Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет «Высшая школа экономики»

Факультет биологии и биотехнологии
Образовательная программа «Клеточная и молекулярная биотехнология»
бакалавриат

Енгибарян Нарек Карапетович

Роль изоформ микроРНК в сетях ко-экспрессии генов

Выпускная квалификационная работа - БАКАЛАВРСКАЯ РАБОТА
по направлению подготовки 06.03.01 Биология
ОП бакалавриат «Клеточная и молекулярная биотехнология»

Рецензент:

к.б.н., ФГБУ «НМИЦ онкологии им. Н.Н. Блохина» Минздрава РФ, заведующий лабораторией

Д.А. Хоченков

Научный руководитель:

д.б.н., профессор

А. Г. Тоневицкий

Консультант:

преподаватель

А. П. Жиянов

Москва – 2023

Содержание

Введение

Состояние отдельных органов и функционирование всего организма на базовом уровне определяются уровнем экспрессии генов. Интенсивность экспрессии генов зависит от присутствия и количества специфических факторов транскрипции [1]; синтез белка во время трансляции может быть подавлен микроРНК-опосредованной РНК-интерференцией [2]. микроРНК - это короткие некодирующие РНК, которые регулируют экспрессию генов путем распознавания 3’-конца мишени с помощью нуклеотидов 2-7, называемыми “seed”-регионом. Иногда изменения в механизмах процессинга микроРНК приводят к созданию изоформ вместо канонической микроРНК – это так называемые isomiR [3]. isomiR — это варианты зрелой микроРНК, которые отличаются друг от друга несколькими нуклеотидами на своих концах. Если такие изменения влияют на 5’-конец микроРНК, то молекула будет иметь другую seed область и, следовательно, она будет нацелена на другую мРНК.

Анализ клеточного транскриптома, то есть набора РНК в клетках, это один из самых распространенных методов изучения клеток *in silico*. Наиболее важной задачей транскриптома клеток является поиск тех генов, которые имеют различные уровни экспрессии в разных группах или состояниях. Так, сравнение нормальных образцов с «больными» позволяет обнаружить гены прогностически [4] и терапевтически [5] значимые гены. Анализ дифференциально экспрессируемых генов поддаётся относительно лёгкой интерпретации, поскольку результатом процедуры является список генов, экспрессия которых изменилась в новом состоянии по сравнению с референсным, а также степень этого изменения. Благодаря этому данный метод является одним из самых распространённых способов анализа транскриптома клеток. Ещё одной достаточно важной причиной распространенности метода является невысокая требовательность

существующих алгоритмов к вычислительным возможностям: анализ может быть проведен даже на персональном компьютере.

Однако при всех своих достоинствах анализ дифференциальной экспрессии не позволяет понять, взаимосвязаны ли изменения в экспрессии нескольких генов или они независимы. Существует анализ дифференциальной ко-экспрессии генов, позволяющий обнаружить изменения в регуляции групп генов в зависимости от условий, например, из-за общности регуляторных путей или факторов транскрипции. Например, в опухолевых клетках меняется характер регуляции множества генов при изменении соотношения микроРНК и её изоформ, поскольку изменяются сами регуляторные сети [6]. Анализ дифференциально ко-экспрессируемых генов даёт возможность обнаружить те регуляторные сети, в которых переопределяется паттерн экспрессии различных генов.

Несмотря на то, что микроРНК играют важную роль в регуляции экспрессии генов, влияние их изоформ на сети экспрессии в раковых клетках изучено недостаточно. Цель этого исследования - обнаружить изоформы микроРНК, которые играют ключевую роль в сетях ко-экспрессии генов различных типов и подтипов рака, и выявить, что определяет соотношение канонических микроРНК и их изоформ.

Основными задачами данного исследования являются:

1. изучить известную информацию о роли микроРНК и их изоформ в различных типах опухолей и в сетях ко-экспрессии генов;
2. определить методы изучения сетей ко-экспрессии генов;
3. выбрать и предобработать наборы данных (датасеты), на которых будет выполняться анализ;

4. проверить, зависит ли взаимодействие изоформ микроРНК с их мишенью от экспрессии генов, которые играют ключевую роль в процессинге микроРНК;
5. найти микроРНК и их изоформы, взаимодействия которых с их мишенями изменяются в опухоли по сравнению с нормальной тканью;
6. охарактеризовать обнаруженные молекулы и предположить возможные причины выявленных закономерностей;
7. сформулировать возможные направления дальнейших исследований.

В центре внимания данной работы находятся сети ко-экспрессии в нормальных и опухолевых тканях; объектом исследования являются изоформы микроРНК, играющие значительную роль в таких сетях. В рамках исследования будет выполнен поиск генов, связанных с экспрессией isomiR, и выявление специфических микроРНК и их изоформ, участвующих в механизмах регуляции экспрессии генов.

1. Обзор литературы

1.1. микроРНК и их роль в онкологических заболеваниях

микроРНК (miRNAs) представляют собой небольшие (длиной 21-23 нуклеотида) одноцепочечные молекулы РНК, которые широко представлены среди эукариот [7]. Будучи вовлеченными во многие клеточные процессы, при раке микроРНК могут действовать как онкогены или супрессоры опухоли, в зависимости от генов-мишеней, которые они регулируют. Нарушение регуляции экспрессии микроРНК связано с различными признаками рака, включая неконтролируемую клеточную пролиферацию, устойчивость к апоптозу или ангиогенез [8]. Они участвуют в регуляции экспрессии генов посредством механизма, называемого РНК-интерференцией. микроРНК в комплексе с белками, называемом RISC (RNA-induced silencing complex, РНК-индуцируемый комплекс выключения гена), использует свои нуклеотиды 2-7, называемые seed-областью, для распознавания специфической мРНК и посттранскрипционного ингибирования экспрессии соответствующего гена [9].

Образование микроРНК – сложный многоступенчатый процесс, в котором участвует множество белков. Наиболее распространенный путь начинается с транскрипции и образования при-микроРНК – РНК, состоящей из двуцепочечной петли-шпильки и длинных одноцепочечных хвостов. Затем фермент под названием Drosha в ассоциации с DGCR8 отсекает эти хвосты у определенных нуклеотидов, оставляя только стебель-петлю, так называемую пре-микроРНК, которая позже транспортируется из ядра. В цитоплазме рибонуклеаза, называемая Dicer, расщепляет шпильку в определенном положении, образуя РНК-дуплекс, одна из нитей которого будет загружена в RISC-комплекс [10].

Показано, что основной белок в комплексе RISC, называемый AGO2 (т.н. «аргонавт»), сверхэкспрессируется при карциномах, включая рак толстой кишки, плоскоклеточный рак головы и шеи, уротелиальную

карциному мочевого пузыря, карциному яичников, карциному желудка и колоректальную карциному, и сверхэкспрессия AGO2 была связана с ростом опухолевых клеток и общей выживаемостью онкологических больных [11]. AGO2 участвует в онкогенезе как через miRNA-зависимые, так и независимые механизмы, и он взаимодействует с хорошо известными опухолевыми факторами, такими как AKT3, EGFR, FAK и P4H.

Согласно исследованию, опубликованному в журнале Nucleic Acids Research [12], нокдаун AGO2 связан с более медленной скоростью роста клеток. Исследовательская группа использовала нокауты вариантов AGO в клетках колоректального рака HCT116 и подтвердила, что нокаут AGO2 и AGO1/2/3 приводит к более медленным темпам роста клеток по сравнению с клетками дикого типа. Это говорит о том, что белок AGO2 оказывает влияние на рост и пролиферацию клеток. Кроме того, исследование обнаружило сильную корреляцию между величиной изменений уровней мРНК в устойчивом состоянии для отдельных генов в нокаутированных клетках AGO1/2/3 и более медленными темпами роста клеток. Таким образом, нокдаун AGO2 может косвенно влиять на изменения экспрессии генов, которые проявляются в замедлении темпов роста клеток.

Другое исследование обнаружило [13], что нарушение регуляции транскрипции Dicer и AGO2 связано с обострением детского острого лимфобластного лейкоза B-клеточной линии (ОЛЛ). Исследование показало, что уровень экспрессии DICER был повышен, в то время как экспрессия AGO2 была снижена у пациентов с ОЛЛ по сравнению со здоровой контрольной группой. Снижение экспрессии AGO2 напрямую коррелировало с прогрессированием заболевания от стадии L1 к стадии L2. Нарушение регуляции факторов, участвующих в биогенезе микроРНК, может влиять на уровень микроРНК, приводя к нарушению клеточного цикла, онкогенеза и апоптоза, способствуя возникновению и развитию лейкемии. Исследование предполагает, что эти факторы можно рассматривать в качестве

потенциальных индикаторов прогрессирования острого лимфобластного лейкоза или даже в качестве прогностических маркеров.

1.2. Изоформы микроРНК

Определено, что иногда Drosha и Dicer расщепляют микроРНК по-разному, образуя молекулы со смещенными нуклеотидами на 5’-конце, 3’-конце или обоих концах одновременно. Такие микроРНК называются изоформами микроРНК (isomiR) и действуют точно так же, как обычные микроРНК [3]. Исключение возникает в случае сдвига на 5’-конце, где расположена seed-область. В случае сдвигов на 5’-конце изменится начало микроРНК, что приведет к нацеливанию на совершенно другой набор мРНК.

Ещё одним вариантом изоформ микроРНК являются isomiRs с точечными заменами нуклеотидов, образующиеся в результате РНК-редактирования канонической микроРНК, например, ферментами ADAR, которые дезаминируют аденин (A), превращая его в инозин (I). Также существует механизм превращения цитозина (C) в урацил (U) с помощью ферментов APOBEC [14]. В некоторых видах рака ферменты семейств ADAR и APOBEC существенно дифференциально экспрессированы по сравнению с нормой. Так, например, в колоректальной аденокарциноме (COAD) ген APOBEC имеет почти в 3 раза большую экспрессию по сравнению с нормой, а уровень экспрессии ADAR1 и ADAR2 понижается в случае колоректальной аденокарциномы, рака молочной железы (BRCA), мультиформной глиобластомы (GBM), аденокарциномы лёгких (LUAD), саркомы (SARC) и других типах опухолей.

Исследование Aristeidis G. Telonis и соавторов показало [15], что при сверхэкспрессии различных изоформ miR-183-5p в клетках MDA-MB-231 изоформы оказывают различное влияние на клеточный транскриптом и могут работать синергически, направляя клетки рака молочной железы к более агрессивному фенотипу. Локус miR-183-5p считается т.н. “oncomiR”, и было обнаружено, что isomiR из этого локуса оверэкспрессированы при тройном

негативном раке молочной железы у белых, но не у чернокожих женщин. Взаимодействие между различными isomiR, берущими начало из одной и той же pri-микроРНК шпильки, и то, как они совместно влияют на количество мРНК, все еще остаётся открытым вопросом.

Исследователи из Университета Томаса Джефферсона провели исследование [16], используя информацию о наличии или отсутствии изоформ микроРНК для классификации 32 типов рака. Изучив 10 271 образец TCGA, они смогли создать классификатор, который опирался исключительно на бинарные профили изоформ, помечая каждую isomiR либо как "присутствующую", либо как "отсутствующую". Классификатор оказался успешным, с чувствительностью в 90% и низким уровнем ложноположительных результатов (FDR) - всего 3%. Анализ также показал, что наиболее важные для классификации изоформы также по-разному экспрессировались между нормальной тканью и раком, что позволяет предположить, что они могут оказаться полезными в качестве биомаркеров рака.

