Water Research 202 (2021) 117384

Contents lists available at ScienceDirect

Water Research

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

Research Paper

Predicting antibiotic resistance gene abundance in activated sludge using shotgun metagenomics and machine learning

Yuepeng Sun [a], Bertrand Clarke [b], Jennifer Clarke [b,c], Xu Li [a,*]

[a] Department of Civil and Environmental Engineering, University of Nebraska-Lincoln, 900N. 16th St, W150D Nebraska Hall, Lincoln, NE 68588-0531, United States
[b] Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE 68583, United States
[c] Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE 68588

ARTICLE INFO

Keywords:
Activated sludge
Antibiotic resistance genes
Machine learning
Random forests
Wastewater treatment plants

ABSTRACT

While the microbiome of activated sludge (AS) in wastewater treatment plants (WWTPs) plays a vital role in shaping the resistome, identifying the potential bacterial hosts of antibiotic resistance genes (ARGs) in WWTPs remains challenging. The objective of this study is to explore the feasibility of using a machine learning approach, random forests (RF's), to identify the strength of associations between ARGs and bacterial taxa in metagenomic datasets from the activated sludge of WWTPs. Our results show that the abundance of select ARGs can be predicted by RF's using abundant genera (*Candidatus* Accumulibacter, *Dechloromonas, Pesudomonas*, and *Thauera*, etc.), (opportunistic) pathogens and indicators (*Bacteroides, Clostridium*, and *Streptococcus*, etc.), and nitrifiers (*Nitrosomonas* and *Nitrospira*, etc.) as explanatory variables. The correlations between predicted and observed abundance of ARGs (*erm*(B), *tet*(O), *tet*(Q), etc.) ranged from medium ($0.400 < R^2 < 0.600$) to strong ($R^2 > 0.600$) when validated on testing datasets. Compared to those belonging to the other two groups, individual genera in the group of (opportunistic) pathogens and indicator bacteria had more positive functional relationships with select ARGs, suggesting genera in this group (e.g., *Bacteroides, Clostridium*, and *Streptococcus*) may be hosts of select ARGs. Furthermore, RF's with (opportunistic) pathogens and indicators as explanatory variables were used to predict the abundance of select ARGs in a full-scale WWTP successfully. Machine learning approaches such as RF's can potentially identify bacterial hosts of ARGs and reveal possible functional relationships between the ARGs and microbial community in the AS of WWTPs.

1. Introduction

Antibiotic resistance is a major threat to public health and the proliferation of antibiotic resistance genes (ARGs) in the environment is believed to contribute to the problem (Martinez, 2008). Wastewater treatment plants (WWTPs) receiving municipal wastewater have been regarded as a key reservoir of ARGs (Bouki et al., 2013). The discharge of treated wastewater and disposal of biosolids from WWTPs can introduce ARGs to water and soil (Jia et al., 2017), altering the magnitude and composition of the resistomes in receiving environments (Xue et al., 2019).

The composition of the resistome in an environment can be strongly correlated to the composition of the microbiome (Forsberg et al., 2014; Yin et al., 2019; Zhang et al., 2016, 2018). The resistome in WWTPs is correlated with the composition of the microbial community therein, which is ultimately determined by the characteristics of the influent to WWTPs and the design and operation of WWTPs (Wu et al., 2018; Yin et al., 2019). The composition of the microbial community can explain 68.2% of the ARG variations among sewage sludge according to redundancy analyses (Zhang et al., 2016). Hence, characterizing the composition of the microbial community may shed light on resistome composition in WWTPs.

Associating ARGs to their bacterial hosts in complex environments is challenging. Efforts have been reported to identify potential bacterial hosts for ARGs using network (Guo et al., 2017) and binning analyses (Liu et al., 2019) on metagenomic data. Network analysis can reveal taxa-ARGs associations by calculating their Spearman's rank correlation coefficient. For instance, using network analyses, Guo and coworkers identified strong Spearman's correlations between seven ARGs and *Dechloromonas* in wastewater (Guo et al., 2017). However, spurious correlations (both false-positive and false-negative correlations) between variables may result when the sample size is small (Guo et al.,

* Corresponding author.
E-mail address: xuli@unl.edu (X. Li).

https://doi.org/10.1016/j.watres.2021.117384
Received 26 March 2021; Received in revised form 6 June 2021; Accepted 21 June 2021
Available online 26 June 2021
0043-1354/© 2021 Elsevier Ltd. All rights reserved.

2017; Rice et al., 2020). By grouping contigs with similar abundance and sequence composition into the same bin, binning analysis can reveal taxa-ARGs association by identifying the genome bins carrying both ARGs and taxonomic marker genes (Liu et al., 2019; Ma et al., 2016). Using binning analysis, Liu et al. (2019) speculated *Mycobacterium, Nitrospira*, and *Nitrosomonas* as multi-drug ARGs hosts in WWTPs treating landfill, municipal and car washing wastewater. For binning analyses, annotation at the genus level may be difficult due to low coverage of draft genome and lack of reference sequences for taxonomy annotation (Liu et al., 2019). Besides, the reconstructed genomes from metagenomics may not capture strain variation. These genomes may miss low-abundance species and introduce biases for quantitative analysis (Ju and Zhang 2015; Rice et al., 2020).

Machine learning provides various alternative methods to search for potential associations between bacterial taxa and ARGs. In particular, random forests (RF's) is a machine learning algorithm that can be used to predict resistome composition based on microbiome data. The variable importance factors of RF's can indicate taxa with higher "importance scores" in predicting individual ARGs. RF's have been developed to identify the association between temperature and microbial composition in WWTPs (Wu et al., 2019) and the correlation of ARGs in wastewater with socioeconomic, health and environmental factors (Hendriksen et al., 2019). Consequently, it is reasonable to apply the machine learning framework to search for the associations between ARGs and taxa (i.e., potential bacterial hosts). Indeed, the increasing number of metagenomic datasets in public repository makes it possible to test the feasibility of this approach in the effort to associate microbiome and resistome in WWTPs.

The objective of this study is to explore the feasibility of using RF's to identify the strength of associations between ARGs and bacterial taxa in metagenomic datasets from the activated sludge of WWTPs. Through systematic review, 21 peer-reviewed publications, corresponding to 248 metagenomic datasets from WWTPs in 10 countries, were selected. Metagenomic datasets were trained using RF's to predict the abundance of select ARGs with explanatory variables of [1] abundant genera; [2] (opportunistic) pathogens and indicator bacteria; and [3] nitrifying bacteria (i.e., nitrifiers). The computed RF's were then validated on testing datasets to assess their performance. Furthermore, the RF's were used to predict ARGs abundance in WWTPs using bacterial taxa data. The findings from this study demonstrate the potential of using a machine learning approach to identify potential bacterial hosts of ARGs in complex environments such as the activated sludge in WWTPs.

2. Material and methods

2.1. Systematic review

Five databases, namely Compendex, Biological Science Research, Web of Science, Pubmed, and Scopus, were searched in August 2019. Search strategies were developed using different keywords and syntax according to the search rules of each database (**Table S1**). Only publications that met the following criteria were included: applying the metagenomic approach to study wastewater, focusing on full-scale wastewater treatment systems, and containing metagenomic sequences that are publically accessible through GenBank or MG-RAST. More details on search strategy and selection criteria are described in SI (**Figure S1**). Information about the selected papers can be found in **Table S2**.

2.2. Bioinformatics analysis

We focused on sequences from activated sludge (AS) in this work because the highest amount of data was available for this sample type compared to other sample types (i.e., influent, effluent, and digested sludge). The AS samples in the original studies were collected from 33 WWTPs in 10 countries.

Raw shotgun metagenomic sequence reads downloaded from public databases were trimmed using Trim_galore (Krueger, 2012). Cutadapt (Martin, 2011) and FastQC (Andrews, 2010) were used to remove low quality reads and adapter sequences. Trimmed reads were used to carry out taxonomy classification with Kaiju (Menzel et al., 2016). Trimmed reads were also annotated for ARGs using Resistance Gene Identifier bwt (RGI bwt) based on reference data from the Comprehensive Antibiotic Resistance Database (CARD) (Alcock et al., 2020). A sequence was annotated as an ARG sequence if it shared 100% sequence identity with a sequence in the database and had an alignment length over 25 amino acids (Kristiansson et al., 2011).

The abundance of ARGs was reported in the unit of "ppm" (i.e., one ARG-like read in one million metagenomic sequencing reads) according to Yang et al. (2013). Abundance of a specific taxon in a sample was calculated using the ratio of the total number of reads matched to taxon by Kaijiu and the total reads within a quality filtered library (Baral et al., 2018). The abundances of genera and ARGs were log-transformed prior to generating RF's.

