**Machine learning–assisted selection of informative loci for strain-level phylogenetics of *Neisseria gonorrhoeae***

## **ВВЕДЕНИЕ**

Метагеномные классификаторы, подробно рассмотренные Marić и соавт. (2024) [<https://doi.org/10.1186/s12859-024-05634-8>], демонстрируют удовлетворительную точность видовой идентификации, однако при переходе к штаммовому уровню сталкиваются с рядом принципиальных ограничений. У k-mer-ориентированных решений Kraken2 [<https://doi.org/10.1186/s13059-019-1891-0>] и Centrifuge [<https://ccb.jhu.edu/software/centrifuge/>] избыточная полиморфность коротких олигонуклеотидов приводит к ложноположительным классификациям. Mapping-подходы Minimap2 [<https://doi.org/10.1093/bioinformatics/bty191>] и MetaMaps [<https://github.com/DiltheyLab/MetaMaps>] 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 [<https://doi.org/10.1093/bioinformatics/btaf139>] сочетает байесовскую оценку долей известных клонов с реконструкцией *de novo* SNP-профилей, устраняя «reference-bias», но надежно работает лишь при покрытии ≥ 20× и чувствителен к рекомбинации. MADRe [<https://doi.org/10.1101/2025.05.12.653324>] уменьшает число ложноположительных детекций за счёт предварительной сборки контигов и автоматического сокращения базы, однако удлиняет время анализа и требует хорошего покрытия. Clasnip [<https://doi.org/10.7717/peerj.14490>] 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** [<https://doi.org/10.1093/bioinformatics/btaa724>], где методом XGBoost отбирается минимальный набор локусов из cgMLST-схем [<https://www.cgmlst.org/ncs>] при сохранении приемлемого разрешения (скорректированный индекс Рэнда 0,4–0,93), активно развиваются и marker-ориентированные метагеномные методы, например StrainPhlAn [<https://doi.org/10.1101/2022.08.22.504593>]. Последний использует видоспецифичные маркеры или пангеномные базы и потенциально может быть адаптирован для анализа изолятов при соответствующей модификации алгоритма.

Тем не менее ни один из существующих инструментов не объединяет высокое штаммовое разрешение, умеренные вычислительные требования и полноценную интеграцию в рутинную диагностику.

#### **Специфика *Neisseria gonorrhoeae***

*Neisseria gonorrhoeae* отличается минимальным межгеномным расстоянием между циркулирующими клонами (≤ 0,001 замен/сайт) и высоким уровнем рекомбинации (r/m ≈ 2,5) [<https://doi.org/10.1099/00221287-139-11-2603>;<https://doi.org/10.7717/peerj.806>;<https://doi.org/10.1186/s12864-020-6511-6>]. Эти особенности ограничивают возможности:

* **MLST** [<https://pubmlst.org/organisms/neisseria-spp/mlst>] — схема из семи housekeeping-генов даёт согласованную, но недостаточно детализированную картину;
* **NG-MAST** [<https://doi.org/10.1099/mgen.0.000076>] — 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**[<https://doi.org/10.1186/s12864-019-5542-3>] — 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 ≥ 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.

#### **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.

## **МАТЕРИАЛЫ И МЕТОДЫ**

### 1. Характеристика и получение геномных данных

В качестве исходных данных использовалась коллекция из 29 референсных штаммов *N. gonorrhoeae*, утверждённая ВОЗ в 2024 г [<https://doi.org/10.1093/jac/dkae176> ]. 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 [<https://www.ncbi.nlm.nih.gov/> ] 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. Определение ортологов и выравнивание

В скачанных CDS данных определялись последовательности однокопийных ортогрупп с использованием **OrthoFinder v2.5.5** [<https://github.com/davidemms/OrthoFinder> ] (режим --dna); полученные последовательности далее использовались для филогенетического анализа. Внутри каждой отдельной ортогруппы проводилось выравнивание с использованием програмного пакета MAFFT (v7.525)(режим --auto) [<https://mafft.cbrc.jp/alignment/software/>] и рассчитывалась средняя энтропия Шеннона, а также другие метрики характеризующие вариабельность ортогруппы.

### 3. Построение эталонной и вариативных филогенетических реконструкций

*Эталонное дерево.* Для оценки качества разработанной системы типирования мы построили эталонное филогенетическое дерево на основании всех одно́кратных ортогрупп, выявленных с помощью OrthoFinder (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. [<https://doi.org/10.1093/molbev/msaa015> ].

