Antimicrobial Resistance in Diverse Urban Microbiomes: Uncovering Patterns and Predictive Markers

1st Rodolfo Brizola Toscan
Małopolska Centre of Biotechnology
Jagiellonian University
Cracow, Poland
https://orcid.org/0000-0002-5652-3666

2th Balakrishnan Subramanian
Institute of Computer Science
University of Białystok
Białystok, Poland
https://orcid.org/0000-0001-8403-167X

3nd Wojciech Lesiński
Institute of Computer Science
University of Białystok
Białystok, Poland
https://orcid.org/0000-0001-6628-9466

4rd Piotr Stomma
Institute of Computer Science
University of Białystok
Białystok, Poland
https://orcid.org/0000-0002-2603-8205

5th Paweł Łabaj
Małopolska Centre of Biotechnology
Jagiellonian University
Cracow, Poland
https://orcid.org/0000-0002-4994-0234

6th Witold R. Rudnicki
Institute of Computer Science
University of Białystok
Białystok, Poland
https://orcid.org/0000-0002-7928-4944

***Abstract*—This study presents a comprehensive analysis of 145 isolates from metagenomic samples across six major U.S. cities, focusing on their resistome, virome, and mobile genetic elements. Using a variety of techniques and mathematical models, we correlated metagenomic data with isolates to predict their origin and map resistome profiles in urban environments. This research, part of the CAMDA 2023 challenge and leveraging the MetaSUB sequencing initiative, evaluates the precision and applicability of resistome profiling tools, providing insights into antimicrobial resistance dynamics and advancing resistome analysis methodologies.**

I. 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 [1]. 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 [2]. Inappropriate use of antibiotics, in conjunction with suboptimal wastewater treatment processes that fail to degrade residual antibiotic agents, has led to a persistent increase in the environmental abundance of antimicrobial compounds [3].

Antimicrobial resistance genes (ARGs) are transmitted either vertically through mitosis or horizontally through horizontal gene transfer (HGT) mechanisms, including conjugation, transformation, and transduction. Transformation involves bacterial uptake of genetic material from their surroundings, while conjugation consists of the direct exchange of genetic material between bacterial cells. Unlike these processes, transduction is mediated by viruses and mobile genetic elements, highlighting the necessity of considering these dynamic entities when investigating in antimicrobial resistance.

Resistome profiling, particularly in hotspots like wastewater treatment facilities, meat processing plants, hospitals, and urban areas, has gained significant attention [4]–[8] as human activities are known to impact the resistome substantially [8]. The global scientific community has thus intensified efforts to understand resistome dynamics [9]. One notable initiative is MetaSUB, which periodically sequences metagenomic material from urban public spaces such as metro stations and bus stops [10]. Metagenomics, advantageous over culturing methods for its independence from known genes and primers, relies on comprehensive reference databases like CARD and Resfam for ARG identification [11], [12].

This study undertakes an extensive analysis of metagenomic samples and antibiotic susceptibility data from 145 isolates, all provided by the CAMDA 2023 challenge. These samples, a subset of the MetaSUB sequencing initiative, were collected from six major U.S. cities [13]. Our investigation encompasses their resistome, virome, and mobile genetic elements, employing a diverse array of techniques both independently and in conjunction. We developed mathematical models to correlate these metagenomic samples with the provided isolates, aiming to predict their origin. This approach allowed us to not only map the resistome profiles across different urban environments but also to evaluate the precision and applicability of various resistome profiling tools in our modeling efforts. Despite facing challenges, our study has yielded insights into the dynamics of antimicrobial resistance in urban settings and has contributed to the advancement of resistome analysis methodologies.

As part of the CAMDA 2023 challenge [14], this study conducts a comprehensive analysis of metagenomic samples from six major U.S. cities, with a particular focus on the qualitative resistome and antibiotic susceptibility profiles of

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

TABLE I: List of cities investigated in this study

145 isolates identified in these cities. Both datasets — the metagenomic fastq files and the isolates' resistome profiles — were provided by the CAMDA 2023 challenge organization team. Notably, the metagenomic samples are an arbitraty subset of the MetaSUB sequencing pool. Our investigation covers the resistome, virome, and mobile genetic elements of these samples, utilizing a variety of techniques both independently and in combination. We developed mathematical models to correlate the metagenomic data with the isolates, aiming to identify their city of origin. The methodology implemented has enabled us not only to delineate the resistome profiles within diverse urban environments but also to assess the precision and applicability of various resistome profiling tools. Despite encountering challenges, our study has provided valuable insights into the dynamics of antimicrobial resistance in urban contexts and has made significant contributions to the advancement of resistome analysis methodologies

***Index Terms*—Microbiomes, PCA, MCV, MDFS, random forest, clustering, antimicrobial resistance, resistome, metagenomics**

II. Methods

A. Data Preparation

The study utilised fastq files obtained from a publicly available database accessible through [15]. A total of 143 libraries from 6 different cities in the United States of America were examined (Table I). To ensure the dataset's quality, the MetaSUB-CAMP metagenomic tool suite [16] was employed. This suite conducted quality control procedures, including the removal of the host (Genome 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 [17] with standard parameters.

