Министерство науки и высшего образования Российской Федерации

Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт радиоэлектроники и информационных технологий – РТФ

Школа бакалавриата

ОТЧЕТ

По проекту
«Проведения исследования на стыке медицины и химии»

по дисциплине «Проектный практикум»

Заказчик: Фамилия И.О.	Ильинский Александр Дмитриевич
Куратор: Фамилия И.О. ученая степень, ученое звание, должность	Ильинский Александр Дмитриевич
Студенты команды Мутаген	Ященко Даниил
Фамилия И.О.	Николаевич
Фамилия И.О.	Булатова Дарья Дмитриевна
Фамилия И.О.	Выборнов Илья Владимирович

Екатеринбург, 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Целью данного проекта является применение методов машинного обучения и статистического анализа для выявления скрытых закономерностей и потенциальных биомаркеров в молекулярных и клинических данных пациентов с раком предстательной железы. В частности, исследование направлено на анализ взаимосвязей между экспрессией miRNA, мРНК и клиническими характеристиками, с целью улучшения диагностики, прогноза и персонализации терапии.

Проект решает следующие задачи:

1) Поиск и отбор открытых наборов данных, содержащих информацию о miRNA и RNA-seq экспрессии, а также сопутствующие клинические параметры (включая данные TCGA).
1) Объединение, предобработка и очистка данных: удаление пропусков, устранение выбросов, нормализация.
2) Анализ признаков, таких как стадия заболевания, возраст пациентов, наличие поражённых лимфатических узлов и статус выживаемости (жив/умер).
3) Формирование и проверка гипотез о связи отдельных miRNA и генов с прогрессией опухоли, рецидивом и выживаемостью пациентов.
4) Группировка пациентов на основе молекулярных признаков и выявление дифференциально экспрессированных miRNA, ассоциированных с неблагоприятным прогнозом.
5) Визуализация результатов (boxplot, volcano plot, PCA, тепловые карты) и документирование рабочих гипотез с использованием инструментов Python, R, Pandas, Bioconductor и других библиотек.

Актуальность проекта определяется растущей потребностью в более точных и ранних методах диагностики онкологических заболеваний, в том числе рака предстательной железы. Традиционные подходы зачастую не учитывают молекулярные особенности пациентов, в то время как использование транскриптомных данных и современных аналитических

методов позволяет обнаружить новые прогностические маркеры и повысить индивидуализацию лечения.

Область применения результатов проекта охватывает биоинформатику, молекулярную онкологию и клинические исследования. Разработанные методики и выявленные биомаркеры могут быть использованы для дальнейших исследований и интеграции в решения поддержки принятия врачебных решений.

Ожидаемые результаты проекта включают:

1) Сформированный набор miRNA и генов, статистически связанных с прогрессией опухоли и выживаемостью;
2) Классификацию/группировку пациентов на основании молекулярных признаков;
3) Комплект визуализаций и аналитических выводов для дальнейшего использования в исследовательской и прикладной практике;
4) Документированные рабочие гипотезы и научно обоснованные рекомендации.

1 Организация и реализация проекта

1.1 1. Распределение ролей в команде

В составе проектной команды было три участника, каждый из которых отвечал за отдельную область работы:

1) Ященко Д.Н. — проектный менеджер. Отвечал за планирование работы, распределение задач, координацию команды, составление документации и финальной презентации проекта. Также занимался поддержкой коммуникации с заказчиком и контролем соблюдения сроков.
2) Булатова Д.Д. — дата-аналитик. Выполняла сбор и очистку молекулярных и клинических данных, предварительную визуализацию, проверку рабочих гипотез, поиск потенциальных биомаркеров и дифференциально экспрессируемых генов и miRNA. Также отвечала за создание графиков и таблиц, подтверждающих статистические закономерности.
3) Выборнов И.В. — специалист по машинному обучению. Отвечал за исследование и реализацию ML-моделей для группировки пациентов по молекулярным подтипам заболевания и классификации образцов по стадиям развития опухоли на основе экспрессии значимых miRNA.

1.2 2. Требования заказчика и пользователей, план работ

Первоначальные требования от заказчика (куратора проекта) заключались в проведении анализа транскриптомных данных с формулировкой и статистической проверкой как минимум двух рабочих гипотез. Команда расширила изначальные цели, чтобы предложить более глубокую и комплексную реализацию. На основании этого был сформирован план разработки (backlog), включающий следующие ключевые этапы:

1) Поиск и загрузка подходящих наборов данных (miRNA-seq, RNA-seq, клинические данные);
2) Очистка и нормализация данных;

3) Формулировка и проверка гипотез о связи молекулярных признаков с клиническими характеристиками;
4) Проведение статистического анализа и визуализация результатов;
5) Разработка и обучение моделей машинного обучения для группировки и классификации пациентов;
6) Интерпретация результатов и подготовка рекомендаций.

Ориентировочная целевая аудитория проекта включает: биоинформатиков**,** молекулярных онкологов**,** аналитиков данных в медицине, а также исследователей в области персонализированной терапии.

1.3 3. Анализ аналогов

В процессе планирования проекта команда провела ознакомительный анализ аналогичных решений и ресурсов, применяемых в современной биоинформатике. Были рассмотрены такие открытые платформы, как:

1) cBioPortal — база данных и визуализатор молекулярных профилей раковых опухолей;
2) GEPIA (Gene Expression Profiling Interactive Analysis) — онлайн-инструмент для анализа RNA-seq данных из TCGA и GTEx;
3) UALCAN — портал для анализа экспрессии генов и выживаемости пациентов.

В отличие от перечисленных систем, которые в первую очередь ориентированы на исследование отдельных генов, наш проект сосредоточен на интегральном подходе: мы анализируем сочетание молекулярных и клинических факторов, осуществляем группировку пациентов по подтипам и акцентируем внимание на выявлении значимых miRNA, связанных с прогрессией рака.

1.4 4. Архитектура программного продукта

Программная реализация проекта выполнялась в среде Google Colab с использованием языка Python и библиотек: Pandas, NumPy, Seaborn, Matplotlib, Scikit-learn и других. Структура проекта включает следующие компоненты:

1) Модуль предобработки данных — чтение исходных файлов, нормализация, фильтрация, устранение пропусков и выбросов;
2) Модуль анализа гипотез — статистическая проверка предположений о роли miRNA как онкогенов или супрессоров, а также связь с клиническими показателями;
3) Модуль визуализации — построение графиков (boxplot, heatmap, PCA, volcano plot);
4) Модуль машинного обучения — кластеризация и классификация пациентов на основе экспрессии значимых признаков;
5) Документационный блок — вывод результатов, формулировка выводов и рекомендаций.

Выбор такой архитектуры обусловлен гибкостью, прозрачностью и простотой воспроизводимости научных вычислений, особенно при работе с биомедицинскими данными.

