



**CHALMERS**  
UNIVERSITY OF TECHNOLOGY

## **Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes**

Downloaded from: <https://research.chalmers.se>, 2021-08-31 11:24 UTC

Citation for the original published paper (version of record):

Berglund, F., Böhm, M., Martinsson, A. et al (2020)

Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes

Microbial Genomics, 6(11)

<http://dx.doi.org/10.1099/mgen.0.000455>

N.B. When citing this work, cite the original published paper.

# Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes

Fanny Berglund<sup>1,2</sup>, Maria-Elisabeth Böhm<sup>2,3</sup>, Anton Martinsson<sup>1,2</sup>, Stefan Ebmeyer<sup>2,3</sup>, Tobias Österlund<sup>1,2</sup>, Anna Johnning<sup>1,2,4</sup>, D. G. Joakim Larsson<sup>2,3</sup> and Erik Kristiansson<sup>1,2,\*</sup>

## Abstract

Tetracyclines are broad-spectrum antibiotics used to prevent or treat a variety of bacterial infections. Resistance is often mediated through mobile resistance genes, which encode one of the three main mechanisms: active efflux, ribosomal target protection or enzymatic degradation. In the last few decades, a large number of new tetracycline-resistance genes have been discovered in clinical settings. These genes are hypothesized to originate from environmental and commensal bacteria, but the diversity of tetracycline-resistance determinants that have not yet been mobilized into pathogens is unknown. In this study, we aimed to characterize the potential tetracycline resistome by screening genomic and metagenomic data for novel resistance genes. By using probabilistic models, we predicted 1254 unique putative tetracycline resistance genes, representing 195 gene families (<70% amino acid sequence identity), whereof 164 families had not been described previously. Out of 17 predicted genes selected for experimental verification, 7 induced a resistance phenotype in an *Escherichia coli* host. Several of the predicted genes were located on mobile genetic elements or in regions that indicated mobility, suggesting that they easily can be shared between bacteria. Furthermore, phylogenetic analysis indicated several events of horizontal gene transfer between bacterial phyla. Our results also suggested that acquired efflux pumps originate from proteobacterial species, while ribosomal protection genes have been mobilized from *Firmicutes* and *Actinobacteria*. This study significantly expands the knowledge of known and putatively novel tetracycline resistance genes, their mobility and evolutionary history. The study also provides insights into the unknown resistome and genes that may be encountered in clinical settings in the future.

## DATA SUMMARY

The data analysed in this study consisted of pre-existing datasets, which are specified in Table 1 and throughout the text. The genes used to construct the hidden Markov models are listed in Table S1 (available with the online version of this article). The new genes predicted in this work are listed in Table S3, together with their respective protein sequences. The fARGene method is publicly available at <https://github.com/fannyhb/fargene>.

## INTRODUCTION

Tetracyclines are broad-spectrum antibiotics widely used to treat a broad range of bacterial infections since their discovery and introduction as clinical agents in the 1940s. Efficiency against many forms of both Gram-negative and Gram-positive bacteria in combination with relatively few side effects have made tetracyclines one of the most extensively used classes of antibiotics, and the most used class for animals [1]. In addition to clinical use, tetracyclines are commonly used in subtherapeutic doses to induce growth-promotion in, for example, swine and poultry [2]. Together, this has resulted in the promotion of resistance in a variety of bacterial species,

Received 23 April 2020; Accepted 27 September 2020; Published 30 October 2020

**Author affiliations:** <sup>1</sup>Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden; <sup>2</sup>Centre for Antibiotic Resistance Research (CARE), University of Gothenburg, Gothenburg, Sweden; <sup>3</sup>Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; <sup>4</sup>Department of Systems and Data Analysis, Fraunhofer-Chalmers Centre, Chalmers Science Park, Gothenburg, Sweden.

**\*Correspondence:** Erik Kristiansson, [erik.kristiansson@chalmers.se](mailto:erik.kristiansson@chalmers.se)

**Keywords:** antibiotic resistance; hidden Markov model; metagenomics; microbiome; resistome; tetracycline resistance.

**Abbreviations:** ARG, antibiotic-resistance gene; HMP, Human Microbiome Project; MFS, major facilitator superfamily; MIC, minimum inhibitory concentration; NCBI, National Center for Biotechnology Information.

**Data statement:** All supporting data, code and protocols have been provided within the article or through supplementary data files. Four supplementary figures and four supplementary tables are available with the online version of this article.

000455 © 2020 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License.

significantly reducing the value of tetracyclines as antibiotics [1, 3]. Increasing resistance has spawned the development of synthetically modified tetracyclines, such as minocycline, doxycycline and tigecycline; however, new forms of resistance encompassing the newer tetracyclines continue to emerge and spread [4].

Tetracycline resistance is often based on three main biochemical mechanisms, alone or in combination: (i) active export of the tetracycline molecules via efflux pumps, (ii) weakening of the interaction between tetracycline and the ribosome (ribosomal protection), and (iii) enzymatic degradation of tetracycline via hydroxylation [1, 4, 5]. Tetracycline resistance genes encoding these mechanisms are often encountered on mobile genetic elements, which makes them easily transferable between bacteria. Indeed, the number of discovered and confirmed tetracycline resistance genes has increased dramatically over the last few decades, mainly due to horizontal gene transfer between bacteria [1]. To date, 62 different tetracycline resistance gene families have been discovered (<80% amino acid sequence identity), where 35 encode efflux proteins, 13 encode ribosomal protection proteins and 14 encode inactivating enzymes [6].

Many antibiotic-resistance genes (ARGs) frequently encountered in clinical isolates are hypothesized not to originate from pathogenic bacteria. In fact, environmental and commensal bacteria maintain large and diverse resistomes from which ARGs can be mobilized into pathogens [7, 8]. A recent example is the gene *tet(X)*, which was discovered in the 1980s and whose product catalyses the degradation of tetracyclines, including tigecycline [9–11]. During the last decade, *tet(X)* has been detected in a wide range of human pathogens, including *Klebsiella pneumoniae* and members of *Pseudomonadaceae*, and today constitutes an increasing clinical problem [12, 13]. The gene was first discovered on a transposable element in *Bacteroides fragilis*, a strict anaerobe, and given that *tet(X)* is dependent on oxygen, it has been suggested that this is not its original host [14]. Another example is the ribosomal protection gene *otr(A)*, which has been found in *Streptomyces* spp. and *Mycobacterium* spp., but is expected to have an environmental origin [4]. This suggests that environmental bacteria harbour a collection – a reservoir – of tetracycline resistance genes that can be transferred into human pathogens over time [15]. The size and diversity of the tetracycline resistome are, however, unclear; therefore, we do not know which resistance genes may be transferred into pathogens in the future. Knowledge about the future forms of tetracycline resistance genes may enable early diagnostics and facilitate the implementation of management strategies to limit the spread of new forms of multiresistant bacteria.

Functional metagenomics – a method where random pieces of DNA from microbial communities are extracted and inserted into a bacterial host that is tested for induced phenotypes – is widely used to identify novel ARGs [16–18]. Analysis of soil using functional metagenomics was, for example, used to identify nine new genes for enzymatic degradation of tetracycline [*tet(47)*–*tet(55)*] [5]. Although functional metagenomics

### Impact Statement

Tetracyclines are one of the most widely used classes of antibiotics and crucial for treating many forms of infection. However, the incidence of tetracycline resistant bacteria is increasing globally, which is to a large extent caused by new variants of tetracycline-resistance genes encountered in pathogens. These genes are hypothesized to originate from an unexplored reservoir of tetracycline resistance genes maintained by environmental bacteria. The size and content of this reservoir are unknown and, thus, we do not know the evolutionary history of the known tetracycline resistance genes and which novel genes we may encounter in the future. In this study, we have characterized a large number of new putative tetracycline resistance genes not yet encountered in clinical settings. Phylogenetic analysis of the predicted and previously known tetracycline resistance genes has given new insights about their evolutionary origin and mobility patterns. This study significantly expands the known tetracycline resistome, and describes which resistance genes may be mobilized and transferred into pathogens in the future.

enables detection of novel ARGs without any prior knowledge about previously described genes, the method has a relatively low throughput (typically, ~10 Gb DNA is incorporated per experiment), which makes it unsuitable for characterization of novel ARGs at a larger scale. An alternative approach is to search for novel ARGs in genomic and shotgun metagenomic data. The great majority of newly discovered ARGs are evolutionarily related to resistance genes that are already known, allowing patterns of previously uncharacterized ARGs to be identified through sequence analysis. We have previously developed fARGene, a highly sensitive computational method based on hidden Markov models, for this particular purpose [19, 20]. By screening large volumes of genomic and metagenomic data we were able, for example, to significantly expand the number of subclass B1 metallo- $\beta$ -lactamases, from 22 to 81 gene families [21]. Given the rapidly increasing volumes of bacterial genomic data present in public repositories, this approach has the potential to explore the diversity of novel ARGs in the bacterial communities at a much larger scale.

