Received: 2 October 2023 | Revised: 31 January 2024 | Accepted: 27 February 2024

DOI: 10.1002/sim.10058

RESEARCH ARTICLE

Statistics in Medicine WILEY

Multiblock partial least squares and rank aggregation: Applications to detection of bacteriophages associated with antimicrobial resistance in the presence of potential confounding factors

Shoumi Sarkar | **Samuel Anyaso-Samuel** | **Peihua Qiu** | **Somnath Datta**

Department of Biostatistics, University of Florida, Gainesville, Florida,

Correspondence
Somnath Datta, Department of Biostatistics, University of Florida, Gainesville, FL 32611, USA.
Email: somnath.datta@ufl.edu

Urban environments, characterized by bustling mass transit systems and high population density, host a complex web of microorganisms that impact microbial interactions. These urban microbiomes, influenced by diverse demographics and constant human movement, are vital for understanding microbial dynamics. We explore urban metagenomics, utilizing an extensive dataset from the Metagenomics & Metadesign of Subways & Urban Biomes (MetaSUB) consortium, and investigate antimicrobial resistance (AMR) patterns. In this pioneering research, we delve into the role of bacteriophages, or "phages"–viruses that prey on bacteria and can facilitate the exchange of antibiotic resistance genes (ARGs) through mechanisms like horizontal gene transfer (HGT). Despite their potential significance, existing literature lacks a consensus on their significance in ARG dissemination. We argue that they are an important consideration. We uncover that environmental variables, such as those on climate, demographics, and landscape, can obscure phage-resistome relationships. We adjust for these potential confounders and clarify these relationships across specific and overall antibiotic classes with precision, identifying several key phages. Leveraging machine learning tools and validating findings through clinical literature, we uncover novel associations, adding valuable insights to our comprehension of AMR development.

KEYWORDS

antimicrobial resistance, bacteriophages, metagenomics, multi-block partial least squares, rank aggregation

1 | INTRODUCTION

The urban environment teems with a multitude of unseen inhabitants that possess the power to reshape our comprehension of microbial interactions. Urban mass transit systems, crucial conduits in modern cities, provide a habitat for a diverse range of microorganisms, forming complex urban microbiomes that intermingle with the lives of millions. Such locations are considered hubs for microbes due to high human traffic, enclosed spaces, shared surfaces, air

circulation, interconnected routes, and diverse passenger demographics.[1] The constant flow of people and close interactions in transit vehicles and stations create favorable conditions for the transmission and exchange of microorganisms. These environments can serve as reservoirs for various microbes, posing potential public health implications. As global urbanization escalates, deciphering the intricacies of these microbial networks assumes paramount importance. The Metagenomics & Metadesign of Subways & Urban Biomes (MetaSUB) consortium has undertaken an ambitious mission to unravel genetic signatures of microbial populations residing within urban communities.[2] Through systematic sampling across diverse urban landscapes worldwide, this consortium continues to conduct an extensive exploration into the realm of metagenomics. In response to the Critical Assessment of Massive Data Analysis[3] (CAMDA) 2023 challenge, the scientific community converges to probe the extensive metagenomic dataset amassed by the MetaSUB consortium, unearthing latent connections previously untapped. This challenge exhorts a comprehensive examination of anti-microbial resistance (AMR) patterns in this vast metagenomic surveillance data, attracting dedicated researchers striving to uncover these complex interactions. Unlike previous studies that have predominantly focused on geolocation prediction[4–6] or spatial modelling[7,8] of such patterns, our research forges a new path by delving into the uncharted territory of bacteriophages' role in orchestrating AMR dissemination. Bacteriophages, also known as phages, are viruses that prey on bacteria.[9] They harbor the capacity to transfer genetic information between microbial hosts through mechanisms such as HGT, potentially catalyzing the transmission of resistance genes.[10] When a phage infects an ARG-harboring bacterium, it might unintentionally include parts of the bacterial DNA, ARGs included, in its replication process. These new phage particles, now equipped with ARGs from their prior hosts, can transfer these genes to other bacteria, thereby contributing to the widespread distribution of resistance genes within bacterial populations.[10] Despite their profound potential to influence microbial dynamics, the existing literature lacks a consensus regarding the magnitude of their association with AMR.[11] While some studies have suggested a limited role for phages in this process,[12] others have emphasized their significant impact on the genetic exchange of antibiotic resistance.[10] We strive to bridge this gap by quantifying these associations, thereby illuminating the intricate nature of these under explored relationships.

The historical trajectory of AMR research has notably underscored the significance of environmental variables–such as sanitation levels, proximity to water bodies, and climate conditions–in the dissemination of antimicrobial-resistant genes (ARGs), also known as resistomes. Evidence gathered from global datasets, with a notable focus on Brazil, contributes to the discourse concerning the potential hazards of sewage and livestock manure in disseminating antibiotic resistance.[13] A range of human-induced activities, including the introduction of contaminated river runoff, the output from wastewater treatment plants, sewage discharges, as well as practices related to aquaculture, promote the spread of ARGs in estuarine and coastal ecosystems.[14] HGT is increased by rising temperatures, in addition, temperature increases generally facilitate bacterial growth.[15] We shift the spotlight away from these established environmental factors and direct it towards the unexplored domain of phages' roles. Concurrently, we strive to assess the relative significance of these factors in contrast to the role of phages when unraveling AMR dynamics.

The effects of such environmental factors also intertwine with the potential impact of phages. We aim to untangle these relationships while accounting for potential confounding effects posed by environmental variables. To accomplish this, we will harness the most relevant machine learning tools, leveraging their analytical prowess to uncover the associations woven within this intricate interplay. This study is one of the few conducted on this important topic and some of our findings are novel. We believe we have used proper statistical tools to reach our conclusions which were lacking in previous literature.

The metagenomic data from the CAMDA 2023 challenge comprises comprehensive metagenomic information that enables us to extract resistome and phage abundances, which we supplement with auxiliary data on pertinent environmental variables on climate, demographics, and landscape collected from publicly available databases. The resistome abundances quantify the ARGs resistant to 17 classes of antibiotics and form a multivariate response. Despite the sparse and high-dimensional nature of the phage abundance data, compounded by the potential confounding effects posed by the auxiliary environmental variables, our analysis is strategically designed to address three principal objectives. These objectives encompass assessing the relative importance of distinct groups of genetic and non-genetic factors, discerning key phages, and validating our method's efficacy through alignment with clinical literature. By segmenting the variables into separate blocks corresponding to phage abundances, climatic factors, demographics, and landscape, we employ multi-block partial least squares regression[16] (MPLS) to elucidate the associations between both the blocks and ARGs. In addition, we devise a two-step strategy that focuses on isolating the contributions of the phages on the ARGs. First, for each class of antibiotic, we partial out the effects of the potential confounding environmental variables. In the second step, a random forest[17] model in conjunction with a weighted rank aggregation[18] approach is utilized to obtain a

consensus ranking ordering of the phages, highlighting the order of the phage importance in the spread of AMR. The selected top phages associated with ARGs are corroborated with findings from clinical literature. This study introduces a viable alternative to traditional, demanding lab methods in identifying phages associated with ARG dissemination.[19] Furthermore, it highlights essential genetic organisms for microbiologists to investigate in lab-based research, enhancing comprehension of ARG spread.

The rest of the paper is organized as follows. Section 2 presents an overview of the dataset utilized in our study. Section 3 goes over the described statistical tools to achieve our objective. Section 4 examines the outcomes of our analyses. The main body of the paper ends with a discussion in Section 5.

2 | DATA

Metagenomic data is provided by the organizers of the CAMDA challenge as a part of their "Anti-Microbial Resistance Prediction and Forensics Challenge 2023". The primary dataset comprises 366 raw whole-genome shotgun (WGS) samples from 16 urban cities across the globe. Table 1 shows the distribution of the samples from the different locations, including a city in Oceania, six cities in North America, four cities in Europe, two cities in South America, two cities in Asia, and one city in Africa. Each sample consists of paired-end sequencing reads in FASTQ format. Independently, we obtain relevant metadata describing the landscape, demographics, and climate for these locations (further elucidated in Section 2.3). To make meaningful inferences from the raw sequence reads, we construct a robust bioinformatics pipeline for the downstream processing of the data. Subsequently, we apply relevant machine learning algorithms to the relative abundances obtained after the pre-processing step. Figure 1 outlines the main steps of our analysis.

2.1 | Bioinformatics pipeline

A total of 366 WGS metagenomic samples are downloaded from the CAMDA host server. To prepare the raw sequence data for downstream statistical analysis, we employ a standard bioinformatic pipeline for pre-processing. In the initial stage of the pipeline, we perform quality checks which necessitate the filtering and trimming of the reads. Next, we carry taxonomic profiling of the reads with acceptable quality scores to obtain phage and resistome abundances. The bioinformatic procedures utilized in this pipeline were performed using the HiPerGator supercomputer located at the University of Florida.

In the initial step of the bioinformatic procedures, we assess the quality of the raw reads using FASTQC[20] and MULTIQC.[21] A significant proportion of low-quality bases and adapter sequences were identified in the raw reads. Consequently, we invoked Trimmomatic[22] from KneadData[23] to trim and remove the adapter sequences from the raw reads. We retained reads with a length of 60 base pairs and a minimum Phred64 quality score of 30. Subsequently, we filtered out host (human) contamination from the trimmed reads by indexing the human reference genome and discarding reads that mapped to it, using BowTie.[24] Any read names that are not identical between pairs due to these preprocessing steps are rectified using the `Repair` function in bbmap.[25]

TABLE 1 Distribution of raw paired-end WGS metagenomic samples obtained across 16 cities.

