frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 29 January 2025
DOI 10.3389/fgene.2025.1460508

OPEN ACCESS

EDITED BY
Christopher Fields,
University of Illinois at Urbana-Champaign, United States

REVIEWED BY
Rodrigo Matheus Pereira,
Federal University of Grande Dourados, Brazil
Yolanda Moreno,
Universitat Politècnica de València, Spain

*CORRESPONDENCE
Rodolfo Brizola Toscan,
rodolfo.toscan@doctoral.uj.edu.pl

RECEIVED 15 August 2024
ACCEPTED 09 January 2025
PUBLISHED 29 January 2025

CITATION
Brizola Toscan R, Lesiński W, Stomma P, Subramanian B, Łabaj PP and Rudnicki WR (2025) Antimicrobial resistance in diverse urban microbiomes: uncovering patterns and predictive markers.
Front. Genet. 16:1460508.
doi: 10.3389/fgene.2025.1460508

Antimicrobial resistance in diverse urban microbiomes: uncovering patterns and predictive markers

Rodolfo Brizola Toscan[1]*, Wojciech Lesiński[2], Piotr Stomma[2], Balakrishnan Subramanian[3], Paweł P. Łabaj[1] and Witold R. Rudnicki[2,3]

[1]Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland, [2]Faculty of Computer Science, University of Białystok, Białystok, Poland, [3]Computational Center, University of Białystok, Białystok, Poland

Antimicrobial resistance (AMR) is a growing global health concern, driven by urbanization and anthropogenic activities. This study investigated AMR distribution and dynamics across microbiomes from six U.S. cities, focusing on resistomes, viromes, and mobile genetic elements (MGEs). Using metagenomic data from the CAMDA 2023 challenge, we applied tools such as AMR++, Bowtie, AMRFinderPlus, and RGI for resistome profiling, along with clustering, normalization, and machine learning techniques to identify predictive markers. AMR++ and Bowtie outperformed other tools in detecting diverse AMR markers, with binary normalization improving classification accuracy. MGEs were found to play a critical role in AMR dissemination, with 394 genes shared across all cities. Removal of MGE-associated AMR genes altered resistome profiles and reduced model performance. The findings reveal a heterogeneous AMR landscape in urban microbiomes, particularly in New York City, which showed the highest resistome diversity. These results underscore the importance of MGEs in AMR profiling and provide valuable insights for designing targeted strategies to address AMR in urban settings.

KEYWORDS

antimicrobial resistance (AMR), feature selection, data science, PCA, random forest, microbiome, resistome modelling

1 Introduction

Antimicrobial resistance (AMR) is a phenomenon that arises when bacteria, viruses, fungi, and parasites undergo genetic changes, rendering them insensitive to the effects of antimicrobial agents, thereby making infections more difficult to treat and increasing the risk of disease transmission, morbidity, and mortality (Sun et al., 2022). The emergence and spread of AMR are inherently driven by anthropogenic factors, and it is estimated that over one million people died due to AMR in 2019 (Murray, 2022). In conjunction with suboptimal wastewater treatment processes that fail to degrade residual antibiotic agents, inappropriate use of antibiotics has led to a persistent increase in the environmental abundance of antimicrobial compounds (Aden and Bashiru, 2022).

Antimicrobial resistance genes (ARGs) are transmitted either vertically through binary fission in bacteria or horizontally through horizontal gene transfer (HGT) mechanisms, including conjugation, transformation, and transduction (Ochman et al., 2000). Transformation involves bacterial uptake of genetic material from their surroundings,

TABLE 1 Overview of sample distribution by City for the study. The table lists each city with its corresponding ID and the number of samples collected.

ID	City	Sample number
BAL	Baltimore	14
DEN	Denver	45
MIN	Minneapolis	6
NYC	New York	46
SAC	Sacramento	16
SAN	San Antonio	16

while conjugation involves the direct exchange of genetic material between bacterial cells. Unlike these processes, transduction is mediated by viruses and mobile genetic elements (MGEs), highlighting the necessity of considering these dynamic entities when investigating antimicrobial resistance (Canchaya et al., 2003; Frost et al., 2005).

Resistome profiling, particularly in hotspots like wastewater treatment facilities, meat processing plants, hospitals, and urban areas, has gained significant attention (Karkman et al., 2018; Honda et al., 2023; Aarestrup, 2012; Mulvey and Simor, 2009; Vassallo et al., 2022). Human activities impact the resistome substantially (Vassallo et al., 2022). The global scientific community has thus intensified efforts to understand resistome dynamics (Cobo-Díaz et al., 2021). One notable initiative is MetaSUB, which periodically sequences metagenomic material from urban public spaces such as metro stations and bus stops (Ryon et al., 2022). Metagenomics is advantageous over culture-based methods because it allows the identification of viable but non-culturable bacteria (VNCB), which are often missed in traditional culturing approaches (Steen et al., 2020; Venter et al., 2004). Additionally, in contrast to molecular methods, metagenomics does not rely on prior knowledge of sequences, enabling comprehensive characterization of microbial communities and their functional potential (Quince et al., 2017; Mason et al., 2014).

This study extensively analyses 143 urban environmental metagenomic samples (see Table 1) and antibiotic susceptibility data from 145 hospital patients and their delivered isolates. The isolates referenced in this study are bacterial isolates collected from patients, cultured, and tested independently. These isolates are not patient-specific data but are derived from clinical samples. The dataset included AMRs commonly associated with clinically relevant pathogens, i.e. *Escherichia coli, Klebsiella pneumoniae*, and *Enterobacter hormaechei*. By comparing AMRs detected in the environmental metagenomes with those identified in the isolates, we assessed the ability of the tools to detect resistance markers of clinical significance.

Both datasets, i.e., the metagenomic fastq files and the isolates' resistome profiles, were provided by the CAMDA (2023) organization team. The metagenomic samples are an arbitrary subset of the MetaSUB sequencing pool, where initial analysis has led to drafting a global metagenomic map of urban microbiomes and antimicrobial resistance (The International MetaSUB Consortium, 2021). These samples were collected from six major U.S. cities (Danko et al., 2021). Our investigation encompasses their resistome, virome, and mobilome, employing a diverse array of techniques both independently and in conjunction (see Figure 1).

We employed mathematical modelling and statistical techniques to analyze the metagenomic data and predict the origins of the samples. We utilized random forest classifiers for feature selection and classification, leveraging algorithms like Boruta and Multi-Dimensional Feature Selector (MDFS) to identify key resistome markers. Additionally, we computed cosine similarities between samples based on their antimicrobial resistance (AMR) profiles and applied clustering algorithms to explore the data structure. Singular Value Decomposition (SVD) was used for dimensionality reduction to enhance the accuracy of similarity calculations (Berry et al., 1995). Furthermore, we investigated the association between MGEs and AMRs, conducting filtering experiments to assess the impact of MGE-associated AMRs on resistome profiles.

These mathematical and machine learning (ML) approaches allowed us to derive meaningful insights into the distribution and dynamics of AMR in urban microbiomes. By mapping resistome profiles across different urban environments and evaluating the precision and applicability of various resistome profiling tools, our study significantly contributed to the advancement of resistome analysis methodologies. Despite facing challenges, our findings underscore the critical role of MGEs in resistome studies and provide valuable insights for future research and public health strategies.

The integration of ML techniques has transformed environmental microbiology, enabling the analysis of large-scale datasets and uncovering patterns and relationships often missed by traditional methods. In this study, ML played a pivotal role in predicting resistome dynamics. Using random forest classifiers and derivative implementations, we identified key resistome markers and their associations with antimicrobial resistance genes (ARGs). These tools handle high-dimensional metagenomic data efficiently, facilitating robust feature selection, classification, and modeling of microbial interactions.

Additionally, ML has been instrumental in predicting sample origins, as demonstrated in global studies like MetaSUB (The International MetaSUB Consortium, 2021). By integrating resistome, virome, and mobilome datasets, ML provides a comprehensive view of microbial ecosystems (Medina et al., 2022; Bhattacharya et al., 2022). Its application in this work not only enhances predictive precision but also underscores ML's critical role in advancing environmental microbiology and shaping public health strategies.

2 Methods

2.1 Data preparation

The study utilized fastq files obtained from a publicly available repository of MetaSUB data (www.metasub.org) accessible through CAMDA (2023) page. A total of 143 libraries from six different cities in the United States were examined (Table 1). To ensure the dataset's quality, the MetaSUB-CAMP metagenomic tool suite (Tierney et al., 2023) was employed. This suite conducted quality control procedures, including the removal of host reads (Genome

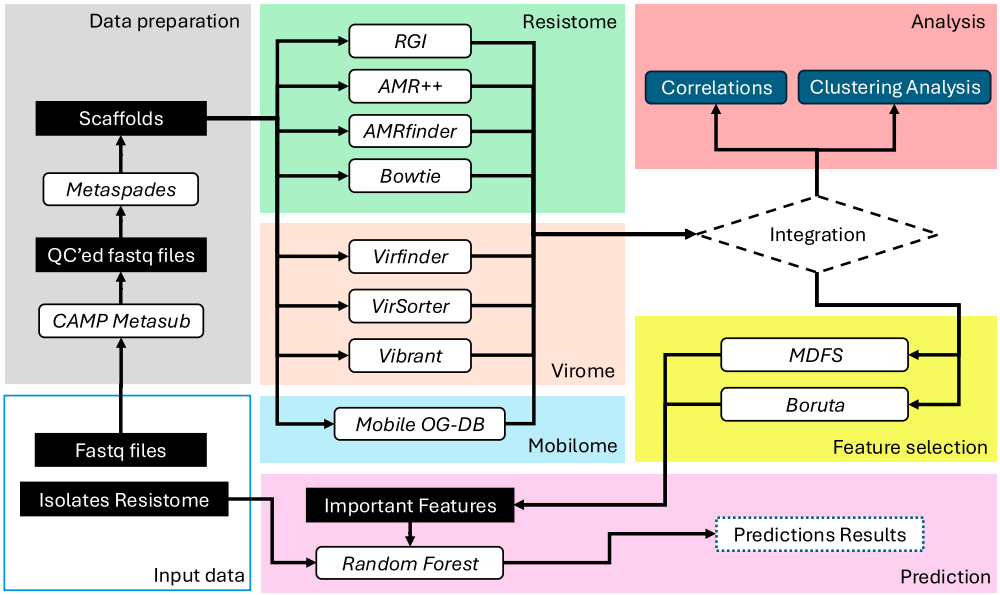

FIGURE 1
This diagram provides a high-level view of the methods workflow used in this study, illustrating the stages from data preparation through to prediction. The process includes quality control of sequencing data, various profiling methods for resistome, virome, and mobilome, followed by integration and analysis using feature selection and machine learning techniques.