In this study, we aimed to expand the number of characterized tetracycline resistance genes and thereby provide a clearer picture of their evolutionary histories and origins. Through large-scale screening of genomic and metagenomic data using highly sensitive and optimized probabilistic gene models, we predicted 195 families of putative tetracycline resistance genes, of which 164 were novel and, thus, previously uncharacterized. Seventeen of the predicted genes were synthetically constructed and expressed in *Escherichia coli*, whereof seven resulted in an induced resistance phenotype. Phylogenetic analysis identified several horizontal gene transfer events

**Table 1.** Predicted tetracycline resistance genes

Dataset	Size (nt)	Efflux		Ribosomal protection		Enzymatic	
		Genes*	Families (new/ total)†	Genes*	Families (new/ total)†	Genes*	Families (new/ total)†
<b>Genomic</b>							
NCBI RefSeq [77]	2.78×10 <sup>10</sup>	137	16/29	232	51/57	20	9/11
NCBI plasmids [77]	1.07×10 <sup>9</sup>	54	2/13	35	3/7	3	0/1
NCBI nt [19]	1.66×10 <sup>11</sup>	289	30/45	563	77/87	76	14/20
NCBI environmental [19]	1.09×10 <sup>11</sup>	23	7/15	317	17/23	17	3/6
HMP genomic [41]	6.83×10 <sup>9</sup>	20	2/8	105	7/12	4	1/3
<b>Metagenomic</b>							
Human gut [47]	2.80×10 <sup>11</sup>	3	0/2	13	0/5	2	0/2
HMP metagenome [45]	3.51×10 <sup>12</sup>	3	0/3	10	3/5	2	0/2
Pig gut [44]	1.74×10 <sup>12</sup>	9	0/7	13	6/10	1	0/1
WWTPs [42]	4.82×10 <sup>11</sup>	11	4/8	14	5/12	3	2/3
Pune river [43]	3.93×10 <sup>11</sup>	5	1/5	19	8/13	1	0/1
Polluted lake [50]	6.76×10 <sup>9</sup>	0	0	0	0	0	0
Polluted river [46]	2.86×10 <sup>10</sup>	3	0/2	2	1/2	1	0/1
Oil spill [49]	3.39×10 <sup>11</sup>	0	0	0	0	0	0
Tara Ocean [48]	4.89×10 <sup>12</sup>	4	2/3	0	0	4	2/3
Well water [51]	7.62×10 <sup>10</sup>	0	0	0	0	0	0
Soil [51]	5.05×10 <sup>10</sup>	0	0	0	0	0	0
<b>Total‡</b>	<b>1.21×10<sup>13</sup></b>	<b>331</b>	<b>38/53</b>	<b>927</b>	<b>106/116</b>	<b>96</b>	<b>20/26</b>

\*Non-redundant genes.

†Amino acid sequence identity cut-off of 70%.

‡Non redundant genes/gene families.

between phyla. Our results are indicative that mobile efflux pumps were mobilized from *Proteobacteria*, while mobile ribosomal protection genes were mobilized from *Firmicutes* and/or *Actinobacteria*. The study presented here significantly expands the potential tetracycline resistome, and increases our knowledge about the diversity and evolutionary history of the tetracycline resistance genes.

## METHODS

### Gene prediction using fARGene

The genes presented in this study were predicted using the recently developed computation method fARGene, which can reconstruct both known and novel genes from fragmented metagenomic data [19]. Briefly, the method

starts by analysing raw reads using HMMER v.3.1b2 [22] and a hidden Markov model optimized for fragmented data. The reads classified as belonging to the gene class of interest are retrieved together with their read-pair and their quality is controlled using Trim Galore! v. 0.4.3 and a Phred score threshold of 30 [23]. The quality-controlled reads are assembled using SPAdes meta [24] and the assembled contigs then proceed to a second classification step where the threshold score is optimized for full-length genes. The contigs in which a gene of interest is predicted are retrieved and searched for ORFs using the National Center for Biotechnology Information (NCBI) ORFfinder (stand-alone version) [25]. The predicted ORFs are analysed with the model once again, in order to retrieve the ORFs with the highest score, and then passed as nucleotide and amino acid

sequences as the final output. In addition to the metagenomic data, the method was also used to analyse genomic data, in which the assembly step is skipped and ORFs are predicted with Prodigal v2.6.3 [26] around the region where the gene of interest is identified. For full details, see our previous work [19].

### Model creation and optimization

For this study, three hidden Markov models were used. Each was built to represent a tetracycline resistance mechanism: efflux pumps of major facilitator superfamily (MFS) group 1, ribosomal protection and enzymatic degradation. Note that efflux pumps of other groups were not included in the study due to the limited number of known genes. The sensitivity for each model was optimized using leave-one-out cross-validation, while the specificity was optimized by applying the model to a set of closely related sequences that lacked the desired phenotype. The efflux model was built using 18 proteins belonging to the tetracycline efflux group 1 [1, 4, 6]. The specificity of the model was estimated using 192 proteins from the MFS [27, 28]. The ribosomal protection gene model was built of 11 confirmed tetracycline resistance ribosomal protection genes [6]. To estimate the specificity, 143 proteins belonging to conserved domain family elongation factor Tu GTP were used (PFAM00009). To build the enzyme model, all proteins verified as tetracycline resistance enzymatic genes [6] were downloaded and used to recreate a phylogenetic tree using ETE NPR v. 0.9.32 with parameters ‘-w standard\_fasttree’ [29]. Based on the tree, the two genes *tet(37)* and *tet(34)* were removed due to their excessive diversion from the rest of the sequences. To enable the model to capture *tet(X)*-like genes, two variants of *tet(X)* were included in the model [*tet(X)* and *tetX3*]. As a negative dataset, 109 genes belonging to the conserved protein domain family UbiH (COG0654) were used. For a full list of genes used to build the three models see Table S1. The threshold scores for a fragment to be classified as positive were set based on manual inspection of the evaluation results, focusing on as high sensitivity as possible while still maintaining the highest possible specificity, with a minimum limit of 0.9. For full-length genes, the threshold scores were set so that both sensitivity and specificity were one for all models except for efflux pumps. For the efflux pumps, the gene *tet(42)* had a significantly lower score compared to the other genes in the cross-validation, and the resulting threshold was set so that all efflux pump genes except *tet(42)* would be captured by the model, resulting in a sensitivity of 0.94. The resulting sensitivity and specificity for the three models can be seen in Table S2.

### Post-processing and phylogenetic analysis

The predicted genes were grouped based on their resistance mechanisms and then clustered at 100% sequence identity using USEARCH v. 8.0.1445 [30], parameters ‘-cluster-fast -id 1’, to remove redundant genes. Then, the genes were clustered together with all previously known tetracycline resistance genes with the same resistance mechanism with

a 70% amino acid sequence identity using USEARCH with parameters ‘-cluster-fast -id 0.7’. The 70% sequence identity cut-off was chosen to ensure that previously known genes clustered consistently and, thus, that no gene family would incorrectly be annotated as new. The clusters that did not contain any previously known gene were called ‘novel family’ (NF) and the number of genes clustering in any of the novel families were counted as well as the number of families. The generated centroid files from the clustering were then aligned using MAFFT v. 7.273 [31] and phylogenetic trees were recreated using FastTree v. 2.1.10 with default settings [32]. To perform the phylum analysis, the bacterial species of each genome from NCBI RefSeq and Human Microbiome Project (HMP) genomic databases carrying at least one tetracycline resistance gene were extracted and counted. Duplicated species were removed and those remaining were organized into groups based on their phyla. The number of unique species in a phylum carrying a tetracycline resistance gene was then compared to the total number of unique species in this phylum in the original dataset using Fisher’s exact test.

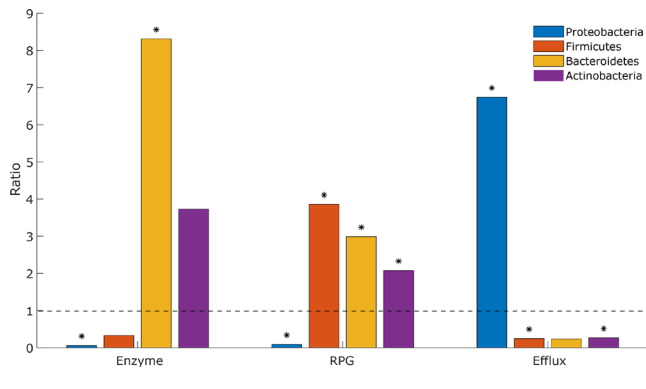
### Genetic context analysis

All genomes available in the GenBank assembly database, as well as all GenBank plasmid sequences (downloaded January 2019) were searched for the novel gene family amino acid sequences using DIAMOND v0.9.24.125 BLASTX with an identity cut-off of 70% [33]. For every hit, up to 20 kb upstream and downstream were extracted and annotated using PROKKA v1.12 [34]. Sequences annotated as hypothetical proteins were searched against the UniprotKB database using DIAMOND BLASTX with a 50% identity cut-off to investigate their potential function. The genetic environments of selected genes were further searched for mobile genetic elements using ISFinder and CDD search [35, 36].