B. Resistome profiling

We constructed resistome profiles using four different analytical methods. For short quality-controlled reads, we have used AMR++ v3.0 [18] and Bowtie. AMR++ makes use of MEGARes v3.0, a comprehensive AMR database with an acyclic hierarchical annotation structure [19]. For the Bowtie approach, we aligned the reads against a custom database using Bowtie2 v2.5.1 [20]. This database was a comprehensive collection of indexed antimicrobial resistance genes, combining sequences from the Comprehensive Antibiotic Resistance Database (CARD) with an additional set of manually curated genes, kindly provided by Dr. Nelly Selém Mojica [21].

For elongated, assembled reads, we utilized two other methods: AMRFinderPlus v3.11.14 [22] and Resistance Gene Identifier (RGI) V3.3.1 [23], developed by and using the CARD initiative [11]. In executing AMR++, AMRFinderPlus, and RGI, we adhered to the standard pipelines without any modifications to the databases or pipeline parameters.

C. Resistome normalization

In this study, we employed various normalization techniques for gene counts. These counts were normalized against several parameters: the quantity of quality-controlled base pairs, the total number of assembled base pairs, the detected SSU count, and the count of SSUs exhibiting a minimum of 50% coverage.

Our examination of assembled contigs for Small Subunit (SSU) identification involved the utilization of bbduk.sh tool from the BBTools suite [24]. We referenced the SILVA Small Subunit database (release 138.1) for this purpose [**?**], [25]. 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.

D. Mobile profiling

Mobile genetic elements serve as the primary conveyors for the dissemination of antibiotic resistance genes. Thus we explored to map these elements to see if they could help us in determining AMR patterns among the cities, especially MGE-AMR co-occurance patterns.

Here is the general outline for the MGE identification and annotation:

1) Preprocessing of Metagenomic Data: The preprocessing step, highlighted in the initial methods section, guarantees the cleanliness of metagenomic data by removing noise and artifacts. Subsequent to the quality check and contig assembly, the preprocessed data is utilized for MGE identification in the pipeline.
2) MGE identification: The identification of MGEs involves aligning the assembled contigs from the previous step with the Mobile orthologous groups database (mobileGo-DB) [26], utilizing advanced algorithms to pinpoint sequence homologies with established MGEs. This step serves as the primary filter, aiding in the refinement of potential MGE candidates.
3) Functional Annotation and Classification: The identified sequences undergo functional annotation and classification to understand the nature and potential impact of the detected MGEs. This step helps in unraveling the functional roles of these elements within the microbial community.
4) Quantification and AMR association: The pipeline integrates quantification tools to estimate the abundance of MGEs, providing insights into their prevalence across different samples. The AMR information around the genomic regions of MGE are explored to study their co-occurnace patterns.

1) MGE identification, Annotation and Classification : We used MobileGo-DB for the MGE identification. MobileGo-DB serves as a meticulously curated database housing well-documented protein sequences of Mobile Genetic Elements (MGEs), encompassing diverse elements such as transposons, plasmids, integrons, and various other mobile entities. MGE identification using mobileGo-DB comprises of three steps:

1) Identification of open reading frames using prodigal [27].
2) Creation of alignment summaries to a mobile orthologous groups database file utilizing diamond [28].
3) Compilation of element-mapping data, providing a summary of matches to proteins from various mobile element classes.

In the initial step, we identify open reading frames within the assembled contigs sourced from US cities. This identification process utilizes the prodigal tool, as outlined in the study conducted by Hyatt et al. in 2010. Progressing to the second stage, we generate summaries illustrating the alignments between the genetic material and a specialized mobile orthologous groups database. The efficiency of this alignment is ensured through the use of the diamond tool, as elucidated in the research by Buchfink et al. in 2015. Finally, in the third phase, we process the results, offering a summary of matches discovered in the samples to proteins from different classes of mobile genetic elements. This compilation of element-mapping data provides valuable insights into the presence and distribution of mobile genetic elements.

2) Quantification and AMR association : We found mobile genetic elements (MGEs) and used the identified MGE gene hits to figure out how often these genes show up in our samples. We focused on MGE genes that were present in more than half of the samples in each city. To explore their connections with antimicrobial resistance (AMR), we traced back our findings to the raw genetic data and looked for AMRs in the areas around the MGE genes. This helped us create a table showing how often MGE genes and AMRs appear together in our data.

E. Analysis of AMR-based city similarity

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 [29] computed using absence-presence tables of AMRs. One could hope that if AMR levels are indeed related to the geographical location of the samples, then by using AMR-based similarity to cluster the samples, one could recover the city assignments of each sample by inspection of the cluster labels. However, after initial tests, we have found that discovered 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.