Глобальные изменения в соотношении isomiR и канонических микроРНК теоретически могут повлиять на всю сеть экспрессии генов клетки, поскольку микроРНК являются существенными регуляторами экспрессии генов. Несмотря на потенциально важную роль изоформ микроРНК в различных патологических состояниях клеток, до сих пор неясно, что может влиять на набор микроРНК и их изоформ в клетке.

1.3. Роль изоформ микроРНК в сетях ко-экспрессии генов

Наиболее часто используемым методом анализа данных РНК-секвенирования является дифференциальная экспрессия генов. Этот метод позволяет обнаружить мРНК или микроРНК, средняя экспрессия которых в одном состоянии (например, при опухоли) изменяется по сравнению с другим состоянием (нормальные ткани) [17][18]. Однако анализ дифференциальной экспрессии не позволяет обнаружить изменения в

регуляции генов в случае, если средние экспрессии молекулы-регулятора или её мишени остаются неизменными. Так, например, в раке предстательной железы экспрессия гена ATG16L1 значимо не изменяется (Рис. 1A), в отличие от уровня экспрессии канонической формы микроРНК hsa-miR-93-5p (Рис. 1B); при этом корреляционный анализ указывает на изменение взаимной экспрессии между этими двумя молекулами (Рис. 1C).

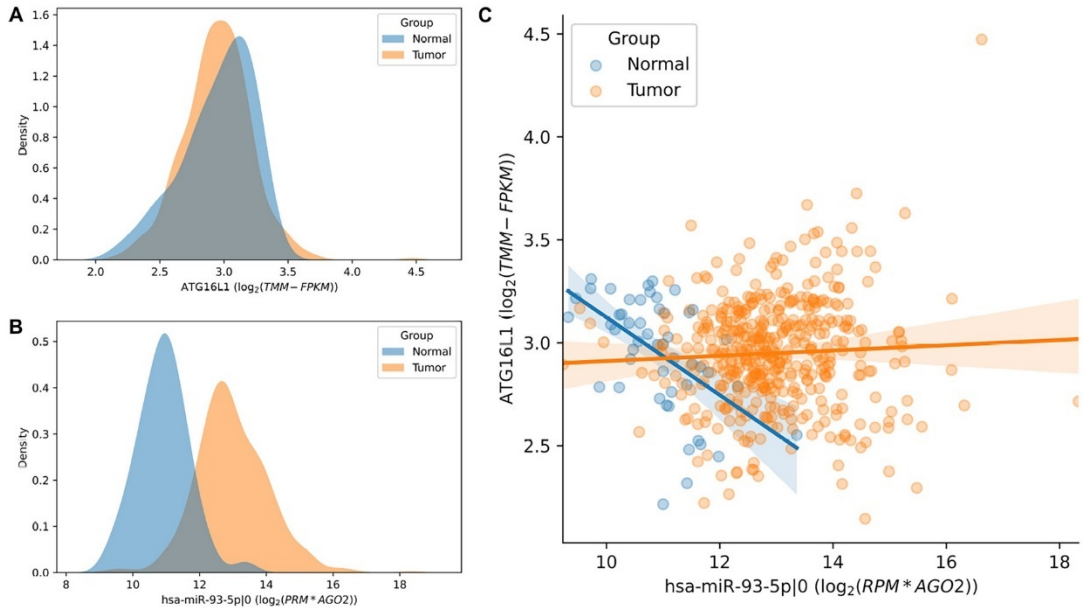


Рисунок 1 - Уровни экспрессий генов ATG16L1 и hsa-miR-93-5p|0 и корреляция между ними в раке предстательной железы [19]

Подобная ситуация описана в работе Keller и соавторов [20]: было показано, что p16 и группа циклинов были отрицательно коррелированы у худых мышей, но положительно коррелировали у мышей с ожирением, что указывает на связанную с ожирением регуляцию клеточного цикла. Примечательно, что p16 и многие циклины не были дифференцированно экспрессированы у мышей с худобой и ожирением и, следовательно, были бы пропущены при анализе дифференциальной экспрессии.

Таким образом, распространенный биоинформатический подход к анализу микроРНК или их изоформ заключается в прогнозировании их мРНК-мишеней [21], и дальнейшем обнаружении значительной отрицательной корреляции между парами микроРНК-мРНК. Такой метод

был использован для демонстрации связывания микроРНК с мишенями при раке молочной железы [22] или раке предстательной железы [23].

Разница в корреляциях между парами регулятор-мишень в двух состояниях, например, в норме и при заболевании, может свидетельствовать о глубоких изменениях в сетях ко-экспрессии генов. Например, было показано, что при раке молочной железы изменения корреляции между фактором транскрипции PTEN и его мишенями связаны с мутациями PTEN [24]. Другое исследование показало потерю отрицательной корреляции между высокоэкспрессируемыми изоформами микроРНК и их мишенями при раке предстательной железы [19]. К сожалению, поведение isomiR в различных типах опухолей изучено недостаточно. Не менее важно изучить их влияние на сети ко-экспрессии генов и найти общие или частные закономерности.

1.4. Методы изучения сетей ко-экспрессии генов

Существует ряд биоинформатических инструментов для анализа сетей дифференциальной ко-экспрессии генов. В целом, методы анализа дифференциальной ко-экспрессии можно разделить на два основных класса. Первый класс имеет дело с попарными ко-экспрессиями и обнаруживает пары генов, взаимодействие которых значительно изменилось в зависимости от условий. Для численной оценки уровня ко-экспрессии используются несколько показателей, такие как корреляция Пирсона или Спирмена и взаимная информация (mutual information). Сравнивая значения этих показателей в разных условиях, можно сделать вывод о существенном изменении ко-экспрессии определенных пар генов. Примерами инструментов, вычисляющих корреляцию для обнаружения изменений ко-экспрессии, могут послужить Differential Gene Correlation Analysis (DGCA) [24], Differential Correlation Network Analysis (DCoNA) [19] и DiffCorr [25]. В указанных алгоритмах авторы использовали основанную на корреляции z-статистику Фишера [26] и проверили гипотезу о том, что корреляции при

заданных условиях равны. В качестве примера некорреляционных методов можно упомянуть PMINR (Pointwise Mutual Information-Based Network Regression) [27], где авторы использовали взаимную информацию и построили регрессионную модель для выявления попарных взаимодействий, связанных с заболеванием.

Второй тип анализа обнаруживает ко-экспрессируемые генные модули или кластеры на основе сходства экспрессии их генов в каждом состоянии. Примерами таких исследований являются Weighted Correlation Network Analysis [28] и Multiscale Embedded Gene Co-expression Network Analysis [29], вычисление статистики перекрытия модулей между условиями [30] или средняя модульная дифференциальная связность [31].

2. Материалы и методы

2.1. Данные РНК-секвенирования TCGA

В этом исследовании использованы данные РНК-секвенирования The Cancer Genome Atlas (TCGA), включающие 33 злокачественные опухоли. Данные для микроРНК и мРНК были получены независимо. Данные mRNA-seq, содержащие 10530 образцов, были загружены с портала UCSC Xena [32]. Данные microRNA-seq (11089 образцов) были загружены с официального веб-сайта инструмента IsoMiRmap [33], и для дальнейшей обработки были отобраны только уникально картированные (exclusive) чтения.

Для нормализации данных использовалась функция `estimateSizeFactors`, которая реализует алгоритм медианы соотношений (the median of ratios) в R пакете DESeq2 [17]. Для создания окончательных нормализованных матриц экспрессии к данным microRNA-seq и mRNA-seq были применены функции `fpm` и `fpkm` соответственно (данные microRNA-seq не были нормализованы на длину транскриптов, поскольку имеют приблизительно равные длины). В качестве заключительного шага предобработки данных к каждому значению экспрессии было применено преобразование $\log_2(x + 1)$.

Поскольку isomiR с одинаковыми 5’-концами и различными 3’-концами должны иметь похожий набор мРНК-мишеней (поскольку такие изменения не затрагивают seed-область микроРНК), были суммированы экспрессии isomiR, происходящих от одной и той же при-микроРНК и имеющих идентичные 5’-концы (функция `isomir_groupby_5prime`, приложение А). Обозначения из исследования Нерсисяна и соавторов [34] используются для 5’-изомиров: число после символа “|” представляет количество нуклеотидов, на которые произошел сдвиг от канонического 5’-конца в направлении 5’-3’.

2.2. Инструмент DCoNA

DCoNA (Differential Co-expression Network Analysis) [19] - это статистический инструмент, позволяющий обнаруживать и измерять дифференциальные корреляции пар генов между двумя состояниями. Он вычисляет корреляции Пирсона или Спирмена для пар генов в двух условиях, например, в “опухолевых” и “нормальных” образцах, и применяет к ним преобразование Фишера, чтобы распределение значений было приближено к нормальному с нулевым средним и единичной дисперсией. Наконец, DCoNA тестирует значения корреляций на равенство между двумя условиями и вычисляет P-value и P-value с поправкой на множественное тестирование гипотезы (метод Беньямини-Хохберга [35]).

Этот инструмент также поддерживает выполнение «перестановочного» теста (permutation test) [36]. Он перемешивает метки образцов и вычисляет P-value на основе эмпирически найденного распределения, что помогает повысить точность, но требует значительного количества перестановок для достижения результата. Перестановочный тест позволяет получить более «щадящие», по сравнению с поправкой Беньямини-Хохберга, значения P-value, но остающиеся при этом статистически значимыми.

Использование гипергеометрического теста позволяет DCoNA агрегировать различия в корреляциях, находя такие молекулы, которые дифференциально коррелируют со высокопредставленной группой генов. Это помогает найти гены, занимающие центральное место в сетях ко-экспрессии: либо микроРНК интенсивно взаимодействуют с широким спектром мРНК, либо наоборот, мРНК подавляется многими молекулами микроРНК.

Для удобства работы с данными TCGA и запусков DCoNA было написано несколько Python-классов и функций, код и описание которых приведено в приложении А.

Во время анализа сравнивались корреляции пар микроРНК-мишень в образцах с низкой экспрессией гена, участвующего в процессинге микроРНК, с образцами с высокой экспрессией этого гена. Дифференциально ко-экспрессированными считались пары микроРНК-мишень, если P-value с поправкой на множественное тестирование было меньше 0.05.

2.3. Предсказания мишеней микроРНК

Для эффективного использования DCoNA следует рассчитывать корреляции не для всех возможных пар молекул, а только между парами микроРНК-транскрипт. Более того, даже среди таких пар не все имеют ценность, поскольку можно взять лишь те, для которых предсказано значимое связывание на основе их нуклеотидной последовательности. Существуют инструменты, позволяющие предсказывать силу и вероятность связывания микроРНК и мишени. Одними из таких программ являются TargetScan [37] и RNA22 [38]. Результаты предсказаний [34] для рака молочной железы были загружены с портала Figshare [39]. Пересечения между результатами RNA22 и TargetScan составляют менее 20% пар микроРНК-мРНК (Рис. 2). Одной из причин этого, вероятно, является то, что TargetScan ищет сайты связывания микроРНК преимущественно в 3'-нетранслируемой области транскрипта, в то время как RNA22 учитывает ещё и кодирующую область (CDS) и 5'-нетранслируемую область.

