



Диссертационная работа на соискание степени доктора философии (PhD)

**Разработка и исследование воспроизводимого
вычислительного конвейера на основе методов
машинного обучения для интегративного анализа
scRNA-seq данных кожи при хронической ране**

Соискатель: **Куникеев Айдын Даулетович**

ОП: **8D06102 — Machine Learning & Data Science**

Научный руководитель: PhD, Омаров Б.

Зарубежный консультант: Азизах Сулайман, Ph.D, Малайзия, профессор, Национальный университет Тенара, Куала-Лумпур

Цель и задачи исследования

Разработать и методологически оценить **воспроизводимый вычислительный конвейер интегративного анализа scRNA-seq данных кожи при хронической ране** с учетом групповой структуры данных и предотвращения утечки при формировании признаков.



1

Задача 1.

Систематизировать методы scRNA-seq, RNA-derived variant analysis, машинного обучения и воспроизводимости; определить научный пробел без утверждений, не подтверждённых источниками.



2

Задача 2.

Описать PRJNA736095: 14 SRR, семь GSM-групп, условия wound edge и unwounded skin; зафиксировать реальную единицу независимости и ограничения метаданных.



3

Задача 3.

Разработать последовательность стадий, интерфейсы входов и выходов, конфигурации, manifests, logs и review artifacts.



4

Задача 4.

Сформировать исследовательскую RNA-derived variant branch: GATK calling, PASS filtering, recurrent loci, variant-to-gene mapping и gene burden.



5

Задача 5.

Сопоставить expression и mutation-derived representations на одинаковых 14 наблюдениях; оценить overlap и combined early-concatenation comparator.



6

Задача 6.

Количественно сравнить SRR-level repeated CV, GSM-aware LOGO и training-only recurrent-loci construction при одинаковой основной модели.



7

Задача 7.

Отделить наблюдаемую внутридатасетную разделимость от независимой обобщаемости; не допускать клинических, причинных и biomarker overclaims.

Объект и предмет исследования



Объект:

вычислительный анализ открытых scRNA-seq данных кожи при хронической ране



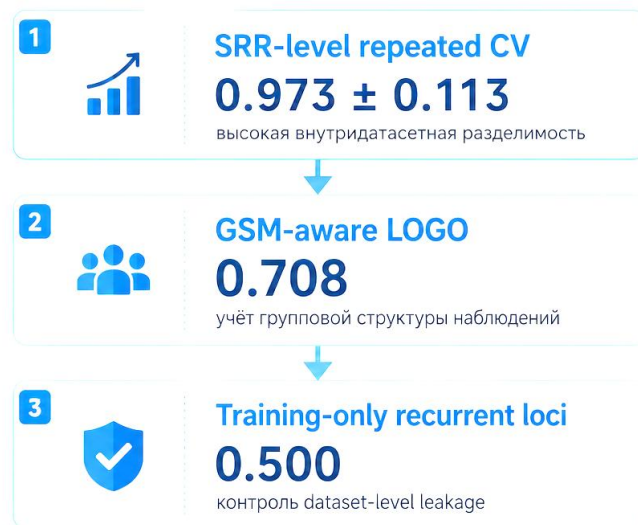
Предмет:

методы формирования признаков, контроля утечки данных, валидации и воспроизводимости ML-результатов

Научная новизна

- Предложен прослеживаемый конвейер «данные → признаки → модель → аудит» для малых иерархических scRNA-seq данных.
- Разделены два уровня утечки: estimator-level leakage и dataset-level leakage.
- Реализовано последовательное сравнение трёх уровней валидации:
SRR-CV
GSM-LOGO
training-only LOGO.

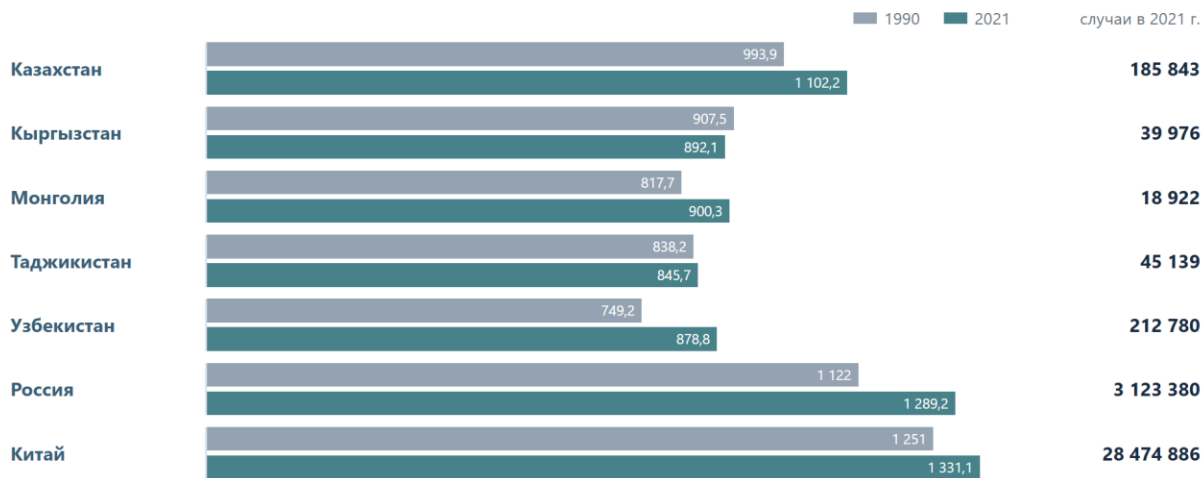
Количественный аудит валидации



 **Усиление контроля валидации:**
0.973 → 0.708 → 0.500

Рост строгости валидации снижает оценку качества и выявляет влияние leakage и групповой структуры данных.