Reference Consortium Human Build 38 - RefSeq assembly accession: GCF_000001405.26) and low-quality sequences. The quality-controlled reads were then subjected to *de novo* assembly using metaSPADES (Nurk et al., 2017) with standard parameters.

2.2 Data generation

2.2.1 Resistome profiling

We constructed resistome profiles using four different analytical methods. For short quality-controlled reads, we used AMR++ v3.0 (Microbial-Ecology-Group, 2023) and Bowtie v2.5.1 (Langmead and Salzberg, 2012). AMR++ makes use of MEGARes v3.0, a comprehensive AMR database with an acyclic hierarchical annotation structure (Bonin et al., 2022). For the Bowtie approach, we aligned the reads against a custom database with Bowtie2 using standard parameters. This database contained a comprehensive collection of indexed antimicrobial resistance genes, combining sequences from the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2022) with an additional set of manually curated genes, kindly provided by Dr. Nelly Selém (Mojica, 2023).

For elongated, assembled reads, we utilized two other methods: AMRFinderPlus v3.11.14 (Feldgarden et al., 2021) and Resistance Gene Identifier (RGI) V3.3.1 (CARD, 2024), developed by the National Center for Biotechnology Information (NCBI) and CARD team, respectively (Alcock et al., 2022). In executing AMR++, AMRFinderPlus, and RGI, we adhered to the standard pipelines without any modifications to the databases or pipeline parameters.

The resistome profile of isolates provided by the CAMDA (2023) consisted of a tabular file listing antimicrobial resistance genes (ARGs) associated with each isolate. This dataset served as a reference for comparative analysis during the study. The table is publicly available and can be freely downloaded from their official webpage or the dedicated repository for this study at https://github.com/rbtoscan/frontiers_camda_2023/blob/main/data/isolates/CAMDA2023_isolates.csv.

2.2.2 Resistome normalization

Considering the contrasting sequencing depth across the dataset, we employed a range of normalization techniques for gene counts.

These counts were normalized against the following parameters: 1) the quantity of quality-controlled base pairs, 2) the total number of assembled base pairs, 3) the detected small subunit ribosomal RNA (SSU) count and 4) the count of SSUs exhibiting a minimum of 50% coverage.

Our examination of assembled contigs for SSU identification involved the utilization of the bbduk.sh tool from the BBTools suite (Bushnell, 2021). We referenced the SILVA Small Subunit database (release 138.1) for this purpose Quast et al. (2013). The detection of SSU fragments was quantified at varying coverage thresholds, including any detectable level (above 0%) and a more stringent criterion of over 50% coverage.

2.2.3 MGE identification, annotation and classification

We used Mobile OG-DB (Brown et al., 2022) for the MGE identification. Mobile OG-DB serves as a meticulously curated database housing well-documented protein sequences of MGEs, encompassing diverse elements such as transposons, plasmids, integrons, and various other mobile entities.

The MGE identification process using Mobile OG-DB involves:

1. Identification of open reading frames (ORFs) using Prodigal (Hyatt et al., 2010) with default parameters as outlined in the MobileGo-DB documentation.

2. Creation of alignment summaries against the mobile orthologous groups database using Diamond (Buchfink et al., 2015) with default parameters as outlined in the MobileGo-DB documentation.
3. Compilation of element-mapping data, summarizing matches to proteins from various MGE classes.

2.2.4 MGE quantification and AMR association

We identified MGEs and quantified their frequency across the dataset by counting the number of samples in which each MGE gene was detected. Specifically, we mapped reads to the MGE database using DIAMOND, identified genes from the alignments, and calculated the presence of each MGE gene as the number of samples in which it was detected. Our analysis focused on MGE genes found in more than 50% of the samples from each city. To investigate the association between MGEs and antimicrobial resistance (AMR), we revisited the original genetic data, searching for AMR genes proximal to the MGE genes. MGEs and resistome markers co-located on the same contig were considered associated and were categorized as mobile antimicrobial resistance markers (mAMRs).

2.2.5 Virome profiling

Virome profiling was conducted using three distinct tools, each chosen for its specific utility in viral genome identification and analysis. VirSorter v2.2.4 (Guo et al., 2021), a widely used tool, employs prophage sequences to identify virus-like signatures in microbial datasets. VirFinder (Ren et al., 2017) v1.1 is known for its statistical learning approach, which assigns a likelihood score to sequences for their viral origin, enhancing detection specificity. Vibrant v1.2.1 (Kieft et al., 2020) utilizes machine learning and known viral databases to annotate and predict viral sequences with high accuracy. The outputs from VirSorter, VirFinder, and Vibrant were combined, and duplicate entries were removed to retain only unique viral sequences. Subsequently, the unique sequences underwent a quality control process using CheckV (Nayfach et al., 2021), which assesses the completeness and quality of the detected viral genomes, ensuring that the data used in further analyses are of high integrity. These tools were integrated into the Snakemake viral investigation pipeline (MetaSUB-CAMP, 2024), which was operated with standard parameters to ensure consistent and reproducible analysis across datasets.

2.2.6 K-mer profiling

K-mer profiling was conducted to assess the diversity and complexity of the metagenomic samples. We used Jellyfish (Marcais and Kingsford, 2011) to compute k-mer statistics, generating a 143 × 12 table, with one row per sample and four columns for each k-mer size (33, 55, and 77). The k-mer sizes of 33, 55, and 77 were chosen to balance sensitivity and specificity in sequence detection. Shorter k-mers (e.g., 33) detect a broader range of sequences and capture small genetic variations, while longer k-mers (e.g., 77) offer higher specificity and reduce false positives by ensuring unique matches. This multi-scale approach leverages the benefits of different k-mer lengths, as supported by previous genomic analysis studies (Chikhi and Medvedev, 2014; Li et al., 2010).

The metrics calculated included unique, distinct, total, and max_count for each k-mer size. The *unique* metric counts k-mers occurring exactly once, serving as a direct indicator of sample diversity. The *distinct* metric counts the number of k-mers while ignoring their multiplicity, representing the cardinality of the set of k-mers. The total metric sums the occurrences of all k-mers, reflecting the overall abundance and richness of the sample. The *max_count* metric identifies the highest occurrence of any single k-mer within a sample, indicating the presence of highly repetitive sequences or dominant species.

2.2.7 Data integration

All intermediate analysis results from the profiling of resistome, mobilome, virome, and k-mer counts were processed and wrangled for subsequent interpretation and analysis through clustering, modeling, and prediction techniques. The results generated by each tool was compiled into matrices and analysed using the R statistical programming language (R Core Team, 2024). The following R packages were employed to support data processing, visualization, feature selection, and modeling:

- Boruta: For feature selection using importance scores derived from random forest algorithms (Kursa and Rudnicki, 2010).
- MDFS: To apply multi-dimensional feature selection based on information theory (Piliszek et al., 2019).
- dplyr, tidyr, stringr: For data wrangling and preprocessing (Wickham et al., 2023a; Wickham et al., 2023c; Wickham, 2023).
- ggplot2 and reshape2: For data visualization and reshaping (Wickham, 2016; Wickham, 2007).
- kableExtra and knitr: To generate reproducible, dynamic tables and reports (Zhu, 2024; Xie, 2024).
- randomForest: For building random forest classification models (Liaw and Wiener, 2002).
- pROC: For ROC curve analysis (Robin et al., 2011).
- mltools and networkD3: To support machine learning tools and visualization of networks (Gorman, 2018; Allaire et al., 2017).
- Readr: For efficient data import (Wickham et al., 2023b).

All packages used are open-source and available through the Comprehensive R Archive Network (CRAN).

2.3 Analysis

2.3.1 Analysis of AMR-based city similarity

The initial goal was to perform an exploratory analysis of the clustering structure of the urban samples derived from their AMR profiles. To this end, clustering-based approaches were tested, based on $L2$ norm cosine similarity matrix (Horn and Johnson, 1985) computed using absence-presence tables of AMRs. It is expected that if AMR levels are related to the geographical location of the samples, then by using AMR-based similarity to cluster the samples, one could recover each sample's city assignments by inspecting the cluster labels. However, after initial tests, we discovered that the clustering structure cannot be easily mapped to the original city labels. Therefore, a more fundamental approach was used, that was

concerned with the AMR-based sample similarities themselves, not the clusters built upon them.

To further validate the AMR profiling results, the resistome profiles of clinical isolates were integrated into this analysis. These profiles served as a baseline for assessing the overlap between clinically significant AMRs and those detected in the environmental samples. By comparing the ARGs identified in isolates with those found in environmental datasets, we evaluated the ability of each profiling tool (AMR++, Bowtie, AMRFinderPlus, and RGI) to capture clinically relevant resistance markers. This integration also allowed us to explore potential patterns of co-occurrence between ARGs and MGEs, further elucidating their role in AMR dissemination across urban environments.

The cosine similarity between samples was computed for each AMR profiling approach: AMRFinderPlus, AMR++, RGI and Bowtie. Then, a statistical analysis of the relationship between the values of the similarities and city labels was performed. Several variants of the cosine-based similarities were computed and compared. Differences between the similarities inside vs. across cities were examined, and their statistical significance was assessed. Our protocol was aimed to answer the question of "Does the AMR-based sample similarity carry information about the geographical origin of the samples?." The general outline of the protocol is stated below:

1. For each tool, compute the sample similarities.
2. For each tool, based on sample similarities, compute summary statistics comparing the inside-city similarity of the samples with the between-city similarity.
3. Assess the statistical significance of the differences and compare the statistics across different AMR finding tools and similarity variants.

2.3.2 Cosine similarity calculation

For each tool, different variants of the cosine similarities were computed. Each variant tested starts with a raw absence/presence table. It is then used to compute plain cosine similarity. Optionally, a combination of the following transformations was applied to the data between those two elementary steps. Details of each step are discussed in the next parts of the manuscript:

- Input markers filtering: use all available markers or only those relevant for the decision variable "city".
- Usage of SVD embedding: either apply SVD embedding (Wall et al., 2003; Berry et al., 1995) on the absence/presence table or not.
- Sparsification of the cosine similarity matrix: either zero out weak connections or not.

Transformations were applied in the order they are listed. Effects of each combination of these transformations were tested.