Location	# of samples	Location	# of samples
Auckland	14	Minneapolis	6
Baltimore	13	New York	46
Berlin	15	Sacramento	16
Bogota	15	San Antonio	16
Denver	44	Sao Paulo	25
Doha	27	Tokyo	49
Ilorin	34	Vienna	16
Lisbon	14	Zurich	16

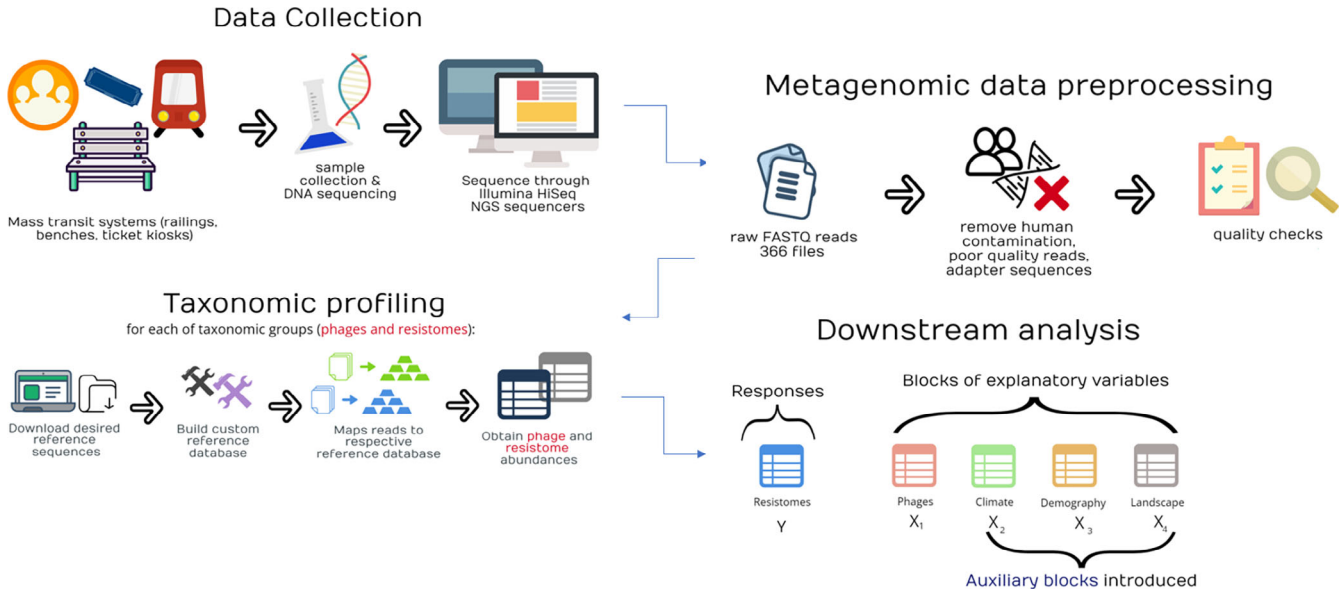

FIGURE 1 Overview of the analysis pipeline.

For taxonomic profiling of bacteriophages, we used the RefSeq[26] database containing viral genetic sequences hosted by NCBI. First, we downloaded the reference sequences of bacteriophages from the NCBI RefSeq database of viral genomes. As of December 2022, the RefSeq database contains approximately 11,555 complete viral genomes, of which 4,293 are phages (bacteria: 4194; archaea: 99). To quantify the abundance of phages in the samples, we indexed the reference sequences in FASTQ format using k-mer mapping[27] (KMA) with an index (k-mer) size of 16. Subsequently, we mapped the pre-processed reads to the RefSeq database of viral genomes using KMA. We obtain the abundance of the phages at the species level from the *.mapstat* file generated by KMA.

Further, to quantify the abundance of resistomes in the WGS samples, we employed the ResFinder[28] database. ResFinder is a curated database that contains information on a wide range of AMR genes, including their nucleotide sequences, annotation, and associated metadata such as antibiotic resistance phenotypes and mobile genetic elements. The database also provides information about these genes' resistance to 17 distinct classes of antibiotics. The antibiotic classes covered by ResFinder include aminoglycoside, beta-lactam, colistin, fosfomycin, fusidicacid, glycopeptide, macrolide, nitroimidazole, oxazolidinone, phenicol, pseudomonicacid, quinolone, rifampicin, sulphonamide, tetracycline, trimethoprim and miscellaneous. For each antibiotic class, we index the corresponding database, then we map the pre-processed reads to the database, using KMA.

2.2 | Pre-processing

We constructed an abundance table of phages–the genetic explanatory variables–by identifying 4,250 operational taxonomic units (OTUs) and applying a filtering threshold of 1% presence or five reads. The resulting high-dimensional data matrix contained the abundances of 1,190 phage species, which we normalized to relative abundances using the `metagenomeSeq`[29] package in R. In parallel, we obtained and normalized abundance data for resistomes corresponding to the 17 antibiotic classes outlined in Section 2.1. These data were arranged in a matrix format, with the samples as rows and the relative abundances for each antibiotic class as columns. The resulting counts, which reflect the relative extent of ARG presence, serve as the multivariate response.

2.3 | Auxiliary variables

In addition to the data on phage relative abundances, we include other data describing the environmental factors of the cities where the samples were obtained. The metadata comprises information relating to the climate (minimum and maximum annual temperatures, minimum and maximum relative humidity, and annual rainfall), demographics

(median annual household income, average age, and percentage of population with access to basic sanitation services), and landscape (Air Quality Index score, elevation above sea level, city land area, proximity to the coast, latitude, longitude, and population), all for the year 2017. Information regarding these environmental variables originates from publicly available sources managed by various national atmospheric research and air quality monitoring organizations corresponding to the diverse sampling locations. These sources include National Institute of Water and Atmospheric Research (NIWA), National Oceanic and Atmospheric Administration (NOAA), Deutscher Wetterdienst (DWD), Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), Qatar Meteorology Department, Nigerian Meteorological Agency (NIMET), Instituto Português do Mar e da Atmosfera (IPMA), Instituto Nacional de Meteorologia (INMET), Japan Meteorological Agency (JMA), Central Institute for Meteorology and Geodynamics (ZAMG), Swiss Federal Office of Meteorology and Climatology (MeteoSwiss). In addition, we obtain publicly available demographic and landscape information from World Bank, World Health Organization (WHO), and the United Nations Children's Fund (UNICEF), and respective national census organizations corresponding to the sample locations. The numeric variables are normalized by centering and scaling.

3 | METHODOLOGY

3.1 | Blocking variables

The explanatory variables form groups, or "blocks", as they pertain to either phages, climate, demographics, or landscape. Let $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_3$, and $\mathbf{X}_4$ denote the $K = 4$ blocks corresponding to phage relative abundances ($p_1 = 1190$ variables), characteristics of climate ($p_2 = 5$ variables), demographics ($p_3 = 3$ variables), and landscape ($p_4 = 7$ variables), respectively. The multivariate response, $\mathbf{Y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_{17}})$ with 17 variables, comprises the resistomes corresponding to each class of antibiotics. Therefore, the data for downstream analysis can be organized into $(K + 1) = (4 + 1)$ blocks.

3.2 | Multi-block partial least squares

Multiblock methods aim to reduce the dimensionality to capture the primary connections between variables and responses. To describe the association between $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4)$ and $\mathbf{Y}$, we utilize MPLS to select explanatory variables that are strongly associated with the multivariate response. In this study, we are interested in determining and ordering the effects of genetic and environmental factors in describing the relative abundances of resistomes. Additionally, we aim to estimate the contribution of each explanatory variable in describing the resistomes. Finally, we aim to obtain a significant set of phages that drives the variation of the resistomes while accounting for the potential confounding effect of the environmental factors.

We wish to obtain an ordered list of the blocks (of explanatory variables) that are sorted by the strength of their effects in explaining the response, as well as an ordered list of all $P = \sum_{k=1}^{K} p_k$ explanatory variables. Both objectives can be readily assessed using MPLS.[16] MPLS is designed to find the underlying relationships between the blocks of data, with the aim of understanding how they are related to each other. It does this by identifying the linear combinations of variables from each block that are most strongly related to the outcome variable. These linear combinations are called latent variables or factors. The goal of MPLS is to find the best set of latent variables that explain the variation in the outcome variables while also taking into account the relationships between the different blocks of data. The method is based on maximizing the covariance between the outcome variable and the linear combinations of variables from each block. At each iteration, MPLS extracts a new set of latent variables that capture the maximum covariance between the blocks. The algorithm then orthogonalizes the data with respect to the new latent variables and repeats the process until a predetermined number of latent variables have been extracted or until convergence is reached.

3.2.1 | Block importance, variable importance, and PLS regression coefficients

MPLS provides several measures to summarize the relationships between the responses and explanatory variables, including block importance, variable importances, and response class-specific variable significance.[30] With a pre-specified

number of dimensions H, the algorithm iterates over dimensions $h = 1, \ldots, H$, and calculates relative importance $\lambda^{(h)}$ of dimension h.[30] Additionally, it computes explanatory loadings $\mathbf{w}^{*(\mathbf{h})}$, which are coefficients or weights representing the contribution of a variable to latent components, as well as block importance coefficient $a_k^{(h)}$ which measures the link between $\mathbf{Y}$ and $\mathbf{X_k}$, $k = 1, 2, 3, 4$. Variable importance $vip^{(h)}$ of a dimension h is the average of the $\mathbf{w}^{*(\mathbf{h})^2}$ weighted by $a_k^{(h)^2}$. Cumulated variable importance, *vipc*, is the average of $vip^{(h)}$, weighted by the importance of each dimension $\lambda^{(h)}$. Cumulated block importance, *bipc*, is the average of block importances $a_k^{(h)^2}$ weighted by $\lambda^{(h)}$. Also, for each component $\mathbf{y_r}$, $r = 1, \ldots, 17$ of the multivariate response, PLS regression coefficients correspond to univariate PLS regression coefficients and measure the links between $\mathbf{X}$ and $\mathbf{y_r}$.

The *vipc* scores are used to measure the global contribution of each variable across all blocks in the model. The *bipc* scores are used to measure the relative importance of the blocks. Further, we use a bootstrap procedure to estimate the 95% confidence interval of the *vipc* and *bipc*. A variable is assessed to be significantly important in predicting the response if the 95% confidence interval of the *vipc* does not contain $\frac{1}{P}$.[30] While a block has significant predictive power if the 95% confidence interval of the *vipc* does not contain $\frac{1}{K}$.[30] PLS regression coefficients correspond to separate univariate PLS analyses to assess the association between phages and antibiotic class-specific resistome relative abundances. A similar 95% bootstrapped confidence interval is constructed for the PLS coefficients of each variable, where the confidence interval not containing 0 indicates that the corresponding variable is significant.

3.3 | Controlling for potential confounders and identifying key phages

In addition, we are interested in determining the contributions of the phages in explaining the variation of the resistomes for separate classes of antibiotics. The goal is to identify a signature of phages for the resistomes while adjusting for covariates (environmental factors). We utilize a two-step strategy: first, we fit separate univariate models *apriori*. In this step, we regress a different response ($\mathbf{Y}_r; r = 1, \ldots, 17$) on all environmental variables, then, we partial out the effects of these variables to obtain the residuals ($\mathbf{Y}_r^*; r = 1, \ldots, 17$). Note that, $\mathbf{Y}_r$ and $\mathbf{Y}_r^*$ are vectors of length n. This corresponds to obtaining the relative abundances of the resistomes with environmental effects removed. In the second step, these residuals are used as the responses in random forest (RF) models, while the predictors are the phage relative abundances. The RF algorithm calculates importance scores based on the mean decrease in accuracy,[31] which we use to obtain an ordered list of phages for each antibiotic class. To obtain an overall top-k ordering of the phages, we utilize the rank aggregation algorithm. The algorithm combines multiple ordered lists into a single overall top-k ordered list, using the respective importance scores obtained from the random forest models as weights. Figure 2 provides a schematic overview of the approaches described in this section for isolating the contributions of the phages in describing the resistance to the antibiotics.

