верификация фолд-специфичной регуляции ауксином на независимых данных

Для подтверждения того, что фолд-специфичная регуляция ауксином не является артефактом, полученным из данных анализируемого RNA-seq эксперимента (Omelyanchuk et al., 2017), мы провели аналогичный анализ на независимых данных полногеномных экспериментов по 4-х, 8-ми и 12-и часовой обработке ауксином корней *Arabidopsis* (Lewis et al., 2013). В результате было выявлено 364 термина, значимо обогащенных в группе дифференциально экспрессирующихся генов, 63% из которых совпали с терминами, обогащенными в нашем RNA-seq эксперименте. Также стоит отметить, что 77% фолд-специфичных терминов, выявленных в независимых экспериментах, совпали с фолд-специфичными терминами, выявленными в нашем RNA-seq эксперименте. Степени изменения экспрессии ДЭГ, выявленные в независимых ДНК-микрочип экспериментах, оказались схожими или были ниже, что можно объяснить тем, что RNA-seq позволяет оценивать большие диапазоны изменения экспрессии, чем ДНК-микрочип эксперименты (Agrawal et al., 2014).

Проведенный нами анализ подтверждает, что транскрипционный ответ на ауксин разделяется на функциональные группы по степени изменения экспрессии генов. Ранее был показан только быстрый и сильный ответ для одной ГО категории – ответа на ауксин (Lavenus et al., 2013; Lewis et al., 2013), в то время как остальные категории и транскрипционный ответ меньшей силы не рассматривался. Данные результаты особенно важны, так как в фолд-специфичном транскрипционном ответе с низким уровнем

изменения экспрессии ДЭГ участвует большое количество генов, ассоциированных с важными физиологическими процессами.

3.4.4. Заключение по главе 3.4.

Проведенный нами анализ показывает, что ауксин реализует сложную стратегию, направленную на экономное распределение ресурсов клетки во время процессов роста и развития. Такая стратегия заключается в поддержании экспрессии энергозатратных процессов на определенном уровне, например, он фолд-специфично и слабо активирует процессы модификации гистонов и трансляции при обширной и не фолд-специфичной активации процессов связанных с делением клетки. Таким образом, разработанный нами метод позволяет анализировать данные транскриптомных экспериментов с большим разрешением чем стандартные методы АФО, выявляя взаимосвязь между функциональностью генов и их транскрипционной активностью.

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

Для апробации FSEA на данных транскриптомного эксперимента, проведенного на клетках человека, мы выбрали данные 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). Известно, что сплайс-вариант AR-V7 ассоциирован с развитием метастаз при раке предстательной железы, поэтому изучение генных сетей, регулируемых AR-V7, является важной задачей для разработки эффективных методов лечения (Magani et al., 2018). Мы проанализировали эти данные при

помощи последней версии метода FSEA (см. главу 2.2.1) с разбиением на 5 квантилей.

3.5.1. Апробация метода на данных эксперимента по экспрессии сплайс-варианта гена AR-V7 в клеточной линии LNCaP

В результате анализа данных по AR-V7, метод FSEA обнаружил 184/197 фолд-специфичных и 167/42 не фолд-специфичных категорий ГО для ДЭГ, активирующих и подавляющих свою экспрессию, соответственно. Большая часть выявленных не фолд-специфичных категорий ГО относилась к обширным изменениям на молекулярном, тканевом уровне и на уровне метаболизма, в то время как некоторые фолд-специфичные категории ГО оказались ассоциированы с процессами, косвенно связанными с переходом к более агрессивной форме рака. Важно отметить, что некоторые из этих категорий ГО, например относящиеся к процессу клеточного цикла (GO:0007049, “cell cycle”), защитного ответа (GO:0006952, “defence response”), репарации ДНК (GO:0006281, “DNA repair”) и клеточной коммуникации (GO:0007154, “cell communication”), не показали значимого обогащения в аннотации для всех дифференциально экспрессирующихся генов, но показали обогащение на отдельных интервалах (Рис. 15). Для детального изучения мы выбрали категорию ГО, описывающую сигнальный путь рецепторов, сопряженных с G-белками (GO:0007187, “G protein-coupled receptor signaling pathway“), ассоциированную с сильной активацией экспрессии генов.

Известно, что рецепторы, связанные с G-белками, играют значительную роль в разнообразных физиологических процессах, на организменном, тканевом и клеточном уровнях (Chou и Elrod, 2002). Анализ полученного набора генов при помощи инструмента GeneMania (Warde-Farley et al., 2010) выявил взаимосвязи между генами, определенные на основе генетических (genetic interactions) и белок-белковых (protein-protein interactions) взаимодействий. Особый интерес вызвал ген сфингозин-1-

фосфатного рецептора (*S1P3*), ассоциированный с процессами ангиогенеза (P. Wang et al., 2019). Нарушение в пути передачи сигнала через *S1P3* рецепторы играет важную роль в переходе к агрессивным формам рака, включая переход к андроген-независимой форме рака предстательной железы (CRPC; Castration Resistant Prostate Cancer) (Brizuela et al., 2014). Однако в публичных базах данных нет информации о связи сигнального пути сфингозин-1-фосфата и экспрессией сплайс варианта *AR-V7*.

Полученный результат говорит о том, что исследователям стоит обращать отдельное внимание на процессы, регулируемые фолд-специфично и детально анализировать вклад ассоциированных генов в изучаемый признак для изучения причин такого координированного поведения.

Рисунок 15. Биологические процессы, ассоциированные с фолд-специфичным изменением экспрессии генов в транскриптомном эксперименте по исследованию экспрессии гена конститутивно активного андрогенового рецептора в клеточной линии рака предстательной железы человека LNCaP (Cottard et al., 2017)**.** Для каждой категории ГО (ось y)

отображен интервал степени изменения экспрессии (ось x), в котором экспрессия генов регулируется фолд-специфично. Активация транскрипции отмечена желтым цветом, подавление транскрипции отмечено синим цветом.

3.5.2. Апробация метода на данных эксперимента по исследованию экспрессии генов в клеточной линии рака предстательной железы человека LNCaP по сравнению с нормальными клетками HPrEC

Сравнение экспрессии генов в нормальных и раковых клетках предполагает выявление значительных отличий в активности генов, отвечающих за самые разные процессы и функции. Поэтому для более подробного анализа результатов работы FSEA мы выбрали эксперимент по сравнению экспрессии генов в клеточной линии рака предстательной железы LNCaP и в нормальных клетках линии HPrEC. Мы также сравнили результаты работы FSEA с результатами двух других широко используемых методов анализа функционального обогащения SEA (Huang et al., 2009) и GSEA (Subramanian et al., 2005). Как выяснилось, результаты работы всех трех методов перекрываются лишь частично (Рис. 16 А, Б), что особенно заметно для категорий ГО, ассоциированных с активацией экспрессии в раковых клетках (Рис. 16 А).

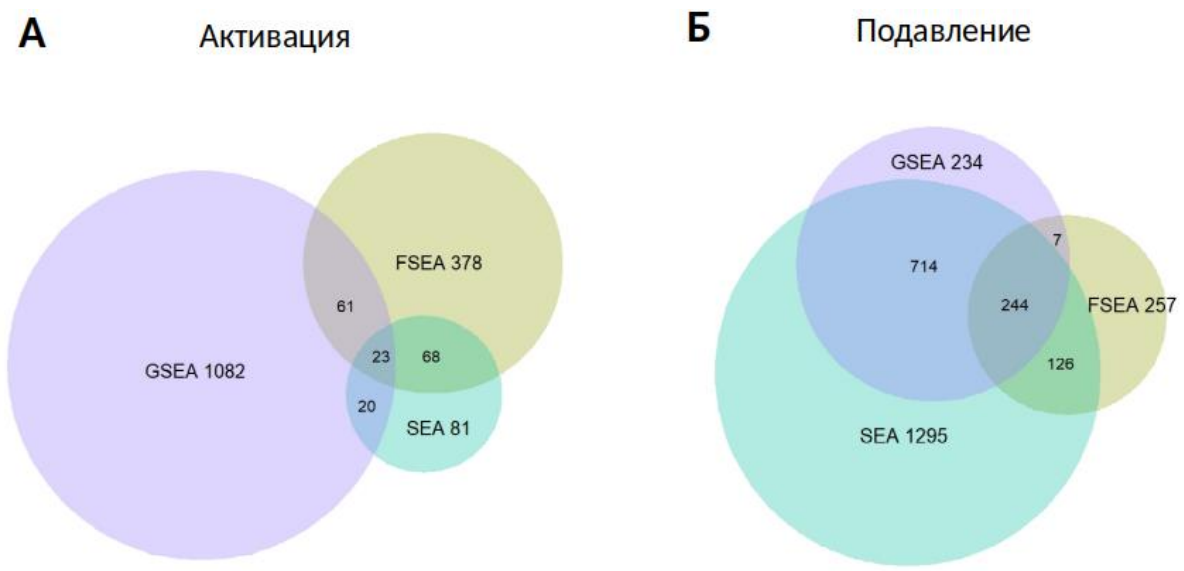

Рисунок 16. Диаграммы Венна, демонстрирующие количества ГО категорий выявленных методами FSEA (бежевый), GSEA (фиолетовый) и SEA (синий) на данных эксперимента по исследованию экспрессии генов в клеточной линии рака предстательной железы LNCaP (GSE70466). (**А**) Количество ГО терминов, выявленных для активируемых генов. (**Б**) Количество ГО терминов, выявленных для подавляемых генов.