2.3. Resistance prediction

2.3.1. Variable preparation

The abundance of bacterial genera from three groups (i.e., [1] abundant bacteria, [2] (opportunistic) pathogens and indicator bacteria, and [3] nitrifiers) and that of select ARGs were defined as *explanatory variables* and *responses*, respectively, for the RF's (**Fig. 1**). Based on the metagenomic datasets of AS, the most abundant genera in activated sludge were narrowed down to 11 genera (Group 1), *Bradyrhizobium, Candidatus* Accumulibacter, *Dechloromonas, Hyphomicrobium, Methyloversatilis, Mycolicibacterium, Nitrosomonas, Nitrospira, Pseudomonas, Streptomyces*, and *Thauera*. Group 2 included 29 (opportunistic) pathogens commonly detected in WWTPs (Cai and Zhang 2013; Li et al., 2015b) and 3 indicator bacteria (i.e., *Clostridium, Enterococcus*, and *Escherichia*). Group 3 contained 7 commonly occurring nitrifying bacterial genera in WWTPs, *Nitrosococcus, Nitrosomonas, Nitrosospira, Nitrobacter, Nitrococcus, Nitrospina*, and *Nitrospira* (Juretschko et al., 1998). Among the one hundred most abundant ARGs conferring resistance to five major antibiotic families (i.e., beta lactams, glycopeptides, macrolides-lincosamides-streptogramines (MLS), sulfonamides, and tetracyclines), we selected the 22 most abundant dominant ARGs based on their occurrence across all AS samples.

2.3.2. RF's development, validation and application

Based on the metagenomic data on AS samples, RF's were developed using each of the three groups of genera and the 22 select ARGs selected in 2.3.1 as variables and responses, respectively (Fig. 1). The computing was carried out using the *caret* package in R, which contains techniques for data splitting, pre-processing, model tuning (a trial-and-error process to determine the best set of hyperparameters), and variable importance evaluation (Kuhn, 2008). RF's average was developed in following steps. First, the data were randomly split into 60% (training) and 40% (testing) subsets using the *createDataPartion* function. Next, with the training dataset, RF's were generated using the *train* function in *caret*. The importance of the variables in each RF's was assessed using the *varImp* function on a 0 - 100 scale. To flag problems of overfitting or selection bias, a five-fold cross-validation step was set up using the *trainControl* function for each RF's by randomly partitioning variables into five sub-datasets of roughly equal sizes followed by estimation of accuracy based on remaining sub-datasets. Third, the RF's were validated on testing datasets. Finally, results of validation on training and testing datasets were gathered by the *gather* function in the *tidyr* package and visualized using the *ggplot* function in the *ggplot* package. Linear regression was then used to assess the accuracy of the RF's. Specifically, the R^2 and Root Mean Square Error (RMSE) values from regressing the predicted on the observed values were calculated to indicate the prediction accuracy of RF's. In general, high values of R^2 (a relative

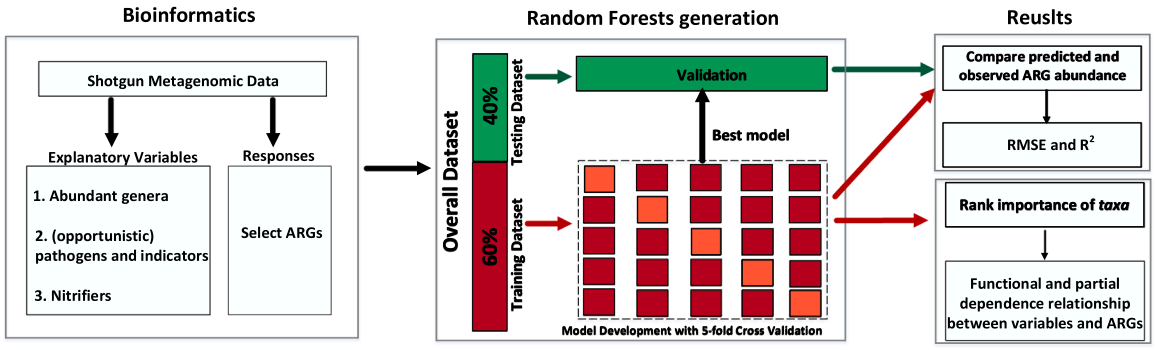

Fig. 1. Steps of Random Forests generation and validation. A five-fold cross-validation step was set up by randomly partitioning explanatory variables into five sub-datasets (red) of roughly equal sizes followed by estimation of accuracy based on remaining sub-datasets (orange).

measure of model fit) and low values of RMSE (an absolute measure of model fit) indicate good fit of the RF's. Because the abundance of individual ARGs (i.e., responses of the RF's) varied greatly within and across sequence libraries, RMSE is a less useful measure than R^2 in assessing model fitness in this study. Hence, while RMSE values are reported, the strength of associations between predicted and observed values in linear regression model context were defined as "weak", "moderate" and "strong" based on R^2 values of 0.3 – 0.4, 0.4 – 0.6, and 0.6 – 1.0, respectively (Hermans et al., 2020).

2.4. Statistical analysis and data visualization

T-tests were used to determine if the mean difference between two variables (e.g., the abundance of ARG families in various sample types) was statistically significant. The abundances of genera and ARGs were visualized in heatmaps using the *pheatmap* package in R. Partial dependence plots were drawn using the *partialPlot* function by in the *randomForest* package. The Nadaraya-Waston (NW) regression estimator was used to identify functional relationships between individual taxa and individual ARGs using the *npregfast* package in R.

3. Results and discussion

3.1. Occurrence and abundance of ARGs in WWTPs

Following the systematic review (**Table S1** and **Figure S1**), 21 publications indexed in the five citation databases met the selection criteria (**Table S2**). A total of 248 shotgun metagenomic datasets were downloaded from the GenBank and MG-RAST databases as FASTQ files in October 2019. Of the 248 datasets, 141 datasets contained DNA sequences on activated sludge (AS), 24 datasets contained DNA sequences on digested sludge (AD), 39 datasets contained DNA sequences on influent, and 44 datasets contained DNA sequences on effluent. The number of DNA reads per sample ranged from 1166,697 to 499,150,364, averaged at 65,736,667. The number of mapped ARG reads ranged from 82 to 168,956 per library, averaged at 19,359. The average ARG abundance in influent and effluent samples was 437.6 and 466.8 ppm, respectively (Fig. 2**a**). By contrast, the average abundance of ARGs in AS and AD were 197.2 and 205.3 ppm, respectively. The ARG abundance in influent was significantly higher than that in AD and AS, and was significantly lower than that in effluent ($p<0.05$). Given that there are

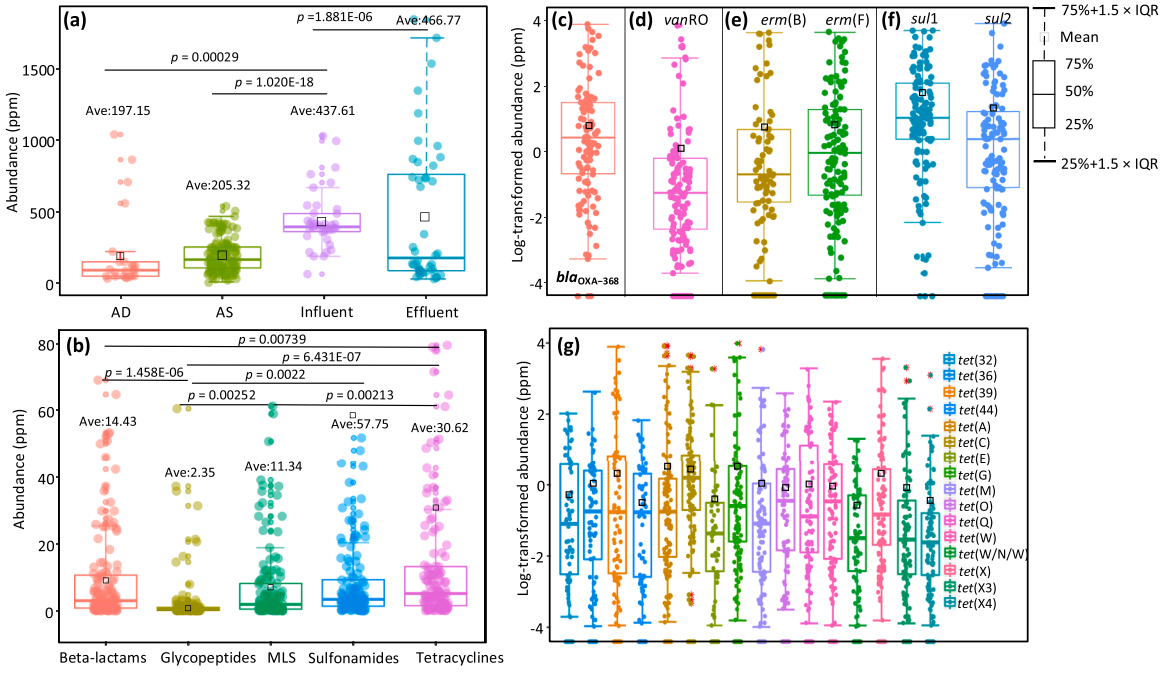

Fig. 2. ARG profiles in WWTPs, (a) total abundance of ARGs in various types of samples in WWTPs; (b) the abundance of select ARG families in activated sludge; log-transformed abundance of select ARGs conferring resistance to (c) beta-lactams, (d) glycopeptides, (e) MLS, (f) sulfonamides, and (g) tetracyclines. IQR represents interquartile range. AD and AS represent samples from anaerobic digester and activated sludge, respectively. p-values are calculated from t-tests. Ave indicates average abundance.

significantly more metagenomic datasets available for AS than for the other sample types, we focused on the metagenomic datasets from AS in this study.