3.3.1 | Stability of rank aggregation results

Following the methodology described in the preceding section, to ensure that the consensus list of top-k ordered phages is superior to a selection based on random chance, we compute "inclusion probabilities" of each phage in the top-k rank aggregation list, and compare them with a threshold that corresponds to the probability of random selection. The inclusion probabilities are calculated through a jackknifing approach: for each of the 366 samples, we leave one sample out at a time and obtain the corresponding top-k rank aggregated list. Therefore, we get 366 separate rank-aggregated lists. Next, we compute the proportion of times the phages in the original rank aggregation list (based on all 366 samples) appear in these leave-one-out lists. We name these proportions as inclusion probabilities which express the stability of our selection. To calculate the probability that a certain phage in the top-k list was selected at random, we employ the hypergeometric distribution: we calculate the probability of selecting the phage from the group of top-k phages, and selecting no phages from the group of the remaining $1190 - k$ phages, which is $\binom{20}{k}\binom{1190-k}{0}/\binom{1190}{1}$. Inclusion probabilities above the computed threshold indicate robustness against selection based on random chance.

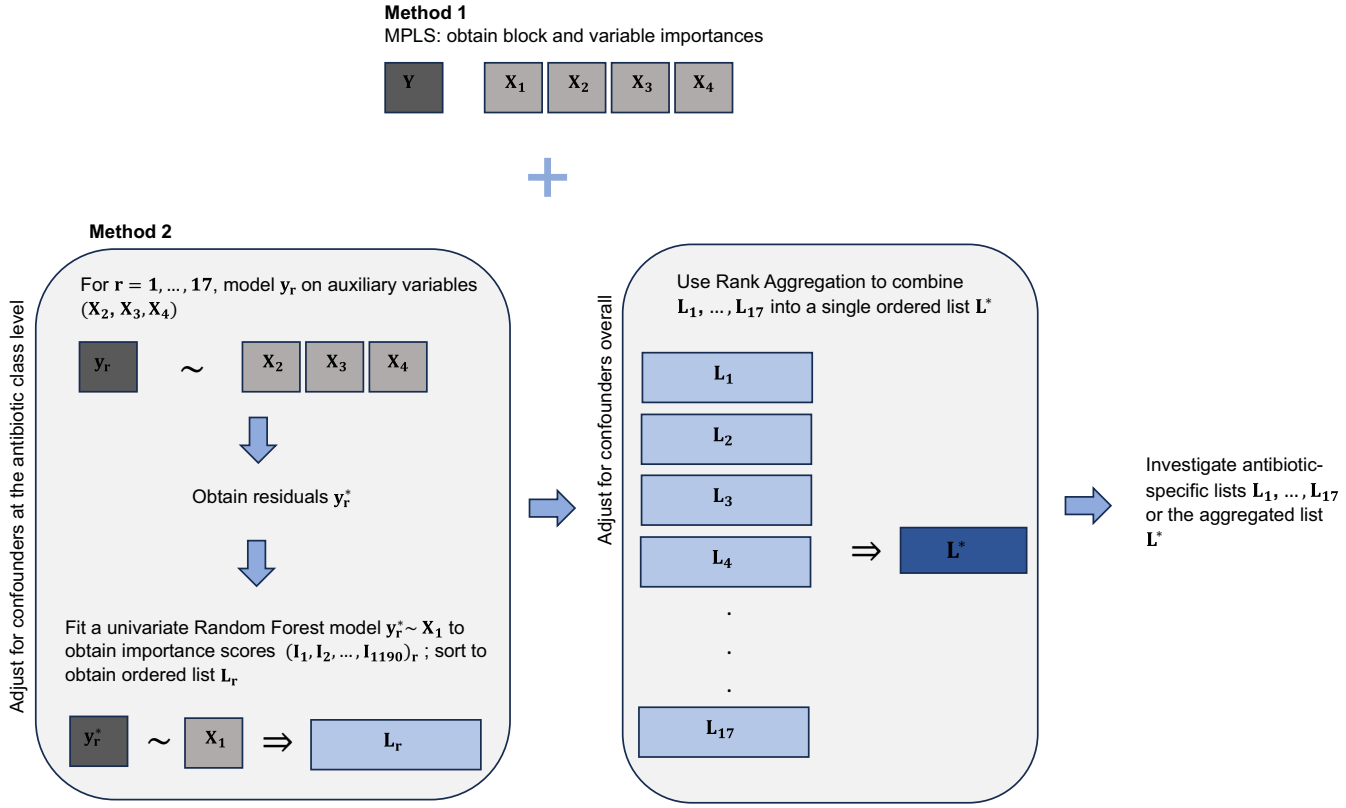

FIGURE 2 Downstream analysis pipeline.

4 | RESULTS

4.1 | Analyses using MPLS

4.1.1 | Blocks of phages in comparison to the blocks of environmental factors

We employ the `ade4`[30] package in R to fit the MPLS model. We obtain *bipc* scores which quantify the relative importance of each block of variables in explaining the (multivariate) resistome abundances. Figure 3a shows the interval plots of the importance scores for each block of predictors. Higher block importance scores indicate that a particular block has a greater impact on the response. Our analysis highlights the significant contribution of phages in explaining the variation in the relative abundances of resistomes, consistent with previous research linking phages to the dissemination of ARGs.[32] The importance plot also reveals the significance of climatic factors, in line with emerging evidence linking the climate crisis to antimicrobial resistance.[33–35] Although not significant, the landscape and demographic factors show importance scores similar to those of the climatic factors.

In line with a valuable suggestion from a reviewer, we also explore block importances at a higher taxonomic level for phages; this approach yields more "balanced" block sizes (details of this additional exploratory analysis are included in Section S1 of the Supplementary File). This assessment encompasses blocks representing phage families, in addition to climate, demographics, and landscape. The environmental blocks are found to assume the anticipated primary importance, while a significant proportion of phage families retain their importance with reduced *bipc* scores. These findings support our hypotheses, suggesting that environmental factors may overshadow the influence of phages. Consequently, prioritizing a more specific analysis of the phage abundance and its relation to antimicrobial resistance is warranted.

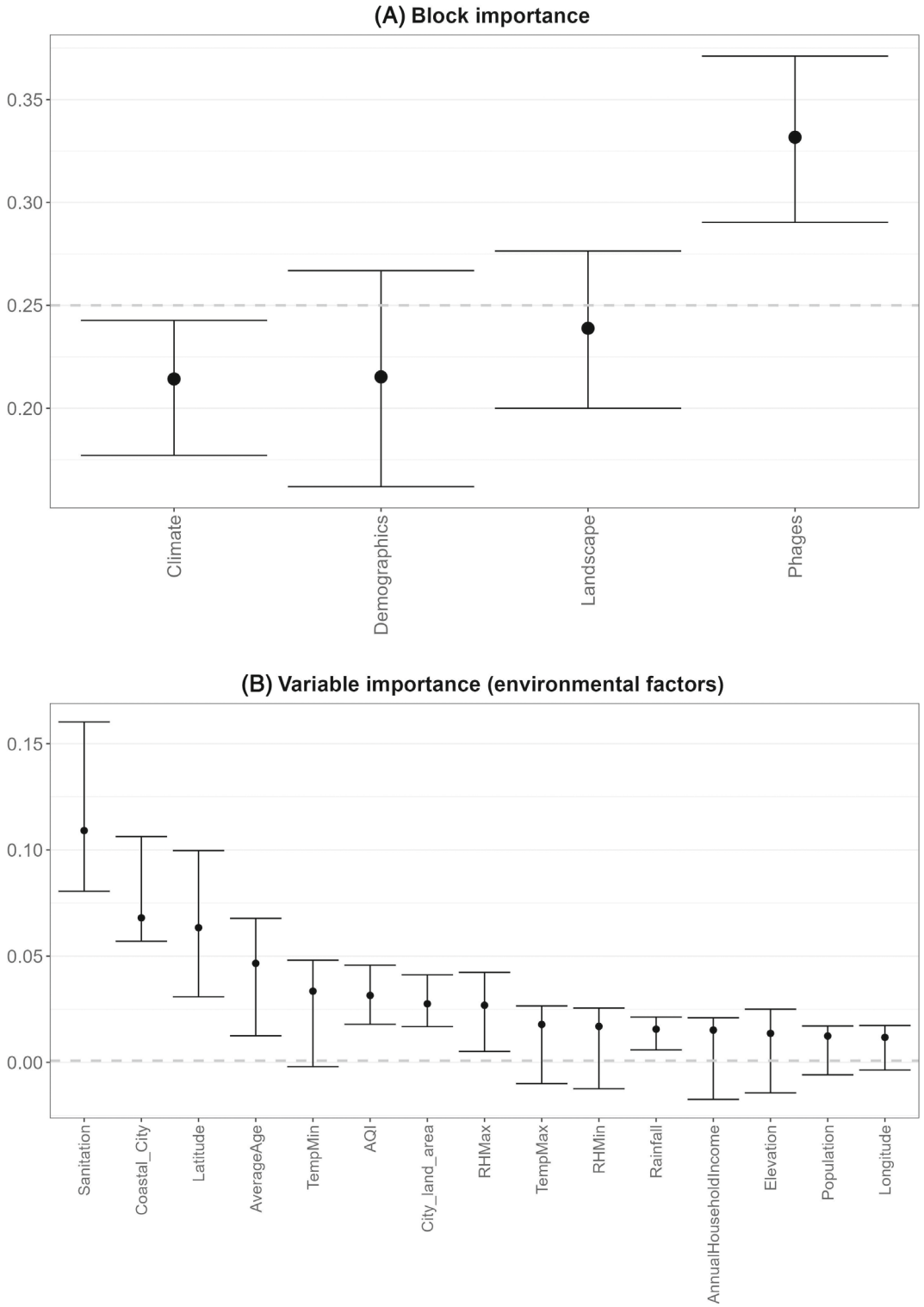

FIGURE 3 *bipc* and *vipc* from multi-block partial least squares, and their associated 95% confidence intervals. In panel (a), the *y* axis and the dotted horizontal line represent the *bipc* value and the $1/K$ value, respectively. Similarly, in panels (b) and (c), they correspond to the *vipc* and the $1/P$ value. The intervals for the first 15 (ordered by the *vipc*) variables are shown in (b). Note that these variables correspond to all the environmental variables considered in this study. While the interval plots in (c) correspond to the next 50 (ordered by the *vipc*) variables, note that these variables correspond to a subset of the phages with higher importance.