1.5 5. Методология и процесс разработки

Проектная команда использовала элементы итеративной методологии**,** с еженедельными встречами для оценки прогресса и корректировки плана. Все задачи фиксировались и отслеживались с помощью Excel-таблицы**,** которая служила внутренним трекером задач и дедлайнов.

В ходе разработки проводилось тестирование гипотез и моделей на промежуточных этапах. Например:

1) Был выявлен ряд miRNA с достоверной разницей в экспрессии между группами пациентов;
2) При обучении моделей использовалась перекрёстная проверка, что позволило минимизировать переобучение;
3) Некоторые изначальные гипотезы не подтвердились, что также было зафиксировано как часть научного анализа.

Были выявлены и устранены ошибки, связанные с некорректной нормализацией и несогласованностью меток в клинических данных.

1.6 6. Планирование и управление проектом

Все задачи распределялись по участникам в соответствии с их специализацией и требованиями этапа проекта. План-график работ составлялся в виде таблицы и включал разбивку по неделям с указанием ответственных лиц. Основное внимание уделялось соблюдению сроков и логической последовательности выполнения задач: от сбора данных — к анализу — к ML — к финальному отчёту и визуализациям.

2 Формулировка гипотез и выбор биомаркеров

Формулировка гипотез осуществлялась на основании анализа предметной области, литературных источников, консультаций с куратором проекта, а также предварительного ознакомления с данными. Основное внимание в исследовании было направлено на выявление микромолекул miRNA и генов, ассоциированных с прогрессией рака предстательной железы, прогнозом выживаемости и стадией заболевания. Ниже представлены ключевые рабочие гипотезы, положенные в основу анализа.

2.1 5.1. Гипотезы на основе miRNA

На основе экспрессии miRNA и её связи с клиническими параметрами (стадия T3/T4) были выделены ключевые кандидаты, потенциально вовлечённые в прогрессию РПЖ. Отрицательная корреляция указывает на пониженную экспрессию в более агрессивных стадиях заболевания. Некоторые микромолекулы miRNA могут выполнять функцию онкосупрессоров или онкогенов в зависимости от их уровня экспрессии, влияя на прогрессию опухоли.
Обоснование: в научной литературе описано участие miRNA в посттранскрипционной регуляции экспрессии генов, включая подавление онкогенов и инактивацию генов-супрессоров. Предполагается, что дисбаланс экспрессии отдельных miRNA может коррелировать с агрессивностью и стадией опухолевого процесса.
Проверка: сравнительный анализ уровня экспрессии miRNA между группами пациентов с различными клиническими признаками (стадия, наличие метастазов, статус выживаемости) с использованием статистических критериев и визуализации (boxplot, volcano plot).

Таблица 1 – miRNA

Название	Корреляция (Fold Change)	Небольшое описание
hsa-mir-30a	Отрицательная (0.973)	Регулирует апоптоз и пролиферацию. Пониженная экспрессия в T3/T4 может быть связана с прогрессией РПЖ.
hsa-mir-133a-1	Отрицательная (0.916)	Супрессор опухолей, ингибирует метастазирование. Снижение экспрессии в T3/T4 указывает на агрессивность.
hsa-mir-222	Отрицательная (0.918)	Регулирует клеточный цикл. Пониженная экспрессия в T3/T4 может способствовать прогрессии РПЖ.
hsa-mir-3676	Отрицательная (0.682)	Малоизученная miRNA, снижение экспрессии в T3/T4 может быть связано с онкогенезом.
hsa-mir-221	Отрицательная (0.942)	Регулирует пролиферацию и апоптоз. Снижение в T3/T4 может указывать на потерю контроля роста.
hsa-mir-891a	Отрицательная (0.777)	Связана с регуляцией сигнальных путей. Пониженная экспрессия в T3/T4 ассоциирована с прогрессией.
hsa-mir-133b	Отрицательная (0.869)	Супрессор опухолей, ингибирует инвазию. Снижение экспрессии в T3/T4 связано с метастазами.
hsa-mir-133a-2	Отрицательная (0.806)	Аналог hsa-mir-133a-1, снижение экспрессии в T3/T4 указывает на агрессивное течение РПЖ.
hsa-mir-217	Положительная (1.155)	ОнкомиР, повышенная экспрессия в T3/T4 связана с пролиферацией и прогрессией РПЖ.
hsa-mir-582	Отрицательная (0.948)	Регулирует иммунный ответ. Снижение экспрессии в T3/T4 может способствовать уклонению от иммунитета.