Стоит отметить, что разработанный нами метод FSEA проверяет гипотезу о взаимосвязи функции генов с конкретными значениями степени изменения экспрессии в контексте условий эксперимента, в то время как методы GSEA и SEA позволяют выявить связь между функцией генов и исследуемым признаком. Следовательно, ситуация при которой обогащение определенной категорией ГО было обнаружено при помощи методов GSEA или SEA и не было обнаружено при помощи FSEA указывает на то, что гены ассоциированные с данной категорией ГО экспрессируются не скоординированно. Таким образом, метод FSEA может использоваться как отдельно, так и совместно с существующими методами анализа функционального обогащения, дополняя их результаты. Далее мы подробно сравним три группы категорий ГО: выявленных как методом FSEA, так и SEA, только SEA и только FSEA.

3.5.2.1. FSEA и SEA: FSEA уточняет результаты SEA

Рассмотрим более детально результаты применения методов FSEA и SEA. Совместно, оба метода обнаружили 91 и 370 категорий ГО для ДЭГ, повышающих и понижающих свою экспрессию, соответственно. Процессы, описываемые данными ГО категориями, обогащены в списке ДЭГ по сравнению с геномом и также обогащены в определенных интервалах степени изменения экспрессии. Например, при помощи FSEA была обнаружена значимая ассоциация процесса окислительного фосфорилирования (GO:0006119, “oxidative phosphorylation”) со слабой активацией экспрессии генов (Рис. 17 А, Б). Известно, что для клеток рака предстательной железы характерен сдвиг метаболизма в сторону окислительного фосфорилирования (Giannoni et al., 2015). В то время как SEA предполагает, что выявленный биологический процесс связан с развитием рака, метод FSEA показывает, что данный процесс относительно слабо активируется в клетках рака предстательной железы.

В качестве примера еще одной категории ГО, обнаруженной как методом FSEA, так и SEA, можно привести категорию, описывающую процесс морфогенеза кровеносных сосудов (GO:0048514, “blood vessel morphogenesis”) (Рис. 17 В, Г). С помощью метода FSEA было обнаружено подавление экспрессии от среднего до очень сильного уровня (интервал 3-5 при разбиении на 5 квантилей) значимой части генов, ассоциированных с данной категорией ГО. При помощи базы данных The Network of Cancer Genes (NCG) (Repana et al., 2019) 56 из 182 генов, аннотированных к ГО категории GO:0048514 и фолд-специфично подавляющихся в клеточной линии рака LNCaP, были идентифицированы как относящиеся к процессам канцерогенеза. Среди них 9 генов оказались онкосупрессорами, например, гены *FBWX7* (Yeh et al., 2018) и *BAX* (Guo et al., 2000; Liu et al., 2016). Известно, что раковым опухолям свойственны нарушения в нормальном развитии кровеносных сосудов (Forster et al., 2017). Результаты, полученные при помощи метода FSEA, позволили выявить гены, ответственные за

развитие кровеносных сосудов, сильное подавление которых, вероятно, связано с развитием раковой опухоли. Такие сокращенные списки ключевых генов уже могут быть исследованы экспериментально для проверки их участия в развитии рака предстательной железы человека. Помимо перечисленных, метод FSEA обнаружил много других фолд-специфичных категорий ГО, ассоциированных с процессами канцерогенеза.

Таким образом, метод FSEA предоставляет информацию, которая может помочь сократить список кандидатных генов, ответственных за развитие исследуемого фенотипа, путем отбора генов, изменяющих свою экспрессию в определенном интервале степени изменения экспрессии и ассоциированных с определенной категорией ГО.

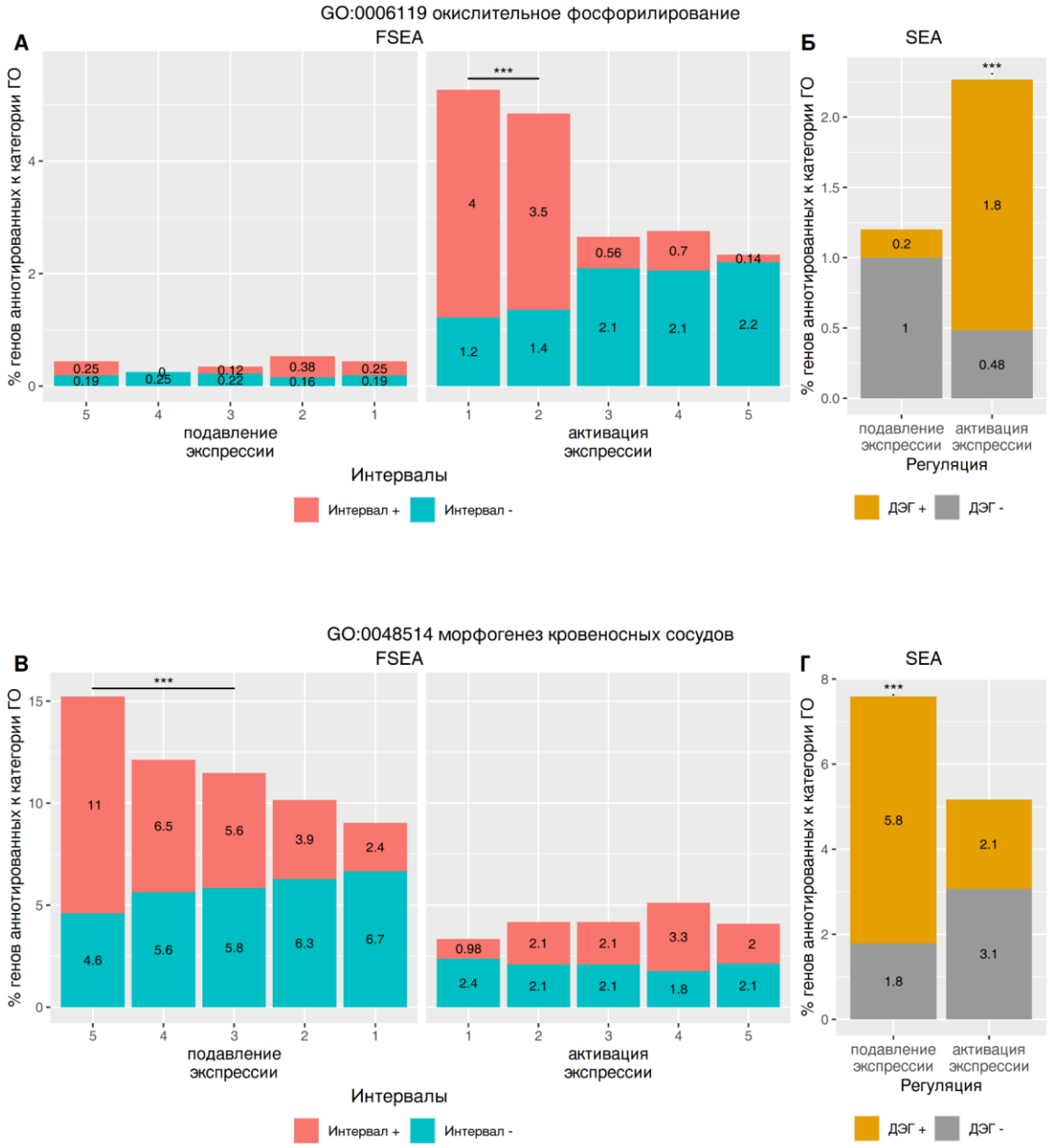

Рисунок 17. Пример категорий ГО, обнаруженных при помощи методов FSEA и SEA на данных эксперимента по исследованию экспрессии генов в клеточной линии рака предстательной железы LNCaP (GSE70466). (А, Б) окислительное фосфорилирование (GO:0006119, “oxidative phosphorylation”); (В, Г) морфогенез кровеносных сосудов (GO:0048514, “blood vessel morphogenesis”). (А, В) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в определенном интервале (ось x) степени изменения экспрессии (красный) и вне данного интервала (синий). (Б, Г) Гистограмма, отображающая процент генов (ось y), аннотированных к

категории ГО в ДЭГ (желтый), активирующих или подавляющих свою экспрессию (ось x), и вне ДЭГ (серый).

3.5.2.2. Функциональные группы генов не обнаруженные методом FSEA: нескоординированный ответ

Обогащение ГО категорий, обнаруженное методом SEA, и не обнаруженное методом FSEA означает, что гены, аннотированные к данной категории ГО, экспрессируются нескоординированно. Кроме того мы обнаружили, что ГО категории, обнаруженные только методом SEA, часто являются очень общим описанием многокомпонентного сложного процесса.

Типичный профиль экспрессии для функциональной группы генов, обнаруженной методом SEA и не обнаруженной методом FSEA, можно показать на примере категории ГО, описывающей процесс катаболизма лизина (GO:0006554, “lysine catabolic process”) (Рис. 18 А, Б). Распределение частот ДЭГ, аннотированных к данной категории ГО, на всех квантилях приблизительно равномерное, что не мешает генам, относящимся к данному процессу, быть обогащенными в списке ДЭГ по сравнению с полным геномом. И, действительно, нескоординированное усиление деградации лизина в клетках рака предстательной железы было показано за счет детекции повышения уровня его метаболитов в опухолевых тканях (Ren et al., 2016).

