Машинное обучение для селекции информативных локусов в штаммоспецифичной филогенетике Neisseria gonorrhoeae

Кочубей Елизавета
Передовая инженерная школа
Университет ИТМО
Санкт-Петербург, Россия
eliz.kochubey@gmail.com

Добровольский Серафим
Передовая инженерная школа
Университет ИТМО
Санкт-Петербург, Россия
serafimin2002@gmail.com

Зенченко Злата
Передовая инженерная школа
Университет ИТМО
Санкт-Петербург, Россия
zlata.zenchenko@mail.ru

17 сентября 2025 г.

Abstract

Эпидемиологический надзор за Neisseria gonorrhoeae осложняется ограничениями существующих методов молекулярного типирования, таких как NG-MAST и MLST, которые либо обладают избыточной вариабельностью, либо недостаточной разрешающей способностью. В настоящем исследовании предложен и апробирован алгоритм на основе машинного обучения (ML) для автоматизированного отбора минимального набора информативных генетических локусов для точной штаммовой классификации. На основе коллекции из 29 референсных геномов N. gonorrhoeae был разработан пайплайн, использующий модели Random Forest и DNABERT-эмбеддинги для формирования оптимальных генных панелей. Результаты показали, что ML-отобранные панели значительно превосходят традиционные схемы: точность филогенетической реконструкции повышается почти вдвое (нормированное расстояние Робинсона-Фоулдса 0.65 против 0.81 у pubMLST), а надежность ветвления превышает 90%. Разработанный подход позволяет существенно снизить вычислительные затраты по сравнению с полногеномным анализом и представляет собой перспективный ресурсо-эффективный инструмент для рутинного эпидемиологического мониторинга, отслеживания путей распространения и выявления штаммов с антибиотикорезистентностью.

1 Введение

Метагеномные классификаторы, подробно рассмотренные Marić и соавт. (2024) [1], демонстрируют удовлетворительную точность видовой идентификации, однако при переходе к штаммовому уровню сталкиваются с рядом принципиальных ограничений. У k-mer-ориентированных решений Kraken2 [2] и Centrifuge [3] избыточная полиморфность коротких олигонуклеотидов приводит к ложноположительным классификациям. Mapping-подходы Minimap2 [4] и MetaMaps [5] зависят от полноты справочника и требуют значительных вычислительных ресурсов, а универсального бенчмарка для длинночитаемых данных пока не существует.

С 2023 по 2025 гг. появились инструменты, частично нивелирующие указанные недостатки. Demixer [6] сочетает байесовскую оценку долей известных клонов с реконструкцией de novo SNP-профилей, устраняя «reference-bias», но надежно работает лишь при покрытии $\geq 20\times$ и чувствителен к рекомбинации. MADRe [foster2024] уменьшает число ложноположительных детекций за счёт предварительной сборки контигов и автоматического сокращения базы, однако удлиняет время анализа и требует хорошего покрытия. Clasnip [7] предлагает HMM-типирование по коротким ампликонам через web-интерфейс, повышая доступность скрининга, но полностью зависит от актуальности базы и не поддерживает смешанные образцы.

Наряду с публикацией minMLST [8], где методом XGBoost отбирается минимальный набор локусов из cgMLST-схем [9] при сохранении приемлемого разрешения (скорректированный индекс Рэнда 0,4–0,93), активно развиваются и marker-ориентированные метагеномные методы, например StrainPhlAn [10]. Тем не менее ни один из существующих инструментов не объединяет высокое штаммовое разрешение, умеренные вычислительные требования и полноценную интеграцию в рутинную диагностику.

1.1 Специфика Neisseria gonorrhoeae

Neisseria gonorrhoeae отличается минимальным межгеномным расстоянием между циркулирующими клонами ($\leq 0,001$ замен/сайт) и высоким уровнем рекомбинации ($r/m \approx 2,5$) [11, 12, 13]. Эти особенности ограничивают возможности:

- MLST [14] — схема из семи housekeeping-генов даёт согласованную, но недостаточно детализированную картину;
- NG-MAST [15] — двухгенная система (porB, tbpB) повышает разрешение, но формирует политомии, не отражающие фенотип устойчивости и географию;
- cgMLST — задействует 100–2000 core-локусов, повышая точность, однако остаётся ресурсоёмким и игнорирует вариабельность аксессорного генома;
- Полногеномный анализ [16] — обеспечивает максимальное филогенетическое разрешение, но требует значительных вычислительных ресурсов, высокой стоимости секвенирования и экспертной интерпретации данных.

Marić и соавт. показали, что отсутствие одного вида в справочнике влечёт $\geq 20\%$ ошибку оценки представленности; на штаммовом уровне это эквивалентно пропуску клинически значимого клона. Кроме того, k-mer-классификаторы склонны генерировать искусственные «штаммы», тогда как ресурсоёмкие мапперы трудно применять в лабораториях с ограниченными вычислительными ресурсами.

1.2 Обоснование исследования

Для эпидемиологического надзора за N. gonorrhoeae необходим специализированный, ресурсно-эффективный пайплайн с курированной базой гонококковых геномов, целевым подмножеством ортологов и интегрированным анализом генов устойчивости.

Наш проект предлагает алгоритм автоматического отбора минимального набора информативных ортогрупп (не ограничиваясь core-генами), последующую оценку топологической стабильности филогенетических деревьев и одновременное определение детерминант антибиотикорезистентности. Такой подход позволяет сохранить штаммовое разрешение, снизить вычислительные затраты и обеспечить практическую применимость метода.

Важность и актуальность нашего исследования обусловлены сохраняющимся высоким уровнем распространения гонококковой инфекции, вызванной N. gonorrhoeae, что по-прежнему представляет собой серьезную проблему общественного здравоохранения. Особенную тревогу вызывает быстрое развитие антибиотикорезистентности у гонококка, требующее постоянного и эффективного эпидемиологического надзора.

Предлагаемый нами подход имеет значительное практическое применение. Разработанный алгоритм способен оптимизировать ПЦР-диагностику, позволяя создавать более точные и экономичные ПЦР-системы, направленные на выявление ключевых эпидемиологических маркеров, включая гены резистентности. Важно отметить, что горизонтальный перенос генов (ГПГ) играет ключевую роль в быстром распространении этих детерминант антибиотикорезистентности среди популяций N. gonorrhoeae, что делает их отслеживание критически важным для контроля. Надеемся, наш пайплайн позволит оперативно выявлять новые штаммы, отслеживать пути их распространения и прогнозировать развитие резистентности, существенно повышая эффективность контроля над заболеванием.

2 Материалы и методы

2.1 Характеристика и получение геномных данных

В качестве исходных данных использовалась коллекция из 29 референсных штаммов N. gonorrhoeae, утверждённая ВОЗ в 2024 г [17].

Коллекция охватывает основные фенотипы и генотипы резистентности, включая клон FC428 (penA-60.001) и мозаичные варианты системы MtrRCDE. Все геномы получены с применением PacBio/Nanopore и последующей коррекцией Illumina, что гарантирует высокое качество сборки. Преимущества данной выборки:

- Контролируемое разнообразие. Штаммы отобраны для репрезентации глобальных клонов, в том числе экстремально резистентных.
- Полнота аннотации. Доступны точные фенотипы чувствительности и геномные маркеры (blaTEM, tetM, 23S rRNA A2059G/C2611T, gyrA S91F/D95G и др.).
- Репрезентативность без избыточности. В отличие от открытых репозиториев, коллекция исключает дублирование и технические артефакты.
- Стандартизация. Штаммы используются в программах GASP и EGASP, обеспечивая сопоставимость результатов.