Актуальность темы исследования

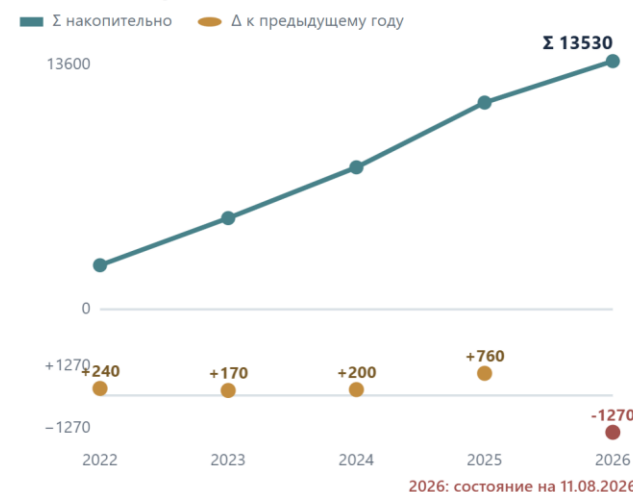


Показатель: возраст-стандартизованная распространённость ЗПА на 100 000 населения. Значения 1990 г. восстановлены из опубликованных значений 2021 г. и процентов изменения GBD; справа — модельная оценка числа распространённых случаев в 2021 г.

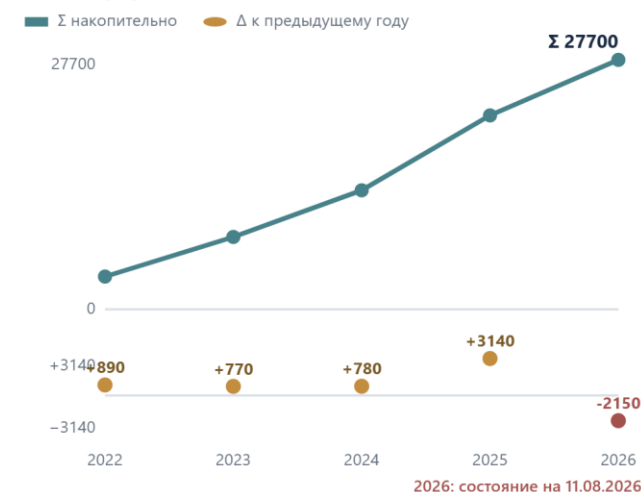
Возраст-стандартизованная распространённость PAD в выбранных странах в 1990 и 2021 годах по данным GBD 2021

Публикационная динамика Google Scholar за 2022-2026 годы: общий запрос по PAD/диабету/хроническим ранам и вычислительный запрос с ML, scRNA-seq или transcriptomics

Общий интерес к заболеванию: 2022–2026



Биоинформатика и ML: 2022–2026



Гипотезы исследования

H1

Объединение expression и RNA-derived признаков может повысить разделимость условий.

$$H1: BA_C > \max(BA_E, BA_M)$$

Проверка: одинаковые samples, labels и GSM-splits

НЕ ПОДТВЕРДИЛАСЬ

Combined-признаки не улучшили grouped Balanced Accuracy

H2

Удержание целых GSM-групп снизит оценку относительно run-level CV.

$$BA_{GSM-LOGO} < BA_{SRR-CV}$$

Проверка: LOGO по 7 GSM-группам

НАПРАВЛЕНИЕ ПОДДЕРЖАНО, НО НЕ ЗНАЧИМО

Balanced Accuracy: 0,973 $\rightarrow$ 0,708; exact $p = 0,257$

H3

Training-only формирование recurrent loci снизит оптимизм глобального feature space.

$$BA_{training-only} < BA_{global-loci}$$

Проверка: nested feature construction

СОГЛАСУЕТСЯ С H3 В АНАЛИЗЕ ЧУВСТВИТЕЛЬНОСТИ

Training-only Balanced Accuracy снизилась до 0,500

Основные научные результаты



1

Разработан воспроизводимый интегративный конвейер анализа scRNA-seq данных.



2

Показано влияние группового пересечения на ML-оценку: **0,973** → **0,708**.



3

Выявлено влияние dataset-level leakage при формировании признаков: **0,708** → **0,500**.



4

Установлено отсутствие подтверждённого group-aware выигрыша от объединения EXP и VAR.

