METAGENOMIC DATA ANALYSIS USING EXTREMELY RANDOMIZED TREE ALGORITHM

By

Suraj Gupta

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Civil Engineering

Peter J. Vikesland, Chair

Marc A. Edwards

Amy J. Pruden-Bagchi

May 2, 2018

Blacksburg, Virginia

Keywords: Antibiotic resistance genes, ARGs, aquatic environments, ensemble learning, extremely randomized trees, wastewater

METAGENOMIC DATA ANALYSIS USING EXTREMELY RANDOMIZED TREE ALGORITHM

Suraj Gupta

ABSTRACT

Many antibiotic resistance genes (ARGs) conferring resistance to a broad range of antibiotics have often been detected in aquatic environments such as untreated and treated wastewater, river and surface water. ARG proliferation in the aquatic environment could depend upon various factors such as geospatial variations, the type of aquatic body, and the type of wastewater (untreated or treated) discharged into these aquatic environments. Likewise, the strong interconnectivity of aquatic systems may accelerate the spread of ARGs through them. Hence a comparative and a holistic study of different aquatic environments is required to appropriately comprehend the problem of antibiotic resistance. Many studies approach this issue using molecular techniques such as metagenomic sequencing and metagenomic data analysis. Such analyses compare the broad spectrum of ARGs in water and wastewater samples, but these studies use comparisons which are limited to similarity/dissimilarity analyses. However, in such analyses, the discriminatory ARGs (associated ARGs driving such similarity/ dissimilarity measures) may not be identified. Consequentially, the reason which drives the dissimilarities among the samples would not be identified and the reason for antibiotic resistance proliferation may not be clearly understood. In this study, an effective methodology, using Extremely Randomized Trees (ET) Algorithm, was formulated and demonstrated to capture such ARG variations and identify discriminatory ARGs among environmentally derived metagenomes. In this study, data were grouped by: geographic location (to understand the spread of ARGs globally), untreated vs. treated wastewater (to see the effectiveness of WWTPs in removing ARGs), and different aquatic habitats (to understand the impact and spread within aquatic habitats). It was observed that there were certain ARGs which were specific to wastewater samples from certain locations suggesting that site-specific factors can have a certain effect in shaping ARG profiles. Comparing untreated and treated wastewater samples from different WWTPs revealed that biological treatments have a definite impact on shaping the ARG profile. While there were several ARGs which got removed after the treatment, there were some ARGs which showed an increase in relative abundance irrespective of location and treatment plant specific variables. On comparing different aquatic environments, the algorithm identified ARGs which were specific to certain environments. The algorithm captured certain

ARGs which were specific to hospital discharges when compared with other aquatic environments. It was determined that the proposed method was efficient in identifying the discriminatory ARGs which could classify the samples according to their groups. Further, it was also effective in capturing low-level variations which generally get over-shadowed in the analysis due to highly abundant genes. The results of this study suggest that the proposed method is an effective method for comprehensive analyses and can provide valuable information to better understand antibiotic resistance.

METAGENOMIC DATA ANALYSIS USING EXTREMELY RANDOMIZED TREE ALGORITHM

Suraj Gupta

GENERAL AUDIENCE ABSTRACT

Antibiotic resistance is a natural and primordial process that predates the use of antibiotics in humans for disease treatment and occurs when a bacterium evolves to render the drugs, chemicals, or other agents meant to cure or prevent infections ineffective. Antibiotic resistance genes (ARGs) conferring resistance to a wide range of antibiotics have been widely found in rivers, surface waters, and hospital and farm wastewater discharges. Even treated wastewater from treatment plants is a concern as ARGs have frequently been detected in effluent discharges which poses questions on the effectiveness of treatment plants in removing ARGs. Since, these systems are interconnected there's a possibility of dissemination and proliferation of ARGs which may pose serious threat to human health. Hence, it is desirable to perform comparative studies among these aquatic habitats. In previous studies, researchers compared different habitats which tells how similar and dissimilar the environments are in terms of ARGs present in these samples. While these analyses are important, it doesn't tell which ARGs are unique or which ARGs are responsible to create those similarities or dissimilarities. This information is crucial in order to understand the water environments in terms of occurrence and presence of ARGs, the risk posed by them, and in identifying factors responsible for resistance gene proliferation. In this research, a methodology was developed which could capture such ARG variations in the environmental samples, using data analysis algorithms. Further the developed methodology was demonstrated using environmental samples such as wastewater samples from different geographical locations (to understand the spread of ARGs globally), untreated vs treated wastewater (to understand the effectiveness of treatment plants in removing ARGs), and different aquatic habitats (to understand the impact and spread of ARGs within these habitats). It was determined that the proposed method was efficient in differentiating samples and identifying discriminatory ARGs. The comparison between environmental samples showed that the samples from different locations have specific ARGs which were unique to wastewater samples from certain locations suggesting that site-specific factors can have certain effect in shaping the ARG profiles. Comparing untreated and treated samples revealed that treatment plants were able to remove certain ARGs but it was also observed