2.3.3 Transformation details

For selecting the markers related to the city indicator, relevance was computed by a chi-squared test based on mutual information (MI) between the decision and the features (Mnich and Rudnicki, 2020). We have used a custom function mimicking the behaviour of the MDFS 1-D feature selector (Piliszek et al., 2019), generalized for handling non-binary decision variables, that is, the "city indicator".

For the SVD embedding, we have used the UD part of the SVD decomposition (Berry et al., 1995) applied to the marker matrix M, where $M = UDV$, where rows of M correspond to samples, and columns to the markers. Each row of UD matrix represents the joint information of the each sample contained in its AMR levels, while first k columns carry the most variation across different samples.

Sparsification step zeroes out weak connections, according to the weight threshold chosen by clique counting on the thresholded graph. A clique is a subset of vertices in a graph such that every two distinct vertices are connected by an edge.

We choose threshold for which the number of observed cliques of size at least 3 is maximal. Such threshold is dependent on the structure of the graph, thus it varies between variants. This is a heuristic we found empirically work well in various scenarios. More details can be found in the Supplementary Material.

2.3.4 Summary statistics of the sample similarities

Here we describe a simple statistic used to assess, on average, how well separated are samples coming from different cities. We have computed the similarity for each of $\frac{1}{2}N(N-1)$ pairs created from N samples.

Set of computed similarities can be partitioned into the set of "inside city" similarities S_{in} and "between city" similarities S_{btw}. Similarities between samples were computed using levels of AMRs. If those levels are overall relevant for the geographical location of the samples, we would expect, on average, for the "inside city" similarity to be bigger than the "between city" similarity. Therefore, we found it meaningful to compute the following summary statistic that summarizes each similarity matrix:

$$\overline{S_{in}} - \overline{S_{btw}} \tag{1}$$

where $\bar{A}$ denotes mean of elements in set A. The greater the value, the greater the similarities between samples from the same city than the similarities between samples from different cities.

2.3.5 Statistical significance assessment

To adjust for possible randomness of the differences between the computed summary statistics, we have utilized both common non-parametric approaches: resampling-based point estimate with uncertainty estimation, as well as permutation-based significance test where applicable (Efron and Tibshirani, 1993).

To estimate the standard error of the statistic, we have used a variation of leave-d-out jackknife (Shao and Wu, 1989). In standard leave-d-out jackknife, one computes the statistic for all (or random sample of all) possible subsets of samples of the original dataset (size N) that have $N - d$ elements, and uses the resulting replicates to compute the spread of original statistic. In our case, we used a stratified variant of such procedure because of a serious class imbalance in the variable "city." We want to leave out $\frac{1}{k}$ samples in each subset. Therefore, to ensure equal treatment of each class of the "city" variable, to compute each replicate, we leave $\frac{1}{k}$ out of each class.

We have also used a permutation test (Efron and Tibshirani, 1993) for the statistic $\hat{\theta} := \overline{S_{in}} - \overline{S_{btw}}$.

TABLE 2 Comparison of AMR detection methods. This table summarizes the total counts, mean values, standard deviations, and variability of antibiotic resistance detection across four methods used to AMRs in this study. Variability was calculated as standard deviation divided by Mean.

Method	Total	Mean	Standard deviation	Variability
AMR++	977	67.8	93.5	1.4
Bowtie	342	67.2	57.1	0.9
RGI	252	12.0	17.1	1.4
AMRFinder	142	9.4	11.4	1.2

More detailed information on specific parameters used in the procedures desribed above can be found in the Supplementary Material.

2.3.6 Feature selection and classification

We used two methods for feature selection: Boruta algorithm (Kursa and Rudnicki, 2010) and Multi Dimensional Feature Selection (MDFS) (Mnich and Rudnicki, 2020; Piliszek et al., 2019). The feature selection and random forest classification focused on predicting the origin of the samples, specifically the "city" indicator. Boruta uses the importance score from multiple runs of the RF algorithm to discover the informative variables. In each iteration, the original data set is extended by adding a randomized copy of each variable. MDFS is based on information theory and considers synergistic interactions between descriptive variables. The algorithm returns binary decisions about variables' relevance and ranking based on Information Gain and p-value.

For modelling, we used Random Forest algorithm Breiman (2001), which is based on decision trees and works well out of the box on most data sets Fernández-Delgado et al. (2014). It also obtains relatively well results with a small number of observations. The classifier was used in three ways: all *versus* all, one *versus* all, and one *versus* one. Models were evaluated by accuracy (multiclass cases) and by Matthews Correlation Coefficient (MCC) Matthews (1975) and area under the receiver operating curve (AUROC or AUC) for binary classification.

TABLE 3 Comparison of prediction results by AMR detection tools. This table presents the accuracy scores for AMRFinder, RGI, Bowtie, and AMR++ in one-versus-all and one-versus-one prediction scenarios across different cities.

	AMRFinder	RGI	Bowtie	AMR++
One versus all				
BAL versus all	0.69	0.71	0.94	0.93
DEN versus all	0.81	0.89	0.92	0.91
MIN versus all	0.64	0.89	0.89	0.97
NYC versus all	0.96	0.92	0.97	0.96
SAC versus all	0.60	0.78	0.94	0.99
SAN versus all	1.00	0.34	0.96	0.86
Average	0.78	0.75	0.94	0.94
One versus one				
DEN BAL	0.66	0.74	0.96	0.92
MIN BAL	0.96	0.96	0.82	0.99
MIN DEN	0.79	0.98	0.89	1.00
NYC BAL	0.69	0.92	0.97	0.92
NYC DEN	0.98	0.98	0.98	0.99
NYC MIN	1.00	0.98	0.93	0.96
SAC BAL	0.99	0.83	0.99	1.00
SAC DEN	0.98	0.93	0.94	0.99
SAC MIN	1.00	0.95	0.99	0.99
SAC NYC	0.91	0.88	0.99	0.99
SAN BAL	0.00	0.76	0.97	0.93
SAN DEN	0.00	1.00	0.86	0.88
SAN MIN	0.00	0.91	0.80	0.98
SAN NYC	0.00	0.92	0.99	0.99
SAN SAC	0.00	0.87	0.91	0.93
Average	0.70	0.89	0.92	0.94

3 Results

3.1 Data description and interpretation

3.1.1 On the resistome

Table 2 provides a summary of AMR detection counts using four different methods: AMR++, Bowtie, RGI, and AMRFinder. AMR++ identified the highest total count of unique genes at 977, with an average (mean) of 68 and a standard deviation of 94, indicating considerable variability in the data. Bowtie detected a total of 342 unique genes, with a mean of 67 and a standard deviation of 57 – indicating more moderate spread around its mean. RGI identified 252 unique genes, with a lower mean of 12 and a standard deviation of 17, reflecting less variability compared to AMR++ and Bowtie. AMRFinder, while detecting the fewest unique genes at 142, had a mean of 9 and a standard deviation of 12, indicating a relatively consistent detection rate.

Overall, AMR++ and Bowtie both show higher mean and standard deviation counts of detected genes when compared to RGI and AMRFinder. AMR++ and RGI both display similar level of variability, as computed by dividing standard deviation by mean values.

We performed a correlation analysis between the number of genes found per sample for each of the tools and the number sequenced nucleotides and assembled basepairs. None of the methods displayed a significant correlation (>55%), suggesting that the detection of AMR markers is largely independent of sample size.

Aiming to identify the most suitable tool and normalization method for achieving our primary goal of detecting the geographic origin of the isolates, we compared the AUC values obtained from four different tools: AMR++, Bowtie, AMRFinderPlus, and RGI (see Table 3) with their standard non-processed values, simply

TABLE 4 Comparative prediction accuracy of AMR++ Tool. This table presents the prediction accuracy results of the AMR++ tool across various normalization modes, including standard, binary, SSU total, SSU cov50, reads, and genome data. The data is shown for both one-versus-all and one-versus-one city comparison scenarios.

	Standard	Binary	SSU total	SSU cov50	Reads	Genome
One versus all						
DEN vs. all	0.92	0.93	0.93	0.94	0.93	0.94
NYC vs. all	0.92	0.97	0.95	0.96	0.95	0.94
SAC vs. all	0.91	0.95	0.93	0.94	0.95	0.94
BAL vs. all	0.86	0.86	0.91	0.89	0.89	0.89
MIN vs. all	0.93	0.98	0.96	0.99	0.96	0.96
SAN vs. all	0.86	0.81	0.91	0.85	0.85	0.85
Average	0.90	0.92	0.93	0.93	0.92	0.92
One versus one						
DEN - NYC	0.99	0.96	0.94	0.97	0.95	0.96
DEN - SAC	0.96	0.97	0.93	0.95	0.94	0.95
DEN - BAL	0.96	0.98	0.97	0.95	0.97	0.97
DEN - MIN	0.94	0.97	0.95	0.95	0.94	0.95
DEN - SAN	0.87	0.86	0.88	0.90	0.89	0.90
NYC - SAC	0.94	0.98	0.96	0.97	0.96	0.96
NYC - BAL	0.97	0.99	0.99	0.98	0.98	0.98
NYC - MIN	0.99	0.99	0.97	0.98	0.98	0.98
NYC - SAN	0.95	0.98	0.96	0.97	0.96	0.97
SAC - BAL	0.97	1.00	0.98	0.99	0.99	0.99
SAC - MIN	0.96	1.00	0.97	0.99	0.99	0.99
SAC - SAN	0.90	0.93	0.91	0.92	0.92	0.92
BAL - MIN	0.93	0.93	0.90	0.92	0.91	0.92
BAL - SAN	0.90	0.93	0.91	0.92	0.91	0.92
MIN - SAN	0.98	1.00	0.98	1.00	1.00	1.00
Average	0.94	0.96	0.94	0.96	0.95	0.95

normalized by sequencing depth. AUC was chosen as it provides a robust measure of classification performance, balancing sensitivity and specificity across all possible thresholds. Based on these comparisons, AMR++ was selected for downstream analysis due to its superior detection of diverse and abundant AMR markers, as well as its consistently higher AUC values across various classification tasks.

Further, to optimize data pre-processing, we compared the AUC values for six different normalization methods (see Table 4). Binary normalization was chosen as the most effective approach, as it simplifies the data representation by focusing on the presence or absence of resistance genes, thus reducing noise and improving classification performance. Other methods, including those based on sequencing depth and SSU counts, did not significantly enhance predictive accuracy.