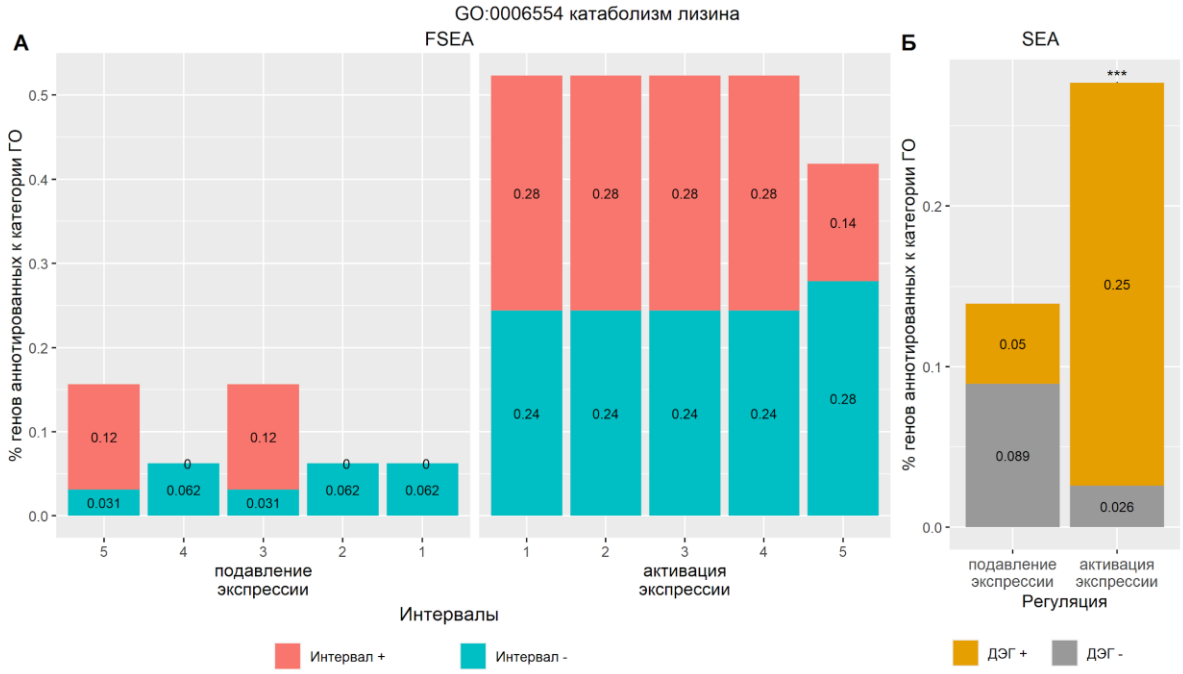

Рисунок 18. Пример категории ГО, обнаруженной только при помощи метода SEA - катаболизм лизина (GO:0006554, “lysine catabolic process”). (**А**) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в определенном интервале (ось x) степени изменения экспрессии (красный) и вне данного интервала (синий). (**Б**) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в ДЭГ (желтый), активирующих или подавляющих свою экспрессию (ось x) и вне ДЭГ (серый).

3.5.2.3. Функциональные группы генов, обнаруженные только методом FSEA: Дискретизированный ответ, невидимый для классических методов функционального обогащения

Метод FSEA обнаружил 439 и 264 ГО категории для активируемых и подавляемых ДЭГ, соответственно, которые не были обнаружены методом SEA и многие из которых косвенно или напрямую связаны процессами канцерогенеза.

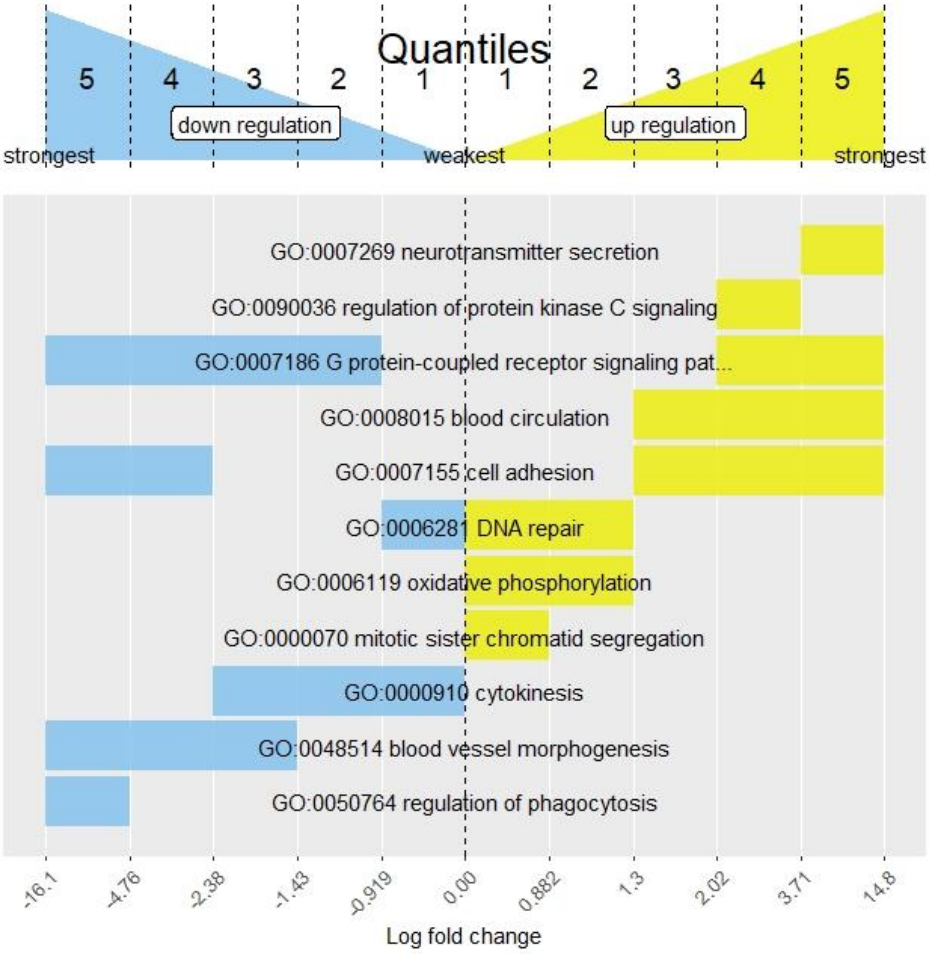

Рисунок 19. Выходная гистограмма веб-сервиса FoldGO (https://webfsgor.sysbio.cytogen.ru), отображающая категории ГО, ассоциированные с процессами канцерогенеза, выявленные на данных эксперимента по исследованию экспрессии генов в клеточной линии рака предстательной железы LNCaP (GSE70466). Данные категории ГО, выявленные только при помощи метода FSEA, показали наибольшую статистическую значимость обогащения в определенных интервалах степени изменения экспрессии.

На рисунке (Рис. 19) изображены профили фолд-специфичной экспрессии генов, ассоциированных с процессами, которые мы отобрали для демонстрации категорий ГО, обнаруженных только методом FSEA. В качестве примера рассмотрим категорию ГО, описывающую процесс регуляции фагоцитоза (GO:0050764, “regulation of phagocytosis”) (Рис. 20 А,

Б). Метод FSEA обнаружил значимое отклонение экспрессии генов, аннотированных к данной категории ГО, в сторону сильного подавления экспрессии. Так как данная категория ГО была обогащена в узком интервале степени изменения экспрессии и не была обогащена по сравнению с геномом, она не была обнаружена методом SEA как обогащенная в списке ДЭГ, снизивших свою экспрессию. Фагоцитоз давно известен как процесс, тесно связанный с развитием опухоли (Schcolnik-Cabrera et al., 2019) и только метод FSEA оказался способен выявить функциональность данного процесса в клеточной линии рака предстательной железы LNCaP.

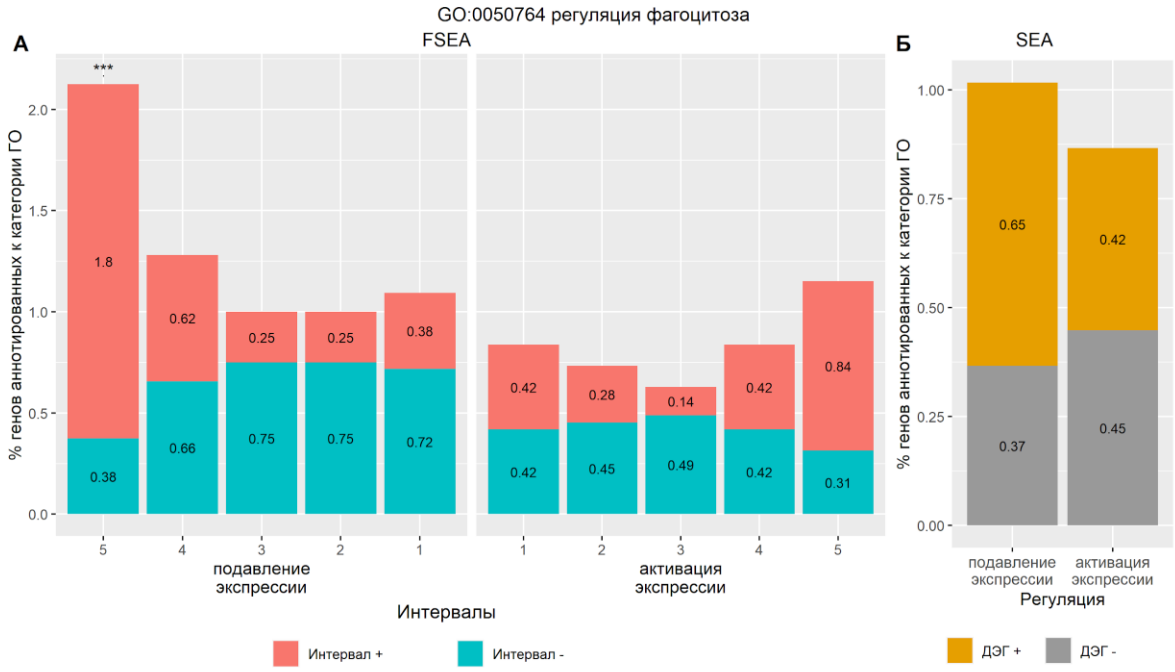

Рисунок 20. Пример категории ГО, обнаруженной только при помощи метода FSEA - регуляция фагоцитоза (GO:0050764, “regulation of phagocytosis”). (**А**) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в определенном интервале (ось x) степени изменения экспрессии (красный) и вне данного интервала (синий). (**Б**) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в ДЭГ (желтый), активирующих или подавляющих свою экспрессию (ось x) и вне ДЭГ (серый).