Этапы исследования

1

Этап 1. Анализ предметной области

Задача 1 — обзор scRNA-seq, RNA-derived variant analysis, ML и воспроизводимости. Задача 2 — анализ набора PRJNA736095, структуры 14 SRR / 7 GSM и единицы независимости.



2

Этап 2. Разработка математической модели

Задача 5 — сопоставление expression и mutation-derived representations на одинаковых 14 наблюдениях; оценка overlap и combined early-concatenation comparator.



3

Этап 3. Разработка воспроизводимого вычислительного конвейера

Задача 3 — разработка последовательности стадий, интерфейсов входов/выходов, конфигураций, manifests, logs и review artifacts. Задача 4 — формирование RNA-derived variant branch: GATK calling, PASS filtering, recurrent loci, variant-to-gene mapping и gene burden.



4

Этап 4. Экспериментальное исследование и тестирование

Задача 5 — формирование и сравнение EXP, VAR и COMBINED представлений. Задача 6 — количественное сравнение SRR-level repeated CV, GSM-aware LOGO и training-only recurrent-loci construction при одинаковой основной модели.



5

Этап 5. Валидация, сравнительный анализ и интерпретация результатов

Задача 6 — сравнительная оценка схем валидации. Задача 7 — отделение наблюдаемой внутридатасетной разделимости от независимой обобщаемости; недопущение клинических, причинных и biomarker overclaims.



Этап 1. Анализ предметной области

Аналогичные исследования по направлению PAD / DFU / scRNA-seq / биоинформатика

1. KZ			клиника PAD/DFU + bioinformatics/omics
Автор	Исследование	Год	
Zhakhina et al.	Temporal trends and mortality patterns in PAD in Kazakhstan	2024	
Shalaeva et al.	T2D + PAD у пациентов после partial foot amputation	2024	
Galiyeva et al.	Epidemiology of type 1 and type 2 diabetes mellitus in Kazakhstan	2022	
Sharip et al.	Transcriptome profiling of ESCC patients from Kazakhstan	2024	
Kairov et al.	Whole-genome sequencing and genomic variant analysis of Kazakh individuals	2022	

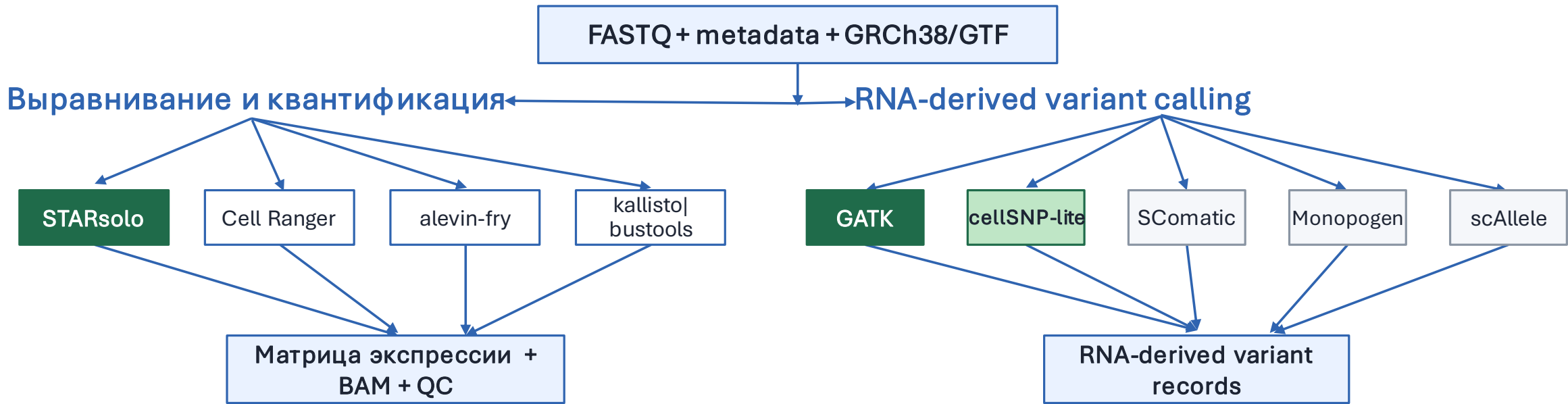
2. Ближнее зарубежье				DFU/PAD и клинические сосудистые осложнения
Автор	Исследование	Год	Страна	
Паршутина et al.	Причины задержки помощи пациентам с синдромом диабетической стопы	2024	Россия	
Оснач et al.	Реваскуляризация и ортопедическая реконструкция стопы при Шарко	2024	Россия	
Aslamov J.K.	Differentiated treatment of diabetic foot syndrome	2023	Узбекистан	

3. Дальнее зарубежье				наиболее близко: DFU/scRNA-seq/bioinformatics/ML
Автор	Исследование	Год	Страна	
Theocharidis et al.	Single-cell transcriptomic landscape of diabetic foot ulcers	2022	США	
Dong et al.	Single-cell RNA-seq in diabetic foot ulcer wound healing	2024	Китай	
Liu et al.	Spatiotemporal single-cell roadmap of human skin wound healing	2025	США/Китай	
Zou et al.	Data mining: association between DFU and PAD	2022	Китай	
Guo et al.	Bioinformatics + machine learning analysis of diabetic foot ulcers	2024	Китай	