Выборка включает 15 новых (2024 г.) и 14 исторических референсных штаммов, что гарантирует временную и географическую репрезентативность. Метаданные содержат информацию о месте и времени выделения, профили MLST, NG-MAST и NG-STAR, облегчая сопоставление с традиционными методами типирования.

Полные и хромосомные сборки N. gonorrhoeae извлекали из RefSeq [18] утилитой ncbi-genome-download с уровнем «complete/chromosome» и форматами GenBank + CDS-FASTA. Из общего списка отфильтровали 29 референсных изолятов серии WHO-2024; дубли исторических штаммов (идентичные клоны, помеченные «_2024») удаляли регулярным фильтром. Для каждой сборки из файла assembly_summary автоматически извлекали идентификаторы BioSample и названия штаммов; эти метаданные добавляли в заголовки CDS-последовательностей собственным скриптом на основе Biopython.

2.2 Определение ортологов и выравнивание

В скачанных CDS данных определялись последовательности однокопийных ортогрупп с использованием OrthoFinder v2.5.5 [19] (режим –dna); полученные последовательности далее использовались для филогенетического анализа. Внутри каждой отдельной ортогруппы проводилось выравнивание с использованием програмного пакета MAFFT (v7.525) [20] (режим –auto) и рассчитывалась средняя энтропия Шеннона, а также другие метрики характеризующие вариабельность ортогруппы.

2.3 Построение эталонной и вариативных филогенетических реконструкций

Эталонное дерево. Для оценки качества разработанной системы типирования мы построили эталонное филогенетическое дерево на основании всех однократных ортогрупп (1776 генов). Нуклеотидные последовательности этих групп были конкатенированы для каждого штамма, после чего дерево реконструировали в IQ-TREE2 методом максимального правдоподобия с автовыбором модели эволюции (MFP), параметрами HKY+F+G4 и 1000 итерациями Ultrafast-bootstrap. [21].

2.4 Эвристический подбор генов

Пошаговая проверка вариабельных локусов. Из ранжированного списка последовательно формировали наборы из 1…N наиболее вариабельных генов, конкатенировали их и для каждого набора строили дерево тем же протоколом; стабильность топологии оценивали по Robinson–Foulds-дистанции (RF) относительно эталона, а также независимой метрике - среднего bootstrap для дерева.

Метрика Robinson–Foulds (RF). RF-дистанция измеряет симметричное различие между двумя нефрактурированными (unrooted) деревьями: это число бифуркационных разрезов («splits»), присутствующих только в одном из сравниваемых деревьев. Для двух идентичных топологий $RF = 0$; если же клады не совпадают вовсе, RF достигает максимума, рассчитывается по формуле 1.

$$RF_{max} = 2(n - 3), \tag{1}$$

где n — это общее число листьев (в данном случае, штаммов) в дереве.

При $n = 29$ штаммах максимальное значение RF будет:

$$RF_{\max} = 2(29 - 3) = 52$$

В работе дополнительно используется нормированное расстояние, которое позволяет сравнивать результаты независимо от размера дерева:

$$RF_{\text{norm}} = \frac{RF}{RF_{\max}} \in [0; 1]$$

Значения RF_{norm} интерпретируются следующим образом:

- $RF_{\text{norm}} \leq 0.20$ рассматривается как «хорошее» совпадение топологий.
- $RF_{\text{norm}} \leq 0.10$ — как практически неотличимое от эталона совпадение.

Контрольные наборы «средней» вариабельности.

Помимо подпоследовательностей из N наиболее вариабельных локусов, для проверки потенциального смещения в сторону гипервариабельных генов формировали эквивалентные наборы из N ортогрупп, расположенных в окрестности медианного значения энтропии (пять локусов выше и четыре ниже) [22]. Процедура конкатенации, реконструкции деревьев (IQ-TREE, 1000 UFBoot) и расчёта нормированного RF-расхождения выполнялась идентично максимально вариабельным наборам, а среднее bootstrap-поддержки использовалось как дополнительный критерий стабильности. Эти контрольные результаты приведены в разделе «Результаты» параллельно с метриками для топ-вариабельных и ML-отобранных панелей.

2.5 ML-подход в подборе генов (подготовка данных)

Генерация случайных комбинаций локусов для тренировочной выборки

Из 200 наиболее вариабельных ортогрупп случайно формировали комбинации по четыре гена (три независимых повторения для каждого гена), пар и троек. Для каждой комбинации создавали конкатенированное выравнивание и генерировали дерево по IQ-TREE. Целевая переменная вычислялась, как нормированное RF-расхождение с эталонной филогенией рассчитывали при помощи ETE3 и независимая метрика оценки надежности дерева - средний бутстреп дерева.

Инженерия признаков для ортогрупп.

Для каждой ортогруппы вычислялось восемь количественных характеристик:

- $Mean\_Edit\_Distance$. Среднее число операций редактирования (замена, вставка, удаление), необходимых, чтобы превратить одну аллельную последовательность в другую. Чем выше значение, тем больше различий между аллелями.
- $TopK\_Coverage$. Доля штаммов (в процентах), охваченных K наиболее частыми аллелями. Например, если 3 самых распространённых аллели встречаются в 23% случаев – то $TopK\_Coverage = 0.23$
- $Mean\_Normalized\_Levenshtein$ Среднее значение расстояния Левенштейна между парами последовательностей, нормированное на их длину (от 0 до 1). 0 – идентичные, 1 – полностью разные.
- $SNP\_Density$ Процент полиморфных позиций (однонуклеотидных замен) в последовательности по отношению к её общей длине.
- $GapFraction$ Доля позиций в множественном выравнивании, представленных пропущенными нуклеотидами (дефисами) из-за вставок/удалений, относительно всей длины выравнивания.
- $LengthVar$ Мера изменчивости длины аллелей в локусе (дисперсия). Более высокое значение указывает на большую разбросанность по длине.
- $UniqueAlleles$ Общее число различных аллельных последовательностей, обнаруженных в выборке для данного локуса. Большее число – больше аллельного богатства.

Эти метрики по ортогруппам далее будут использованы для обучения модели с целью минимизации целевой переменной RF.

Эмбеддинг подход - использование DNABERT.

Для эмбеддинг-подхода из 200 наиболее вариабельных ортогрупп формируют все возможные пары и для каждой пары конкатенируют множественное выравнивание. Последовательность, если она длиннее

512 нуклеотидов, разбивается на токены по 512 знаков(требование к работе предобученной модели DNABERT). Каждый фрагмент пропускают через модель DNABERT [23], после чего применяют усреднение по всем k-мерам и по всем фрагментам, получая единый 768-мерный вектор (embedding) для каждой пары ортогрупп.

2.6 ML-подход в подборе генов (обучение моделей)

Обработка комбинаций

На этапе обучения модели использовались данные, полученные из случайных комбинаций ортогрупп. Для каждой комбинации из двух, трёх или четырёх генов (с тремя независимыми повторениями) формировались конкатенированные выравнивания, по которым строились деревья в IQ-TREE. В качестве целевой переменной использовалось нормированное расстояние Робинсона–Фоулза (RF) между деревом комбинации и эталонной филогенией, рассчитанное с помощью ETE3. Дополнительно учитывалась метрика устойчивости дерева — средний bootstrap.