В качестве еще одного примера ГО категории, обнаруженной при помощи метода FSEA и не обнаруженной методом SEA, можно привести категорию, описывающую процесс секреции нейромедиатора (GO:0007269, “neurotransmitter secretion”), которая оказалась значимо ассоциирована с очень сильной активацией экспрессии, аннотированных к ней генов (Рис. 21 А, Б).

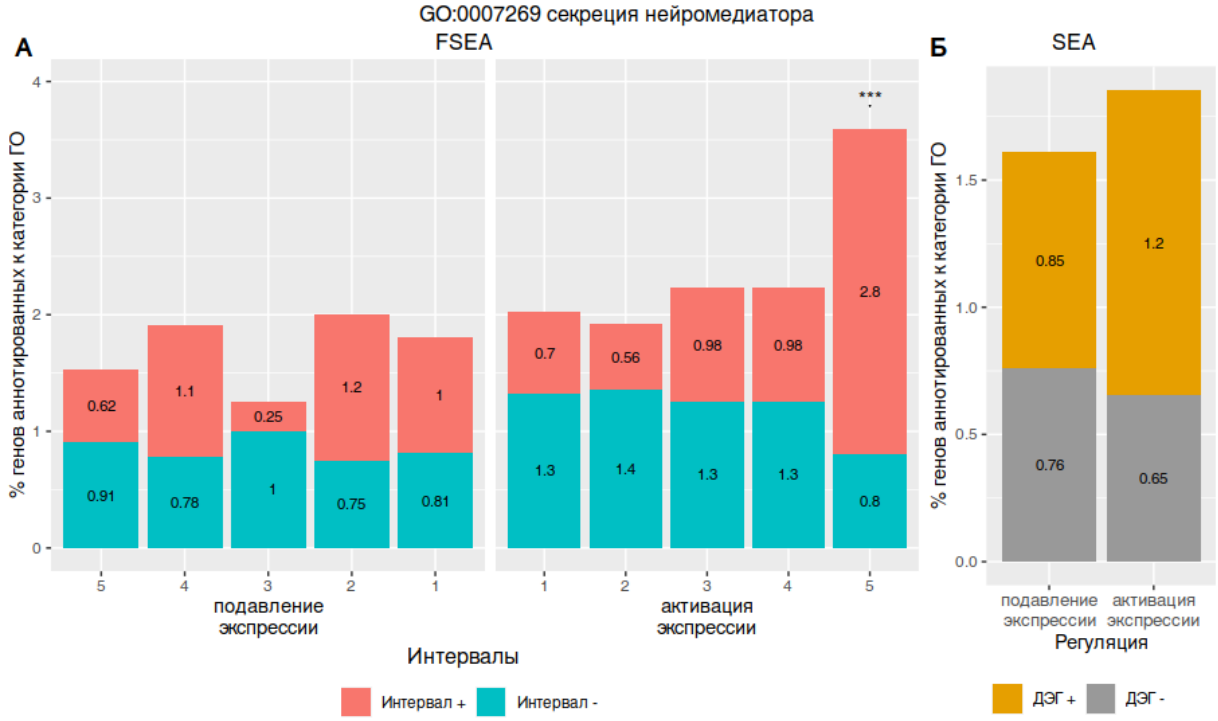

Рисунок 21. Пример категории ГО, обнаруженной только при помощи метода FSEA - секреция нейромедиатора, (GO:0007269, “neurotransmitter secretion”). (**А**) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в определенном интервале (ось x) степени изменения экспрессии (красный) и вне данного интервала (синий). (**Б**) Гистограмма, отображающая процент генов (ось y), аннотированных к категории ГО в ДЭГ (желтый), активирующих или подавляющих свою экспрессию (ось x) и вне ДЭГ (серый).

Совсем недавно изучение роли нейромедиаторов в прогрессии раковой опухоли стало одним из направлений, которому уделяется пристальное

внимание научного сообщества (Kissick et al., 2015). В качестве примера, демонстрирующим связь канцерогенеза и нейромедиаторов, можно привести онкоген *ERG* из группы генов обнаруженных методом FSEA как связанные с секрецией нейромедиаторов и сильно повышающих свою экспрессию в клетках LNCAP. Транскрипционный фактор кодируемый геном *ERG* вызывает сверхэкспрессию никтоинового ацетилхолинового рецептора (*nAChRs*) в клетках рака предстательной железы, который, в свою очередь, при воздействии никотина индуцирует пролиферацию раковых клеток (Kissick et al., 2015).

Таким образом, детальный анализ категорий ГО обнаруженных только методом FSEA показал, что биологические процессы, которые не обогащены в списке ДЭГ и не выявлены классическими методами функциональной аннотации, могут иметь специфичный профиль экспрессии и могут быть важны для лучшего понимания молекулярных механизмов, лежащих в основе изучаемого явления.

3.6. Применение метода FSEA к различным характеристикам генов

Метод FSEA является универсальным и может быть применен не только к Генной Онтологии, но и к любым другим категориальным характеристикам генов, таким, например, как наличие сайта связывания в промоторе или участие в сигнальном пути. Мы применили метод FSEA для анализа обогащения ДНК-мотивами и наборами онкогенов в данных по исследованию экспрессии сплайс-варианта *AR-V7* в клеточной линии LNCaP (GSE71334) (Cottard et al., 2017). Наборы генов C3, содержащие определенные мотивы, и наборы генов C6, ассоциированные с процессами канцерогенеза, были взяты из базы данных MSigDB (Liberzon et al., 2011). В результате метод FSEA выявил слабую активацию генов, содержащих в своих промоторных районах сайты связывания белков NRF1 и ELK1.

Транскрипционный фактор ELK1 связан с различными патологическими состояниями и связывается с сайтом связывания SRE (Serum Response Element) в промоторе протоонкогена *c-Fos* (Y. Chai et al., 2001). Примечательно, что транскрипционный фактор NRF1 участвует в процессе трансактивации гена андрогенового рецептора *AR* (Schultz et al., 2014). Среди результатов анализа, проведенного с помощью метода FSEA, на наборах генов, связанных с канцерогенезом, наиболее интересным было обнаружение ассоциации генов, связанных с экспрессией энхансер-связывающего фактора LEF1 (Lymphoid Enhancer-binding factor 1) с очень высокой степенью изменения экспрессии. Известно, что сверхэкспрессия фактора LEF1 стимулирует развитие опухоли по андроген-независимому пути и может привести к кастрационно-резистентному раку предстательной железы (CRPC) (Y. Li et al., 2009). Все наборы генов, для которых было выявлено фолд-специфическое обогащение в списке ДЭГ, представлены в приложениях в Таблице П1.

4. Заключение

Тысячи генов скоординированно действуют, объединяясь в генные сети для обеспечения жизненно важных процессов и функций в клетках живого организма. В настоящий момент для изучения таких сетей активно используются данные полученные из транскриптомных экспериментов. При помощи транскриптомных экспериментов мы можем достаточно точно оценить с какой силой активируются или подавляются гены в ответ на определенный фактор, а используя анализ функционального обогащения с какой функцией они связаны. Как правило, информацию о силе экспрессии генов используют лишь для выявления дифференциально-экспрессирующихся генов (ДЭГ). Однако, пренебрегая информацией о схожести степеней изменения экспрессии генов объединенных одной функцией и анализируя ДЭГ как единое целое мы упускаем важную информацию об исследуемом объекте.

В рамках данной работы мы разработали метод анализа фолд-специфичного обогащения (FSEA), который позволяет выявлять ГО категории, ассоциированные с определенным интервалом степени изменения экспрессии (фолд-специфичные ГО категории). Данный метод был реализован в виде пакета FoldGO для языка программирования R и веб-сервиса. Апробация метода была проведена на данных RNA-seq (Omelyanchuk et al., 2017) и ДНК-микрочип (Lewis et al., 2013) экспериментов по обработке ауксином корней *Arabidopsis thaliana*, а также по исследованию экспрессии генов в клеточной линии аденокарциномы предстательной железы (LNCaP) по сравнению с нормальными клетками эпителия предстательной железы (клеточная линия HPrEC) и по сравнению с клетками с экспрессией конститутивно активного андрогенового рецептора *AR-V7* (Cottard et al., 2017).

Были произведены оценки доли ложноположительных результатов, которую дает применение метода FSEA на данных транскриптомных

экспериментов, и чувствительности метода. Оценка доли ложноположительных результатов показала, что при правильном выборе параметров для процедуры коррекции на множественное тестирование метод FSEA показывает долю ложноположительных результатов меньше 5 %. Оценка чувствительности метода показала, что FSEA обнаруживает более 70 % групп генов, состоящих из 10 и более генов со схожей степенью изменения экспрессии.

Метод FSEA был протестирован в анализе десятков разных транскриптомов. При условии достаточного количества ДЭГ этот метод всегда находил фолд-специфичные ГО категории. Более подробный анализ фолд-специфичных категорий был проведен с тремя разными транскриптомами. В результате апробации на данных эксперимента по обработке ауксином корней *Arabidopsis thaliana* (Omelyanchuk et al., 2017) было выявлено 225 ГО категорий, ассоциированных с повышением экспрессии ДЭГ, и 307 ГО категорий, ассоциированных с понижением экспрессии ДЭГ, среди которых 82 категории (36%) и 36 категории (12%) соответственно показали ассоциацию с фолд-специфичным изменением экспрессии. Детальный анализ выявленных ГО категорий показал, что механизмы действия ауксина очень сбалансированы. Можно сказать, что обработка ауксином позволяет сохранять жизненно важные ресурсы в период активного роста и развития тканей корня. Ауксин фолд-специфично подавляет процессы катаболизма и активирует процессы первичного метаболизма. Также фолд-специфичная регуляция была выявлена для процессов, связанных с экспрессией генов, таких как модификация гистонов и трансляция, причем последний оказался наиболее специфично регулируемым ауксином среди остальных процессов. При верификации на независимых данных результат анализа подтвердился. Результаты апробации на данных RNA-seq эксперимента по обработке ауксином корней *Arabidopsis*

thaliana опубликованы в журнале Nature Scientific Reports (Omelyanchuk et al., 2017).