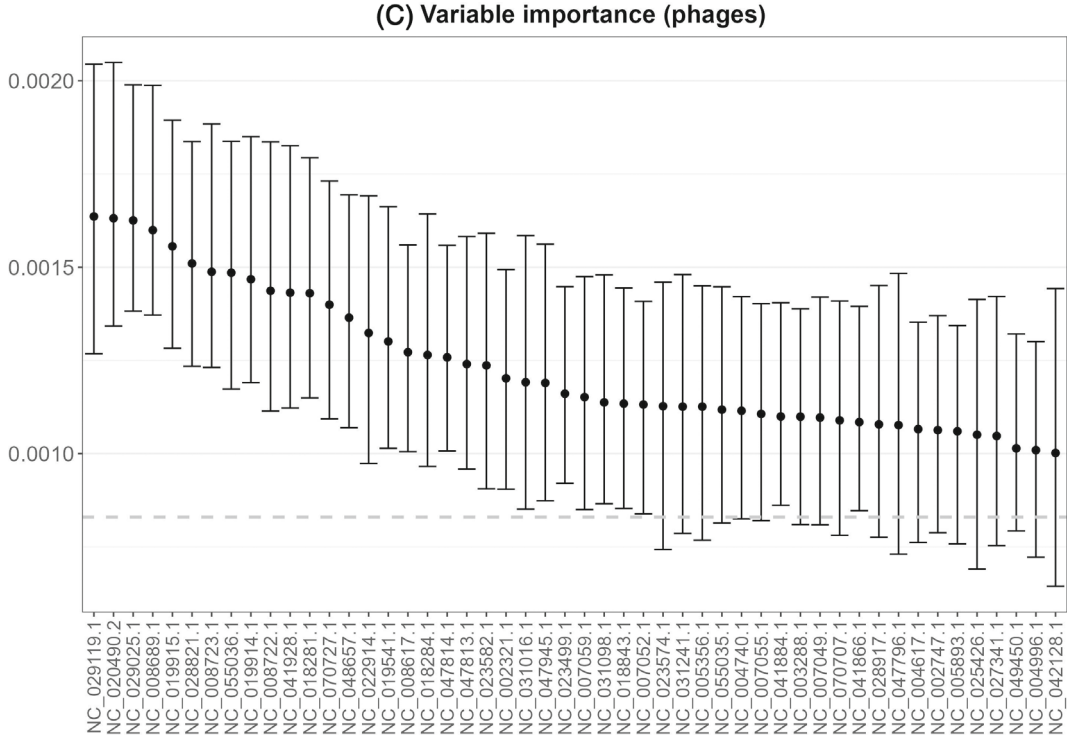

FIGURE 3 (Continued)

4.1.2 | Ordering variables by global importance

Our second objective is to order all predictor variables in terms of their global importance. We achieve this by using the *vipc* scores obtained from the fitted MPLS model. These scores measure the overall contribution of each variable in explaining the variation in the resistome block. Variables with higher importance scores have a stronger association with the multivariate response. Interestingly, we found that all variables corresponding to environmental factors had higher importance scores than those corresponding to phages. Figure 3b shows the interval plots of the importance scores for the first 15 predictors (ordered by *vipc*). This set of predictor variables corresponds to the $p_2 + p_3 + p_4$ variables on the environmental factors. Figure 3c shows the next 50 variables (ordered by *vipc*); the variables correspond to a subset of phages with increasing magnitude of *vipc* scores, less than the scores corresponding to the environmental variables.

From Figure 3b, we notice that the percentage of the population (where the sample was collected) with basic access to sanitation had the largest significant impact in explaining the variation in the block of resistomes. Other variables that have a significant impact on explaining the variation in the block of resistome include the indicator of whether the sample was collected from a coastal city, latitude, the average age of the city's population, air quality index, city land area, maximum relative humidity, and annual rainfall. Their effects are further elucidated in Section 4.3.2.

4.1.3 | Quantifying relationships between explanatory variables and resistomes

Now, we present the results where we measure the association between each explanatory variable in **X** and each component of the multivariate response **Y**. As seen in Figure 4, *macrolide*, *beta-lactam*, and *tetracycline* exhibit the highest relative resistome abundances. We prioritize these specific antibiotic classes as they might be more prone to resistance development in the sampled urban environments.[36–39] Panels (a), (b), and (c) of Figure 5 show the estimated regression coefficient (and 95% confidence interval) from the three respective models where the resistome abundance corresponding to the *macrolide*, *beta-lactam*, and *tetracycline* class of antibiotics are regressed on all the explanatory variables. For ease of visualizing the results corresponding to the three antibiotic classes, Figure 5a–c show only the top 20 variables (ordered by the magnitude of the estimated coefficient) from the respective models. From these plots, the variables

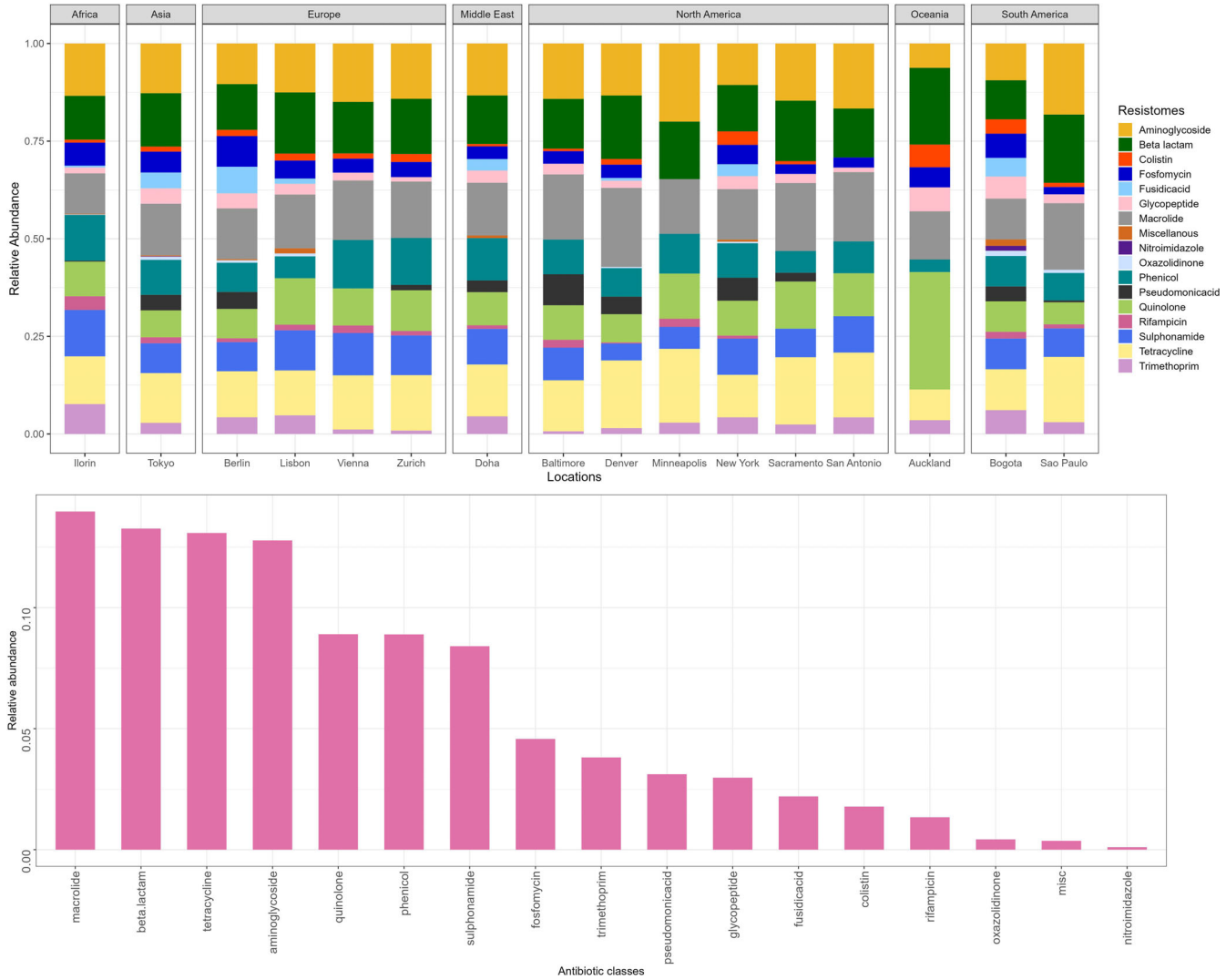

FIGURE 4 Top panel: Distribution of resistome abundances by geographical location; bottom panel: Overall relative abundances of resistomes across the 17 antibiotic classes.

corresponding to the environmental factors have stronger effects in explaining the variation in the resistomes corresponding to the given classes of antibiotics, despite phages forming the most important block (as shown in Figure 3a). These disparate results may be due to the confounding of the association of phages with resistomes by environmental factors.

The phages in these plots are labeled by their corresponding NCBI accession numbers, rather than their longer complete phage names, for the sake of optimal presentation. Among phages, we find that *Staphylococcus* phage SPbeta-like (NC_029119.1) has the highest magnitude of regression coefficient across the three antibiotic classes, as well as globally in terms of *vipc* score, as seen in Figures 3c and 5. *Staphylococcus* phage StB12 (NC_020490.2), *Staphylococcus* phage IME-SA4 (NC_029025.1), *Staphylococcus* phage StB20 (NC_019915.1), *Staphylococcus* phage StB27 (NC_019914.1), *Staphylococcus* phage StB20-like (NC_028821.1), *Staphylococcus* phage virus 108PVL (NC_008689.1), *Staphylococcus* phage PH15 (NC_008723.1), *Staphylococcus* phage IME1348_01 (NC_055036.1), and *Staphylococcus* phage Ipla5 (NC_018281.1) also consistently appear in these shortlists, and have been previously identified for encoding ARGs.[40–42] Interestingly, the top phages identified based on *vipc* scores and PLS regression coefficients almost completely consist of *Staphylococcus* phages, which is puzzling as the latter make up only about 4% of all phages. This outcome may be attributed to potential interfering effects of the auxiliary (environmental) variables at play.

