Unraveling the origins of mobile antibiotic resistance genes using random forest classification of large-scale genomic data

Citation for the original published paper (version of record):
Ebmeyer, S., Kristiansson, E., Larsson, D. (2025). Unraveling the origins of mobile antibiotic resistance genes using random forest classification
of large-scale genomic data. Environment International, 198.
http://dx.doi.org/10.1016/j.envint.2025.109374

Environment International 198 (2025) 109374

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

Full length article

Unraveling the origins of mobile antibiotic resistance genes using random forest classification of large-scale genomic data

Stefan Ebmeyer [a,b], Erik Kristiansson [a,c], D. G. Joakim Larsson [a,b,*]

[a] Center for Antibiotic Resistance Research in Gothenburg (CARe), SE-40530 Göteborg, Sweden
[b] Department of Infectious Diseases, Institute of Biomedicine, University of Gothenburg, SE-41346 Göteborg, Sweden
[c] Department of Mathematical Sciences, Chalmers University of Technology and the University of Gothenburg, SE-41296 Göteborg, Sweden

ARTICLE INFO

Keywords:
Antimicrobial resistance
AMR
Antibiotic resistance
Evolution
Wastewater
Antibiotic resistance genes

ABSTRACT

Understanding in which environments and under what conditions chromosomal antibiotic resistance genes (ARGs) acquire increased mobility is crucial to effectively mitigate their emergence in and dissemination among pathogens. In order to identify the conditions and environments facilitating these processes, it is valuable to know from which bacterial species mobile ARGs were mobilized initially, before their dissemination to other species. In this study, we used data generated from > 1.5 million publicly available bacterial genome assemblies to train a random forest classifier to identify the origins of mobile genes. Analysis of the models' predictions revealed the previously unknown origins of 12 mobile ARG groups, which confer resistance to 4 different classes of antibiotics. This included ARGs conferring resistance to tetracyclines, an antibiotic class for which, to the best of our knowledge, no recent origins of ARGs have previously been convincingly demonstrated. All identified origin species in this study are known opportunistic pathogens, and some are the origin of multiple mobile ARGs. An analysis of public metagenomes from different sources indicates that most of the origin species are particularly abundant in municipal wastewaters, a few were highly abundant in animal feces and three were most common in environments polluted with waste from antibiotic manufacturing. This study highlights environments where these origin species thrive and where there is a need for limiting antibiotic selection pressures.

1. Introduction

Resistance of bacterial pathogens to treatment with antibiotics is a fundamental threat to modern health care. Apart from being intrinsically resistant to certain antibiotics, bacteria can acquire resistance determinants through mutations of preexisting DNA or horizontal gene transfer. By association with mobile genetic elements (MGEs) such as plasmids or insertion sequences (IS), mobile ARGs can move horizontally between bacterial cells, and may confer their host with resistance to antibiotics of any class, even for antibiotics that today are considered as a 'last resort'. During recent years, the number of described mobile ARGs has steadily increased, and novel mobile resistance genes are described frequently (Lund et al., 2022).

Where these mobile ARGs come from in the first place, and how they make their way into human pathogens is crucial in order to understand where to focus efforts that aid the mitigation of the emergence of mobile ARGs that to date have not been observed in clinical settings. A widespread hypothesis is the 'producer hypothesis', which attributes the presence of mobile ARGs in Gram-negative pathogens to transfer events of genes from antibiotic-producing bacteria, such as streptomyces or actinomyces (Jiang et al., 2017). However, sequence identities between mobile ARGs and potential progenitors in antibiotic producers are usually relatively low, ruling out recent transfer events.

The original hosts (the species from which the ARG has been mobilized prior to its dissemination among pathogens) of several notorious ARGs that are today widely circulating in Gram-negative pathogens, such as the CTX-M beta-lactamases Humeniuk et al., 2002; Poirel et al., 2002; the quinolone resistance determinant QnrA (Poirel et al., 2005), the colistin resistance gene MCR-2 (Poirel et al., 2017) and more, have been identified in the past two decades. While there is some evidence that the presence of chromosomal ARGs in some origin species are the results of ancient transfer events (Ebmeyer et al., 2018), in most cases, the chromosomal ARGs and parts of the adjacent sequences in the respective origin are nearly identical to the mobile ARG loci in nucleotide sequence identity, suggesting evolutionarily recent transfer events prior to their dissemination. A *meta*-analysis of the to-date proposed

* Corresponding author.
E-mail address: joakim.larsson@fysiologi.gu.se (D.G.J. Larsson).

https://doi.org/10.1016/j.envint.2025.109374
Received 9 July 2024; Received in revised form 9 January 2025; Accepted 12 March 2025
Available online 15 March 2025

origin species and the respective mobilized ARGs suggested criteria by which the recent origins of these ARGs can be identifiable through detailed comparison of their genomic context. The criteria were: 1. *Absence/presence of mobile genetic elements associated with the ARG in the origin/host of the mobile ARG*, 2. *Conserved synteny between mobile and chromosomal ARG loci*, 3. *High nucleotide sequence identity between the origin and mobile ARG loci* (>=95 %), and 4. *Presence of similar (but divergent in nucleotide identity) ARG loci in other members of the origin genus*. The study concluded that there was enough evidence to assign taxonomic origins to about 30 groups of ARGs (Ebmeyer et al., 2021), and, interestingly, all of these origin taxa were Gram-negative Pseudomonadota (previously Proteobacteria), none of them known antibiotic-producers. Though these findings do not exclude other bacterial phyla as the sources of mobile resistance genes, they highlight Pseudomonadota as an important source of clinically relevant, mobile ARGs.

To date, the origins are only known for of about ~ 4 % of mobile ARGs, with the origins of mobile ARGs for some antibiotic classes (e.g. tetracyclines) completely unknown. Utilizing the rapidly increasing amount of genomic data available in public databases, such as the NCBI assembly database (containing > 1.5 million genome assemblies at time of writing), it is likely that more origins could be identified using the patterns described above. However, visual comparison of ARG loci from hundreds of thousands of genomes using available comparative genomic tools is tedious and time consuming at best.

Machine learning algorithms are efficient tools for processing large amounts of data and identifying underlying patterns, and are able to effectively leverage biological data to generate insights (Lund et al., 2023). Decision tree classifiers classify datapoints based on criteria defined by the paths from the root of the tree to the leaves. The dataset is recursively split into subsets according to the value of the feature that maximizes homogeneity within the subsets according to some criterion (i.e. information gain, impurity etc.). The split datasets then represent branches of the tree, and are further split according to the next feature until a stopping criterion is met. Single decision trees however are prone to overfitting, which leads to suboptimal performance on previously unseen data. A Random Forest, constituted by an ensemble of multiple decision trees, alleviates this problem through random selection of feature subsets when splitting the dataset, so each tree encounters different sets of features at the split points. Once all trees in the ensemble are trained, final class predictions are assigned to an entry through majority voting, meaning the entry is assigned the class that the majority of trees assigned it to. Random forests are, furthermore, not dependent on the assumptions of a parametric distribution and are able to handle non-linear relationships between dependent and independent variables (Schonlau and Zou, 2020).

Using random forest classification on data generated from the > 1.5 million genome assemblies publicly available at the NCBI Assembly database (Kitts et al., 2016), enabled the identification of the origins of 10 groups of mobile ARGs that to the best of our knowledge were previously unknown. All identified origins were opportunistic pathogens, and an analysis of public metagenomes identified wastewater, domesticated animal feces and antibiotic polluted freshwaters as environments in which certain of these origin species were particularly abundant. As the model was trained exclusively on Pseudomonadota origins (as these are to the best of our knowledge the only well documented origins of mobile ARGs), this approach might miss origin species from bacterial phyla other than Pseudomonadota. Nevertheless, these results further highlight the role of Pseudomonadota as the recent origins of mobile antibiotic resistance genes.

2. Material and methods

2.1. Sequence processing

2.1.1. ARG identification in assemblies

All available bacterial assemblies (n = 1,549,614, April 2023) were downloaded from the NCBI assembly database. A custom database of mobile ARGs (created through searching the CARD databases protein homolog model Jia et al., 2017; v3.0.5, entries against the ResFinder database v2.0.3 (Zankari et al., 2012) using DIAMOND (Buchfink et al., 2014) at 95 % sequence identity) was created to obtain a database containing the sequences of mobile ARGs (as present in ResFinder) with orderly CARD annotations. Special characters in ARG names such as brackets or hyphens were removed from gene names, (as software used later in the analysis has difficulties handling these) GEnView (Ebmeyer et al., 2022) was then used to search the assemblies against the ARG database (80 % identity cutoff) The resulting database contained all hits to ARG-like genes with 10 kb upstream and downstream extracted from the locus of the identified ARG, as well as the annotated open reading frames (ORFs) generated by GEnView (n > 10 million).