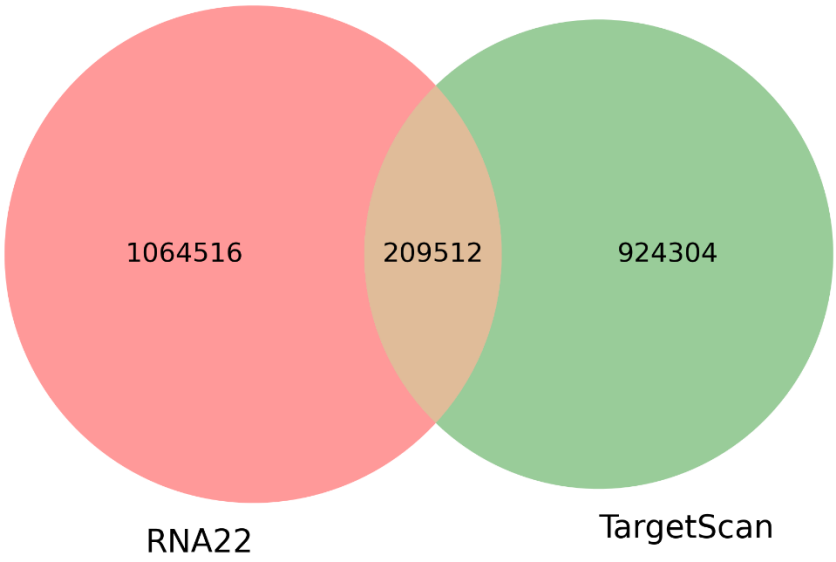


Рисунок 2 - Диаграмма Венна для предсказанных с помощью RNA22 и TargetScan пар микроРНК-мРНК

2.4. Отбор данных для анализа

Для анализа были выбраны образцы рака молочной железы (TCGA-BRCA), поскольку РМЖ имеет удобное разделение на подтипы по чётко определенным критериям [40,41]; а также по той причине, что в данных TCGA-BRCA есть достаточное количество образцов для получения статистически значимого результата при анализе с помощью DCoNA.

Использовались данные экспрессии мРНК на уровне транскриптов. Предобработка включала отбор образцов определенного подтипа (например, Normal или Luminal A), после чего следовала фильтрация (функция `filter_by_median`, приложение А): сперва отбирались транскрипты с медианным уровнем экспрессии больше нуля, затем рассчитывались кумулятивные суммы значений FPKM, и отбирались транскрипты, составляющие 95% суммарной экспрессии образца.

Данные экспрессии микроРНК фильтровались с помощью той же функции, однако в качестве «отсечки» использовалось 99% экспрессии: в результате из 1216 5’-изоформ микроРНК было отобрано 142 наиболее экспрессированных.

2.5. Подготовка аннотаций для DCoNA

Поскольку основной задачей является изучение влияния генов, участвующих в процессинге микроРНК, на сети ко-экспрессии генов, было решено сравнивать взаимодействия микроРНК с мишенями в двух состояниях: с высоким и низким уровнями экспрессии выбранного гена в определенном подтипе рака молочной железы. Для этого были отобраны четыре гена и соответствующие транскрипты:

1. AGO2 (ENST00000220592) – главный белок комплекса RISC, осуществляющий РНК-интерференцию;
2. DROSHA (ENST00000513349) – нуклеаза, участвующая в процессинге pri-miRNA в ядре клетки с образованием pre-miRNA;
3. DGCR8 (ENST00000351989) – белок, стабилизирующий pri-miRNA для процессинга с помощью DROSHA;
4. DICER1 (ENST00000393063) – нуклеаза, осуществляющая отрезание петли pre-miRNA с образованием дуплекса микроРНК;

Для каждого гена были отобраны образцы, относящиеся к тому или иному подтипу РМЖ, и из них выбирались те, в которых уровень экспрессии данного гена наиболее высок (метка «High_expr») или низок (метка «Low_expr»). В отношении всех генов данная процедура была проведена для подтипов РМЖ Normal и Luminal A, а для AGO2 дополнительно были отобраны образцы Luminal B и Basal-like. Таким образом, было получено 10 аннотаций (Рис. 3). Количество образцов в каждой аннотации указано в таблице 1.

Таблица 1. Количество образцов, с высокой и низкой экспрессией ключевых генов

Ген	Подтип РМЖ	Всего образцов	Отобрано образцов с высокой экспрессией	Отобрано образцов с низкой экспрессией
AGO2	Normal	104	31	31
	Luminal A	561	57	56
	Luminal B	210	42	42
	Basal-like	183	37	37
DROSHA	Normal	104	31	31
	Luminal A	561	75	75
DGCR8	Normal	104	31	31
	Luminal A	561	75	75
DICER1	Normal	104	31	31
	Luminal A	561	75	75

При запуске DCoNA использовались матрицы экспрессий, полученные способом, описанным в предыдущем разделе. В качестве групп, между которыми сравнивается корреляция, выступают образцы с высокой и низкой экспрессией ключевого гена. Наконец, каждая аннотация (кроме DGCR8 Luminal A) комбинируется с двумя таблицами предсказанных взаимодействий: RNA22 и TargetScan. Таким образом, было проведено 19 запусков.

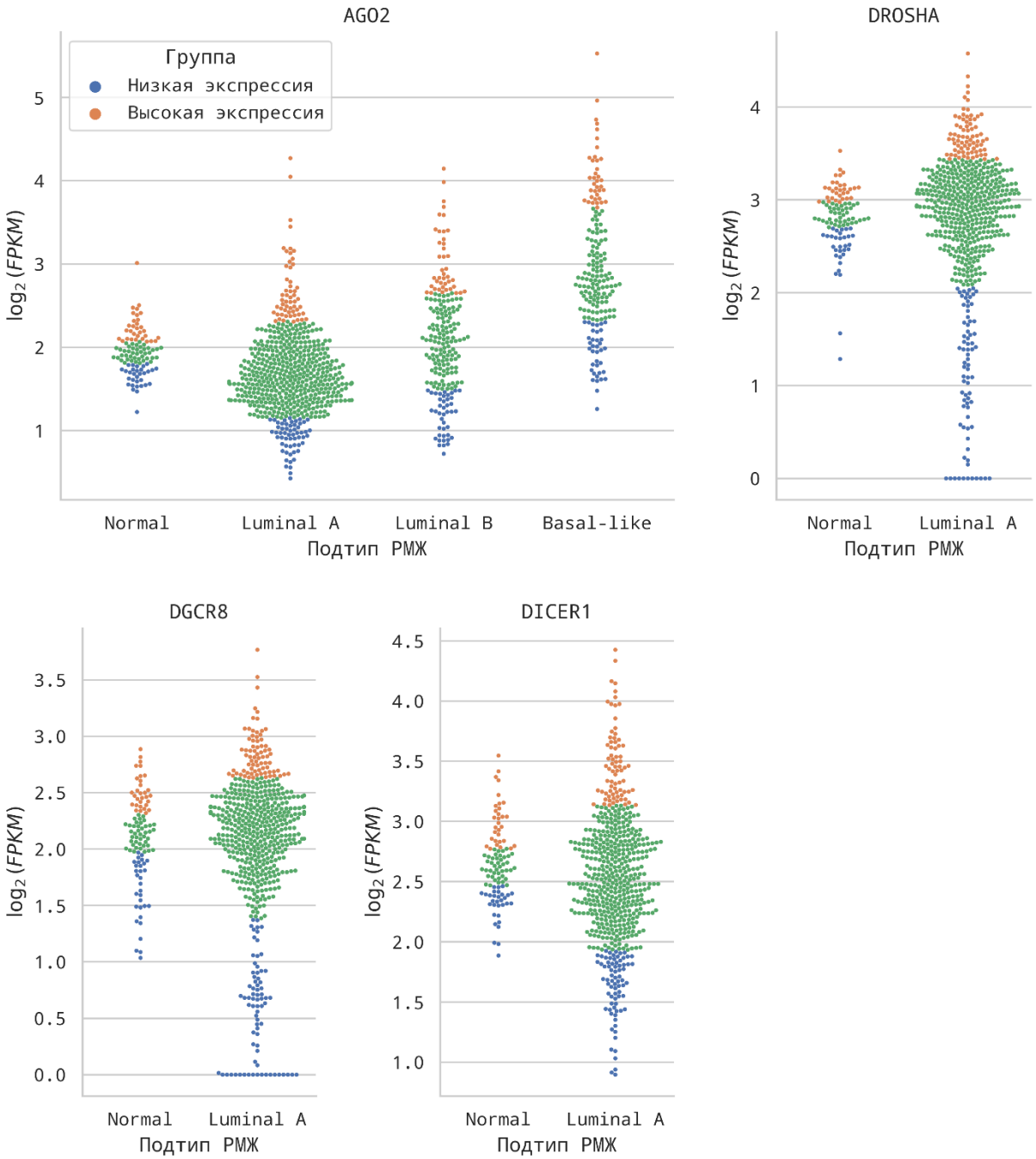


Рисунок 3 - распределение уровней экспрессии ключевых в процессинге микроРНК генов и отобранные для анализа группы образцов

2.6. Анализ дифференциальной экспрессии генов

В качестве дополнительного способа проанализировать данные был использован обычный анализ дифференциальной экспрессии микроРНК между группами с высокой и низкой экспрессиями ключевого гена, то есть были использованы те же аннотации, что и для запуска DCoNA (п. 2.5).

Для анализа использовался пакет DESeq2 [17]; образцы с низкой экспрессией гена, участвующего в процессинге микроРНК, сравнивались с образцами с высокой экспрессией этого гена. В качестве матрицы экспрессий использовались данные экспрессий 5’-изоформ микроРНК, полученные с помошью агрегации экспрессией функцией `isomir_groupby_5prime` (приложение А). Дальнейшая фильтрация по уровню экспрессий не применялась. микроРНК считались дифференциально экспрессированными, если p-value с поправкой на множественное тестирование [35] было меньше 0.05.

3. Анализ дифференциальной ко-экспрессии в зависимости от уровня экспрессии ключевого гена

3.1. Поиск пар микроРНК-мРНК с изменившейся взаимной корреляцией

Ко-экспрессии микроРНК и транскриптов сравнивались при помощи утилиты DCoNA между двумя состояниями: в случае высокой экспрессии ключевого гена и в случае низкой экспрессии. Мерой ко-экспрессии является корреляция Спирмена между микроРНК и их мишенями, предсказанными при помощи TargetScan или RNA22. Обобщенные результаты приведены в таблице 2.