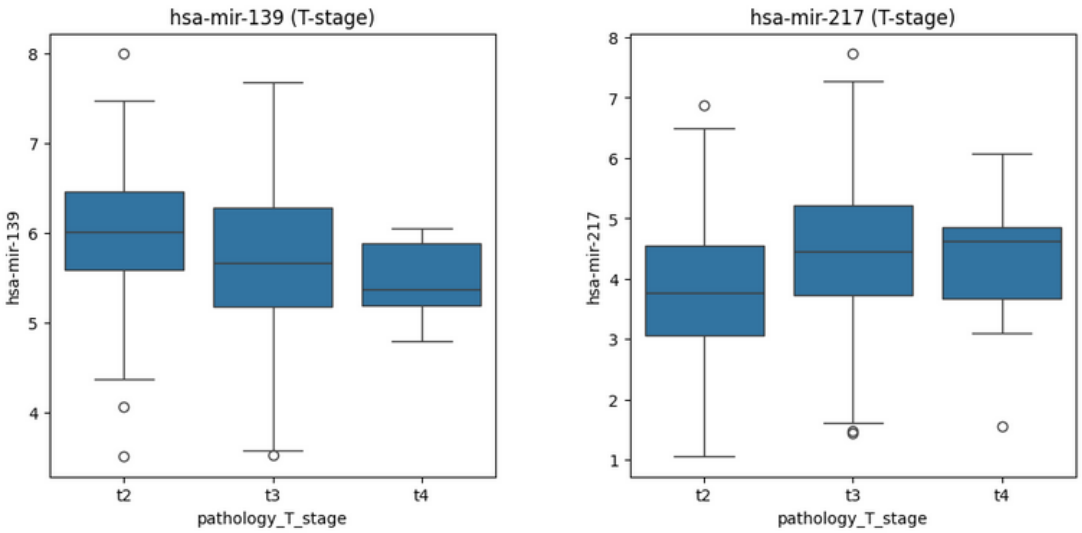

Рисунок 1 – miRNA boxplot

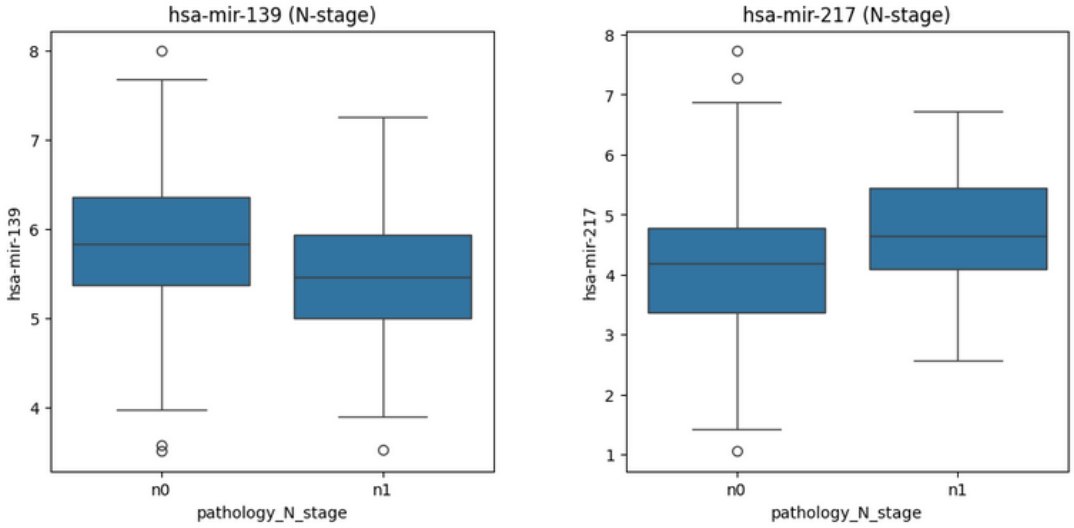

Рисунок 2 – miRNA boxplot

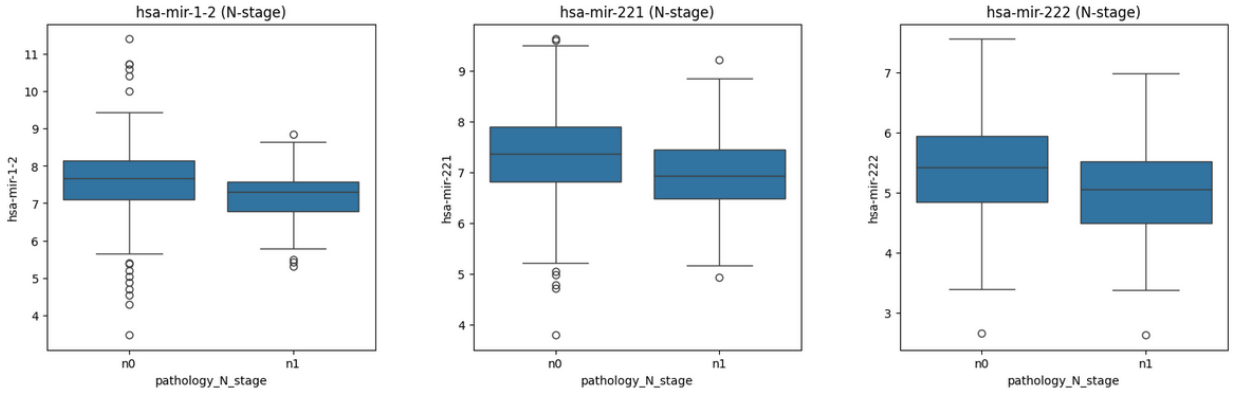

Рисунок 3 – miRNA boxplot

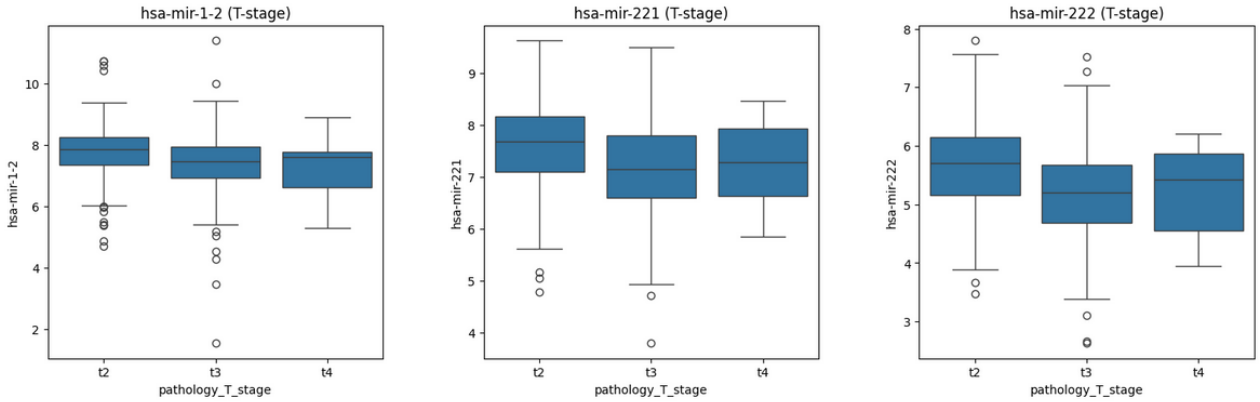

Рисунок 4 – miRNA boxplot

2.2 5.2. Гипотезы на основе экспрессии генов

Выявлены гены с экспрессией, значимо ассоциированной с выживаемостью. Транскрипционные факторы семейства ZNF, а также гены ZNHIT2 и ZP1, показали потенциальную прогностическую значимость. Существует набор сверхэкспрессированных или подавленных генов (на уровне RNA-seq), специфичных для поздней стадии рака предстательной железы.

Обоснование: развитие опухоли сопровождается нарушением транскрипции определённых генов. Предполагается, что гены, проявляющие устойчивые изменения экспрессии у пациентов с продвинутыми стадиями заболевания, могут служить потенциальными биомаркерами прогрессии. Проверка: нормализация данных RNA-seq, группировка по стадии заболевания и анализ дифференциальной экспрессии с использованием log2 fold change и поправки на множественные сравнения (FDR).

Таблица 2 – экспрессия генов

Название	Корреляция (LogFC)	Небольшое описание
ZNHIT2	Отрицательная (-0.225)	Регулирует транскрипцию, часть комплекса SNARP. Пониженная экспрессия связана с худшим выживанием в РПЖ.
ZP1	Положительная (0.580)	Белок зоны пеллюциды, роль в РПЖ неясна. Высокая экспрессия может быть связана с прогрессией.
ZNF77	Отрицательная (-0.073)	Транскрипционный фактор, регулирует экспрессию генов. Снижение экспрессии связано с плохим прогнозом.
ZNF419	Отрицательная (-0.076)	Транскрипционный фактор семейства C2H2. Пониженная экспрессия может указывать на прогрессию РПЖ.
ZNF565	Отрицательная (-0.057)	Регулятор транскрипции. Снижение экспрессии связано с онкогенезом в РПЖ.
ZNF431	Отрицательная (-0.065)	Транскрипционный репрессор. Пониженная экспрессия ассоциирована с худшим выживанием.
ZNF418	Отрицательная (-0.097)	Регулятор генной экспрессии. Снижение экспрессии может способствовать прогрессии РПЖ.