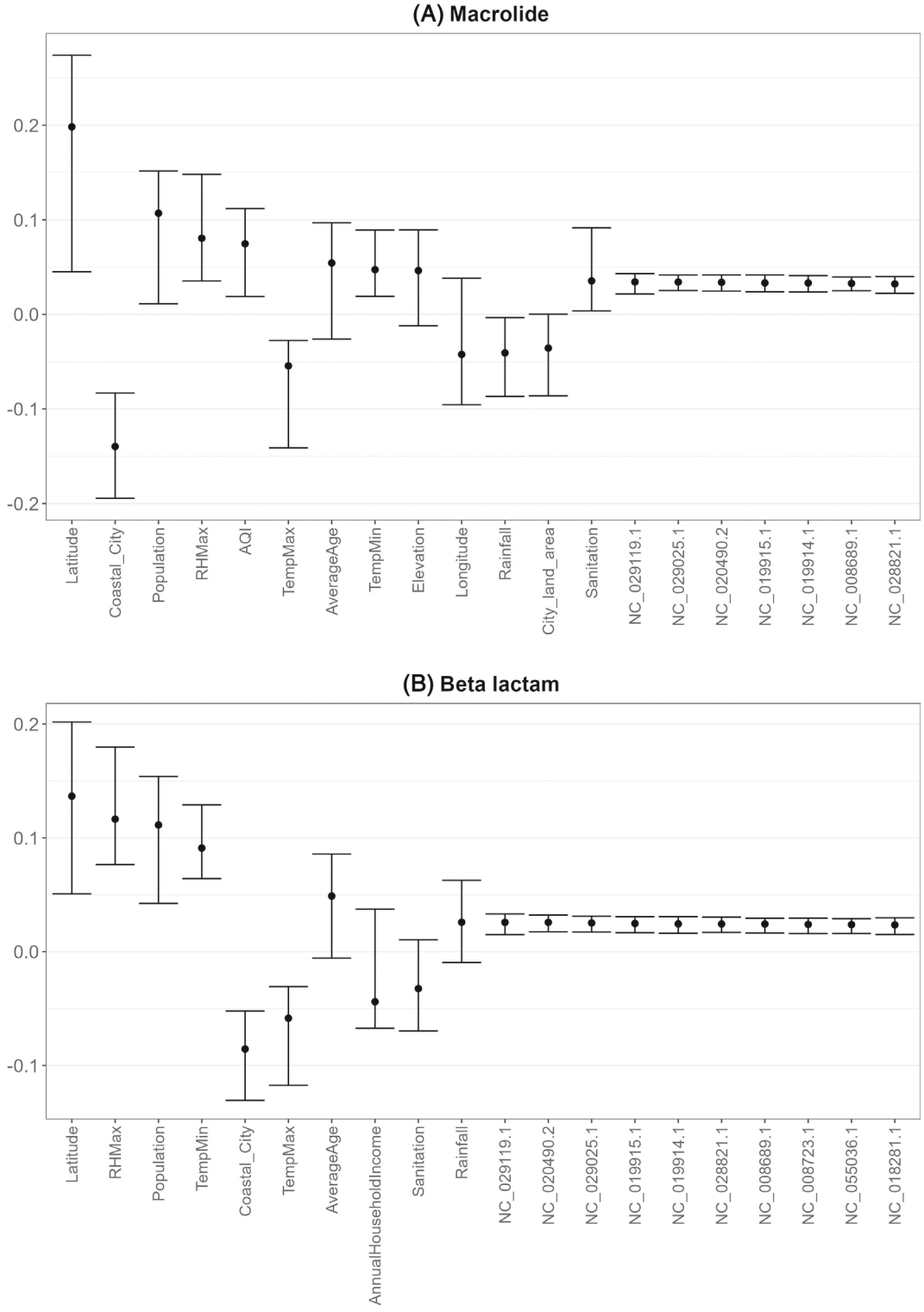

FIGURE 5 Estimated regression coefficient (and 95% confidence interval) for 20 variables (chosen by order of magnitude) from the set of all explanatory variables associated with (a) *macrolide*, (b) *beta lactam*, and (c) *tetracyline* classes of antibiotics, respectively.

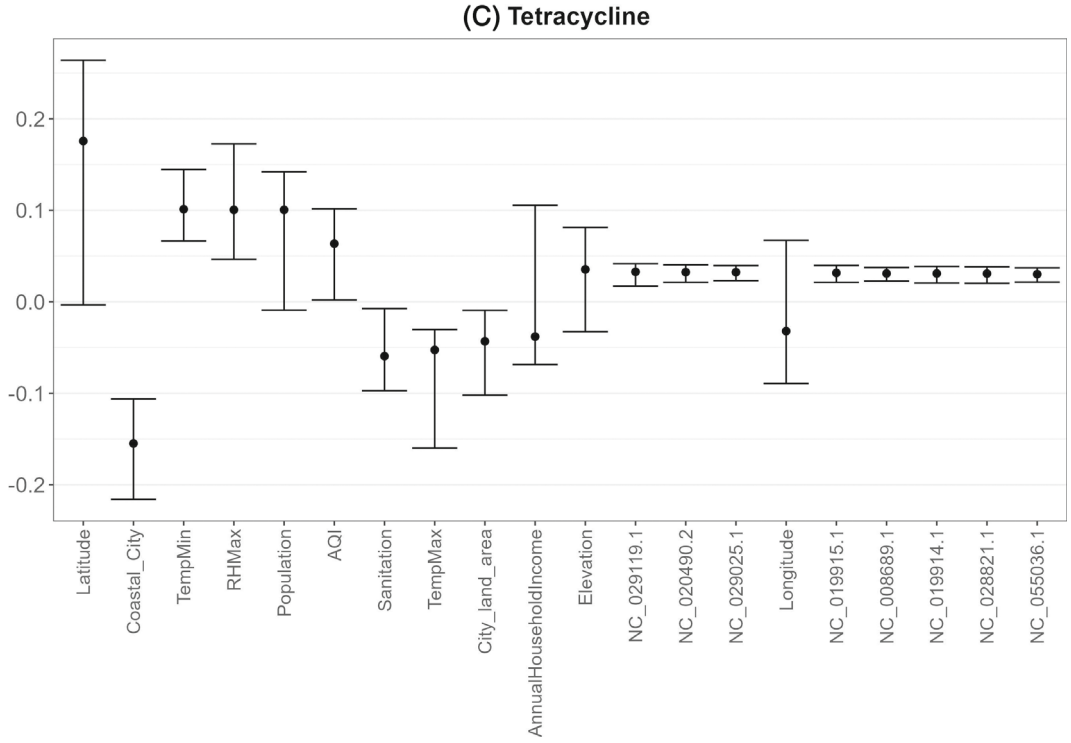

FIGURE 5 (Continued)

4.2 | Controlling for auxiliary variables to identify top phages associated with ARGs

Figure 3a validates our objective of prioritizing phages for a closer analysis. As explained in Section 3.3, we fit 17 random forest models, one for each antibiotic class, to identify phages that are strongly associated with resistomes. For each antibiotic class, the model produces a list of phages ranked according to their association with ARGs. Top importance scores from the 17 class-specific univariate random forest models are shown in Table 2, while Figure 6 summarizes the most "important" phages found for *macrolide*, *beta-lactam* and *tetracycline*, the antibiotics with the highest ARG abundances. In contrast to the previous scenario detailed in Section 4.1.3, where potential confounding variables were not adjusted for, leading to an overrepresentation of *Staphylococcus* phages, the current compilation of top phages demonstrates a more balanced distribution, encompassing a diverse array of phage types.

However, since the ordering of the phages varies across different antibiotic classes, we use rank aggregation to obtain a consensus ranking for the top k phages. This is implemented through the `RankAggreg`[18] package in R. The top 20 phages species for aggregated across all antibiotic classes are shown in Table 3. Some of the identified key phages overlap with those found in Section 4.1.3. Similar to the results obtained from the random forest analysis specific to each antibiotic class, the consensus list through random forest analysis, this list exhibits a diverse array of phages, extending beyond those exclusively targeting *Staphylococcus* bacteria. They also correspond to phages that have already been established in clinical literature as having a significant association with antibiotic resistance (elucidated in Section 4.3.1). The inclusion probabilities of all these phages are well over the threshold of 0.0168. Therefore, the inclusion of each phage in the consensus list is not due to random chance.

4.3 | Validation from clinical literature

4.3.1 | Phages

The leading phages (controlling for potential confounders) identified in Section 4.2 have a notable presence in current research, despite the limited availability of relevant literature in this area. For instance, *Propionibacterium* phage P100_A,

TABLE 2 Top phages associated with AMR based on ordered importance scores of each of the 17 antibiotic class-specific random forest models.