2.1.2. Creation of ARG groups and sequence filtering

Many mobile ARGs are part of a family of closely related genes, sometimes containing dozens of variants differing by just a single amino acid. In order to summarize all sequences for closely related ARGs under a single name, all ARG loci were grouped (based on clustering of the ARG sequences at 90 % AA identity using cd-hit v4.8.1 (Li and Godzik, 2006), such that sequences containing the ARG, or one within a 90 % amino acid identity range, were grouped together, hereafter referred to as ARG group. In the following steps, redundant or non-informative ARG containing sequences were filtered in order to significantly reduce the number of sequences to processable levels for downstream analysis. The loci in each previously identified ARG group were clustered at 95 % amino acid identity using cd-hit-est, and only the centroids of each cluster were used for further analysis. As not to discard sequences that may represent potential origins, if a cluster included sequences of < 3 distinct bacterial genera and had an average sequence length of > 17kbp up to seven additional sequences were randomly selected from the cluster and used for further analysis (As clusters containing an origin species were expected to contain long sequences from ideally only one genus (the origin), so we allow for misclassification of sequences in public databases by using 3 genera as a cutoff here). If after clustering the number of sequences per ARG group exceeded 5000 sequences, the sequences were filtered further according to the following procedure: Using diamond blastx, the sequences were searched against the mobileOG database (Brown et al., 2022) (−id 90, −scov 90, excluding phage associated entries) in order to identify putative mobile elements within the sequences. All sequences with hits were marked as potentially mobile. If the total number of centroid sequences for a group exceeded 10,000, all sequences shorter than 10kbp were marked as potentially mobile as well. This was done to reduce the number of sequences to process later on – though this classification is imperfect, short sequences may more likely be derived from mobile elements than from chromosomal sequences, due to the repetitive nature of many MGEs found in gram-negatives, which can disrupt the assembly process and result in short contigs. If the total number of sequences classified as mobile within an ARG group exceeded 300, 300 sequences were randomly sampled and utilized for further analysis together with all longer sequences without any hits for mobile genetic elements. This filtering step based on the classification of sequence mobility was necessary, as the preliminary analysis showed that multiple ARG groups contained thousands of sequences classified as mobile, the majority of which were identical or nearly identical (i.e genetic rearrangements of the same MGE, incorporations of single novel ARGs or transposases in the same MGE, etc.). Through filtering out the majority of these sequences, we retain the information within the non-filtered mobile sequences together with the sequences classified as chromosomal, and significantly decrease the cpu hours necessary for the analysis (note that all filtered out sequences are still available for the analysis of our classifiers prediction, described in section 2.5).

2.1.3. Alignment block creation

In the next step, the sequences within an ARG group were further grouped into subgroups, so called 'alignment blocks', to identify distinct subgroups of sequences where the ARG was present in similar genetic contexts. Our aim during this step was to identify blocks of sequences forming long alignments with each other, as chromosomal sequences without MGEs (such as the origins of ARGs) often are part of longer contigs. Blastn (Altschul et al., 1990) v2.14.0 (−perc_id 70 −strand both −task blastn) was used to align all sequences in a group to one another, and alignments that overlapped completely were removed. The lengths of HSPs (high-scoring pairs) between two sequences were then added to obtain the total length of the alignment. Then, all aligned sequence pairs were iterated. For each pair, all alignments within the group containing one of the two sequences where the query sequence was at least 15kbp long, and the alignment spanned at least 13kbp, the respective sequences were grouped into the same block (subgroup). Sequences that not fulfilled the criteria to be part of any block, were grouped into a separate block containing these 'ungroupable' sequences, as to not lose taxonomic information provided by these sequences. All sequences in all blocks were then searched for mobile genetic elements as described during the filtering step. As preliminary analyses showed that many shorter sequences were falsely classified as non-mobile, all sequences shorter than 10kbp were marked as mobile. For each block, the number of sequences, unique species, unique genera, percent identity range to the reference ARG and percentage of sequences containing at least one MGE were calculated. Thus, several blocks of similar sequences were formed for each ARG group.

2.2. Data labeling

Alignment blocks for ARG groups where the origin has been verified in Ebmeyer et al. 2021 were then labeled manually as either containing the origin, or not containing the origin, following the process described next. As, in some cases, genomes of the origin species can also carry mobilized variants of the chromosomal gene, whether a block for a ARG group with described origin was labeled as origin depended on the percent identity range to the reference ARG within the block (origin label: range between 93 and 100 % AA identity), the presence/absence of sequences classified as mobile/chromosomal (origin label: both mobile and chromosomal blocks should be among all alignment blocks for the ARG group), and the number of sequences from different bacterial genera within the block (origin label: max. two genera, including the origin genus, to account for taxonomic misclassifications). In addition to the blocks that did not contain the origin species from ARG groups with known origin, randomly chosen ARG groups with no described origin in Ebmeyer et al. 2021 were labeled as 'no origin', if all blocks contained exclusively genes classified as mobile or chromosomal within a single genus (supplementary file 3). This last step was included to increase the variety of the negative dataset.

2.3. Feature generation and selection

To serve as input for the classifier, the following features were calculated for each block (subgroup of sequences) within an ARG: Minimum AA percent identity to reference ARG, maximum AA percent identity to reference ARG, AA identity range towards reference within alignment block, presence of both chromosomal and mobile blocks in all alignment blocks of the respective ARG group, average taxonomic distance within alignment block (where taxonomic distance is expressed as number of taxonomic levels until common taxonomic level between each two sequences in a given block – e.g average taxonomic distance for a block containg a sequence from *Klebsiella pneumoniae, Klebsiella varicola* and *Escherichia coli* each would be $\frac{1+2+2}{3} = 1.67$), the number of genera and species within a block, percentage of mobile sequences within a block and the mean sequence length within a block. Correlation analysis revealed moderate correlations between the number of genera/species and the mean taxonomic distance. As mean taxonomic distance was generated as a measure of taxonomic diversity within a block, that is more robust to misclassifications of the original sequence hosts, the 'number of genera/species' features were excluded from further analysis. Furthermore, features with gini importance (as calculated by scikit-learn.feature_importances_) below 0.05 (of a total of 1 for the sum of gini importances for all features), were dropped due to their low predictive power. This removed the'mean sequence length within block' feature from the feature list. The features thus selected to train the final classifier were minimum AA percent identity towards reference, maximum AA percent identity towards reference, percent of mobile sequences in block, mean taxonomic distance within block and presence of both mobile and chromosomal blocks in all blocks for the respective ARG group.

2.4. Classifier training and origin prediction

A random forest classifier was trained on the labelled data (as described in 2.2) (n = 613) using Python v3.8.13 and scikit-learn v.1.2.2 (Pedregosa et al., 2018). As the dataset was imbalanced with regards to the labels ($n_{origin} = 51$, $n_{non\text{-}origin} = 562$), the class_weight parameter was set to 'balanced_subsample', which adjusts the weights of the training data based on the class (label) proportion for each trained tree in the ensemble, increasing the importance of the minority label (origin) during the splitting process. Classifier performance was assessed through leave-one-out cross-validation.

The trained classifier was then used to predict the unlabeled data in order to identify potential blocks containing the origin species of a gene group, and all unique ARG groups from the positive (i.e origin) prediction were analyzed further.

2.5. Manual assessment of positive predictions

Sequences containing the respective ARG, were extracted as described in section 2.1. Metadata and visualizations were created using GEnView, and the respective contexts and nucleotide similarities were compared manually using blastn. After visualization, ARG groups for which the only 'mobile' sequences were as short as to only contain the respective ARG-like sequence, without other evidence of mobility (e.g transposable elements, plasmid associated genes), were discarded due to uncertainty about their actual mobility. For a species to be assigned as the origin of a mobile ARG, the majority of criteria described in section 1 had to be fulfilled, and the nucleotide sequence identity between the mobile and putatively chromosomal locus had to be at least 95 %. In cases where phylogenies were created (see results section and supplementary file 2), all genomes of the respective genus/species were downloaded separately from NCBI assembly. Marker gene protein sequences (RpoB, genbank accession CAA23625.1 and DnaK, genbank accession NP_414555.1) were identified in those genomes using diamond blastx (−-id 70 –subject-cover 70 –max-target-seqs 1 –ultra-sensitive) and aligned using mafft (−-auto −-reorder). Phylogenies were created using Fasttree v 2.1.11 and visualized using the ete3 python library (Huerta-Cepas et al., 2016). Final visualizations sequence comparisons for representative sequences and alignments were automatized using the pyGenomeViz Python library (Shimoyama, 2024). Sequence annotations in the figures were derived from diamond blastx (−-id 90 –subject-cover –0.9 more_sensitive) against the NCBI protein and ISFinder databases (NCBI Resource Coordinators, 2013; Siguier et al., 2006). Global average nucleotide identities were calculated using ANI-calculator_v1 (Varghese et al., 2015). Preliminary analysis of the predicted origins showed that in many cases ARG loci in Gram-positive bacteria were not clearly identifiable as mobile or non-mobile in the manual verification step. To date known ARG origins (which are exclusively Pseudomonadota species) are clearly identifiable due to the IS-mediated 'copy and paste' mobilization mechanism. These patterns

between the predicted Gram-positive origins and the respective mobile genes could not be clearly identified in the data – hinting at potentially different recombination and mobilization mechanisms used in Gram-positive and Gram-negative bacteria. To be certain that the origins reported in this manuscript are strongly supported by our data and the previously defined criteria, we excluded ARG groups for which at least 40 % of the host genera were Gram-positives.