Для каждой ортогруппы заранее вычислялись восемь числовых признаков, отражающих её структурно-эволюционные характеристики (расстояния, разнообразие, плотность SNP, фракция пропусков и др.). На основе этих признаков для каждого гена в составе комбинации создавался объединённый вектор, описывающий комбинацию. Для предсказания точности дерева по комбинациям генов была обучена модель Random Forest с использованием заранее вычисленных признаков по ортогруппам. Данные разделялись на тренировочную и тестовую выборки (например, 80/20), а внутри обучения применялась 5-кратная кросс-валидация. Каждый пример представлялся вектором признаков, агрегированных по входящим в комбинацию локусам. Целевая переменная — нормированное RF-расхождение с эталонным деревом.

На основе важности признаков и предсказаний RF-модели был составлен топ отдельных генов, чьи характеристики статистически ассоциировались с низкими RF-значениями (т.е. высоким соответствием эталонной филогении).

В параллельном подходе аналогичная модель обучалась на всех возможных парах ортогрупп из тех же 200. Для каждой пары рассчитывались объединённые признаки, и модель обучалась предсказывать RF-расхождение. Далее отбирались пары с минимальными предсказанными RF, и по этим парам строились деревья, чтобы проверить точность такого ранжирования. Таким образом, ML-подход позволил выявить как одиночные локусы, так и пары, обладающие наибольшим потенциалом для точного восстановления филогении.

Обработка эмбеддингов на парах ортогрупп

Для каждой из C(200,2) пар ортогрупп получили embedding через DNABERT Вместо обучения модели мы вычислили между всеми embedding-векторами Евклидовы расстояния и отобрали пары с максимальным удалением (т.е. «неперекрывающиеся» по признакам). Эти пары затем использовали для конкатенации выравниваний, построения деревьев в IQ-TREE и расчёта нормированного RF с эталоном. Идея в том, что наибольшая дистанция в embedding-пространстве отражает комплементарную информацию локусов и ведёт к более низким RF-значениям без дополнительной ML-предсказания.

2.7 Проверка ML-панели и сопоставление с классическими схемами MLST

Отобранные локусы по вариабельности конкатенировали по схеме «1…10 генов»; для каждого объёма строили дерево (IQ-TREE, 1000 UFBoot) и рассчитывали RF-дистанцию к эталону. Стабильность оценивали также средним bootstrap-значением по ветвям.

NG − MAST. Для каждого штамма извлекали CDS-фрагменты porB и tbpB, конкатенировали, выравнивали (MAFFT) и строили двухгенное дерево максимального правдоподобия. Топологию сравнивали с эталонной филогенией по метрике Robinson–Foulds.

pubMLST (семь housekeeping-генов). По аналогичной процедуре обрабатывали схему abcZ, adk, aroE, fumC, gdh, pdhC, pgm: для каждого гена извлекали внутренний фрагмент, определённый оригинальной MLST-схемой, выполняли множественные выравнивания, конкатенацию семи аллелей и реконструкцию дерева с теми же параметрами максимального правдоподобия и bootstrap-поддержки. Полученную топологию сопоставляли с эталоном, используя ту же дистанционную метрику.

Таким образом, обе исторически используемые схемы (двухгенная NG-MAST и семигенная pubMLST) были оценены в параллеле с варьируемыми наборами ортологов и ML-отобранной панелью, что обеспечило единый сравнительный контекст для всех методов типирования.

Полная воспроизводимая реализация всех описанных шагов — от загрузки геномов до расчёта филогенетических метрик и машинного отбора локусов — размещена в открытом репозитории GitHub [24], включающем все скрипты, настройки окружения и примеры команд для запуска конвейера.

2.8 Анализ генов антибиотикорезистентности (ARG)

Для оценки клинической значимости филогенетических кластеров параллельно проводили скрининг генов устойчивости с помощью платформы CARD. Полногеномные сборки каждого штамма (всего их 29) обрабатывали модулем Resistance Gene Identifier (RGI) в режимах «perfect» + «strict», что позволяет идентифицировать гены устойчивости к антибиотикам (ARGs) путем сравнения последовательностей с тщательно составленной базой данных CARD.

Выходные текстовые таблицы RGI для каждого штамма агрегировали и обрабатывали с использованием библиотек pandas, seaborn и matplotlib.pyplot в среде Python (версия 3.9.12). В частности, названия штаммов извлекались из имен файлов, а записи об антибиотиках очищались и нормализовались по словарю CARD, включая обработку множественных антибиотиков в одной записи. Механизмы (("β-лактамаза"), «эффлюкс», «изменение мишени» и т. д.) классифицировались согласно онтологии базы CARD. Для анализа в Colab были взяты текстовые файлы.

3 Результаты

3.1 ML-подход

Средняя абсолютная ошибка обучения с кроссвалидацией составила MAE = 0.042 RF-ед. Это свидетельствует о том, что обученный случайный лес надёжно предсказывает топологическое расхождение (Robinson–Foulds, RF) между деревьями, построенными по произвольным комбинациям локусов, и эталонной филогенией. Для количественной оценки вклада алгоритмической селекции мы сопоставили показатели филогенетических деревьев, построенных из «топ-вариабельных» генов, с аналогичными метриками панелей, сформированных моделью машинного обучения. Результаты сведены в табл. 1.

Таблица 1: Сравнение «топ-вариабельных» и ML-предсказанных наборов.

Кол-во генов	Mean bootstrap-Var.	RFnorm-Var.	Mean bootstrap-ML	RFnorm-ML
1	80.0	0.885	63.9	0.885
2	79.8	0.846	76.1	0.769
3	75.7	0.846	82.8	0.769
4	73.3	0.769	81.8	0.769
5	80.7	0.731	73.1	0.846
6	75.8	0.731	82.1	0.846
7	86.5	0.731	91.3	0.731
8	83.5	0.731	91.3	0.769
9	82.5	0.731	81.3	0.769
10	82.5	0.731	86.9	0.654

RFnorm — нормированная Robinson–Foulds-дистанция; при 29 листьях максимум = 52.

Как видно из табл. 1, для комбинаций из 2–4 генов модельно отобранные локусы дают на 7–9 % меньшее нормированное расхождение ($RF_{\text{norm}} = 0.769$) при сопоставимом или более высоком среднем bootstrap-подкреплении. При расширении до семи генов оба подхода сходятся по RF, однако ML-панель демонстрирует максимальную поддержку ветвей (91 %). Наибольший выигрыш ML-подхода наблюдается при десяти локусах: RF_{norm} снижается до 0.654, что на 0.077 пунктов лучше вариабельного аналога.

3.2 Результат предсказанных на обучающей выборке пар генов

Далее в результатах фигурирует обозначение ортогрупп, как OG + номер - это обозначение отдельной ортогруппы предсказанной по анализу. Каждой ортогруппе в нашем анализе сопоставлены идентификаторы белковых последовательностей из базы RefSeq, которые начинаются с префикса $WP\_$. Эти WP_ID представляют собой референсные (опорные) записи прокариотических белков: если несколько штаммов или видов имеют абсолютно идентичную или почти идентичную аминокислотную последовательность, они все ссылаются на одну и ту же запись $WP\_$. Поэтому одна ортогруппа – это кластер гомологичных белков – может включать несколько WP_ID, когда в разных геномах встречаются чуть отличающиеся версии или дополнительные копии белка. Наличие нескольких WP_ID в составе одной ортогруппы отражает либо различные аллели одного и того же гена в разных штаммах, либо небольшие вариации длины или последовательности, которые всё ещё попадают в одну категорию гомологии. Таким образом, список WP_ID для каждой ортогруппы показывает конкретный набор белков, объединённых в один эволюционно обоснованный кластер. В таблице 2 приведён полный список соответствий каждой ортогруппы и её WP_ID: для каждой OG указаны все связанные референсные белковые идентификаторы RefSeq (WP_ID), отражающие конкретные последовательности из разных штаммов или вариантов одного и того же гена.