### Experimental validation

Potential tetracycline resistance genes were synthesized by Thermo Fisher Scientific, using the GeneArt gene synthesis service and provided in the vector pMK. Each candidate gene sequence was equipped with the promoter  $P_{bla}$  plus ribosomal binding site upstream and the *rrnB* terminator T1 downstream to ensure constitutive expression. The negative control consisted of the expression elements without any gene, the positive controls were *tet(X)* [37] and *tet(A)* [38]. The pMK plasmids containing the resistance gene candidates were used to transform chemically competent *E. coli* TOP10 (Invitrogen, Thermo Fisher Scientific). The functionality of candidate genes was verified against both tetracycline (Sigma Aldrich, Merck) and tigecycline (Glentham Life Sciences). Minimum inhibitory concentrations (MICs) were determined by broth microdilution in cation-adjusted Mueller-Hinton medium. In accordance with European Committee on Antimicrobial Susceptibility Testing (EUCAST) guidelines (clinical breakpoint tables v. 9.0), medium for tigecycline tests was prepared freshly on the day of use. Serial dilutions of the tested antibiotics were





**Fig. 1.** Phylum analysis of genomes carrying tetracycline resistance genes. The analysis was based on genomes present in NCBI RefSeq and HMP genomic databases. The significance of the ratios was assessed using Fisher's exact test and results with  $P < 0.001$  are marked by asterisks. RPG, Ribosomal protection genes.

prepared in triplicate in 96-well plates and inoculated with  $5 \times 10^5$  cells  $\text{ml}^{-1}$  in each well [39] at a final volume of 200  $\mu\text{l}$ . After 24 h incubation at 37 °C and 180 r.p.m., optical density ( $\text{OD}_{650}$ ) was measured (Spectramax 340PC 384; Molecular Devices) and the MIC was defined as the lowest concentration of an antimicrobial that reduced growth to an  $\text{OD}_{650} \leq 0.2$  [40].

Growth behaviour of selected constructs was determined at different tetracycline concentrations using the OmniLog system (Biolog). Bacteria were grown in 96-well plates as described above with the addition of redox dye A. During 24 h incubation at 37 °C, metabolic activity was measured every 15 min in the form of a colour change caused by the reduction of the dye (OmniLog units). Growth curves and standard deviations were calculated from the mean of three independent experiments, each consisting of three technical replicates. The reported  $P$  values were calculated using a one-sided two-sample  $t$ -test with the values taken from the 24th hour of incubation at a tetracycline concentration of 0.45  $\mu\text{g ml}^{-1}$ . Since the current nomenclature of tetracycline resistance genes requires that the putative gene should be shown to be functional in its original host, none of the experimentally verified genes could be given an official name.

## RESULTS

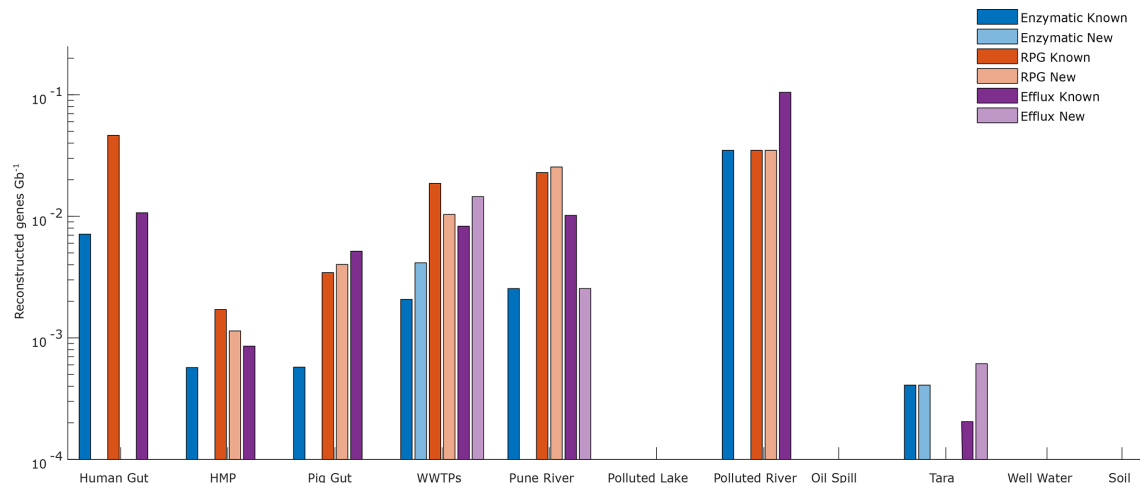
### Creation and evaluation of tetracycline resistance gene models

Genomic and metagenomic data were analysed for tetracycline resistance genes (both known and new) using fARGene, a computational method that applies probabilistic gene models to find ARGs (see our previously published work [19]). Three gene models were developed, corresponding to each of the three major mechanisms of tetracycline resistance: the MFS efflux pumps of type group 1, ribosomal protection genes and enzymatic degradation genes [1]. The models were trained using previously known tetracycline resistance genes,

and the sensitivity and specificity were optimized for both short reads and full-length genes using cross-validation. The resulting models all had a specificity of 1 for full-length genes, while the full-length sensitivity was 1 for all but the efflux pump model, which had a sensitivity of 0.94. The specificity for fragments (100 nucleotides) was at least 0.90 for all models (0.93, 0.90 and 0.97 for enzymatic, ribosomal protection and efflux pump genes, respectively), while the sensitivity ranged from 0.78 (ribosomal protection genes) to 0.96 (enzymatic degradation genes) (see Methods and Table S2 for more details). Using the gene models, fARGene was applied to more than 12 Tb of genomic and fragmented metagenomic data. This resulted in 1354 unique predicted tetracycline resistance genes, which were clustered into 195 gene families, of which 164 families were novel (amino acid sequence identity  $< 70\%$  to any previously known tetracycline resistance gene) (Tables 1 and S3). Genes encoding ribosomal protection genes were most common (927 genes in 116 families, of which 106 were novel), followed by efflux genes (331 genes in 53 families, of which 38 were novel) and enzymatic degradation genes (96 genes in 26 families, of which 20 were novel).

### Prediction of tetracycline-resistance genes in genome sequence data

The genomic data (NCBI genome, NCBI plasmids, NCBI nt, NCBI environmental, HMP genomic [41]) yielded a prediction of 1278 unique tetracycline resistance genes (Table 1). Among the 7376 complete genomes present in the NCBI genome database, 542 (7.3%) carried at least one predicted tetracycline-resistance gene, whereof 51 (0.7%) and 3 (0.04%) carried two and three genes, respectively. Ribosomal protection genes were most commonly predicted (314 genomes, 4.3%), followed by efflux pumps (244 genomes, 3.3%) and degradation enzymes (32 genomes, 0.4%). In addition, 458 (4.0%) plasmids present in the NCBI plasmids database carried at least one predicted tetracycline resistance gene (Table 1). Enrichment analysis, where the observed proportion of predicted resistant species in each phylum in NCBI genome (7376 genomes) and HMP genomic (1271 genomes) was compared to what was expected by chance, showed that the distribution of tetracycline-resistance genes was associated with taxonomy (see Methods). In particular, *Proteobacteria* were underrepresented carriers of both ribosomal protection genes (ratio 0.09 with  $P < 10^{-15}$ ) and enzymatic degradation genes (ratio 0.06,  $P = 1.4 \times 10^{-4}$ ), while efflux pump genes were overrepresented (ratio 6.74,  $P < 10^{-15}$ ) (Fig. 1). In contrast, *Firmicutes* showed an overrepresentation of ribosomal protection genes (ratio 3.86,  $P < 10^{-15}$ ), while *Bacteroidetes* showed an overrepresentation of both ribosomal protection and enzymatic degradation genes (ratio 2.99,  $P = 5.8 \times 10^{-11}$ , and ratio 8.31,  $P = 9.0 \times 10^{-6}$ , respectively). Finally, *Actinobacteria* showed an overrepresentation of ribosomal protection genes and an underrepresentation of efflux pump genes (ratio 2.08,  $P = 5.3 \times 10^{-7}$ , and ratio 0.27,  $P = 2.3 \times 10^{-4}$ , respectively).



**Fig. 2.** Number of reconstructed genes  $\text{Gb}^{-1}$  for each metagenomic dataset. A reconstructed gene was classified as novel if it had a sequence identity of  $<70\%$  to any previously known tetracycline resistance gene. RPG, Ribosomal protection genes.

### Prediction of tetracycline resistance genes in metagenomic data

From the metagenomic data, which consisted of human and animal-related sources, as well as samples from polluted and more pristine environments (Table 1), 98 non-redundant genes (123 genes in total) were predicted. Similar to the genomic data, ribosomal protection genes were the most common (61 unique genes corresponding to 24 families, of which 17 were novel), followed by efflux pumps (28 genes in 16 families, of which 7 were novel) and enzymatic degradation genes (9 genes in 8 families, of which 5 were novel) (Fig. 2). This corresponded to 0.006, 0.0024 and 0.0012 reconstructed genes  $\text{Gb}^{-1}$  for ribosomal protection genes, efflux pumps and degradation enzymes, respectively. Putative tetracycline resistance genes were found in all but four of the metagenomic datasets, but their abundance varied considerably between communities. The highest number of genes was found in the metagenomes from Swedish wastewater treatment plants (WWTPs) [42], sediments from Pune river (Pune, India) [43] and in pig microbiome (pig gut) [44]. Furthermore, the wastewater treatment plant metagenome was the only dataset that contained novel genes of all three resistance mechanisms. Efflux pump genes were especially abundant in wastewater treatment plants and the pig microbiome.