2.6. Metagenomic analysis

To investigate the presence of the origin species identified in this study in different environments, the approach presented in Berglund et al. 2023 (Berglund et al., 2023) was employed. A subset of the short-read metagenomes (read length 75–250 bp) used in Berglund et al. 2023 from different environments (14 projects, 7 environment types, 1697 samples), was downloaded (supplementary Table 2), with the goal of representing different environments that have been deemed important for emergence and transmission of mobile ARGs. Only samples with at least 20 million fragments were included. The Samples were quality controlled and adapters were removed using BBduk v38.86 (BBMap software) (Bushnell, xxxx). Reads with phred score < 20 were removed. For taxonomic classification of the reads, kraken2 (Wood et al., 2019), a tool specifically developed to assign bacterial taxonomy based on short reads, was used. A custom kraken2 database, containing kraken2's standard bacterial reference database (from which the here identified origin species were removed, see data availability), the univec database, the viral reference database and a custom plasmid database (containing all plasmids available at NCBI) were created. To ensure accurate identification of the origin species, all complete genomes of each origin species identified in this study, containing the respective resistance gene with > 90 % nucleotide similarity were downloaded and compared with respect to gANI (which describes the average nucleotide identity between all homologous genes between two genomes, using ANIcalculator_v1). Genomes that were dissimilar (>=4% dissimilarity) to the majority of the other genomes of the respective species were excluded from the database. As *Providencia stuartii* and *Providencia thailandensis* were indistinguishable based on gANI, *Providencia thailandensis* genomes were excluded from the database (to not diminish the amount of reads potentially assigned to *P. stuartii*). Subsequently, all assemblies for established species in the same genus (excluding uncultivated, unclassified or genus genomes lacking a species classification) as the origin species were downloaded and added to the database. All contigs < 5,000 bp were removed from the previously described assemblies in order to avoid incorporating and misclassifying small plasmids as origin species. To assess false positive rates (FPR) for each origin, the database was tested using kraken2 with confidence value of 0.3 against simulated paired short read metagenomes (read length 150 bp, created using art_illumina –ss HS25 –f 1 –i infile –o outfile –l 150 –-sdev 35 –paired) from randomly chosen genomes of closely related species only (one genome per species, origin species not present). In cases where closely related species were misclassified as origin species, more genomes of the respective non-origin species were assessed via gANI and, if closely matching (gANI>=98 %) other genomes of that species, added to the database (if more genomes were available) in order to increase the resolution between origin and non-origin species. The accession numbers of genomes included in the database is given in supplementary file 2 and the estimated false positive rates for each origin species are provided in supplementary Table 3.

The metagenomes were searched against the database using kraken2 with a confidence score of 0.3. Species hits were normalized by the number of reads classified as originating from bacteria multiplied by one million (bacterial reads per million). Samples containing less than 50,000 reads classified as bacterial were excluded. To assess the fraction of samples per environment in which each origin species were present, rarefication of the metagenomes was simulated through drawing 50,355 (sample with lowest amount of reads classified as bacterial) reads from a hypergeometric distribution (sampling without replacement), repeated 1000 times for each origin species and sample (Fig. 5). All described analyses in the method section were conducted through custom python scripts, which are publicly available (see data availability statement).

3. Results

3.1. Sequence processing, model performance and model limitations

Processing of the > 1.5 million bacterial assemblies from the NCBI assembly database yielded ~ 5.3 million sequences containing ARGs or ARG-like sequences, which were concentrated into 11,567 blocks of aligned sequences, from 612 distinct groups of ARGs (supplementary file 3). Of these, 51 were labeled as containing an origin of the specific ARG group and 552 were labeled as not containing any origin. The remaining 10,954 blocks remained unlabelled (Fig. 1).

The labeled data were used to train a random forest classifier, as described in section 2.4. Leave-one-out cross-validation was performed in order to assess the best possible performance of the classifier. The balanced accuracy score, as a measure of how well both labels are predicted, was 0.93, while precision and recall were 0.84 and 0.88, respectively (see confusion matrix, supplementary Table 4). This suggests that the classifier's positive (origin) predictions were correct 84 % of the time, but only 88 % of all origins were classified as such. For the remaining unlabeled 10,954 ARG blocks (representing 525 ARG groups), 87 were predicted as origins and were selected for further analysis (37 origins predicted in Gram-positives were removed, see section 2.4). Out of these, we managed to manually confirm (as described in section 2.4.1) the recent origins of 12 mobile groups of ARGs.

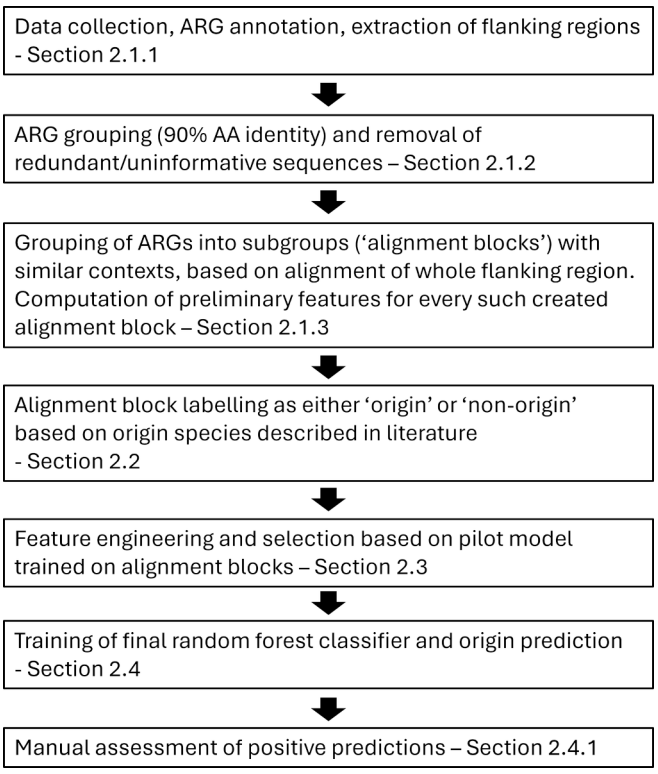

Fig. 1. Schematic overview over the process of creating classifier input features from genomic data. The details of every step are described in the respective section referenced in each box.

3.2. Identified origins

The 12 groups of mobile ARGs for which a recent origin has been identified confered resistance to four different classes of antibiotics: aminoglycosides, beta-lactams, chloramphenicol and tetracyclines. Literature research showed that three of the groups for which an origin was predicted and that passed the manual analysis, *bla*OXA-427, *bla*HER and *bla*KLUC-5, had been reported chromosomally in the respective origin species (*Aeromonas media, Atlantibacter hermannii* – previously *Escherichia hermannii* and *Kluyvera cryocrescens*, respectively (Beauchef et al., 2003; Bogaerts et al., 2017; Decousser et al., 2001) before (Supplementary Figs. 1-3). The *bla*CdiA beta-lactamase gene had previously been described chromosomally in *Citrobacter amalonaticus* (Underwood and Avison), but not in mobile contexts, which we report here. The origins of the remaining eight groups, AAC(6')-Ian, CatI, CatII, CatIII, Tet(B), Tet(D), Tet(H) and Tet(59), with origins in *Pseudochrobactrum asaccharolyticum, Atlantibacter hermanii, Morganella morganii, Providencia stuartii, Providencia stuartii, Morganella morganii, Proteus terrae* and *Providencia rettgeri*, respectively (Table 1, Supplementary Figs. 1-17), have to the best of our knowledge not been previously reported. To illustrate the principles applied in this study, the identification of the origin of the widely disseminated tetracycline resistance gene tet(B) will be demonstrated below (Fig. 2). Due to the large number of figures and detail needed to present the evidence for every single origin identified in this study, detailed analyses and figures on each case are supplied in the supplementary material, whereas the results of these analyses are summarized in Table 1.

The tetracycline resistance gene *tet(B)* encodes a tetracycline MFS efflux pump providing resistance to tetracycline, doxycycline and minocycline. It was detected in 76,758 unique genome assemblies. Analysis of the *tet(B)* loci revealed that 81 of 89 *Providencia stuartii* assemblies (June 2023) harbored a *tet(B)* locus from which MGEs were largely absent, and which was highly conserved among *P. stuartii* isolates with regards to synteny. Nucleotide identities between different *P. stuartii tet (B)* loci differed however considerably – while the majority of *P. stuartii tet(B)* loci were > 99 % similar to one another (these isolates are hereafter referred to as group 1) over the whole studied area (20kpb), other *P. stuartii tet(B)* loci differed up to 22 % from these loci (hereafter referred to as group 2), though the synteny of these loci was partly still conserved. Furthermore, the *P. stuartii* group 1 *tet(B)* locus was basically indistinguishable from the *Providencia thailandensis tet(B)* locus, whereas the *P. stuartii* group 2 locus was similar to the *tet(B)*-like locus in *Providencia vermicola* assemblies (Fig. 2). This may either indicate that some of these isolates have been misclassified, or that the taxonomy of *Providencia* does not describe the genomic diversity of this taxon. To investigate this, we analyzed the global average nucleotide identities (gANI) between randomly selected *P. stuartii, P. thailandensis* and *P. vermicola* assemblies. The analysis showed that *P. stuartii* (group 1) and *P. thailandensis* had gANIs of >=99 %, showing that the two species are indistinguishable from one another based on their nucleotide identities of the set of shared genes (supplementary Fig. 12), explaining the extreme similarities of the two species *tet(B)* loci. The gANI between group 2*P. stuartii* and group 2*P. vermicola* assemblies (also including a *P. rettgeri* assembly from the same branch, GCA_028062415.1) was >=99 %, whereas the gANI between group 1 and group 2*P.* stuartii was only ~ 83–84 %, indicating that the two groups are evolutionarily distinct (supplementary Fig. 13). This strongly indicates that the *Providencia* taxonomy does not reflect the evolutionary diversification of *Providencia* spp. Phylogenetic analysis based on the sequences of the marker gene *rpoB* in all *Providencia* assemblies revealed that the great majority of *Providencia* species harbored a gene at least 50 % identical to tet(B) (Fig. 3). As expected for a chromosomal gene, different branches of the phylogeny, largely representing different *Providencia* species complexes, harbored differential *tet(B)*-like genes. Exceptions were some *P. rettgeri* assemblies, which harbored tet(B) genes with 90–100 % nucleotide identity towards the mobile and *P. stuartii tet(B)* genes – as opposed to the majority of *P. rettgeri* isolates, which harbored *tet(B)*-like genes (*tet (57)/tet(59)*) with 50–80 % identity towards the mobile and *P. stuartii tet (B)* genes. Visual analysis of the *tet(B)* loci in these assemblies revealed these *tet(B)* loci to be mobile through association with IS elements in the respective genomes. The mobile *tet(B)* loci in e.g *S. enterica* or *Shigella flexneri* were >=99 % similar in nucleotide identity over several thousand basepairs, including several ORFs from the *P. stuartii tet(B)* locus (Fig. 2). In summary, these results strongly suggest that the *tet(B)* locus is native to *P. stuartii*, and thus that *P. stuartii* is the recent origin of mobile *tet(B)* genes.