For each of the tools: AMRFinderPlus, AMR++, RGI and Bowtie, cosine similarity between samples was computed. 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, with their statistical significance assessed. Our protocol was aimed to answer a question of ”Does the AMR-based sample similarity carry information about the geographical origin of the samples?”. 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 different similarity variants.

1) Methods: cosine similarity calculation: For each tool, different variants of the cosine similarities were computed. Basically, in each variant tested, one starts with a raw absence/presence table. Then one uses it to compute plain cosine similarity. In between those two elementary steps, one optionally applies transformations to the data. Effects of each combination of transformations listed below were tested. Transformations were applied in the order they are listed. Details of each step are discussed in the next parts of the manuscript:

- **Input markers filtering**: either use all markers available or only those relevant for decision variable ‘city‘.
- **Usage of SVD embedding**: either apply SVD embedding [30], [31] on the absence/presence table or don’t.
- **Sparsification of the cosine similarity matrix**: Either zero out weak connections or don’t.

2) Methods: transformation details: For the selection of the markers related to the city indicator, relevance was computed by chi-squared test based on mutual information (MI) between decision and the features [32]. We have used custom function mimicking the behavior of the MDFS 1-D feature selector [33] (generalized for handling non-binary decision variable, that is, the ’city indicator’).

For the SVD embedding, we have used the UD part of the SVD decomposition [31] 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.

After computing the cosine similarity, in part of the variants, weaker connections were zeroed out using threshold chosen according to the criterion of maximum number of cliques of size at least 3 found on the thresholded similarity matrix. This is a heuristic for choosing the sparsification threshold which we found empirically to work well in various scenarios, which is dependent on the topology characteristics of the network induced by each specific similarity matrix. Basically, for a dataset of N samples, we get a $N \times N$ similarity matrix S

of non-negative entries, where S_{ij} encodes the degree of the similarity between objects i, j. Taking S as a weight matrix of the weighted graph, we find it's partition into heavily connected disjoint cliques by an efficient greedy algorithm and count them. We count such cliques on a thresholded S^{thr} version of the similarity matrix S. Finally, we pick the final sparsification threshold thr as the one which maximizes the number of cliques of size at least 3, from all of the tested candidate thresholds.

3) Methods: summary statistics of the sample similarities: For N samples, we get $\frac{1}{2}N(N-1)$ pairs. For each pair, we have computed their similarity. Therefore, we can partition the set of computed similarities in 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 following summary statistic, that summarizes each similarity matrix:

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

where $\overline{A}$ denotes mean of elements in set A. The bigger such value, the bigger the similarities between samples coming from the same city are from the similarities between the samples coming from different cities.

4) Methods: 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) [34].

For assessing the standard error [34] of the statistic, we have used a variation of leave-d-out jackknife [35]. In standard leave-d-out jackknife, one computes the statistic for all (or random sample of all) possible subsets of samples of original dataset (size N) that have $N-d$ elements, and use the resulting replicates to compute the spread of original statistic. In our case, because of serious class imbalance for the variable 'city', we have used a stratified variant of such procedure. We state that 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 (so in total we approximately leave out $\frac{1}{k}$ samples, still).

We have also used a permutation test [34] for the statistic $\widehat{\theta} := \overline{S_{in}} - \overline{S_{btw}}$. Null distribution in a permutation test is a distribution over all possible rearrangements of city labels. Exact p-value of a statistic $\widehat{\theta_{real}}$ calculated on a real dataset can then be calculated as (for one-sided test) [34]:

$$\frac{\#\text{rearrangements with } \widehat{\theta_{rearrangement}} \geq \widehat{\theta_{real}}}{\#\text{all rearrangements}} \tag{2}$$

If specific value of $\widehat{\theta_{real}}$ is not special, then any other rearrangement would be likely to produce a statistic of the size like $\widehat{\theta_{real}}$ or bigger (in the case when big value of $\widehat{\theta_{real}}$ is meaningful as in the case with the AMR based inside vs between city similarity difference). Then the number of rearrangements with a value of the statistic exceeding the $\widehat{\theta_{real}}$ would be big.

If the $\widehat{\theta_{real}}$ estimate we have from the data is non-random, such that the difference between AMR based similarities truly comes from geographical differences, then number of such rearrangements would be small. Then such a p-value is also small.

5) Computational details: For selecting relevant variables in the 'data limitation before cosine calculation' step, we have used a threshold of significance of 0.05 with a Benjamini–Yekutieli correction.

For SVD embedding dimension, we have arbitrarily chosen the $\frac{1}{3}$ of the maximum possible dimension for each dataset. It has to be noted here that cosine similarity between two vectors v, w is meaningful only if both of those vectors have norm > 0. Therefore, before computing the cosine similarity itself, one has to filter out samples which have zero norm. In some of the variants, transformations can induce the zero norm on some of the original samples themselves. For that reason, number of used samples for each dataset and similarity variant can differ slightly. In case of the tool 'AMRFinderPlus' and limitation of variables to the subset significantly related to the 'city' variable, the number of non-zero norm samples was reduced to 52 (vs 92 originally available). Since these variants turned out to be difficult to compare with the rest, we do not report statistics for them.