Таблица 2. Количество пар, статистически значимо изменивших взаимную корреляцию, по результатам DCoNA ztest

Ген	Подтип РМЖ	Источник пар микроРНК-мРНК	Всего пар	Число пар с $FDR < 0.05$
AGO2	Normal	TargetScan	488477	0
		RNA22	513659	0
	Luminal A	TargetScan	455889	766
		RNA22	483291	848
	Luminal B	TargetScan	446066	0
		RNA22	470265	0
	Basal-like	TargetScan	429421	0
		RNA22	448281	1
DROSHA	Normal	TargetScan	488477	0
		RNA22	513659	0
	Luminal A	TargetScan	455889	236
		RNA22	483291	822
DGCR8	Normal	TargetScan	488477	193
		RNA22	513659	118
	Luminal A	TargetScan	455889	743
DICER1	Normal	TargetScan	488477	0
		RNA22	513659	16
	Luminal A	TargetScan	455889	0
		RNA22	483291	0

Как видно физиологические различия в экспрессии DICER1 мало влияют на корреляции микроРНК и их мишеней. Какие-либо общие

закономерности выделить трудно. Можно выдвинуть предположение, что в случае рака молочной железы есть общая тенденция к дифференциальной ко-экспрессии в Luminal A по сравнению с нормальными тканями. Однако это не подтверждается результатами анализа корреляций в зависимости от экспрессий DGCR8 и DICER1.

Интересным вопросом является то, есть ли смысл в использовании предсказаний мишеней как TargetScan, так и RNA22. Уже исходя из того, что сами таблицы взаимодействий микроРНК-мишень пересекаются между двумя инструментами достаточно слабо (рис. 2), можно сделать вывод о необходимости учитывать результаты обоих инструментов.
Это подтверждается тем, что в результатах анализа дифференциальной корреляции при различающихся таблицах взаимодействий и прочих равных достаточно небольшое количество пересекающихся пар. Например, для подтипа Luminal A и при сравнении корреляций при разных экспрессиях AGO2 наблюдается менее 25% общих пар микроРНК-транскрипт (рис. 4).

Рисунок 4 - количество значимо дифференциально коррелированных пар микроРНК-мишень, предсказанных соответствующим инструментом и общие пары

3.2. Обобщение результата вычислений с помощью гипергеометрического теста

Использование гипергеометрического теста позволяет обнаружить микроРНК, у которых изменились корреляции со значимо большим числом мишеней. Рассмотрим результаты анализа для каждого гена.

3.2.1. AGO2

Исходя из функций AGO2 можно предположить, что изменение уровня экспрессии этого белка должно повлечь за собой системные изменения в ко-экспрессии микроРНК с появлением множества отрицательных корреляций между микроРНК и их мишенями, поскольку именно AGO2 участвует во взаимодействии между ними.

В целом, большая часть из тех микроРНК, что изменили корреляции, в случае высокой экспрессии AGO2 действительно приобретает новые отрицательные корреляции (Рис. 5). Однако остаётся непонятной причина исчезновения отрицательной корреляции со множеством мишеней у некоторых микроРНК.

Также стоит отметить, что количество мишеней, изменивших корреляции с микроРНК, составляют в среднем меньше, чем 5% от всех взаимодействий этой микроРНК (Приложение Б). При этом существенные изменения ко-экспрессии обнаружились только в подтипе Luminal A. В Luminal B и в случае нормальной ткани вовсе не было пар микроРНК-мишень, изменивших корреляцию, а в Basal-like наблюдался лишь один случай. Это даёт возможность сделать вывод, что появление корреляций в данном случае носит несистемный характер и, по крайней мере прямо, не зависит от физиологических изменений уровня экспрессии AGO2.

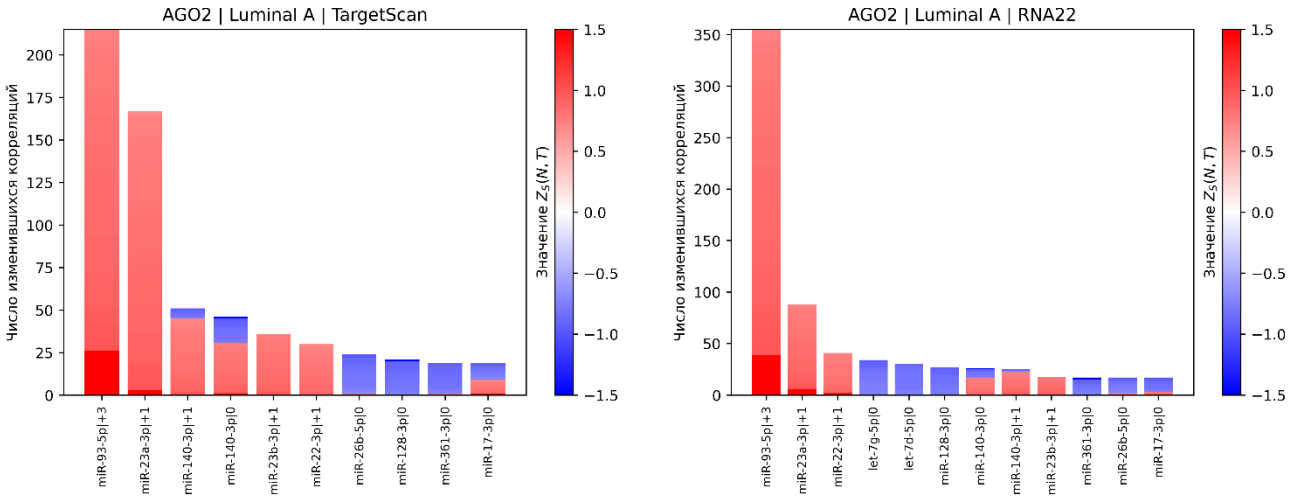


Рисунок 5 - количество значимо изменившихся корреляций. Синий цвет отображает потерю отрицательной корреляции при высоком уровне экспрессии AGO2, красный – приобретение новых отрицательных корреляций.

3.2.2. DROSHA и DGCR8

Поскольку DROSHA и DGCR8 совместно в ядре занимаются процессингом микроРНК, можно выдвинуть предположение, что при высокой экспрессии этих генов должны появляться новые корреляции микроРНК-мишень. Также возможны специфичные изменения в корреляциях изоформ микроРНК, поскольку именно во время работы DROSHA появляются изоформы.

Однако при анализе результатов (Рис. 6) сложно обнаружить подтверждения этих гипотез. Даже в пределах одной микроРНК наблюдаются как потеря, так и приобретение новых отрицательных корреляций. Разницы между и канонической формой и изоформами тоже нет: например, miR-140-3p и её изоформа miR-140-3p|+1 изменяют ко-экспрессию со своими мишенями похожим образом. Как и в случае с AGO2 изменения корреляций составляют малый процент от всех взаимодействий микроРНК с мишенями (Приложение Б).

Таким образом, изменения в экспрессии DROSHA и DGCR8 прямо не влияют ни на изменения в корреляциях, ни на соотношение канонической формы и изоформы.

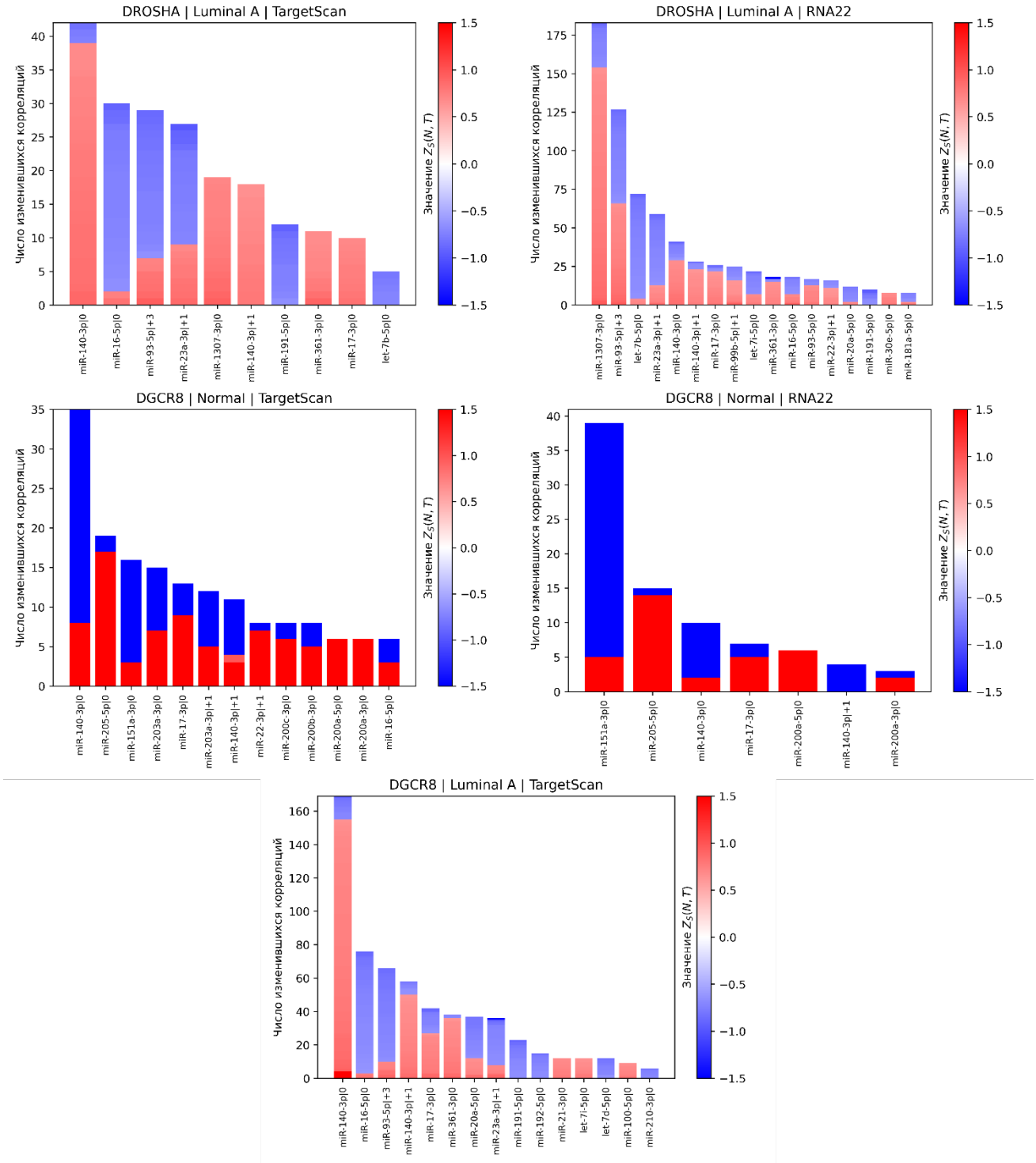


Рисунок 6 - количество значимо изменившихся корреляций. Синий цвет отображает потерю отрицательной корреляции при высоком уровне экспрессии DROSHA или DGCR8, красный – приобретение новых отрицательных корреляций.

3.2.3. DICER1

Поскольку DICER1 также участвует в процессинге микроРНК, можно предположить, что уровень его экспрессии будет влиять на ко-экспрессию микроРНК и их мишеней. Однако в случае DICER1 дифференциально коррелирующих пар почти не было выявлено, не считая 15 штук в одном из экспериментов (Рис.6). Очевидно, физиологические изменения в экспрессии DICER1 очень слабо влияют на ко-экспрессию микроРНК и мРНК.