Этап 1. Анализ предметной области

Анализ существующих методов обработки scRNA-seq данных



Сравнительная таблица

Уровень	Реализовано в работе	Альтернативы	Критерий выбора
Alignment/ quantification	STARsolo	Cell Ranger, alevin-fry, kallisto bustools	BAM с CB/UB и матрица экспрессии
Variant calling	GATK + cellSNP-lite	SComatic, Monopogen, scAllele	Sample-level и cell-level представления

Обозначения

- основной реализованный метод
- дополнительная реализованная ветвь
- изучено по литературе
- будущий benchmarking



10 STARsolo и GATK выбраны как единая воспроизводимая основа для параллельного формирования экспрессионных и RNA-derived признаков.

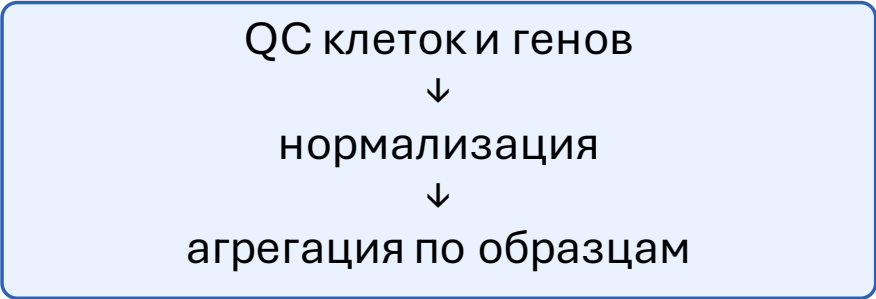
STARsolo [60] Cell Ranger [24]
GATK [43,61] SComatic [71]
Monopogen [72] scAllele [74]

Этап 1. Анализ предметной области

Сравнение способов формирования признаков

Обработанные scRNA-seq данные:
expression matrices + BAM/VCF + metadata

А. Экспрессионные признаки



EXP
18 860 признаков



COMBINED = EXP + VAR
46 090 признаков

В. RNA-derived признаки



Одинаковые GSM groups + splits + preprocessing + metrics

Пространство	Размерность	Содержание	Экспериментальный вывод
EXP	18 860	Экспрессия генов	Сравнительно устойчивее
VAR	27 230	RNA-derived gene burden	Чувствительно к loci formation
COMBINED	46 090	Конкатенация EXP и VAR	Нет подтверждённого group-aware выигрыша



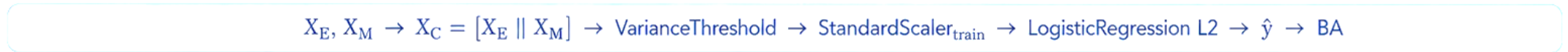
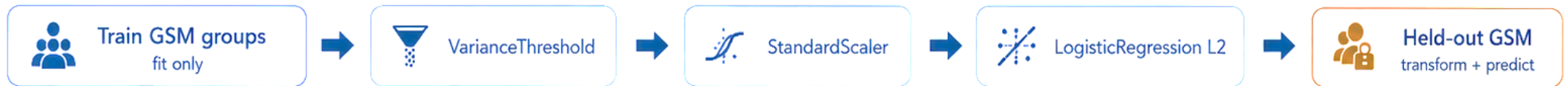
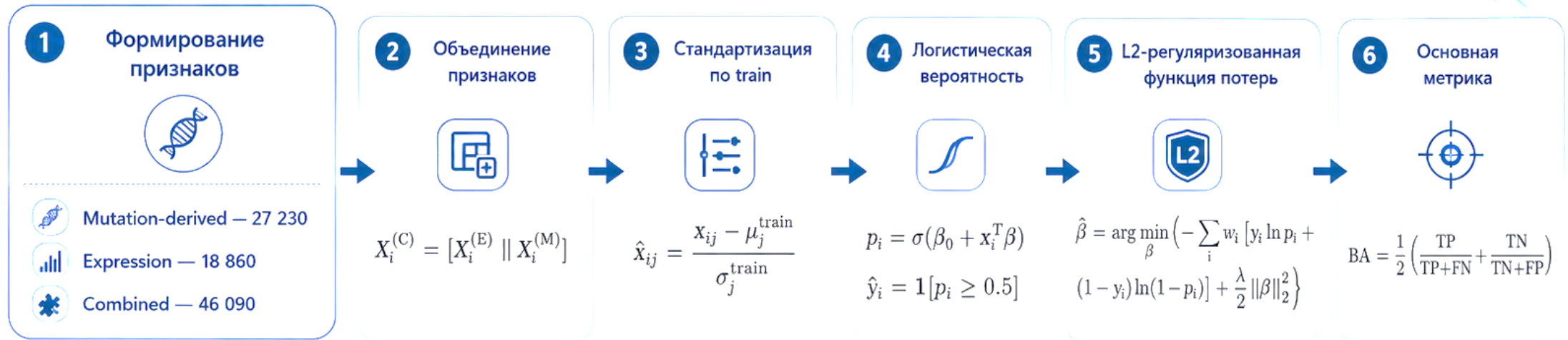
Больше признаков не означает больше обобщаемой информации: простая конкатенация EXP и VAR не улучшила grouped validation.