### 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.

, (1)

где *n* — число листьев. При *n = 29* штаммах . В работе дополнительно используем **нормированное расстояние** . Значения рассматриваются как «хорошее» совпадение топологий, а — как практически не отличимое от эталона.

*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) [<https://doi.org/10.1186/1471-2164-14-879>]. 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.

### 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 [<https://doi.org/10.1093/bioinformatics/btab083>], после чего применяют усреднение по всем k‑мерам и по всем фрагментам, получая единый 768‑мерный вектор (embedding) для каждой пары ортогрупп.

### 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.

### 7. Проверка ML-панели

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.

### 8. Сопоставление с классическими схемами MLST

*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 (URL будет указан при публикации), включающем все скрипты, настройки окружения и примеры команд для запуска конвейера.

### 9. 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.

## **РЕЗУЛЬТАТЫ**

### 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 |

*— 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 (=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: ​ снижается до 0.654, что на 0.077 пунктов лучше вариабельного аналога.

#### 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\_), отражающие конкретные последовательности из разных штаммов или вариантов одного и того же гена.

Таблица 2. Таблица соответсвия 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,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 ≈ 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.

#### 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 ~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.

#### 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 (abcZ, adk, aroE, fumC, gdh, pdhC, pgm) | 7 | 64 | 42 | 0.808 |

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

Двухгенная **NG-MAST** панель воспроизводит лишь ≈ 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 (= 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, демонстрирующей вдвое меньшее RF-расхождение при сопоставимой или более высокой поддержке клад.

#### 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. 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 |

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

* Для комбинаций ≤ 4 генов 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.

#### Ограничения подхода

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

*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.

*Методологические ограничения*

Точность определения ортологов критически зависит от качества автоматической аннотации генов в исходных сборках. Ошибки в предсказанных CDS или не точные границы в последовательностях могут приводить к возникновению артефактных ортогрупп и снижению надежности филогенетической реконструкции.

Ограничения существующих инструментов. Алгоритм кластеризации ортологов, реализованного в рамках программного пакета OrthoFinder, может ошибочно группировать паралогичные гены или разделять ортологи с высокой вариабельностью последовательностей.

*Технические ограничения*

Вычислительные требования DNABERT. Эмбеддинг-подход требует значительных вычислительных ресурсов и ограничен длиной последовательности (512 нуклеотидов), что может приводить к потере информации при анализе длинных генов или регионов с множественными вставками/удалением.

Ограничения длины последовательностей. Фрагментация последовательностей для соответствия требованиям DNABERT может нарушать локальные структурные мотивы и снижать качество эмбеддингов, особенно для генов с выраженной доменной организацией.

*Биологические ограничения*

Отсутствие независимой валидации. 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.

Ограниченность RF-метрики. Robinson-Foulds дистанция оценивает топологические различия деревьев, характеризующее разделение штаммов, но не учитывает длины ветвей и может недооценивать биологически значимые различия между близкородственными штаммами.

### Анализ генов антибиотикорезистентности

В ходе исследования был проведён комплексный анализ профилей антибиотикорезистентности с использованием вычислительного подхода, что позволило выявить ключевые генетические варианты, определить распространённость различных механизмов устойчивости и идентифицировать наиболее значимые классы антибиотиков, к которым наблюдается резистентность.

#### Используемые инструменты и методы

Для анализа антибиотикорезистентности использовалась база данных CARD (версия 3.2.5 от 15 мая 2024 года). CARD - это высококачественная биоинформатическая база данных, которая связывает последовательности генов с механизмами резистентности и соответствующими антибиотиками, обеспечивая всестороннюю информацию о детерминантах антибиотикорезистентности. Все вычислительные процессы выполнялись с использованием Python версии 3.9.12.

#### Значимые генетические варианты и их распределение

Среди значимых вариантов, влияющих на антибиотикорезистентность, были обнаружены изменения как в коровых генах, так и в акцессорных генах. К коровым генам относятся 16S rRNA, PBP1 (ponA), PBP2 (penA), gyrA, parC, mtrC, mtrR, porB, rpsJ. Акцессорные гены включали TEM-1, TEM-135, macA, macB, mtrA, tet(M).