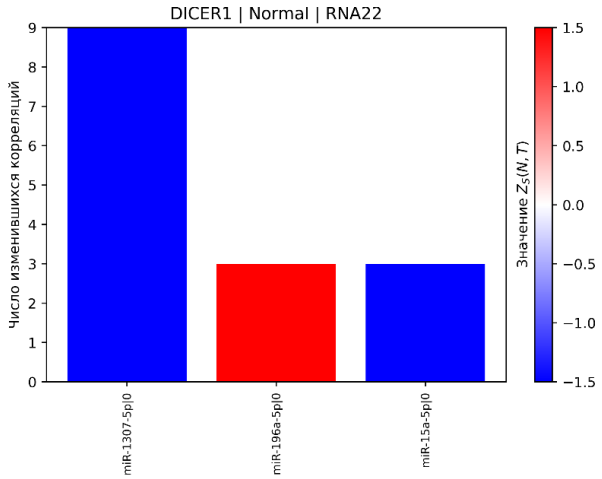


Рисунок 7 - количество значимо изменившихся корреляций. Синий цвет отображает потерю отрицательной корреляции при высоком уровне экспрессии DICER1, красный – приобретение новых отрицательных корреляций.

3.3. Обсуждение

Как видно, экспрессия ключевых генов в клетках не влияет прямым образом на ко-экспрессию микроРНК, однако некоторые бессистемные изменения, тем не менее, происходят. Стоит при этом отметить, что анализ учитывает лишь физиологические уровни экспрессии генов – скорее всего, при нокауте или нокдауне анализируемых генов мы бы увидели иной результат. Также стоит иметь в виду, что учитывались лишь наиболее представленные транскрипты генов AGO2, DROSHA, DICER1 и DGCR8. На ко-экспрессию микроРНК может оказывать альтернативный сплайсинг, что не учитывалось.

При этом необходимо отметить, что некоторые изоформы микроРНК изменяли свою корреляцию при изменении уровня экспрессии нескольких ключевых генов. Так, например, miR-93-5p|+3, miR-23a-3p|+1 и miR-140-3p|+1 оказывались значимо дифференциально ко-экспрессированы при разных уровнях экспрессии DROSHA, DGCR8, AGO2. Какая за этим стоит природа, непонятно – необходимы дальнейшие исследования.

4. Анализ дифференциальной экспрессии генов

Проводя анализ дифференциальной экспрессии и сравнивая образцы с высокой экспрессией и низкой экспрессией генов, участвующих в жизненном цикле микроРНК, можно предполагать, что высокая экспрессия AGO2, DROSHA, DGCR8 и DICER1 может быть связана с большей концентрацией зрелых молекул микроРНК в цитоплазме клетки, и что разница будет заметна для подавляющего большинства микроРНК.

Обобщённый результат анализа представлен в таблице 3. В некоторых экспериментах результат соответствует ожиданиям: в образцах с большей экспрессией «ключевого» гена наблюдается много генов, увеличивших экспрессию. Однако сделать вывод о прямой связи между уровнем экспрессии генов, участвующих в процессинге микроРНК, с уровнем экспрессии самих микроРНК невозможно по нескольким причинам. Во-первых, дифференциальная экспрессия носит не системный характер: в здоровых (Normal) образцах в случае AGO2 и DICER1 она практически не наблюдается. Во-вторых, во всех случаях есть внушительный процент микроРНК, показывающих обратную картину: уменьшение экспрессии микроРНК при увеличении экспрессии гена. Не наблюдается ни зависимости от подтипа рака, ни стабильного изменения экспрессии в одну сторону для подавляющего большинства микроРНК. Таким образом, экспрессия генов, принимающих участие в процессинге микроРНК, может оказывать влияние на уровень их экспрессии, однако не носит системный характер и не связана с уровнем экспрессии микроРНК напрямую.

Таблица 3. Количество микроРНК, увеличивших или уменьшивших экспрессию по результатам анализа дифференциальной экспрессии в различных подтипах РМЖ. Сравнивались образцы с разным уровнем экспрессии гена в первом столбце

Ген	Ткань	Увеличение экспр. микроРНК при высокой экспр. гена	Уменьшение экспр. микроРНК при высокой экспр. гена
AGO2	Normal	2	1
AGO2	Luminal A	230	52
AGO2	Luminal B	82	28
AGO2	Basal-like	52	43
DROSHA	Normal	48	42
DROSHA	Luminal A	228	120
DICER1	Normal	0	1
DICER1	Luminal A	96	82
DGCR8	Luminal A	174	70
DGCR8	Normal	285	152

Заключение

Учитывая ключевую роль микроРНК в сетях ко-экспрессии генов, определение места микроРНК и их изоформ в таких сетях и выявление того, от чего зависит их формирование, может быть важным для понимания процессов, происходящих в клетках человека.

Исследование показало, что ни ко-экспрессия микроРНК с их мишенями, ни экспрессия микроРНК не находятся в прямой зависимости от уровней экспрессии основных генов, участвующих в процессинге или действии микроРНК, коими являются DROSHA, DGCR8, DICER1, AGO2. Не было обнаружено строгих закономерностей в уровнях экспрессии микроРНК и isomiR.

Однако очевидно, что может существовать косвенная связь между уровнями экспрессии «ключевых» генов и микроРНК. Об этом говорит то, что некоторые isomiR оказались дифференциально ко-экспрессированы сразу при нескольких условиях. Так, hsa-miR-93-5p|+3, hsa-miR-23a-3p|+1 и hsa-miR-140-3p|+1 изменили корреляцию с большим числом своих мишеней при изменении уровней экспрессии DROSHA, DGCR8 и AGO2.

Для более глубокого понимания процессов, связывающих уровень экспрессии микроРНК и isomiR с уровнем экспрессии участвующих в их образовании генов, необходимы дальнейшие исследования. Например, проведённый в данной работе анализ может быть расширен на другие подтипы рака молочной железы. Также может быть полезной попытка учесть наличие у рассмотренных «ключевых» генов альтернативного сплайсинга или гомологов (например, семейство AGO1/2/3). Не менее важным является проведение аналогичной работы с использованием данных о других опухолях, например, TCGA-PRAD или TCGA-COAD.

Не менее ценной может быть попытка изменить дизайн эксперимента. Например, можно сравнивать корреляции не между группами с разным

уровнем экспрессии определенного гена, а между здоровой тканью и опухолью или между подтипами рака. Подобное исследование уже проведено для рака простаты [19], и эту идею можно расширить, применив в других видах опухолей.

Список использованных источников

1. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018;172:650–65.

2. Chipman LB, Pasquinelli AE. miRNA Targeting: Growing beyond the Seed. Trends Genet. 2019;35:215–22.

3. Tomasello L, Distefano R, Nigita G, Croce CM. The MicroRNA Family Gets Wider: The IsomiRs Classification and Role. Front Cell Dev Biol. 2021;9:1–15.

4. Chen J, Zhao X, Yuan Y, Jing J-J. The expression patterns and the diagnostic/prognostic roles of PTPN family members in digestive tract cancers. Cancer Cell Int. 2020;20:238.

5. Ma H-P, Chang H-L, Bamodu OA, Yadav VK, Huang T-Y, Wu ATH, et al. Collagen 1A1 (COL1A1) Is a Reliable Biomarker and Putative Therapeutic Target for Hepatocellular Carcinogenesis and Metastasis. Cancers (Basel). 2019;11:786.

6. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. Nature Publishing Group; 2012;8:1–9.

7. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. Cell. 2004;116:281–97.

8. Kloosterman WP, Plasterk RHA. The Diverse Functions of MicroRNAs in Animal Development and Disease. Dev Cell. 2006;11:441–50.

9. Iwakawa H oki, Tomari Y. Life of RISC: Formation, action, and degradation of RNA-induced silencing complex. Mol Cell. Elsevier Inc.; 2022;82:30–43.

10. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. Nature Publishing Group; 2014;15:509–24.

11. Ye ZL, Jin HJ, Qian QJ. Argonaute 2: A novel rising star in cancer research. J Cancer. 2015;6:877–82.

12. Chu Y, Kilikevicius A, Liu J, Johnson KC, Yokota S, Corey DR. Argonaute binding within 3′-untranslated regions poorly predicts gene repression. Nucleic Acids Res. Oxford University Press; 2020;48:7439–53.

13. Piroozian F, Bagheri Varkiyani H, Koolivand M, Ansari M, Afsa M, AtashAbParvar A, et al. The impact of variations in transcription of DICER and AGO2 on exacerbation of childhood B-cell lineage acute lymphoblastic leukaemia. Int J Exp Pathol. 2019;100:184–91.

14. de Sousa MC, Gjorgjieva M, Dolicka D, Sobolewski C, Foti M. Deciphering miRNAs' action through miRNA editing. Int J Mol Sci. 2019;20.

15. Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I. Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. Nucleic Acids Res. 2015;43:9158–75.

16. Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I. Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. Nucleic Acids Res. 2017;45:2973–85.

17. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.

18. Nersisyan S, Galatenko A, Chekova M, Tonevitsky A. Hypoxia-Induced miR-148a Downregulation Contributes to Poor Survival in Colorectal Cancer. Front Genet. 2021;12:1–9.

19. Zhiyanov A, Engibaryan N, Nersisyan S, Shkurnikov M, Tonevitsky A. Differential co-expression network analysis with DCoNA reveals isomiR targeting aberrations in prostate cancer. Bioinformatics. 2023;39.

20. Keller MP, Choi YJ, Wang P, Davis DB, Rabaglia ME, Oler AT, et al. A

gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res. 2008;18:706–16.

21. Riolo G, Cantara S, Marzocchi C, Ricci C. miRNA targets: From prediction tools to experimental validation. Methods Protoc. 2021;4:1–20.

22. Shkurnikov M, Nikulin S, Nersisyan S, Poloznikov A, Zaidi S, Baranova A, et al. LAMA4-Regulating miR-4274 and Its Host Gene SORCS2 Play a Role in IGFBP6-Dependent Effects on Phenotype of Basal-Like Breast Cancer. Front Mol Biosci. 2019;6:1–7.

23. Magee RG, Telonis AG, Loher P, Londin E, Rigoutsos I. Profiles of miRNA Isoforms and tRNA Fragments in Prostate Cancer. Sci Rep. Springer US; 2018;8:1–13.

24. McKenzie AT, Katsyv I, Song W-M, Wang M, Zhang B. DGCA: A comprehensive R package for Differential Gene Correlation Analysis. BMC Syst Biol. 2016;10:106.

25. Fukushima A. DiffCorr: An R package to analyze and visualize differential correlations in biological networks. Gene. 2013;518:209–14.

26. Fisher RA. Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. Biometrika. 1915;10:507.

27. Lin W, Ji J, Zhu Y, Li M, Zhao J, Xue F, et al. PMINR: Pointwise Mutual Information-Based Network Regression – With Application to Studies of Lung Cancer and Alzheimer's Disease. Front Genet. 2020;11:1–12.

28. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.

29. Song WM, Zhang B. Multiscale Embedded Gene Co-expression Network Analysis. PLoS Comput Biol. 2015;11.

30. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? PLoS Comput Biol. 2011;7.

31. Zhang B, Gaiteri C, Bodea LG, Wang Z, McElwee J, Podtelezhnikov AA, et al. Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell. Elsevier Inc.; 2013;153:707–20.

32. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38:675–8.

33. Loher P, Karathanasis N, Londin E, Bray PF, Pliatsika V, Telonis AG, et al. IsoMiRmap: Fast, deterministic and exhaustive mining of isomiRs from short RNA-seq datasets. Bioinformatics. 2021;37:1828–38.