Ribosomal protection genes were only found in the human and animal-related datasets together with sediments from two polluted rivers (human gut, HMP metagenome, WWTPs, pig gut, Pune river and polluted river (Isakavagu, India)) [45–47]. Compared to efflux pump genes and enzymatic degradation genes, ribosomal protection genes were highly abundant in the human and pig microbiomes. Furthermore, only 4 out of 23 (17%) ribosomal protection genes found in the human microbiome (human gut and HMP metagenome) were novel. The human gut metagenome (human gut) only contained previously known genes, while the HMP metagenome did neither contain any novel enzymatic nor novel efflux pump

genes. Enzymatic degradation genes were rare overall, but had the largest proportion of novel genes (5 out of 9, 56%), which were all found in the wastewater treatment plant and marine environment (WWTPs and Tara Ocean [48]) metagenomes. The previously known tetracycline resistance genes found in the most environments were *tet(A)* (efflux), *tetB(P)* (ribosomal) and *tet(X)* (enzymatic), which all were found in five environments, respectively (Table S4). Finally, the analysis of data from oil-contaminated deep-sea water (oil spill) [49], the Kazipally lake in India polluted with wastewater from pharmaceutical manufacturing (polluted lake) [50], and bacterial communities in Indian soil and well water [51] did not yield any reconstructed full-length tetracycline resistance genes.

### Experimental verification

Next, we investigated whether some of the predicted genes induced a resistance phenotype in *E. coli*. In total, 17 predicted novel genes – 5 enzymatic degradation genes, 6 ribosomal protection genes and 6 efflux genes – were chosen from different parts of their respective phylogenetic tree. The genes were synthesized and inserted into an *E. coli* host, after which the MIC was determined (Table 2, Methods). For three genes, the growth rate in the presence of tetracycline was also measured (G231 in NF54 and G242 in NF55 ribosomal protection; G241 in NF21 efflux). For seven of the tested genes, expression in *E. coli* resulted in a resistance phenotype (Table 2). The highest increase in MIC was  $>256$ -fold and was observed for the efflux gene G256 in NF26. MIC assays were inconclusive for two of the ribosomal protection genes (increase in MIC between 1 and 4), but a more detailed analysis of growth rate over time showed that both genes increased growth significantly compared to the control (Fig. S1). Ten genes did not induce a resistance phenotype in *E. coli* (Table 2). Non-functionality in the tested host is, however, not evidence for lack of resistance function, as the genes may be functional in their native context and species

**Table 2.** Summary of the functional verification in *E. coli* of 17 predicted genes

Mechanism	Gene ID	Gene family	Fold-change tetracycline MIC	Increased growth rate*	Species	Most similar previously known	Amino acid identity to most similar known (%)
Enzyme	G8	NF3	–	ND	<i>Labilithrix luteola</i>	<i>tet</i> (55)	40.3
Enzyme	G35	NF11	–	ND	<i>Bacillus simplex</i>	<i>tet</i> (55)	39.7
Enzyme	G38	NF13	–	ND	<i>Legionella lansingensis</i>	<i>tet</i> (49)	41.5
Enzyme	G42	NF17	128	ND	Metagenome	<i>tet</i> (47)	54.0
Enzyme	G47	NF20	–	ND	<i>Niabella ginsenosidivorans</i>	<i>tet</i> (X)	63.9
Ribosomal protection	G139	NF28	–	ND	<i>Jiangella alkaliphila</i>	<i>otr</i> (A)	52.0
Ribosomal protection	G161	NF35	–	ND	<i>Cloacibacillus porcorum</i>	<i>tetB</i> (P)	34.7
Ribosomal protection	G231	NF54	1–4	p=1.6×10 <sup>-4</sup>	<i>Ensifer adhaerens</i> (P)†	<i>otr</i> (A)	44.5
Ribosomal protection	G242	NF55	1–4	p=5.8×10 <sup>-3</sup>	<i>Pseudomonas chlororaphis</i>	<i>otr</i> (A)	43.9
Ribosomal protection	G307	NF60	16	ND	<i>Clostridium saccharolyticum</i>	<i>tet</i> (44)	51.4
Ribosomal protection	G404	NF81	–	ND	<i>Rhizobium</i> sp. LPU83 (P)†	<i>otr</i> (A)	44.7
Efflux	G131	NF12	–	ND	<i>Burkholderia</i> sp. PAMC26561 (P)†	<i>tet</i> (A)	49.7
Efflux	G160	NF17	–	ND	<i>Ensifer adhaerens</i> (P)†	<i>tet</i> (41)	50.7
Efflux	G241	NF21	–	p=0.26	<i>Acinetobacter baumannii</i>	<i>tet</i> (41)	44.6
Efflux	G256	NF26	256–512	ND	Uncultured bacterium	<i>tet</i> (G)	62.8
Efflux	G281	NF34	128–256	ND	<i>Ochrobactrum anthropi</i>	<i>tet</i> (G)	55.3
Efflux	G296	Tet(A)/Tet(C)	2–16	ND	<i>Achromobacter xylosoxidans</i>	<i>tet</i> (A)	76.0

ND, Not determined.

\*P values calculated at hour 24 with a tetracycline concentration of 0.45 µg ml<sup>-1</sup>.

†Plasmid.

more closely related to their actual hosts. Even in *E. coli*, the resulting MIC for a given ARG can be highly strain-dependent [52]; hence, we put more emphasis on increases in MICs rather than on the absolute MIC values.

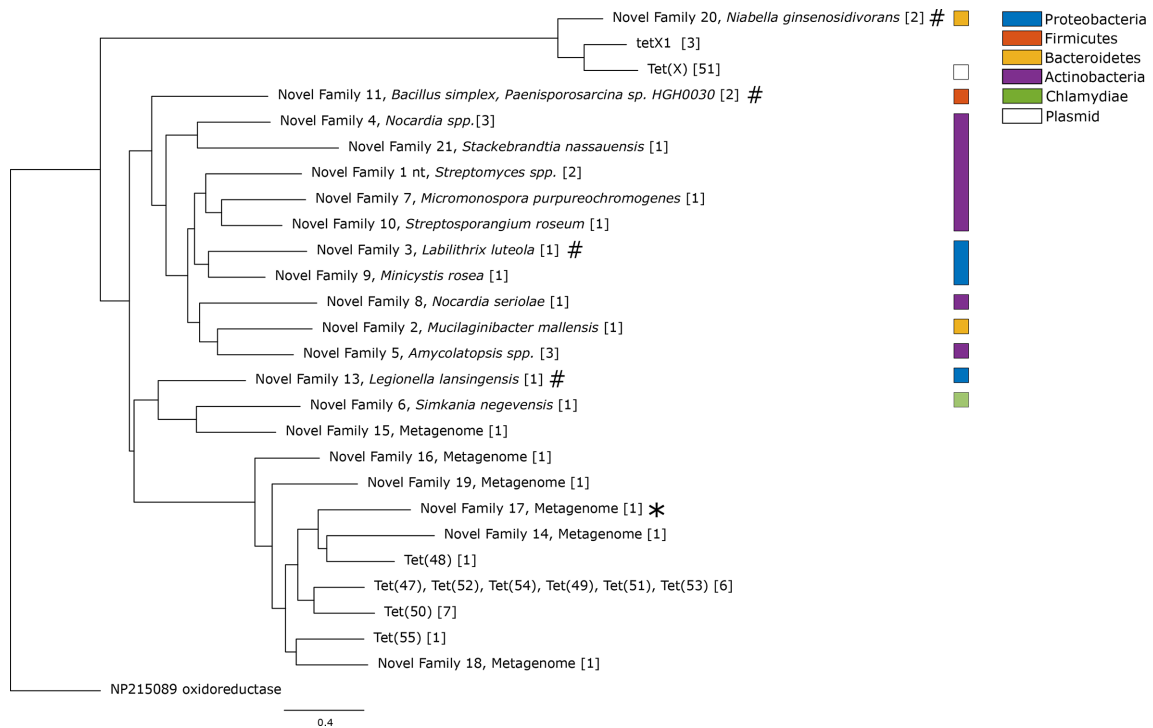
### Phylogenetic analysis

Next, all predicted tetracycline resistance genes (1354 unique genes, corresponding to 195 families) were used to recreate three phylogenetic trees, one for each resistance mechanism (see Methods). The phylogenetic tree of the enzymatic degradation genes, consisting of 26 gene families (Figs 3 and S2), was separated into two monophyletic groups. The first group consisted of the known *tet*(X), a full-length variant of the *tet*(X) homologue *tetX1* [53] and a novel gene family (NF20). In addition to the predicted novel genes, we investigated whether the full-length variant of *tetX1* induced a resistance phenotype in *E. coli*. However, similar to previous reports on

the truncated variant of *tetX1*, we did not see any increase in MIC against either tetracycline or tigecycline for either NF20 (G47) or the full-length *tetX1* variant [10]. The second group contained all other previously known enzymatic degradation genes [*tet*(47) to *tet*(55)] and was more diverse, with genes from five phyla and a wide range of metagenomic datasets (Fig. 3). Interestingly, the clade where *tet*(47) to *tet*(55) were located contained only two genes found in genomes, which were both pathogens: the proteobacterial species *Legionella lansingensis* and the *Chlamydiae* species *Simkania negevensis* [54, 55].