The UpSet plot on Figure 2 displays the overlap of antimicrobial resistance genes (ARGs) shared among the six cities included in the study. The horizontal bars represent the total number of ARGs detected in each city, while the vertical bars denote the size of intersections between sets of cities. Notably, a majority of ARGs are unique to individual cities, with NYC exhibiting the largest set size. The accompanying pie chart highlights the distribution of ARG classes identified, with resistance to Beta-lactams being the most predominant (11.5%). ARG classes with less than 1.5% representation were excluded from the chart for clarity. This figure underscores the heterogeneity of ARG distributions across urban environments and the presence of metal-associated resistance, which has been shown to foster the development and spread of antimicrobial resistance by interacting with antibiotics, reducing their bioactivity, and promoting resistance gene selection under selective pressure (Sutradhar et al., 2023).

3.1.2 On the virome

A total of 70,839 viral contigs were detected; however, only 394 contigs (0.56%) were considered relevant by CheckV

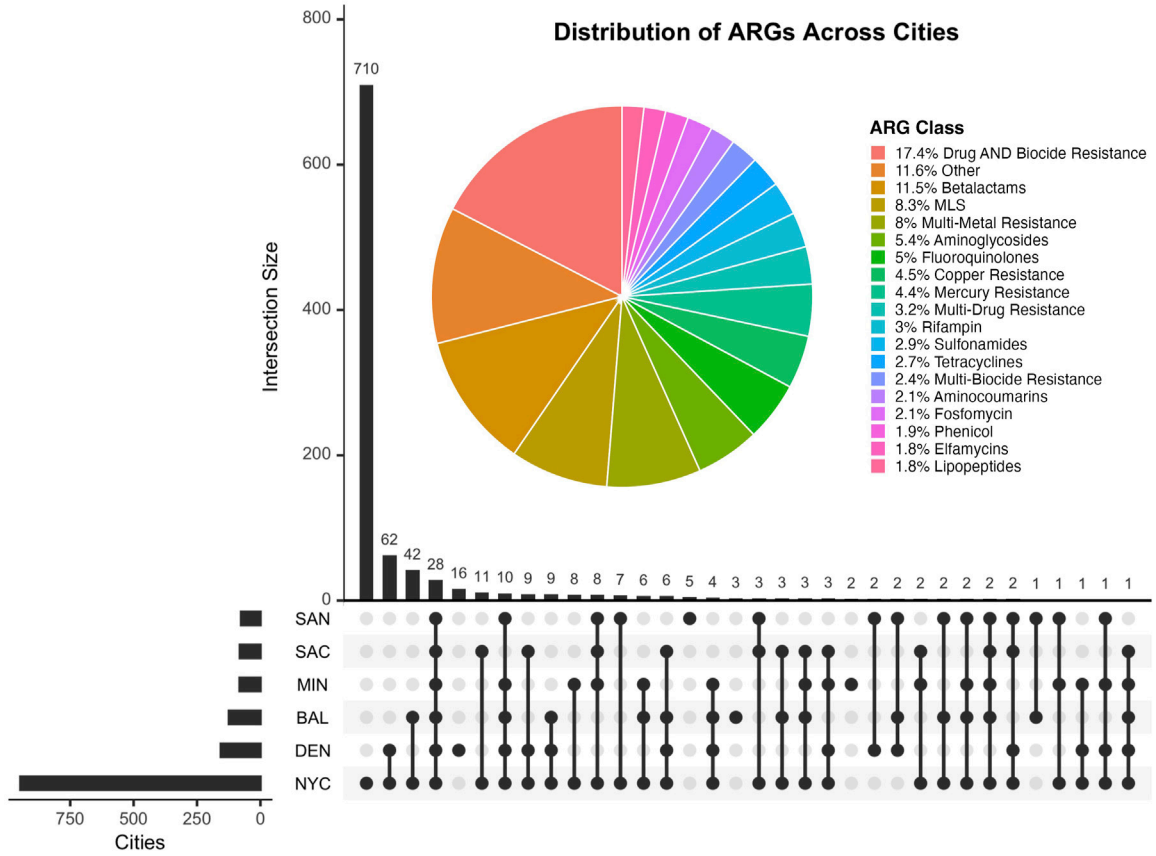

FIGURE 2
The UpSet plot illustrates the shared antimicrobial resistance genes (ARGs) between cities in this study. The accompanying pie chart displays the distribution of ARG classes, annotated using the MEGARes v3 database. The Other category comprises ARG classes that ocurr less than 1.5% and were thus removed from the plot.

TABLE 5 Distribution of MGE genes across cities. This table shows the number of samples, genes, and protein hits (homologs) for each city analyzed in the study.

City	Samples	Genes	Protein_hits (homologs)
DEN	45	924	8,279
BAL	14	910	8,184
SAC	16	867	7,248
NYC	46	1,577	32798
SAN	16	666	5,051
MIN	6	532	3,943

(Medium-quality, High-quality, and Complete). Upon investigation, it was observed that the number of viral contigs was affected by the small-sized contigs generated by the assembly. VirFinder was responsible for over 95% of the total number of viral contigs detected, with a high false positive rate. We inferred the high false positive rate based on the quality control performed by the CheckV tool. Putative viral contigs detected by VirFinder that lacked viral genes were tagged as non-viral and counted as falsely identified as viruses.

3.1.3 On the mobilome

Table 5 illustrates the distribution of MGE genes across all samples. We identified a total of 1660 MGE genes distributed across six US cities. Figure 3 highlights the overlap of these genes among the cities. Specifically, 394 MGE genes were shared across all six cities, suggesting a conserved core set of genes. In contrast, 382 genes were city-specific, appearing in only one city, indicating localized variation in MGE composition.

Among the MGE genes we identified, 438 MGE's genes co-occur with 325 AMR markers across the US cities, and a total of 12 of the MGE-AMR gene marker patterns were found across all cities (Table 6). A detailed table about the co-occurrence between the MGE-AMR across US cities with city and sample information can be found in the Supplementary Material. The Table 6 shows most common MGE gene and AMR gene that co-occured together among all 6 cities of the US.

3.2 Clustering

Figure 4 shows the point estimates of the similarity comparison statistic. Table 7 contains the FWER corrected p-values for the $S_{in} - S_{btw}$ statistic. Overall, the results suggest that based on AMR levels

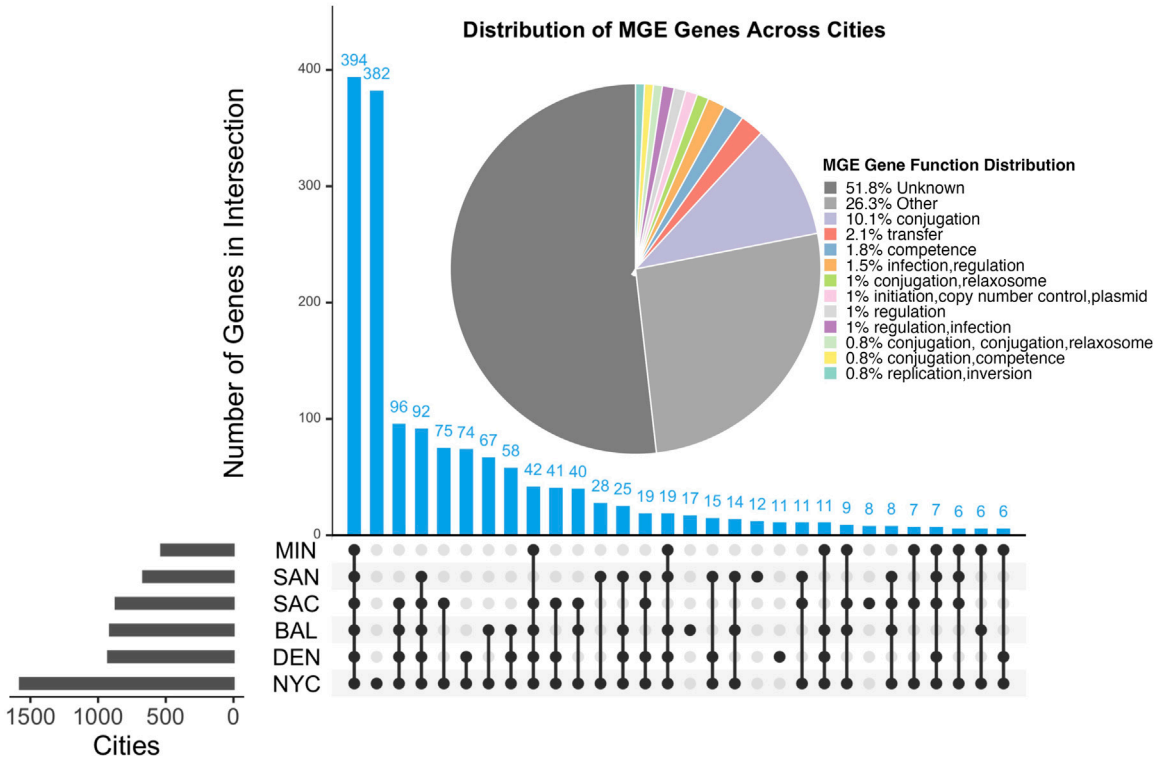

FIGURE 3
UpSet plot showing the overlap of MGE genes across six U.S. cities. Blue bars indicate the number of overlapping MGE genes between cities, while gray bars represent the total number of unique MGE genes identified in each city. Horizontal bars on the left show the total number of MGE genes identified per city. The pie chart displays the functional distribution of MGE genes based on the Mobile OG-DB. The Other category comprises MGE gene functions that ocurr less than 0.5% and were thus removed from the plot.

TABLE 6 Prevalence of MGE genes and antibiotic resistance markers (AMR) Across US Cities. This table lists the most common MGE and AMR genes, along with the number of cities, samples, and sequencing reads in which they were identified. Verification ensures that genes listed are distinct despite similar names.

MGE genes	AMR	Cities	Samples	Fastq reads
gyrA	GYRA	6	116	3,819
parC	PARC	6	66	1,271
parE	PARE	6	93	2,628
gyrB	GYRB	6	73	3,601
dnaK	DNAK	6	75	288
gyrB	PARY	6	33	75
tnpA_IS6100	VEB	6	9	19
tnpA	VEB	6	41	151
gyrB	GYRBA	6	69	1,559
thyA	THYA	6	20	44
ruvB	RUVBM	6	18	37
gyrB	GYRBZ	6	7	7

obtained from both AMR++ and Bowtie derived datasets we can find that in fact similarities of samples coming from the same cities are greater than similarities of samples coming from different cities. Comparing the point estimates for those tools (first row of Figure 4) shows that in terms of finding markers relevant for geographic location, AMR++ performs slightly better. Significant difference for the sparsified variants also suggests that indeed, the strongest similarities form along samples from the same cities, at least for data obtained with AMR++ and Bowtie methods.