34. Nersisyan S, Zhiyanov A, Engibaryan N, Maltseva D, Tonevitsky A. A novel approach for a joint analysis of isomiR and mRNA expression data reveals features of isomiR targeting in breast cancer. Front Genet. 2022;13:1–10.

35. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57:289–300.

36. Onghena P. Randomization Tests or Permutation Tests? A Historical and Terminological Clarification. Randomization, Mask Alloc Concealment. Boca Raton : Taylor & Francis, a CRC title, part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic division of T&F Informa plc, 2018.: Chapman and Hall/CRC; 2017. p. 209–28.

37. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;

38. Loher P, Rigoutsos I. Interactive exploration of RNA22 microRNA target predictions. Bioinformatics. 2012;28:3322–3.

39. Nersisyan S. RNA22 and TargetScan predictions, negative controls

[Internet]. 2022.

40. Chia SK, Bramwell VH, Tu D, Shepherd LE, Jiang S, Vickery T, et al. A 50-Gene Intrinsic Subtype Classifier for Prognosis and Prediction of Benefit from Adjuvant Tamoxifen. Clin Cancer Res. 2012;18:4465–72.

41. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

Приложения

Приложение А

Python-классы и функции, использованные в работе

Использовался Python 3.9.14

0. Библиотеки, импортируемые для работы нижеописанных классов и функций:

```
import multiprocessing # Для запуска dcona в фоновом процессе
import gspread_pandas # Для интеграции Pandas с Google Sheets
import telegram_send # Для отправки уведомлений в telegram-бот
import pandas as pd
import numpy as np
import pathlib
import dcona
from datetime import datetime # Для фиксации затраченного на расчёты времени
```

1. Класс `DataProcessor`. Используется для быстрой предобработки и фильтрации «сырых» данных. При инициализации экземпляра на вход подаются три Pandas датафрейма: с экспрессией транскриптов, с аннотацией образцов (включающей подтипы рака) и с экспрессией микроРНК. Далее последовательно вызываются методы, фильтрующие матрицы экспрессий: `cutoff_expressions` и `cutoff_mirnas`. Наконец, метод `final_data` принимает на вход аннотацию для DCoNA и таблицу предсказаний мишеней микроРНК, а возвращает объединенную матрицу экспрессий транскриптов и микроРНК, итоговую аннотацию DCoNA и таблицу предсказаний мишеней. Данный класс использует некоторые вспомогательные функции, приведенные в пункте 5.

```
class DataProcessor:
    def __init__(self, expression_df: pd.DataFrame, annotation_df: pd.DataFrame,
mirna_df: pd.DataFrame):
        self.annotation = annotation_df.copy()
        self.gene_exp = expression_df.copy()
        self.mirna = mirna_df.copy()

    def cutoff_expressions(self, subtypes: list, cutoff: float = 95):
        ids = self.annotation.loc[self.annotation['Sample
type'].isin(subtypes)].index.to_numpy()
        exp_all = extract_samples(self.gene_exp, ids, additional_columns=['gene
symbol'])
```

```
        exp = filter_by_median(exp_all, additional_columns=['gene symbol'],
cutoff=cutoff)
        self.cutted_exp = remove_transcript_version(exp)
        print(self.cutted_exp.shape[0], 'genes and', self.cutted_exp.shape[1],
'samples are chosen')

    def cutoff_mirnas(self, cutoff: float = 99):
        mirna_5prime = isomir_groupby_5prime(self.mirna)
        self.mirna_5prime = filter_by_median(mirna_5prime,
additional_columns=['median'], cutoff=cutoff)
        print(self.mirna_5prime.shape[0], 'miRNAs are chosen')

    def final_data(self, description_df: pd.DataFrame, raw_interaction_df:
pd.DataFrame):
        self.expr_united =
pd.concat(tcga_match_samples(self.cutted_exp.drop('gene symbol', axis=1),
self.mirna_5prime))
        self.interaction_df = raw_interaction_df[['isomir',
'transcript']].drop_duplicates()
        self.interaction_df.columns = ['Source', 'Target']
        return brush_data_before_dcona(self.expr_united, description_df,
self.interaction_df)

    def remove_transcript_version(expr):
        expr.index = expr.index.map(lambda x: x.split('.')[0])
        return expr

    def tcga_match_samples(df1, df2):
        tech_cols1 = [c for c in df1.columns if not c.startswith("TCGA")]
        tech_cols2 = [c for c in df2.columns if not c.startswith("TCGA")]
        common_samples = sorted(list(set(df1.columns) & set(df2.columns)))
        return df1[tech_cols1 + common_samples], df2[tech_cols2 + common_samples]
```

2. Класс `CuttingSamples`. Необходим для разделения выделения образцов TCGA по уровню экспрессии выбранного гена. При создании экземпляра класса ему в качестве аргумента передаётся Pandas-датафрейм, содержащий матрицу экспрессий. Далее вызывается метод `CuttingSamples.cut()`, аргументами которого являются название транскрипта, по уровню экспрессии которого будут отобраны образцы, и количество квантилей, на которые нужно разбить экспрессию этого гена (по-умолчанию, q=10). В результате вызова метода возвращается список квантилей с диапазонами экспрессий и количеством попавших в каждый диапазон образцов, а также отрисовывается гистограмма

экспрессии этого гена. Вызвав метод `CuttingSamples.extract_samples()`, можно указать, сколько верхних и нижних квантилей будет выбрано – результатом является таблица, содержащая аннотацию образцов с метками `Low_expr` для слабоэкспрессированных образцов («нижних» квантилей) и `High_expr` для высокоэкспрессированных образцов. Наконец, для сохранения этой таблицы можно вызвать метод `CuttingSamples.save_annotation()`, передав название гена, транскрипт которого был использован для разделения образцов, и путь к папке для сохранения аннотации.

```
class CuttingSamples:
    def __init__(self, exp_df, extra_columns=None):
        self._exp_df = exp_df
        self._extra_columns = extra_columns

    def cut(self, transcript, q=10):
        self.__transcript = transcript
        df = self._exp_df.loc[[transcript]].drop(self._extra_columns, axis=1)
        series = df.T.squeeze()
        self._qcut_df = pd.qcut(series, q=10, duplicates='drop')
        df.T.plot.hist(bins=20, alpha=0.5)
        print(self._qcut_df.value_counts(sort=False))

    def extract_samples(self, left_num=1, right_num=1):
        left_categories = self._qcut_df.cat.categories[:left_num]
        right_categories = self._qcut_df.cat.categories[-right_num:]
        left_samples = self._qcut_df.loc[self._qcut_df.isin(left_categories)]
        right_samples = self._qcut_df.loc[self._qcut_df.isin(right_categories)]
        left_df = self._df_sample_group(left_samples, "Low_expr")
        right_df = self._df_sample_group(right_samples, "High_expr")
        self.__cutted_samples = pd.concat([left_df,
right_df]).reset_index(drop=True)
        return self.__cutted_samples

    def _df_sample_group(self, samples, group):
        df = pd.DataFrame({
            "Sample": list(samples.index),
            "Group": [group]*len(samples)
        })
        return df

    def save_annotation(self, gene_name,
directory="/home/jovyan/shared/narek/outputs/sample_cuts"):
        filename =
f"{gene_name}_{self.__transcript}_{len(self.__cutted_samples)}_samples.csv"
```

```
        path = pathlib.Path(directory)/pathlib.Path(filename)
        if path.exists():
            print('File already exists:', path)
            return
        self.__cutted_samples.to_csv(path)
        print('Annotation is saved at:', path)
```

3. Класс `MetaExporter`. Используется для синхронизации информации о каждом запуске в таблицах Google Sheets. Для первичной настройки необходимо создать сервисный аккаунт в Google Cloud и предоставить ему доступ к Google-таблице. Атрибут `MetaExporter.df` даёт возможность получить доступ к таблице, а методы `MetaExporter.create_row()` и `MetaExporter.send_row()` позволяют создать пустую строку по шаблону таблицы и записать строку в таблицу, соответственно.

```
class MetaExporter:
    def __init__(self,
account_file='/home/jovyan/diploma_scripts/.service_account.json'):
        my_config =
gspread_pandas.conf.get_config('/'.join(account_file.split('/')[:-1]),
account_file.split('/')[-1])
        self.spread = gspread_pandas.Spread('DCoNA_runs', config=my_config)
        self.df

    @property
    def df(self):
        df = self.spread.sheet_to_df()
        df.index = df.index.astype(int)
        self._df_cache = df
        return df

    def create_row(self, dct={}):
        row = pd.DataFrame(data=dct, index=[0], columns=self._df_cache.columns)
        return row

    def send_row(self, row, row_number):
        df = self.df
        row.index = [row_number]
        if row_number in df.index:
            df.loc[row_number] = row.to_numpy()[0]
        else:
            df = pd.concat([df, row], ignore_index=False)
        df = df.sort_index()
        self.spread.df_to_sheet(df, freeze_headers=True)
        return df
```

4. Класс `Experiment`. Хранит данные о проводимом эксперименте, синхронизируя их с Google-таблицей через экземпляр `MetaExporter` и позволяет запускать DCoNA внутри Jupyter Notebook в фоновом режиме, не блокируя выполнение других ячеек с кодом. При создании экземпляра класса в качестве аргумента передаются первичные данные: номер эксперимента, тип(ы) и подтип(ы) рака, источник таблицы взаимодействий микроРНК-мРНК и другие метаданные. С помощью метода `Experiment.run_ztest()` можно запустить функцию *ztest()* пакета *dcona*, передав аргументы к функции в словаре `ztest_kwargs`. При этом для уведомления о завершении расчётов используется пакет *telegram-send*, отправляющий сообщения в преднастроенный telegram-бот.

```
class Experiment:
    def __init__(self, *args, **kwargs):
        self._init_args = args
        self._init_kwargs = kwargs
        self.gspread = MetaExporter()
        self._run_number = (int(self.gspread._df_cache.index.max()) + 1 if
len(self.gspread._df_cache) else 1)
        self._meta_row = self.gspread.create_row(kwargs)
        self.replace_in_row('status', 'New')

    def ztest_wrapped(self, *args, **kwargs):
        self._ztest_started = datetime.now()
        self.replace_in_row('ztest_start', self._ztest_started)
        telegram_send.send(messages=[f'Ztest started, experiment
{self._init_kwargs["experiment"]}'])
        try:
            args_to_export = ['reference_group', 'experimental_group']
            self.replace_in_row('ztest_args', {key: kwargs['ztest_kwargs'][key]
for key in args_to_export})
            self.replace_in_row('status', 'Running')
            self.ztest_result = dcona.ztest(**kwargs['ztest_kwargs'])
            self._ztest_finished = datetime.now()
            self._ztest_time_spent = (self._ztest_finished) -
(self._ztest_started)
            dirpath =
pathlib.Path(f'/home/jovyan/shared/narek/outputs/experiments/{self._init_kwargs["
experiment"]}')
            dirpath.mkdir(parents=True, exist_ok=True)
            filepath = dirpath/kwargs["filename"]
```

```
            self.ztest_result.to_csv(filepath)
            telegram_send.send(messages=[f'Ztest performed (exp
{self._init_kwargs["experiment"]}), time spent {self._ztest_time_spent}'])
            self.replace_in_row('ztest_path', filepath)
            self.replace_in_row('ztest_end', self._ztest_finished)
            self.replace_in_row('ztest_time', self._ztest_time_spent)
            self.replace_in_row('status', 'Finished')
        except Exception as exc:
            self.replace_in_row('status', 'Error')
            telegram_send.send(messages=[f'Error occured:\n {exc}'])

    def run_ztest(self, *args, **kwargs):
        multiprocessing.Process(target=self.ztest_wrapped, args=args,
kwargs=kwargs).start()

    def replace_in_row(self, column, value):
        self._meta_row.iloc[0].loc[column] = value
        self.gspread.send_row(self._meta_row, self._run_number)
```