For the jackknife standard error estimation, to compute each replicate we have left out 0.25 of the original dataset (in a stratified manner described before). Choice of such fraction follows both from the theory [35] (such that number of left out samples, d, is contained in $(\sqrt{(N)}, N)$) and class distribution of the 'city' variable (smallest size of the smallest class across all variants is 4). Number of replicates for the jackknife estimation was equal to 500.

Lastly, for the permutation test, we have used 5000 rearrangements and we have limited ourselves to the variants where we use all of the available markers for cosine calculation. Reason for the latter is that when we permute the labels of the 'city' variable, it is very unlikely to find any marker relevant for it. For that rearrangements one cannot compute the similarity and this makes the comparisons between variants difficult to execute fairly.

F. Feature selection and classification

We used two methods for feature selection: Boruta algorithm [36] and Multi Dimensional Feature Selector (MDFS) [32], [33]. 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 [37], which is based on decision trees and works well *out of the box* on most data sets [38]. 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) [39] and area under the receiver operating curve (AUROC or AUC) for binary classification.

III. Results

A. Data description and interpretation

TABLE II: Total number of Antimicrobial Resistance detected for each of the cities.

City	Total	Mean	Std. Dev.
BAL	2721	48.59	63.92
DEN	3345	18.58	33.22
MIN	794	33.08	27.08
NYC	32640	177.39	294.88
SAC	2242	35.03	45.80
SAN	1265	19.77	30.15

1) On the resistome: AMRFinderPlus detected a total count of 1863 resistome markers, with an average (mean) of around 13.03 and a standard deviation of about 38.15, indicating some variability in the data. RGI's mean is higher at 47.45 with a greater standard deviation, suggesting more spread in its results. Bowtie and AMR++ have means of approximately 123.27 and 117, respectively, with Bowtie showing a very tight spread around its mean (sd of 123.13) compared to AMR++'s much wider variability (sd of 310.80).

TABLE III: Summary AMR detection counts

Method	Total	Mean	Standard Deviation
AMRFinderPlus	1863	13.0	38.1
RGI	6785	47.4	121.6
Bowtie	17628	123.2	123.1
AMR++	16731	117.0	310.8

The correlation analysis indicates that 'AMRFinderPlus', 'RGI', and 'AMR++' methods are highly consistent with each other in detecting viral elements, while 'Bowtie' differs significantly. None of the methods show a strong correlation with the number of base pairs in a sample, suggesting that the detection of viral elements is largely independent of sample size.

TABLE IV: Correlation calculation for number of AMRs per sample per method

	AMRFinderPlus	RGI	AMR++	AMR++	bases
AMRFinderPlus	1.00	0.96	0.35	0.99	0.40
RGI	0.96	1.00	0.35	0.97	0.39
AMR++	0.35	0.35	1.00	0.33	0.81
AMR++	0.99	0.97	0.33	1.00	0.37
bases	0.40	0.39	0.81	0.37	1.00

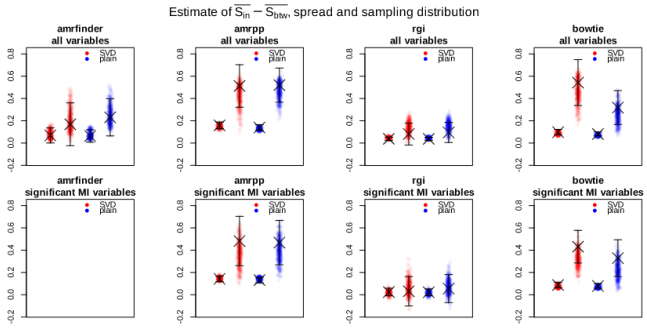

Fig. 1: Estimates of $\overline{S_{in}} - \overline{S_{btw}}$, for datasets generated by different tools (columns) and similarity variants (colors). Y axis of each plot is the value of the statistic. Each X position corresponds to each variant. Left of each color corresponds to statistic computed on 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.

similarity variant	AMRFinderPlus	AMR++	RGI	Bowtie
all markers,SVD	0.025	0	0.0014	0
all markers,plain	0.025	0	0	0
sparse,all markers,SVD	0.025	0	0.025	0
sparse,all markers,plain	0.0084	0	0.025	0

TABLE V: Holm corrected p-values of the permutation test on the $S_{in} - S_{btw}$ statistic

B. Clustering

Figure 1 shows the point estimates of the similarity comparison statistic.

Table V contains the FWER corrected p-values for the $S_{in} - S_{btw}$ statistic. Overall, results suggest that based on AMR levels obtained from both AMR++ and Bowtie derived datasets we can find that in fact similarities of samples coming from same cities are bigger than similarities of samples coming from different cities. Comparing the point estimates for those tools (first row of fig. 1) shows that in terms of finding markers relevant for geographic location, both tools perform on a similar level. Significant difference for the sparsified variants also suggests that indeed, the strongest similarities form along samples from the same cities.