ZNF670	Отрицательная (-0.078)	Транскрипционный фактор. Пониженная экспрессия связана с агрессивным течением РПЖ.
ZNF692	Отрицательная (-0.070)	Регулятор транскрипции. Снижение экспрессии может быть маркером плохого прогноза.
ZNF594	Отрицательная (-0.050)	Транскрипционный фактор. Пониженная экспрессия ассоциирована с прогрессией РПЖ.

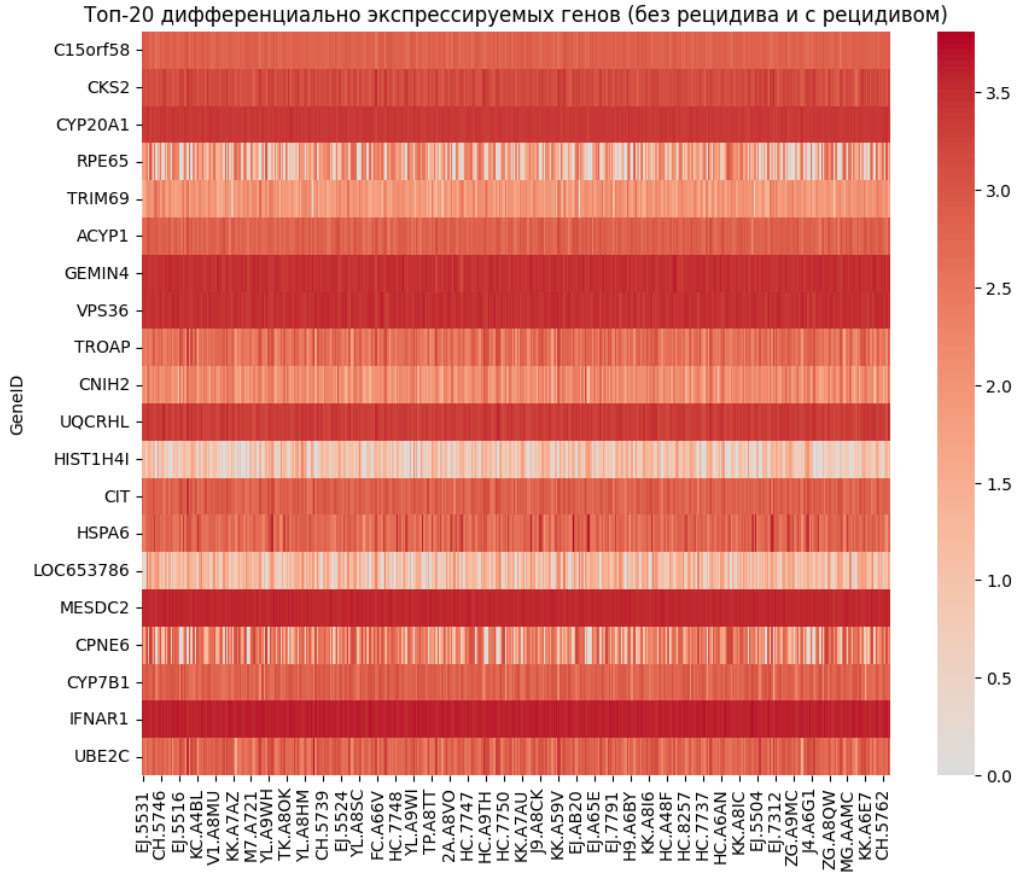

Рисунок 5 – дифференциальная экспрессия (без рецидива\рецидив)

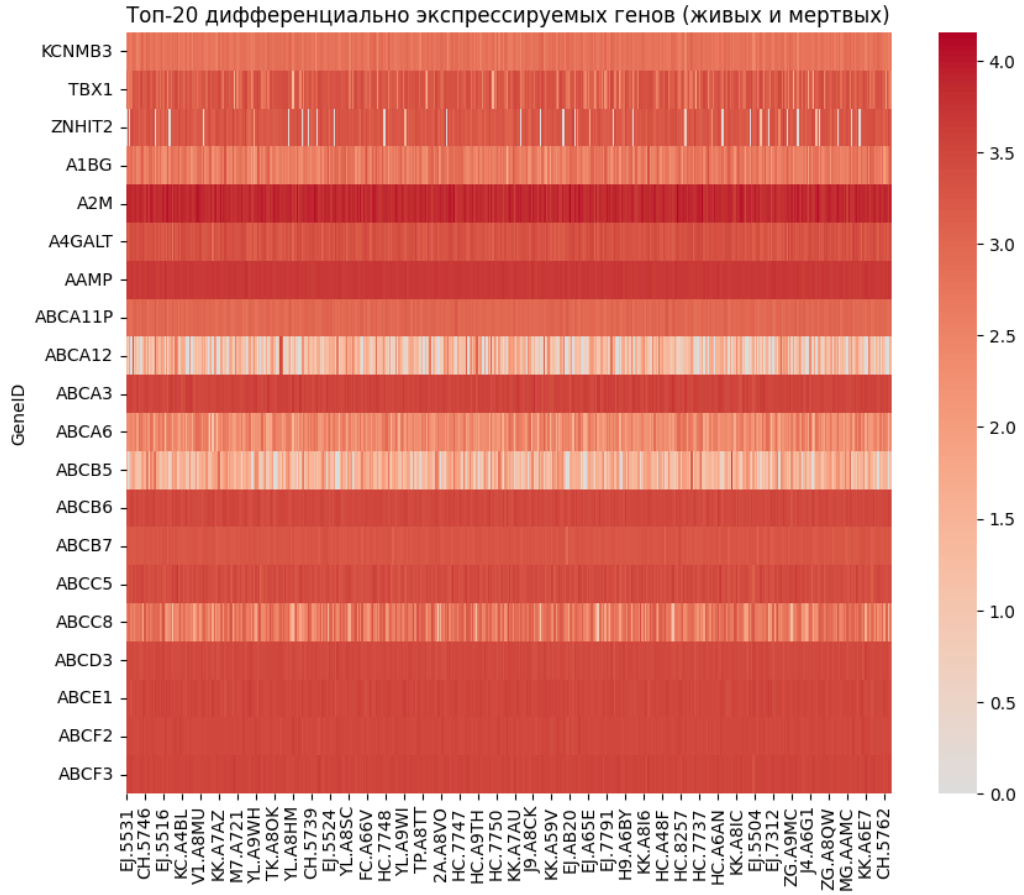

Рисунок 6 – дифференциальная экспрессия (жив\мертв)

2.3 5.3. Гипотезы на основе мутационного профиля

Проанализированы частоты соматических мутаций. Особое внимание уделено генам TP53 и SPOP, часто вовлеченным в патогенез РПЖ.