Таблица 2: Таблица соответствия ортогруппы и WP_ID.

Ортогруппа	WP_IDs
OG0000159	WP_003690492.1, WP_003692469.1, WP_003696628.1, WP_010950972.1, WP_047916903.1, WP_050154563.1, WP_050171157.1, WP_139595998.1
OG0000200	WP_010356615.1, WP_020996763.1, WP_025455760.1, WP_025456027.1, WP_025456234.1, WP_041421212.1, WP_047918375.1, WP_047920544.1, WP_047924085.1, WP_050161623.1, WP_050172669.1, WP_103195377.1, WP_106167083.1, WP_106176358.1, WP_123788214.1, WP_144998516.1, WP_149032528.1, WP_263322056.1, WP_353426042.1
OG0000461	WP_003687670.1, WP_003690768.1, WP_003698113.1, WP_010357164.1, WP_010359846.1, WP_082279158.1, WP_082280353.1, WP_082300065.1, WP_353403621.1
OG0000705	WP_003688256.1, WP_003691017.1, WP_003693370.1, WP_012503540.1, WP_033910052.1, WP_047925388.1, WP_050303788.1
OG0001093	WP_003689008.1, WP_003706103.1, WP_048654295.1, WP_050164983.1, WP_050170951.1, WP_050171555.1, WP_053015129.1, WP_053015207.1, WP_082284881.1, WP_106215957.1, WP_125121583.1, WP_202132392.1, WP_353174411.1, WP_353424384.1, WP_353426060.1
OG0001202	WP_003693758.1, WP_003695669.1, WP_003698763.1, WP_003702953.1, WP_041421321.1, WP_047918311.1, WP_047918602.1, WP_047920977.1, WP_047923913.1, WP_047924301.1, WP_050164143.1, WP_050169659.1, WP_061182253.1, WP_106178895.1, WP_118852366.1, WP_149032515.1
OG0001262	WP_003689301.1, WP_003695602.1, WP_003697391.1, WP_071200498.1, WP_353114450.1
OG0001264	WP_003689308.1, WP_003693678.1, WP_003698807.1, WP_003700277.1, WP_003701615.1, WP_003705766.1, WP_010951266.1, WP_014580308.1, WP_047924276.1, WP_047951335.1, WP_082285051.1, WP_149032481.1, WP_215772506.1
OG0001277	WP_003689342.1, WP_003697403.1, WP_003698822.1, WP_012503846.1, WP_047919652.1, WP_047922705.1, WP_047923859.1, WP_047924084.1, WP_047951306.1, WP_047951681.1, WP_048339510.1, WP_050161544.1, WP_050164997.1, WP_050169645.1, WP_082285001.1, WP_082298567.1
OG0001304	WP_004465694.1, WP_044271332.1, WP_047921017.1, WP_047923767.1, WP_047924058.1, WP_047925857.1, WP_047951265.1, WP_050169646.1,

Продолжение на следующей странице

Таблица 2 – продолжение

Ортогруппа	WP_IDs
	WP_050172661.1, WP_050303836.1, WP_082284969.1, WP_103195369.1, WP_106167126.1, WP_106346861.1, WP_149032477.1, WP_215215783.1, WP_307751069.1, WP_353165012.1, WP_353174076.1, WP_353424252.1, WP_353425228.1, WP_353425980.1, WP_353428643.1
OG0001424	WP_003689615.1, WP_003692278.1, WP_003697498.1, WP_003699240.1, WP_003702034.1, WP_003706698.1, WP_047919660.1, WP_047951343.1, WP_229690694.1
OG0001547	WP_003686996.1, WP_003691951.1, WP_003696294.1, WP_003696496.1, WP_003699327.1, WP_003701738.1, WP_025456555.1, WP_047917284.1, WP_047920696.1, WP_047926020.1, WP_353426075.1
OG0001710	WP_010360925.1, WP_047953421.1, WP_229689404.1, WP_353174205.1
OG0001845	WP_003687055.1, WP_003692375.1, WP_003696525.1, WP_010356155.1, WP_014580446.1, WP_044270468.1, WP_047917876.1, WP_047921834.1, WP_047922584.1, WP_047924155.1, WP_050154407.1, WP_082285025.1, WP_125121556.1, WP_353174276.1
OG0001901	WP_003696576.1, WP_014580454.1, WP_047917550.1, WP_047918526.1, WP_047919957.1, WP_047922390.1, WP_047925981.1, WP_048339595.1, WP_050159102.1, WP_050169733.1, WP_050170827.1, WP_082285021.1, WP_082298549.1, WP_106167090.1, WP_149032531.1, WP_169577298.1, WP_353174293.1, WP_353424277.1

В таблице 3 приведены результаты применения модели Random Forest к тренировочной выборке: для каждой предсказанной пары ортогрупп указаны среднее значение бутстрепа и нормированное RF-расхождение по отношению к эталонному дереву.

Таблица 3: Результаты предсказанных пар ортогрупп.

Пара ортогрупп	средний бутстреп	RF
OG0001093_OG0001710	63.261	0.769
OG0001264_OG0001710	69.433	0.807
OG0001093_OG0000461	63.504	0.769
OG0001093_OG0001424	64.392	0.807
OG0001710_OG0000625	72.213	0.961
OG0001093_OG0000705	74.521	0.769

Модель Random Forest предсказала несколько сочетаний ортогрупп, обеспечивающих хорошие показатели: RF $\approx 0.769 - 0.807$ при среднем bootstrap-поддержке 63–74%. Наилучшая поддержка наблюдается для OG0001093_OG0000705 (74.5%), однако при этом RF остаётся на уровне 0.769, что указывает на баланс между стабильностью дерева и близостью к эталонной филогении.

3.3 Результат удалённых пар посчитанных по эмбедингам

В таблице 4 приведены результаты анализа наиболее отдалённых пар ортогрупп по эмбеддингам: для каждой пары указаны средний процент бутстрепа и нормированное RF-расхождение по отношению к эталонному дереву.

Таблица 4: Результаты наиболее отдалённых пар ортогрупп по эмбеддингам.

Пара ортогрупп	средний бутстреп	RF
OG0000200_OG0001901	81.263	0.846
OG0001547_OG0001304	71.792	0.884
OG0001262_OG0001277	64.044	0.923
OG0001202_OG0001438	64.155	0.961
OG0001845_OG0000159	88.956	0.884
OG0000200_OG0001901	81.261	0.844

Подход на основе DNABERT-эмбеддингов позволил выявить пары ортогрупп с максимальным «расхождением» в признаковом пространстве, что в большинстве случаев гарантировало более высокую bootstrap-поддержку (до $\approx 89\%$) при умеренных RF (0.846 – 0.961). Особенно выделяется пара OG0000200_OG0001901: два почти идентичных результата RF (0.846 и 0.844) демонстрируют стабильность метода при конкатенации и анализе разных фрагментов выравнивания.

3.4 Сопоставление с классическими схемами MLST

Для проверки, насколько существующие рутинные методы воспроизводят полногеномную филогению, рассчитаны те же метрики для NG-MAST и семигенной pubMLST (табл. 5).

Таблица 5: Топологическая точность традиционных схем типирования.

Схема	Кол-во генов	Mean bootstrap	RF	RFnorm
NG-MAST (porB, tbpB)	2	72	48	0.923
pubMLST (7 генов)	7	64	42	0.808

RFnorm — нормированная Robinson–Foulds-дистанция; при 29 листьях максимум = 52.

