УДК 577.218:004.032.26

Определение множественной лекарственной устойчивости и классификация доминирующих в РФ генетических линий *Mycobacterium tuberculosis* с применением методов глубокого обучения на программно-аппаратном комплексе Onsiteseq

Р. А. Горбенко[1]

[1]Московский физико-технический институт (национальный исследовательский университет)

В работе представлена реализация полного цикла (End-to-End) обработки данных секвенирования третьего поколения для клинической диагностики *Mycobacterium tuberculosis* (МБТ) с использованием программно-аппаратного комплекса (ПАК) Onsiteseq. Разработанный ПАК объединяет автоматизированный биоинформатический конвейер и гибридные нейросетевые архитектуры. Комплекс решает две фундаментальные задачи: высокоточную классификацию специфичных для Российской Федерации генетических линий возбудителя (в частности, Lineage 2 / Beijing) на базе свёрточного автоэнкодера с остаточными связями и прогнозирование множественной лекарственной устойчивости (МЛУ-ТБ) с применением гибридной архитектуры собственной разрботки, объеденяющей такие ML-подходы, как 1D-CNN, Self-Attention и графовые сети GCN для моделирования эпистаза и 3D-структуры белков. Разработанный графический интерфейс (GUI) позволяет транслировать сырые данные с нанопорового NGS-секвенатора (Нанопорус, Россия) непосредственно в интерпретируемый клинический отчет в формате, исключая необходимость использования зарубежных веб-серверов и обеспечивая полностью автономную портативную диагностику (Point-of-Care) и технологический суверенитет.

Эпидемиологическая ситуация по туберкулезу в Российской Федерации имеет ярко выраженную специфику: она характеризуется значительным доминированием семейства Beijing (Lineage 2), обладающего высокой вирулентностью. В то же время ключевым препятствием для эффективной терапии становится растущая множественная лекарственная устойчивость (МЛУ-ТБ) [1].

Существующие зарубежные инструменты (например, алгоритмы TB Profiler) требуют ручной выгрузки данных на внешние веб-серверы и недостаточно адаптированы к специфике российских изолятов. Для внедрения технологий геномного надзора непосредственно в клиническую практику необходим интегрированный подход. Целью данной работы являлось создание ПАК Onsiteseq - автономной системы, способной принимать «сырые» данные с портативного нанопорового секвенатора (Нанопорус, Россия) и с помощью методов глубокого обучения выдавать клинически значимые данные.

В основе обработки данных лежит масштабируемый конвейер на базе Snakemake. Пайплайн автоматически осуществляет фильтрацию сырых прочтений (reads), выравнивание на эталонный геном (H37Rv), полировку однонуклеотидных полиморфизмов (SNP) и сборку консенсусной последовательности. Управление жизненным циклом моделей реализовано через строгую методологию BioMLOps на базе системы контроля версий данных DVC, что решает проблему воспроизводимости при пополнении клинических баз и обеспечивает безопасное независимое обновление версий оркестратора и ML-воркеров.

Для определения генетической линии используется анализ паттернов SNP по всему геному. Из-за неравномерного распределения данных в мировых базах(например, The CRyPTIC Consortium [2]) был применен метод искусственной молекулярной эволюции для аугментации обучающей выборки. Архитектура модели представляет собой 1D-CNN автоэнкодер с остаточными связями (ResNet) и транспонированными свёртками. Латентный вектор передается в полносвязный классификатор, обучаемый с комбинированной функцией потерь (реконструкция + классификация)

Прогнозирование фенотипа резистентности к антимикобактериальным препаратам осуществляется на основе локусов мишеней (*rpoB*, *katG*, *pncA*). Базовая архитектура одномерных свёрточных сетей ограничена локальным рецептивным полем. Для преодоления этого ограничения была разработана гибридная мультимодальная архитектура Onsiteseq-Epiformer-3D (Epistatic Transformer with 3D Graph integration), состоящая из трех контуров. Модуль локального экстрактора (1D-CNN) ищет специфичные k-меры; механизм самовнимания (Multi-Head Self-Attention) вычисляет матрицу внимания

для учета дистальных эпистатических взаимодействий (например, компенсаторных мутаций). Наконец, модуль пространственного контекста (Graph Convolutional Network, GCN) строит 3D-граф белка, где узлы - это аминокислоты, а ребра образуются при физическом расстоянии между Cα-атомами менее 5 Å. Это решает проблему выявления резистентности в сложных генах (таких как *pncA*), где мутации распределены линейно, но функционально сгруппированы в активном 3D-центре фермента.

Для взаимодействия с врачом-клиницистом разработано графическое приложение на базе фреймворка PyQt. Ход выполнения анализа транслируется в дружественный лог выполнения, а результатом работы является автоматическая генерация двух отчетов: технического (QC) для биоинформатика и финального медицинского бланка с профилем резистентности и определенным субтипом.

Таблица 1. Ключевые показатели эффективности компонентов ПАК Onsiteseq

Модуль	Архитектура	Метрика	Достигнутый результат
TB-Lineage-Detector (Классификация генетических линий)	1D-CNN Автоэнкодер+ ResNet	Точность (Accuracy) / F1-score для линии Lineage 2 / Beijing	Accuracy: 0.998 / F1-score: 0.997
TB-Res-Detector (Детекция МЛУ-ТБ)	Onsiteseq-Epiformer-3D (Epistatic Transformer with 3D Graph integration)	«ROC-AUC / Чувствительность (Sensitivity)»	Базовые (RIF/INH): ROC-AUC ~0.85 / Чувств. ~64% / Спец. ~90% Резерв (BDQ): ROC-AUC 0.841 / Чувств. 68.7% / Спец. 95.0%
Биоинформатический конвейер (End-to-End подход)	Snakemake	Успешность сборка генома и определение значимых мутаций	Успешно (выравнивание на H37Rv и точная экстракция SNP)

Литература

1. Merker M. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage / M. Merker, C. Blin, S. Mona [et al.] // Nature Genetics. — 2015. — Vol. 47, № 3. — P. 242–249. — URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11044984/ (дата обращения: 21.03.2026).
2. The CRyPTIC Consortium. A data compendium associating the genomes of 12,289 *Mycobacterium tuberculosis* isolates with quantitative resistance phenotypes to 13 antibiotics // PLoS Biology. — 2022. — Vol. 20, № 8. — P. e3001721. — DOI: 10.1371/journal.pbio.3001721.
3. Горбенко Р. А. TB-Lineage-Detector [Электронный ресурс] // GitVerse : [платформа для работы с исходным кодом]. — URL: https://gitverse.ru/onsiteseq/TB-Lineage-Detector (дата обращения: 25.03.2026).
4. Горбенко Р. А. TB-Res-Detector [Электронный ресурс] // GitVerse : [платформа для работы с исходным кодом]. — URL: https://gitverse.ru/onsiteseq/TB-Res-Detector (дата обращения: 25.03.2026).