Таблица 3 – мутирующих генов

Название	Частота мутаций (%)	Роль в РПЖ и клиническая значимость
TTN	12.65	Кодирует титин, структурный белок. Высокая частота мутаций из-за большой длины гена. Роль в РПЖ неясна, вероятно, пассажирская мутация.
TP53	11.45	Супрессор опухолей. Мутации нарушают контроль клеточного цикла, связаны с агрессивным РПЖ и плохим прогнозом. Потенциальная мишень для терапии (например, APR-246).
SPOP	11.45	Регулятор деградации белков (убиквитин-лигаза). Мутации в субстрат-связывающем домене (F133, W131) стабилизируют онкогены (например, AR). Связаны с чувствительностью к ингибиторам андрогенов.
MUC16	7.43	Кодирует CA-125, мембранный гликопротеин. Мутации могут влиять на иммунный ответ. Роль в РПЖ требует изучения.

MUC17	6.22	Муцин, участвует в защите эпителия. Мутации могут способствовать метастазированию. Роль в РПЖ неясна.
MLL2 (KMT2D)	5.82	Гистонметилтрансфераза, регулирует эпигенетику. Мутации нарушают экспрессию генов, связанных с РПЖ. Потенциальная мишень для эпигенетической терапии.
MLL3 (KMT2C)	5.82	Аналог MLL2, регулирует эпигеном. Мутации связаны с прогрессией РПЖ.
FOXA1	5.62	Транскрипционный фактор, коактиватор андрогенового рецептора (AR). Мутации усиливают AR-сигнализацию, связаны с резистентностью к терапии.
SPTA1	5.22	Компонент цитоскелета эритроцитов. Роль в РПЖ неясна, возможно, пассажирская мутация.
SYNE1	5.22	Ядерный мембранный белок. Мутации могут нарушать ядерную архитектуру. Роль в РПЖ требует изучения.