Тепловая карта "Присутствие генов резистентности по штаммам" (рис. 1) представляет собой бинарную панель присутствия/отсутствия генов резистентности для каждого исследуемого штамма. Зелёные ячейки указывают на присутствие конкретного гена у данного штамма, белые – на его отсутствие. Данная визуализация позволяет быстро оценить наличие антибиотикорезистентности у каждого штамма и сопоставить эти данные с кладами штаммов. Однако важно отметить, что эта панель не предоставляет информации об источнике резистентности (конкретном гене) или её механизме действия.

#### Распределение генов антибиотикорезистентности (ARG) по классам антибиотиков

Анализ распределения частот генов антибиотикорезистентности (ARG) по классам антибиотиков позволил получить общую картину резистентности в исследуемой выборке и выявить наиболее распространённые гены и классы антибиотиков, к которым наблюдается устойчивость. Наиболее часто встречающимися классами антибиотиков оказались бета-лактамные антибиотики и макролиды (рис. 2). Среди конкретных генов, связанных с резистентностью, наиболее распространёнными были mtrA, penA, rpsJ и mtrC.

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.

#### Влияние генов на антибиотикорезистентность и распространённость штаммов

Тепловая карта "Упоминания антибиотиков в отчетах CARD" (рис. 4) представляет собой тепловую карту, где каждая ячейка отображает количество упоминаний конкретного антибиотика в отчётах CARD для каждого штамма. Интенсивность синего цвета указывает на более высокую степень резистентности или на наличие большего количества генов, ассоциированных с резистентностью к данному антибиотику у конкретного штамма. Эта карта 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.

Построена гистограмма (рис. 5), отображающая штаммы по убыванию количества обнаруженных у них генов резистентности. Это позволило идентифицировать штаммы с наиболее обширным профилем резистентности. Наиболее массовыми по количеству генов (по 10 генов) оказались штаммы WHO\_Q и WHO\_beta.

#### Анализ SNP-аллелей и механизмы устойчивости

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.

График "Частота механизмов резистентности по антибиотикам" (рис. 6) представляет собой тепловую карту, которая демонстрирует частоту встречаемости различных механизмов резистентности для каждого класса антибиотиков. Цветовая шкала справа показывает частоту, от низкой (светло-желтый) до высокой (тёмно-зелёный). Числовые значения в ячейках указывают на точное количество случаев обнаружения данного механизма резистентности к определенному антибиотику. Например, для эритромицина и азитромицина механизм "antibiotic efflux" является доминирующим (64 и 67 случаев соответственно). Это позволяет понять, какие механизмы резистентности наиболее распространены для конкретных антибиотиков.

#### Топ-15 наиболее частых комбинаций "Антибиотик/Ген/Механизм"

Данный рисунок представляет собой горизонтальную столбчатую диаграмму (рис. 7), которая отображает 15 самых распространенных комбинаций антибиотик/ген/механизм устойчивости, обнаруженных в исследуемой выборке. Ось Y показывает конкретную комбинацию, а ось X – её частоту встречаемости.

Интерпретация гистограммы:

* Доминирование эффлюкса: Наиболее распространёнными комбинациями оказались "erythromycin/RND efflux pump/efflux" и "azithromycin/RND efflux pump/efflux". Это свидетельствует о том, что эффлюкс (выкачивание антибиотиков из клетки белковыми насосами), опосредованный RND-насосами, является ключевым механизмом устойчивости к макролидам (эритромицину и азитромицину) в данной выборке. Макролиды — это важный класс антибиотиков, широко используемых для лечения бактериальных инфекций. Их устойчивость через эффлюксные насосы представляет серьезную клиническую проблему.
* Роль Tet. Rib. Prot. и PBP мутаций: Комбинация "tetracycline / Tet. Rib. Prot. / target protection" указывает на то, что устойчивость к тетрациклину (широко используемому антибиотику, ингибирующему синтез белка) часто обусловлена защитой мишени рибосомными защитными белками. Мутации в PBP (пенициллин-связывающих белках), как видно из комбинаций с цефтриаксоном и цефиксимом, являются значимым механизмом изменения мишени для бета-лактамных антибиотиков (цефалоспоринов), что приводит к снижению их аффинности к целевой мишени.
* Снижение проницаемости: Для тетрациклина также отмечена комбинация "tetracycline / Bacterial Porin (red. perm.) / reduced permeability", что указывает на вклад пониженной проницаемости бактериальной оболочки за счёт изменений в поринах как механизма резистентности.
* Гены gyrA и parC в резистентности к фторхинолонам: Комбинации с ситафлоксацином и ципрофлоксацином показывают, что устойчивость к фторхинолонам (таким как ципрофлоксацин и ситафлоксацин, ингибирующие бактериальные топоизомеразы) часто связана с изменениями мишени (генов *gyrA* и *parC*).