that some ARGs proliferated after the treatment irrespective of location and treatment plant specific variables. Analyzing different environments, the approach was able to identify certain ARGs which were specific to certain environments. The results of this study suggest that the proposed method is an effective method for comprehensive analyses and can provide valuable information to better understand antibiotic resistance. In essence, it is a valuable addition for improved surveillance of antibiotic resistance pollution and for the framing of best management practices.

Acknowledgements

I would like to extend my sincere gratitude to my advisor Dr Peter Vikesland, for his immense support, guidance, and boundless patience. I'd especially like to thank him for the opportunity he provided for me to realize my goals and interests, his belief on me, and all the advices he provided both at personal and professional levels. I would like to thank Dr Amy Pruden, my committee member, for her invaluable time, support and inputs during the duration of this project. I also thank Dr Marc Edwards for his continuous support during this project. I'd also like to thank Dr Liqing Zhang for her invaluable inputs in this study. I consider myself blessed to have worked with all of them.

I'd like to convey my heartfelt appreciation to Dr. Virginia Riquelme for sharing her kindness and vast knowledge, and for offering help every time I asked for it. Thanks for teaching me so much about microbes! Next, I'd like to thank Gustavo Arango without whom this project would have been impossible to complete. Thanks for helping me out with all the data science doubts. I feel so fortunate to have learned from you both.

I'd like to thank Dr Weinan Leng for his useful support in the lab. I'd also like to thank my friends: Himanki, Pranav, Jayesh, Akshay, Morgan, Mariah, Nevetha, Jacob, Gregory, Samantha, Salil, Romit, Rahul, Pururaj, Akash, Kris, members of Vikesland and Pruden group, for all the fun, support and love.

Last but not the least, I'd like to thank my Mom, Dad and Sister for their immense love. Thank you for always being there for me and supporting my endeavors.

Table of Contents

Table of Contents

List of Figures

List of Tables

Attributions

Gustavo Arango-Argoty, Ph.D Student, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Dr Liqing Zhang, Ph.D., Associate Professor, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Dr Amy Pruden, Ph.D., Professor, Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Dr Peter J Vikesland, Ph.D, Professor, Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Gustavo Arango-Argoty, Drs. Zhang, Pruden and Vikesland are duly credited as the co-author in Chapter-2 of this document. Dr. Vikesland and Dr. Pruden served as the PI and co-PI for this project.

All the authors contributed equally to the ideas and the direction of this research. Gustavo Arango-Argoty contributed in development of the methodology and review of the manuscript. Drs. Zhang, Pruden and Vikesland provided their inputs and contributed to the review of the manuscript.

Chapter 1

Introduction

The Centers for Disease Control and Prevention, and World Health Organization, identify dissemination of antibiotic resistance genes (ARGs) as a serious threat for public health[1, 2]. The increased appearance of ARGs in pathogens is a major concern and could challenge the effectiveness of antibiotics for the treatment of infectious diseases, in both humans and animals. The emergence of antibiotic resistance is a highly convoluted process which is yet to be fully understood with regard to the significance of bacterial communities, antibiotics, and other anthropogenic stressors. The growing concern of antibiotic resistance genes' dissemination and proliferation has fueled a great deal of research in this area. Aquatic systems can act as a reservoir, recipient, and source of ARGs[3]. There is thus an increased focus towards urban water systems, natural water bodies, soil, and ocean metagenomes in order to characterize the occurrence, diversity, and habitat of ARGs [4-7]. Wastewater treatment plants which form a critical node in discharging water into the aquatic environment have been thought to be potential hotspots for ARG dissemination[8]. Thus it becomes imperative to study the efficacy of these treatment plants in mitigating ARGs in their final discharge. Research also suggests that ARG profiles depend on anthropogenic stressors which in turn depend on the geographical location of concern. Therefore, a study across locations would present an understanding about the reasons for ARG dissemination in a particular location.

With the advancement in the biomolecular field, metagenomic sequencing has emerged as a powerful tool which allows access to the collective genome of an environmental sample[9, 10]. It is an approach which is used to sequence the microorganisms in an environmental sample without a prior need of isolation and cultivation. While sequencing has provided the in-depth detailed investigation of environmental samples in the context of antibiotic resistance, it has been difficult to analyze the data. The highly correlated nature of the variables poses significant challenges in statistical analysis of this data.