Апробация метода FSEA на данных экспериментов по исследованию экспрессии генов в клеточной линии аденокарциномы предстательной железы (LNCaP) и сравнение результатов анализа с результатами, полученными при помощи методов SEA (Huang et al., 2009) и GSEA (Subramanian et al., 2005), выявила более 1000 категорий ГО, значимо обогащенных в определенных интервалах степени изменения экспрессии, большая часть из которых не была обнаружена при помощи классических подходов - SEA и GSEA. Мы установили, что помимо самостоятельного применения, FSEA дополняет классические подходы к анализу функционального обогащения, предоставляя важную информацию о скоординированности экспрессии функционально связанных генов. Качественный анализ ГО категорий, обнаруженных методом FSEA, показал, что многие из них тесно связаны с процессом канцерогенеза. Например, только при помощи метода FSEA была обнаружена слабая активация экспрессии генов, ассоциированных с сегрегацией сестринских хроматид (Pallai et al., 2015) в процессе митоза, и высокий уровень активации генов, ассоциированных с регуляцией передачи сигнала при помощи протеинкиназы С (Cornford et al., 1999; Tanaka et al., 2003). Стоит отметить, что данные категории ГО не были выявлены при помощи классического анализа функционального обогащения. Данные результаты и детальное описание метода FSEA опубликованы в журнале MDPI Genes (Wiebe et al., 2020).

Результаты полученные в рамках данной работы показывают, что разработанный нами метод FSEA достоверно выявляет биологические процессы, у которых гены скоординированно изменяют свою экспрессию в ответ на исследуемый фактор, даже в тех случаях когда исследуемые процессы не обогащены в ДЭГ. Таким образом, метод FSEA может как

дополнять существующие методы анализа функционального обогащения, так и использоваться в качестве самостоятельного инструмента, предоставляя исследователям возможность выявлять закономерности, релевантные исследуемому фактору.

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 литературы

1. Agrawal, A. A., McLaughlin, K. J., Jenkins, J. L., & Kielkopf, C. L. Structure-guided U2AF^{65} variant improves recognition and splicing of a defective pre-mRNA // Proceedings of the National Academy of Sciences. - 2014 - Т. - № 111(49) - С.17420–17425.

2. Ala, U., Piro, R. M., Grassi, E., Damasco, C., Silengo, L., Oti, M., Provero, P., & Di Cunto, F. Prediction of Human Disease Genes by Human-Mouse Conserved Coexpression Analysis // PLoS Computational Biology - Т. 4 - № 3

3. Alexa, A., & Rahnenfuhrer, J. *t*opGO: Enrichment Analysis for Gene Ontology // Bioconductor - R package version 2.38.1 - 2019.

4. Andrews, S. FastQC: a quality control tool for high throughput sequence data [Электронный ресурс] // 2010 - режим доступа: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

5. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. . Gene Ontology: Tool for the unification of biology // Nature Genetics - 2000 - Т. 25 - № 1 - С.25–29.

6. Auer, P. L., & Doerge, R. W. A Two-Stage Poisson Model for Testing RNA-Seq Data // Statistical Applications in Genetics and Molecular Biology - 2011 - Т. 10 - № 1.

7. Balashanmugam, M. V., Shivanandappa, T. B., Nagarethinam, S., Vastrad, B., & Vastrad, C. (). Analysis of Differentially Expressed Genes in Coronary Artery Disease by Integrated Microarray Analysis // Biomolecules - 2019 -

T. 10 - № 1 - C. 35.

8. Belmonte, M. F., Kirkbride, R. C., Stone, S. L., Pelletier, J. M., Bui, A. Q., Yeung, E. C., Hashimoto, M., Fei, J., Harada, C. M., Munoz, M. D., Le, B. H., Drews, G. N., Brady, S. M., Goldberg, R. B., & Harada, J. J. Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed // Proceedings of the National Academy of Sciences of the United States of America - 2013 - T. 110 - № 5 - C. 435-444

9. Benjamini, Y., & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing // Journal of the Royal Statistical Society: Series B (Methodological) - 1995 - T. 57 - №1 C. 289–300

10. Benjamini, Y., & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency // The Annals of Statistics - 2001 - T. 29 - № 4 - C. 1165–1188.

11. Bland, J. M., & Altman, D. G. Statistics notes: Multiple significance tests: the Bonferroni method // BMJ - 1995 - T. 310 - № 6973 - C. 170.

12. Brizuela, L., Martin, C., Jeannot, P., Ader, I., Gstalder, C., Andrieu, G., Bocquet, M., Laffosse, J.-M., Gomez-Brouchet, A., Malavaud, B., Sabbadini, R. A., & Cuvillier, O. Osteoblast-derived sphingosine 1-phosphate to induce proliferation and confer resistance to therapeutics to bone metastasis-derived prostate cancer cells // Molecular Oncology - 2014 - T. 8 - № 7 - C. 1181–1195.

13. Carlson, M. Org.Hs.eg.db // Bioconductor - 2017

14. Céraline, J., Cruchant, M.D., Erdmann, E., Erbs, P., Kurtz, J.E., Duclos, B., Jacqmin, D., Chopin, D., Bergerat, J.P. Constitutive activation of the androgen receptor by a point mutation in the hinge region: a new mechanism

for androgen-independent growth in prostate cancer // Int J Cancer - 2004 - T. 108 - № 1 - C.152-7.

15. Chai, N., Haney, M. S., Couthouis, J., Morgens, D. W., Benjamin, A., Wu, K., Ousey, J., Fang, S., Finer, S., Bassik, M. C., & Gitler, A. D. Genome-wide synthetic lethal CRISPR screen identifies FIS1 as a genetic interactor of ALS-linked C9ORF72 // Brain Research - 2020 - T.1728 - C.146601.

16. Chai, Y., Chipitsyna, G., Cui, J., Liao, B., Liu, S., Aysola, K., Yezdani, M., Reddy, E. S., & Rao, V. N. C-Fos oncogene regulator Elk-1 interacts with BRCA1 splice variants BRCA1a/1b and enhances BRCA1a/1b-mediated growth suppression in breast cancer cells // Oncogene - 2001 - T. 20 - № 11 - C. 1357–1367.

17. Chen, Y., Souaiaia, T., & Chen, T. PerM: Efficient mapping of short sequencing reads with periodic full sensitive spaced seeds // Bioinformatics - 2009 - T. 25 - № 19 - C. 2514–2521.

18. Chou, K.-C., & Elrod, D. W. Bioinformatical analysis of G-protein-coupled receptors // Journal of Proteome Research - 2002 - T. 1 - №5 - C. 429–433.

19. Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., & Grimmond, S. M. Stem cell transcriptome profiling via massive-scale mRNA sequencing // Nature Methods - 2008 - T. 5 - №7 -C. 613–619.

20. Clough, E., & Barrett, T. The Gene Expression Omnibus Database. In E. Mathé & S. Davis (Eds.) // Statistical Genomics - 2016 - T. 1418 C. 93–110.

21. Cornford, P., Evans, J., Dodson, A., Parsons, K., Woolfenden, A., Neoptolemos, J., & Foster, C. S. Protein Kinase C Isoenzyme Patterns Characteristically Modulated in Early Prostate Cancer // The American

Journal of Pathology - 1999 - T. 154 - № 1 - C.137–144.

22. Cottard, F., Madi-Berthélémy, P. O., Erdmann, E., Schaff-Wendling, F., Keime, C., Ye, T., Kurtz, J.-E., & Céraline, J. Dual effects of constitutively active androgen receptor and full-length androgen receptor for N-cadherin regulation in prostate cancer // Oncotarget - 2017 - T. 8 - № 42

23. Edgar, R., Domrachev, M., & Lash, A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository // Nucleic Acids Research - 2002 - T. 30 - №1 - C. 207–210.

24. Falcon, S., & Gentleman, R. Using GOstats to test gene lists for GO term association // Bioinformatics - 2007 - T. 23 - №2 - C. 257–258.

25. Forster, J., Harriss-Phillips, W., Douglass, M., & Bezak, E. A review of the development of tumor vasculature and its effects on the tumor microenvironment // Hypoxia - 2017 - T. 5 C. 21–32.

26. Gehan, M. A., Park, S., Gilmour, S. J., An, C., Lee, C.-M., & Thomashow, M. F. Natural variation in the C-repeat binding factor cold response pathway correlates with local adaptation of Arabidopsis ecotypes // The Plant Journal: For Cell and Molecular Biology - 2015 - T. 84 - № 4 - C. 682–693.

27. Gelmann, E.P. Molecular biology of the androgen receptor // J Clin Oncol - 2002 - T. 20 - № 13 - C. 3001-15.

28. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements // Nucleic Acids Res. - 2010. - T. 331- №. 5.

29. Gene Ontology Consortium. Gene Ontology annotations and resources // Nucleic Acids Res. - 2013. - T. 530 - №.5

30. Giannoni, E., Taddei, M. L., Morandi, A., Comito, G., Calvani, M., Bianchini, F., Richichi, B., Raugei, G., Wong, N., Tang, D., & Chiarugi, P. Targeting stromal-induced pyruvate kinase M2 nuclear translocation impairs

oxphos and prostate cancer metastatic spread // Oncotarget - 2015 - T. 6 - № 27 - C. 24061–24074.