The abundance of five commonly studied ARG families (i.e., those corresponding to beta-lactams, glycopeptides, MLS, sulfonamides, and tetracyclines) in AS is shown in Fig. 2**b.** Tetracycline and sulfonamide resistance genes are the most abundant ARG families, with average abundances of 30.6 and 57.8 ppm, respectively. Significant difference in abundance was observed between ARGs conferring resistance to beta-lactams vs. glycopeptides, beta-lactams vs. tetracyclines, and glycopeptides vs. MLS ($p<0.05$), etc. According to metagenomic studies that were not used in this work, the ARG abundance in AS ranges between 24 – 708 ppm (Christgen et al., 2015; Li et al., 2015a; Tang et al., 2016; Yang et al., 2014). The ARG abundance obtained from this study fell within this range.

Among the 100 most abundant ARGs belonging to the five ARG families in the datasets (**Figure S2**), 22 ARGs were selected for further analyses based on their abundance across all 141 metagenomic libraries on AS (Fig. 2**c - 2 g**): 1 beta-lactam resistance gene $bla_{OXA-368}$ (Fig. 2**c**); 1 glycopeptide resistance gene *van*RO (Fig. 2**d**); 2 MLS resistance genes *erm*(B) and *erm*(F) (Fig. 2**e**); 2 sulfonamide resistance genes *sul*1 and *sul*2 (Fig. 2**f**); 16 tetracycline resistance genes (i.e., *tet*(32), *tet*(36), *tet*(39), *tet*(44), *tet*(A), *tet*(C), *tet*(E), *tet*(G), *tet*(M), *tet*(O), *tet*(Q), *tet*(W), *tet*(W/N/W), *tet*(X), *tet*(X3), *tet*(X4)) (Fig. 2**g**). The mean abundance of the *tet* genes ranged from 0.3 for *tet*(W/N/W) to 4.2 ppm for *tet*(A). The mean abundance of *sul*1 and *sul*2 were 36.2 and 24.1 ppm, respectively, while the mean abundance was 1.7 ppm for *van*RO, 6.3 ppm for $bla_{OXA-368}$, 2.6 ppm for *erm*(B), and 4.8 ppm for *erm*(F).

3.2. Association of ARGs with abundant genera

The relationship between ARGs and abundant genera were investigated by RF's with the group of abundant genera as explanatory variables and individual ARGs as responses. The top 100 genera were identified according to their relative abundance across all metagenomic libraries (**Figure S3**). Among them, 17 genera had a mean relative

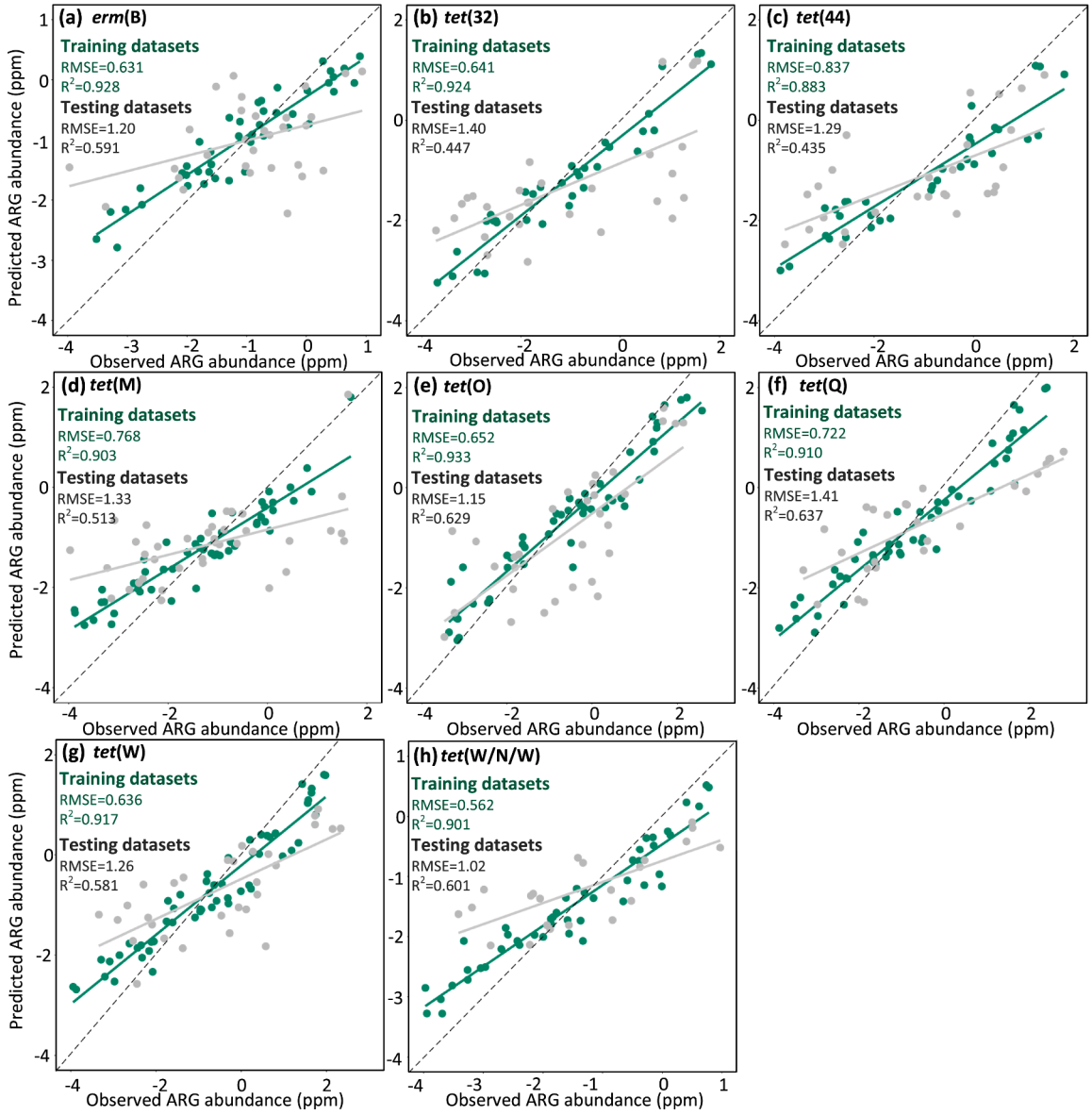

Fig. 3. Observed ARG abundance was plotted against the ARG abundance predicted using the Random Forests generated with abundant genera as explanatory variables. Only the eight ARGs with R^2 higher than 0.400 in the testing datasets are shown (a - h). The values shown in the plots are log-transformed abundance. Dashed lines indicate the theoretical lines for perfect predictions. Data (dots) and models (line) were separately plotted for the training datasets (green) and the testing datasets (gray). RMSE and adjusted R^2 are reported in the panels.

abundance higher than 0.3%. To avoid multicollinearity, 6 of the 17 genera were excluded, as they were highly correlated to each other with correlation coefficients higher than 0.6. The remaining 11 genera were used in RF's (**Figure S4**).

Our results show that for the training dataset the group of 11 abundant genera could explain over 88% of the variations in ARGs abundance with R^2 ranging 0.883 – 0.938 and RMSE ranging 0.547 – 0.874 (**Table S3**). When applied to the testing dataset, the RF's exhibited a wider range of R^2 values ranging from 0.0216 to 0.637 (**Table S3**). For 8 of 22 ARGs tested, moderate to strong associations were shown between predicted to observed ARGs abundance with R^2 ranging 0.435 – 0.637 and RMSE ranging 1.02 – 1.41 (Fig. 3). Particularly, for the RF's developed for *tet*(O), *tet*(Q), and *tet*(W/N/W), the associations between predicted and observed ARGs abundance had R^2 values higher than 0.600 and RMSE lower than 1.41, indicating strong associations between these ARGs and the abundant genera tested.

According to the importance score, the most important genera within the explanatory variables were *Candidatus* Accumulibacter, *Nitrosomonas, Nitrospira, Dechloromonas, Pseudomonas*, and *Thauera* (**Figure S4**), which were previously reported as potential ARG hosts (**Table 1**) (Guo et al., 2017; Sui et al., 2018; Xia et al., 2019; Zhou et al., 2019). For genus-ARG pairs with importance factors higher than 90 in RF's (**Figure S4**), partial dependence plots were generated to show the functional relationship between an individual genus and the predicted ARG abundance (**Figure S5**). For the genus-ARG pairs included in the partial dependence analysis, most of the predicted ARG abundance exhibited positive dependence on the individual genera tested, except for *Candidatus* Accumulibacter vs. *tet*(44) and *tet*(M), as well as *Dechloromonas* vs. *tet*(36) and *tet*(X). Similarly, for genus-ARG pairs with importance factors higher than 90 in RF's (**Figure S4**), the abundance of individual genus was regressed on the observed ARG abundance using the Nadaraya-Waston (NW) estimator (**Figure S6**). The NW plots show nonlinear functional relationships between observed ARGs and individual abundant genera, e.g., observed abundance of $bla_{OXA-368}$ and *tet*(A) increased with *Pseudomonas*, and the observed abundance of $bla_{OXA-368}$, *tet*(X), *sul*1 and *sul*2 increased with *Thauera*.