Many researchers have recently begun to use metagenomic sequencing to obtain the broad spectrum of ARGs in environmental samples [11-13]. The previous metagenomic data analysis comparing ARG profiles in different urban water systems and natural water bodies, influent and

effluent samples of WWTPs, and at different sampling locations have been limited to feature projection methods[14]. While these analyses are beneficial in obtaining a measure of similarity or dissimilarity between the samples, it doesn't identify the discriminatory ARGs associated with the dissimilarities in the ARG profiles. Identification of discriminatory ARGs is essential to improve surveillance of antibiotic resistance pollution, and this would help to identify the source of contamination and frame best management practices to mitigate the dissemination of ARGs.

The aim of this study was to develop a methodology which could map the variations in ARG profiles and identify associated ARGs using the environmentally-derived metagenomes of water/wastewater samples categorized according to the user defined groups (for example: geographic location, untreated vs. treated wastewater, different aquatic environments). Due to the global concern of the dissemination of ARGs and assessing the role of WWTPs in mitigating this spread. This study seeks to extend this investigation to the global scale to better understand the magnitude and variation of this issue. Hence, influent and effluent samples of WWTPs were collected from six different countries. Influent samples were compared to understand the variation in ARG profiles globally and identify associated discriminatory ARGs. Influent and effluent samples were used to understand the effectiveness of WWTPs in mitigating the dissemination of ARGs. Further, various water and wastewater metagenomes such as farm source water, river source water, hospital effluents were retrieved from public databases to draw a comparison among different aquatic systems.

The methodology was developed using Extremely Randomized Trees (ET) algorithm[15]. This was used to identify the discriminatory ARGs using the concept of ranking the variables according to the Gini importance measure. Further, hierarchical clustering and NMDS was used to validate the clustering of samples into their assigned groups. The variations in the relative abundance of identified discriminatory ARGs across the groups was visualized using heatmaps. The results of this thesis suggest that ET is an effective method to study and analyze the metagenomic data.

This thesis contains three chapters. The first chapter is a brief introduction of the thesis, followed by the second chapter which is a manuscript of the study describing the methodology development and its application in analyzing environmentally derived metagenomic data. The third chapter is the conclusion of the study.

References

1. Leidl, P., *Overcoming Antimicrobial Resistance: World Health Organization Report on Infectious Diseases 2000*. 2000: World health organization (WHO).
2. Lushniak, B.D., *Antibiotic resistance: a public health crisis.* Public Health Reports, 2014. **129**(4): p. 314-316.
3. D'costa, V.M., et al., *Sampling the antibiotic resistome.* Science, 2006. **311**(5759): p. 374-377.
4. Fick, J., et al., *Contamination of surface, ground, and drinking water from pharmaceutical production.* Environmental Toxicology and Chemistry, 2009. **28**(12): p. 2522-2527.
5. Kristiansson, E., et al., *Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements.* PloS one, 2011. **6**(2): p. e17038.
6. LaPara, T.M., et al., *Tertiary-treated municipal wastewater is a significant point source of antibiotic resistance genes into Duluth-Superior Harbor.* Environmental science & technology, 2011. **45**(22): p. 9543-9549.
7. Pruden, A., M. Arabi, and H.N. Storteboom, *Correlation between upstream human activities and riverine antibiotic resistance genes.* Environmental science & technology, 2012. **46**(21): p. 11541-11549.
8. Gao, P., M. Munir, and I. Xagoraraki, *Correlation of tetracycline and sulfonamide antibiotics with corresponding resistance genes and resistant bacteria in a conventional municipal wastewater treatment plant.* Science of the Total Environment, 2012. **421**: p. 173-183.
9. Lindgreen, S., K.L. Adair, and P.P. Gardner, *An evaluation of the accuracy and speed of metagenome analysis tools.* Scientific reports, 2016. **6**: p. 19233.
10. Simon, C. and R. Daniel, *Metagenomic analyses: past and future trends.* Applied and environmental microbiology, 2011. **77**(4): p. 1153-1161.
11. Ju, F., et al., *Metagenomic analysis on seasonal microbial variations of activated sludge from a full-scale wastewater treatment plant over 4 years.* Environmental microbiology reports, 2014. **6**(1): p. 80-89.

12. Li, B., et al., *Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes.* The ISME journal, 2015. **9**(11): p. 2490.
13. Rizzo, L., et al., *Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review.* Science of the total environment, 2013. **447**: p. 345-360.
14. Ramette, A., *Multivariate analyses in microbial ecology.* FEMS microbiology ecology, 2007. **62**(2): p. 142-160.
15. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.