31. Grossmann, S., Bauer, S., Robinson, P. N., & Vingron, M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis // Bioinformatics - 2007 - T. 23 - № 22 - C. 3024–3031.

32. Guo, B., Cao, S., Tóth, K., Azrak, R. G., & Rustum, Y. M. Overexpression of Bax Enhances Antitumor Activity of Chemotherapeutic Agents in Human Head and Neck Squamous Cell Carcinoma // Clinical Cancer Research - 2000 - T. 6 - № 2 - C. 718–724.

33. Guo, Z., Yang, X., Sun, F., Jiang, R., Linn, D.E., Chen, H., Chen, H., Kong, X., Melamed, J., Tepper, C.G., Kung, H.J., Brodie, A.M., Edwards, J., Qiu, Y. A novel androgen receptor splice variant is up-regulated during prostate cancer progression and promotes androgen depletion-resistant growth // Cancer Res - 2009 - T. 69 - № 6 - C. 2305-13

34. Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification // Cell Reports - 2012 - T. 2 - № 3 - C. 666–673.

35. Heller, M. J. DNA Microarray Technology: Devices, Systems, and Applications // Annual Review of Biomedical Engineering - 2002 - T. 4 - № 1 - C. 129–153.

36. Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J. F., Zhu, J.-K., Cushman, J. C., Gollery, M., & Girke, T. Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis // Plant Physiology - 2008 - T. 147 - № 1 - C. 41–57.

37. Horoszewicz, J.S., Leong, S.S., Kawinski, E., Karr, J.P., Rosenthal, H., Chu, T.M., Mirand, E.A., Murphy, G.P. LNCaP model of human prostatic carcinoma // Cancer Res - 1983 - T. 43 - № 4 - C. 1809-18.

38. Huang, C.G., Li, F.X., Pan, S., Xu, C.B., Dai, J.Q., Zhao, X.H. Identification of genes associated with castration-resistant prostate cancer by gene expression profile analysis // Mol Med Rep - 2017 - Т. 16 - № 5 - С. 6803-6813.

39. Huang, D. W., Sherman, B. T., & Lempicki, R. A. . Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists // Nucleic Acids Research - 2009 - Т. 37 - № 1 - С. 1–13.

40. Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., & Lempicki, R. A. DAVID Bioinformatics Resources: Expanded annotation database and novel algorithms to better extract biology from large gene lists // Nucleic Acids Research - 2007 - Т. 35 - С. 169–175.

41. Jia, Y., Ding, Y., Shi, Y., Zhang, X., Gong, Z., & Yang, S. The cbfs triple mutants reveal the essential functions of CBFs in cold acclimation and allow the definition of CBF regulons in Arabidopsis. The New Phytologist - 2016 - Т. 212 - № 2 - С. 345–353.

42. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., Wu, A. Y. A local search approximation algorithm for k-means clustering // Computational Geometry - 2004 - Т. 28 - № 2 - С. 89–112.

43. Kissick, H. T., On, S. T., Dunn, L. K., Sanda, M. G., Asara, J. M., Pellegrini, K. L., Noel, J. K., & Arredouani, M. S. The transcription factor ERG increases expression of neurotransmitter receptors on prostate cancer cells // BMC Cancer - 2015 - Т. 15 - № 1 - С. 604.

44. Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. GOATOOLS: A Python library for Gene Ontology analyses // Scientific

Reports - 2018 - T. 8 - № 1 - C.10872.

45. Kuppuswamy, U., Ananthasubramanian, S., Wang, Y., Balakrishnan, N., & Ganapathiraju, M. K. Predicting gene ontology annotations of orphan GWAS genes using protein-protein interactions // Algorithms for Molecular Biology - 2014 - T. 9 - №1 - C. 10.

46. Lai, K.P., Yamashita, S., Huang, C.K., Yeh, S., Chang, C. Loss of stromal androgen receptor leads to suppressed prostate tumourigenesis via modulation of pro-inflammatory cytokines/chemokines // EMBO Mol Med - 2012 - T. 4 - № 8 - C.791-807.

47. Langfelder, P., & Horvath, S. WGCNA: An R package for weighted correlation network analysis // BMC Bioinformatics - 2008 - T. 9 - № 1 -C. 559.

48. Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // Genome Biology - 2009 - T. 10 - № 3 - C. 25.

49. Lavenus, J., Goh, T., Roberts, I., Guyomarc'h, S., Lucas, M., De Smet, I., Fukaki, H., Beeckman, T., Bennett, M., & Laplaze, L. Lateral root development in Arabidopsis: Fifty shades of auxin // Trends in Plant Science - 2013 - T. 18 - № 8 - C. 450–458.

50. Lewis, D. R., Olex, A. L., Lundy, S. R., Turkett, W. H., Fetrow, J. S., & Muday, G. K. A Kinetic Analysis of the Auxin Transcriptome Reveals Cell Wall Remodeling Proteins That Modulate Lateral Root Development in Arabidopsis // The Plant Cell - 2013 - T. 25 - № 9 - C. 3329–3346.

51. Li, H., & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics - 2009 - T. 25 - № 14 - C. 1754–1760.

52. Li, H., Ruan, J., & Durbin, R. Mapping short DNA sequencing reads and

calling variants using mapping quality scores // Genome Research - 2008 - Т. 18 - № 11 - С. 1851–1858.

53. Li, W.-X., He, K., Tang, L., Dai, S.-X., Li, G.-H., Lv, W.-W., Guo, Y.-C., An, S.-Q., Wu, G.-Y., Liu, D., & Huang, J.-F. Comprehensive tissue-specific gene set enrichment analysis and transcription factor analysis of breast cancer by integrating 14 gene expression datasets // Oncotarget - 2017 - Т. 8 - № 4 - С. 6775–6786.

54. Li, Y., Wang, L., Zhang, M., Melamed, J., Liu, X., Reiter, R., Wei, J., Peng, Y., Zou, X., Pellicer, A., Garabedian, M. J., Ferrari, A., & Lee, P. LEF1 in androgen-independent prostate cancer: Regulation of androgen receptor expression, prostate cancer growth, and invasion // Cancer Research - 2009 - Т. 69 - № 8 - С. 3332–3338.

55. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. Molecular signatures database (MSigDB) 3.0 // Bioinformatics - 2011 - Т. 27 - № 12 - С. 1739–1740.

56. Liesecke, F., De Craene, J.-O., Besseau, S., Courdavault, V., Clastre, M., Vergès, V., Papon, N., Giglioli-Guivarc'h, N., Glévarec, G., Pichon, O., & Dugé de Bernonville, T. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation // Scientific Reports - 2019 - Т. 9 - № 1 - С. 14431.

57. Liu, Z., Ding, Y., Ye, N., Wild, C., Chen, H., & Zhou, J. Direct Activation of Bax Protein for Cancer Therapy: DIRECT ACTIVATION OF Bax FOR CANCER THERAPY // Medicinal Research Reviews - 2016 - Т. 36 - № 2 - С. 313–341.

58. Ljung K, Hull AK, Celenza J, Yamada M, Estelle M, Normanly J, Sandberg G. Sites and regulation of auxin biosynthesis in Arabidopsis roots // Plant Cell - 2005 - Т. 17 - № 4 - С. 1090-104.

59. Love, M. I., Huber, W., & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 // Genome Biology - 2014 - T. 15 - № 12 - C. 550.

60. Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. Transcriptomics technologies // PLOS Computational Biology - 2017 - T. 13 - № 5.

61. Maere, S., Heymans, K., & Kuiper, M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks // Bioinformatics - 2005 - T. 21 - № 16 - C. 3448–3449.

62. Magani, F., Bray, E. R., Martinez, M. J., Zhao, N., Copello, V. A., Heidman, L., Peacock, S. O., Wiley, D. J., D'Urso, G., & Burnstein, K. L. Identification of an oncogenic network with prognostic and therapeutic value in prostate cancer // Molecular Systems Biology - 2018 - T. 14 - №8

63. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays // Genome Research - 2008 - T. 18 - № 9 - C. 1509–1517.

64. Massey, F. J. The Kolmogorov-Smirnov Test for Goodness of Fit // Journal of the American Statistical Association - 1951 - T. 46 - № 253 - C. 68–78.

65. Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools // Nucleic Acids Research - 2019 - T. 47 - № 1 - C. 419–426.

66. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq // Nature Methods - 2008 - T. 5 - № 7 - C. 621–628.

67. Nelson, N. J. Microarrays Have Arrived: Gene Expression Tool Matures //

JNCI Journal of the National Cancer Institute - 2001 - T. 93 - № 7 - C. 492–494.

68. Nie, Y., Yu, S., Li, Q., Nirala, N. K., Amcheslavsky, A., Edwards, Y. J. K., Shum, P. W., Jiang, Z., Wang, W., Zhang, B., Gao, N., & Ip, Y. T. Oncogenic Pathways and Loss of the Rab11 GTPase Synergize To Alter Metabolism in Drosophila // Genetics - 2019 - T. 212 - № 4 - C. 1227–1239.

69. Omelyanchuk, N. A., Wiebe, D. S., Novikova, D. D., Levitsky, V. G., Klimova, N., Gorelova, V., Weinholdt, C., Vasiliev, G. V., Zemlyanskaya, E. V., Kolchanov, N. A., Kochetov, A. V., Grosse, I., & Mironova, V. V. Auxin regulates functional gene groups in a fold-change-specific manner in Arabidopsis thaliana roots // Scientific Reports - 2017 - T. 7 - № 1 - C.2489.

70. Ozsolak, F., & Milos, P. M. RNA sequencing: Advances, challenges and opportunities // Nature Reviews Genetics - 2011 - T. 12 - № 2 - C. 87–98.