The general lines of evidence presented above for *P. stuartii* as the origin of mobile *tet(B)* genes, as described in the introduction section, were applied to all candidate origin species to determine whether they likely are the origin of the respective mobile ARG.

3.3. Abundance of origins in microbial communities

To identify the distribution of origin species found in this study in metagenomic samples, and to assess in which environments these species may be abundant, we created a custom kraken2 database. The database was created especially for identifying these species (methods section 2.6) and distinguishing them from closely related species of the same genus.

The false positive rates generated for each origin species, generated from testing the custom database against simulated reads from reference genomes of all species in the same genus, are shown in supplementary

Table 1
Overview of origins of mobile ARGs identified in this study.

Resistance determinant group	Antibiotic class	Nucleotide identity origin/mge	Mge in origin loci	Chromosomal arg loci in other taxa in genus	Origin species (presence/genomes)	IS
aac(6')-ian	Aminoglycosides	87–100 %	Absent	Yes	*Pseudochrobactrum asaccharolyticum* (2/2)	IS*Kpn*18 ISVsa3
blaCdia	β-lactams	88–100 %	Absent	Yes	*Citrobacter amalonaticus (104/112)*	IS*Ec*9
blaHer(a)	β-lactams	98–99 %	Absent	No	*Atlantibacter hermannii* (11/16)	–
blaKluc	β-lactams	98–99 %	Absent	Yes	*Kluyvera cryocrescens (10/12)*	IS*Kpn*8
blaOxa-427	β-lactams	90–99 %	Absent	Yes	*Aeromonas media (31/33)*	IS1326
cati	Chloramphenicol	97–99 %	Absent	No	*Atlantibacter/Escherichia hermannii* (10/16)	
catii	Chloramphenicol	83–100 %	Absent	Yes	*Morganella morganii (209/376)*	–
catiii	Chloramphenicol	99–100 %	Absent	No	*Providencia stuartii (64/89)*	IS*Sf*1
tet(B)	Tetracycline	78–99 %	Absent	Yes	*Providencia stuartii (81/89)*	–
tet(D)	Tetracycline	92–97 %	Absent	Yes	*Morganella morganii (87/376)*	–
tet(H)	Tetracycline	95–99 %	Absent	Yes	*Proteus terrae (32/46)*	IS*Pa*14
tet(59)	Tetracycline	78–99 %	Absent	Yes	*Providencia rettgeri (147/328)*	IS*Vsa*3

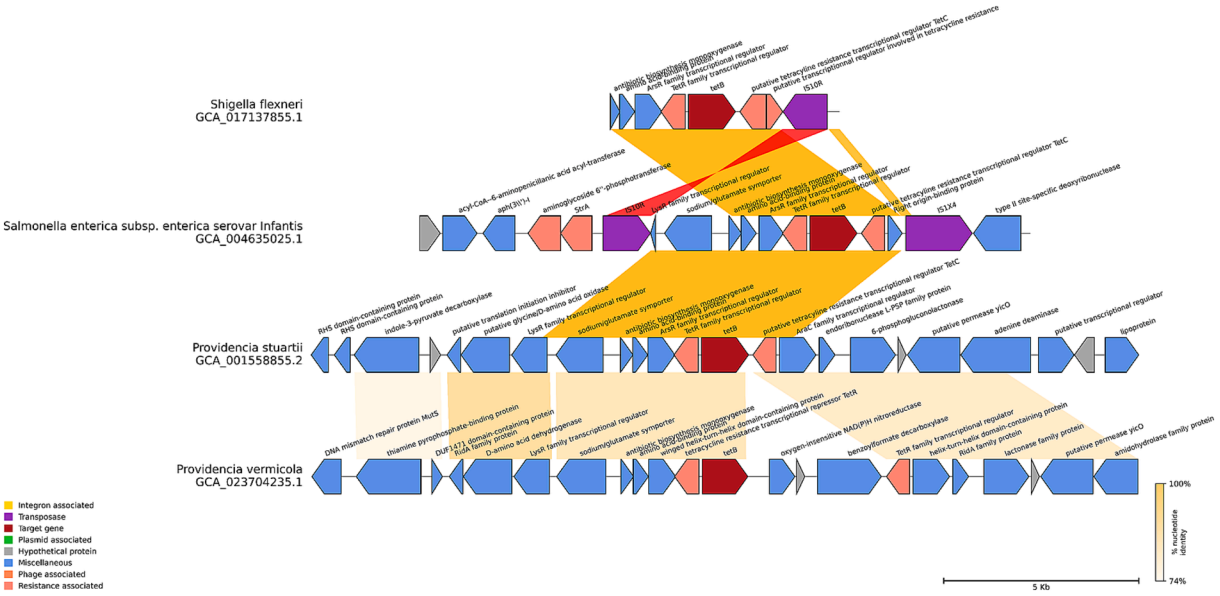

Fig. 2. Sequence comparison between putative chromosomal Tet(B) locus in *Providencia stuartii/vermicola* and mobile loci in *Salmonella enterica* and *Shigella flexneri*. Open reading frames are represented by boxes, arrows on boxes represent orf orientation. Box color represent orf type based on ncbi protein database annotation – Red: *tet(B)*, salmon: antibiotic resistance associated genes, purple: IS based on ISFinder annotation, grey: hypothetical proteins, blue: miscancellous. Elements between genomic loci indicate aligning regions between two sequences. Orange color intensity correlates with nucleotide identity over the aligning region, red color intensity and hourglass shape represents inverted gene orientations between the sequence alignments. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

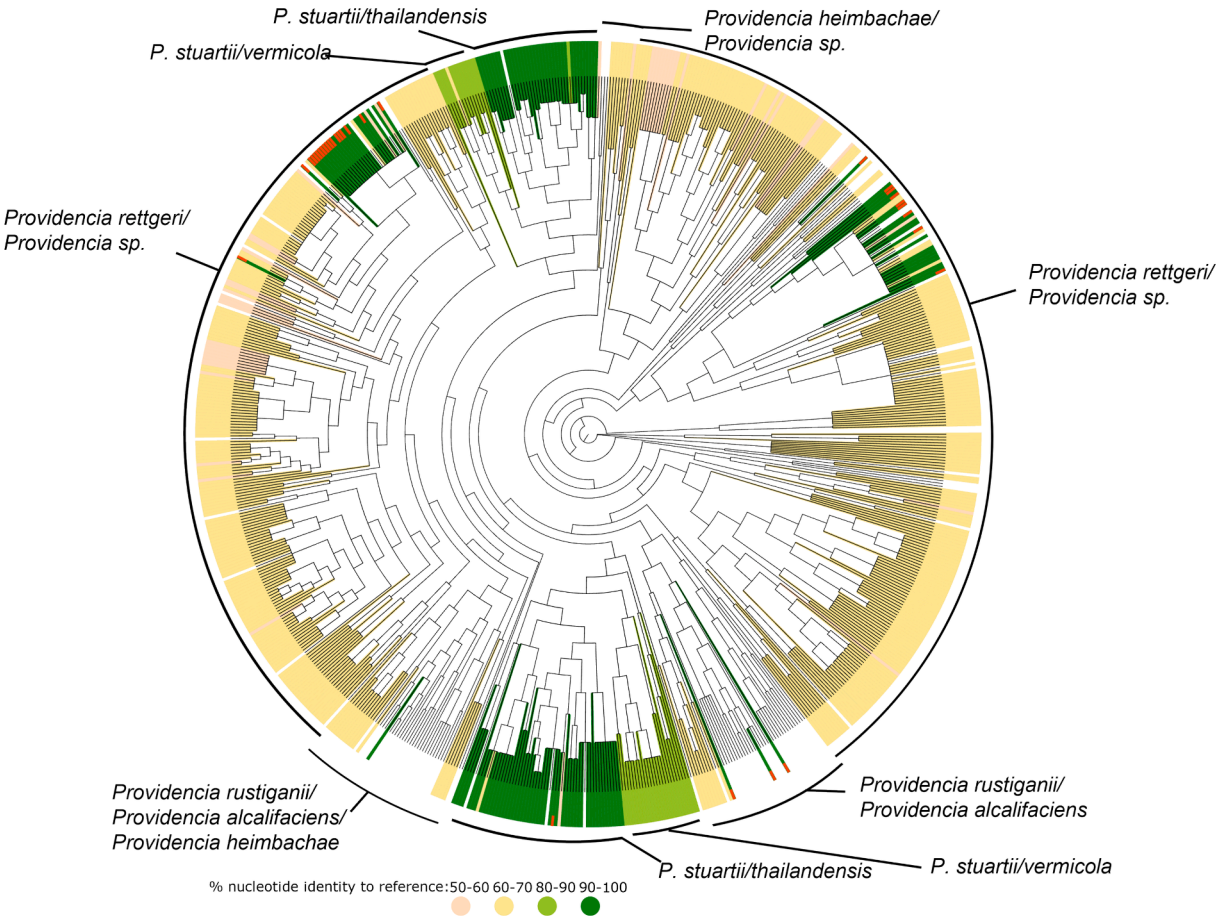

Fig. 3. *rpoB*-based phylogeny of *Providencia* assemblies. Color annotations are based on sequence similarity of the lowest identity *tet(B)*-like gene towards the mobile *tet(B)* reference gene. Orange rectangles denote assemblies carrying a gene >=90 % similar to tet(B) in which a MGE was identified within up to 10kbp up- or downstream of the tet(B)-like gene. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3, and were <= 0.5 % of all classified reads for all origin species in the respective mock community, with the exception of *M. morganii*, which had a false positive rate of ~ 3 %.