5. Были также использованы вспомогательные функции:

a. Для извлечения из матрицы экспрессий необходимых образцов по списку. Возможно также указание дополнительных вспомогательных столбцов в отдельном списке.

```
def extract_samples(df, sample_ids, additional_columns=None):
    if additional_columns:
        return df.loc[:, np.concatenate((additional_columns, sample_ids))]
    return df.loc[:, sample_ids]
```

b. Для фильтрации матрицы экспрессий по медиане. По умолчанию отсекаются гены/транскрипты, с медианой равной нулю. Однако при необходимости есть возможность ввести более строгие критерии. Например, можно отобрать транскрипты, составляющие N% всей экспрессии. Для определения этой отсечки используется кумулятивная сумма по медиане. Также есть возможность вручную указать максимальное количество транскриптов – тогда будет отобран топ наиболее экспрессированных молекул.

```
def filter_by_median(df, additional_columns=None, max_rows=None, cutoff=None):
    filtering_df = df.copy(); df_out = df.copy()
    if additional_columns:
        filtering_df = filtering_df.drop(additional_columns, axis=1)
    df_out['median'] = filtering_df.median(axis=1)
    df_out = df_out.loc[df_out['median'] > 0]
    df_out.sort_values(['median'], ascending=False, inplace=True)
```

```
    if cutoff:
        temp_df = df_out.drop(additional_columns, axis=1)
        temp_df = 2**temp_df - 1
        cutoff_value =
temp_df.median(axis=1).sort_values(ascending=False).cumsum()[-1] * float(cutoff)
/ 100
        df_out =
df_out.loc[temp_df.median(axis=1).sort_values(ascending=False).cumsum() <
cutoff_value]
    if max_rows:
        df_out = df_out.iloc[:max_rows, :]
    try:
        df_out.drop('median', axis=1, inplace=True)
    except: pass
    return df_out
```

c. Для агрегации значений экспрессий изоформ микроРНК с одинаковым 5’-концом и разными 3’-концами. Необходимо для упрощения работы с данными IsoMiRmap при условии, что мы не рассматриваем 3’-вариации изоформ.

```
def isomir_groupby_5prime(isomir):
    isomir["median"] = isomir.iloc[:, 3:].median(axis=1)
    isomir = isomir.sort_values("median", ascending=False)
    # isomir = isomir.loc[isomir["median"] > 0]
    isomirs_with_mirdb =
set(open("/home/jovyan/shared/miRNA_predictions_BRCA/isomiRs_with_miRDB.txt").rea
d().strip().split())
    isomir.iloc[:, 3:] = 2**isomir.iloc[:, 3:] - 1
    mature_5p = []
    for mm in isomir["mature"]:
        if type(mm) != str:
            mature_5p.append(None)
            continue

        has_mirdb = False
        for m in mm.split(", "):
            m_5p = "|".join(m.split("|")[:-1])
            if m_5p in isomirs_with_mirdb:
                has_mirdb = True
                mature_5p.append(m_5p)
                break

        if not has_mirdb:
            mature_5p.append(None)

    isomir["mature"] = mature_5p
    isomir = isomir.loc[isomir["mature"].notna()]
    isomir = isomir.drop(columns=["hairpin", "repeat", "median"])
```

```
    isomir = isomir.groupby("mature").sum()
    isomir = np.log2(isomir + 1)
    isomir["median"] = isomir.median(axis=1)
    isomir = isomir.loc[isomir["median"] > 0]
    isomir = isomir.sort_values("median", ascending=False)
    return isomir
```

d. Для приведения данных для ztest (таблиц экспрессии, аннотаций и взаимодействий) в соответствие друг другу. Во всех таблицах отбираются только общие транскрипты и образцы.

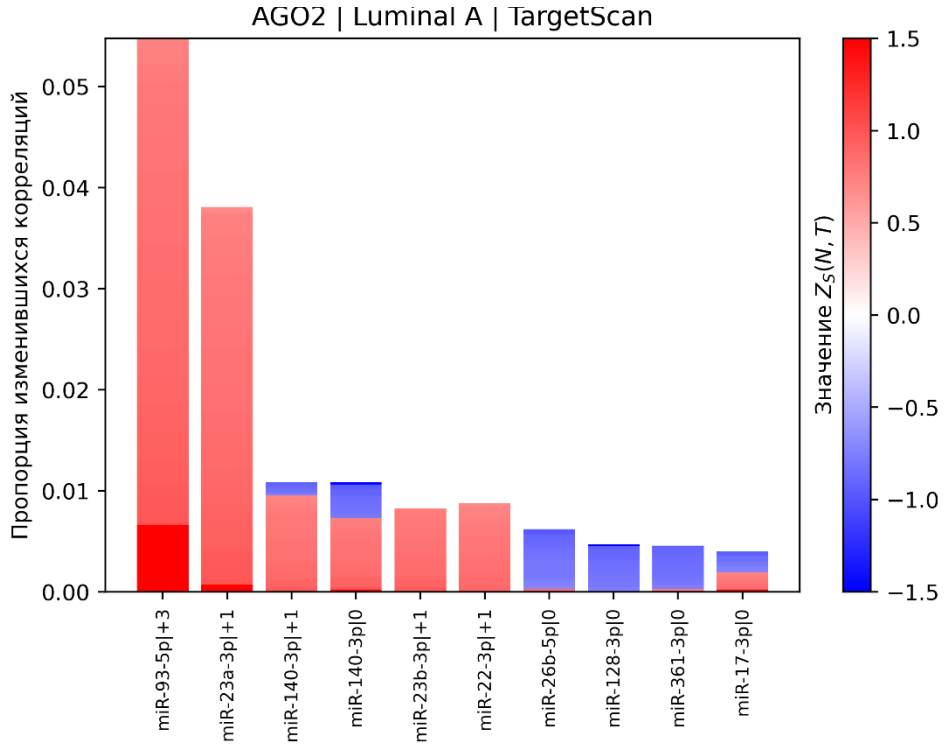
```
def brush_data_before_dcona(expr, annotation, interaction):
    expr, annotation, interaction = expr.copy(), annotation.copy(),
interaction.copy()
    common_samples = list(set(expr.columns) & set(annotation["Sample"]))
    common_genes = list((
            set(interaction['Source'].unique()) |
            set(interaction['Target'].unique())
        ) & set(expr.index))
    common_mirnas = [c for c in common_genes if c.startswith("hsa")]
    common_transcripts = [c for c in common_genes if c.startswith("ENS")]
    interaction = interaction.loc[interaction['Source'].isin(common_mirnas) &
interaction['Target'].isin(common_transcripts)]
    annotation = annotation.loc[annotation['Sample'].isin(common_samples)]
    expr = expr.loc[common_genes, common_samples]
    return expr, annotation, interaction
```

Приложение Б

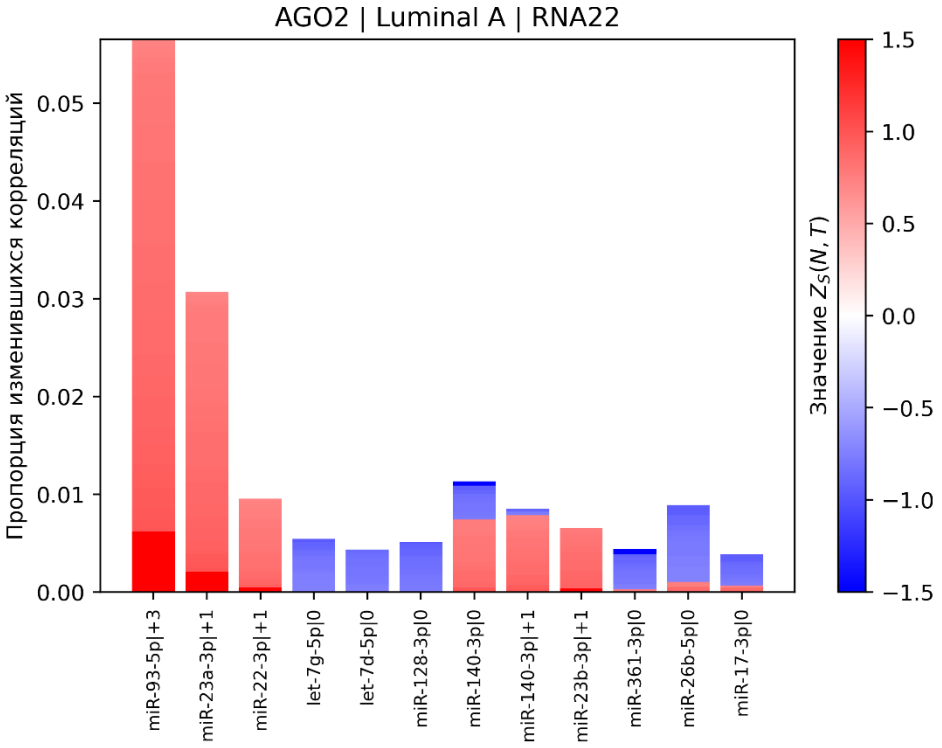
Число изменившихся корреляций, нормализованное на общее количество мишеней

Синий цвет на графиках отображает потерю отрицательной корреляции при высоком уровне экспрессии ключевого гена. Красный цвет отражает приобретение новых отрицательных корреляций.