Antibiotic class	Top phages identified by random forest
Aminoglycoside	Cellulophaga phage phi46:3, Enterobacteria phage P7, Prochlorococcus phage P-TIM68, Escherichia phage RCS47, Sinorhizobium phage phiLM21, Propionibacterium phage PAC1, Prochlorococcus phage P-SSM2, Pseudomonas phage Lu11, Pseudomonas phage JBD25, Escherichia phage Av-05, Caulobacter phage Sansa, Acinetobacter phage vB_AbaM-IME-AB2, Pseudomonas phage EL, Prochlorococcus phage P-SSP10, Synechococcus phage ACG-2014j, Pseudomonas phage MD8, Streptococcus phage phiARI0131-2, Pseudomonas phage PollyC, Escherichia phage SH2026Stx1, Synechococcus phage S-CAM9
Beta lactam	Acinetobacter phage vB_AbaS_TRS1, Acinetobacter phage YMC11/11/R3177, Escherichia phage PA2, Lactobacillus phage SAC12B, Pseudomonas phage B3, Acinetobacter phage YMC/09/02/B1251, Enterobacteria phage YYZ-2008, Burkholderia phage phiE125, Pseudomonas phage phiPSA1, Microbacterium phage Appa, Mycobacterium phage Cambiare, Prochlorococcus phage P-SSP10, Staphylococcus phage 2638A, Mycobacterium phage Phabba, Cronobacter phage vB_CsaM_GAP31, Staphylococcus virus 108PVL, Lactobacillus phage LBR48, Bacillus phage IEBH, Staphylococcus phage phiN315, Bacillis phage phi4J1
Colistin	Enterobacter phage Tyrion, Pantoea phage vB_PagM_AAM37, Gordonia phage MelBins, Gordonia phage Phinally, Streptococcus phage A25, Enterobacter phage phiT5282H, Propionibacterium phage PHL037M02, Gordonia phage EMoore, Pseudomonas phage PMBT3, Brucella phage BiPBO1, Propionibacterium phage Keiki, Psychrobacter phage pOW20-A, Lactobacillus phage Bromius, Streptomyces phage Lannister, Acinetobacter phage YMC/09/02/B1251, Gordonia phage Blueberry, Aeromonas phage LAh_7, Streptococcus phage CHPC950, Staphylococcus phage JS01, Streptococcus phage SW1
Fosfomycin	Propionibacterium phage P100_A, Staphylococcus phage StB20-like, Salmonella phage 118970_sal3, Staphylococcus phage phiN315, Salmonella phage SPN3UB, Escherichia phage vB_EcoP-CHD5UKE1, Escherichia phage HK022, Pseudomonas phage JD18, Salmonella phage Fels-1, Staphylococcus phage StB12, Propionibacterium phage Pacnes 2012-15, Bacillus phage phBC6A52, Staphylococcus phage vB_SepS_SEP9, Enterobacteria phage YYZ-2008, Brochothrix phage NF5, Streptococcus phage VS-2018a, Lactococcus phage AM4, Staphylococcus phage Ipla5, Staphylococcus phage IME1354_01, Staphylococcus phage phiSLT
Fusidicacid	Propionibacterium phage P100_A, Staphylococcus phage StB27, Propionibacterium phage Pacnes 2012-15, Staphylococcus phage SPbeta-like, Staphylococcus virus 108PVL, Escherichia phage vB_EcoP-CHD5UKE1, Propionibacterium phage QueenBey, Staphylococcus phage StB12, Staphylococcus phage phiSA_BS1, Staphylococcus phage vB_SepS_SEP9, Staphylococcus phage StB20-like, Staphylococcus phage StB20, Staphylococcus phage S25-4, Propionibacterium phage PHL010M04, Lactobacillus phage LJ, Escherichia phage DE3, Lactococcus phage P1532, Lactobacillus phage phiAQ113, Staphylococcus phage IME1348_01, Propionibacterium phage PHL199M00
Glycopeptide	Propionibacterium phage P100_A, Staphylococcus phage JS01, Staphylococcus phage Ipla7, Escherichia phage N15, Escherichia phage RCS47, Escherichia phage Lyz12581Vzw, Escherichia phage vB_EcoP-CHD5UKE1, Staphylococcus phage CNPH82, Staphylococcus phage phi 11, Staphylococcus phage vB_SepS_SEP9, Streptococcus phage phiARI0468-4, Acinetobacter phage vB_AbaS_TRS1, Staphylococcus phage B236, Propionibacterium phage Pacnes 2012-15, Staphylococcus phage StB20, Lactobacillus phage phiAT3, Cronobacter phage ENT39118, Propionibacterium phage PHL037M02, Lactobacillus phage LfeSau, Escherichia phage HK446
Macrolide	Staphylococcus phage StB20-like, Staphylococcus phage phiN315, Prochlorococcus phage Syn1, Ralstonia phage Raharianne, Listeria phage LP-030-3, Xanthomonas phage OP2, Staphylococcus phage phiRS7, Streptomyces phage Lannister, Lactococcus phage r1t, Lactococcus phage 66901, Clostridium phage phiSM101, Staphylococcus phage IME1348_01, Enterobacteria phage P7, Staphylococcus phage IME-SA4, Klebsiella phage ST147-VIM1phi7.1, Staphylococcus phage SA12, Aurantimonas phage AmM-1, Escherichia phage DE3, Lactococcus phage phi145, Staphylococcus phage 37
Oxazolidinone	Escherichia phage DE3, Propionibacterium phage P100_A, Lactococcus phage BM13, Lactococcus phage ul36, Lactococcus phage 28201, Lactococcus phage TP901-1, Propionibacterium phage Pacnes 2012-15, Lactococcus phage 62503, Lactococcus phage bIL309, Staphylococcus phage B166, Lactococcus phage 96401, Clostridium phage phiCP39-O, Staphylococcus phage PVL, Staphylococcus phage SPbeta-like, Ralstonia phage Heva, Enterobacter phage phiT5282H, Faecalibacterium phage FP_Mushu, Lactococcus phage P1532, Lactococcus phage Q33, Pseudomonas phage VCM

(Continues)

TABLE 2 (Continued)

Antibiotic class	Top phages identified by random forest
Phenicol	Microbacterium phage Min1, Propionibacterium phage Moyashi, Propionibacterium phage SKKY, Propionibacterium phage PHL071N05, Propionibacterium phage Pacnes 2012-15, Propionibacterium phage P105, Propionibacterium phage Lauchelly, Lactobacillus phage phiAT3, Propionibacterium phage PA1-14, Propionibacterium phage PHL037M02, Propionibacterium phage PHL082M03, Erwinia phage pEp_SNUABM_08, Mycobacterium phage Myrna, Propionibacterium phage PHL092M00, Propionibacterium phage PHL095N00, Propionibacterium phage P101A, Eggerthella phage PMBT5, Propionibacterium phage PFR2, Propionibacterium phage PHL009M11, Bacillus phage Deep Blue
Pseudomonicacid	Staphylococcus phage Ipla7, Staphylococcus phage StB20, Propionibacterium phage P100_A, Staphylococcus phage StB27, Staphylococcus phage vB_SepS_SEP9, Staphylococcus phage StB20-like, Escherichia phage vB_EcoP-CHD5UKE1, Staphylococcus phage IME-SA4, Corynebacterium phage P1201, Propionibacterium phage Pirate, Staphylococcus virus 108PVL, Staphylococcus phage PVL, Propionibacterium phage Pacnes 2012-15, Staphylococcus phage MCE-2014, Propionibacterium phage ATCC29399B_T, Staphylococcus phage phiRS7, Staphylococcus phage SPbeta-like, Staphylococcus phage IME1354_01, Cellulophaga phage phi48:2, Pseudomonas phage PollyC
Quinolone	Escherichia phage 500465-1, Enterobacteria phage HK225, Escherichia phage 500465-2, Acinetobacter phage YMC/09/02/B1251, Enterobacteria phage YYZ-2008, Mycobacterium phage Kumao, Erwinia phage ENT90, Escherichia phage DE3, Mycobacterium phage Myrna, Escherichia phage HK97, Pseudomonas phage YMC12/01/R24, Propionibacterium phage P104A, Rhodococcus phage Sleepyhead, Escherichia phage RCS47, Propionibacterium phage PHL067M10, Pseudomonas phage MD8, Mycobacterium phage Enkosi, Propionibacterium phage PFR2, Brochothrix phage NF5
Rifampicin	Enterobacter phage phiEap-2, Streptococcus phage VS-2018a, Gordonia phage Phinally, Streptococcus phage CHPC577, Gordonia phage EMoore, Aeromonas phage phiO18P, Cronobacter phage vB_CsaM_GAP161, Enterobacteria phage phiP27, Propionibacterium phage PHL117M01, Staphylococcus phage IME1354_01, Propionibacterium phage SKKY, Edwardsiella phage pEt-SU, Rheinheimera phage Barba5S, Escherichia phage vB_EcoP-CHD5UKE1, Propionibacterium phage PHL116M00, Propionibacterium phage Stormborn, Rhodobacter phage RcapNL, Gordonia phage Horus, Enterobacteria phage mEp043 c-1
Sulphonamide	Pseudomonas phage YMC12/01/R24, Escherichia phage RCS47, Ralstonia phage Raharianne, Synechococcus virus S-ESS1, Lactococcus phage r1t, Propionibacterium phage P100_A, Microbacterium phage Min1, Ralstonia phage RSK1, Lactococcus phage phiL47, Escherichia phage RB69, Escherichia phage Lambda, Salmonella phage SJ46, Burkholderia phage phiE125, Lactococcus phage 28201, Propionibacterium phage PHL301M00, Propionibacterium phage PHL141N00, Staphylococcus phage StB20-like, Cellulophaga phage phiST, Klebsiella phage ST13-OXA48phi12.1, Propionibacterium phage Keiki
Tetracycline	Staphylococcus phage SPbeta-like, Lactococcus phage 28201, Synechococcus phage S-SM2, Staphylococcus phage StB20-like, Erwinia phage Ea35-70, Staphylococcus phage IME1361_01, Salmonella phage 118970_sal3, Bacillus phage G, Staphylococcus phage StB12, Lactococcus phage 4268, Staphylococcus phage Ipla5, Staphylococcus phage phiIPLA-C1C, Enterobacteria phage YYZ-2008, Staphylococcus virus 108PVL, Enterococcus phage vipetofem, Lactococcus phage TP901-1, Lactobacillus phage 521B, Propionibacterium phage ATCC29399B_T, Synechococcus phage ACG-2014j, Propionibacterium phage PFR2
Trimethoprim	Streptococcus phage IC1, Enterobacteria phage YYZ-2008, Lactobacillus phage phiAQ113, Salmonella phage SJ46, Pseudomonas phage PS-1, Escherichia phage RCS47, Propionibacterium phage P100_A, Cronobacter phage ENT47670, Acinetobacter phage WCHABP12, Salmonella phage P22, Arthrobacter phage Wheelbite, Klebsiella phage ST147-VIM1phi7.1, Pseudomonas phage YMC12/01/R24, Enterococcus phage EF62phi, Lactococcus phage AM1, Propionibacterium phage Pacnes 2012-15, Synechococcus phage S-SM2, Escherichia phage RB69, Klebsiella phage ST13-OXA48phi12.1, Escherichia phage 500465-2
Miscellaneous	Phage Gifsy-1, Enterococcus phage phiEf11, Gordonia phage Cucurbita, Enterococcus phage EFP01, Staphylococcus phage phiIPLA-C1C, Escherichia phage Sortsne, Bacillus phage phi105, Aeromonas phage LAh_7, Caulobacter phage Seuss, Staphylococcus phage 42E, Pseudomonas phage gh-1, Staphylococcus phage StB12, Lactobacillus phage Lenus, Escherichia phage Stx2 II, Staphylococcus phage SPbeta-like, Cronobacter phage phiES15, Propionibacterium phage Ouroboros, Staphylococcus phage 29, Streptococcus phage CHPC950, Streptococcus phage P0095

Note: The first 20 most "important" phages are listed for each group of antibiotics.