11 QC клеток и генов [62–64]
агрегация по образцам / pseudobulk [33,67]
GSM-level validation / group-aware CV [29–34,36–38]

нормализация / интеграция / batch correction [14,15,65,66]
RNA-derived variant calling / gene burden [41–44,47,49–52,61,71–74]
EXP / VAR / COMBINED и ML-аудит [46]

Ограничение: RNA-derived variant-признаки — вычислительный исследовательский слой, не подтверждённые соматические мутации или клинические биомаркеры.

Этап 2. Разработка математической модели

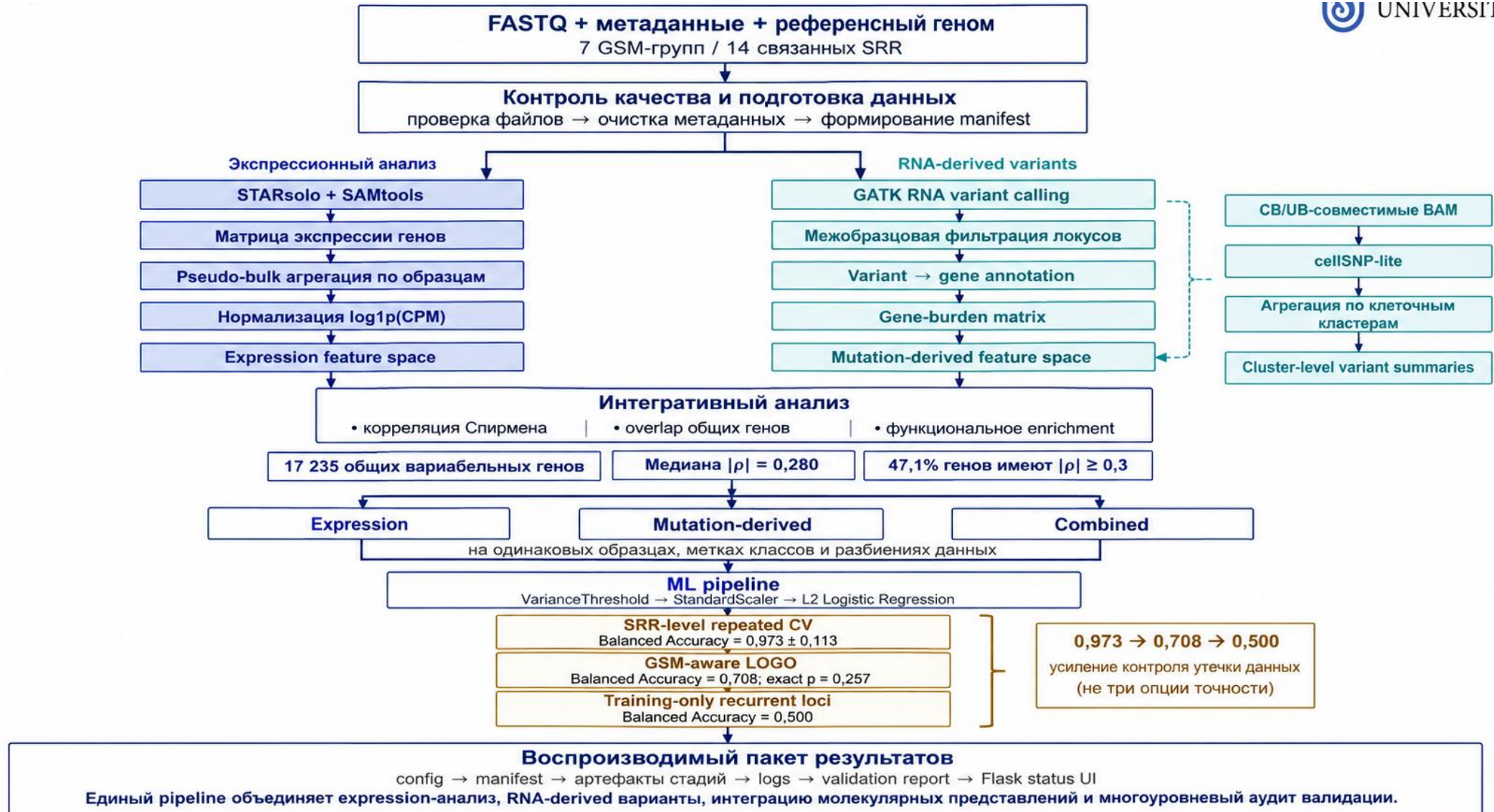


i Объединённое пространство — comparator, а не доказанная multi-omic generalization.