71. Pallai, R., Bhaskar, A., Barnett-Bernodat, N., Gallo-Ebert, C., Nickels, J. T., & Rice, L. M. Cancerous inhibitor of protein phosphatase 2A promotes premature chromosome segregation and aneuploidy in prostate cancer cells through association with shugoshin // Tumor Biology - 2015 - T. 36 - № 8 - C. 6067–6074.

72. Paponov, I.A., Paponov, M., Teale, W., Menges, M., Chakrabortee, S., Murray, J.A., Palme, K. Comprehensive transcriptome analysis of auxin responses in Arabidopsis // Mol Plant - 2008 - T. 1 - № 2 - C. 321-37.

73. Paque, S., Weijers, D. Q&A: Auxin: the plant molecule that influences almost anything // BMC Biol - 2016 - T. 14 - № 67

74. Petersson, S.V., Johansson, A.I., Kowalczyk, M., Makoveychuk, A., Wang, J.Y., Moritz, T., Grebe, M., Benfey, P.N., Sandberg, G., Ljung, K. An auxin gradient and maximum in the Arabidopsis root apex shown by high-resolution cell-specific analysis of IAA distribution and synthesis // Plant

Cell - 2009 - T. 21 - № 6 - C. 1659-68.

75. Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., & Pritchard, J. K. Understanding mechanisms underlying human gene expression variation with RNA sequencing // Nature - 2010 - T. 464 - № 7289 - C. 768–772.

76. Rachid Zaim, S., Kenost, C., Berghout, J., Vitali, F., Zhang, H. H., & Lussier, Y. A. Evaluating single-subject study methods for personal transcriptomic interpretations to advance precision medicine // BMC Medical Genomics - 2019 - T. 12 - № 5 - C. 96.

77. Ren, S., Shao, Y., Zhao, X., Hong, C. S., Wang, F., Lu, X., Li, J., Ye, G., Yan, M., Zhuang, Z., Xu, C., Xu, G., & Sun, Y. Integration of Metabolomics and Transcriptomics Reveals Major Metabolic Pathways and Potential Biomarker Involved in Prostate Cancer // Molecular & Cellular Proteomics - 2016 - T. 15 - № 1 - C. 154–163.

78. Repana, D., Nulsen, J., Dressler, L., Bortolomeazzi, M., Venkata, S. K., Tourna, A., Yakovleva, A., Palmieri, T., & Ciccarelli, F. D. The Network of Cancer Genes (NCG): A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens // Genome Biology - 2019 - T. 20 - № 1 - C. 1.

79. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. Limma powers differential expression analyses for RNA-sequencing and microarray studies // Nucleic Acids Research - 2015 - T. 43 - № 7 - C. 47.

80. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data // Bioinformatics - 2010 - T. 26 - № 1 - C. 139–140.

81. Robinson, M. D., & Smyth, G. K. Small-sample estimation of negative

binomial dispersion, with applications to SAGE data // Biostatistics - 2008 - T. 9 - № 2 - C. 321–332.

82. Robinson, Mark D., & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance // Bioinformatics - 2007 - T. 23 - №21 - C. 2881–2887.

83. Roche, S., O'Neill, F., Murphy, J., Swan, N., Meiller, J., Conlon, N. T., Geoghegan, J., Conlon, K., McDermott, R., Rahman, R., Toomey, S., Straubinger, N. L., Straubinger, R. M., O'Connor, R., McVey, G., Moriarty, M., & Clynes, M. Establishment and Characterisation by Expression Microarray of Patient-Derived Xenograft Panel of Human Pancreatic Adenocarcinoma Patients // International Journal of Molecular Sciences - 2020 - T. 21 - № 3.

84. Saad, F., Hotte, S.J. Guidelines for the management of castrate-resistant prostate cancer // Can Urol Assoc J - 2010 - T. 4 - № 6 - C. 380-4.

85. Salleh, M. S., Mazzoni, G., Höglund, J. K., Olijhoek, D. W., Lund, P., Løvendahl, P., & Kadarmideen, H. N. RNA-Seq transcriptomics and pathway analyses reveal potential regulatory genes and molecular mechanisms in high- and low-residual feed intake in Nordic dairy cattle // BMC Genomics - 2017 - T. 18 - № 1 - C. 258.

86. Schcolnik-Cabrera, A., Oldak, B., Juárez, M., Cruz-Rivera, M., Flisser, A., & Mendlovic, F. Calreticulin in phagocytosis and cancer: Opposite roles in immune response outcomes // Apoptosis - 2019 - T. 24 - № 3 - C.245–255.

87. Schlaen, R. G., Mancini, E., Sanchez, S. E., Perez-Santángelo, S., Rugnone, M. L., Simpson, C. G., Brown, J. W. S., Zhang, X., Chernomoretz, A., & Yanovsky, M. J. The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms // Proceedings of the National Academy of Sciences of the United States of

America - 2015 - T. 112 - № 30 - C. 9382–9387.

88. Schultz, M. A., Hagan, S. S., Datta, A., Zhang, Y., Freeman, M. L., Sikka, S. C., Abdel-Mageed, A. B., & Mondal, D. Nrf1 and Nrf2 transcription factors regulate androgen receptor transactivation in prostate cancer cells // PloS One - 2014 - T. 9 - №1 - C. 87204.

89. Shannon, P. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks // Genome Research - 2003 - T. 13 - № 11 - C. 2498–2504.

90. Sharma, E., Jain, M., & Khurana, J. P. Differential quantitative regulation of specific gene groups and pathways under drought stress in rice // Genomics - 2019 - T. 111 - № 6 - C. 1699–1712.

91. Siegel, R.L., Miller, K.D., Jemal, A. Cancer statistics, 2016 // CA Cancer J Clin. - 2016 - T. 66 - № 1 - C. 7-30.

92. Singer, G. A. C., Lloyd, A. T., Huminiecki, L. B., & Wolfe, K. H. Clusters of co-expressed genes in mammalian genomes are conserved by natural selection // Molecular Biology and Evolution - 2005 - T. 22 - № 3 - C. 767–775.

93. Song, Z., Huang, Y., Zhao, Y., Ruan, H., Yang, H., Cao, Q., Liu, D., Zhang, X., Chen, K. The Identification of Potential Biomarkers and Biological Pathways in Prostate Cancer // J Cancer - 2019 - T. 10 - № 6 - C. 1398-1408.

94. St. Laurent, G., Shtokalo, D., Tackett, M. R., Yang, Z., Vyatkin, Y., Milos, P. M., Seilheimer, B., McCaffrey, T. A., Kapranov, P. On the importance of small changes in RNA expression // Methods - 2013 - T. 63 - № 1 - C. 18–24.

95. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S.,

& Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles // Proceedings of the National Academy of Sciences - 2005 - Т. 102 - № 43 - С. 15545–15550.

96. Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., & Yaspo, M.-L. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome // Science - 2008 - Т. 321 - № 5891 - С. 956–960.

97. Sun, J., Li, S., Wang, F., Fan, C., & Wang, J. Identification of key pathways and genes in PTEN mutation prostate cancer by bioinformatics analysis // BMC Medical Genetics - 2019 - Т. 20 - № 1 - С. 191.

98. Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T. L., Buske, F. A., Sabir, K., Gould, C. M., Zotenko, E., Bert, S. A., Giles, K. A., Bauer, D. C., Smyth, G. K., Stirzaker, C., O'Donoghue, S. I., & Clark, S. J. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations // Genome Research - 2016 - Т. 26 - № 6 - С. 719–731.

99. Takayama, K., Inoue, S. Transcriptional network of androgen receptor in prostate cancer progression // Int J Urol - 2013 - Т. 20 - № 8 - С. 756-68.

100. Tanaka, Y., Gavrielides, M. V., Mitsuuchi, Y., Fujii, T., & Kazanietz, M. G. Protein Kinase C Promotes Apoptosis in LNCaP Prostate Cancer Cells through Activation of p38 MAPK and Inhibition of the Akt Survival Pathway // Journal of Biological Chemistry - 2003 - Т. 278 - № 36 - С. 33753–33762.

101. Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. Data quality aware analysis of differential

expression in RNA-seq with NOISeq R/Bioc package // Nucleic Acids Research - 2015 - T. 711.

102. Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., & Su, Z. agriGO v2.0: A GO analysis toolkit for the agricultural community, 2017 update // Nucleic Acids Research - 2017 - T. 45 - № 1 - C122–129.

103. Tripathi, S., Glazko, G. V., & Emmert-Streib, F. Ensuring the statistical soundness of competitive gene set approaches: Gene filtering and genome-scale coverage are essential // Nucleic Acids Research - 2013 - T. 41 - № 7 - C. 82.

104. Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., & Altman, R. B. Nonparametric methods for identifying differentially expressed genes in microarray data // Bioinformatics - 2002 - T. 18 - № 11 - C. 1454–1461.

105. Wang, P., Yuan, Y., Lin, W., Zhong, H., Xu, K., & Qi, X. Roles of sphingosine-1-phosphate signaling in cancer // Cancer Cell International - 2019 - T.19 - № 295.

106. Wang, Z., Gerstein, M., & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics // Nature Reviews Genetics - 2009 - T. 10 - № 1 - C. 57–63.

107. Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., Franz, M., Grouios, C., Kazi, F., Lopes, C. T., Maitland, A., Mostafavi, S., Montojo, J., Shao, Q., Wright, G., Bader, G. D., & Morris, Q. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function // Nucleic Acids Research - 2010 - T. 38 - C. 214–220.

108. Watson, P.A., Chen, Y.F., Balbas, M.D., Wongvipat, J., Socci, N.D., Viale, A., Kim, K., Sawyers, C.L. Constitutively active androgen receptor

splice variants expressed in castration-resistant prostate cancer require full-length androgen receptor // Proc Natl Acad Sci U S A - 2010 - T. 107 - № 39 - C. 16759-65.