The origin species identified in this study were detected in different types of environments (Fig. 4). Generally, whereas all origin species were detected in wastewater (all wastewater metagenomes in this study represent influents to wastewater treatment plants) from different geographical regions, only a fraction was identified in human feces, cow feces, soil and non-polluted fresh- or saltwater environments. *Morganella morganii*, *Pseudochrobactrum asaccharolyticum, Kluyvera cryocrescens, Atlantibacter hermannii, Citrobacter amalonaticus* and *Aeromonas media* were identified as most abundant in wastewater – though abundance was shown to vary between geographic locations. *Proteus terrae, Providencia rettgeri* and *Providencia stuartii* were by far most abundant in Kazipally lake, a lake in India polluted by waste from antibiotic manufacturing (Bengtsson-Palme et al., 2014). Interestingly, in contrast to the other origin species identified in this study, these species were more abundant in European poultry feces than in wastewater samples.

An analysis of the fraction of samples in which each respective species could be identified after rarefication to the smallest metagenome

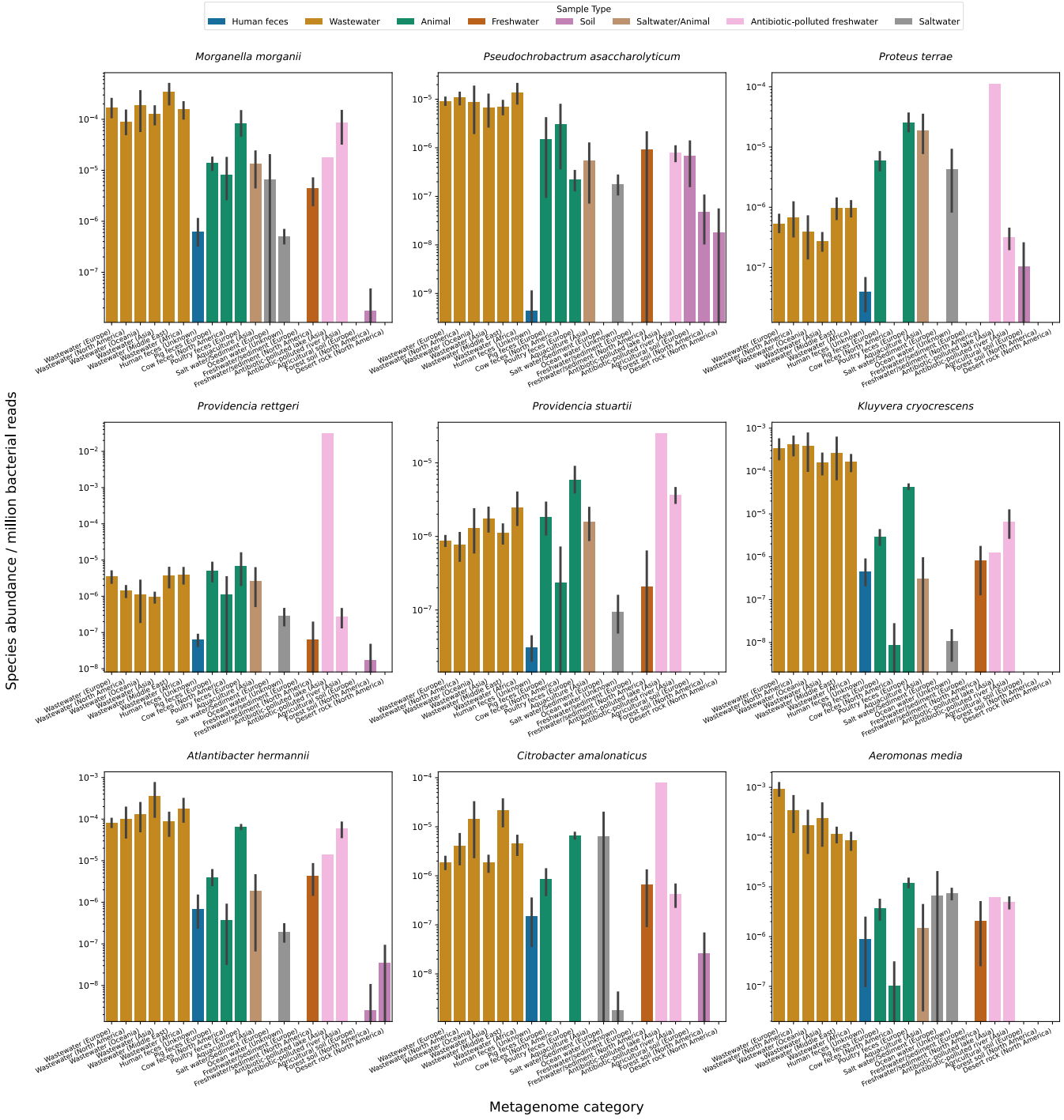

Fig. 4. Average abundances of origin species identified in this study in 1697 metagenomic samples from different environments. Error bars represent the 95% confidence interval. N samples number per environment type – Human feces: 538, Animal feces: 390, Wastewater: 273, Saltwater: 249, Soil: 176, Antibiotic-polluted freshwater: 63, Freshwater: 53, Saltwater/Animal: 13.

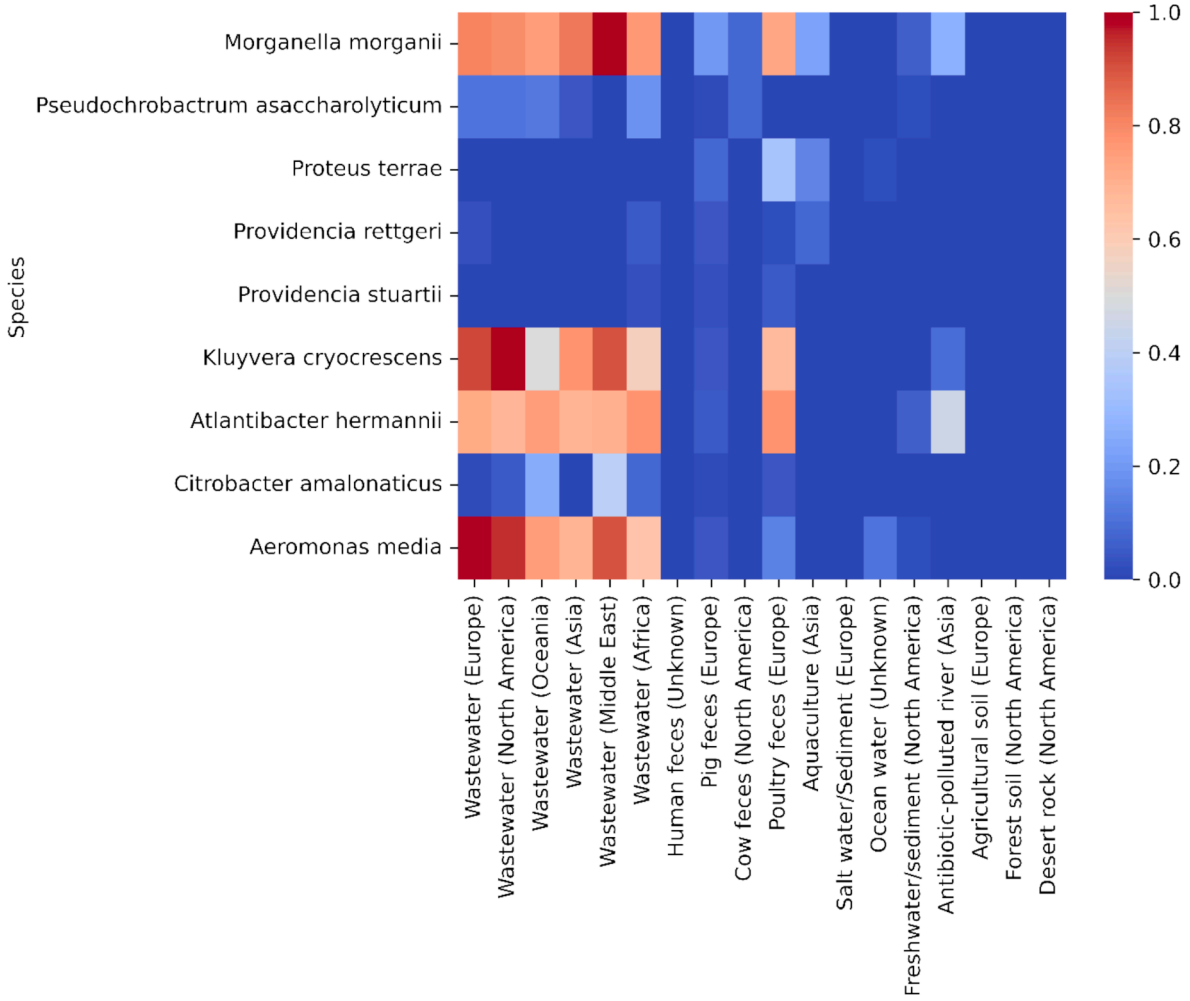

Fig. 5. Prevalence of the in this study identified origin species as average fraction of samples in which the species could be detected after rarefication for each environment. Environment types with < 3 samples are excluded from the plot.

sample size (50,304 reads) showed that *M. morganii, K, cryocrescens*, *A. hermannii* and *A. media* could be detected in the majority of all wastewater samples (n = 273, species present in 58–100 % of samples). *C. amalonaticus* and *P. asaccharolyticum* were only detected in a low fraction of the wastewater samples (species present in 0–25 % of samples). The highest detected fractions of *P. terrae, P. rettgeri and P. stuartii* were detected in poultry feces, aquaculture water and pig feces respectively. No origin species were recovered from Swedish lake samples (Freshwater/Sediment (Europe)), human feces or any type of soil samples after rarefication.