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 contained 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 MI variables does not seem

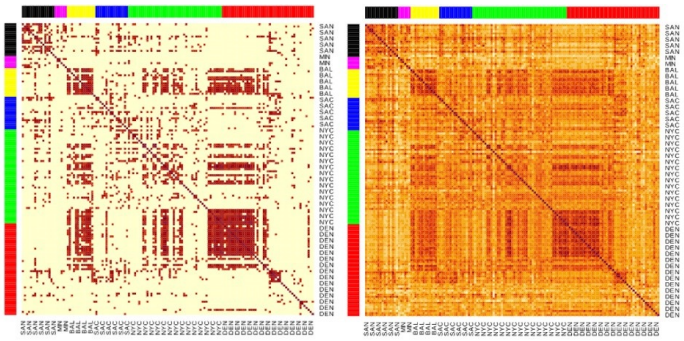

Fig. 2: Heatmaps of the sample similarities for the bowtie derived dataset, using the SVD embedding applied on all of the markers. Left shows sparsified matrix, right – the plain one. Each row/column corresponds to each sample. Row/columns are arranged by the city labels, which are also marked by the side colors. Cities are arranged in that specific order: SAN, MIN, BAL, SAC, NYC, DEN

to bring much difference. In the end, Bowtie in conjunction with the SVD transform stands out as the best approach (but corresponding AMR++ score is really close).

We show the visualization of the similarities for the SVD embedded samples derived from the bowtie dataset (fig. 2). We see that samples from different cities still have strong connections to each other, but blocks along the diagonal are noticeable. Interestingly, we can see that the Baltimore samples do tend to cluster with themselves (yellow colors), although there are some strong connections noticeable to Denver and NYC samples.

C. Classification

Four tools (AMR++, Bowtie, RGI and AMRFinderPlus) and six normalization methods were tested. Multi-class models generate poor results. The likely reason was that the large number of classes with a small number of in-class samples. Therefore, the next step was pair classification. Classifications were done in one-versus-all and one-versus-one manner.

Best prediction results were obtained by models built on data converted by AMR++ (see Table VI). In most cases, we achieved an AUC greater than 0.9. Similar results were obtained for the bowtie tool VII. Other methods were less efficient.

D. MGE AMR co-occurrence

Table VIII shows the distribution of MGE genes across all samples. We identified a total of 1660 MGE genes across 6 US cities, among which 438 MGE's genes co-occur with 325 AMR's across the US cities, and totally 12 of the MGE-AMR gene marker patterns were found across all cities (Table IX). Detailed table about the co-occurrence between the MGE-AMR across US cities with city and sample information can be found in the supplementary materials.

	binary	cov50	count std	Genome	reads	std.dev
one versus all						
DEN vs all	0.92	0.90	0.89	0.90	0.91	0.90
NYC vs all	0.95	0.92	0.92	0.93	0.93	0.93
SAC vs all	0.99	0.92	0.92	0.93	0.92	0.91
BAL vs all	0.91	0.91	0.89	0.88	0.90	0.89
MIN vs all	0.92	0.91	0.92	0.95	0.93	0.93
SAN vs all	0.87	0.83	0.78	0.85	0.84	0.86
one versus one						
DEN - NYC	0.99	0.96	0.96	0.97	0.96	0.97
DEN - SAC	0.99	0.91	0.92	0.93	0.92	0.92
DEN - BAL	0.92	0.91	0.89	0.87	0.90	0.89
DEN - MIN	1	0.99	0.90	0.99	0.90	0.92
DEN - SAN	0.93	0.71	0.71	0.73	0.73	0.76
NYC - SAC	0.99	0.97	0.98	0.98	0.98	0.97
NYC - BAL	0.94	0.95	0.95	0.95	0.96	0.93
NYC - MIN	0.93	0.92	0.96	0.93	0.94	0.91
NYC - SAN	0.99	0.96	0.97	0.96	0.97	0.96
SAC - BAL	1	0.99	0.99	0.99	0.99	0.99
SAC - MIN	1	0.96	0.96	0.97	0.97	0.96
SAC - SAN	0.96	0.90	0.90	0.90	0.92	0.89
BAL - MIN	0.93	0.91	0.90	0.92	0.92	0.89
BAL - SAN	0.92	0.94	0.93	0.94	0.93	0.91
MIN - SAN	0.84	0.76	0.82	0.83	0.82	0.80

TABLE VI: Prediction results for data converted by amrpp tool in various normalization modes.