Chapter 2

Manuscript

Title: Analysis of Environmentally-Derived Metagenomic Data using Extremely Randomized Tree Algorithm

S. Gupta[a*], G Arango-Argoty[b], Liqing Zhang[b], A. Pruden[a], P. J. Vikesland[a]

a: Department of Civil and Environmental Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

b: Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

*Corresponding author

Abstract

Antibiotic Resistance Genes (ARGs) conferring resistance to a wide range of antibiotics have been widely found in rivers, surface waters, and hospital and farm effluents. Even treated wastewater from wastewater treatment plants (WWTPs) is a concern as high microbial activity in the biological treatment processes and the presence of co-selecting factors may result in horizontal gene transfer and in proliferation of resistance genes and resistant bacteria. Furthermore, these aquatic systems are interconnected and may influence each other in the proliferation and dissemination of resistance genes. Such a complication necessitates comparative studies among these aquatic habitats. In an attempt to understand these issues, various studies compared the broad spectrum of ARGs in water and wastewater samples, but these studies use comparisons which are limited to similarity/dissimilarity analyses. Although these analyses are important, it is crucial to identify the ARGs (hereafter referred as discriminatory ARGs) driving the measured similarities and dissimilarities. This information would provide a better understanding of the water environment in terms of occurrence and presence of ARGs, the risk posed by them, and in identifying factors responsible for resistance gene proliferation. The author of this study formulated and demonstrated a method to capture such ARG variations in environmental samples when categorized into their groups, using the Extremely Randomized Tree Algorithm (ET). In this study, data was grouped by: geographic location (to understand the spread of ARGs globally), untreated vs. treated wastewater (to see the effectiveness of WWTPs in removing ARGs), and different aquatic habitats (to understand the impact and spread within aquatic habitats). It was determined that the proposed method was efficient in differentiating samples and identifying discriminatory ARGs according to their groups. It was observed that there were certain ARGs which were specific to wastewater samples from certain locations suggesting that site-specific factors can have certain effect in shaping the ARG profiles. Comparing untreated and treated wastewater samples from different WWTPs revealed that biological treatments have definite impact on shaping the ARG profile. While there were several ARGs which got removed after the treatment, there were some ARGs which showed increase in relative abundance irrespective of location and treatment plant specific variables. On comparing different aquatic environments, the algorithm identified ARGs which were specific to certain environments. The algorithm also captured certain ARGs which were specific to hospital discharges when compared with other aquatic environments. The algorithm was able to capture low-level variations and models an

effective way to analyze and visualize metagenomics data. In essence, it is a valuable addition for improved surveillance of antibiotic resistance pollution and for the framing of best management practices.

Keywords: Antibiotic resistance genes, ARGs, aquatic environments, ensemble learning, extremely randomized trees, wastewater

Introduction

Antibiotic resistance, as recognized by the World Health Organization (WHO) and other international organizations, poses a serious threat to public health and challenges the effectiveness of antibiotics for the treatment of infectious diseases [1, 2]. This phenomenon is being widely studied across the globe because it has been identified as a major setback in the treatment of many diseases [3]. Antibiotic resistance is a natural and primordial process that predates the use of antibiotics in humans for disease treatment and occurs when a bacterium evolves to render the drugs, chemicals, or other agents meant to cure or prevent infections ineffective. Susceptible bacteria can be damaged or inhibited by antibiotic action, whereas resistant bacteria are unaffected and may be enriched in the presence of therapeutic levels of an antibiotic. Background resistance occurs irrespective of human interference, as bacteria can produce and use antibiotics against other bacteria, prompting a low level of natural selection for resistance to antibiotics [4]. However, there has been a recent increase in the emergence of resistant bacteria and the genes conferring resistance. Selective pressures exerted by antibiotic residues, along with other co-selecting factors such as metals and surfactants, in the environment are thought to sustain and exacerbate antibiotic resistance [5, 6].

A major concern is the occurrence of antibiotic resistance genes (ARGs) in clinical pathogens, which are severely endangering the effective use of antibiotics in human and veterinary medicines [7, 8]. The ARGs found in pathogenic bacteria often originate after the long-term evolution of resistance mechanisms in non-pathogenic bacteria [9]. The environmental microbiome contains a rich diversity of microorganisms and is believed to serve as a reservoir and source of resistance genes [10, 11]. Under favorable conditions these genes can be transferred to pathogenic bacteria via horizontal gene transfer, thus extending resistance to new bacteria [12]. While these conditions have yet to be identified, the selective pressure exerted by antibiotics discharged into the environment with other co-selecting factors is thought to aid in this process. Hence, analyzing the impact of anthropogenic activities on the augmentation and dissemination of ARGs is crucial.