4. Discussion

In this study, we identified the origins of 12 mobile ARGs at species level using a random forest classifier on features calculated from > 1.5 million bacterial genomes. The ARGs for which origins were identified confer resistance to aminoglycoside, beta-lactam, chloramphenicol, and noteworthily, tetracycline antibiotics. Our results show that several species are the origin of multiple mobile ARGs. Metagenomic analysis revealed that, while the majority of the in this study identified origin species were most highly abundant in wastewaters, other origin species (especially those that were found to be the origin of mobile tetracycline resistance genes) were most abundant in poultry feces, aquaculture samples and an Indian lake polluted with waste from antibiotic manufacturing. In human feces and soil samples, origin species were generally identified at much lower abundances compared to wastewater/animal/polluted water samples. These results suggest that wastewater, poultry husbandry and aquaculture farming, as well as aquatic environments subjected to extreme pollution with antibiotics, may have been involved in the mobilization of mobile ARGs, and may thus be risk environments for the mobilization ARGs in the future.

4.1. Identified origins

As the previously known origins, which made up the training dataset, were exclusively Pseudomonadota, it is not surprising that Pseudomonadota species are identified as recent origins of mobile ARGs from the data presented here as well. This homogeneity may indicate the limitations of the criteria used to identify origins, as they may not capture all evolutionary processes by which ARGs could be mobilized, and thus miss the origin of certain genes even if they were present in the data. Of note is that Gram-positive bacteria were excluded from our analysis, as the molecular mechanisms of gene mobilization appear to differ from those of Gram-negatives, and we have not been able to manually confirm any Gram-positive origin using the previously defined criteria. As to other bacterial phyla and genes, if the mobilization and spread of the respective genes were not mediated evolutionarily recently by transposable elements, our model will not capture it, as it will not correspond to the pattern of a recently transposase-mobilized ARG in the same manner as the training data. For example, all mobile ARGs for which origins could be identified in this study, are widespread in Pseudomonadota. We hypothesize that focussing on mobile ARGs that are most

prevalent in other bacterial phyla might very well lead to the discovery of these ARGs origins in the respective phylum. Nevertheless, the results show that this strategy is effective in identifying at least a subset of the recent origins of mobile resistance genes.

Importantly, the results also show that not all criteria have to be fulfilled rigidly in order to assign an origin. The genes coding for *catII*, *catIII*, *tet(D)* and *tet(59)* were not present in parts of the respective origin genus (as discussed in supplementary sections 1.7, 1.8, 1.10 and 1.12). The reasons for this can vary from high genetic within-taxon variation to low-quality assemblies with missing sequences, and have to be investigated carefully in each specific case, in order to keep the number of falsely assigned origins as low as possible.

Curiously, some of the here identified origin taxa have been previously identified as the origins of other mobile ARGs as well. These include *Aeromonas* Ebmeyer et al., 2019, Ebmeyer et al., 2019*; Citrobacter* Jacoby et al., 2011, Barlow and Hall, 2002*; Kluyvera* (Poirel et al., 2002) and *Morganella* (Barnaud et al., 1998). Remarkably*; Morganella morganii,* previously identified as the origin of the DHA-family beta-lactamases, was in this study identified as the origin of two additional mobile ARGs, the chloramphenicol and tetracycline efflux pumps *catII* and *tet(D)*. *Providencia stuartii* and *Atlantibacter* hermanii have independently been identified as the origins of two mobile ARG groups each (Table 1). Yet, similar cases have been previously reported: e.g. *Klebsiella pneumoniae* is the origin of SHV-family beta-lactamases (Ford and Avison, 2004), the fosfomycin resistance genes fosA5/6 (Guo et al., 2016) and the mobile OqxAB efflux pumps (Kim et al., 2009), the *Citrobacter freundii* complex is the origin of CMY-2 family beta-lactamases and the *qnrB* fluoroquinolone resistance gene. This poses the question of why several genes are mobilized from certain species, but not from others, despite the latter harboring a plethora of ARGs effective against antibiotics. Perhaps there are certain traits that favor the mobilization and spread of genes from certain taxa, such as the ability to exchange DNA with a wide variety of other taxa, or high permissiveness towards exogenous DNA, e.g. allowing these taxa the uptake of a variety of transposable elements from the environment that in turn can mediate gene mobilization. What traits favor the mobilization of chromosomal resistance genes from certain species is important to understand in order to estimate risks associated with, as of now, exclusively chromosomal ARGs and requires further study.

4.2. Metagenomic analysis and prevalence of origin species in different environments

The testing of the custom kraken2 database created in this study resulted in low positive rates (<3%) for all origin species, indicating that it is useful to create reliable abundance estimates of these species in metagenomic samples. Though it is possible that unknown, closely related species exist that are falsely classified as origin species, the created database identifies the origin species reliably based on the genome data that are available to date.

Based on the database used in this study, the origin species identified in this study could be detected in a multitude of environments. Six of these species (*A. media, A. hermannii, C. amalonaticus, K. cryocrescens*, *M. morganii, P. asaccharolyticum*) were detected at the highest abundances (for the respective species) in the influent of wastewater treatment plants (Fig. 4). This result is concordant with Berglund et al. 2023 (Berglund et al., 2023), where the great majority of 22 studied origin species was found to be most abundant in wastewaters – an environment which not only harbors a wide variety of Pseudomonadota, but also contains exactly those mobile genetic elements that are suspected to have been involved in the mobilization of the origins' respective chromosomal ARGs (Berglund et al., 2023). It has been shown that hospital wastewaters (Kraupner et al., 2021) can select for resistance to antibiotics. Although often at somewhat lower concentrations, municipal wastewaters also contain many antibiotics (Novo et al., 2013), and it is plausible that many of these, in particular influents, could provide sufficient selection pressures to promote resistance development. A blend of selection pressures, a multitude of MGEs, origin species and recipients from various environments might hence make wastewaters of different kinds key sites for the mobilization and horizontal transmission of novel mobile ARGs.

Current evidence does not point towards a single, but several independent mobilization events for several different ARG variants (Ford and Avison, 2004; Ribeiro et al., 2015). The high nucleotide similarities between mobile ARGs and their chromosomal counterparts in their origins (>=95 % nt identity) further suggest that the mobilizations of these ARGs are evolutionary recent events. This indicates that the respective origin species is likely present in environments that repeatedly, if not constantly, contain the above-described blend of factors that promote the mobilization of ARGs, and that the mobilization even of ARGs that are already disseminated to human pathogens is a reoccurring process in those environments. The finding that some bacterial species are origins of several ARGs (*M. morganii* for example is the origin of DHA-1, catII and tet(D), *Citrobacter freundii* is the origin of *CMY-2* and *qnrB*, etc) suggests that these species thrive in environments providing conditions that effectively promote ARG mobilization. Wastewaters appear to be the environment type that most often fullfills these criteria (Berglund et al., 2023). This does, of course, not exclude that other environments, like the human/animal gut or other external environments, may fulfill these criteria too occasionally. Indeed, the great majority of today's known origin species (including those identified in this study), are opportunistic pathogens in humans (Ebmeyer et al., 2021) and even if they are rare in humans, opportunities for mobilization and fixation may arise when subjected to selection pressure during treatment of infections. However, it appears more likely that mobilization happens in environments where the origin species and other factors needed for ARG mobilization are constantly present. From the data presented here and in the literature, wastewaters can be concluded to represent such an environment.

Interestingly, not all origin species were most abundant in wastewater. The species *P. terrae*, *P, rettgeri* and *P. stuartii* were each highly abundant in metagenomes derived from European poultry feces. These three species are the origins of the mobile tetracycline resistance genes tet(H), tet(59) and the notorious tet(B). It is therefore intriguing to speculate about a causative relation between the emergence of these genes and the high use of tetracyclines in animal farming, including poultry (Grave et al., 2012; Morello et al., 2021). Furthermore, the analysed poultry feces samples were previously shown to contain high abundances of different IS, compared to the IS content in other (non-wastewater) environments (Berglund et al., 2023), which may contribute to the mobilization of chromosomal genes. In concordance with this, *M. morganii*, here identified as the origin of tet(D) (but also of the DHA beta-lactamases (Barnaud et al., 1998) and the catII chloramphenicol resistance gene), was identified in poultry feces in similar abundances as in wastewater. The highest abundances of all three species were however detected in the metagenome from Kazipally lake, an Indian lake polluted with exceptionally high concentrations of antibiotics through wastewater from antibiotic manufacturing (Fick et al., 2009) – The lake accordingly harbors a microbiota characterized by exceptional abundances of resistance genes towards all major antibiotic classes (Bengtsson-Palme et al., 2014). The high abundances of several origin species, as shown here, provide additional evidence for the role of industrial antibiotic pollution in the evolution/emergence of clinically important antibiotic resistance. Together these findings should further motivate actions to limit industrial antibiotic pollution, for example by adopting the newly developed WHO guidance on wastewater and solid waste magament for manufacturing of antibiotics in different contexts (World Health Organization, 2024). These may may include adopting pollution criteria during procurement, in subsidy decisions, in environmental legislation and in decisions to invest in antibiotic manufacturing, to name a few (Larsson and Flach, 2022). The recently adopted political declaration at the United Nation Global Assembly

accordingly stress the value of limiting pollution from antibiotic manufacturing and adopting pollution standards, as well as the need to reduce current use of antibiotics in animal production systems (World Health Organization, 2024).