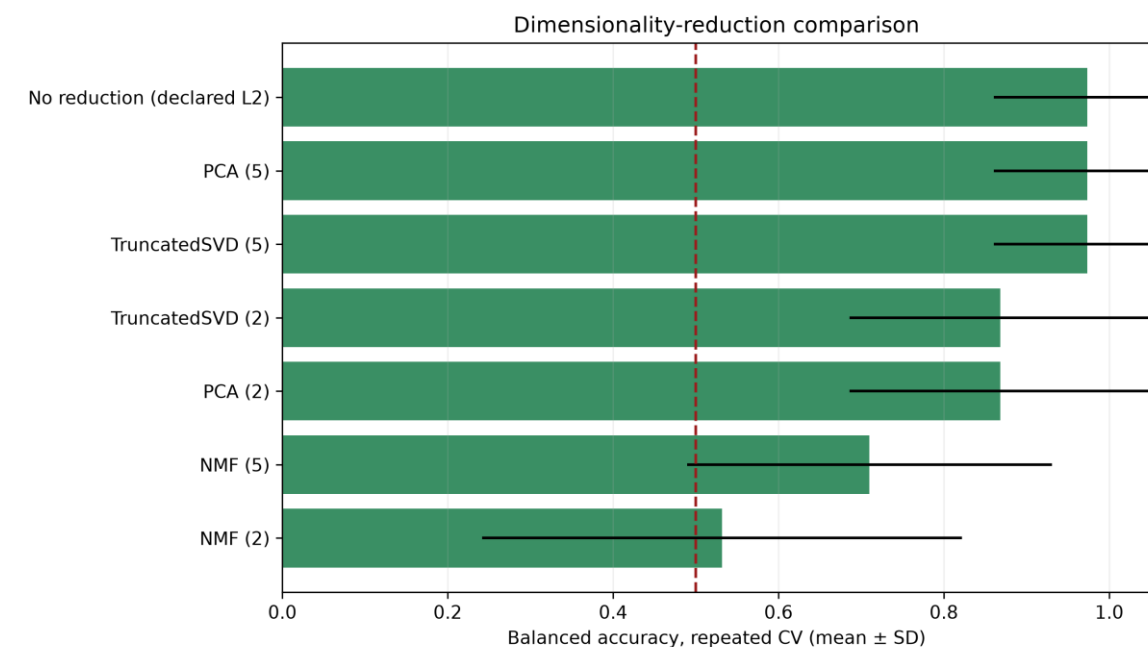
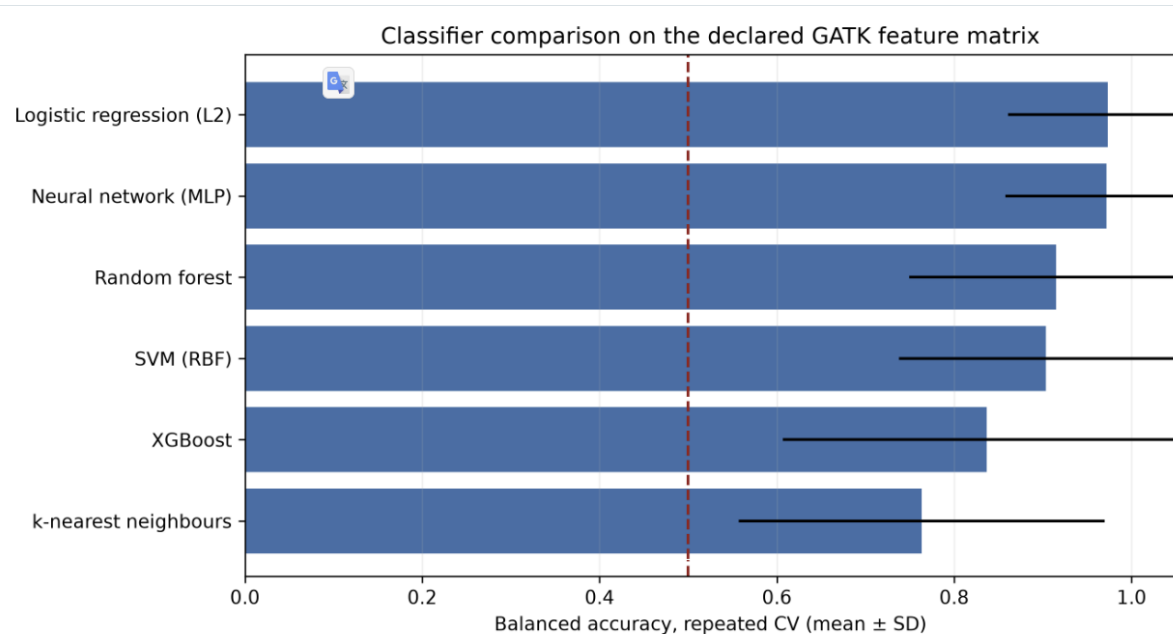
Разработка воспроизводимого вычислительного конвейера

Архитектура вычисления



Экспериментальное исследование и тестирование

Сравнение моделей внутри набора (run-level CV)