Nitrosomonas and *Nitrospira* are nitrifying bacteria, while *Candidatus* Accumulibacter (Wu et al., 2019), *Dechloromonas* (Wang et al., 2020), *Pseudomonas* (Scherson et al., 2013) and *Thauera* (Wang et al., 2020) are denitrifying bacteria, suggesting ARGs may be linked to bacteria involved in nitrogen transformation in WWTPs (Wang et al., 2020). *Thauera* can survive the pressure of tetracycline and kanamycin below the minimal inhibitory concentrations (Zhao et al., 2019), and correlate with *sul*2, *tet*(A), *tet*(O), *tet*(W) (Du et al., 2019), and *tet*(X) (Wang et al., 2020). *Candidatus* Accumulibacter and *Dechloromonas* belong to the global core bacterial community in WWTPs and play important roles in organics and nitrogen transformation (Wu et al., 2019). These two genera could be enriched under long-term exposure to tetracycline and sulfamethoxazole in lab scale reactors treating wastewater, and their abundance were correlated with *sul*1, *tet*(A), *tet*(C), *tet*(L), *tet*(O), and *tet*(X) at the end of enrichment (Du et al., 2019), suggesting the ability to acquire ARGs under selective pressure (Wang et al., 2019).

Table 1
Association between genera and select ARGs from published papers and this study.

Taxa	ARGs	Methodology	References
Nitrosomonas	*dfr*K, *pen*A, *van*HAc2, and *van*R-F	Network	Guo et al., 2017
Nitrosomonas	*bac*A, *sul*2	Binning	Liu et al., 2019
Nitrosomonas	bla_{TEM}, *ere*A, *erm*(B), *erm*(F), *sul*2, and *tet*(X)	Network	Sui et al., 2018
Nitrosomonas	*sul*1, *tet*(32), *tet*(36), *tet*(A), *tet*(O), *tet*(W), *tet*(W/N/W), and *tet*(X)	Machine learning	This study
Nitrospira	*sul*1, *tet*(G)	Network	Sui et al., 2018
Nitrospira	$bla_{OXA-368}$, *erm*(B), *tet*(32), *tet*(C), *tet*(O), *tet*(W), *tet*(W/N/W), and *tet*(X4)	Machine learning	This study
Pseudomonas	*van*R-C	Network	Guo et al., 2017
Pseudomonas	*sul*1, *tet*(A), *tet*(C), and *tet*(O)	Network	Du et al., 2019
Pseudomonas	$bla_{OXA-368}$, *erm*(F), *tet*(E) and *tet*(A)	Machine learning	This study
Candidatus Accumulibacter	*Peb-EC*, SFO-1, *van*R-B, and *van*R-C,	Network	Guo et al., 2017
Candidatus Accumulibacter	*sul*1, *sul*2, *tet*(A), *tet*(C), *tet*(L), *tet*(O), and *tet*(X)	Network	Du et al., 2019
Candidatus Accumulibacter	$bla_{OXA-368}$, *tet*(36), *tet*(44), *tet*(C), and *tet*(M)	Machine learning	This study
Thauera	*acrA, MacA*, and NPS-1	Network	Zhao et al., 2019
Thauera	*sul*2, *tet*(A), *tet*(O), and *tet*(W)	Network	Du et al., 2019
Thauera	*sul*1, *sul*2, *tet*(G), and *tet*(X)	Machine learning	This study

3.3. Association of ARGs with (opportunistic) pathogens and indicator bacteria

RF's were developed using the group of (opportunistic) pathogens (e.g., *Acinetobacter, Bacillus, and Bordetella*, etc.) and indicator bacteria (i.e., *Clostridium, Enterococcus*, and *Escherichia*) as explanatory variables and individual ARGs as responses. Our results show that for the training dataset the group of (opportunistic) pathogens and indicators explained over 91% of the variations in ARGs abundance with R^2 ranging 0.910 – 0.964 and RMSE ranging 0.455 – 0.821 (**Table S3**). When applied to the testing dataset, the RF's exhibited a wider range of R^2 values, ranging from 0.0123 to 0.654, between predicted and observed ARGs (**Table S3**). The predicted ARG abundance was strongly associated with the observed abundance of *erm*(B), *tet*(39), *tet*(M), and *tet*(Q) with R^2 higher than 0.600 (Fig. 4). Additionally, moderate associations were observed between predicted to observed abundance for *sul*2, *tet*(32), *tet*(36), *tet*(44), *tet*(A), *tet*(W), and *van*RO with R^2 ranging 0.400 – 0.600 (Fig. 4).

As shown in **Figure S7**, *Bacteroides, Clostridium, Escherichia, Enterococcus, Eubacterium, Klebsiella* and *Streptococcus* were the genera with importance scores higher than 90 for multiple ARGs, e.g., *tet*(32), *tet*(44), *tet*(M), *tet*(O), *tet*(W) and *tet*(W/N/W). For the 7 genera and the ARGs that they had high importance scores for, partial dependence plots show that the predicted abundance of all the ARGs included in this analysis exhibited positive dependence on the genera included (**Figure S8**). These genus-ARG pairs were further analyzed using the NW estimator (Fig. 5). The functional relationships between observed ARG abundance and the genera of (opportunistic) pathogens and indicators were obviously non-linear. The abundance of observed ARG abundance exhibited generally positive relationships with the abundance of individual genera, with the exception of *Enterococcus* vs *tet*(E) (Fig. 5). The NW curves are very steep in some cases. Possible threshold effects may be further investigated by fitting a single tree model using techniques such as those reported in Chipman et al. (1998).

The functional relationship between ARGs and genera identified in this section have also been reported in studies employing other approaches. Network analyses showed strong co-occurrence between *tet*(Q) and *Bacteroides* as well as *Escherichia* in fecal environmental samples (Li et al., 2015c); between *Clostridium* and *tet*(32) (Li et al., 2015c); *Streptococcus* and *erm*(B) in AS samples (Ju et al., 2016). Using the Mantel test and canonical correspondence analysis, Jia and co-authors (2017) reported that tetracycline resistance genes were mainly carried by *Bacteroides, Streptococcus*, and *Clostridium* in livestock wastewater. Lee et al. (2020) reported that the relative abundances of fecal bacteria including *Bacteroides* and *Clostridium* were linearly correlated with ARG abundance ($R^2 = 0.21$) in river water.