1. Ген – AGO2
   1.1. Подтип РМЖ - Luminal A
      1.1.1. Источник мишеней микроРНК – TargetScan



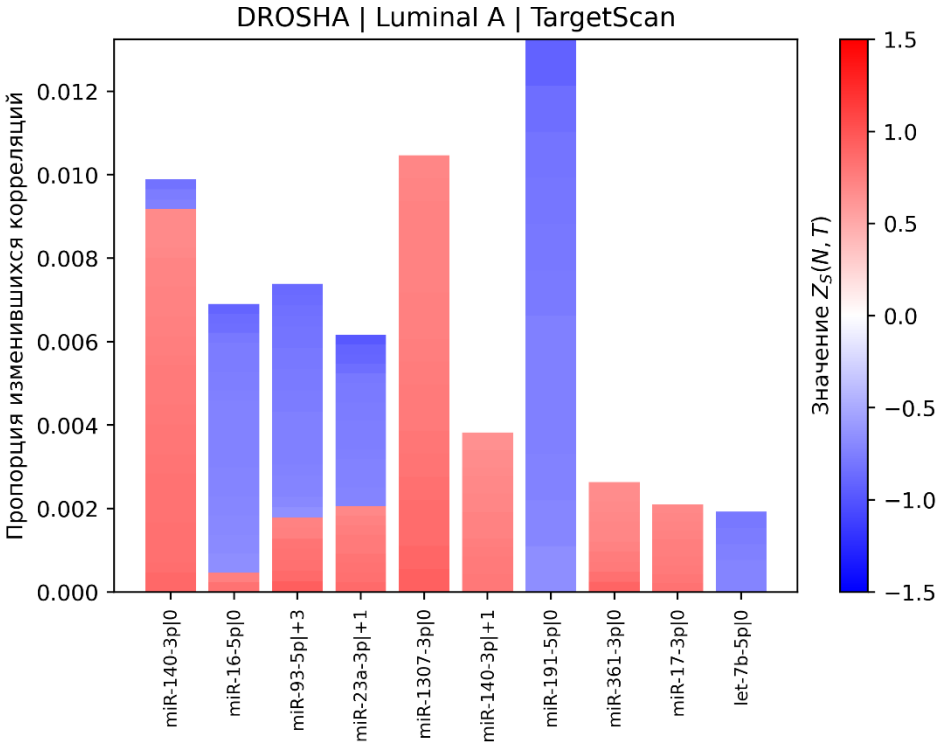
1.1.2. Источник мишеней микроРНК – RNA22



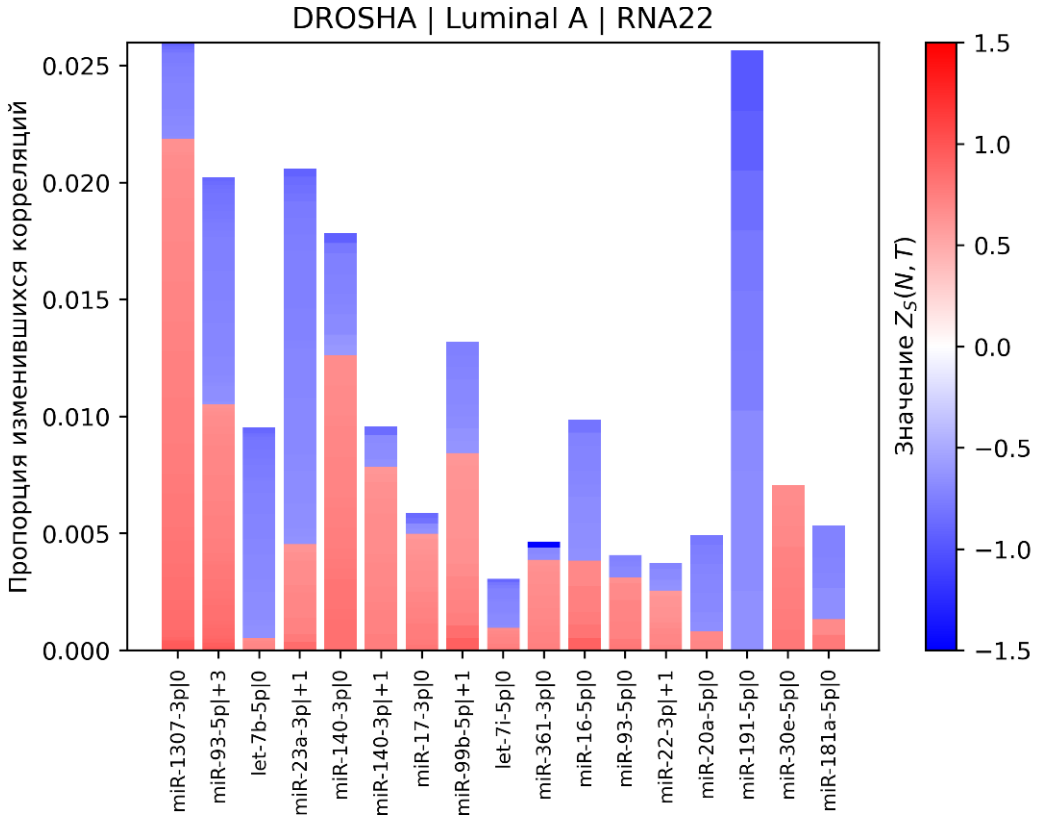
2. Ген – DROSHA

2.1.Подтип РМЖ - Luminal A

2.1.1. Источник мишеней микроРНК – TargetScan



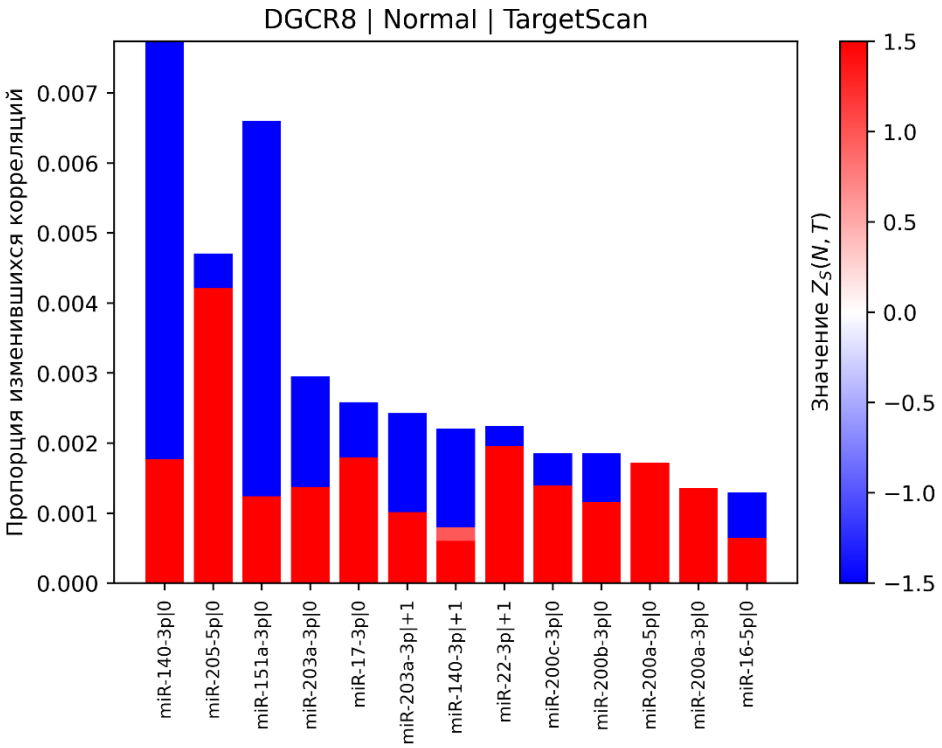
2.1.2. Источник мишеней микроРНК – RNA22



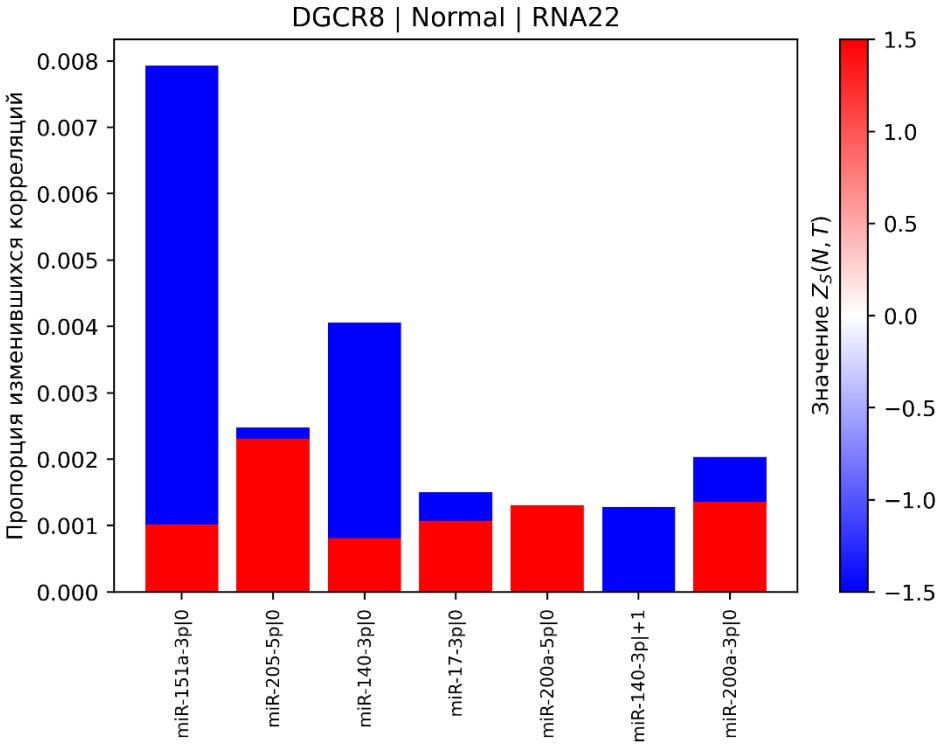
3. Ген – DGCR8

3.1.Подтип РМЖ – Normal

3.1.1. Источник мишеней микроРНК – TargetScan

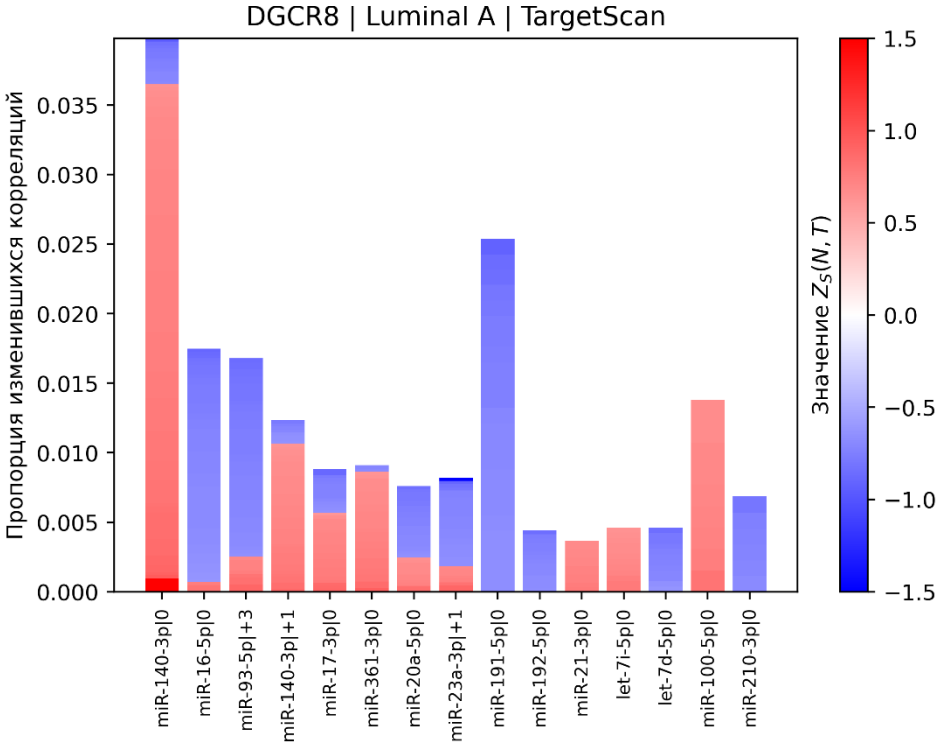


3.1.2. Источник мишеней микроРНК – RNA22



3.2.Подтип РМЖ - Luminal A

3.2.1. Источник мишеней микроРНК – TargetScan



4. Ген – DICER1

4.1.Подтип РМЖ - Luminal A

4.1.1. Источник мишеней микроРНК – RNA22