Экспериментальное исследование и тестирование

Экспериментальное сравнение протоколов машинного обучения

EXP / VAR / COMBINED + labels + GSM groups

Одинаковые данные, основная L2 Logistic Regression, balanced accuracy и random-state policy

SRR-CV

0,973 ± 0,113

- split: SRR-запуск
- feature formation: global
- средний GSM overlap: 0,884
- интерпретация: внутридатасетная разделимость

GSM-LOGO

0,708

- split: целая GSM-группа
- feature formation: global
- GSM overlap: 0
- интерпретация: внутренняя проверка на удержанной GSM-группе

Training-only LOGO

0,500

- split: целая GSM-группа
- recurrent loci: только train
- GSM overlap: 0
- VAR/COMBINED: пространство признаков строится без test-группы

усиление контроля независимости

0,973

0,708

0,500

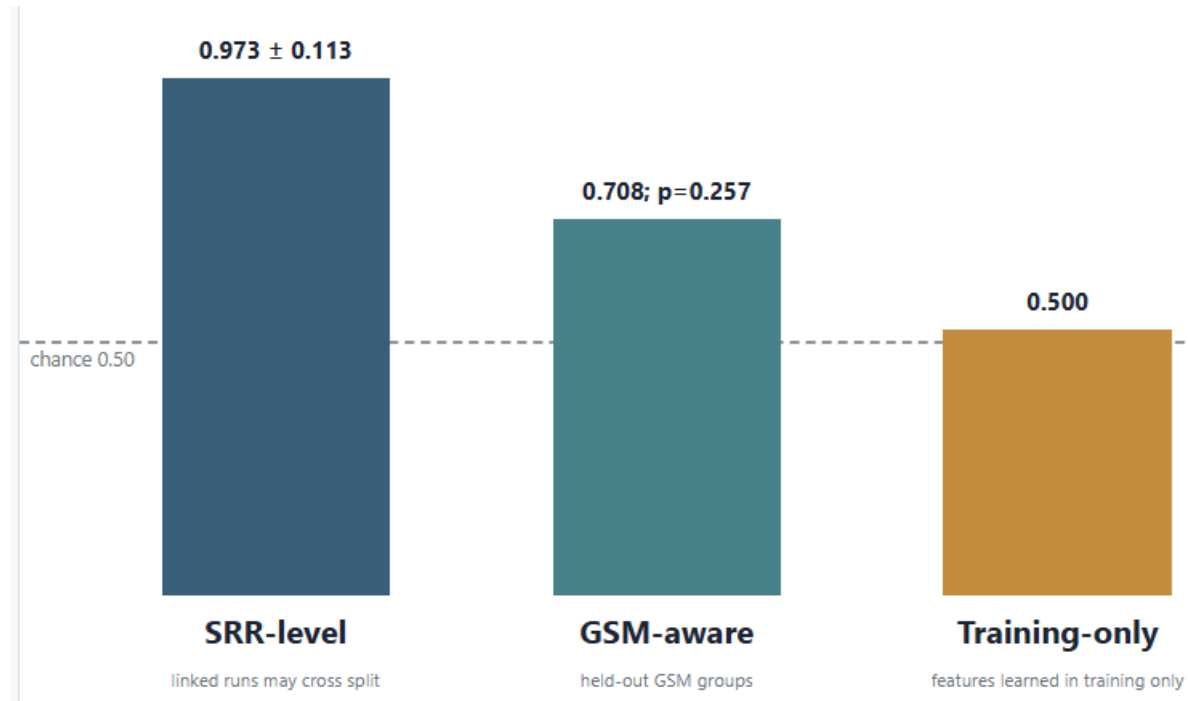
Logistic Regression | SVM | Random Forest | XGBoost | MLP

Смена алгоритма не устраняет малое число независимых групп, псевдорепликацию или upstream leakage.

Достоверность результата определяется единицей независимости и моментом формирования признаков, а не только выбором классификатора.

Экспериментальное исследование и тестирование

Аудит валидации: 0,973 → 0,708 → 0,500



design	Accuracy	BA	ROC-AUC	overlap
Run-level CV	0.964±0.150	0.973±0.113	0.980±0.140	0.884
GSM LOGO runs	0.714	0.708	0.750	0
GSM LOGO groups	0.714	0.708	0.750	0
Exact permutation	—	0.708; p=.257	0.750; p=.286	—

Flask-мониторинг pipeline

12 стадий workflow • статусы • журналы • артефакты
Снимки: idle/not_started — активного запуска не было

Pipeline Status

Last refresh: 2026-08-11 14:28:47 Auto-refresh every 10 seconds

Runner

not_started

Sequential stages continue automatically while the runner is alive.

Watchdog

not_started

No sequence violations detected.

Stage Counts

finished 0 running 0 error 0 idle 12

Total events: 0

Changed means this stage was updated in the current pipeline version.