The phylogenetic analysis of the ribosomal protection genes was based on a total of 116 gene families (Figs 4 and S3). Similarly to the enzymatic degradation genes, the tree could be divided into two major groups. The first group contained all previously known ribosomal protection genes except for



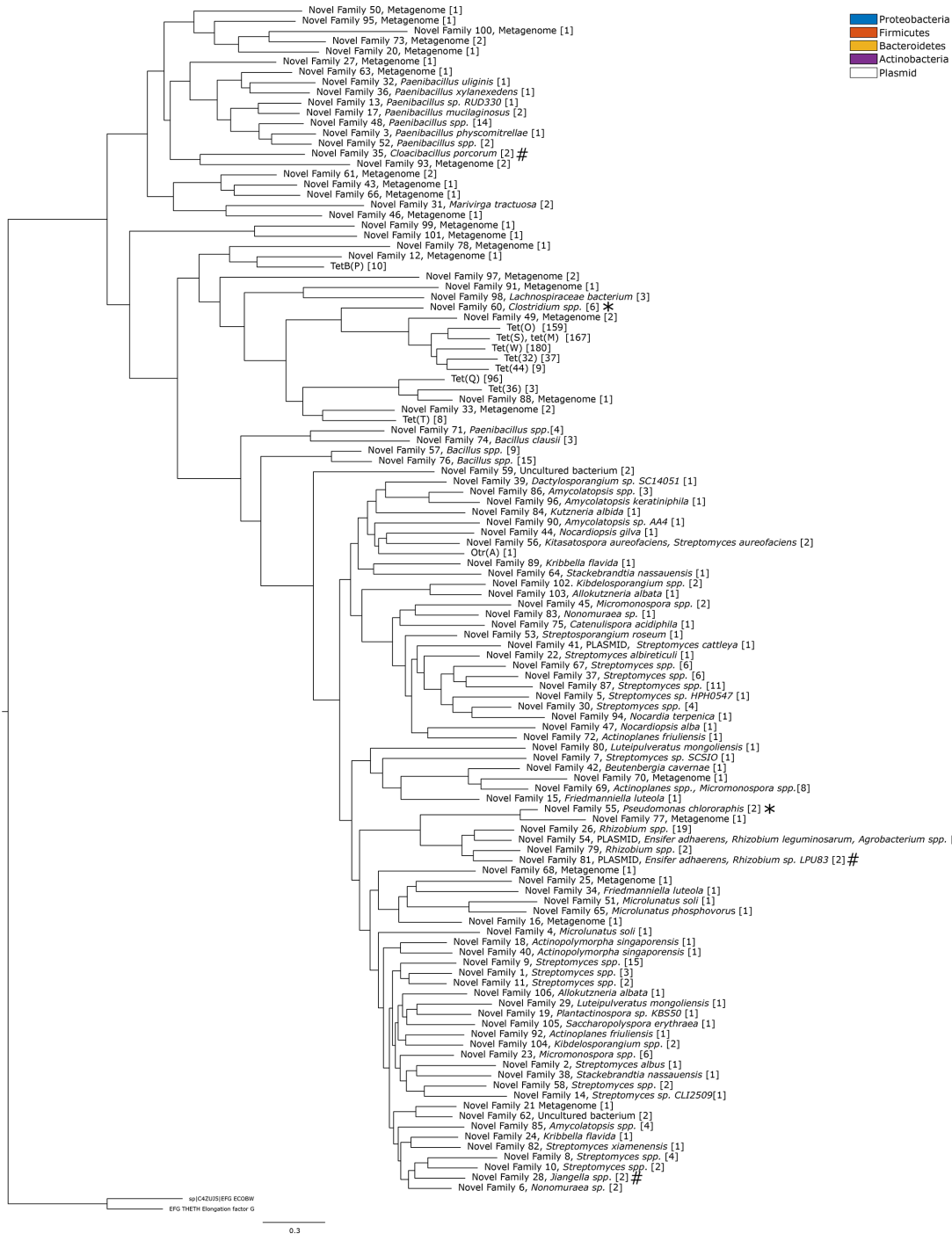


**Fig. 3.** Phylogenetic tree of the enzymatic tetracycline-resistance genes predicted in this study. The tree was recreated from all previously known enzymatic tetracycline-resistance gene families, together with gene families predicted in this study. Each gene family contains genes with >70% amino acid sequence identity, and the number of unique genes in each gene family is presented within the square brackets. The gene functional in *E. coli* is indicated by an asterisk and the gene families for which the tested genes did not function in *E. coli* are indicated by hash signs. The scale bar indicates number of substitutions per site.

*otr(A)*. The first group could be further separated into two distinct clades. Clade one in group one contained all of the previously known genes together with genes reconstructed from the bacterial communities from river sediment (Pune river) and wastewater treatment plants (WWTPs). The second clade in group one contained only novel genes, of which several were found in the pig gut metagenome. Relatively few of the genes in the first group were identified in genomic data, and this could not be associated with any host. However, two gene families that were the closest related to the previously known genes were both found in *Firmicutes* species: NF60, identified in *Clostridium saccharolyticum* and the pathogen *Clostridium sphenoides* [56]; and NF98, identified in *Lanchospiraceae* sp. The NF60 gene variant G307 in *Clostridium saccharolyticum* was functional when expressed in *E. coli*. Genes from NF60 were also commonly detected in metagenomes from both the environment and the human microbiome, suggesting that these genes are widely spread. Furthermore, several species from *Paenibacillus* – a genus responsible for the American foulbrood disease in honeybees and commonly treated with tetracycline – carried predicted novel families of ribosomal protection genes [57]. The second group of the phylogenetic tree contained the known *otr(A)* gene, which was first discovered in an oxytetracycline-producing *Streptomyces* spp. [4, 58, 59]. This part of the tree consisted mainly of genes located in other *Actinobacteria* (56 novel families). A small

clade within the second group with novel gene families from proteobacterial species contained two genes located on plasmids (G232 in NF54, *Ensifer adhaerens*, and G404 in NF81, *Rhizobium* sp. LPU83) indicative of horizontal gene transfer across bacterial phyla. In the same clade, another resistance gene was present in *Pseudomonas chlororaphis*. This species is root-colonizing and has previously been shown to promote plant growth and to have biopesticide properties [60]. When expressed in *E. coli*, the gene located on the plasmid in *Ensifer adhaerens* and the one located in *Pseudomonas chlororaphis* did not result in an increased MIC, but both showed significantly increased growth rates in the presence of tetracycline compared to the negative control (Table 2, Fig. S1).

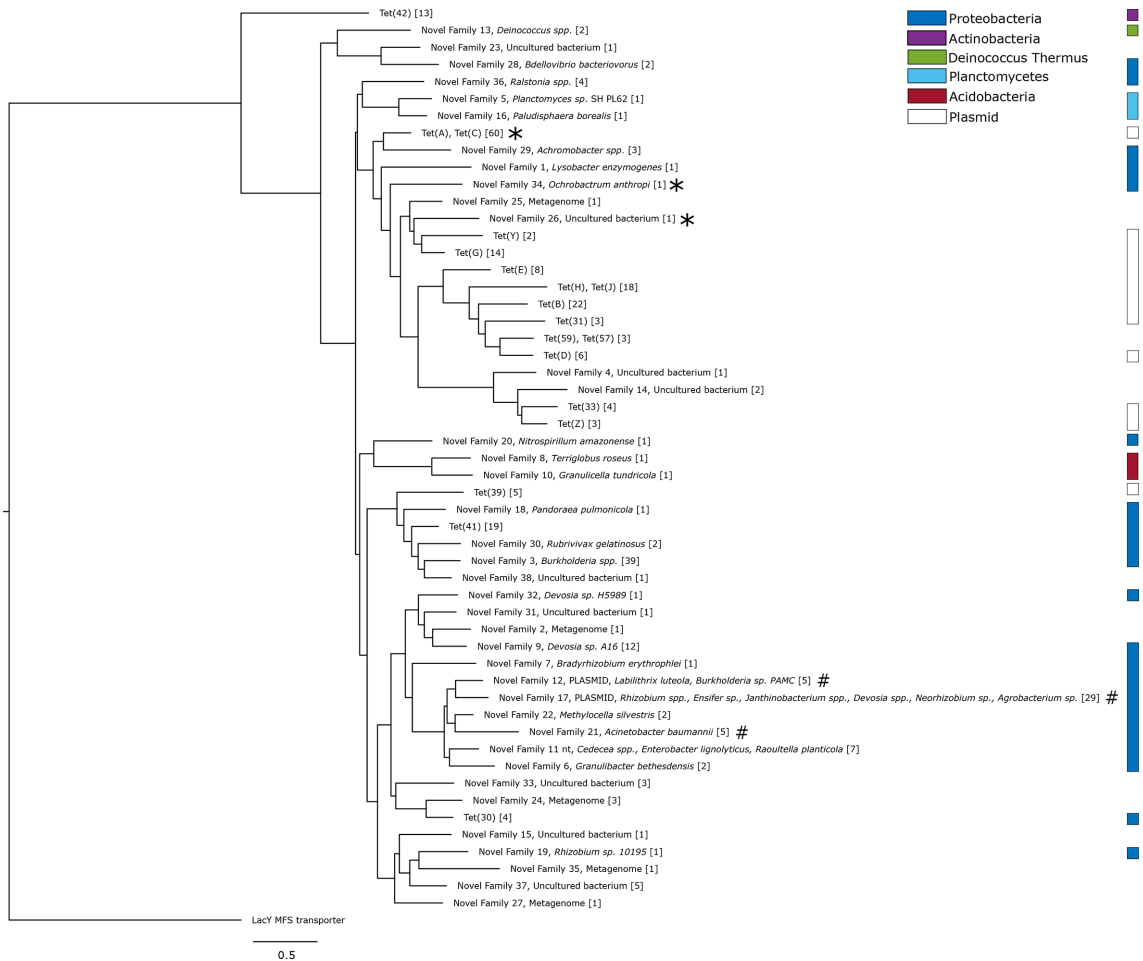
The phylogenetic tree of the efflux pump genes consisted mainly of genes found in genomic data. Most genes appeared in *Proteobacteria*, but some were also found in species of *Acidobacteria*, *Actinobacteria*, *Planctomycetes* and *Deinococcus-Thermus* (Figs 5 and S4). The known efflux pump genes were distributed over the entire tree, but several accumulated in a clade containing novel genes from *Alphaproteobacteria* and *Betaproteobacteria* (including the pathogens *Ochrobactrum anthropi* and *Achromobacter* spp.). Furthermore, we identified two novel families (NF12, NF17) with genes located on plasmids. The plasmid carrying a variant of NF17 was found in *Ensifer adhaerens*, but located on a different