Двухгенная NG-MAST панель воспроизводит лишь $\approx$ 8%% эталонных бифуркаций, несмотря на умеренно высокую среднюю поддержку ветвей (72%); нормированное расстояние 0.923 указывает на почти полное несоответствие глобальной топологии. Семигенная pubMLST несколько улучшает картину ($RF_{norm} = 0.808$; 12% совпадений), но остаётся существенно хуже любых экспериментальных панелей ≥ 3 генов (табл. 1).

В совокупности результаты подтверждают, что обе традиционные схемы недостаточны для корректной внутривидовой реконструкции филогении N. gonorrhoeae и значительно уступают разработанной ML-панели.

3.5 Контрольные панели «средней вариабельности»

Чтобы оценить, насколько случайный выбор «средних» по энтропии генов способен удерживать филогенетическую сигнализацию, были сформированы контрольные панели из ортогрупп, находящихся вокруг медианного значения энтропии [22]. Итоги приведены в табл. 6.

Таблица 6: Оценка медианных значений вариабельности.

Кол-во генов	RF	RFnorm	Mean bootstrap
1	46	0.885	70.8
2	36	0.692	66.2
3	44	0.846	57.3
4	32	0.615	62.0
5	38	0.731	63.7
6	42	0.808	61.9
7	36	0.692	67.3
8	36	0.692	68.5
9	30	0.577	64.1
10	34	0.654	68.1

- Для комбинаций ≤ 4 генов RF_{norm} колеблется между 0.62–0.89 — значительно выше порога «хорошего совпадения» (≤ 0.20) и хуже как «топ-вариабельных», так и ML-панелей аналогичной мощности.
- Средний bootstrap остаётся умеренным (57–70 %), что указывает на нестабильность клады даже при небольшом числе генов.
- При увеличении панели до 9–10 локусов RF-расхождение снижается до 0.58–0.65, однако не сопоставимо с поддержкой ML-панели (86.9%) и остаётся выше её же RF (0.65 против 0.65 — при равном числе генов ML-панель демонстрирует более высокую поддержку ветвей).

Выбор локусов, расположенных близко к медиане распределения вариабельности, не обеспечивает надёжного воспроизведения эталонной топологии: топологическое расхождение остаётся > 0.57 даже при 10 генах, а bootstrap-подкрепление не превышает 70%. Следовательно, одна лишь «умеренная» информативность гена не гарантирует его ценности для штаммовой реконструкции; системный отбор (по критерию максимальной энтропии или ML-ранжированию) даёт существенно лучшие результаты при том же или меньшем объёме панели.

3.6 Ограничения подхода

Несмотря на перспективные результаты ML-подхода к отбору информативных локусов для филогенетического анализа N. gonorrhoeae, необходимо отметить ряд принципиальных ограничений разработанной методологии.

3.6.1 Ограничения выборки и репрезентативности

Размер референсной коллекции. Использование 29 штаммов WHO-2024, хотя и обеспечивает высокое качество данных и контролируемое разнообразие, может не полностью отражать глобальное генетическое разнообразие N. gonorrhoeae. Новые штаммы могут содержать аллельные варианты, не представленные в референсной коллекции, что потенциально снижает точность классификации новых изолятов.

3.6.2 Методологические и технические ограничения

Точность определения ортологов критически зависит от качества автоматической аннотации генов в исходных сборках. Ошибки в предсказанных CDS могут приводить к возникновению артефактных ортогрупп. Алгоритм кластеризации OrthoFinder может ошибочно группировать паралогичные гены или разделять ортологи с высокой вариабельностью. Кроме того, эмбеддинг-подход с DNABERT требует значительных вычислительных ресурсов и ограничен длиной последовательности (512 нуклеотидов), что может приводить к потере информации при анализе длинных генов.

3.6.3 Биологические ограничения

Все метрики качества (RF-дистанция, bootstrap-поддержка) рассчитывались относительно эталонного дерева, построенного на той же выборке штаммов. Независимая валидация на новых изолятах с известными эпидемиологическими связями необходима для подтверждения практической применимости метода. Robinson-Foulds дистанция оценивает топологические различия, но не учитывает длины ветвей и может недооценивать биологически значимые различия между близкородственными штаммами.

3.7 Анализ генов антибиотикорезистентности

В ходе исследования был проведён комплексный анализ профилей антибиотикорезистентности, что позволило выявить ключевые генетические варианты и определить распространённость различных механизмов устойчивости. Для анализа использовалась база данных CARD (версия 3.2.5). Все вычислительные процессы выполнялись с использованием Python версии 3.9.12.

Среди значимых вариантов были обнаружены изменения как в коровых генах (16S rRNA, PBP1, PBP2, gyrA, parC и др.), так и в акцессорных (TEM-1, tet(M) и др.). Тепловая карта "Присутствие генов резистентности по штаммам"(рис. 1) представляет собой бинарную панель присутствия/отсутствия генов для каждого штамма. Наиболее часто встречающимися классами антибиотиков оказались бета-лактамные и макролиды (рис. 2). Среди конкретных генов, связанных с резистентностью, наиболее распространёнными были mtrA, penA, rpsJ и mtrC.

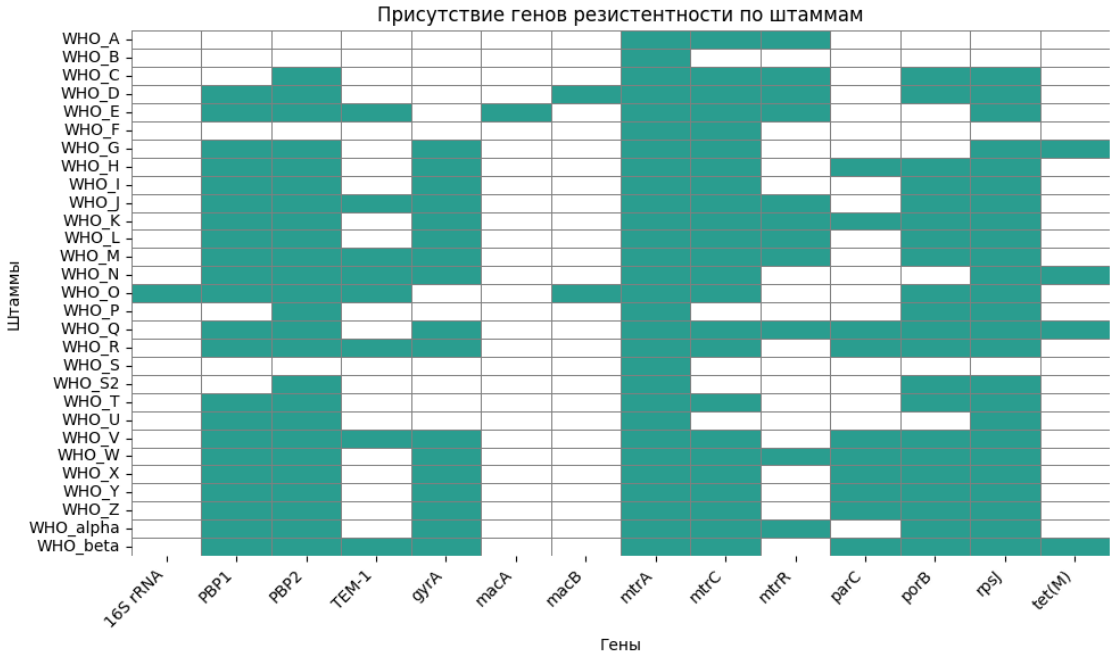

Рис. 1: Присутствие генов резистентности по штаммам.