AMRFinderPlus and RGI in that test performed worse, which is both visible in p-values of the permutation test as well as on the point estimate plots. For those tools, the difference between sparsified and normal similarity variant is within the bounds of standard error, which reinforces that observation. 'RGI' came up as the worst from all of the used tools. Limitation to the significant mutual information (MI) variables does not seem to bring much difference. In the end, AMR++ stands out as the best approach, without any significant difference between SVD based similarity and plain one.

We show the visualization of the similarities for the SVD embedded samples derived from the AMR++ dataset (Figure 5). We see that samples from different cities still have strong connections to each other, but blocks along the diagonal are noticeable. Overall, unsupervised analysis highlighted the need for including information external to the clustering procedure.

SAC samples (blue colors) form a noticable cluster, as seen on both sparsified and plain version of the heatmap plot. This corresponds to the high classification score by ML model for that particular city, as we show in the next section. We can also spot a well separated cluster in the middle that corresponds to portion of

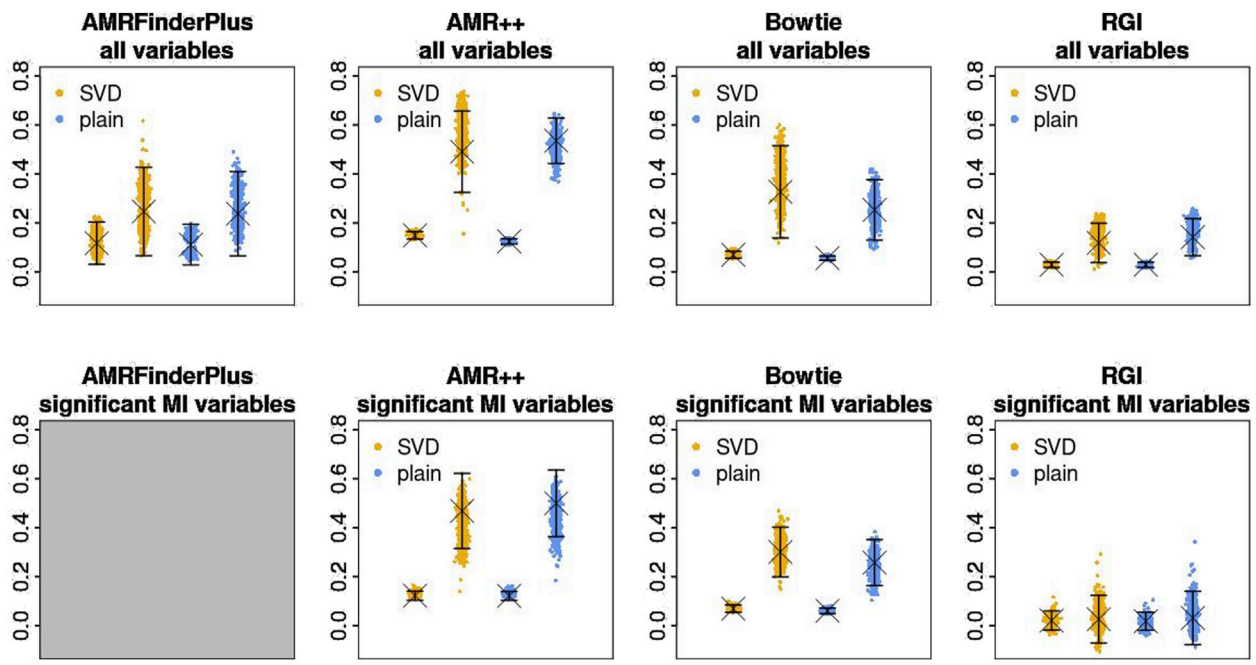

FIGURE 4
Estimates of $\overline{S_{in}} - \overline{S_{btw}}$, for datasets generated by different tools (columns) and similarity variants (colors). *Y*-axis of each plot is the value of the statistic. Each X position corresponds to each variant. Left of each color corresponds to statistic computed on whole similarity matrix. Right corresponds to statistic computed on sparsified matrix. "X" marks the value computed on real data, jitter plots show distribution of the jackknife replicates. Error bars signify standard errors. Top row shows results of calculation using all markers, bottom one – limited to markers significant to "city" variable by MI test.

TABLE 7 Holm corrected p-values of the permutation test on the $S_{in} - S_{btw}$ statistic. All similarity variants were calculated using all markers (without filtering relevant ones).

Similarity variant	AMRFinderPlus	AMR++	Bowtie	RGI
SVD, sparse	0.005	0	0	0.005
Plain, sparse	0.005	0	0	0.0028
SVD, non-sparse	0.0016	0	0	0.005
Plain, non-sparse	0.0036	0	0	0.005

samples coming from NYC (green colors). Baltimore group (yellow) contains several outliers visibly distinct from the rest. Both of these features turned out to be partly explained by k-mer content of each sample as which we show later.

3.3 K-mer profiling results

K-mer profiling revealed significant insights into the diversity and complexity of the metagenomic samples. As shown in Figure 6, each row and column corresponds to a sample. The line plots on top of the figure represent the number of unique k-mers at each X coordinate (sample). The yellow, orange, and red curves represent k-mer sizes 33, 55, and 77, respectively. The colors in the margins represent samples from different cities: DEN (red), NYC (green), SAC (blue), BAL (yellow), MIN (magenta), and SAN (black).

As observed, the peak of k-mer counts is particularly pronounced for samples from New York City (NYC). NYC samples are split into two distinct clusters, one with high complexity and one with low complexity. The high complexity cluster is exclusively composed of NYC samples, while the low complexity cluster includes samples from Sacramento (SAC) and potentially Baltimore (BAL). This clustering pattern suggests that elevated k-mer counts from NYC samples may be due to shared genetic elements or contamination, leading to unexpected clustering across different cities.

3.4 Modelling

The four tools and the six normalization approaches were tested. The multi-class models yielded poor results. The likely reason was the large number of classes with a small number of in-class samples. Consequently, we have focused on pairwise classification. Those were done in one-vs-all and one-vs-one manner and can be seen in Table 3. Notably, Bowtie and AMR++ displayed the best results,

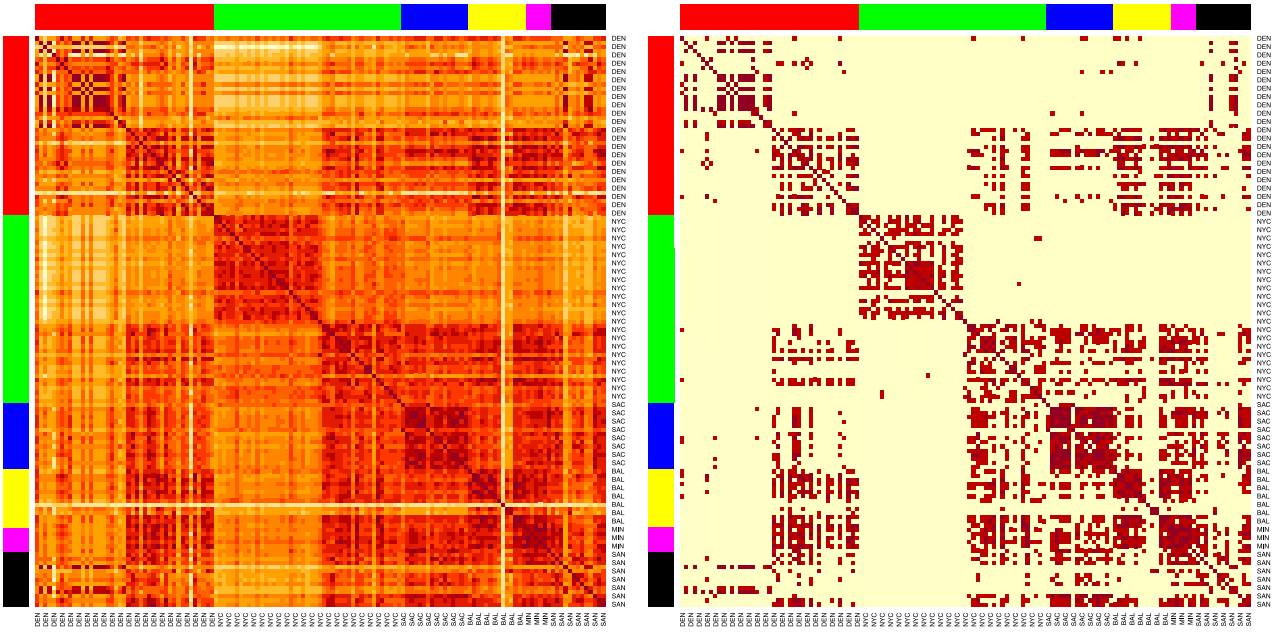

FIGURE 5
Heatmaps of the sample similarities for the AMR++ derived dataset, using the SVD embedding applied on significant MI markers. Right shows sparsified matrix, left – the plain one. Each row/column corresponds to each sample. Row/columns are arranged by the city labels, which are also marked by the side colors. Cities are arranged in following order (left to right): DEN, NYC, SAC, BAL, MIN, SAN.