Stage	Description	Status	Start	End	Duration (s)	Estimate error (s)	Last error
01_input_data	Input validation and metadata cleaning	idle					
02_starsolo	STARsolo alignment (CB/UB BAM) Changed	idle					
03_gatk_call	GATK RNA variant calling (filtered VCF)	idle					
04_cohort_filter	Cohort-common filter	idle					
05_variant_to_gene	Variant to gene annotation Changed Variant-to-gene annotation logic was replaced with a more robust interval-based mapping step.	idle					
06_gene_burden	Gene-burden matrix	idle					
07_ml_control_vs_disease	ML classification + permutation	idle					
08_cellseq	cellseq-lite per-cell allele counts Changed cellseq-lite stage was prepared for real per-cell allele counting instead of placeholder reuse.	idle					
09_cluster_aggregation	Cluster aggregation	idle					
10_mutational_analysis	Mutational analysis summaries Changed Mutational summaries were hardened to support broader VCF filename patterns and richer burden metrics.	idle					
11_correlation	Correlation analysis Changed Correlation was expanded into integrated sample-level expression-mutation analysis with metadata, cluster, and STARsolo-aware summaries.	idle					
12_integrated_interpretation	Collect report bundle Changed Final reporting now includes agentic stage reports, validation artifacts, and final bundle-level interpretation.	idle					

[Back to pipeline status](#)

01_input_data - Input validation and metadata cleaning

Runner: not_started

Watchdog: not_started

Status: idle | Estimate (s): 60 | Started: | Finished:

Last duration (s): | Estimate error (s):

Last error: None

Events

Type	Timestamp	Message
------	-----------	---------

Scanned Files

Path	Size (bytes)	Timestamp	Preview
------	--------------	-----------	---------



Заключение

Основные результаты исследования

1. Анализ современных подходов

Систематизированы методы анализа scRNA-seq, RNA-derived вариантов и машинного обучения. Определён научный пробел: отсутствие единой воспроизводимой pipeline с контролем утечки данных.

2. Структура набора данных

Проанализирован PRJNA736095: 14 запусков SRR объединены в 7 групп GSM, используемых как приближение независимых биологических единиц для сравнения wound edge и unwounded skin.

3. Воспроизводимый конвейер

Разработан двенадцатистадийный вычислительный pipeline, обеспечивающий прослеживаемость входных данных, параметров, логов выполнения, промежуточных артефактов и итоговых отчётов.

4. RNA-derived variants

Реализована обработка RNA-derived вариантов: GATK variant calling, межобразцовая фильтрация, сопоставление вариантов с генами и формирование матрицы геной нагрузки.

5. Интегративный анализ признаков

Сформированы пространства expression, mutation-derived и combined. Выполнены корреляционный анализ, 114 overlap-тестов и функциональное обогащение; FDR-значимых пересечений не выявлено.

6. Аудит машинного обучения

Проведён многоуровневый аудит ML: Balanced Accuracy снизилась с 0,973 (SRR-level validation) до 0,708 (GSM-aware LOGO) и до 0,500 (training-only recurrent loci).

0,973 → 0,708 → 0,500

усиление контроля
утечки данных

7. Границы доказательности

Полученные показатели характеризуют методологические свойства исследуемого pipeline и малой выборки, но не подтверждают клиническую диагностику PAD, причинные связи или готовую биомаркерную панель.

Итог: предложен воспроизводимый интегративный pipeline для анализа scRNA-seq с многоуровневым контролем независимости и утечки данных.



Список научных трудов

№ п/п	Наименование	Характер издания	Выходные данные	Кол стр.	Соавторы
1	2	3	4	5	6
Публикации в международных рецензируемых научных журналах, входящие в БД Scopus/Web of Science					
1	Integrating Mutation-Derived and Expression Features from Single-Cell RNA Sequencing: Pitfalls of Standard Cross-Validation in Small-Cohort Settings (статья)	электронный	International Journal of Molecular Sciences, ISSN 1422-0067, 2026, Том 27, номер 14 Scopus (2025) Computer Science Applications – 86% процентиль	20	Amankeldi A. Salybekov, A.Y., Omarov B
2	Latest Advances in Endothelial Progenitor Cell-Derived Extracellular Vesicles Translation to the Clinic (обзор)	электронный	Frontiers in Cardiovascular Medicine, 2021, Том 8 Scopus Cardiology and Cardiovascular Medicine – 87% процентиль	13	Amankeldi A., Shuzo K, Takayuki A
Статьи в изданиях, рекомендованных Комитетом по обеспечению качества в сфере науки и высшего образования МНВО РК					
3	A Review of Tools, Methodologies, And Techniques For Processing, Pre-Processing, And Clustering Analysis Of Genetic Data (статья)	электронный	Қазақстан- Британ Техникалық Университетінің Хабаршысы, ISSN 2959-8109 (Online), 2024, Т 21, н 4, С 32-45 https://doi.org/10.55452/1998-6688-2024-21-4-45-57	14	Kunikeyev A., Yerimbetova A., Satybaldiyeva R.

По теме диссертации: публикации № 1 и № 3; № 2 — смежная научная работа.

Спасибо за внимание

Вопросы и обсуждение