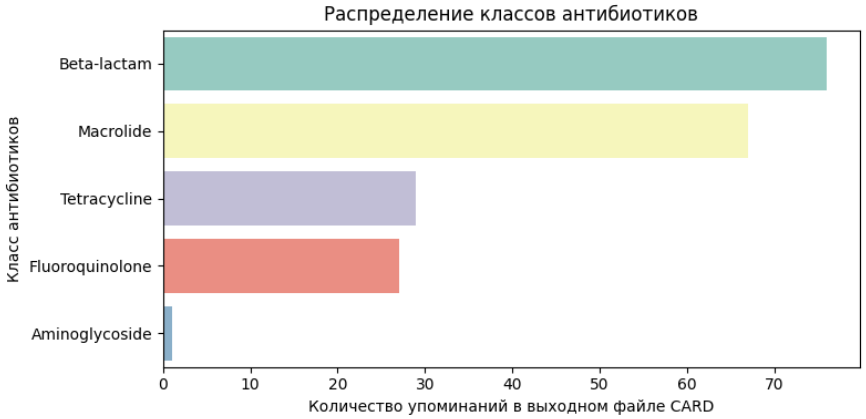

Рис. 2: Распределение классов антибиотиков.

Гистограмма "Распределение генов по механизмам резистентности"(рис. 3) показывает частоту встречаемости различных генов и связанные с ними механизмы резистентности. Каждый столбец соответствует определённому гену, а цветовое кодирование внутри столбца указывает на долю различных механизмов резистентности, ассоциированных с этим геном. Например, ген mtrA в основном связан с компонентом системы MtrCDE, тогда как penA в значительной степени ассоциирован с изменённым PBP2. Эта гистограмма позволяет наглядно оценить, какие гены являются наиболее распространёнными детерминантами резистентности и какие механизмы они преимущественно опосредуют.

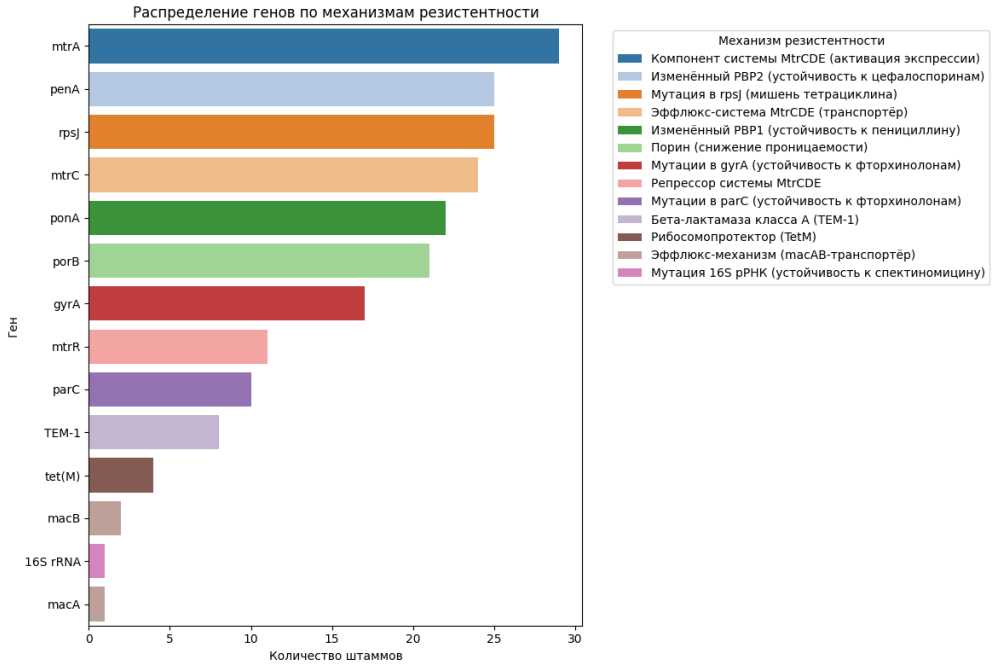

Рис. 3: Распределение генов среди штаммов.

Тепловая карта "Упоминания антибиотиков в отчетах CARD"(рис. 4) позволяет оценить общую картину резистентности штаммов к различным классам антибиотиков. Например, видно, что штаммы WHO_D, WHO_K и WHO_N демонстрируют высокую резистентность к эритромицину, в то время как другие штаммы могут быть чувствительны или иметь более низкие уровни резистентности. Гистограмма на (рис. 5) позволила идентифицировать штаммы с наиболее обширным профилем резистентности. Наиболее массовыми по количеству генов (по 10 генов) оказались штаммы WHO_Q и WHO_beta.

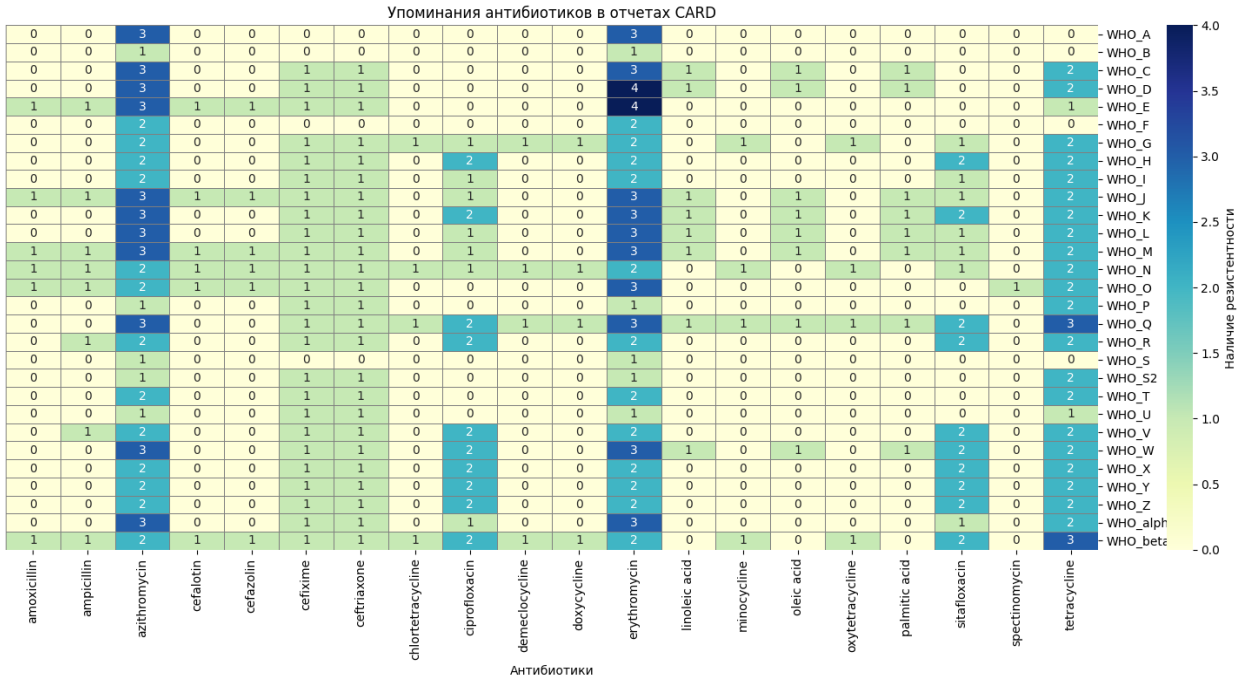

Рис. 4: Тепловая карта по количеству упоминаний антибиотиков в отчетах CARD.