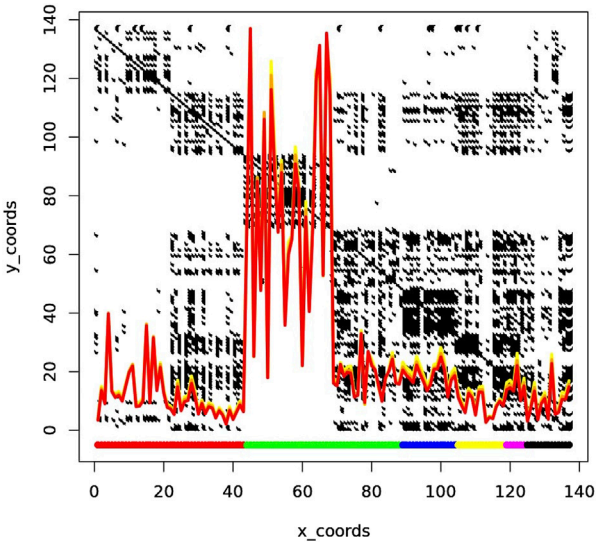

FIGURE 6
The line plots represent the number of unique k-mers at each X coordinate, with the yellow, orange, and red curves corresponding to k-mer sizes 33, 55, and 77, respectively. The margin colors indicate samples from different cities: DEN (red), NYC (green), SAC (blue), BAL (yellow), MIN (magenta), and SAN (black). The peak k-mer counts are notably higher for New York City (NYC) samples, which are divided into two clusters: one with high complexity containing only NYC samples, and one with low complexity, shared with samples from Sacramento (SAC) and potentially Baltimore (BAL).

indicating that reads-based resistome profiling is more accurate than assembled-based profiling. Bowtie and AMR++ consistently demonstrated the highest performance across the pairwise classifications. Specifically, in the one-vs-all evaluations, Bowtie and AMR++ achieved average AUC scores of 0.94. Similarly, in the one-vs-one comparisons, Bowtie and AMR++ maintained high performance, with average AUC scores of 0.92 and 0.94, respectively. This indicates that reads-based resistome profiling, utilized by Bowtie and AMR++, tends to be more accurate than assembly-based approaches, as evidenced by their superior AUC values. The results reveal a significant deviation from randomness (an AUC of 0.5), demonstrating the effectiveness of Bowtie and AMR++ in accurately profiling the resistome.

The resistome profile generated by AMR++ was used for further experiments. The reason was manyfold: 1) It had the highest and most diverse number of AMRs detected (Table 2), 2) the highest number of AMRs detected that matched the AMRs present in the isolates, 3) the highest inter-city dissimilarity, and 4) higher AUC values for the different modelling approaches. Table 4 displays the AUC values for different normalization approaches using AMR++ data. The values found on upper part of the table display how distinguishable is a given city from the rest of the dataset while the values on the lower part display how distinguishable is a city from another, pairwise. Each column represents a different pre-processing approach. The “Standard” column shows the values directly outputted by AMR++. The “binary” column represents the binary version of these values, where the data is converted into binary form, indicating the presence or absence of features. The best-preprocessing approach was the binarization. It has shown the highest values across all comparisons, followed by Standard ouput. Normalization by the number of SSU units (SSU total and SSU cov50) did not significantly improve modelling results. The same was observed for normalization based on sequencing and assembly nucleotides.

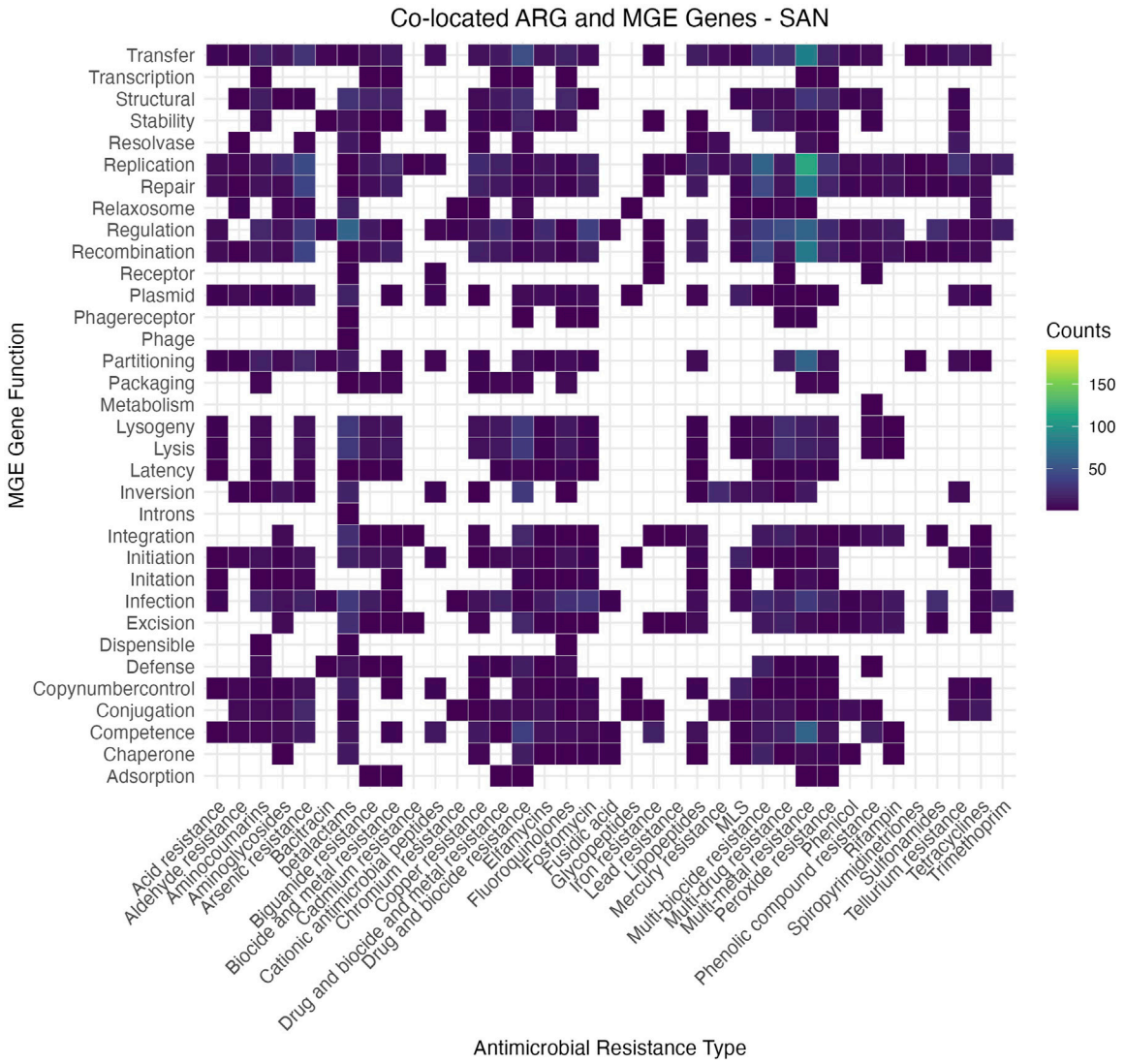

FIGURE 7
Heatmap illustrating the co-occurrence matrix between antimicrobial resistance (AMR) gene classes (*x*-axis) and MGE functions (*y*-axis) across the dataset. The intensity of the color indicates the frequency of association, with brighter colors representing stronger associations. This visualization highlights key AMR-MGE pairings, suggesting potential mechanisms of horizontal gene transfer and their role in antimicrobial resistance dissemination.

3.5 Linking AMRs to MGEs and viruses

Aiming to observe how MGEs and viruses influence AMR-based modelling, we excluded antimicrobial resistance genes (AMRs) associated with these elements and repeated the classification process. This involved removing virus-associated AMRs (vAMRs) and MGE-associated AMRs (mAMRs) through a two-step process. First, we used Bowtie2 mapping to identify reads linked to either viruses or MGEs. Then, we examined the remaining reads using the AMR++ tool. Out of nearly 70,000 viral contigs detected, only 4 were associated with resistance markers, all belonging to New York City. Therefore, we decided to utilize only mAMRs in the downstream analysis.

The data were processed to identify co-localization patterns between AMRs and MGEs across urban microbiomes. ARG subclasses, MGE functions, and city-specific counts were combined into a single dataset, with redundant entries removed and MGE descriptions standardized. ARG-MGE associations were filtered to retain only those appearing at least three times. Each MGE function was further split into its components for detailed mapping to ARG subclasses.

The heatmap (Figure 7) illustrates the co-occurrence patterns between ARG subclasses and MGE functions across the dataset. These visualizations reveal distinct trends, where functions such as replication, transfer, and regulation consistently co-occur with a broad range of ARGs, suggesting their pivotal role in resistance dissemination. Other MGE functions, such as lysogeny and excision, appear more specific to certain ARG subclasses, reflecting potential niche adaptation or targeted mechanisms of horizontal gene transfer. The overall patterns highlight the significant influence of MGEs on the dissemination and maintenance of AMR genes in urban microbiomes, underscoring the complexity of these interactions. The individual heatmaps for each city, illustrating city-specific

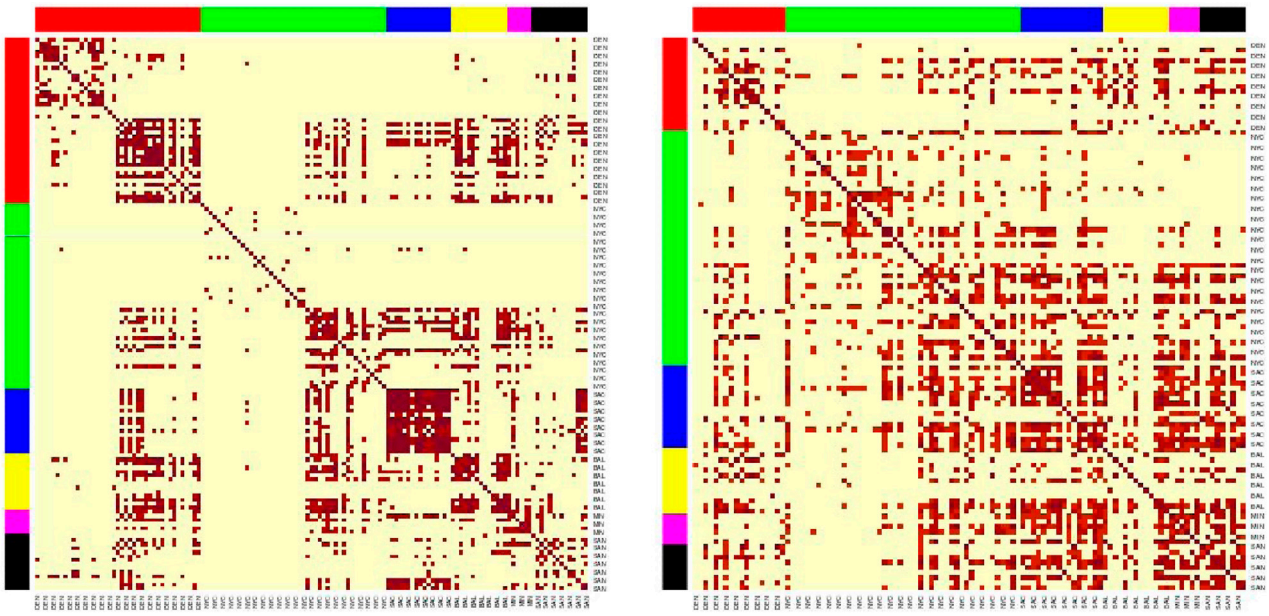

FIGURE 8
Heatmaps of the sample similarities for the AMR++ derived dataset, using the SVD embedding applied on significant MI markers. Left (50% MGE filtering) shows clear clusters, while right (100% MGE filtering) no longer shows clusters after removing mAMRs. Each row/column corresponds to each sample. Row/columns are arranged by the city labels, which are also marked by the side colors. Cities are arranged in following order (left to right): DEN, NYC, SAC, BAL, MIN, SAN.