	binary	cov50	count std	Genome	reads	std.dev
one versus all						
DEN vs all	0.9	0.85	0.85	0.88	0.85	0.9
NYC vs all	0.95	0.91	0.93	0.93	0.94	0.95
SAC vs all	0.93	0.63	0.6	0.64	0.69	0.93
BAL vs all	0.8	0.74	0.76	0.65	0.56	0.81
MIN vs all	0.74	0.54	0.72	0.73	0.77	0.77
SAN vs all	0.86	0.54	0.51	0.51	0.57	0.87
one versus one						
DEN - NYC	0.98	0.94	0.93	0.95	0.93	0.98
DEN - SAC	0.93	0.8	0.8	0.85	0.82	0.92
DEN - BAL	0.88	0.8	0.79	0.62	0.63	0.89
DEN - MIN	0.81	0.81	0.84	0.84	0.83	0.78
DEN - SAN	0.81	0.42	0.58	0.55	0.62	0.81
NYC - SAC	0.98	0.81	0.78	0.83	0.85	0.98
NYC - BAL	0.86	0.91	0.91	0.92	0.9	0.85
NYC - MIN	0.92	0.78	0.78	0.77	0.8	0.93
NYC - SAN	0.98	0.9	0.89	0.93	0.91	0.98
SAC - BAL	0.92	0.64	0.71	0.75	0.6	0.94
SAC - MIN	0.9	0.47	0.46	0.62	0.56	0.85
SAC - SAN	0.93	0.48	0.49	0.58	0.55	0.9
BAL - MIN	0.73	0.62	0.6	0.57	0.62	0.58
BAL - SAN	0.93	0.76	0.67	0.63	0.92	0.93
MIN - SAN	0.88	0.47	0.49	0.69	0.62	0.88

TABLE VII: Prediction results for data converted by bowtie tool in various normalization modes.

City	samples	genes	protein_hits (gene_copies)
DEN	45	924	8279
BAL	14	910	8184
SAC	16	867	7248
NYC	46	1577	32798
SAN	16	666	5051
MIN	6	532	3943

TABLE VIII: City wise distribution of MGE genes

MGE	AMR	Cities	Samples	Fastq reads
gyrA	GYRA	6	116	3819
parC	PARC	6	66	1271
parE	PARE	6	93	2628
gyrB	GYRB	6	73	3601
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	1559
thyA	THYA	6	20	44
ruvB	RUVBM	6	18	37
gyrB	GYRBZ	6	7	7

TABLE IX: Most common MGE and AMR markers found in all US city samples.

IV. Closure

A. Conclusion and Discussion

AR exhibits a heterogeneous distribution across the dataset, with varying resistome profiles that do not correlate with sample depth. Investigated samples did not present nearly half of the AMRs 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.

All of the used clustering methods are adaptive, do not assume prior number of clusters and use different strategies to form clusters. High similarity between partitions (ARI>0.6) from different approaches suggests that discovered clustering structure is non-arbitrary. Unfortunately, it does not seem to be connected with the city distribution, which revealed that a more targeted supervised classification should be used.

Classification proved to be a difficult task. Multiclass prediction generated many misclassifications. AMR-based features may require further refinement through additional filtering. The mobilome related analysis carried on the data showed cetrain corelations between mobilome and AMR. So if further feature selection and filtering can be applied using mobilome and virome information in the data, this might help us further in identifying distiunct AMR features for each geographical region.

In conclusion, our study provides a comprehensive analysis of metagenomic libraries from urban environments, focusing on resistome profiling through various tools and employing mathematical models, notably random forest, to ascertain the most effective tool for determining the city origin of each sample. Additionally the exploration of mobilome in this analysis has extended the scope of the study to further look at other genetic elements that can influence the AMRs across cities. We have successfully identified distinct patterns in antimicrobial resistance and explored the presence of viruses and mobile genetic elements, laying the groundwork for future studies. This work is part of a broad, systematic approach to investigate the microbial dynamics in urban settings. Having settled the best methods for resistome profiling, we aim, moving forward, to delve deeper into the synergy and influence of viruses and mobile genetic elements on antimicrobial resistance patterns. There are still questions that remain unanswered and orbit around the impact of mobile and virome elements in resistome modelling. Does it have any impact and can we quantify it? Can viruses and MGEs be used as resistome indicators? Future work will focused towards answering those questions, offering new insights into managing and mitigating antimicrobial resistance.

V. Acknowledgements

Thanks to the SONATA BIS 10 Funding Grant, provided by the National Science Centre - Poland. We also gratefully acknowledge Poland's high-performance Infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH, PCSS, CI TASK, WCSS) for providing computer facilities and support.

References

[1] G. Sun, Q. Zhang, Z. Dong, D. Dong, H. Fang, C. Wang, Y. Dong, J. Wu, X. Tan, P. Zhu, and Y. Wan, "Antibiotic resistant bacteria: A bibliometric review of literature," *Frontiers in Public Health*, vol. 10, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2022.1002015

[2] A. R. Collaborators, "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," vol. 339, 2022. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02724-0/fulltext%20

[3] M. Aden and G. Bashiru, "How misuse of antimicrobial agents is exacerbating the challenges facing somalia's public health," vol. 339, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9480883/

[4] A. Karkman, T. T. Do, F. Walsh, and M. P. J. Virta, "Antibiotic-resistance genes in wastewater," *Trends Microbiol*, vol. 26, no. 3, pp. 220–228, 2018.