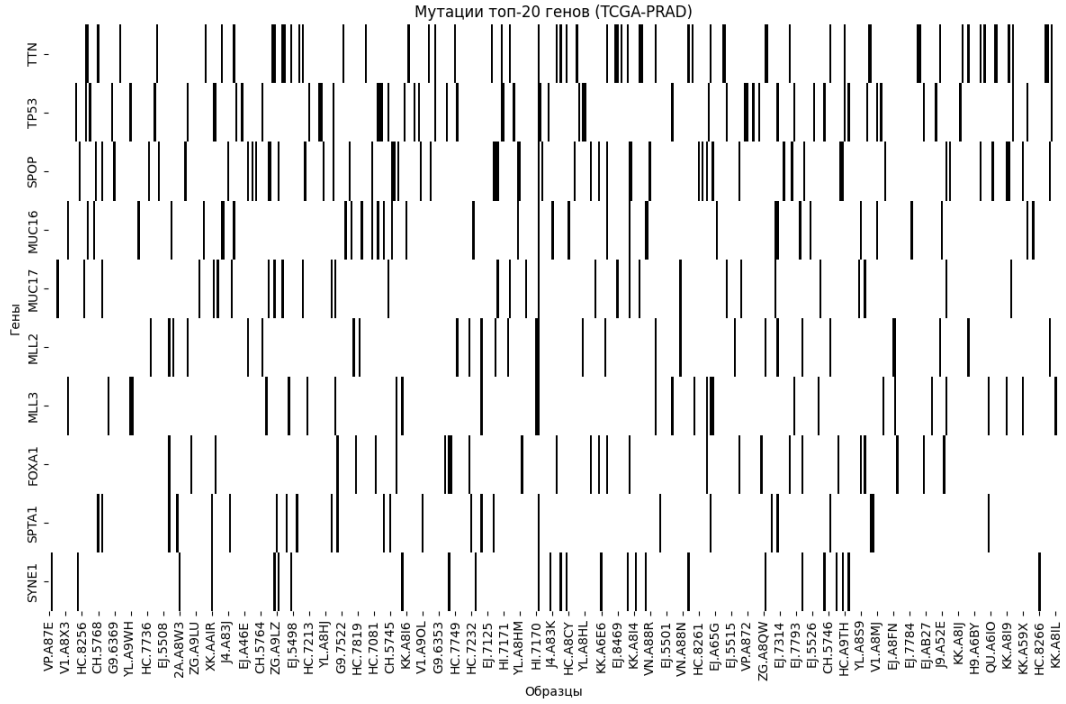

Рисунок 7 – топ 20 мутаций генов

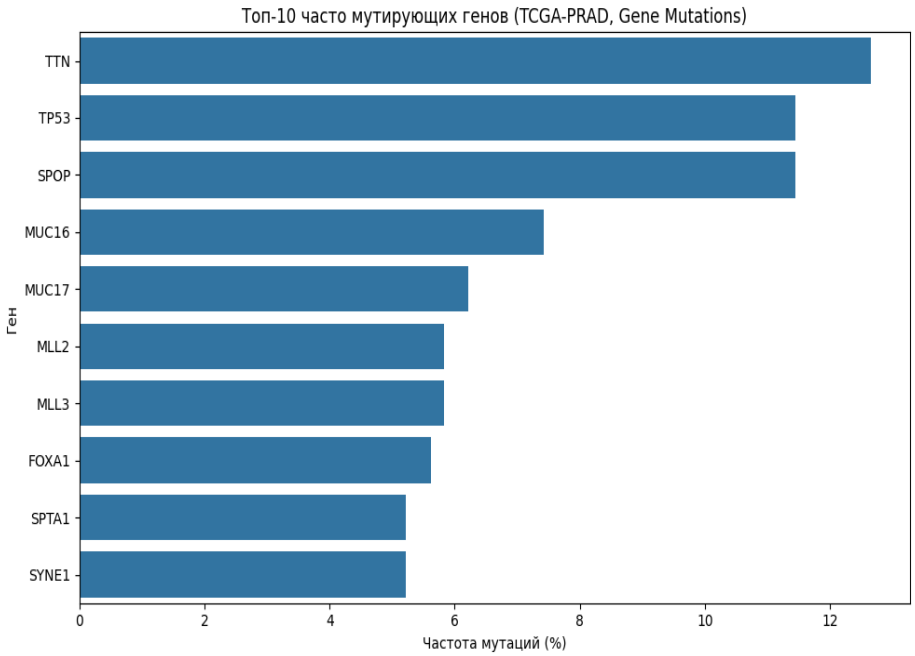

Рисунок 8 – топ частых мутирующих генов

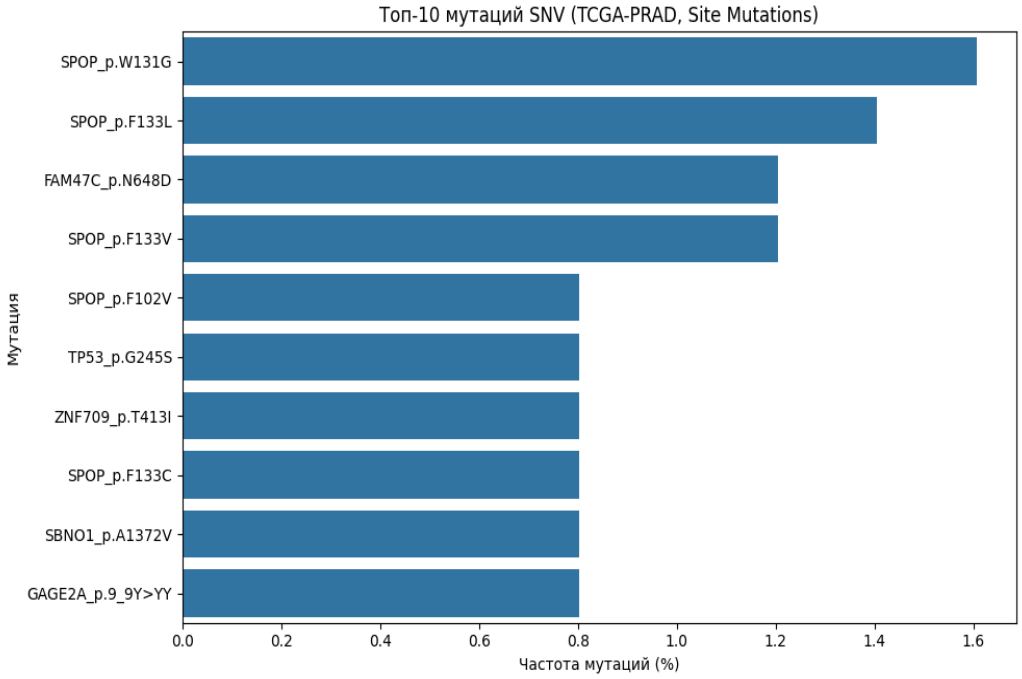

Рисунок 9 – топ мутаций SNV

3 Ограничения и допущения

В ходе анализа данных из проекта TCGA-PRAD пришлось учитывать ряд ограничений, связанных как с качеством и полнотой данных, так и с особенностями применяемых методов. В первую очередь, в клинической части набора данных наблюдалось большое количество пропущенных значений. В частности, часто отсутствовали данные о расе и этнической принадлежности пациентов, а также о статусе выживаемости — например, было недостаточно информации, чтобы надёжно сравнивать группы живых и умерших пациентов. Такие пропуски существенно ограничивали возможности проведения стратифицированного анализа и вынуждали либо удалять соответствующие строки, либо вовсе исключать переменные из дальнейшего рассмотрения. Это, в свою очередь, могло повлиять на полноту модели и снизить её обобщающую способность.

Кроме того, данные экспрессии генов (включая miRNA и miRNA seq) изначально имели высокую размерность, что создавало риски переобучения моделей. Также эти данные подвержены биологическому и техническому шуму, наличию выбросов и мультиколлинеарности между признаками. Многие из применённых статистических и машинных методов, такие как регрессионные модели, методы снижения размерности и алгоритмы классификации, чувствительны к подобным эффектам. Поэтому на этапе предварительной обработки данных была проведена нормализация признаков, стандартизация шкал и удаление переменных с низкой информативность. Также был реализован отбор признаков по дисперсии и корреляции, чтобы уменьшить влияние мультиколлинеарности и исключить сильно скоррелированные переменные, которые могли бы исказить работу моделей.

Ещё одним ограничением стало то, что часть гипотез и выбор признаков для анализа основывались не только на статистических критериях, но и на литературных источниках и предварительных знаниях. Это могло повлечь за собой эффект предвзятости — так называемый confirmation bias — при котором подтверждаются уже известные связи, в то время как потенциально

важные, но новые или неожиданные находки могут быть пропущены. Также стоит отметить, что при множественном сравнении (например, при тестировании связи каждого miRNA с клиническими исходами) не во всех случаях проводилась строгая коррекция на множественные проверки, что увеличивает вероятность ложноположительных результатов.

Наконец, важным ограничением является и сама структура данных TCGA. Поскольку эта база содержит в основном ретроспективные данные пациентов из США, преимущественно европеоидной расы, полученные результаты могут не быть универсально применимыми к другим популяциям. Для полноценной валидации моделей и обобщения выводов необходима проверка на внешних, независимых когортах, например, из базы GEO или других международных проектов.

Таким образом, при интерпретации результатов необходимо учитывать, как ограничения самой выборки, так и допущения, связанные с методами обработки и анализа данных. Несмотря на предпринятые меры по очистке и нормализации, итоговые выводы требуют осторожности и дальнейшей проверки.

4 Перспективы развития проекта

Проведённое исследование позволяет выделить несколько направлений для дальнейшего развития проекта, как с точки зрения расширения источников данных, так и в плане усложнения применяемых аналитических подходов. Одним из наиболее перспективных шагов является интеграция дополнительных слоёв молекулярной информации — в частности, данных по метилированию ДНК и копийному числу генов (copy number variation, CNV). Это позволит перейти от одномодального анализа (основанного исключительно на экспрессии miRNA) к более комплексной multi-omics модели. Такой подход способен дать более полное представление о механизмах регуляции генов и их роли в развитии и прогрессировании рака предстательной железы.

Также представляется целесообразным рассмотреть внедрение более сложных моделей машинного обучения, включая глубокие нейронные сети. Нейросетевые архитектуры, такие как автоэнкодеры или рекуррентные сети, могут быть особенно полезны при работе с последовательными биологическими данными и при попытках выявить скрытые паттерны в высокоразмерных омических признаках. Однако применение таких моделей потребует более масштабной предварительной подготовки данных и обеспечения достаточной обучающей выборки, в том числе за счёт расширения используемых биобанков.

Ещё одним возможным направлением является проверка разработанных моделей и аналитических гипотез на других типах онкологических заболеваний. Такой перенос моделей и сравнение закономерностей между различными типами опухолей позволят оценить степень специфичности выявленных биомаркеров и их потенциальную универсальность. Особенно интересным может быть сравнение с опухолями, имеющими схожую природу или механизм гормональной регуляции, такими как рак молочной железы.