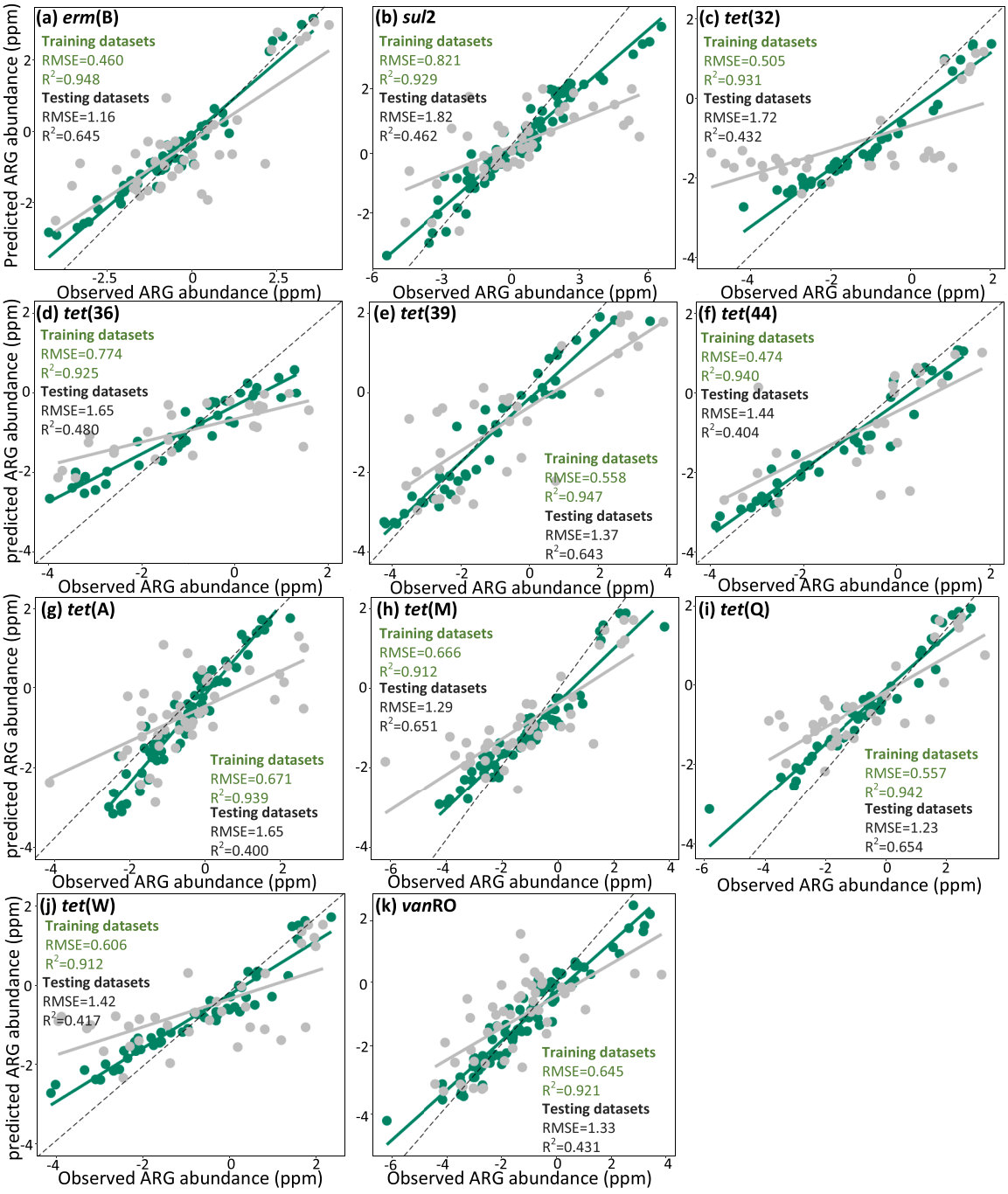

Fig. 4. Observed ARG abundance was plotted against the ARG abundance predicted using the Random Forests generated with (opportunistic) pathogens and indicators as explanatory variables. Only the eleven ARGs with R^2 higher than 0.400 in the testing datasets are shown (a - l). The values shown in the plots are log-transformed abundance. Dashed lines indicate the theoretical lines for perfect predictions. Data (dots) and models (line) were separately plotted for the training datasets (green) and the testing datasets (gray). RMSE and adjusted R^2 are reported in the panels.

Our findings are also supported by studies that employed pure cultures. *Bacteroides* isolates from WWTPs exhibited a high percentage of resistance to tetracyclines (80%) and tested positive for *tet*(Q) and *tet*(M) using PCR (Niestepski et al., 2019; Salyers et al., 2004). ARGs *tet*(M) and *tet*(Q) were carried by integrative and conjugative elements from *Bacteroides* and *Streptococcus*, respectively (Che et al., 2019). In addition, *Bacteroides* species may acquire *erm*(B), *tet*(Q) and *tet*(M) from *Streptococcus* spp, *Clostridium* spp, and *Entrococcus* spp in human intestines (Salyers et al., 2004). Clinical *Bacteroides* isolates from hospital were confirmed to possess *erm*(B) (Johnsen et al., 2017). Similarly, some *Streptococcus* strains isolated from WWTPs and human fecal specimens were resistant to ampicillin, tetracycline, kanamycin, penicillin and vancomycin (Limayem et al., 2019), and the genome of *Streptococcus* strains isolated from a throat swab of a child contained *erm*(B) and *tet*(M) (Huang et al., 2020). Moreover, *Eubaeterium* isolates from patients with periodontal disease harbored the *tet*(M) gene (Olsvik et al., 1995).

3.4. Associations of ARGs with the nitrifiers

RF's were developed for a group of nitrifiers as explanatory variables and individual ARGs as responses. These nitrifiers includes *Nitrosococcus, Nitrosomonas, Nitrosospira, Nitrobacter, Nitrococcus, Nitrospina,*

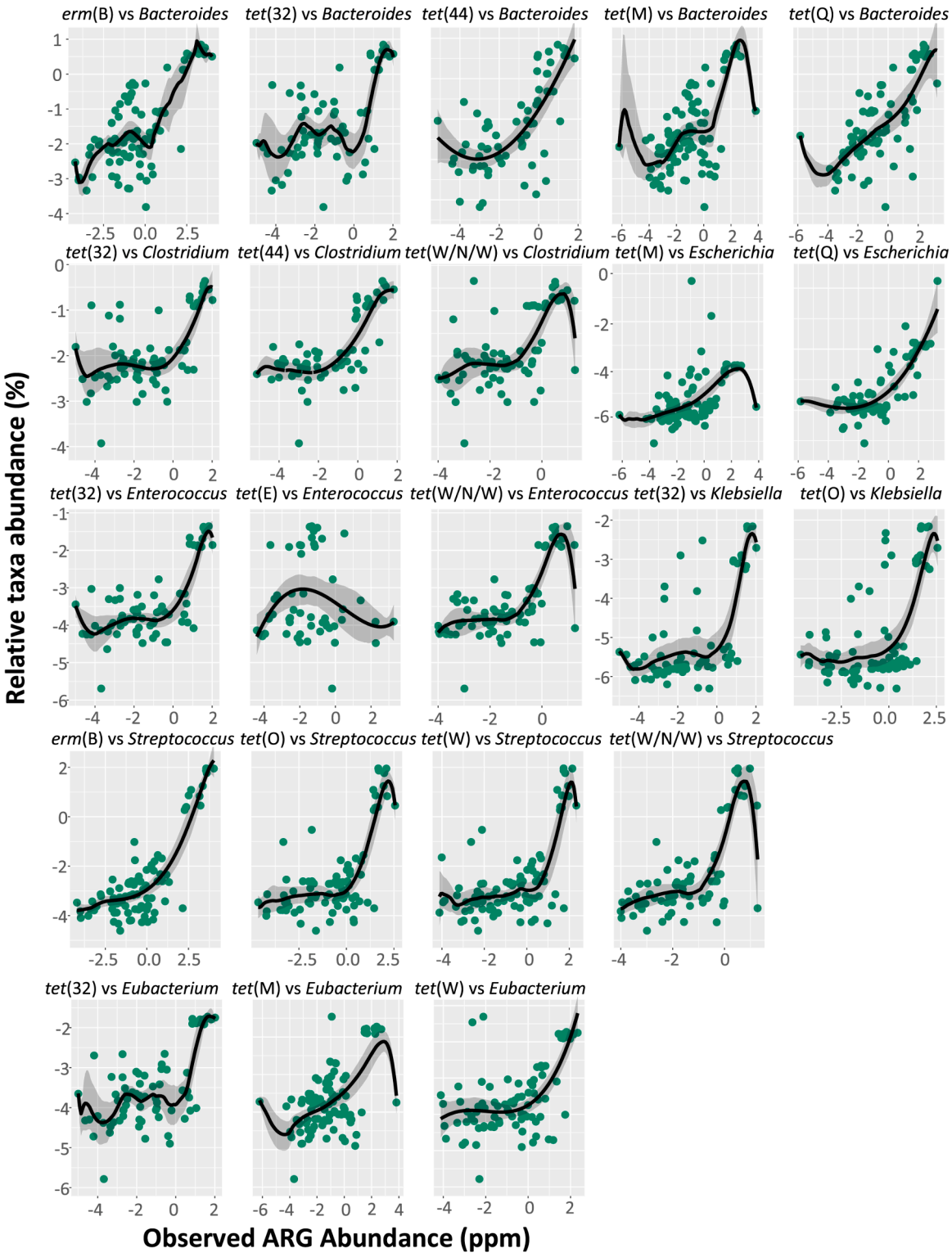

Fig. 5. Abundance plots for ARGs and the most important opportunistic pathogens and ARGs abundance. Each panel of this figure has a horizontal axia representing the abundance of ARGs (ppm), the corresponding vertical axis shows the relative abundance of the opportunistic pathogens (%). Data are smooth by the Nadaraya-Waston estimator (black line) with bootstrap-based 95% confidence band (shaded area). Abundance of ARGs and taxa were log-transformed.

and *Nitrospira*, as they have been consistently detected in the metagenomic libraries. Nitrifying bacteria that were not consistently detected in metagenomic libraries were not included in this analysis.

Within the training dataset, the RF's explained over 88% of the variations in ARG abundance, with R^2 ranging 0.880 – 0.949 and RMSE ranging 0.519 – 0.983 (**Table S3**). When validated using the testing dataset, the RF's exhibited a wider range of R^2 values from 0.0313 to 0.718 (**Table S3**). For example, the RF's exhibited strong associations ($R^2 > 0.600$) between predicted and observed abundance for *tet*(32), *tet*(W), and *tet*(W/N/W), as well as moderate associations (R^2 ranging 0.400 – 0.600) for *erm*(B), *tet*(44), *tet*(M), *tet*(O) and *tet*(Q).

As shown in **Figure S9**, *Nitrosomonas* and *Nitrospira* had importance

scores higher than 90 with more ARGs than did the other nitrifiers. That is, *Nitrosomonas* was important in predicting the abundance of *sul*1, *tet* (32), *tet*(36), *tet*(A), *tet*(O) and *tet*(X), while *Nitrospira* was important in predicting *erm*(B), *tet*(C), *tet*(W), *tet*(W/N/W) *tet*(X3) and *tet*(X4). For all genus-ARG pairs with importance score higher than 90, partial dependence plots show that most of the ARGs included in this analysis exhibited positive dependence on the nitrifiers (**Figure S10**). Our results also show nonlinear functional relationships between ARGs and individual nitrifier genera (**Figure S11**). For instance, abundance of $bla_{\text{OXA-368}}$ and *tet*(A) increases with the abundance of *Nitrococcus*.