**Fig. 4.** Phylogenetic tree of the ribosomal protection genes predicted in this study. The tree was recreated from all previously known ribosomal protection gene families, together with gene families predicted in this study. Each gene family contains genes with >70% amino acid sequence identity, and the number of unique genes in each family is presented within the square brackets. The gene families functional in *E. coli* are indicated by asterisks and the gene families for which the tested genes did not function in *E. coli* are indicated by hash signs. The scale bar indicates number of substitutions per site.

plasmid separate from the ribosomal protection gene present in the same species. The gene family NF17 also contained genes from five additional genera of *Alphaproteobacteria*: *Rhizobium*, *Janthinobacterium*, *Devosia*, *Agrobacterium* and

*Neorhizobium*. Finally, a plasmid harbouring NF12 was found in *Burkholderia* sp. PAMC 26561 (G131), while other members of the same gene family were found in *Labilithrix luteola* and the marine metagenome (Tara Ocean).



**Fig. 5.** Phylogenetic tree of the efflux pump genes predicted in this study. The tree describes the previously known efflux gene families of MFS group 1, together with gene families predicted in this study. Each gene family contains genes with >70% amino acid sequence identity, and the number of unique genes in each group is presented within the square brackets. The gene families functional in *E. coli* are indicated by asterisks and the gene families for which the tested genes did not function in *E. coli* are indicated by hash signs. The scale bar indicates number of substitutions per site.

## DISCUSSION

In this study, we have presented a large-scale prediction of novel tetracycline resistance genes. This was performed by using sensitive probabilistic models developed to specifically identify novel tetracycline resistance genes, even when they have low sequence identity to known resistance genes [19]. Large amounts of genomic and metagenomic data were screened, and the predicted genes were used to infer their evolutionary history. Our study identified all previously known tetracycline resistance genes of the three modelled main mechanisms, together with 164 predicted novel gene families. These results expand the potential tetracycline resistome significantly. In addition, experimental validation of 17 predicted genes showed that 7 of these resulted in a resistance phenotype when expressed in *E. coli*, indicating that they are functional in this host.

The number of predicted genes differed considerably between the three main resistance mechanisms investigated in this

study: efflux pump genes, ribosomal protection genes and enzyme degradation genes. Most common were ribosomal protection genes (927), followed by efflux pumps genes (331) and enzymatic degradation genes (96). Efflux pump genes were found in a larger number of phyla (nine), compared to six and five phyla for ribosomal protection genes and enzymatic degradation genes, respectively. There were also clear taxonomic differences between genes from the three resistance mechanisms. Efflux pump genes were overrepresented in *Proteobacteria* with more than 8% of the unique proteobacterial species in HMP genomic and NCBI RefSeq carrying an efflux pump gene, while enzymatic degradation and ribosomal protection genes were underrepresented in *Proteobacteria*. These resistance mechanisms were more common in other phyla, in particular *Bacteroidetes*, and for ribosomal protection genes, also *Actinobacteria* and *Firmicutes*. In contrast, only three species of *Bacteroidetes* were carrying efflux pump genes. Note, however, that the enzymatic degradation genes

found in *Bacteroidetes* belonged to only four gene families, which included *tet(X)*, *tetX1* and novel family 20 that is closely related to *tet(X)* and *tetX1*. A similar pattern could be seen for the ribosomal protection genes, where all genes located in *Bacteroidetes* were either previously known or from a single novel family (NF31). Thus, even though enzymatic degradation genes and ribosomal protection genes were common in species from *Bacteroidetes*, these numbers represent mainly known mobile genes and, thus, only a limited part of the full diversity of these gene classes. However, it should be pointed out that the genomic data used in this study is likely biased towards well-studied organisms, in particular those related to human health; thus, the results are not representative of the full diversity of the tetracycline resistome.

The relative number of reconstructed tetracycline resistance genes varied between environments. The human and pig microbiomes, as well as environments that are potentially polluted by human faecal matter, all contained high levels of genes from all three resistance mechanisms. In these environments, the enzymatic genes were heavily dominated by *tet(X)* and the previously studied *tet(X)*-like gene *tetX1* [10, 50]. A similar pattern was observed for ribosomal protection and efflux pumps genes, where the human microbiome contained almost exclusively previously known genes. Here, the abundance was especially high in the intestinal tract where *Bacteroidetes* is dominant [45], adding additional support to the hypothesis that bacteria from this phylum seem to carry a relatively limited set of mostly known tetracycline resistance genes. The environments that are likely to contain less faecal matter contained no tetracycline resistance genes at all, except for marine microbial communities in the Tara Ocean dataset, which contained both known and novel efflux pumps and enzymatic genes. An environment with a high abundance and diversity of novel resistance genes may indicate a higher risk for transfer events into recipients that will enable spread to pathogens. Based on our analysis, such environments include the human and animal-associated microbiomes, as well as the bacterial communities from the wastewater treatment plant (Table 1). Note, however, that aside from the Tara Ocean data, the environmental datasets used in this study were also the least deeply sequenced (especially the polluted lake dataset) and considering that the abundance needs to be fairly high for a gene to be reconstructed from a metagenomic dataset, it is likely that some resistance genes were below the limit of detection [61, 62]. Thus, we cannot exclude that there is a much larger diversity of novel tetracycline resistance genes in these communities than what is possible to detect with existing data.

The phylogenetic analysis showed the presence of potentially mobile novel tetracycline resistance genes. For example, the phylogenetic group containing the known ribosomal protection gene *otr(A)*, which was dominated by *Actinobacteria*, contained a small clade of gene families from proteobacterial species (Fig. 4). One of these gene families (NF81) comprised a plasmid-borne gene from *Rhizobium* sp. LPU83 (G404), while another gene family (NF54) harboured a gene on a plasmid from *Ensifer adhaerens* (G231). Other members of

this family are located in the genomes of *Rhizobium* sp. and *Agrobacterium* spp. None of these genes showed a consistent increase in MIC when expressed in *E. coli*. However, the gene located on the *Ensifer adhaerens* plasmid was tested in the growth assay and allowed significantly faster growth than the control, indicating that it is indeed functional. One additional gene family (NF55) from the same clade was found in *Pseudomonas chlororaphis*, and was also shown to induce a resistance phenotype in *E. coli*. The members of the genus *Pseudomonas* are known to have highly plastic genomes and to commonly harbour conjugative elements carrying ARGs [63]. Interestingly, the gene was located on a ~13 kb long insert present in three *Pseudomonas chlororaphis* strains, but absent in three other strains and all other *Pseudomonas* species. This insert also contained distant homologues to fosfomycin-resistance genes (33.6% amino acid identity to FosA) and macrolides (34.9% amino acid identity to OleC). Taken together, this indicates that there has been a horizontal gene transfer event of ribosomal protection genes from species within *Actinobacteria* to *Proteobacteria*. Since the genes are present on mobile genetic elements or, as in the case of *Pseudomonas chlororaphis*, highly plastic genomic regions, there is an apparent risk that they can spread further, including to human pathogens.

There were also indications of new mobile efflux pump genes. For example, two of the novel gene families had genes located on plasmids, one in *Burkholderia* sp. PAMC 26561 (G131 in NF12) and one in *Ensifer adhaerens* (G160 in NF17), although not the same strain as the one carrying a predicted ribosomal protection gene. Furthermore, genes within NF17 were found in the genomes of *Janthinobacterium* spp. and several species within the order *Rhizobiales* (*Agrobacterium* sp., *Rhizobium* spp., *Devosia* spp. and *Neorhizobium* sp.). Many of these species are symbiotic nitrogen-fixating bacteria associated with root nodules, and are known to often carry plasmids or megaplasmids harbouring genes essential for symbiosis with the host plant [64, 65]. Functional verification showed, however, that these genes did not result in an increased MIC when expressed in *E. coli*. Moreover, several of the predicted new efflux pump genes were located in opportunistic or pathogenic bacteria, e.g. *Acinetobacter baumannii*, *Achromobacter insolitus*, *O. anthropi*, *Pandoraea pulmonicola* and *Burkholderia cenocepacia* [66–70]. Among these, *Acinetobacter baumannii* is a notorious pathogen harbouring a wide range of mobile resistance genes against clinically relevant antibiotics, including tetracyclines [71]. The gene family identified in *Acinetobacter baumannii* was only present in 25 of 155 *Acinetobacter baumannii* strains in NCBI RefSeq. However, when a member of this gene family (G241) was expressed in *E. coli*, we did not observe any increase in MIC. Closer analysis of the regions surrounding the identified gene in the 25 strains of *Acinetobacter baumannii* revealed an insertion sequence (IS<sub>Aba6</sub>) located in direct connection to the gene in one of the strains, suggesting that the gene could potentially be mobile [72].