Наконец, накопленный материал и полученные результаты могут быть использованы для подготовки научной публикации. Это может быть препринт

на платформе bioRxiv или аналогичной, либо доклад на студенческой или межвузовской конференции, что позволит представить проект научному сообществу, получить обратную связь и улучшить качество работы на основе рецензирования и обсуждений. Подобная научная активность также может стать основой для дальнейшего развития проекта в рамках дипломного исследования или более масштабного грантового направления.

Таким образом, проект обладает значительным потенциалом для дальнейшего углубления и масштабирования, а его развитие может внести вклад в биоинформатику, онкологию и персонализированную медицину.

ЗАКЛЮЧЕНИЕ

Разработанный программный и аналитический продукт в полной мере отвечает поставленным задачам и требованиям со стороны предполагаемого пользователя — исследователя, занимающегося поиском молекулярных маркеров в онкологии. Проведённая интеграция и предварительная обработка данных, реализация визуализаций, а также анализ связи экспрессии miRNA с клиническими показателями позволяют не просто продемонстрировать отдельные функции системы, но и на практике использовать её для поиска биомаркеров и выдвижения биологических гипотез. Система была разработана с учётом специфики биомедицинских данных, и это позволило избежать избыточных или нефункциональных решений. Несмотря на сложность входных данных, в том числе наличие пропусков и шумов, продукт показал устойчивость и надёжность при выполнении анализа.

По результатам тестирования и опытного использования можно сделать вывод, что качество продукта находится на высоком уровне. Выявленные в ходе тестирования дефекты (например, чувствительность к неполным данным или необходимость ручной фильтрации некоторых признаков) не оказывают критического влияния на работоспособность и аналитическую ценность. Напротив, они подчеркнули важность качественной предварительной подготовки данных и позволили уточнить этапы очистки, что также отражено в системе. Все обнаруженные ошибки, связанные с импортом данных или некорректной визуализацией на отдельных этапах, были оперативно устранены. Отдельно стоит отметить стабильность архитектуры при работе с различными подмножествами выборки и адекватность вычислительных затрат при обработке даже высокоразмерных матриц экспрессии.

Тем не менее, существуют направления, в которых продукт может быть улучшен. В первую очередь — расширение спектра данных, включая геномные мутации, метилирование и CNV, что позволит перейти к многослойному мультиомическому анализу. Также перспективным является внедрение более мощных алгоритмов, таких как ансамбли моделей или

нейросетевые подходы, что повысит чувствительность в выявлении закономерностей. С точки зрения пользовательского опыта, возможна разработка более удобного интерфейса для визуального анализа и автоматизированной отчётности. Все эти предложения имеют под собой как технические предпосылки (реализованная модульность и масштабируемость), так и научные, заложенные в текущую структуру проекта.

Таким образом, результаты проекта свидетельствуют о его успехе как с точки зрения научной значимости, так и прикладной реализуемости. Созданный инструмент способен эффективно решать задачи первичного и углублённого анализа биомедицинских данных, служит основой для более сложных моделей и в будущем может быть масштабирован под задачи клинической биоинформатики.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology, 19(1A), A68–A77. https://doi.org/10.5114/wo.2014.47136

2. Weinstein, J. N., Collisson, E. A., Mills, G. B., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45, 1113–1120. https://doi.org/10.1038/ng.2764

3. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15, 550. https://doi.org/10.1186/s13059-014-0550-8

4. Ritchie, M. E., Phipson, B., Wu, D., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47. https://doi.org/10.1093/nar/gkv007

5. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

6. The Cancer Genome Atlas Program (TCGA). National Cancer Institute. https://www.cancer.gov/tcga

7. GEO — Gene Expression Omnibus. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/geo/

8. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

9. McKinney, W. (2010). Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

10. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

11. Seaborn library. https://seaborn.pydata.org/

12. Python Software Foundation. Python Language Reference. https://www.python.org/

13. Bioconductor Project. https://www.bioconductor.org/

ПРИЛОЖЕНИЕ А

Дополнительные материалы по проекту

Графики клинических данных

В рамках предварительного анализа были построены визуализации, иллюстрирующие распределение клинических параметров в исследуемой выборке. На графиках представлены:

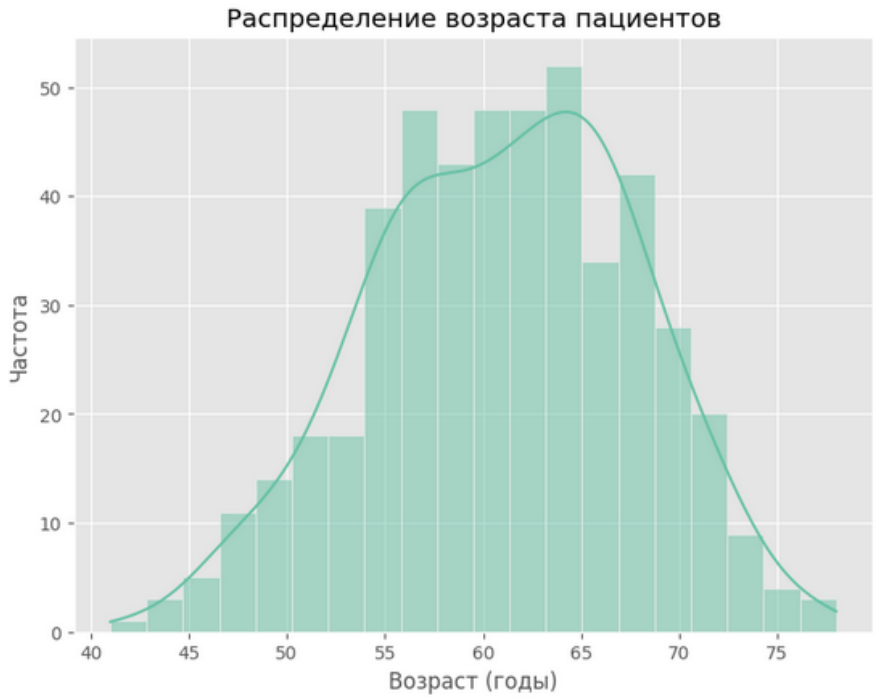

Рисунок 10 – Возраст пациентов

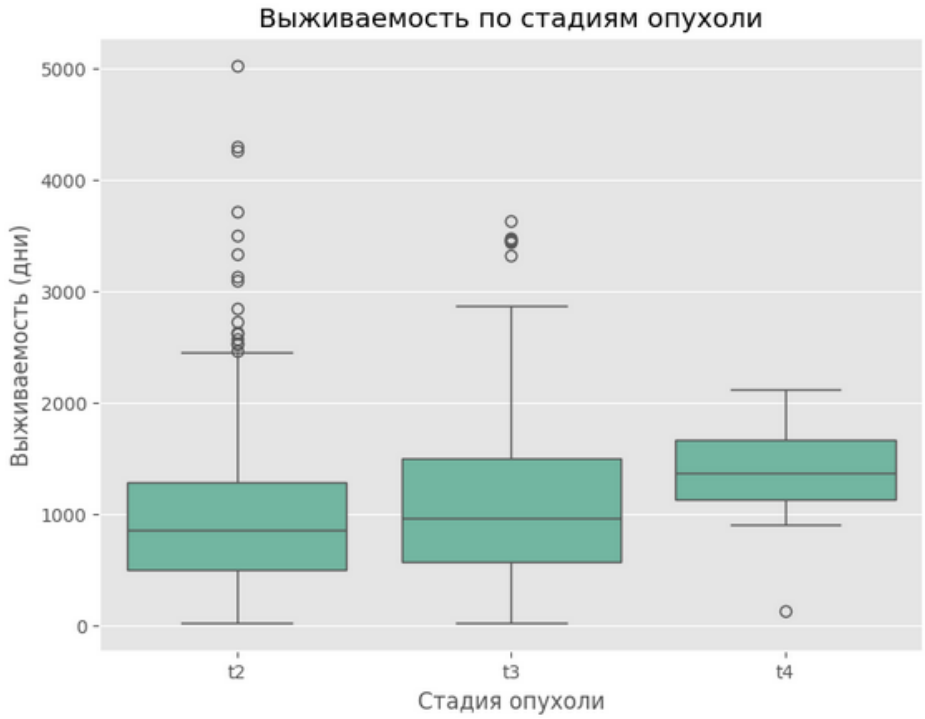

Рисунок 11 – Выживаемость по стадиям опухоли

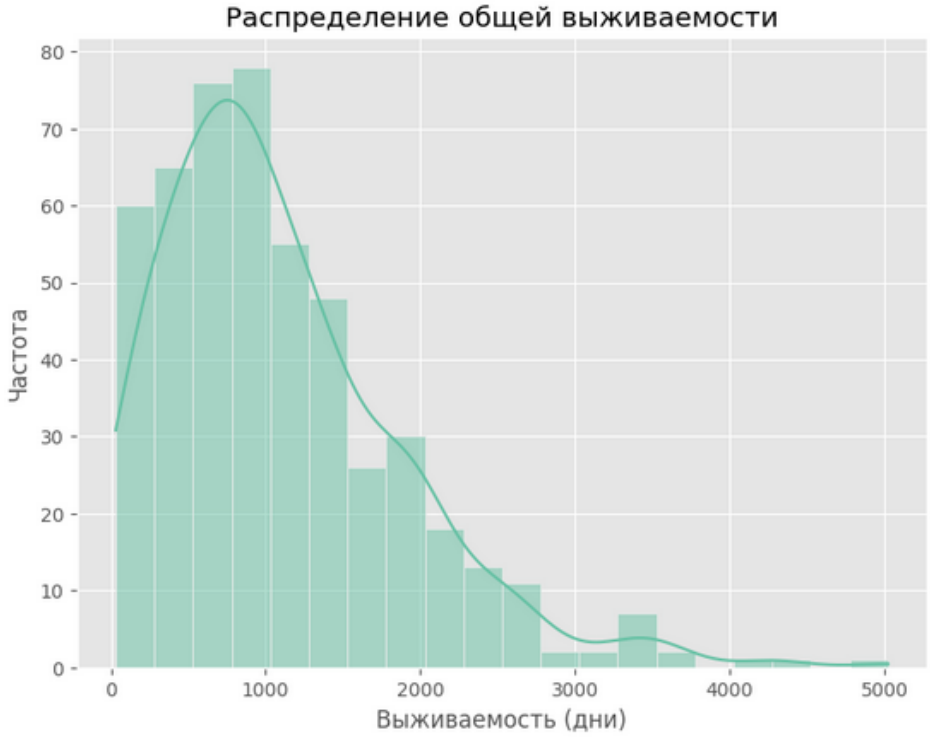

Рисунок 12 – Выживаемость

Рисунок 13 – Стадии лимфатических узлов

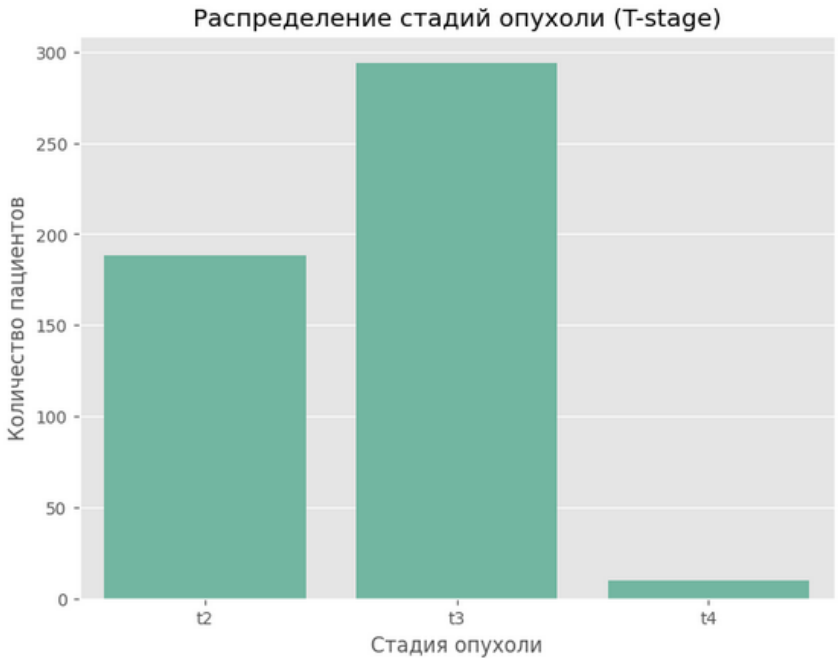

Рисунок 14 – Стадии опухоли

Рисунок 15 – Чистота опухоли

Графики представлены в виде диаграмм плотности. Они служили основой для первичного выявления возможных связей между клиническими параметрами и профилем экспрессии генов и miRNA.

Таблица 4 – Словарь терминов и аббревиатур

Термин / аббревиатура	Расшифровка / пояснение
miRNA	MicroRNA — короткие некодирующие РНК, регулирующие экспрессию генов
RNA - seq	RNA sequencing — технология для анализа уровня экспрессии РНК
PCA	Principal Component Analysis — метод главных компонент для снижения размерности
TCGA	The Cancer Genome Atlas — открытая база данных генетической информации по опухолям
Volcano plot	Визуализация, объединяющая значимость (p-value) и уровень изменений экспрессии (log2FC)
Boxplot	Диаграмма размаха, отображающая медиану, квартили и выбросы
Log-rank test	Статистический тест для сравнения выживаемости между группами
Confusion matrix	Матрица ошибок для оценки качества классификации
ROC-кривая	Receiver Operating Characteristic — график чувствительности и специфичности модели