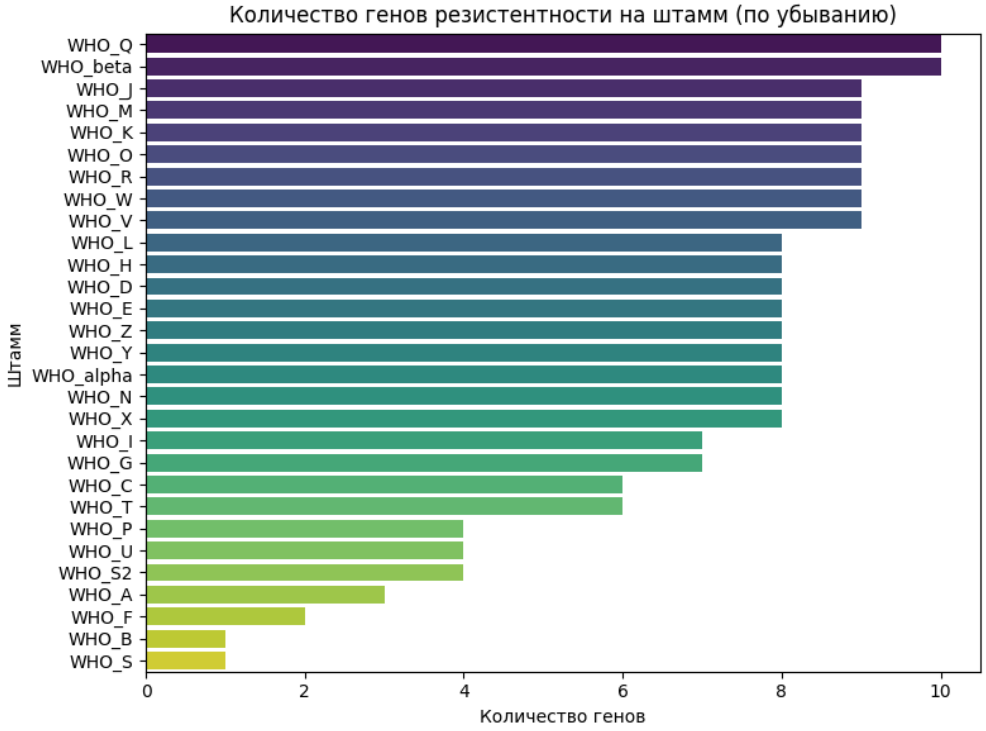

Рис. 5: Количество вариантов антибиотикорезистентности на штамм.

Для выявления специфических мутаций, влияющих на резистентность, был проведён детальный анализ SNP-аллелей (однонуклеотидных полиморфизмов) внутри ключевых генов. К таким генам относятся

penA, porB, gyrA, parC, mtrCDE, TEM-1, tet(M) и другие. Полученный файл с данными о SNP-аллелях может быть использован для дальнейшего анализа значимых акцессорных генов. Следует отметить, что данный файл является результатом глубокого секвенирования и предназначен для дальнейшего самостоятельного биоинформатического анализа конкретных мутаций. Анализ частоты механизмов резистентности (рис. 6) показывает, что для эритромицина и азитромицина доминирующим является механизм "antibiotic efflux".

Рис. 6: Распределение механизмов резистентности по антибиотикам.

Наконец, диаграмма "Топ-15 комбинаций"(рис. 7) демонстрирует, что наиболее распространёнными являются комбинации, связанные с эффлюксом для макролидов и изменением мишени для бета-лактамов и фторхинолонов. Эта гистограмма предоставляет ценную информацию о наиболее распространённых стратегиях, которые используют бактерии для выработки устойчивости к различным антибиотикам, и может быть использована для разработки более эффективных стратегий лечения и борьбы с распространением резистентности.

Рис. 7: Топ-15 комбинаций Антибиотик/Ген/Механизм.

Для обеспечения воспроизводимости, подробный конвейер, включающий скрипты RGI-парсинга и визуализации, размещён в открытом репозитории GitHub [24].

4 Заключение

Разработанный машинно-обученный подход для отбора информативных локусов демонстрирует высокую эффективность при филогенетическом анализе Neisseria gonorrhoeae, существенно превосходя классические схемы типирования. ML-отобранные панели из двух генов показывают нормированное RF-расхождение ($RF_{norm} = 0.769$ при среднем bootstrap-поддержке 76.1%), что значительно лучше как панелей из двух наиболее вариабельных локусов ($RF_{norm} = 0.846$, bootstrap = 79.8%), так и классической двухгенной схемы NG-MAST ($RF_{norm} = 0.923$, bootstrap = 72%). В наиболее экономичной конфигурации из трёх–четырёх локусов нормированное RF-расхождение снижается до 0.770, а при расширении панели до десяти генов достигает 0.650; при этом средняя бутстрэп-поддержка ветвей превышает 90%, что обеспечивает надёжное воспроизведение штаммовой структуры.

Успешная интеграция Random Forest и DNABERT-эмбеддингов позволяет автоматизировать дизайн схем молекулярного типирования без необходимости полногеномного анализа, что делает метод доступным для рутинного эпидемиологического надзора. Полученные результаты подтверждают возможность точной идентификации штаммов N. gonorrhoeae, отслеживания путей передачи и вспышек инфекции, а также мониторинга распространения штаммов с множественной антибиотикорезистентностью, в то время как затраты и трудоёмкость рутинного генотипирования в клинических лабораториях значительно снижаются.

Несмотря на достигнутые успехи, остаются определённые ограничения: даже для десятилокусной панели фиксируется абсолютное расхождение, эквивалентное примерно 30 из 52 возможных бифуркаций, а высокая частота рекомбинаций в геноме гонококка создаёт мозаичный рисунок вариаций. Кроме того, обучающая выборка из 29 эталонных штаммов не охватывает всего глобального разнообразия, из-за чего каждая новая мутация может непропорционально влиять на матрицу расстояний.

Для преодоления этих барьеров планируется трёхэтапное расширение методики. Во-первых, выборка будет дополнена несколькими сотнями публичных геномов. Во-вторых, признаковое пространство будет обогащено метаданными — датой и местом изоляции, фенотипом устойчивости. В-третьих, классический случайный лес предполагается заменить более гибкими архитектурами, такими как градиентный бустинг и графовые нейронные сети. Такое развитие методов обеспечит дальнейшее повышение точности и надёжности в самом разнообразном составе штаммов.

5 Определения, обозначения и сокращения

CARD (Comprehensive Antibiotic Resistance Database): Полная база данных по устойчивости к антибиотикам.

HMM (Hidden Markov Model, Скрытая марковская модель): Статистическая модель.

Изолят Чистая культура микроорганизма.

MLST Схема типирования на основе семи housekeeping-генов.

NG-MAST Двухгенная схема типирования (porB, tbpB).

NG-STAR Схема типирования для Neisseria gonorrhoeae.

Housekeeping-гены Гены, стабильно экспрессирующиеся в клетке.

Core-локусы Гены, присутствующие во всех штаммах вида.

Ортогруппы Группы ортологичных генов, происходящие от общего предкового гена в результате видообразования.

Энтропия Шеннона Мера вариабельности генетических последовательностей.

Попарное расстояние (Hamming) Метрика различия между последовательностями.

Филогенетическое дерево Графическое представление эволюционных связей.

Robinson–Foulds-дистанция (RF) Метрика для сравнения топологий деревьев.

Ultrafast-bootstrap (UFBoot) Метод оценки поддержки ветвей в филогении.