AMR-MGE associations, are provided in the Supplementary Material of this submission for further reference.

3.6 Filtering mAMRs

We analyzed MGEs observed in more than one city, focusing on their proportions in each location. When selecting MGEs to filter, we considered two criteria: 1) MGEs must be observed in more than one city, and 2) each city must have a minimum presence of 30%, 50%, 80%, or 100%. Additionally, we selected MGEs for removal by randomly removing reads equal to the number removed during the 50% filtering criteria from each sample. These filtering criteria were used to identify potential candidates for mAMRs. The mAMRs were then filtered at the FASTQ read level by removing reads containing the selected MGEs. Subsequently, we have computed a new resistome profile using the AMR++ tool.

We utilized the newly generated antibiotic resistance matrix from the AMR++ detection to reassess AMR city similarity analysis and modeling (Figure 8). Among the four MGE filtering percentages, the 30% criterion resulted in minimal changes, while the 100% criterion led to a complete reshuffling of antibiotic resistance (AMR) clusters and deterioration in model performance (Table 8). The 50% criterion showed some improvements in the models, although the AMR clusters exhibited minimal variation, as our analysis was based on binarized information. Random removal of MGEs did not affect the results of AMR city similarity analysis and models.

These results suggest that certain resistome markers are inherently connected with MGEs. The removal of these markers significantly alters the resistome profile, impacting the efficiency of resistome modeling. Therefore, careful consideration must be given when filtering MGEs, as their presence or absence can drastically change the resistome profile and affect the accuracy of AMR detection and modeling. Our findings highlight the importance of preserving key MGEs in resistome studies to maintain the integrity and predictive power of resistome profiling models.

4 Closure

4.1 Conclusion and discussion

AMR exhibits a heterogeneous distribution across the dataset, with varying resistome profiles that do not correlate with sample depth. The investigated samples did not present nearly half of the ARGs presented in the isolates, indicating that either: 1) the sequencing depth of the urban samples was insufficient, 2) the isolated species were not dominant in the urban dataset, or 3) the classification methods were limited by incomplete reference databases.

Statistical analysis of the similarities suggested that the link between the origin of the sample and its AMR levels is non-random. In regards to how much of location-relevant markers each tool finds, we can see non-negligible variability between approaches, highlighting AMR++ and Bowtie as most informative in that aspect. Regardless of all of these general considerations, in practice we still see pairs of samples coming from different cities which are more strongly connected than pairs from the same city, which means that inferring their origin based on the clustering would not be effective. For some of the cities, such as DEN and SAC we can see that all of the samples are

TABLE 8 Prediction accuracy for AMR++ with MGE removal scenarios. This table compares the accuracy when mobile antimicrobial Resistance Markers (mAMRs) are removed under two conditions: "Removal MGE 50cov" where mAMRs in at least 50% present, and "Removal all MGE" where all mAMRs are excluded. Results are shown for one-versus-one and one-versus-all city comparisons to assess the impact of MGE on predictive accuracy.

	Removal MGE 50cov	Standard	Removall all MGE
One versus one			
DEN BAL	0.9247	0.9869	0.9231
MIN BAL	1.0000	0.9556	0.9487
MIN DEN	1.0000	0.9941	0.9259
NYC BAL	0.8947	0.9327	0.8979
NYC DEN	1.0000	1.0000	0.9221
NYC MIN	0.9211	0.9636	0.9316
SAC BAL	1.0000	1.0000	0.7835
SAC DEN	0.9924	0.9958	0.8750
SAC MIN	1.0000	0.9714	1.0000
SAC NYC	1.0000	1.0000	0.8329
SAN BAL	0.8951	0.9877	0.9402
SAN DEN	0.8615	1.0000	0.9568
SAN MIN	0.9692	0.9556	0.9815
SAN NYC	0.9798	1.0000	0.8913
SAN SAC	0.9256	1.0000	0.9583
Average	0.9577	0.9861	0.9181
One versus all			
BAL versus all	0.8709	0.9520	0.5887
DEN versus all	0.9359	0.9088	0.8129
MIN versus all	0.9571	0.9890	0.8660
NYC versus all	0.9611	0.9748	0.8683
SAC versus all	0.9831	0.9944	0.7378
SAN versus all	0.8957	0.9474	0.6891
Average	0.9340	0.9611	0.7605

tightly connected. However, for clusters cross city boundaries and inside some of the cities we can spot subgroups not related to each other at all (e.g., DEN and NYC). This fact seems to correspond to different counts of total k-mer content, especially evident by split within NYC group. Interestingly, removal of MGE related reads leads to disappearance of the clusters, which highlights their importance to the study of AMRs.

4.1.1 Resistome profiles

Beta-lactam resistance genes were the most abundant across all cities, comprising 11.5% of the total ARGs detected. New York City exhibited the highest diversity of ARGs, while cities like Minneapolis and San Antonio showed fewer unique ARGs, potentially reflecting differences in urban anthropogenic activities. The overlap of ARGs between environmental samples and clinical isolates highlights the clinical relevance of Beta-lactam resistance genes, which were predominantly associated with key pathogens. One of the main challenges of this study was modeling the resistome given that most antimicrobial resistance genes originated from New York City samples, resulting in a significant sample imbalance that affected downstream analyses.

4.1.2 Virome profiles

A total of 70,839 viral contigs were identified, with only 394 (0.56%) passing quality control thresholds. Most viral contigs were detected in New York City, possibly due to unique viral dynamics or contamination. Viral-AMR associations were rare, with only four vAMRs identified, all exclusively in New York City, suggesting the limited role of viruses in ARG dissemination across the studied urban environments. Surprisingly, such a small proportion of viral particles passed quality control. We hypothesize that this is due to the low sequencing depth of the samples, which were purposefully selected for the CAMDA challenge to emphasize analytical challenges.

4.1.3 Mobilome profiles

We identified 1,660 MGE genes, with 394 shared across all six cities and 382 being city-specific. Key MGE-AMR associations included gyrA and parC, found in all cities, which facilitate resistance to quinolones and highlight the pivotal role of MGEs in horizontal gene transfer. Functional analysis revealed replication and transfer as dominant MGE activities co-occurring with ARGs, indicating their significant contribution to AMR dissemination. City-specific variations in MGE profiles suggest localized selective pressures influencing resistome dynamics.

4.1.4 Mobilome-resistome interaction

Across all cities, MGEs associated with transfer, replication, and recombination functions exhibited the strongest connections to ARGs, particularly beta-lactam and multidrug resistance genes. These interactions emphasize the pivotal role of MGEs in horizontal gene transfer, facilitating the spread of clinically relevant resistance markers across diverse urban settings. A subset of these interactions was shared across all cities, indicating conserved mechanisms of resistance dissemination that transcend local environmental differences.

While specific patterns varied between cities, the overall trends reflect a common framework of resistance spread driven by MGEs. Detailed heatmaps for each city and combined analyses are provided in the Supplementary Material, offering insights into localized variations and broader patterns. These findings contribute to our understanding of how MGEs mediate ARG dissemination in urban environments, reinforcing the importance of addressing MGEs as a central component of antimicrobial resistance mitigation strategies.

4.1.5 Modelling

AMR++ has proven to be the best tool for resistome modeling. It relies on quality-controlled reads rather than assemblies, allowing it to detect the highest and most diverse number of AMR markers. Additionally, it achieves the highest AUC scores in random forest prediction of sampling sites. This is encouraging news for newcomers in the field, as the assembly process can be a significant bottleneck due to its high demand for computational resources.

Binarization appears to be the most effective data pre-processing approach because it simplifies the data, reduces noise and overfitting, and enhances the detection of critical signals by focusing on the presence or absence of resistance genes rather than their abundance. This method generally achieved higher AUC scores, suggesting improved classification performance and predictive power.

4.1.6 Isolates prediction

Classification proved to be challenging, with multiclass prediction resulting in many misclassifications. AMR++ was the only tool that produced noteworthy results, but they were still incorrect. Limitations of the database and the uneven number of samples for each city skewed predictions towards NYC. Efforts to clean the dataset by removing AMR markers associated with viruses or MGEs influenced the modeling but did not improve the predictions.

4.2 Limitations and future work

While this study has provided valuable insights into the distribution and diversity of AMRs in urban microbiomes, future studies could achieve more robust results by including a larger sample size, deeper sequencing, and a wider range of AMR identification tools and databases, along with additional metadata. These efforts could offer a more comprehensive understanding of the AMR landscape in urban environments and the effects of anthropogenic influence, thereby fostering the development of effective prevention strategies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, PRJNA732392; http://camda2023.bioinf.jku.at/contest_dataset, direct sftp download.

Author contributions

RB: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing–original draft, Writing–review and editing. WL: Formal Analysis, Writing–original draft, Writing–review and editing, Validation, Data curation. PS: Formal Analysis, Investigation, Writing–original draft, Writing–review and editing, Data curation, Visualization, Supervision. BS: Data curation, Formal Analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review and editing. PŁ: Funding acquisition, Project administration, Resources, Supervision, Writing–original draft, Writing–review and editing. WR: Project administration, Supervision, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. RB, PS, BS, and PL are fully or partialy funded through the NCN Sonata BIS grant number 2020/38/E/NZ2/00598.