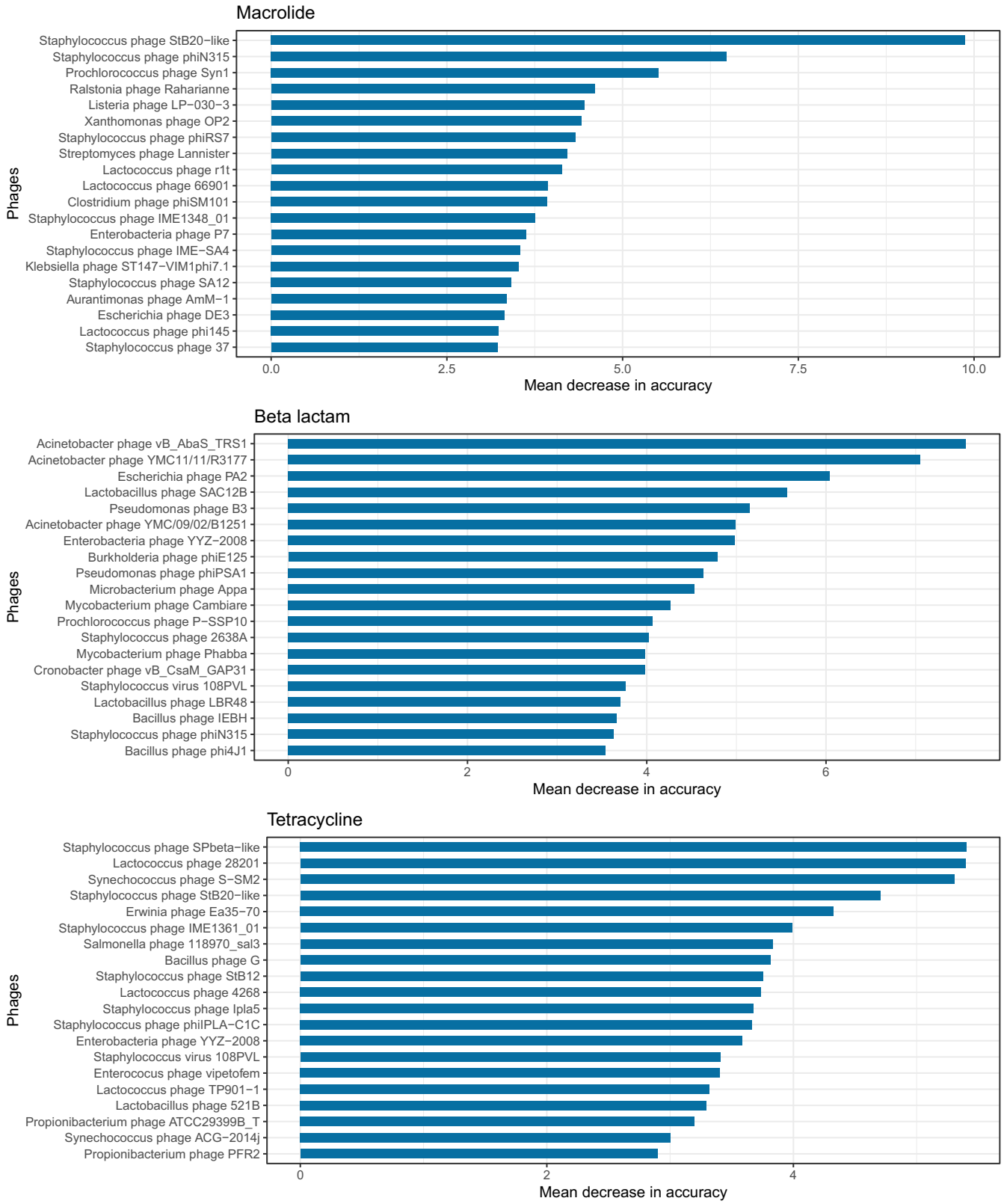

FIGURE 6 Phages with top importance scores based on the random forest for *macrolide*, *beta lactam*, and *tetracycline* antibiotic classes, where ARGs were most abundant.

TABLE 3 Rank aggregated list of top 20 phages associated with ARG dissemination.

Aggregated rank	Species	Inclusion probability
1	Propionibacterium phage P100_A	1.00
2	Staphylococcus phage StB20-like	1.00
3	Staphylococcus phage SPbeta-like	0.97
4	Escherichia phage RCS47	0.99
5	Staphylococcus phage StB12	0.69
6	Enterobacteria phage YYZ-2008	0.73
7	Staphylococcus phage Ipla7	0.95
8	Propionibacterium phage Pacnes 2012-15	0.55
9	Escherichia phage vB_EcoP-CHD5UKE1	0.57
10	Lactococcus phage 28201	0.97
11	Acinetobacter phage vB_AbaS_TRS1	0.38
12	Staphylococcus phage vB_SepS_SEP9	0.39
13	Synechococcus phage S-SM2	0.11
14	Pseudomonas phage YMC12/01/R24	0.95
15	Escherichia phage DE3	0.38
16	Microbacterium phage Min1	0.59
17	Enterobacter phage Tyrion	0.28
18	Escherichia phage 500465-1	0.41
19	Phage Gifsy-1	0.19
20	Streptococcus phage IC1	0.48

which targets clindamycin-, erythromycin-, and tetracycline-resistant *Propionibacterium acnes*,[43,44] is the top phage in the consensus list (Table 3), as well as in several of the class-specific lists (Table 2). The *Escherichia* phage RCS47, identified as a carrier of the SHV-2 Extended-Spectrum β-Lactamase from an *Escherichia coli* strain,[45] has been found to be a highly associated phage for beta lactam, fusidic acid, pseudomonic acid, and tetracycline antibiotics. Other crucial phages identified include the *Staphylococcus* virus 108PVL, which carries the Panton-Valentine leukocidin (PVL) gene linked to increased virulence and antibiotic resistance in *S. aureus*;[46] the *Staphylococcus* phage vB_SepS_SEP9, a phage of *Staphylococcus epidermidis*, which is known for its high rates of antibiotic resistance and biofilm formation, can potentially spread of ARGs among the Staphylococcal population through mechanisms like HGT;[47,48] and *Staphylococcus* phages like StB12 and StB20, targeting methicillin-resistant *Staphylococcus aureus* (MRSA). MRSA is a strain that is challenging to combat with conventional antibiotics such as glycopeptides, cephalosporins, and quinolones, all of which are associated with an increased risk of MRSA colonization.[49–52]

4.3.2 | Environmental variables

The prominent environmental variables–including sanitation, coastal location, temperature, latitude, population age, and Air Quality Index–uncovered in our analysis are substantiated by a wealth of existing research on antimicrobial resistance.

Sanitation practices play a pivotal role, as evidenced by studies indicating that poor sanitation practices can heighten exposure to antimicrobial-resistant bacteria and their genetic elements. Inadequate sanitation can lead to the proliferation of bacterial pathogens, potentially increasing antibiotic use and thereby fostering antibiotic resistance.[53] Conversely, improved sanitation practices have been associated with a reduction in the spread of antimicrobial-resistant bacteria and their genes.

Coastal areas, characterized by their proximity to water bodies and heightened human activity, have been the subject of attention. They often exhibit features of urbanization, including high population density and pollution, leading to increased antibiotic usage.[54] The discharge of untreated sewage and agricultural runoff into coastal waters further contributes to the dissemination of ARGs.

Temperature is another influential factor, with a temperature increase of 10°C linked to elevated antibiotic resistance levels in common pathogens such as *Escherichia coli*, *Klebsiella pneumoniae*, and *Staphylococcus aureus* across diverse regions.[33] This temperature-dependent effect can be attributed to the propensity of microorganisms in anaerobic or high-temperature environments to exchange genes horizontally, including ARGs.[34,35] Variations in genetic diversity, ecosystems, and human activities across latitudes contribute to this phenomenon, with central latitudes being identified as hotspots for the emergence of highly mobile genetic elements associated with AMR.[55–57]

Population age is a noteworthy proponent, with our findings aligning with existing research. Younger age groups, characterized by a lack of awareness about AMR, tend to engage in inappropriate antibiotic usage.[58]

Regarding the Air Quality Index, environmental bacteria can act as biological aerosols when they adhere to fine particulate matter. This phenomenon leads to the long-range dispersion of airborne ARGs, eventually returning to the Earth's surface through precipitation. This intricate process creates a global "environmental ARG loop".[59,60]

5 | DISCUSSION

We present a strategic method to unravel the primary drivers of AMR by focusing on the pivotal role of phages. We acknowledge that various environmental factors may exert confounding effects that are intricately intertwined with the spread of ARGs. In the initial analysis, we explore existing relationships as they stand, without addressing these potential confounders. We employ MPLS to unveil significant factors both at the block and global levels. While phages emerge as crucial explanatory variables in elucidating AMR dynamics based on the *bipc* criteria, *vipc* suggest that several environmental factors–including sanitation, coastal location, temperature, latitude, population age, and Air Quality Index–play the most pivotal roles at the global level. This observation might seem paradoxical, as phages, despite having the highest block importance, exhibit smaller global variable importances. In addition, *Staphylococcus* phages appear to be the leading phages in the orderings despite making up only a very small fraction of the available lists of phages. The involvement of environmental factors as potential confounders in the association between phages and resistomes may underlie these observations, suggesting the need to adjust resistomes for the influence of these environmental variables. This adjustment leads to informative lists that are in agreement with clinical literature (Section 4.3.1). We reach this conclusion using random forest models to identify pivotal phages for each of the 17 distinct antibiotic classes, as well as across the broader spectrum encompassing all antibiotic classes through rank aggregation. We ensure the stability of our overall list by assessing the likelihood of including phages in the consensus list by chance, comparing it against a threshold calculated using the hypergeometric distribution. All phages in the consensus list significantly exceed this threshold, affirming the robustness of our methodology.

The identification of these significant phages provides crucial insights into the driving factors behind antimicrobial resistance, laying the foundation for future targeted interventions and management strategies. Our study represents a fundamental step toward a deeper understanding of the intricate interplay between genetic and environmental factors that influence the dissemination of ARGs. While we have adjusted for auxiliary variables that could impact the distribution of ARGs, it is important to acknowledge the limitations of observational studies in making causal interpretations. Our research, guided by the metagenomic dataset from the CAMDA 2023 challenge, primarily addressed DNA phages, necessitating the exclusion of RNA phages due to the absence of metatranscriptomic data. This limitation, while inherent to our dataset, points to a potential research avenue involving RNA phages, which are likely significant in microbial ecosystems and antibiotic resistance mechanisms. To evaluate how much our metagenomic data explains resistome variation, we have detailed R^2 values for our comprehensive MPLS model in the Supplementary file. Further, our study involved normalizing absolute abundances to relative ones, with adjustments for sequencing depth, effectively tackling scale disparities. A future research opportunity lies in adjusting abundances by genome lengths. Additionally, our approach involves controlling for covariates, although some of these could potentially be confounders within the model itself. It is worth noting that there exist alternative statistical methods for controlling confounders, and future studies might find value in comparing and contrasting these methods. We hold the hope that our novel findings will be subject to validation through future

experimental research. Such advancements can provide valuable insights for the development of effective antimicrobial therapies, ultimately contributing to the mitigation of the severe consequences associated with AMR.

ACKNOWLEDGEMENTS

Our sincere thanks go to the reviewers for their invaluable feedback on our earlier draft. A special note of appreciation is extended to the second reviewer, whose insightful suggestion to expand the MPLS analysis to a broader taxonomic level of phages greatly enriched our study. The analyses presented in this paper are based on the raw WGS metagenomics data provided as part of the 2023 CAMDA Anti-Microbial Resistance Prediction and Forensics Challenge. The data is publicly available on the challenge's website. We participated in this challenge and presented our results at the 2023 Intelligent Systems for Molecular Biology (ISMB) conference.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Anti-Microbial Resistance Prediction and Forensics at http://camda2023.bioinf.jku.at/data_download.