Studies have revealed that aquatic environments serve as a recipient, reservoir, and source of ARGs [13]. Wastewater discharges, with high ARG loadings, can directly affect various aquatic environments such as rivers and surface waters and can aid in augmenting the resistance gene pool [14]. Extensive use of antibiotics in clinical and agricultural settings has established hospital wastewater and farm wastewater effluents as a major source of antibiotic pollution [15-17].

Surface water and river water, which are often treated to use for drinking purposes, have often been characterized for the presence of ARGs [18-22]. The presence and possible transfer of these ARGs to pathogens could serve as a serious threat to public health. Thus, assessing the risk posed by different aquatic environments remains a crucial endeavor. Wastewater treatment plants (WWTPs) serve as a critical node for either mitigation or dissemination of resistance genes. Wastewater from various sources may contain different antibiotics with varying concentration levels depending on the antibiotic consumption/usage pattern, which could lead to different microbial flora, ARGs, and ARG loadings present in the wastewater [23]. Wastewater influents and sludge discharges are often found to be rich in ARGs and other co-selecting factors [24, 25]. Hence, there is increased attention towards the characterization of these influents and effluents in the context of antibiotic resistance. The dissemination of ARGs is a complex process and targeted studies are required to address this issue. Due to the global nature of this concern, many questions remain to be addressed, such as (1) How do ARG distributions vary globally? (2) How effective are WWTPs in removing ARGs present in wastewater? (3) Are there any differences in ARG profiles between different aquatic systems? (4) How can wastewater discharges affect natural water bodies in terms of ARG composition?

Many researchers have used quantitative polymerase chain reaction (qPCR) to investigate the profile and abundance of specific ARGs in environmental sample [26, 27]. However, qPCR results are constrained by many factors, such as reaction chemistry, PCR machines, and matrix effects of the DNA extracts, and provide limited information on the range of ARGs present in the sample [28]. With recent advancements in the biomolecular field, metagenomic sequencing has emerged as a powerful tool that could reveal the broad spectrum of ARGs in samples [29, 30]. There are studies focused on characterizing and comparing the ARG profiles in different urban water systems and natural water bodies [31], influent and effluent samples from WWTPs, and ARGs in different WWTPs [32][33]. The present literature on the analysis of metagenomics data for ARGs has primarily focused on feature projection methods such as principal component analysis (PCA), non-metric multi-dimensional scaling (NMDS), and principal coordinate analysis (PCoA) [34]. The problem with these analyses is that they only provide measures of similarity or dissimilarity between samples rather than identifying the discriminatory ARGs associated with the dissimilarities. Another problem with metagenomics sequencing and data analysis is that massive sequencing projects are expensive and the data produced is difficult to analyze [35]. The large

scale of genomic data presents significant challenges for statistical and bioinformatics data analysis. In particular, the highly correlated nature of the variables in genomic data renders the independent assumptions required by many statistical models invalid. Hence, a key research need is to develop a methodology that could better explore the dissimilarities among samples and identify the variation in ARG profiles. Such a methodology would improve the surveillance of antibiotic resistance pollution, and would help to identify the source of contamination and frame best management practices to mitigate the spread of ARGs. This study seeks to extend this investigation to the global scale to better understand the magnitude and variation of this issue.

Ensemble learning methods have recently been developed to handle high-dimensional problems such as metagenomic data. The Extremely Randomized Tree (ET) algorithm is one of the widely used ensemble learning methods in the field of machine learning [36]. It uses a similar approach to Random Forests (RF) [37] to build an ensemble of trees, but with two major differences: (a) instead of using bagging features, it uses full datasets to grow and learn the trees, and (2) the node split is picked randomly, as compared to RF where best splits are chosen within the random subset sampled. The ET algorithm is efficient in handling correlations and interactions among variables and gives effective data inference. ET algorithms can also be used to rank features by variable importance measures and better differentiate the class based on the feature variables (see Section III). This property of ET algorithms was used in this study to find discriminatory ARGs to classify the samples according to their groups.

The objective of this study was to (1) formulate a methodology to classify samples based on their categorized groups (for example: geographic location, untreated vs. treated wastewater, different aquatic environments) using the ET algorithm, (2) identify discriminatory ARGs to differentiate untreated wastewater samples based on their geographic location which encompasses a broad-range of site-specific variables, (3) identify discriminatory ARGs among treated and untreated wastewater samples, and (4) compare and identify discriminatory ARGs among different aquatic environments such as river water, hospital and farm discharges, and WWTP influent and effluent.