[5] R. Honda, N. Matsuura, S. Sorn *et al.*, "Transition of antimicrobial resistome in wastewater treatment plants: Impact of process configuration, geographical location and season," *npj Clean Water*, vol. 6, p. 46, 2023. [Online]. Available: https://doi.org/10.1038/s41545-023-00261-x

[6] F. Aarestrup, "Get pigs off antibiotics," *Nature*, vol. 486, no. 7404, pp. 465–466, 2012. [Online]. Available: https://doi.org/10.1038/486465a

[7] M. R. Mulvey and A. E. Simor, "Antimicrobial resistance in hospitals: how concerned should we be?" *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*, vol. 180, no. 4, pp. 408–15, 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2638041/

[8] A. Vassallo, S. Kett, D. Purchase, and M. Marvasi, "The bacterial urban resistome: Recent advances," *Antibiotics (Basel)*, vol. 11, no. 4, p. 512, 2022.

[9] J. F. Cobo-Díaz, A. Alvarez-Molina, E. A. Alexa *et al.*, "Microbial colonization and resistome dynamics in food processing environments of a newly opened pork cutting industry during 1.5 years of activity," *Microbiome*, vol. 9, p. 204, 2021. [Online]. Available: https://doi.org/10.1186/s40168-021-01131-9

[10] K. A. Ryon, B. T. Tierney, A. Frolova, A. Kahles, C. Desnues, C. Ouzounis, C. Gibas, D. Bezdan, Y. Deng, D. He, E. Dias-Neto, E. Elhaik, E. Afshin, G. Grills, G. Iraola, H. Suzuki, J. Werner, K. Udekwu, L. Schriml, M. Bhattacharyya, M. Oliveira, M. M. Zambrano, N. H. Hazrin-Chong, O. Osuolale, P. P. Łabaj, P. Tiasse, S. Rapuri, S. Borras, S. Pozdniakova, T. Shi, U. Sezerman, X. Rodo, Z. H. Sezer, and C. E. Mason, "A history of the metasub consortium: Tracking urban microbes around the globe," *iScience*, vol. 25, no. 11, p. 104993, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2589004222012652

[11] McMaster University, "Comprehensive Antibiotic Resistance Database," https://card.mcmaster.ca/home, accessed: [insert date of access here].

[12] M. K. Gibson, K. J. Forsberg, and G. Dantas, "Improved annotation of antibiotic resistance functions reveals microbial resistomes cluster by ecology," *The ISME Journal*, 2014, doi:ISMEJ.2014.106.

[13] D. Danko, D. Bezdan, E. E. Afshin, S. Ahsanuddin, C. Bhattacharya, D. J. Butler, K. R. Chng, D. Donnellan, J. Hecht, K. Jackson, K. Kuchin, M. Karasikov, A. Lyons, L. Mak, D. Meleshko, H. Mustafa, B. Mutai, R. Y. Neches, A. Ng, O. Nikolayeva, T. Nikolayeva, E. Png, K. A. Ryon, J. L. Sanchez, H. Shaaban, M. A. Sierra, D. Thomas, B. Young, O. O. Abudayyeh, J. Alicea, M. Bhattacharyya, R. Blekhman, E. Castro-Nallar, A. M. Cañas, A. D. Chatziefthimiou, R. W. Crawford, F. De Filippis, Y. Deng, C. Desnues, E. Dias-Neto, M. Dybwad, E. Elhaik, D. Ercolini, A. Frolova, D. Gankin, J. S. Gootenberg, A. B. Graf, D. C. Green, I. Hajirasouliha, J. J. Hastings, M. Hernandez, G. Iraola, S. Jang, A. Kahles, F. J. Kelly, K. Knights, N. C. Kyrpides, P. P. Łabaj, P. K. Lee, M. H. Leung, P. O. Ljungdahl, G. Mason-Buck, K. McGrath, C. Meydan, E. F. Mongodin, M. O. Moraes, N. Nagarajan, M. Nieto-Caballero, H. Noushmehr, M. Oliveira, S. Ossowski, O. O. Osuolale, O. Özcan, D. Paez-Espino, N. Rascovan, H. Richard, G. Rätsch, L. M. Schriml, T. Semmler, O. U. Sezerman, L. Shi, T. Shi, R. Siam, L. H. Song, H. Suzuki, D. S. Court, S. W. Tighe, X. Tong, K. I. Udekwu, J. A. Ugalde, B. Valentine, D. I. Vassilev, E. M. Vayndorf, T. P. Velavan, J. Wu, M. M. Zambrano, J. Zhu, S. Zhu, C. E. Mason, N. Abdullah, M. Abraao, A.-h. Adel, M. Afaq, F. S. Al-Quaddoomi, I. Alam, G. E. Albuquerque, A. Alexiev, K. Ali, L. E. Alvarado-Arnez, S. Aly, J. Amachee, M. G. Amorim, M. Ampadu, M. A.-F. Amran, N. An, W. Andrew, H. Andrianjakarivony, M. Angelov, V. Antelo, C. Aquino, Aranguren, L. F. Araujo, H. F. Vasquez Arevalo, J. Arevalo, C. Arnan, L. E. Alvarado Arnez, F. Arredondo, M. Arthur, F. Asenjo, T. S. Aung, J. Auvinet, N. Aventin, S. Ayaz, S. Baburyan, A.-M. Bakere, K. Bakhl, T. F. Bartelli, E. Batdelger, F. Baudon, K. Becher, C. Bello, M. Benchouaia, H. Benisty, A.-S. Benoiston, J. Benson, D. Benítez, J. Bernardes, D. Bertrand, S. Beurmann, T. Bitard-Feildel, and Bitt, "A global metagenomic map of urban microbiomes and antimicrobial resistance."