Analysis of the phylogenetic tree in the neighbourhood around the previously known mobile tetracycline resistance



genes provided further insights into their recent evolutionary history. Phylogenetic analysis of the ribosomal protection genes showed, for example, that *otr(A)* was most closely related to genes found in various *Actinobacteria*, including many species from the genus *Streptomyces* (Fig. 4). The chromosomal gene closest to *otr(A)* was located in the species *Amycolatopsis* sp., but the evolutionary distance to *otr(A)* was still substantial. The *otr(A)* gene has previously been found in *Streptomyces* spp. and *Mycobacterium* spp. [4]. and our results suggest, in accordance with the literature, that this gene has likely been mobilized from an unknown *Actinobacteria* species [4, 73]. Indeed, the majority of the clinically used antibiotics are produced by *Actinobacteria*, and several classes of resistance genes are hypothesized to originate from this phylum [74, 75]. The other known mobile ribosomal protection genes were located in a single clade where only two of the novel gene families were found in isolated species, both *Firmicutes* (NF98, *Lachnospiraceae* spp., and NF60, *Clostridium* spp.). Ribosomal protection genes were in general common in *Bacteroidetes*, but no species in this phylum contained a chromosomal non-mobile gene that was close evolutionary to any mobile ribosomal protection gene. Therefore, it is plausible that the mobile ribosomal protection genes originated elsewhere – based on the current evidence most likely *Firmicutes* – and were horizontally transferred to *Bacteroidetes*. It should be pointed out that a wide range of species from both phyla are naturally occurring in animal and human microbiomes, suggesting that there is both ecological connectivity and tetracycline selection pressure that may enable horizontal gene transfer.

The phylogenetic tree of the efflux pump genes (Fig. 5) was heavily dominated by genes from *Proteobacteria* and all of the known mobile efflux pumps had their closest relative within this phylum. This strongly suggests that mobile efflux pump genes were mobilized from *Proteobacteria*. However, the evolutionary distances were large and no mobile efflux pump gene with high sequence identity to a non-mobile efflux pump gene was found, suggesting that the origin – or origins – are species with yet uncharacterized genomes. Note, however, that one of the genes closest related to any previously known mobile efflux pump gene was located in *O. anthropi* (*Betaproteobacteria*), an opportunistic pathogen known to thrive in a large variety of habitats [76]. Furthermore, the *O. anthropi* gene showed high resistance against tetracycline when expressed in *E. coli*.

The enzymatic resistance genes had few matches in the genomic data, and based on the available information, it is difficult to assess their origin with certainty. We did, however, identify a new family related to *tet(X)* (NF20), which, like the *tetX1* and the *tet(X)* family, was found in *Bacteroidetes* species (Fig. 3). Furthermore, the *tetX1* genes were always found together with a copy of *tet(X)*, except for a reconstructed *tetX1* gene from the pig microbiome. Interestingly, this was not the case for the genes in NF20, which appeared both in *Niabella ginsenosidivorans* and in a metagenome from a wastewater treatment plant. We were not able to identify any elements associated with known mechanisms for horizontal gene

transfer, suggesting that these genes may be chromosomal. Considering the high degree of mobility among *tet(X)*-like genes in *Bacteroidetes* [12], it is hard to assess if this is the origin of the gene or if the genes were horizontally transferred to this part of the taxonomic tree.

The method, fARGene, used in this study applies probabilistic gene models in the form of hidden Markov models to identify novel resistance genes [22]. The models have been optimized to accurately discriminate between homologues with and without a resistance phenotype to keep the number of false positives low [19]. Nevertheless, all genes identified in this study are predictions and, therefore, it is necessary to experimentally verify if they can induce resistance phenotypes and, if that is the case, in what hosts they are functional. In this study, we selected 17 of the predicted genes for verification in *E. coli* and among those 7 were found to be functional. Note, however, that even some well-known tetracycline resistance genes that are functional in other species do not induce a significant resistance phenotype in *E. coli*. In particular, ribosomal protection genes, which are rare in *Proteobacteria*, confer only low levels of resistance when expressed in *E. coli* [4]. Indeed, several ribosomal protection genes showed only a small increase in MIC, but the functionality of two of these genes could be verified with the considerably more sensitive growth assay. Therefore, it cannot be excluded that other genes which did not result in an increased MIC in *E. coli* may be functional and, potentially, be much more potent in other hosts. In addition, experimental verification was performed based on the exact predicted protein sequences without codon optimization; therefore, an impaired expression may explain the lack of function in *E. coli*. Furthermore, we cannot exclude that some of the predicted genes may have other molecules as their natural substrate and, thus, demonstrate a lower efficacy for tetracyclines. Also, we cannot disregard the possibility that some of the predicted genes may be false positives; thus, they should be considered putative until further experimental verification has been made.

## Conclusion

Screening of more than 12 Tb of genomic and metagenomic data was used to identify 195 families of tetracycline resistance genes, of which 164 families were novel predictions and previously uncharacterized. Phylogenetic analysis suggested that mobile tetracycline resistance genes were mobilized from different parts of the taxonomical tree, where ribosomal protection genes originated from *Firmicutes* and *Actinobacteria*, while efflux pump genes originated from proteobacterial species. Enzymatic genes were scarce; therefore, their evolutionary origin remains unclear. Several of the novel mobile genes showed patterns of horizontal transfer between bacterial phyla. This study expands the known tetracycline resistome, and describes the reservoir of resistant determinants maintained by environmental and commensal bacteria. If mobilized and transferred to pathogens, these genes could reduce the efficacy of existing and future tetracycline antibiotics and thereby pose a significant threat to our ability to treat bacterial infections.



### Funding information

This research was supported by the Swedish Research Council (VR) (2016–06512, 2018–02835, 2019–03482) and the Wallenberg Foundation. Funding sources took no part in the design, analysis nor interpretation of the results.

### Author contributions

F. B., A. M., T. Ö., A. J., D. G. J. L. and E. K. designed the study and developed the approach. F. B. and A. M. collected the data. F. B. and A. M. created and optimized the probabilistic models. F. B. ran the analysis pipeline and recreated the phylogenetic trees. F. B., S. E. and E. K. performed the genetic context analysis. F. B., M.-E. B., T. Ö., A. J., D. G. J. L. and E. K. analysed the results. M.-E. B. performed the experimental verification. All authors discussed the results and its implications. F. B. and E. K. drafted the manuscript. All authors edited and approved the final manuscript.

### Conflicts of interest

The authors declare that there are no conflicts of interest.

### References

1. Thaker M, Spanogiannopoulos P, Wright GD. The tetracycline resistome. *Cell Mol Life Sci* 2010;67:419–431.
2. Hao H, Cheng G, Iqbal Z, Ai X, Hussain HI et al. Benefits and risks of antimicrobial use in food-producing animals. *Front Microbiol* 2014;5:288.
3. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2017. *EFSA J* 2019;17:e05598.
4. Chopra I, Roberts M. Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev* 2001;65:232–260.
5. Forsberg KJ, Patel S, Wenciewicz TA, Dantas G. The tetracycline destructases: a novel family of tetracycline-inactivating enzymes. *Chem Biol* 2015;22:888–897.
6. Roberts MC. GenBank numbers; 2018. <https://faculty.washington.edu/marilynr/>
7. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MO et al. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* 2012;337:1107–1111.
8. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J et al. Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* 2010;8:251–259.
9. Moore IF, Hughes DW, Wright GD. Tigecycline is modified by the flavin-dependent monooxygenase TetX. *Biochemistry* 2005;44:11829–11835.
10. Yang W, Moore IF, Koteva KP, Bareich DC, Hughes DW et al. TetX is a flavin-dependent monooxygenase conferring resistance to tetracycline antibiotics. *J Biol Chem* 2004;279:52346–52352.
11. Guiney DG, Hasegawa P, Davis CE. Expression in *Escherichia coli* of cryptic tetracycline resistance genes from *Bacteroides* R plasmids. *Plasmid* 1984;11:248–252.
12. Aminov RI. Evolution in action: dissemination of tet(X) into pathogenic microbiota. *Front Microbiol* 2013;4:192.
13. Leski TA, Bangura U, Jimmy DH, Ansumana R, Lizewski SE et al. Multidrug-resistant tet(X)-containing hospital isolates in Sierra Leone. *Int J Antimicrob Agents* 2013;42:83–86.
14. Speer BS, Bedzyk L, Salyers AA. Evidence that a novel tetracycline resistance gene found on two *Bacteroides* transposons encodes an NADP-requiring oxidoreductase. *J Bacteriol* 1991;173:176–183.
15. D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic resistome. *Science* 2006;311:374–377.
16. Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J* 2009;3:243–251.
17. Marathe NP, Janzon A, Kotsakis SD, Flach CF, Razavi M et al. Functional metagenomics reveals a novel carbapenem-hydrolyzing mobile beta-lactamase from Indian river sediments contaminated with antibiotic production waste. *Environ Int* 2018;112:279–286.
18. Riesenfeld CS, Goodman RM, Handelsman J. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* 2004;6:981–989.
19. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Ostell J et al. Genbank. *Nucleic Acids Res* 2018;46:D41–D47.
20. Boulud F, Berglund F, Flach CF, Bengtsson-Palme J, Marathe NP et al. Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. *BMC Genomics* 2017;18:682.
21. Berglund F, Marathe NP, Osterlund T, Bengtsson-Palme J, Kotsakis S et al. Identification of 76 novel B1 metallo- $\beta$ -lactamases through large-scale screening of genomic and metagenomic data. *Microbiome* 2017;5:134.
22. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol* 2011;7:e1002195.
23. Krueger F. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. Cambridge: The Babraham Institute; 2015.
24. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
25. NCBI. ORFfinder; 2018. <https://www.ncbi.nlm.nih.gov/orffinder/>
26. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;11:119.
27. Paulsen IT, Brown MH, Skurray RA. Proton-dependent multidrug efflux systems. *Microbiol Rev* 1996;60:575–608.
28. Reddy VS, Shlykov MA, Castillo R, Sun EI, Saier MH. The major facilitator superfamily (MFS) revisited. *FEBS J* 2012;279:2022–2035.
29. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python environment for tree exploration. *BMC Bioinformatics* 2010;11:24.
30. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;26:2460–2461.
31. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
32. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;5:e9490.
33. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;12:59–60.
34. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.
35. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 2006;34:D32–D36.
36. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res* 2015;43:D222–D226.
37. Speer BS, Shoemaker NB, Salyers AA. Bacterial resistance to tetracycline: mechanisms, transfer, and clinical significance. *Clin Microbiol Rev* 1992;5:387–399.
38. Allmeier H, Cresnar B, Greck M, Schmitt R. Complete nucleotide sequence of Tn1721: gene organization and a novel gene product with features of a chemotaxis protein. *Gene* 1992;111:11–20.
39. NCCLS. *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically, Approved Standard*. Wayne, PA: National Committee for Clinical Laboratory Standards; 2006.
40. Schuurmans JM, Nuri Hayali AS, Koenders BB, ter Kuile BH. Variations in MIC value caused by differences in experimental protocol. *J Microbiol Methods* 2009;79:44–47.
41. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC et al. A catalog of reference genomes from the human microbiome. *Science* 2010;328:994–999.