Supervised Ensemble Learning

Extremely Randomized Trees Algorithm

The Extremely Randomized Trees (ET) Algorithm is a tree-based ensemble method that is traditionally used for supervised classification and regression problems. The ensemble method is

a process by which the outcomes from many decision trees are averaged to get a final output. ET is used to deduce useful information from a labeled set of data. The labeled dataset contains 'features' (also called attributes or input variables) and 'classes' (output variable) and then the algorithm is used to build a mapping function between Attributes and Class Variables. The motive is to identify the variables associated with different classes and to predict the class variable based on the given attribute data. Attributes are a form of description that can collectively define a particular object, element, or a sample. Simply, attributes are a set of specific values that together describe an object; for example, weight, height and gender are all attributes that could describe a classroom of ten students. However, if we want to categorize the students into groupings of males and females, gender is the controlling factor that best describes the data into the desired categorizations. As applied in this study, the ET algorithm works to separate the samples based on the user defined group (class variable) using ARG relative abundance (attribute) as the controlling factor.

The basic unit of the ET algorithm (i.e., a decision tree) is typically constructed using the Classification and Regression Trees (CART) methodology [38]. Decision trees are grown by splitting the input dataset into subsets using simple decision rules deduced from the attribute information. The subsets are partitioned recursively until homogeneity or near homogeneity is achieved within the nodes (i.e., class variables are separated using decision rules on attributes). ET uses hundreds to thousands of such trees to build an ensemble. The obtained ensemble of trees is then used to build the association between the variables and the labeled class. The averaged decision based on the ensemble of trees reduces the variance of the model, without increasing the bias; and gives better classification. This technique largely overcomes overfitting problems associated with the single classification tree method. The difference between ET and other tree-based ensemble approaches is that it splits nodes using randomly generated thresholds for each feature. Then, the best threshold is chosen as the splitting rule. The randomness in choosing cut-point thresholds of the attributes reduces the variance. The introduction of more randomness in selecting the cut-point threshold and attributes reduces the variance effectively when combined with ensemble averaging. Another difference is that it uses full dataset to build the trees where other methods adopt a bootstrapping approach to sample the dataset. In the bootstrapping method, a part of sample is used to make the trees which could lead to high bias towards the classification. Using whole dataset helps in further reducing this bias.

The ET algorithm estimates the important variables for the classification. It computes the variable "importance measure" using either the mean decrease in accuracy or the mean decrease in Gini. The variable importance measure indicates which attributes are most efficient in effectively classifying the class variables and contribute the most in building strong decision trees. In the context of this study, the attributes are the **relative abundance metrics of the resistance genes** and the classes are the **labels given by the users**. The objective is to map the resistance genes with the labels (e.g., sampling location, habitats) and identify the ARGs associated with different labels. The obtained knowledge of discriminatory ARGs could help in framing targeted studies and associating the genes to the specific environmental and anthropogenic stressors, and ultimately in identifying the factors stimulating the proliferation of the resistance genes.

<u>Variable Importance: Ranking</u>

The concept of ranking the variables by "importance measure" was used to select the discriminatory ARGs. The variable importance was calculated by estimating the Gini importance. As proposed by Breiman [39], the importance of a variable X_m for predicting a class Y could be estimated by adding up the weighted impurity decreases $p(t)\ \Delta i(s_t, t)$ for all nodes t where X_m is used for dividing the node, averaged over number of total trees in the ensemble:

$$Imp(\mathrm{Xm}) = \frac{1}{\mathrm{NT}} \sum_{T} \sum_{t \in T: v(st) = Xm} p(t)\ \Delta i(st, t)$$

$p(t)$ is the proportion, N_t/N of samples reaching t and $v(s_t)$ is the variable used in split s_t.

When Gini index is used as the impurity function to measure the importance of the variable, that importance is called Gini Importance or Mean decrease in Gini.

The Gini importance is a measure of the purity of the node achieved when a particular variable is used to split the node. The Gini importance of the descendent node is calculated and compared with the preceding node. The smaller value of the Gini index indicates relatively purer node and thus less likelihood of misclassification. Gini index equal to 0 resembles the purest nodes whereas 1 is not. The Gini importance gives a measure of variable importance that can be used to rank variables. This approach proves useful for high dimensional data such as metagenomics data. The variable importance provides the features, best suitable for the classification of class variable.