Acknowledgments

We gratefully acknowledge Poland's high-performance Infrastructure PLGrid - ACK Cyfronet AGH for providing computer facilities and support within computational grant no PLG/2023/016299. Special thanks also go to the high-performance cluster at the Małopolska Centre of Biotechnology of Jagiellonian University, and to its dedicated system administrator, Wojciech Pilch, for his invaluable support in our computational endeavors.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2025.1460508/full#supplementary-material

References

Aarestrup, F. (2012). Sustainable farming: get pigs off antibiotics. *Nature* 486, 465–466. doi:10.1038/486465a

Aden, M., and Bashiru, G. (2022). How misuse of antimicrobial agents is exacerbating the challenges facing Somalia's public health 339. doi:10.1016/S0140-6736(21)02724-0

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A., Wlodarski, M. A., et al. (2022). CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* 51, D690–D699. doi:10.1093/nar/gkac920

Allaire, J., Gandrud, C., Russell, K., and Yetman, C. (2017). networkD3: D3 JavaScript network graphs from R. *R. package* (0.4). doi:10.32614/CRAN.package.networkD3

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Rev.* 37, 573–595. doi:10.1137/1037127

Bhattacharya, C., Tierney, B. T., Ryon, K. A., Bhattacharyya, M., Hastings, J. J. A., Basu, S., et al. (2022). Supervised machine learning enables geospatial microbial provenance. *Genes. (Basel)* 13, 1914. doi:10.3390/genes13101914

Bonin, N., Doster, E., Worley, H., Pinnell, L. J., Bravo, J. E., Ferm, P., et al. (2022). Megares and amr++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Res. gkac1047* 51, D744–D752. Epub ahead of print. doi:10.1093/nar/gkac1047

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Brown, C. L., Mullet, J., Hindi, F., Stoll, J. E., Gupta, S., Choi, M., et al. (2022). Mobileog-db: a manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements. *Appl. Environ. Microbiol.* 88, e0099122–22. doi:10.1128/aem.00991-22

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176

Bushnell, B. (2021). Bbmap: a fast, accurate, splice-aware aligner. Available at: https://sourceforge.net/projects/bbmap/(Accessed March 06, 2024).

CAMDA (2023). Camda 2023 challenge page. Available at: http://camda2023.bioinf.jku.at (Accessed March 06, 2024).

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L., and Brüssow, H. (2003). Prophage genomics. *Microbiol. Mol. Biol. Rev.* 67, 238–276. doi:10.1128/MMBR.67.2.238-276.2003

CARD (2024). Resistance gene identifier (rgi). Available at: https://github.com/arpcard/rgi.

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30, 31–37. doi:10.1093/bioinformatics/btt310

Cobo-Díaz, J. F., Alvarez-Molina, A., Alexa, E. A., Walsh, C. J., Mencía-Ares, O., Puente-Gómez, P., et al. (2021). Microbial colonization and resistome dynamics in food processing environments of a newly opened pork cutting industry during 1.5 years of activity. *Microbiome* 9, 204. doi:10.1186/s40168-021-01131-9

Danko, D., Bezdan, D., Afshin, E. E., Ahsanuddin, S., Bhattacharya, C., Butler, D. J., et al. (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance, , 184, 3376, 3393.e17. doi:10.1016/j.cell.2021.05.002

Efron, B., and Tibshirani, R. J. (1993). "An introduction to the bootstrap," in *No. 57 in monographs on statistics and applied probability*. Boca Raton, FL, USA: Chapman and Hall/CRC.

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., et al. (2021). Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance. *stress response, virulence* 183. doi:10.1038/s41598-021-91456-0

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res.* 15, 3133–3181. doi:10.5555/2627435.2697065

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* 3, 722–732. doi:10.1038/nrmicro1235

Gorman, B. (2018). Mltools: machine learning tools. *R. package* (0.3.5). doi:10.32614/CRAN.package.mltools

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., et al. (2021). Virsorter2: a multi-classifier, expert-guided approach to detect diverse dna and rna viruses. *Microbiome* 9, 37. doi:10.1186/s40168-020-00990-y

Honda, R., Matsuura, N., Sorn, S., Asakura, S., Morinaga, Y., Van Huy, T., et al. (2023). Transition of antimicrobial resistome in wastewater treatment plants: impact of process configuration, geographical location and season. *npj Clean. Water* 6, 46. doi:10.1038/s41545-023-00261-x

Horn, R. A., and Johnson, C. R. (1985). *Norms for vectors and matrices*. Cambridge University Press, 257–342. doi:10.1017/CBO9780511810817.007

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* 11, 119. doi:10.1186/1471-2105-11-119

Karkman, A., Do, T. T., Walsh, F., and Virta, M. P. J. (2018). Antibiotic-resistance genes in wastewater. *Trends Microbiol.* 26, 220–228. doi:10.1016/j.tim.2017.09.005

Kieft, K., Zhou, Z., and Anantharaman, K. (2020). Vibrant: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 8, 90. doi:10.1186/s40168-020-00867-0

Kursa, M. B., and Rudnicki, W. R. (2010). Feature selection with the boruta package. *J. Stat. Softw.* 36, 1–13. doi:10.18637/jss.v036.i11

Langmead, B., and Salzberg, S. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923

Li, R. e. a., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* 20, 265–272. doi:10.1101/gr.097261.109

Liaw, A., and Wiener, M. (2002). Classification and regression by randomforest. *R. News* 2, 18–22. doi:10.32614/CRAN.package.randomForest

Marcais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. (first published online January 7, 2011). doi:10.1093/bioinformatics/btr011

Mason, C. E., Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., et al. (2014). The metagenomics and metatranscriptomics of the human microbiome. *Genome Biol.* 15, 524. doi:10.1186/s13059-014-0524-1

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica Biophysica Acta* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Medina, R. H., Kutuzova, S., Nielsen, K. N., Johansen, J., Hansen, L. H., Nielsen, M., et al. (2022). Machine learning and deep learning applications in microbiome research. *ISME Commun.* 2, 98. doi:10.1038/s43705-022-00182-9

MetaSUB-CAMP (2024). CAMP virus-phage detect. Available at: https://github.com/MetaSUB-CAMP/camp_virus-phage-detect.

Microbial-Ecology-Group (2023). Amrplusplus: a bioinformatic pipeline for antimicrobial resistance analysis. Available at: https://github.com/Microbial-Ecology-Group/AMRplusplus.

Mnich, K., and Rudnicki, W. R. (2020). All-relevant feature selection using multidimensional filters with exhaustive search. *Inf. Sci.* 524, 277–297. doi:10.1016/j.ins.2020.03.024

Mojica, N. S. (2023). ccm-bioinfo/cambda2023

Mulvey, M. R., and Simor, A. E. (2009). Antimicrobial resistance in hospitals: how concerned should we be? *CMAJ Can. Med. Assoc. J. = J. de l'Association medicale Can.* 180, 408–415. doi:10.1503/cmaj.080239

Murray, C. J. L. e. a. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. , 399, 629, 655. doi:10.1016/S0140-6736(21)02724-0

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N. C. (2021). Checkv assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 39, 578–585. doi:10.1038/s41587-020-00774-7

Nurk, S., Meleshko, D., Korobeynikov, A., and Pa, P. (2017). Metaspades: a new versatile metagenomic assembler. *Genome Res.* 27, 824–834. doi:10.1101/gr.213959.116

Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304. doi:10.1038/35012500

Piliszek, R., Mnich, K., Migacz, S., Tabaszewski, P., Sułecki, A., Polewko-Klim, A., et al. (2019). MDFS: MultiDimensional feature selection in R. *R J.* 11, 198. doi:10.32614/RJ-2019-019

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucl. Acids Res.* 41, D590–D596. doi:10.1093/nar/gks1219

Quince, C., Trubl, G., Hyman, P., Roux, S., and Abedon, S. T. (2017). Metagenomics comes of age. *Nat. Rev. Genet.* 18, 315–329. doi:10.1038/nrg.2017.25

R Core Team (2024). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A., and Sun, F. (2017). Virfinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5, 69. doi:10.1186/s40168-017-0283-5

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinforma.* 12, 77. doi:10.1186/1471-2105-12-77

Ryon, K. A., Tierney, B. T., Frolova, A., Kahles, A., Desnues, C., Ouzounis, C., et al. (2022). A history of the metasub consortium: tracking urban microbes around the globe. *iScience* 25, 104993. doi:10.1016/j.isci.2022.104993

Shao, J., and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Ann. Statistics* 17, 1176–1197. doi:10.1214/aos/1176347263

Steen, A. D., Crits-Christoph, R. E., and Pernthaler, J. (2020). Viable but non-culturable bacteria: a persistent and enigmatic state. *Nat. Rev. Microbiol.* 18, 157–167. doi:10.1038/s41579-019-0292-0

Sun, G., Zhang, Q., Dong, Z., Dong, D., Fang, H., Wang, C., et al. (2022). Antibiotic resistant bacteria: a bibliometric review of literature. *Front. Public Health* 10, 1002015. doi:10.3389/fpubh.2022.1002015

Sutradhar, I., Kalyan, P., Chukwu, K., Abia, A. L. K., Mbanga, J., Essack, S., et al. (2023). Metal ions and their effects on antimicrobial resistance development in wastewater. *bioRxiv*. doi:10.1101/2023.06.16.545339

The International MetaSUB Consortium (2021). A global metagenomic map of urban microbiomes and antimicrobial resistance 183. doi:10.1016/j.cell.2021.05.002

Tierney, B., Mak, L., Toscan, R., Meleshko, D., Ronkowski, C., Henriksen, J., et al. (2023). Metasub-camp

Vassallo, A., Kett, S., Purchase, D., and Marvasi, M. (2022). The bacterial urban resistome: recent advances. *Antibiot. (Basel)* 11, 512. doi:10.3390/antibiotics11040512

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the sargasso sea. *Science* 304, 66–74. doi:10.1126/science.1093857

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). Singular value decomposition and principal component analysis

Wickham, H. (2007). Reshaping data with the reshape package. *J. Stat. Softw.* 21, 1–20. doi:10.18637/jss.v021.i12

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. New York: Springer-Verlag.

Wickham, H. (2023). Stringr: simple, consistent wrappers for common string operations. *R. package version* 1 (5.1). doi:10.32614/CRAN.package.stringr

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023a). Dplyr: a grammar of data manipulation. *R. package*. doi:10.32614/CRAN.package.reshape

Wickham, H., Hester, J., and Bryan, J. (2023b). Readr: read rectangular text data. *R. package version* 2 (1.4). doi:10.32614/CRAN.package.readr

Wickham, H., Vaughan, D., and Girlich, M. (2023c). Tidyr: tidy messy data. *R. package*. doi:10.32614/CRAN.package.tidyr

Xie, Y. (2024). Knitr: a general-purpose package for dynamic report generation in R. *R. package version* 1.48. doi:10.32614/CRAN.package.knitr

Zhu, H. (2024). kableExtra: construct complex table with kable and pipe syntax. *R. package version* 1 (4.0). doi:10.32614/CRAN.package.kableExtra