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.

#### Графические представления и воспроизводимость данных

Все вышеупомянутые графические представления были построены и активно использовались во «вторичном анализе». Эти визуализации сыграли ключевую роль в интерпретации эпидемиологически значимых кластеров штаммов и в выработке рекомендаций по созданию таргетных ПЦР-панелей. Они позволили эффективно обнаруживать кластеры штаммов со схожими профилями резистентности, что является крайне важным для эпидемиологических исследований и принятия решений в области общественного здравоохранения.

Для обеспечения воспроизводимости и открытости исследования, подробный воспроизводимый конвейер, включающий скрипты RGI-парсинга, визуализации и сводные таблицы, размещён в открытом репозитории GitHub. Ссылка на репозиторий будет указана в Приложении.

## **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 (RFnorm = 0.769 при среднем bootstrap‑поддержке 76.1 %), что значительно лучше как панелей из двух наиболее вариабельных локусов (RFnorm = 0.846, bootstrap = 79.8 %), так и классической двухгенной схемы NG‑MAST (RFnorm = 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, при котором локусы с «вертикальным» сигналом быстро насыщаются параллельными заменами. 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, а агрегированные статистики Random Forest не всегда различают конвергентные и истинно филогенетические изменения.

Для преодоления этих барьеров планируется трёхэтапное расширение методики. Во-первых, выборка будет дополнена несколькими сотнями публичных геномов, что обеспечит модели представленность редкими и географически ограниченными кластерами. Во-вторых, признаковое пространство будет обогащено метаданными — датой и местом изоляции, фенотипом минимальной подавляющей концентрации, типом детерминант устойчивости, — позволяющими отделить сигналы селективного давления от вертикальной эволюции. В-третьих, классический случайный лес предполагается заменить более гибкими архитектурами: сочетания градиентного бустинга, учитывающего взаимосвязи между ортогруппами, и графовых нейронных сетей, способных работать с рекомбинационно‑информативными метриками. Такое развитие методов обеспечит дальнейшее повышение точности и надёжности в самом разнообразном составе штаммов.

## **ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ**

В настоящем отчете о НИР применяют следующие сокращения и обозначения:

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

**HMM** (Hidden Markov Model, Скрытая марковская модель): это 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

**Сore-локусы**: Гены, присутствующие во всех штаммах вида.

**Ортогруппы:** Группы ортологичных генов - происходят от общего предкового гена в результате видообразования

**Энтропия Шеннона:** Мера вариабельности генетических последовательностей.

**Парсимониально-информативные позиции:** Позиции в выравнивании, информативные для филогении.

**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, которые обычно подавляют их рост или убивают.

**Коровые гены:** Гены, присутствующие у большинства или всех представителей определённой группы организмов, обычно необходимые для их выживания и базовых клеточных функций.

**Акцессорные гены (дополнительные гены):** Гены, которые могут быть приобретены или утрачены организмами, не являясь строго необходимыми для выживания, но часто придающие адаптивные преимущества, такие как антибиотикорезистентность.

**Бинарная панель присутствия/отсутствия:** Визуальное представление данных, где наличие или отсутствие определённого признака (в данном случае, гена резистентности) отображается в виде бинарного значения (например, цветом).

**Клады штаммов:** Группы штаммов, имеющие общего предка, определяемые на основе филогенетического анализа.

**ARG (Antibiotic Resistance Gene):** Ген антибиотикорезистентности – генетическая последовательность, кодирующая белок или РНК, которая обеспечивает устойчивость к одному или нескольким антибиотикам.

**Тепловая карта (Heatmap):** Графическое представление данных, где значения в матрице представлены цветами. Используется для визуализации закономерностей и кластеров в больших массивах данных.

**Гистограмма:** Графическое представление распределения числовых данных.

**SNP-аллели (Single Nucleotide Polymorphism alleles):** Различные варианты последовательности ДНК, отличающиеся одним нуклеотидом в определённой позиции.

**Эффлюкс:** Активный механизм выведения веществ (в данном случае, антибиотиков) из клетки с помощью специализированных белковых насосов.

**Изменение мишени:** Механизм резистентности, при котором структура белка-мишени антибиотика изменяется, снижая или исключая его связывание с антибиотиком.