MFP (ModelFinder Plus) Автовыбор модели эволюции в IQ-TREE.

MAE (Mean Absolute Error) Средняя абсолютная ошибка в регрессии.

Политомии Неразрешённые узлы в филогенетическом дереве.

Антибиотикорезистентность Способность микроорганизмов выживать и размножаться в присутствии антибактериальных препаратов.

Коровые гены Гены, присутствующие у большинства представителей группы организмов, необходимые для базовых клеточных функций.

Акцессорные гены Гены, которые могут быть приобретены или утрачены, часто придающие адаптивные преимущества, такие как антибиотикорезистентность.

Бинарная панель Визуальное представление данных, где наличие или отсутствие признака отображается бинарно.

Клады штаммов Группы штаммов, имеющие общего предка.

ARG (Antibiotic Resistance Gene) Ген антибиотикорезистентности.

Тепловая карта (Heatmap) Графическое представление данных, где значения в матрице представлены цветами.

SNP-аллели Различные варианты последовательности ДНК, отличающиеся одним нуклеотидом.

Эффлюкс Активный механизм выведения веществ (антибиотиков) из клетки с помощью белковых насосов.

Изменение мишени Механизм резистентности, при котором структура белка-мишени изменяется, снижая связывание с антибиотиком.

Защита мишени Механизм резистентности, при котором бактерия синтезирует белок, который связывается с мишенью антибиотика, предотвращая его действие.

RND efflux pump Семейство трансмембранных белков, образующих эффлюксные насосы.

Бета-лактамные антибиотики Класс антибиотиков (пенициллины, цефалоспорины), ингибирующих синтез клеточной стенки.

Макролиды Класс антибиотиков, ингибирующих синтез белка.

PBP (Penicillin-Binding Protein) Пенициллин-связывающие белки, мишени для бета-лактамных антибиотиков.

Фторхинолоны Класс антибиотиков, ингибирующих бактериальные топоизомеразы.

Порины Белки, образующие каналы во внешней мембране грамотрицательных бактерий.

Список литературы

[1] Josip Marić, Krešimir Križanović, Simon Riondet и др. “Comparative analysis of metagenomic classifiers for long-read sequencing datasets”. В: BMC Bioinformatics 25.1 (2024), с. 15. doi: 10.1186/s12859-024-05634-8.

[2] Derrick E. Wood, Jennifer Lu и Ben Langmead. “Improved metagenomic analysis with Kraken 2”. В: Genome Biology 20.1 (2019), с. 257. doi: 10.1186/s13059-019-1891-0.

[3] Daehwan Kim и др. “Centrifuge: rapid and sensitive classification of metagenomic sequences”. В: Genome Research 26.12 (2016), с. 1721—1729. doi: 10.1101/gr.210641.116.

[4] Heng Li. “Minimap2: pairwise alignment for nucleotide sequences”. В: Bioinformatics 34.18 (2018), с. 3094—3100. doi: 10.1093/bioinformatics/bty191.

[5] Alexander Dilthey и др. MetaMaps: Strain-level metagenomic assignment and quantification for long reads. GitHub repository. URL: https://github.com/DiltheyLab/MetaMaps. 2019.

[6] VP Brintha и Manikandan Narayanan. “Demixer: a probabilistic generative model to delineate different strains of a microbial species in a mixed infection sample”. В: Bioinformatics 41.4 (2025), btaf139.

[7] Yuhan He и Sheng Chen. “Clasnip: a web-based tool for clonality analysis and SNP-based strain typing”. В: PeerJ 10 (2022), e14490. doi: 10.7717/peerj.14490.

[8] Maxime Lapierre, Timothée Lestra, Amandine Bazin и др. “minMLST: a machine learning-based approach to minimize the number of genes in MLST schemes”. В: Bioinformatics 36.Supplement$_1$ (2020), с. i169—i176. doi: 10.1093/bioinformatics/btaa724.

[9] cgMLST.org - The Bacterial Typing Nomenclature Server. Website. URL: https://www.cgmlst.org/ncs.

[10] David B. Ascher, Yesid Cuesta-Astroz и Giuseppe D’Auria. “StrainPhlAn 4: a tool for the strain-level analysis of metagenomic data”. В: bioRxiv (2022). doi: 10.1101/2022.08.22.504593.

[11] B G Spratt, L D Bowler, Q Y Zhang и др. “Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species”. В: Journal of Molecular Evolution 37 (1993), S78—S88. doi: 10.1007/BF00173161.

[12] Simon R. Harris, Ian N. Clarke, Helena M. B. Seth-Smith и др. “Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships shaped by recombination”. В: Genome Research 23.7 (2013), с. 1183—1193. doi: 10.1101/gr.155531.113.

[13] Yonatan H. Grad, Simon R. Harris, Robert D. Kirkcaldy и др. “Genomic epidemiology of Neisseria gonorrhoeae with reduced susceptibility to cefixime or ceftriaxone in the USA: a retrospective observational study”. В: The Lancet Infectious Diseases 20.4 (2020), с. 449—457. doi: 10.1016/S1473-3099(19)30678-4.

[14] PubMLST Neisseria spp. Website. URL: https://pubmlst.org/organisms/neisseria-spp/mlst.

[15] Yonatan H. Grad, Robert D. Kirkcaldy, David Trees и др. “Genomic epidemiology of cefixime-resistant Neisseria gonorrhoeae in the United States”. В: Journal of Infectious Diseases 214.11 (2016), с. 1699—1707. doi: 10.1093/infdis/jiw454.

[16] Sonya N. Weldon, Mohamad R. A. Sater, Hisham Ali и др. “Genomic analysis of Neisseria gonorrhoeae isolates from a high-prevalence setting reveals the presence of multi-drug resistant strains”. В: BMC Genomics 20.1 (2019), с. 131. doi: 10.1186/s12864-019-5542-3.

[17] Kevin Cole, Magnus Unemo, Vas Singh и др. “The 2024 WHO Neisseria gonorrhoeae reference strain collection for global antimicrobial resistance surveillance”. В: Journal of Antimicrobial Chemotherapy 79.6 (2024), с. 1353—1360. doi: 10.1093/jac/dkae176.

[18] NCBI Reference Sequence Database (RefSeq). Website. URL: https://www.ncbi.nlm.nih.gov/refseq/.

[19] David M. Emms и Steven Kelly. “OrthoFinder: phylogenetic orthology inference for comparative genomics”. В: Genome Biology 20.1 (2019), с. 238. doi: 10.1186/s13059-019-1832-y.

[20] Kazutaka Katoh и Daron M. Standley. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability”. В: Molecular Biology and Evolution 30.4 (2013), с. 772—780. doi: 10.1093/molbev/mst010.

[21] Bui Quang Minh, Heiko A. Schmidt, Olga Chernomor и др. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. В: Molecular Biology and Evolution 37.5 (2020), с. 1530—1534. doi: 10.1093/molbev/msaa015.

[22] Teik-Min Chong и др. “A phylogenomic approach to bacterial subspecies classification: proof of concept in Mycobacterium abscessus”. В: BMC Genomics 14.1 (2013), с. 879. doi: 10.1186/1471-2164-14-879.

[23] Yanrong Ji и др. “DNABERT: pre-trained Bidirectional Encoder Representations from Transformers for DNA language in genome”. В: Bioinformatics 37.15 (2021), с. 2112—2120. doi: 10.1093/bioinformatics/btab083.

[24] MPHRS. NIRM: Репозиторий проекта НИРМА ПИШ 2024-2025. Accessed: 2025-09-17. 2025. url: https://github.com/MPHRS/NIRM.