109. Wei, C., Li, J., & Bumgarner, R. E. Sample size for detecting differentially expressed genes in microarray experiments // BMC Genomics - 2004 - T. 5 - № 1 - C. 87.

110. Weng, J., Wang, J., Hu, X., Wang, F., Ittmann, M., Liu, M. PSGR2, a novel G-protein coupled receptor, is overexpressed in human prostate cancer // Int J Cancer - 2006 - T. 118 - № 6 - C. 1471-80.

111. 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 - T. 11 - № 4 - C. 434.

112. Xiong, H., Guo, H., Xie, Y., Zhao, L., Gu, J., Zhao, S., Li, J., & Liu, L. RNAseq analysis reveals pathways and candidate genes associated with salinity tolerance in a spaceflight-induced wheat mutant // *S*cientific Reports - 2017 - T. 7 - № 1 - C. 2731.

113. Yeh, C.-H., Bellon, M., & Nicot, C. FBXW7: A critical tumor suppressor of human cancers // Molecular Cancer - 2018 - T. 17 - № 1 - C. 115.

114. Zhang, B., & Horvath, S. A general framework for weighted gene co-expression network analysis // Statistical Applications in Genetics and Molecular Biology - 2005 - T. 4 - № 17.

115. Zhang, J.-B., Pan, Z.-X., Lin, F., Ma, X.-S., & Liu, H.-L. Biochemical methods for the analysis of DNA-protein interactions: Biochemical methods for the analysis of DNA-protein interactions // Hereditas (Beijing) - 2009 -

T. 31 - №3 - C. 325–336.

116. Zhao, X., Hu, H., Lin, H., Wang, C., Wang, Y., & Wang, J. Muscle Transcriptome Analysis Reveals Potential Candidate Genes and Pathways Affecting Intramuscular Fat Content in Pigs // Frontiers in Genetics - 2020 - T. 11 - № 877.

117. Zucconi, B. E., Makofske, J. L., Meyers, D. J., Hwang, Y., Wu, M., Kuroda, M. I., & Cole, P. A. Combination Targeting of the Bromodomain and Acetyltransferase Active Site of p300/CBP // Biochemistry - 2019 - T. 58 - № 16 - C. 2133–2143.

118. Zyla, J., Marczyk, M., Weiner, J., & Polanska, J. Ranking metrics in gene set enrichment analysis: Do they matter? // BMC Bioinformatics - 2017 - T. 18 - № 1 - C. 256.

Приложение 1. Применение метода FSEA к наборам генов содержащих определенные регуляторные элементы в промоторах или относящихся к процессам канцерогенеза

Таблица П1. Результат применения метода FSEA для анализа обогащения ДНК-мотивами (набор C3) и наборами онкогенов (набор C6) взятых из базы данных MSigDB (Liberzon et al., 2011) в данных по исследованию экспрессии сплайс-варианта *AR-V7* в клеточной линии LNCaP (GSE71334) (Cottard et al., 2017).

Идентификатор набора генов	Идентификатор MSigDB	интервал	*p-value*	*q-value*
ELK1_02	C3	1-2	4.67E-08	2.7E-04
TATA_01	C3	4-5	3.11E-08	2.7E-04
SCGGAAGY_ELK1_02	C3	1-2	1.02E-07	3.9E-04
TATAAA_TATA_01	C3	3-5	7.35E-07	2.14E-03
TGCGCANK_UNKNOWN	C3	1-3	1.21E-06	2.81E-03
RCGCANGCGY_NRF1_Q6	C3	1-2	3.19E-06	6.18E-03
NRF1_Q6	C3	1-2	5.29E-06	8.79E-03
KRAS.600_UP.V1_UP	C6	5	9.61E-06	1.01E-02
KRAS.LUNG_UP.V1_DN	C6	4-5	1.15E-05	1.01E-02
LEF1_UP.V1_UP	C6	5	4.61E-06	1.01E-02

Приложение 2. Категории ГО обнаруженные только при помощи метода FSEA в данных эксперимента по исследованию экспрессии генов в клеточной линии LNCaP

Таблица П2. Категории ГО, относящиеся к процессам канцерогенеза, обнаруженные в списках активируемых (акт.) или подавляемых (под.) генов, только при помощи метода FSEA в эксперименте по исследованию экспрессии генов в клеточной линии рака предстательной железы (клеточная линия LNCaP) по сравнению с нормальными клетками (клеточная линия HPrEC) (GSE70466).

Регуляция	Идентификатор ГО	Название ГО	SEA *p-value*	SEA *p-adjusted*	FSEA *p-value*	FSEA *p-adjusted*	Фолд-специфичный интервал
акт.	GO:0043410	positive regulation of MAPK cascade	0.941	1	6.53E-06	0.005	4-5
акт.	GO:0008283	cell proliferation	0.996	1	6.49E-06	0.005	4-5
акт.	GO:0000070	mitotic sister chromatid segregation	0.983	1	1.11E-05	0.007	1
акт.	GO:0090036	regulation of protein kinase C signaling	0.055	0.772	4.58E-05	0.023	4

акт.	GO:0048870	cell motility	1	1	5.64E-05	0.027	2-5
акт.	GO:0070848	response to growth factor	0.999	1	5.66E-05	0.027	4-5
акт.	GO:0017157	regulation of exocytosis	0.079	0.916	7.14E-05	0.032	5
акт.	GO:0040012	regulation of locomotion	0.999	1	7.26E-05	0.033	2-5
акт.	GO:0050900	leukocyte migration	0.999	1	9.96E-05	0.043	4
акт.	GO:0031570	DNA integrity checkpoint	0.962	1	1.08E-4	0.047	1-2
акт.	GO:0001503	ossification	0.4	1	1.17E-4	0.049	5
акт.	GO:0007269	neurotransmitter secretion	0.001	0.058	2.14E-05	0.012	5
акт.	GO:0007186	G protein-coupled receptor signaling pat...	0.205	0.916	7.27E-10	1.60E-6	4-5
акт.	GO:0008015	blood circulation	0.242	0.946	3.83E-07	4.14E-4	3-5
акт.	GO:0007155	cell adhesion	0.999	1	3.21E-10	8.15E-7	3-5
акт.	GO:0006119	oxidative phosphorylation	4.2e-13	1.08E-9	1.06E-11	4.65E-8	1-2

акт./под.	GO:0006281	DNA repair	0.99/1	1/1	1.92E-07/1.04E-11	2.24E-4/4.22E-8	1-2/1
под.	GO:0006302	double-strand break repair	0.999	1	1.02E-05	0.006	1
под.	GO:0044782	cilium organization	0.998	1	1.49E-05	0.008	1-4
под.	GO:0000910	cytokinesis	0.025	0.111	7.49E-07	6.21E-4	1-3
под.	GO:0048514	blood vessel morphogenesis	6.0e-30	1.395E-27	3.41E-10	7.97E-7	3-5
под.	GO:0050764	regulation of phagocytosis	0.038	0.158	1.35E-4	0.048	5

Приложение 3. Категории ГО обнаруженные при помощи метода FSEA в данных транскриптомных экспериментов из Таблицы 2.

Таблица П3. Категории ГО, показавшие наибольшую значимость в списках активируемых (акт.) генов в 30 экспериментах отобранных для апробации метода FSEA (Таблица 2) для трех словарей ГО (БП - биологические процессы; КК - клеточные компоненты; МФ - молекулярные функции).

Идентификатор ГО	Название ГО	Фолд-специфичный интервал	FSEA p-adjusted	Порядковый номер эксперимента из Таблицы 2	Словарь ГО
GO:0001510	RNA methylation	5	5.08E-46	23	БП
GO:0009451	RNA modification	5	4.85E-35	23	БП
GO:0006412	translation	1	7.21E-29	11	БП
GO:0006518	peptide metabolic process	1	7.21E-29	11	БП
GO:0043043	peptide biosynthetic process	1	7.21E-29	11	БП
GO:0006412	translation	4-5	1.22E-28	23	БП
GO:0043043	peptide biosynthetic process	4-5	1.22E-28	23	БП
GO:0043604	amide biosynthetic process	4-5	1.67E-28	23	БП
GO:0043604	amide biosynthetic process	1	1.74E-28	13	БП
GO:0006518	peptide metabolic	4-5	3.90E-28	23	БП

	process				
GO:0003735	structural constituent of ribosome	4-5	5.03E-51	23	КК
GO:0005198	structural molecule activity	4-5	1.48E-43	23	КК
GO:0003723	RNA binding	1-2	2.27E-30	17	КК
GO:0003735	structural constituent of ribosome	1	3.88E-25	13	КК
GO:0005198	structural molecule activity	1	6.77E-25	13	КК
GO:0003735	structural constituent of ribosome	1-2	8.49E-24	22	КК
GO:0003723	RNA binding	1-4	6.01E-20	25	КК
GO:0005198	structural molecule activity	1-2	6.35E-20	22	КК
GO:0003735	structural constituent of ribosome	1-3	8.54E-19	27	КК
GO:0003735	structural constituent of ribosome	4	3.28E-18	18	КК
GO:0044445	cytosolic part	4-5	4.95E-54	23	МФ
GO:0022626	cytosolic ribosome	4-5	1.46E-53	23	МФ
GO:0005840	ribosome	4-5	2.70E-51	23	МФ
GO:0044391	ribosomal subunit	4-5	8.56E-51	23	МФ
GO:1990904	ribonucleoprotein complex	4-5	5.83E-45	23	МФ

GO:0043228	non-membrane-bounded organelle	4-5	1.32E-39	23	МФ
GO:0043232	intracellular non-membrane-bounded organelle	4-5	1.32E-39	23	МФ
GO:0032991	protein-containing complex	1-2	1.81E-37	22	МФ
GO:0022625	cytosolic large ribosomal subunit	5	1.66E-36	23	МФ
GO:0005829	cytosol	1	3.61E-35	13	МФ