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

**RND efflux pump (Resistance-Nodulation-Cell Division efflux pump):** Семейство трансмембранных белков, образующих эффлюксные насосы, которые активно выкачивают широкий спектр антибиотиков и других токсичных соединений из бактериальной клетки.

**Бета-лактамные антибиотики:** Широкий класс антибиотиков, включающий пенициллины, цефалоспорины, карбапенемы и монобактамы, которые действуют, ингибируя синтез бактериальной клеточной стенки.

**Макролиды:** Класс антибиотиков, ингибирующих синтез белка бактериями путём связывания с 50S субъединицей рибосом.

**Tetracycline Ribosomal Protection Protein (Tet. Rib. Prot.):** Белки, которые связываются с рибосомами и предотвращают действие тетрациклиновых антибиотиков.

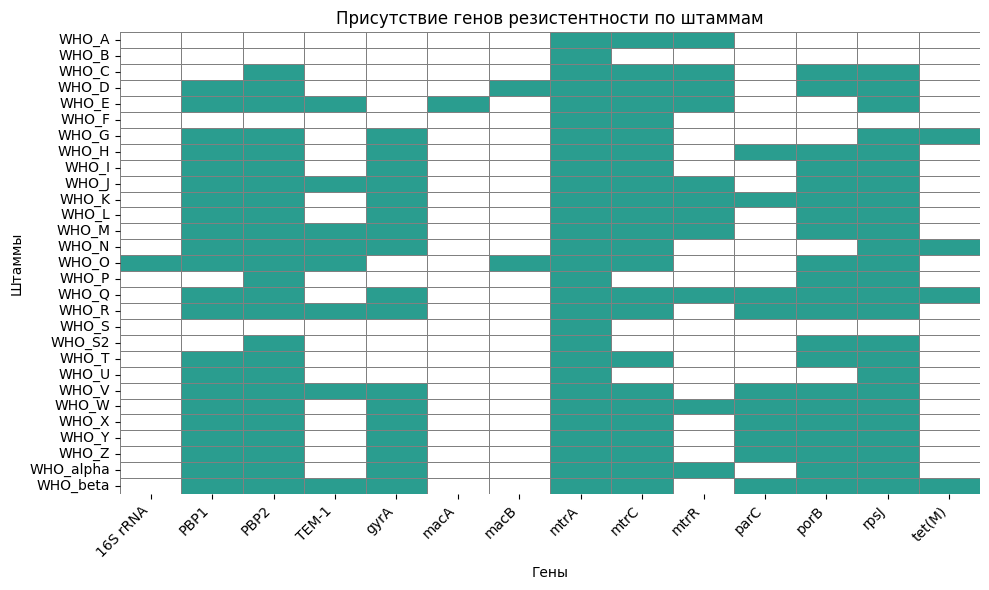
**PBP (Penicillin-Binding Protein):** Пенициллин-связывающие белки – ферменты, участвующие в синтезе пептидогликана клеточной стенки бактерий, являющиеся мишенями для бета-лактамных антибиотиков. Мутации в этих белках могут приводить к резистентности.

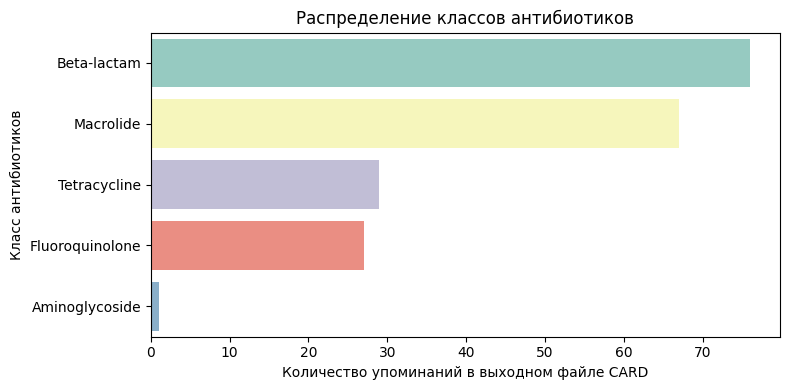
**Фторхинолоны:** Класс антибиотиков, ингибирующих бактериальные топоизомеразы (ДНК-гиразу и топоизомеразу IV), необходимые для репликации и деления бактериальной ДНК.

**Порины:** Белки, образующие каналы во внешней мембране грамотрицательных бактерий, через которые могут проникать гидрофильные молекулы, включая антибиотики. Изменения в поринах могут приводить к снижению проницаемости и резистентности.

