УДК 577.218:004.032.26

Определение резистентности и классификация доминирующих в РФ субтипов ВИЧ-1 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 ВИЧ-1 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 субтипов вируса (A6, CRF63_02A6) на базе свёрточного автоэнкодера с остаточными связями и прогнозирование лекарственной устойчивости с применением архитектуры 1D-CNN + Self-Attention для моделирования эпистаза. Разработанный графический интерфейс (GUI) позволяет транслировать сырые данные с секвенатора (Нанопорус, Россия) непосредственно в интерпретируемый клинический отчет, исключая необходимость использования зарубежных веб-серверов и обеспечивая технологический суверенитет.

Эпидемия вируса иммунодефицита человека (ВИЧ-1) в Российской Федерации имеет ярко выраженную специфику: она носит преимущественно монофилетический характер со значительным доминированием субтипа A6 (около 83%) и его рекомбинантных форм (CRF63_02A6). В то же время ключевым препятствием для эффективной антиретровирусной терапии (АРТ) становится растущая лекарственная устойчивость (ЛУ) вируса [1].

Существующие зарубежные инструменты (например, алгоритмы Stanford HIVDB [2]) требуют ручной выгрузки данных на внешние веб-серверы и недостаточно адаптированы к специфике российских субтипов вируса. 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 ВИЧ-1 с портативного секвенатора (Нанопорус, Россия) и с помощью методов глубокого обучения выдавать клинически значимые данные.

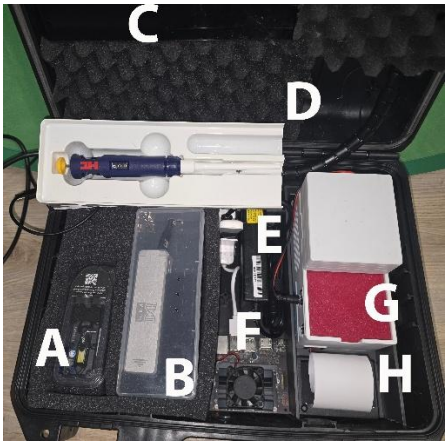

Рис. 1. Прототип ПАК Onsiteseq. Слева направо: (A) Кассета Flongle (B) Нанопоровый сенвенатор “Нанопорус” (C) Сенсорный экран (D) Шприц-дозатор (E) Коммуникационный блок Wi-Fi/4G (F) CPU/GPU процессор на базе платформы Nvidia Jetson (G) Амплификатор (H) Термопринтер

В основе обработки данных лежит масштабируемый конвейер на базе Snakemake. Пайплайн автоматически осуществляет фильтрацию сырых прочтений (reads), выравнивание на эталонный геном (HXB2), полировку однонуклеотидных полиморфизмов (SNP) и сборку консенсусной последовательности. Для экстракции генов-мишеней (PR, RT, IN) реализован алгоритм локального выравнивания Смита-Ватермана, устойчивый к шуму нанопорового секвенирования и встроенным стоп-кодонам. Для определения субтипа используется ген оболочки *env*. Из-за неравномерного

распределения данных в мировых базах (например, преобладание субтипа B) был применен метод искусственной молекулярной эволюции (синонимичные/несинонимичные мутации, индели, рекомбинация) 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.

Прогнозирование фенотипа резистентности к АРВТ осуществляется на основе последовательностей гена pol. Базовая архитектура одномерных свёрточных сетей ограничена локальным рецептивным полем. Для преодоления этого архитектура была дополнена механизмом самовнимания (Self-Attention) . Это позволило модели динамически вычислять матрицу внимания и учитывать дистальные эпистатические взаимодействия (например, компенсаторное влияние мутации L74I на дефекты репликации от G118R в гене интегразы)

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. Интегрированный комплекс был валидирован на независимых данных из российской базы RuHIV [3] и "сырых" данных секвенирования взятых из исследования PRJDB17699 [4]. Полученные данные проверялись на серверах Stanford HIVDB [2] и Comet HIV-1 [5]

Детекция эпистаза**:** In silico тестирование доказало, что разработанная гибридная модель (CNN+Attention) более чем в 2 раза эффективнее классической CNN улавливает синергию мутаций L74I и G118R (Δ уверенности +0.102 против +0.047) для доминирующего в РФ субтипа A6. Общие результаты представлены в Таблице 1.

Таблица 1. Ключевые показатели эффективности компонентов ПАК Onsiteseq

Модуль	Архитектура	Метрика	Достигнутый результат
HIV-1-M-Env-Rus (Классификация субтипов)	1D-CNN Автоэнкодер+ ResNet	Точность (Accuracy) / F1-score для субтипа A6	99.64% / 1.00 (Абсолютная точность для штаммов РФ)
HIV-1-Resist-Rus (Детекция ЛУ-ВИЧ)	CNN + Self-Attention(Сравнение с базовой CNN)	Прирост уверенности к резистентности (Δ Prob)	0.102 (в 2.1 раза чувствительнее базовой CNN)
Биоинформатический конвейер (End-to-End подход)	Snakemake + алгоритм Смита-Ватермана	Успешность сборка генома и определение значимых мутаций	Успешно (точная экстракция генов RT, PR и IN)

Литература

1. Кириченко А.А., Киреев Д.Е., Сидорина Ю.Н., Абашина Н.Д., Брусенцева Е.Е., Акимкин В.Г. Пилотное исследование по изучению особенностей распространения резистентных вариантов ВИЧ-1 с помощью молекулярных кластеров // Журнал микробиологии, эпидемиологии и иммунобиологии. 2024. Т. 101. № 5. С. 581–593. doi:10.36233/0372-9311-56
2. Tang, M. W. The HIVdb System for HIV-1 Genotypic Resistance Interpretation / M. W. Tang, T. F. Liu, R. W. Shafer // Intervirology. 2012. Vol. 55, № 1. P. 47–53. DOI: 10.1159/000331998.
3. Киреев Д.Е., Кириченко А.А., Лопатухин А.Э., Шлыкова А.В., Галкин Н.Ю., Савельев Е.В., Глазов М.Б., Покровский В.В., Акимкин В.Г. Российская база данных устойчивости ВИЧ к антиретровирусным препаратам. Журнал микробиологии, эпидемиологии и иммунобиологии. 2023;100(2):219–227. doi: https://doi.org/10.36233/0372-9311-345
4. PRJDB17699 [Электронный ресурс] NCBI BioProject. – 2024. – URL: https://www.ncbi.nlm.nih.gov/bioproject/PRJDB17699 (дата обращения: 22.02.2026)
5. Daniel Struck; Glenn Lawyer; Anne-Marie Ternes; Jean-Claude Schmit; Danielle Perez Bercoff. COMET: adaptive context-based modeling for ultrafast HIV-1 subtype identification. Nucleic Acids Research 2014; doi: 10.1093/nar/gku739