The three *Providencia* species were also detected in wastewater samples from around the globe (albeit *P. terrae* abundances were orders of magnitudes higher in poultry feces than in wastewater). Further research focusing on to what extent antibiotics in wastewater cause selection pressure on microbial communities is needed in order to assess how likely the emergence of certain ARGs is in wastewater vs other environments. Certainly, if, for example, tetracycline selection pressure was present in both wastewater and poultry husbandry/aquaculture associated environments, all of these might be potential risk environments for the emergence of mobile ARGs, as the origin species thriving in these environments would be constantly subjected to antibiotic selection pressure, which in turn increases the chance of the respective 'native' ARGs to be mobilized (Lartigue et al., 2006).

The abundances of the origin species described here were considerably lower in soil environments (and for most species, human faeces) compared to other environments, which is also what Berglund et al. observed in their recent study on where previously known origins of ARGs thrive (Berglund et al., 2023). Especially soil has been suggested as a potential source of novel mobile ARGs (Davies and Davies, 2010; Han et al., 2022); to a large extent in relation to the producer hypothesis. The to date available data however, suggest that the abundance of Gram-negative, Pseudomonadota origin species in soil (with the exception of *P. asaccharolyticum*) is low compared to these species abundance in other environments. In addition, some studies (Song et al., 2017) suggest that the bioavailability, and thus the exerted selection pressure, of antibiotics in soil may be limited. Furthermore, the overlap between both known and predicted ARGs between soil and the human gut is low (Inda-Díaz et al., 2023). Thus, current evidence does not suggest soil environments as a source for the mobile ARGs where an origin is known. Whether this observation holds for mobile ARGs originating from other types of bacteria remains to be investigated. Further research, especially on diverse, novel metagenomes from soil and other natural environments is needed to see whether the high abundances of origin species in wastewater in comparison to those natural environments is a universal phenomenon.

4.3. Classifier performance

The precision (0.84) and comparably recall (0.88) values produced by the classifier may be a result of the small amount of ARG origins that are known to date and that data are available on, which likely only in part reflects the mobilization mechanisms present in nature.While this means that the model will miss origins that deviate from the learned pattern, it shows that if the origin species of ARGs following this pattern were present in the new data, there is a good chance of identifying them. Focusing on a model with high precision makes sense in this case, as a higher recall would increase the amount of false positives as well – thus the amount of manual work to analyze potential origins would increase. The discrepancy between the precision obtained on the training data using leave one out cross validation (84 %) and the approximate precision based on the new data, only ~ 14 %, is quite large. This is expected, since the proportion of non-origin species is likely much larger, compared to the training set, when taking all bacterial species available at the NCBI assembly database into account. Another factor influencing the classifiers precision on the new data is data quality – in many cases, some blocks contained shorter regions of putatively chromosomal origin which were, due to the automatized processing, classified as mobile. Therefore, sequences that only were present on the chromosomes of certain species were falsely classified as being mobile, and thus as having a chromosomal origin. A second factor is human error. In many cases, distinguishing between mobile and putatively chromosomal loci is not straightforward, and further research on the respective whole assemblies would be required (i.e gANI analyses, phylogeny etc) to obtain the required evidence, which is difficult with thousands of genomes for several hundred ARG groups each. Furthermore, some ARG groups are identified in hundreds or thousands of assemblies – this requires, at least when using the approach presented here, random subsampling of assemblies to visualize (even with predictions on what species may be the origin of the respective gene), which opens up for missing important sequences. So, while the applied approach somewhat alleviates the difficulty of searching through millions of sequences, certain challenges remain.

5. Conclusion

The results presented in this study show that the origins of multiple mobile ARGs are non-antibiotic-producing, Gram-negative Pseudomonadota species, many of them known to be opportunistic pathogens of humans and animals. Furthermore, the results of the metagenomic analysis show the presence of these origin species in environments that contain the necessary factors to drive the mobilization and dissemination of their chromosomal ARGs, such as wastewater, poultry feces and sites polluted by antibiotics. While these findings should be validated by further, large-scale studies using diverse sets of metagenomes, these results indicate that these environments could very well act as matrices for the mobilization of novel ARGs, they present a risk to human health and should thus be included in risk estimations and management associated with antibiotic resistance. Limiting the antibiotic exposure of microbial communities at these sites is likely crucial in order to mitigate the risk for the emergence of novel mobile ARGs.

While this study shows the potential of machine learning techniques to unveil the origins of mobile ARGs utilizing ever increasing amounts of data, the capabilities of the model were limited by the few origins known to date (the training data), which were, in turn, biased towards mobile ARGs in Pseudomonadota. Further research is needed to understand how different kinds of ARGs are mobilized in different bacterial phyla, in order to gain a more complete understanding of from which bacteria mobile ARGs originate and where these taxa thrive. The finding that almost all identified origin species have been identified in sewage supports recent research pointing toward wastewaters as a potential environment where ARGs can be mobilized and disseminated.

Code and data availability

All assemblies used in this study are publicly available at the NCBI Assembly database, accession numbers for assemblies used in specific analyses are provided in the respective text/figures. All code for this manuscript is available at https://github.com/EbmeyerSt/origin_rfc.

CRediT authorship contribution statement

Stefan Ebmeyer: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Erik Kristiansson:** Writing – review & editing, Supervision, Resources, Methodology, Investigation, Funding acquisition, Conceptualization. **D. G. Joakim Larsson:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

This research was funded by the Swedish Research Council VR (grant numbers 2020–06648, 2023–03891, 2022–00945, 2018–05771, 2018–02835), Swedish Research Council FORMAS 2021–00949. Open access funding was provided by the University of Gothenburg.

Author Contribution (CREDIT statement) for the manuscript "Unraveling the origins of mobile antibiotic resistance genes using random forest classification of large-scale genomic data" submitted for

publication in Environment International.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [D. G. Joakim Larsson reports financial support was provided by Swedish Research Council. Erik Kristiansson reports financial support was provided by Swedish Research Council. D. G. Joakim Larsson reports financial support was provided by Swedish Research Council Formas. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2025.109374.

Data availability

All assemblies are available at the NCBI Assembly database, accession numbers are provided in the respective text/figures. All code is available at https://github.com/EbmeyerSt/origin_rfc.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Barlow, M., Hall, B.G., 2002. Origin and evolution of the AmpC beta-lactamases of Citrobacter freundii. *Antimicrob. Agents Chemother.* 46, 1190–1198.

Barnaud, G., Arlet, G., Verdet, C., Gaillot, O., Lagrange, P.H., Philippon, A., et al., 1998. Salmonella enteritidis: AmpC plasmid-mediated inducible beta-lactamase (DHA-1) with an ampR gene from Morganella morganii. Antimicrob. Agents Chemother. 42, 2352–2358.

Beauchef, A., Arlet, G., Gautier, V., Labia, R., Grimont, P., Philippon, A., et al., 2003. Molecular and biochemical characterization of a novel class A β-lactamase (HER-1) from *Escherichia hermannii*. Antimicrob. Agents Chemother. 47, 2669–2673.

Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E., Larsson, D.G.J., 2014. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Front. Microbiol.* 5, 648.

Berglund, F., Ebmeyer, S., Kristiansson, E., Larsson, D.G.J., 2023. Evidence for wastewaters as environments where mobile antibiotic resistance genes emerge. *Commun. Biol.* 6.

Bogaerts, P., Naas, T., Saegeman, V., Bonnin, R.A., Schuermans, A., Evrard, S., Bouchahrouf, W., Jove, T., Tande, D., de Bolle, X., Beyrouthy, R., Huang, T.D., Glupczynski, Y., et al., 2017. OXA-427, a new plasmid-borne carbapenem-hydrolysing class D β-lactamase in Enterobacteriaceae. J. Antimicrob. Chemother. 72, 2469–2477.

Brown, C.L., Mullet, J., Hindi, F., Stoll, J.E., Gupta, S., Choi, M., Keenum, I., Vikesland, P., Pruden, A., Zhang, L., et al., 2022. mobileOG-db: a manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements. Appl. Environ. Microbiol. https://doi.org/10.1128/AEM.00991-22.

Buchfink, B., Xie, C., Huson, D.H., 2014. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60.

Bushnell B. BBMap. sourceforge.net/projects/bbmap/.

Davies, J., Davies, D., 2010. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev. MMBR* 74, 417–433.

Decousser, J.W., Poirel, L., Nordmann, P., 2001. Characterization of a chromosomally encoded extended-spectrum class A β-lactamase from *Kluyvera cryocrescens*. *Antimicrob. Agents Chemother.* 45, 3595–3598.

Ebmeyer, S., Kristiansson, E., Larsson, D.G.J., 2018. PER extended-spectrum β-lactamases originate from Pararheinheimera spp. *Int. J. Antimicrob. Agents*. https://doi.org/10.1016/J.IJANTIMICAG.2018.10.019.