**Таргетные ПЦР-панели:** Наборы праймеров для полимеразной цепной реакции (ПЦР), разработанные для обнаружения специфических генов или мутаций, связанных с антибиотикорезистентностью.

## **ПРИЛОЖЕНИЕ**

Рисунок 1 - Присутствие генов резистентности по штаммам

Рисунок 2 - Распределение классов антибиотиков

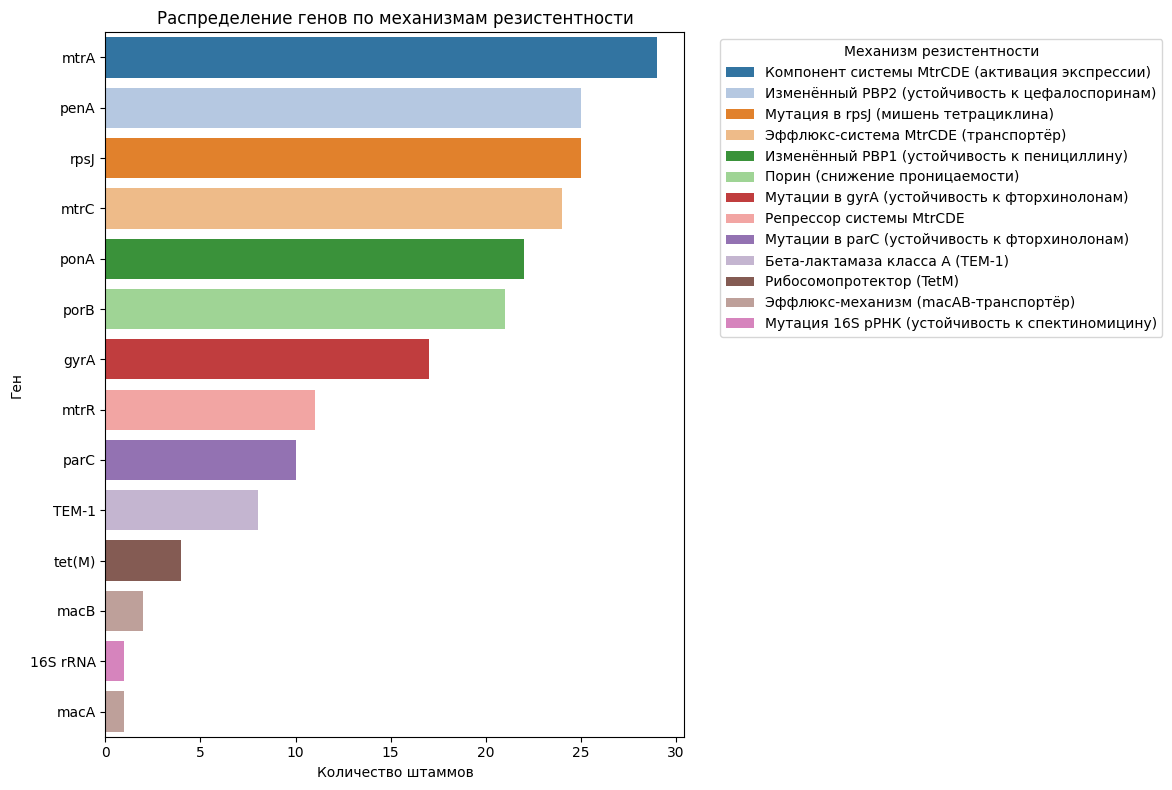


Рисунок 3 - Распределение генов среди штаммов

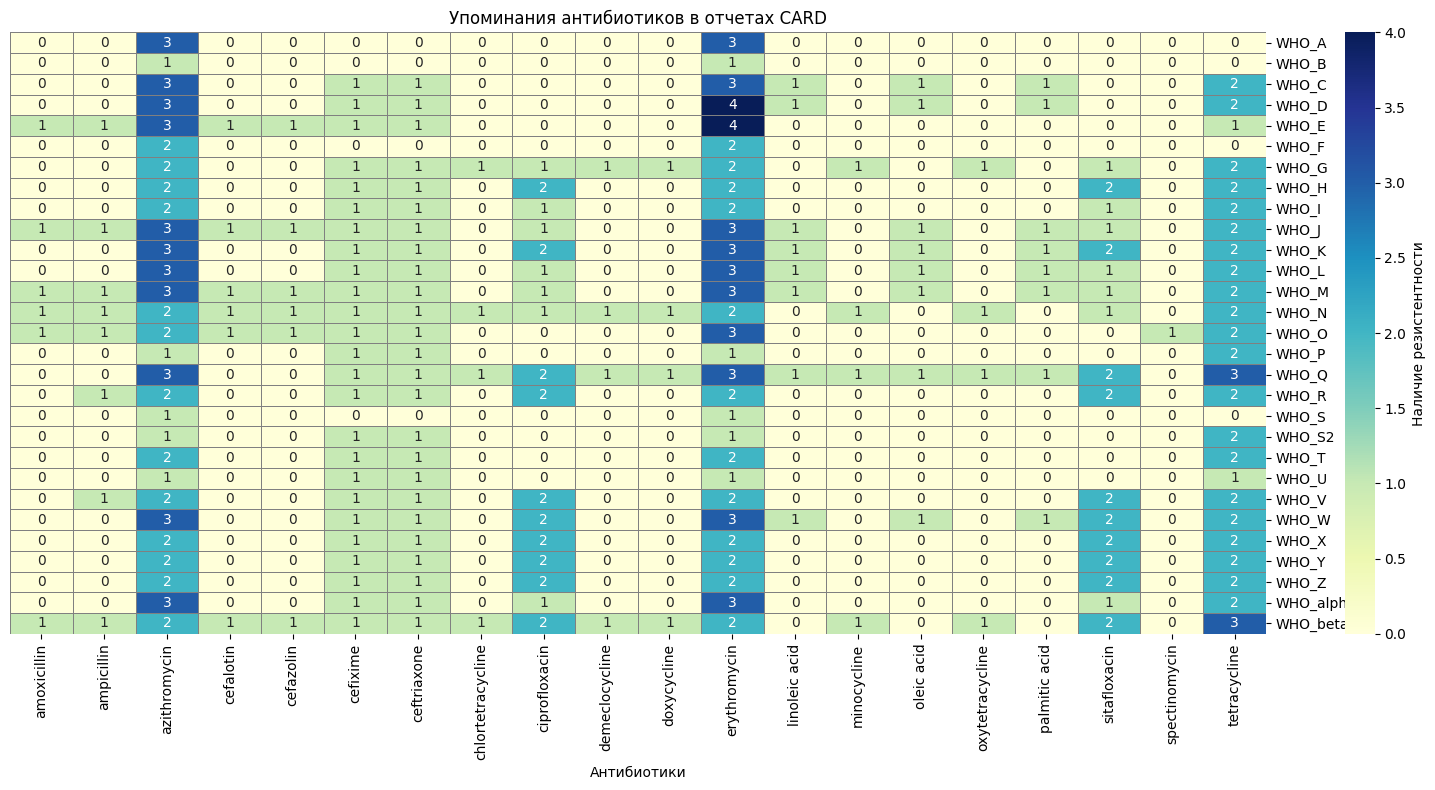


Рисунок 4 - Тепловая карта по количеству упоминаний антибиотиков в отчетах CARD

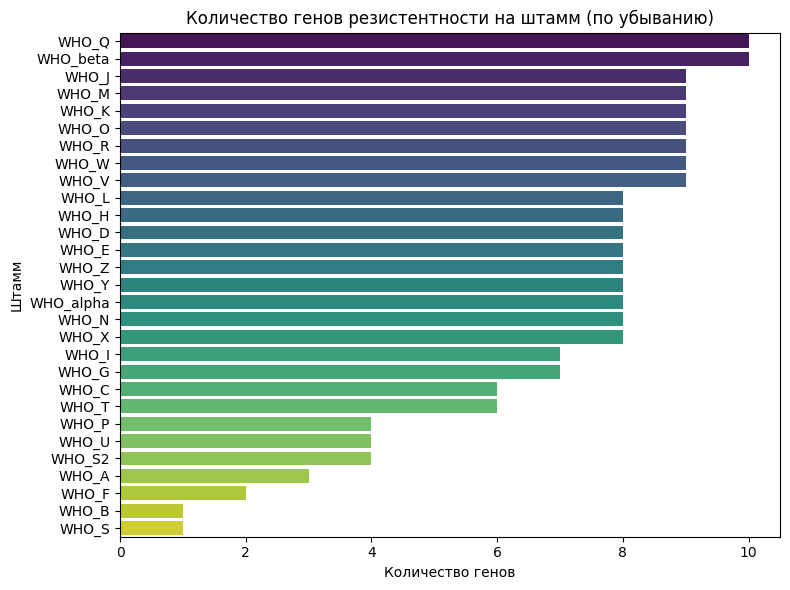