Studies using network or binning analyses suggest the associations between nitrifying bacteria and ARGs (**Table 1**). *Nitrosomonas* and *Nitrospira* abundance was suggested to indicate the fluctuation of ARGs abundances in AS reactors (Zhao et al., 2019) and partial-nitritation biofilters (Gonzalez-Martinez et al., 2018; Zhao et al., 2019) following antibiotic addition. Because there is a strong Spearman correlation with class 1 integron-integrase *intI1, Nitrosomonas* spp. may be involved in the horizontal gene transfer of ARGs (Wu et al., 2020). *Nitrosomonas* and *Nitrospira* could survive antibiotic treatment in the reactors and therefore were speculated to be ARG hosts or antibiotic degraders (Zhao et al.,

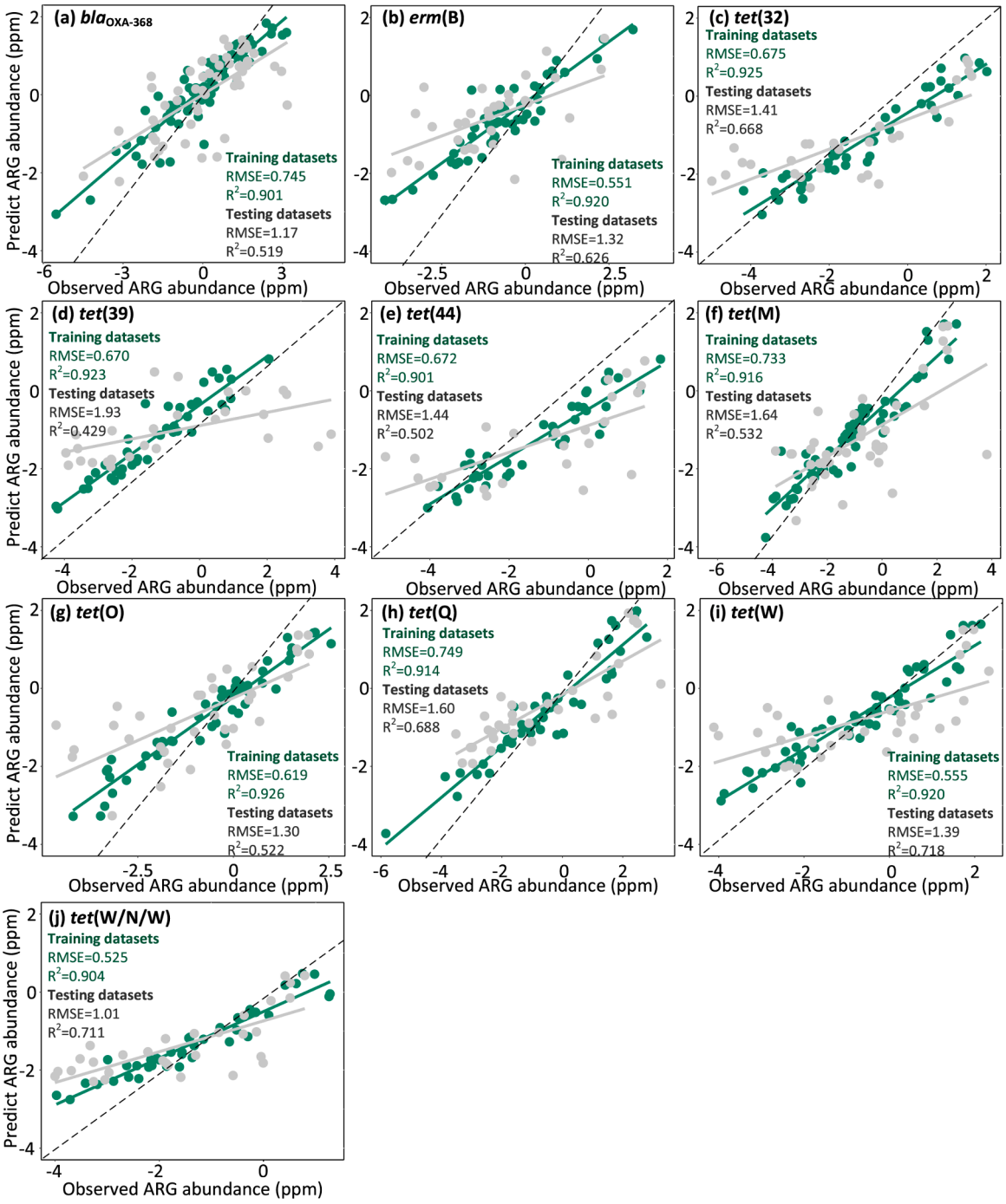

Fig. 6. Observed ARG abundance was plotted against the ARG abundance predicted using the Random Forests generated with nitrifiers as explanatory variables. Only the ten ARGs with R^2 higher than 0.400 in the testing datasets are shown (a - k). The values shown in the plots are log-transformed abundance. Dashed lines indicate the theoretical lines for perfect predictions. Data (dots) and models (line) were separately plotted for the training datasets (green) and the testing datasets (gray). RMSE and adjusted R^2 are reported in the panels.

2019).

3.5. Prediction of ARGs using taxa for a WWTP

Given their superior performance in the testing dataset (Fig. 4 **vs 3 and 6; Fig. 5 vs S6 and S11**), the RF's with (opportunistic) pathogens and indicator bacteria as explanatory variables were used to predict the concentrations of select ARGs in a WWTP in Hong Kong, China (Yin et al., 2019). The shotgun metagenomic data of the WWTP were not included in the 141 datasets used to develop the RF's.

RF's were able to predict the ARGs abundance in the AS of this WWTP in Hong Kong reasonably well (**Figure S12**). The RF's predicted the abundance of *tet*(M), *tet*(Q), and *tet*(W) with R^2 ranging 0.421 - 0.472 and RMSE ranging 0.486 - 0.564. The relatively good agreement based on the R^2 value (i.e., $0.4 \leq R^2 \leq 0.6$) demonstrates the feasibility of the RF's developed in this study to predict ARG concentrations in WWTPs over time. The predicted ARGs abundance of the WWTP ranged from 0.04 - 0.7 ppm for *tet*(M); 0.01 - 0.17 ppm for *tet*(Q); 0.01 - 0.07 ppm for *tet*(W) in present study was consistent with 0.02 - 0.60 ppm observed in Yang et al. (2014).

3.6. Implications and perspectives

RF's were previously used to predict ARG abundance in wastewater using socioeconomic, health, and environmental factors (Hendriksen et al., 2019). In this work, we linked the abundance of select ARGs in AS with the abundance of three bacterial populations and individual genera within the populations. By employing RF's, we have demonstrated that certain bacterial populations exhibit strong associations with select ARGs. In addition, the NW estimator indicates that the abundance of select ARGs increases with the abundance of certain taxa. These functional relationships may be used to develop hypotheses about certain genera being the potential bacterial hosts of ARGs and to estimate ARGs abundance based on microbiome composition in WWTP AS.

One major challenge in studying the environmental resistome is to identify the bacterial hosts of ARGs. In addition to identifying the associations between explanatory variables and responses, RF's can rank the relative importance of individual variables in predicting responses (Cai et al., 2019; Chang et al., 2017; Sun et al., 2018; Yeo et al., 2020). This capability of RF's has been used in applications, including ranking variables (e.g. nanoparticle loading, membrane pore size and relative water contact angle) in regulating water permeability of reverse osmosis membranes (Yeo et al., 2020), identifying associations between the antibiotic resistance in wastewater and socioeconomic variables (Hendriksen et al., 2019), linking feed substrate to the microbiome in microbial fuel cells (Cai et al., 2019), and connecting environmental parameters to the nitrogen fixation related genes (Sun et al., 2020) and microbial diversity (Sun et al., 2018) in soil. In this work, we found that ARGs abundance were predicted with higher accuracy using RF's with the group of (opportunistic) pathogen and indicator bacteria as explanatory variables than those with the groups of abundant genera and nitrifiers. We also tested individual genera from three groups of bacteria, and observed positive dependence with select ARGs. In particular, the functional relationship between ARGs and (opportunistic) pathogens and indicators warrants further investigation (Fig. 5). Any hypothesis about hosts derived from the RF's will still need to be validated using culture-based methods.

As date from next generation sequencing becomes more available, the amount of microbial taxa information is likely to expand quickly. RF's like the ones developed in this study can be used to estimate the abundance of certain ARGs based on microbial community composition. Hermans et al. (2020) emphasized the association between bacterial taxa and soil physico-chemical variables using RF's with R^2 of 0.35 - 0.73 (validated on the testing dataset). Wu et al. (2019) demonstrated the performance of RF's to correlate taxa composition and temperature of wastewater treatment plant, with R^2 of 0.47 (validated on the testing dataset). RF's developed in this study explained over 40% of the variation in the abundance of 8, 11, 10 ARGs in testing datasets for abundant genera, (opportunistic) pathogens and indicators, and nitrifiers, respectively.