[14] "CAMDA 2023," http://camda2023.bioinf.jku.at, accessed: [insert date here].

[15] T. I. M. Consortium, "A global metagenomic map of urban microbiomes and antimicrobial resistance," vol. 183, 2021. [Online]. Available: https://www.cell.com/action/showPdf?pii=S0092-8674%2821%2900585-7

[16] B. Tierney, L. Mak, R. Toscan, D. Meleshko, C. Ronkowski, J. Henriksen, M. Toomey, S. Fumagalli, and C. Fu, "Metasubcamp," 2023. [Online]. Available: https://github.com/MetaSUB-CAMP

[17] S. Nurk, D. Meleshko, A. Korobeynikov, and P. PA, "Metaspades: a new versatile metagenomic assembler," 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5411777/

[18] Microbial-Ecology-Group, "Amrplusplus: A bioinformatic pipeline for antimicrobial resistance analysis," https://github.com/Microbial-Ecology-Group/AMRplusplus, 2023.

[19] N. Bonin, E. Doster, H. Worley, L. J. Pinnell, J. E. Bravo, P. Ferm, S. Marini, M. Prosperi, N. Noyes, P. S. Morley, and C. Boucher, "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.*, vol. gkac1047, 2022, epub ahead of print.

[20] B. Langmead and S. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012. [Online]. Available: https://doi.org/10.1038/nmeth.1923

[21] N. S. Mojica, "ccm-bioinfo/cambda2023," 2023. [Online]. Available: https://github.com/ccm-bioinfo/cambda2023/tree/main/06$_amr_resistance$

[22] M. Feldgarden, V. Brover, N. Gonzalez-Escalona, J. G. Frye, J. Haendiges, D. H. Haft, M. Hoffmann, J. B. Pettengill, A. B. Prasad1, G. E. Tillman, G. H. Tyson, and W. Klimke, "Amrfinderplus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence," vol. 183, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8208984/

[23] Comprehensive Antibiotic Resistance Database (CARD), "Resistance gene identifier (rgi)," https://github.com/arpcard/rgi, accessed: [insert date of access here].

[24] DOE Joint Genome Institute, "Bbtools," https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/, accessed: [insert date of access here].

[25] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The silva ribosomal rna gene database project: improved data processing and web-based tools," *Nucl. Acids Res.*, vol. 41, no. D1, pp. D590–D596, 2013.

[26] C. L. Brown, J. Mullet, F. Hindi, J. E. Stoll, S. Gupta, M. Choi, I. Keenum, P. Vikesland, A. Pruden, and L. Zhang, "MobileOG-db: A manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements," *Appl. Environ. Microbiol.*, vol. 88, no. 18, p. e0099122, Sep. 2022.

[27] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Mar. 2010.

[28] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nat. Methods*, vol. 12, no. 1, pp. 59–60, Jan. 2015.

[29] R. A. Horn and C. R. Johnson, *Norms for vectors and matrices*. Cambridge University Press, 1985, p. 257–342.

[30] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, "Singular value decomposition and principal component analysis," 2003.

[31] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Review*, vol. 37, no. 4, pp. 573–595, 1995. [Online]. Available: https://doi.org/10.1137/1037127

[32] K. Mnich and W. R. Rudnicki, "All-relevant feature selection using multidimensional filters with exhaustive search," *Information Sciences*, vol. 524, pp. 277 – 297, 2020.

[33] R. Piliszek, K. Mnich, S. Migacz, P. Tabaszewski, A. Sułecki, A. Polewko-Klim, and W. Rudnicki, "MDFS: MultiDimensional Feature Selection in R," *The R Journal*, 2019. [Online]. Available: https://doi.org/10.32614/RJ-2019-019

[34] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, ser. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993, no. 57.

[35] J. Shao and C. F. J. Wu, "A General Theory for Jackknife Variance Estimation," *The Annals of Statistics*, vol. 17, no. 3, pp. 1176 – 1197, 1989. [Online]. Available: https://doi.org/10.1214/aos/1176347263

[36] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software*, vol. 36, pp. 1–13, 2010.

[37] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[38] M. Fernández-Delgado *et al.*, "Do we need hundreds of classifiers to solve real world classification problems," *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 3133–3181, 2014.

[39] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442—-451, 1975.