Рисунок 5 - Количество вариантов антибиотикорезистентности на штамм

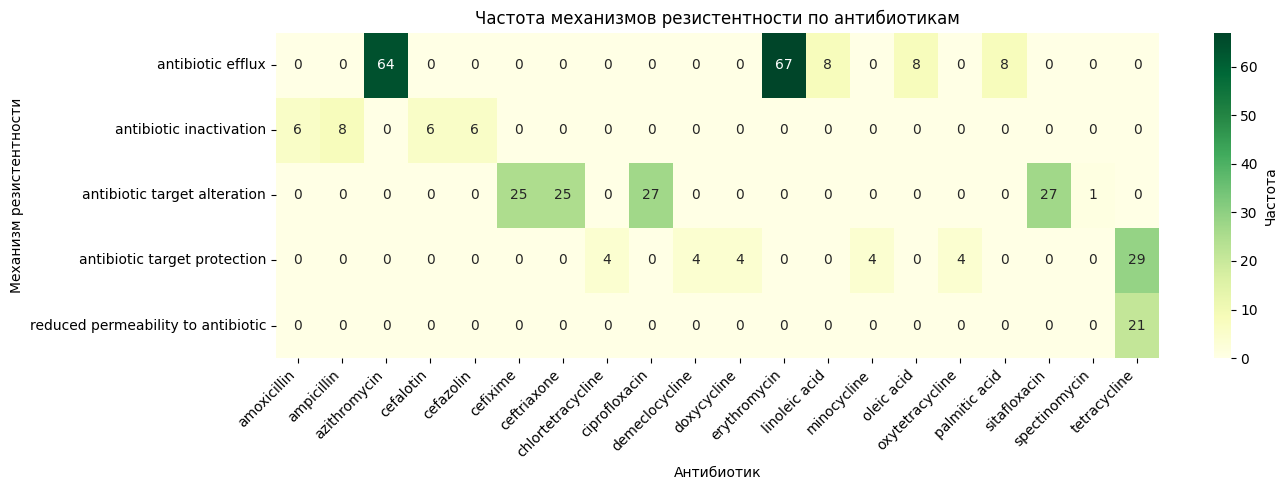


Рисунок 6 - Распределение механизмов резистентности по антибиотикам

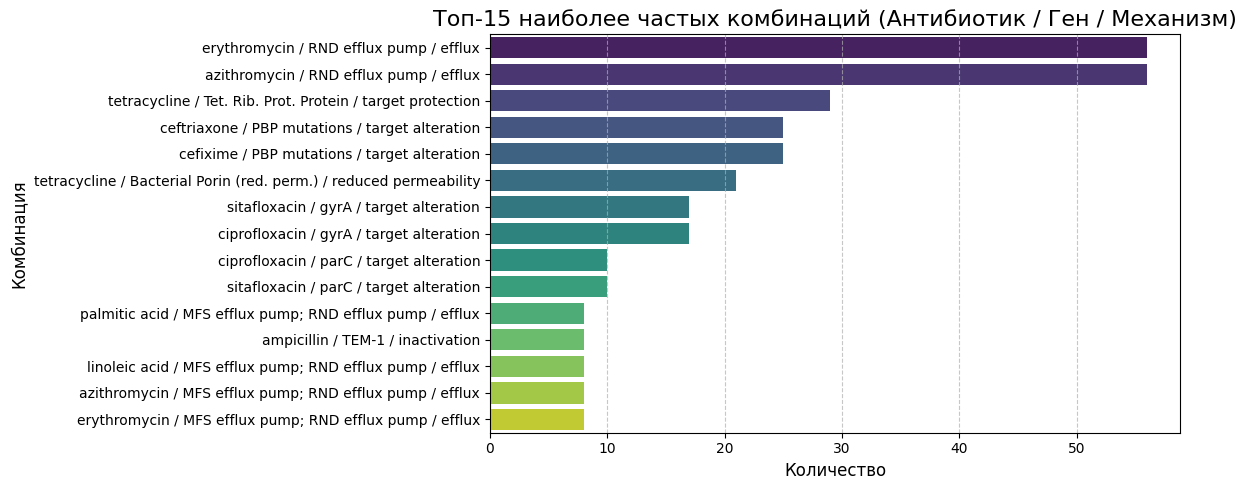


Рисунок 7 - Топ-15 комбинаций Антибиотик/Ген/Механизм