Ebmeyer, S., Kristiansson, E., Larsson, D.G.J., 2019. The mobile FOX AmpC beta-lactamases originated in Aeromonas allosaccharophila. *Int. J. Antimicrob. Agents* 54, 798–802.

Ebmeyer, S., Kristiansson, E., Larsson, D.G.J., 2019. CMY-1/MOX-family AmpC β-lactamases MOX-1, MOX-2 and MOX-9 were mobilized independently from three *Aeromonas* species. *J. Antimicrob. Chemother.* https://doi.org/10.1093/jac/dkz025.

Ebmeyer, S., Kristiansson, E., Larsson, D.G.J., 2021. A framework for identifying the recent origins of mobile antibiotic resistance genes. *Commun. Biol.* 4, 1–10.

Ebmeyer, S., Coertze, R.D., Berglund, F., Kristiansson, E., Larsson, D.G.J., 2022. GEnView: a gene-centric, phylogeny-based comparative genomics pipeline for bacterial genomes and plasmids. *Bioinformatics* 38, 1727–1728.

Fick, J., Söderström, H., Lindberg, R.H., Phan, C., Tysklind, M., Larsson, D.G.J., et al., 2009. Contamination of surface, ground, and drinking water from pharmaceutical production. Environ. Toxicol. Chem. 28, 2522–2527.

Ford, P.J., Avison, M.B., 2004. Evolutionary mapping of the SHV -lactamase and evidence for two separate IS26-dependent blaSHV mobilization events from the Klebsiella pneumoniae chromosome. *J. Antimicrob. Chemother.* 54, 69–75.

Grave, K., Greko, C., Kvaale, M.K., Torren-Edo, J., Mackay, D., Muller, A., Moulin, G., ESVAC Group, et al., 2012. Sales of veterinary antibacterial agents in nine European countries during 2005-09: trends and patterns. J. Antimicrob. Chemother. 67, 3001–3008.

Guo, Q., Tomich, A.D., McElheny, C.L., Cooper, V.S., Stoesser, N., Wang, M., Sluis-Cremer, N., Doi, Y., et al., 2016. Glutathione- *S* -transferase FosA6 of *Klebsiella pneumoniae* origin conferring fosfomycin resistance in ESBL-producing *Escherichia coli*. J. Antimicrob. Chemother. 71, 2460–2465.

Han, B., Ma, L., Yu, Q., Yang, J., Su, J., Hilal, N., Li, L., Zhang, T., Li, H., et al., 2022. The source, fate and prospect of antibiotic resistance genes in soil: a review. Front. Microbiol. 13, 976657.

Huerta-Cepas, J., Serra, F., Bork, P., 2016. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638.

Humeniuk, C., Arlet, G., Gautier, V., Grimont, P., Labia, R., Philippon, A., et al., 2002. Beta-lactamases of Kluyvera ascorbata, probable progenitors of some plasmid-encoded CTX-M types. Antimicrob. Agents Chemother. 46, 3045–3049.

Inda-Díaz, J.S., Lund, D., Parras-Moltó, M., Johnning, A., Bengtsson-Palme, J., Kristiansson, E., et al., 2023. Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes. Microbiome 11, 44.

Jacoby, G.A., Griffin, C.M., Hooper, D.C., 2011. Citrobacter spp. as a source of qnrB Alleles. *Antimicrob. Agents Chemother.* 55, 4979–4984.

Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N., Doshi, S., Courtot, M., Lo, R., Williams, L.E., Frye, J.G., Elsayegh, T., Sardar, D., Westman, E.L., Pawlowski, A.C., Johnson, T.A., Brinkman, F.S.L., Wright, G.D., McArthur, A.G., et al., 2017. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res 45, D566–D573.

Jiang, X., Ellabaan, M.M.H., Charusanti, P., Munck, C., Blin, K., Tong, Y., Weber, T., Sommer, M.O.A., Lee, S.Y, 2017. Dissemination of antibiotic resistance genes from antibiotic producers to pathogens. Nat. Commun. 8, 15784.

Kim, H.B., Wang, M., Park, C.H., Kim, E.-C., Jacoby, G.A., Hooper, D.C., et al., 2009. oqxAB encoding a multidrug efflux pump in human clinical isolates of Enterobacteriaceae. Antimicrob. Agents Chemother. 53, 3582–3584.

Kitts, P.A., Church, D.M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R. G., Tatusova, T., Xiang, C., Zherikov, A., DiCuccio, M., Murphy, T.D., Pruitt, K.D., Kimchi, A., 2016. Assembly: a resource for assembled genomes at NCBI. Nucleic Acids Res. 44, D73–D80.

Kraupner, N., Hutinel, M., Schumacher, K., Gray, D.A., Genheden, M., Fick, J., Flach, C.-F., Larsson, D.G.J., et al., 2021. Evidence for selection of multi-resistant E. coli by hospital effluent. Environ. Int. 150, 106436.

Larsson, D.G.J., Flach, C.-F., 2022. Antibiotic resistance in the environment. Nat. Rev. Microbiol. 20, 257–269.

Lartigue, M.-F., Poirel, L., Aubert, D., Nordmann, P., 2006. In vitro analysis of IS *Ecp1B* -mediated mobilization of naturally occurring β-lactamase gene *bla* $_{CTX-M}$ of *Kluyvera ascorbata*. *Antimicrob. Agents Chemother.* 50, 1282–1286.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* 22, 1658–1659.

Lund, D., Kieffer, N., Parras-Molto, M., Ebmeyer, S., Berglund, F., Johnning, A., Larsson, D.G.J., Kristiansson, E., 2022. Large-scale characterization of the macrolide resistome reveals high diversity and several new pathogen-associated genes. Microb. Genomics 8, 770.

Lund, D., Coertze, R.D., Parras-Moltó, M., Berglund, F., Flach, C.-F., Johnning, A., Larsson, D.G.J., Kristiansson, E., et al., 2023. Extensive screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens. Commun. Biol. 6, 812.

Morello, S., Pederiva, S., Avolio, R., Amato, G., Zoppi, S., Di Blasio, A., Abete, M.C., Casalone, C., Desiato, R., Ru, G., Marchis, D., 2021. Tetracyclines in processed animal proteins: a monitoring study on their occurrence and antimicrobial activity. Foods 10 (696), 696. https://doi.org/10.3390/foods10040696.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41**, D8–D20 (2013).

Novo, A., André, S., Viana, P., Nunes, O.C., Manaia, C.M., 2013. Antibiotic resistance, antimicrobial residues and bacterial community composition in urban wastewater. *Water Res.* 47, 1875–1887.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2018. Scikit-learn: Machine Learning in Python. Preprint at. http://arxiv.org/abs/1201.0490.

Poirel, L., Kämpfer, P., Nordmann, P., 2002. Chromosome-encoded Ambler class A beta-lactamase of Kluyvera georgiana, a probable progenitor of a subgroup of CTX-M extended-spectrum beta-lactamases. *Antimicrob. Agents Chemother.* 46, 4038–4040.

Poirel, L., Rodriguez-Martinez, J.-M., Mammeri, H., Liard, A., Nordmann, P., 2005. Origin of plasmid-mediated quinolone resistance determinant QnrA. *Antimicrob. Agents Chemother.* 49, 3523–3525.

Poirel, L., Kieffer, N., Fernandez-Garayzabal, J.F., Vela, A.I., Larpin, Y., Nordmann, P., et al., 2017. MCR-2-mediated plasmid-borne polymyxin resistance most likely originates from Moraxella pluranimalium. J. Antimicrob. Chemother. 72, 2947–2949.

Ribeiro, T.G., Novais, Â., Branquinho, R., Machado, E., Peixe, L., 2015. Phylogeny and comparative genomics unveil independent diversification trajectories of qnrB and genetic platforms within particular citrobacter species. *Antimicrob. Agents Chemother.* 59, 5951–5958.

Schonlau, M., Zou, R.Y., 2020. The random forest algorithm for statistical learning. *Stata J. Promot. Commun. Stat. Stata* 20, 3–29.
Shimoyama, Y.. pyGenomeViz: A genome visualization python package for comparative genomics. https://github.com/moshi4/pyGenomeViz.
Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., Chandler, M., 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* 34, D32–D36.
Song, J., Rensing, C., Holm, P.E., Virta, M., Brandt, K.K., 2017. Comparison of metals and tetracycline as selective agents for development of tetracycline resistant bacterial communities in agricultural soil. *Environ. Sci. Technol.* 51, 3040–3047.
Underwood, S., Avison, M.B., 2004. Citrobacter koseri and Citrobacter amalonaticus isolates carry highly divergent -lactamase genes despite having high levels of biochemical similarity and 16S rRNA sequence homology. J. Antimicrob. Chemother. 53, 1076–1080.
Varghese, N.J., Mukherjee, S., Ivanova, N., Konstantinidis, K.T., Mavrommatis, K., Kyrpides, N.C., Pati, A., et al., 2015. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 43, 6761–6771.
Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 1–13.
World Health Organization, 2024. Guidance on wastewater and solid waste management for manufacturing of antibiotics, Geneva, ISBN 978-92-4-009725-4, p. 79pp.
Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M., Larsen, M.V., et al., 2012. Identification of acquired antimicrobial resistance genes. J. Antimicrob. Chemother. 67, 2640.

Further reading

Roberts, M.C., 2005. Update on acquired tetracycline resistance genes. *FEMS Microbiol. Lett.* 245, 195–203.
United Nations General, Assembly, 2024. Political declaration of the high-level meeting on antimicrobial resistance. A/79/L.5. https://documents.un.org/doc/undoc/ltd/n24/278/35/pdf/n2427835.pdf.