Several factors can affect the performance of RF's, such as outliers in datasets, size and number of trees, and folds and times of cross validation. RF's can yield bias in regression problems when extreme observations are estimated using the averages of response values. Large values may be underestimated and small values may be overestimated (Zhang and Lu, 2012). More work is needed to corroborate the accuracy of the RF's and further correct any biases. RF's can be improved by supplementing metagenomic data with other explanatory variables, such as wastewater characteristics (e.g., pH, temperature, wastewater types, and nutrient concentrations) and operational parameters (e.g., hydraulic retention time, sludge retention time, dissolved oxygen, and organic loading rate). Any associations identified between ARG abundance and individual operational parameters can be used to guide the optimization of operation to minimize ARG spreading.

4. Conclusions

In this work, RF's were used to estimate the relationships between the abundance of select ARGs and three groups of bacteria: abundant genera, (opportunistic) pathogens and indicators, and nitrifiers. For RF's with abundant genera as variables, *Pesudomonas* and *Thauera* showed strong associations with multiple ARGs ($bla_{\text{OXA-368}}$, *sul*1, *tet*(X) etc.). For RF's with (opportunistic) pathogens and indicators as variables, *Bacteroides, Clostridium*, and *Streptococcus* exhibited strong associations with *tet* and *erm* genes. RF's with nitrifiers as variables suggest that nitrifiers associate with ARGs abundance, particularly *Nitrosomonas* and *Nitrospira*. Among the three groups of explanatory variables, the group of (opportunistic) pathogens and indicators exhibited more positive functional relationships between individual genera and ARGs than did the other two groups, suggesting members of taxa within this group as potential hosts of these ARGs. Finally, RF's developed based on the (opportunistic) pathogens and indicators could predict ARGs temporal profiles for a full-scale WWTP successfully.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

This study was supported by the National Science Foundation (CBET-1351676 and CBET-1805990). The authors also thank Bing Wang and Ted Naylor for their help with the systematic review. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

Author contributions

The original concept was conceived by XL. The systematics review, bioinformatic analysis, and statistical analyses were completed by YS. The statistical analyses were overseen by BC and JC. The manuscript was written by YS and XL and revised by BC and JC.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2021.117384.

References

Alcock, B.P., Raphenya, A.R., Lau, T.T.Y., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.V., Cheng, A.A., Liu, S., Min, S.Y., Miroshnichenko, A.,

Tran, H.K., Werfalli, R.E., Nasir, J.A., Oloni, M., Speicher, D.J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A.N., Bordeleau, E., Pawlowski, A.C., Zubyk, H.L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G.L., Beiko, R.G., Brinkman, F.S.L., Hsiao, W.W.L., Domselaar, G.V., McArthur, A.G., 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. 48 (D1), D517–D525.

Andrews, S., 2010. FastQC: a Quality Control Tool For High Throughput Sequence data, Babraham Bioinformatics. Babraham Institute, Cambridge, United Kingdom.

Baral, D., Dvorak, B.I., Admiraal, D., Jia, S., Zhang, C., Li, X., 2018. Tracking the sources of antibiotic resistance genes in an urban stream during wet weather using shotgun metagenomic analyses. Environ. Sci. Technol. 52 (16), 9033–9044.

Bouki, C., Venieri, D., Diamadopoulos, E., 2013. Detection and fate of antibiotic resistant bacteria in wastewater treatment plants: a review. Ecotoxicol Environ. Saf. 91, 1–9.

Cai, L., Zhang, T., 2013. Detecting human bacterial pathogens in wastewater treatment plants by a high-throughput shotgun sequencing technique. Environ. Sci. Technol. 47 (10), 5433–5441.

Cai, W., Lesnik, K.L., Wade, M.J., Heidrich, E.S., Wang, Y., Liu, H., 2019. Incorporating microbial community data with machine learning techniques to predict feed substrates in microbial fuel cells. Biosen. Bioelectron. 133, 64–71.

Chang, H.-X., Haudenshield, J.S., Bowen, C.R., Hartman, G.L., 2017. Metagenome-wide association study and machine learning prediction of bulk soil microbiome and crop productivity. Front. Microbiol. 8, 519.

Che, Y., Xia, Y., Liu, L., Li, A.D., Yang, Y., Zhang, T., 2019. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. Microbiome 7 (1), 44.

Chipman, H.A., George, E.I., McCulloch, R.E., 1998. Bayesian CART model search. J. Am. Stat. Assoc. 93 (443), 935–948.

Christgen, B., Yang, Y., Ahammad, S.Z., Li, B., Rodriquez, D.C., Zhang, T., Graham, D.W., 2015. Metagenomics shows that low-energy anaerobic-aerobic treatment reactors reduce antibiotic resistance gene levels from domestic wastewater. Environ. Sci. Technol. 49 (4), 2577–2584.

Du, B., Yang, Q., Wang, R., Wang, R., Wang, Q., Xin, Y., 2019. Evolution of Antibiotic Resistance and the Relationship between the Antibiotic Resistance Genes and Microbial Compositions under Long-Term Exposure to Tetracycline and Sulfamethoxazole. Int. J. Environ. Res. Public Health 16 (23), 4681.

Forsberg, K.J., Patel, S., Gibson, M.K., Lauber, C.L., Knight, R., Fierer, N., Dantas, G., 2014. Bacterial phylogeny structures soil resistomes across habitats. Nature 509 (7502), 612–616.

Gonzalez-Martinez, A., Margareto, A., Rodriguez-Sanchez, A., Pesciaroli, C., Diaz-Cruz, S., Barcelo, D., Vahala, R., 2018. Linking the effect of antibiotics on partial-nitritation biofilters: performance, microbial communities and microbial activities. Front. Microbiol. 9, 354.

Guo, J., Li, J., Chen, H., Bond, P.L., Yuan, Z., 2017. Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. Water. Res. 123, 468–478.

Hendriksen, R.S., Munk, P., Njage, P., van Bunnik, B., McNally, L., Lukjancenko, O., Roder, T., Nieuwenhuijse, D., Pedersen, S.K., Kjeldgaard, J., Kaas, R.S., Clausen, P., Vogt, J.K., Leekitcharoenphon, P., van de Schans, M.G.M., Zuidema, T., de Roda Husman, A.M., Rasmussen, S., Petersen, B., Global Sewage Surveillance project, c., Amid, C., Cochrane, G., Sicheritz-Ponten, T., Schmitt, H., Alvarez, J.R.M., Aidara-Kane, A., Pamp, S.J., Lund, O., Hald, T., Woolhouse, M., Koopmans, M.P., Vigre, H., Petersen, T.N., Aarestrup, F.M., 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat. Commun. 10 (1), 1124.

Hermans, S.M., Buckley, H.L., Case, B.S., Curran-Cournane, F., Taylor, M., Lear, G., 2020. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome 8 (1), 79.

Huang, Y., Wen, Y., Jia, Q., Wang, L., Cheng, Q., Liu, W., Huang, T., Xie, L., 2020. Genome analysis of a multidrug-resistant *Streptococcus sanguis* isolated from a throat swab of a child with scarlet fever. J. Glob. Antimicrob. Resist. 20, 1–3.

Jia, S., Zhang, X.X., Miao, Y., Zhao, Y., Ye, L., Li, B., Zhang, T., 2017. Fate of antibiotic resistance genes and their associations with bacterial community in livestock breeding wastewater and its receiving river water. Water Res 124, 259–268.

Johnsen, B.O., Handal, N., Meisal, R., Bjørnholt, J.V., Gaustad, P., Leegaard, T.M., 2017. *erm* gene distribution among Norwegian *Bacteroides* isolates and evaluation of phenotypic tests to detect inducible clindamycin resistance in *Bacteroides* species. Anaerobe 47, 226–232.

Ju, F., Li, B., Ma, L., Wang, Y., Huang, D., Zhang, T., 2016. Antibiotic resistance genes and human bacterial pathogens: co-occurrence, removal, and enrichment in municipal sewage sludge digesters. Water Res 91, 1–10.

Ju, F., Zhang, T., 2015. Experimental design and bioinformatics analysis for the application of metagenomics in environmental sciences and biotechnology. Environ. Sci. Technol. 49 (21), 12628–12640.

Juretschko, S., Timmermann, G., Schmid, M., Schleifer, K.-H., Pommerening-Röser, A., Koops, H.-P., Wagner, M., 1998. Combined molecular and conventional analyses of nitrifying bacterium diversity in activated sludge: *nitrosococcus mobilis* and *Nitrospira*-like bacteria as dominant populations. Appl. Environ. Microbiol. 64 (8), 3042–3051.

Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., Weijdegård, B., Söderström, H., Larsson, D.J., 2011. Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. PLoS ONE 6 (2), e17038.

Krueger, F. (2012) Trim Galore: a Wrapper Tool Around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ files, With Some Extra Functionality For MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) Libraries. URL http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/. (Date of access: 28/04/2016).

Kuhn, M., 2008. Building Predictive Models in R Using the Caret Package. J. Stat. Softw. 28 (5), 1–26.