42. Bengtsson-Palme J, Hammarén R, Pal C, Östman M, Björleinius B et al. Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci Total Environ* 2016;572:697–712.
43. Marathe NP, Pal C, Gaikwad SS, Jonsson V, Kristiansson E et al. Untreated urban waste contaminates Indian river sediments with resistance genes to last resort antibiotics. *Water Res* 2017;124:388–397.
44. Xiao L, Estellé J, Kiilerich P, Ramayo-Caldas Y, Xia Z et al. A reference gene catalogue of the pig gut microbiome. *Nat Microbiol* 2016;1:16161.
45. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–214.
46. Kristiansson E, Fick J, Janzon A, Grabic R, Rutgersson C et al. Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS One* 2011;6:e17038.
47. Qin J, Li Y, Cai Z, Li S, Zhu J et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 2012;490:55–60.
48. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C et al. A holistic approach to marine eco-systems biology. *PLoS Biol* 2011;9:e1001177.
49. Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J* 2012;6:1715–1727.
50. Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, Larsson DGJ. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Front Microbiol* 2014;5:648.
51. Rutgersson C, Fick J, Marathe N, Kristiansson E, Janzon A et al. Fluoroquinolones and qnr genes in sediment, water, soil, and human fecal flora in an environment polluted by manufacturing discharges. *Environ Sci Technol* 2014;48:7825–7832.
52. Rosano GL, Ceccarelli EA. Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front Microbiol* 2014;5:172.
53. Whittle G, Hund BD, Shoemaker NB, Salyers AA. Characterization of the 13-kilobase *ermF* region of the *Bacteroides* conjugative transposon CTnDOT. *Appl Environ Microbiol* 2001;67:3488–3495.
54. Thacker WL, Dyke JW, Benson RF, Havlichek DH, Robinson-Dunn B et al. *Legionella lansingensis* sp. nov. isolated from a patient with pneumonia and underlying chronic lymphocytic leukemia. *J Clin Microbiol* 1992;30:2398–2401.
55. Vouga M, Baud D, Greub G. *Simkania negevensis*, an insight into the biology and clinical importance of a novel member of the *Chlamydiales* order. *Crit Rev Microbiol* 2017;43:62–80.
56. Kelesidis T, Tsiodras S. *Clostridium sphenoides* bloodstream infection in man. *Emerg Infect Dis* 2011;17:156–158.
57. Alippi AM, León IE, López AC. Tetracycline-resistance encoding plasmids from *Paenibacillus larvae*, the causal agent of American foulbrood disease, isolated from commercial honeys. *Int Microbiol* 2014;17:49–61.
58. Doyle D, McDowall KJ, Butler MJ, Hunter IS. Characterization of an oxytetracycline-resistance gene, *otrA*, of *Streptomyces rimosus*. *Mol Microbiol* 1991;5:2923–2933.
59. Dittrich W, Schrempp H. The unstable tetracycline resistance gene of *Streptomyces lividans* 1326 encodes a putative protein with similarities to translational elongation factors and Tet(M) and Tet(O) proteins. *Antimicrob Agents Chemother* 1992;36:1119–1124.
60. Kupferschmied P, Maurhofer M, Keel C. Promise for plant pest control: root-associated pseudomonads with insecticidal activities. *Front Plant Sci* 2013;4:287.
61. Gweon HS, Shaw LP, Swann J, De Maio N, AbuOun M et al. The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples. *Environ Microbiome* 2019;14:7.
62. Lanza VF, Baquero F, Martínez JL, Ramos-Ruiz R, González-Zorn B et al. In-depth resistome analysis by targeted metagenomics. *Microbiome* 2018;6:11.
63. Pohl S, Klockgether J, Eckweiler D, Khaledi A, Schniederjans M et al. The extensive set of accessory *Pseudomonas aeruginosa* genomic components. *FEMS Microbiol Lett* 2014;356:235–241.
64. Pérez Carrascal OM, VanInsberghe D, Juárez S, Polz MF, Vinuesa P et al. Population genomics of the symbiotic plasmids of sympatric nitrogen-fixing *Rhizobium* species associated with *Phaseolus vulgaris*. *Environ Microbiol* 2016;18:2660–2676.
65. Alexandre A, Laranjo M, Oliveira S. Natural populations of chickpea rhizobia evaluated by antibiotic resistance profiles and molecular methods. *Microb Ecol* 2006;51:128–136.
66. Li G, Zhang T, Yang L, Cao Y, Guo X et al. Complete genome sequence of *Achromobacter insolitus* type strain LMG 6003T, a pathogen isolated from leg wound. *Pathog Dis* 2017;75:ftx037.
67. Alnor D, Frimodt-Møller N, Espersen F, Frederiksen W. Infections with the unusual human pathogens *Agrobacterium* species and *Ochrobactrum anthropi*. *Clin Infect Dis* 1994;18:914–920.
68. Kanjee R, Koreishi AF, Tanna AP, Goldstein DA. Chronic post-operative endophthalmitis after cataract surgery secondary to vancomycin-resistant *Ochrobactrum anthropi*: case report and literature review. *J Ophthalmic Inflamm Infect* 2016;6:25.
69. Degand N, Lotte R, Decondé Le Butor C, Segonds C, Thouverez M et al. Epidemic spread of *Pandoraea pulmonicola* in a cystic fibrosis center. *BMC Infect Dis* 2015;15:583.
70. Rosales-Reyes R, Sánchez-Gómez C, Ortiz-Navarrete V, Santos-Preciado JI. *Burkholderia cenocepacia* induces macrophage apoptosis to enter macrophages. *Biomed Res Int* 2018;2018:4271560.
71. Peleg AY, Seifert H, Paterson DL. *Acinetobacter baumannii*: emergence of a successful pathogen. *Clin Microbiol Rev* 2008;21:538–582.
72. Mahillon J, Chandler M. Insertion sequences. *Microbiol Mol Biol Rev* 1998;62:725–774.
73. Benveniste R, Davies J. Aminoglycoside antibiotic-inactivating enzymes in actinomycetes similar to those present in clinical isolates of antibiotic-resistant bacteria. *Proc Natl Acad Sci USA* 1973;70:2276–2280.
74. van der Heul HU, Bilyk BL, McDowall KJ, Seipke RF, van Wezel GP. Regulation of antibiotic production in Actinobacteria: new perspectives from the post-genomic era. *Nat Prod Rep* 2018;35:575–604.
75. Tolba S, Egan S, Kallifidas D, Wellington EM. Distribution of streptomycin resistance and biosynthesis genes in streptomycetes recovered from different soil sites. *FEMS Microbiol Ecol* 2002;42:269–276.
76. Romano S, Aujoulat F, Jumas-Bilak E, Masnou A, Jeannot JL et al. Multilocus sequence typing supports the hypothesis that *Ochrobactrum anthropi* displays a human-associated subpopulation. *BMC Microbiol* 2009;9:267.
77. O'Leary NA, Wright MW, Brister JR, Ciufu S, Haddad D et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;44:D733–D745.