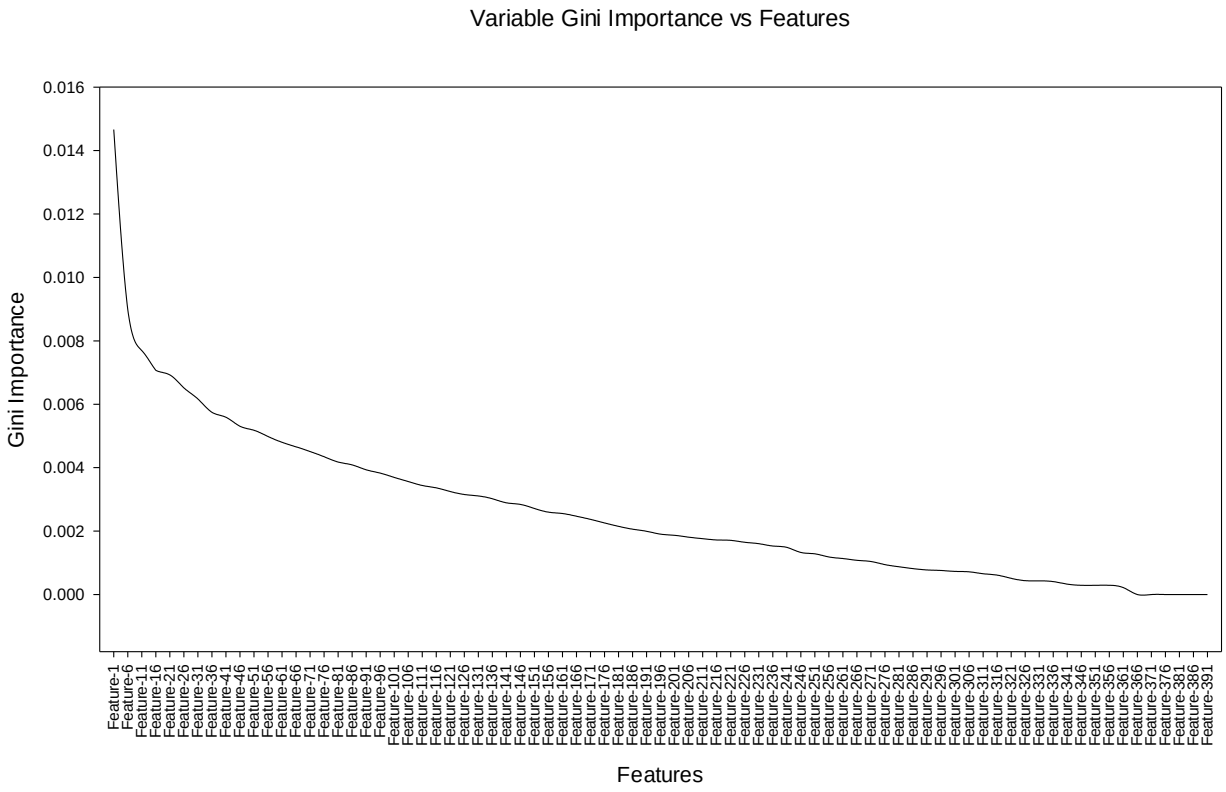


Figure 2.1: Variable importance determined by the ET algorithm.

This is a model example to demonstrate the output of ET Algorithm. The Y axis represents the Gini importance value and X axis contains features (in this study, ARGs) sorted in ascending order of their Gini importance values. The feature with highest Gini importance would be best suitable for separating the user labeled samples into the categorized groups and would be ranked first in the list. Similarly, all other features would be ranked based on their Gini importance. This plot represents the concept of ranking the variables.

Experimental Section

Data Sources

In total, 42 metagenomes were used in this study covering numerous environmental samples. These included 9 hospital effluents, 3 river source waters, 4 farm effluent, 14 WWTP influents and 12 WWTP effluents. Among these samples, the hospital effluent, river water, farm effluent metagenomes were downloaded from EBI and NCBI metagenome databases. WWTP influent and effluent samples were collected from WWTPs situated in the United States of America,

Switzerland, the Philippines, Sweden, Hong Kong and India. The detailed information about the metagenomes retrieved from databases is in Table 1, and influent and effluent samples are presented in Table 2. Figure 2.2 represents the map of sampling locations.

Figure 2.2: Map of sampling locations

Table 2.1: Metadata of different environmental samples obtained from public databases