Lee, K., Kim, D.-W., Lee, D.-H., Kim, Y.-S., Bu, J.-H., Cha, J.-H., Thawng, C.N., Hwang, E.-M., Seong, H.J., Sul, W.J., 2020. Mobile resistome of human gut and pathogen drives anthropogenic bloom of antibiotic resistance. Microbiome 8 (1), 1–14.

Li, A.D., Li, L.G., Zhang, T., 2015a. Exploring antibiotic resistance genes and metal resistance genes in plasmid metagenomes from wastewater treatment plants. Front. Microbiol. 6, 1025.

Li, B., Ju, F., Cai, L., Zhang, T., 2015b. Profile and Fate of Bacterial Pathogens in Sewage Treatment Plants Revealed by High-Throughput Metagenomic Approach. Environ. Sci. Technol. 49 (17), 10492–10502.

Li, B., Yang, Y., Ma, L., Ju, F., Guo, F., Tiedje, J.M., Zhang, T., 2015c. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. ISME J 9 (11), 2490–2502.

Limayem, A., Wasson, S., Mehta, M., Pokhrel, A.R., Patil, S., Nguyen, M., Chen, J., Nayak, B., 2019. High-Throughput Detection of Bacterial Community and Its Drug-Resistance Profiling From Local Reclaimed Wastewater Plants. Front. Cell. Infect. Microbiol. 9, 303.

Liu, Z., Klumper, U., Liu, Y., Yang, Y., Wei, Q., Lin, J.G., Gu, J.D., Li, M., 2019. Metagenomic and metatranscriptomic analyses reveal activity and hosts of antibiotic resistance genes in activated sludge. Environ. Int. 129, 208–220.

Ma, L., Xia, Y., Li, B., Yang, Y., Li, L.-G., Tiedje, J.M., Zhang, T., 2016. Metagenomic assembly reveals hosts of antibiotic resistance genes and the shared resistome in pig, chicken, and human feces. Enviro. Sci. Technol. 50 (1), 420–427.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 17 (1), 10–12.

Martinez, J.L., 2008. Antibiotics and antibiotic resistance genes in natural environments. Science 321 (5887), 365–367.

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat. Commun. 7, 11257.

Niestepski, S., Harnisz, M., Korzeniewska, E., Aguilera-Arreola, M.G., Contreras-Rodriguez, A., Filipkowska, Z., Osinska, A., 2019. The emergence of antimicrobial resistance in environmental strains of the *Bacteroides fragilis* group. Environ. Int. 124, 408–419.

Olsvik, B., Olsen, I., Tenover, F., 1995. Detection of tet (M) and tet (Q) using the polymerase chain reaction in bacteria isolated from patients with periodontal disease. Oral Microbiol. Immunol. 10 (2), 87–92.

Rice, E.W., Wang, P., Smith, A.L., Stadler, L.B., 2020. Determining Hosts of Antibiotic Resistance Genes: a Review of Methodological Advances. Environ. Sci. Technol. Lett. 7 (5), 282–291.

Salyers, A.A., Gupta, A., Wang, Y., 2004. Human intestinal bacteria as reservoirs for antibiotic resistance genes. Trends Microbiol 12 (9), 412–416.

Scherson, Y.D., Wells, G.F., Woo, S.-G., Lee, J., Park, J., Cantwell, B.J., Criddle, C.S., 2013. Nitrogen removal with energy recovery through N 2 O decomposition. Energy Environ. Sci. 6 (1), 241–248.

Sui, Q., Jiang, C., Zhang, J., Yu, D., Chen, M., Wang, Y., Wei, Y., 2018. Does the biological treatment or membrane separation reduce the antibiotic resistance genes from swine wastewater through a sequencing-batch membrane bioreactor treatment process. Environ. Int. 118, 274–281.

Sun, W., Xiao, E., Haggblom, M., Krumins, V., Dong, Y., Sun, X., Li, F., Wang, Q., Li, B., Yan, B., 2018. Bacterial Survival Strategies in an Alkaline Tailing Site and the Physiological Mechanisms of Dominant Phylotypes As Revealed by Metagenomic Analyses. Environ. Sci. Technol. 52 (22), 13370–13380.

Sun, X., Kong, T., Haggblom, M.M., Kolton, M., Li, F., Dong, Y., Huang, Y., Li, B., Sun, W., 2020. Chemolithoautotropic Diazotrophy Dominates the Nitrogen Fixation Process in Mine Tailings. Environ. Sci. Technol. 54 (10), 6082–6093.

Tang, J., Bu, Y., Zhang, X.X., Huang, K., He, X., Ye, L., Shan, Z., Ren, H., 2016. Metagenomic analysis of bacterial community composition and antibiotic resistance genes in a wastewater treatment plant and its receiving surface water. Ecotoxicol. Environ. Saf. 132, 260–269.

Wang, Q., Li, X., Yang, Q., Chen, Y., Du, B., 2019. Evolution of microbial community and drug resistance during enrichment of tetracycline-degrading bacteria. Ecotoxicol. Environ. Saf. 171, 746–752.

Wang, Z., Yuan, S., Deng, Z., Wang, Y., Deng, S., Song, Y., Sun, C., Bu, N., Wang, X., 2020. Evaluating responses of nitrification and denitrification to the co-selective pressure of divalent zinc and tetracycline based on resistance genes changes. Bioresour. Technol. 314, 123769.

Wu, D., Dolfing, J., Xie, B., 2018. Bacterial perspectives on the dissemination of antibiotic resistance genes in domestic wastewater bio-treatment systems: beneficiary to victim. Appl. Microbiol. Biotechnol. 102 (2), 597–604.

Wu, J.W., Wu, C.R., Zhou, C.S., Dong, L.L., Liu, B.F., Xing, D.F., Yang, S.S., Fan, J.N., Feng, L.P., Cao, G.L., You, S.J., 2020. Fate and removal of antibiotic resistance genes in heavy metals and dye co-contaminated wastewater treatment system amended with beta-cyclodextrin functionalized biochar. Sci. Total. Environ. 723, 137991.

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., Zhang, Q., Brown, M.R., Li, Z., Van Nostrand, J.D., Ling, F., Xiao, N., Zhang, Y., Vierheilig, J., Wells, G.F., Yang, Y., Deng, Y., Tu, Q., Wang, A., Global Water Microbiome, C., Zhang, T., He, Z., Keller, J., Nielsen, P.H., Alvarez, P.J.J., Criddle, C.S., Wagner, M., Tiedje, J.M., He, Q., Curtis, T.P., Stahl, D.A., Alvarez-Cohen, L., Rittmann, B.E., Wen, X., Zhou, J., 2019. Global diversity and biogeography of bacterial communities in wastewater treatment plants. Nat. Microbiol. 4 (7), 1183–1195.

Xia, J., Sun, H., Zhang, X.-x., Zhang, T., Ren, H., Ye, L., 2019. Aromatic compounds lead to increased abundance of antibiotic resistance genes in wastewater treatment bioreactors. Water Res 166, 115073.

Xue, G., Jiang, M., Chen, H., Sun, M., Liu, Y., Li, X., Gao, P., 2019. Critical review of ARGs reduction behavior in various sludge and sewage treatment processes in wastewater treatment plants. Critical Rev. Environ. Sci. Technol. 49 (18), 1623–1674.

Yang, Y., Li, B., Ju, F., Zhang, T., 2013. Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. Environ. Sci. Technol. 47 (18), 10197–10205.

Yang, Y., Li, B., Zou, S., Fang, H.H., Zhang, T., 2014. Fate of antibiotic resistance genes in sewage treatment plant revealed by metagenomic approach. Water Res 62, 97–106.

Yeo, C.S.H., Xie, Q., Wang, X., Zhang, S., 2020. Understanding and optimization of thin film nanocomposite membranes for reverse osmosis with machine learning. J. Membr. Sci., 118135

Yin, X., Deng, Y., Ma, L., Wang, Y., Chan, L.Y.L., Zhang, T., 2019. Exploration of the antibiotic resistome in a wastewater treatment plant by a nine-year longitudinal metagenomic study. Environ. Int. 133 (Pt B), 105270.

Zhang, G., Lu, Y., 2012. Bias-corrected random forests in regression. J. Appl. Stat. 39 (1), 151–160.

Zhang, J., Chen, M., Sui, Q., Tong, J., Jiang, C., Lu, X., Zhang, Y., Wei, Y., 2016. Impacts of addition of natural zeolite or a nitrification inhibitor on antibiotic resistance genes during sludge composting. Water Res 91, 339–349.

Zhang, J., Yang, M., Zhong, H., Liu, M., Sui, Q., Zheng, L., Tong, J., Wei, Y., 2018. Deciphering the factors influencing the discrepant fate of antibiotic resistance genes in sludge and water phases during municipal wastewater treatment. Bioresour. Technol. 265, 310–319.

Zhao, R., Feng, J., Liu, J., Fu, W., Li, X., Li, B., 2019. Deciphering of microbial community and antibiotic resistance genes in activated sludge reactors under high selective pressure of different antibiotics. Water Res 151, 388–402.

Zhou, S., Zhu, Y., Yan, Y., Wang, W., Wang, Y., 2019. Deciphering extracellular antibiotic resistance genes (eARGs) in activated sludge by metagenome. Water Res 161, 610–620.