ORCID

Shoumi Sarkar https://orcid.org/0000-0002-8511-3914
Samuel Anyaso-Samuel https://orcid.org/0000-0002-3323-2655
Peihua Qiu https://orcid.org/0000-0003-4439-9466
Somnath Datta https://orcid.org/0000-0003-4381-1842

REFERENCES

1. Guevarra RB, Hwang J, Lee H, et al. Metagenomic characterization of bacterial community and antibiotic resistance genes found in the mass transit system in Seoul. *South Korea Ecotoxicol Environ Saf*. 2022;246:114176.
2. Ryon KA, Tierney BT, Frolova A, et al. A history of the MetaSUB consortium: tracking urban microbes around the globe. *iScience*. 2022;25(11):104993.
3. Johnson K, Lin S. Call to work together on microarray data analysis. *Nature*. 2001;411(6840):885.
4. Anyaso-Samuel S, Sachdeva A, Guha S, Datta S. Metagenomic geolocation prediction using an adaptive ensemble classifier. *Front Genet*. 2021;12:642282.
5. Chappell T, Geva S, Hogan JM, Lovell D, Trotman A, Perrin D. Metagenomic geolocation using read signatures. *Front Genet*. 2022;13:643592.
6. Zhang R, Ellis D, Walker AR, Datta S. Unraveling city-specific microbial signatures and identifying sample origins for the data from CAMDA 2020 metagenomic geolocation challenge. *Front Genet*. 2021;12:659650.
7. Zhelyazkova M, Yordanova R, Mihaylov I, et al. Origin sample prediction and spatial modeling of antimicrobial resistance in metagenomic sequencing data. *Front Genet*. 2021;12:642991.
8. Tsonev S, Vassilev D. Bioinformatics and Biostatistical Models for Analysis and Prognosis of Antimicrobial Resistance. 658: 53 2023.
9. d'Herelle F. *The Bacteriophage and its Behavior*. Baltimore, Maryland: Williams & Wilkins; 1926.
10. Balcazar JL. Bacteriophages as vehicles for antibiotic resistance genes in the environment. *PLoS Pathog*. 2014;10(7):e1004219.
11. Lekunberri I, Subirats J, Borrego CM, Balcázar JL. Exploring the contribution of bacteriophages to antibiotic resistance. *Environ Pollut*. 2017;220:981-984.
12. Enault F, Briet A, Bouteille L, Roux S, Sullivan MB, Petit MA. Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *Isme J*. 2017;11(1):237-247.
13. Manaia C, Donner E, Vaz-Moreira I, Hong P. Antibiotic resistance, sanitation, and public health. In: Hong P, ed. *Antibiotic Resistance in the Environment. The Handbook of Environmental Chemistry*. Vol 91. Cham: Springer; 2020:189-216.
14. Zheng D, Yin G, Liu M, et al. A systematic review of antibiotics and antibiotic resistance genes in estuarine and coastal environments. *Sci Total Environ*. 2021;777:146009.
15. Burnham JP. Climate change and antibiotic resistance: a deadly combination. *Ther Adv Infect Dis*. 2021;8:2049936121991374.
16. Bougeard S, Qannari EM, Lupo C, Hanafi M. From multiblock partial least squares to multiblock redundancy analysis. *A Contin Approach Informat*. 2011;22(1):11-26.
17. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
18. Pihur V, Datta S, Datta S. RankAggreg, an R package for weighted rank aggregation. *BMC Bioinform*. 2009;10:1-10.
19. Galhano BS, Ferrari RG, Panzenhagen P, J, Conte-Junior CA. Antimicrobial resistance gene detection methods for bacteria in animal-based foods: a brief review of highlights and advantages. *Microorganisms*. 2021;9(5):923.
20. Andrews S. FastQC. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
21. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048.
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120.

23. McIver LJ, Abu-Ali G, Franzosa EA, et al. bioBakery: a meta'omic analysis environment. *Bioinformatics*. 2018;34(7):1235-1237.
24. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9(4):357-359.
25. Bushnell B. *BBMap: a Fast, Accurate, Splice-Aware Aligner*. Tech. rep. Berkeley, CA (United States): Lawrence Berkeley National Lab.(LBNL); 2014.
26. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745.
27. Clausen PT, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinform*. 2018;19:1-8.
28. Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*. 2012;67(11):2640-2644.
29. Paulson JN, Olson ND, Braccia DJ, et al. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. 2013 Bioconductor package.
30. Bougeard S, Dray S. Supervised multiblock analysis in R with the ade4 package. *J Stat Softw*. 2018;86:1-17.
31. Calle ML, Urrea V. Stability of random Forest importance measures. *Brief Bioinform*. 2011;12(1):86-89.
32. Strange JE, Leekitcharoenphon P, Møller FD, Aarestrup FM. Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. *Sci Rep*. 2021;11(1):1-11.
33. MacFadden DR, McGough SF, Fisman D, Santillana M, Brownstein JS. Antibiotic resistance increases with local temperature. *Nat Clim Change*. 2018;8(6):510-514.
34. Fuchsman CA, Collins RE, Rocap G, Brazelton WJ. Effect of the environment on horizontal gene transfer between bacteria and archaea. *PeerJ*. 2017;5:e3865.
35. Banerjee G, Ray AK, Kumar R. Effect of temperature on lateral gene transfer efficiency of multi-antibiotics resistant bacterium. *Alcaligenes Faecalis Sains Malays*. 2016;45:909-914.
36. Pawlowski AC, Stogios PJ, Koteva K, et al. The evolution of substrate discrimination in macrolide antibiotic resistance enzymes. *Nat Commun*. 2018;9(1):112.
37. Miklasińska-Majdanik M. Mechanisms of resistance to macrolide antibiotics among Staphylococcus aureus. *Antibiotics*. 2021;10(11):1406.
38. Babic M, Hujer AM, Bonomo RA. What's new in antibiotic resistance? *Focus on Beta-Lactamases Drug Resist Updat*. 2006;9(3):142-156.
39. Grossman TH. Tetracycline antibiotics and resistance. *Cold Spring Harb Perspect Med*. 2016;6(4):a025387.
40. Benz F. *Evolutionary and Genetic Drivers of and Barriers to the Spread of Antibiotic-Resistance Plasmids*. PhD thesis. Zurich, Switzerland: ETH Zurich; 2022.
41. Martins BTF, Meirelles dJL, Omori WP, et al. Comparative genomics and antibiotic resistance of Yersinia enterocolitica obtained from a pork production chain and human clinical cases in Brazil. *Int Food Res J*. 2022;152:110917.
42. Chang SC, Lin LC, Lu JJ. Comparative genomic analyses reveal potential factors responsible for the ST6 oxacillin-resistant staphylococcus lugdunensis endemic in a hospital. *Front Microbiol*. 2021;12:3546.
43. Marinelli LJ, Fitz-Gibbon S, Hayes C, et al. Propionibacterium acnes bacteriophages display limited genetic diversity and broad killing activity against bacterial skin isolates. *MBio*. 2012;3(5):10-1128.
44. Eady E, Gloor M, Leyden J. Propionibacterium acnes resistance: a worldwide problem. *Dermatology*. 2003;206(1):54-56.
45. Billard-Pomares T, Fouteau S, Jacquet ME, et al. Characterization of a P1-like bacteriophage carrying an SHV-2 extended-spectrum β-lactamase from an Escherichia coli strain. *Antimicrob Agents Chemother*. 2014;58(11):6550-6557.
46. Diep BA, Carleton HA, Chang RF, Sensabaugh GF, Perdreau-Remington F. Roles of 34 virulence genes in the evolution of hospital-and community-associated strains of methicillin-resistant Staphylococcus aureus. *J Infect Dis*. 2006;193(11):1495-1503.
47. Melo LD, Sillankorva S, Ackermann HW, Kropinski AM, Azeredo J, Cerca N. Characterization of Staphylococcus epidermidis phage vB_SepS_SEP9–a unique member of the Siphoviridae family. *Res Microbiol*. 2014;165(8):679-685.
48. Fišarová L, Botka T, Du X, et al. Staphylococcus epidermidis phages transduce antimicrobial resistance plasmids and mobilize chromosomal islands. *Msphere*. 2021;6(3):10-1128.
49. Maleki F, Hadadi MH, Rezaei F, Mohamadi HR, Khosravi A, Nasser A. Classification and replication mechanism of staphylococcus phage. *Biosci Biotechnol Res Asia*. 2015;12(1):481-486.
50. Zyl vLJ, Abrahams Y, Stander EA, et al. Novel phages of healthy skin metaviromes from South Africa. *Sci Rep*. 2018;8(1):12265.
51. Tacconelli E, De Angelis G, Cataldo MA, Pozzi E, Cauda R. Does antibiotic exposure increase the risk of methicillin-resistant Staphylococcus aureus (MRSA) isolation? A systematic review and meta-analysis. *J Antimicrob Chemother*. 2008;61(1):26-38.
52. Muto CA, Jernigan JA, Ostrowsky BE, et al. SHEA guideline for preventing nosocomial transmission of multidrug-resistant strains of Staphylococcus aureus and enterococcus. *Infect Control Hosp Epidemiol*. 2003;24(5):362-386.
53. Bürgmann H, Frigon D, Gaze WH, et al. Water and sanitation: an essential battlefront in the war on antimicrobial resistance. *FEMS Microbiol Ecol*. 2018;94(9):fiy101.
54. Fresia P, Antelo V, Salazar C, et al. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. *Microbiome*. 2019;7(1):1-9.
55. Liu YY, Wang Y, Walsh TR, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect Dis*. 2016;16(2):161-168.
56. Poirel L, Potron A, Nordmann P. OXA-48-like carbapenemases: the phantom menace. *J Antimicrob Chemother*. 2012;67(7):1597-1606.
57. Walsh TR, Weeks J, Livermore DM, Toleman MA. Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study. *Lancet Infect Dis*. 2011;11(5):355-362.

58. Guo H, Hildon ZJL, Lye DCB, Straughan PT, Chow A. The associations between poor antibiotic and antimicrobial resistance knowledge and inappropriate antibiotic use in the general population are modified by age. *Antibiotics*. 2021;11(1):47.
59. Zhu G, Wang X, Yang T, et al. Air pollution could drive global dissemination of antibiotic resistance genes. *Isme J*. 2021;15(1):270-281.
60. Zhou XY, Li H, Zhang YS, Su JQ. City-scale distribution of airborne antibiotic resistance genes. *Sci Total Environ*. 2023;856:159176.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sarkar S, Anyaso-Samuel S, Qiu P, Datta S. Multiblock partial least squares and rank aggregation: Applications to detection of bacteriophages associated with antimicrobial resistance in the presence of potential confounding factors. *Statistics in Medicine*. 2024;43(13):2527-2546. doi: 10.1002/sim.10058