Sample ID	Sample	Location	Description	Database	Accession Number	Reference
R.Farm1 R.Farm2 R.Farm3 R.Farm4	Farm Effluent	Cambridge, UK	Samples were collected from the effluent lagoon of a dairy farm	EMBL-EBI	ERR1193297 ERR1193298 ERR1193300 ERR1193301	Rowe et al. [16]
R.Hospital1 R.Hospital2 R.Hospital3 R.Hospital4 R.Hospital5 R.Hospital6	Hospital Effluent	Cambridge, UK	Samples were collected from the combined wastewater effluents of the main wards of University Hospital	EMBL-EBI	ERR1191817 ERR1191818 ERR1191819 ERR1191820 ERR1191821 ERR1191822	
R.River1 R.River2 R.River3	River Source Water	Cambridge, UK	Samples were collected within the River Cam Catchment	EMBL-EBI	ERR1193292 ERR1193293 ERR1193294	
N.Hospital1	Hospital Wastewater Effluent	Singapore	Sample was collected from General Ward	NCBI	SRR5997540	Ng, Charmaine, et al. [40]
N.Hospital2 N.Hospital3	Hospital Wastewater Effluent	Singapore	Samples were collected from Clinical Isolation ward	NCBI	SRR5997541 SRR5997548	

Table 2.2: Sampling information: WWTP Influent and Effluent Samples

Sample ID	Sample Type	Sampling Country	Coordinates	Location	WWTP Name	Sampling Date	Total Reads	Annotated Read
CHE1-P1-FE1	Final Effluent	Switzerland	47.089453, 8.319829	Lucerne	Emmen	17-May-16	10489905	2283
CHE1-P1-IN1	Influent	Switzerland	47.089453, 8.319829	Lucerne	Emmen	17-May-16	13932488	14136
CHE1-P2-FE1	Final Effluent	Switzerland	47.405586, 8.597585	Zurich	Dubendorf	18-May-16	14849493	2386
CHE1-P2-IN1	Influent	Switzerland	47.405586, 8.597585	Zurich	Dubendorf	18-May-16	15314202	18311
HKG1-P1-IN1	Influent	Hong Kong	22.406709, 114.213706	Hong Kong	Sha Tin	14-Jul-16	15763560	14979
HKG1-P2-IN1	Influent	Hong Kong	22.509115, 114.119869	Hong Kong	Shek Wu Hui	14-Jul-16	13174430	14938
HKG2-P1-FE1	Final Effluent	Hong Kong	22.406709, 114.213706	Hong Kong	Sha Tin	6-Dec-16	14867246	3083
HKG2-P1-IN1	Influent	Hong Kong	22.406709, 114.213706	Hong Kong	Sha Tin	6-Dec-16	9883170	6434
HKG2-P2-FE1	Final Effluent	Hong Kong	22.509115, 114.119869	Hong Kong	Shek Wu Hui	6-Dec-16	11626313	2143
HKG2-P2-IN1	Influent	Hong Kong	22.509115, 114.119869	Hong Kong	Shek Wu Hui	6-Dec-16	13707015	14578
IND1-P1-IN1	Influent	India	13.036238, 80.193738	Chennai	Nesapakkam	10-Mar-16	13045504	15421
IND1-P1-FE1	Final Effluent	India	13.036238, 80.193738	Chennai	Nesapakkam	10-Mar-16	13451894	3769
IND1-P2-FE1	Final Effluent	India	12.956851, 80.233880	Chennai	Perungudi	14-Mar-16	14368769	4131
IND1-P2-IN1	Influent	India	12.956851, 80.233880	Chennai	Perungudi	14-Mar-16	14414214	15375

PHL1-P2-IN1	Influent	Philippines	14.592113, 121.058931	Mandalayong City	Robinson's Gallera STP	29-Nov-16	14332977	20267
PHL1-P1-FE1	Final Effluent	Philippines	14.275940, 121.060087	Binan	Laguna Water	2-Dec-16	15018872	4761
PHL1-P1-IN1	Influent	Philippines	14.275940, 121.060087	Binan	Laguna Water	2-Dec-16	12222754	12309
PHL1-P2-FE1	Final Effluent	Philippines	14.592113, 121.058931	Mandalayong City	Robinson's Gallera STP	29-Nov-16	8985967	6579
SWE1-P1-FE1	Final Effluent	Sweden	57.704713, 12.926666	Boras	Gasslosa	8-Jun-16	17467346	2755
SWE1-P1-IN1	Influent	Sweden	57.704713, 12.926666	Boras	Gasslosa	8-Jun-16	11801763	10149
SWE1-P2-FE1	Final Effluent	Sweden	58.390452, 13.879866	Skovde	Skovde	9-Jun-16	15019026	1204
SWE1-P2-IN1	Influent	Sweden	58.390452, 13.879866	Skovde	Skovde	9-Jun-16	14524181	13441
USA1-P1-FE1	Final Effluent	USA	37.156200, -80.469794	Christiansburg	Christiansburg	1-Nov-16	9948362	1499
USA1-P1-IN1	Influent	USA	37.156200, -80.469794	Christiansburg	Christiansburg	1-Nov-16	13310751	14135
USA1-P2-FE1	Final Effluent	USA	37.201889, -76.447378	Hampton Roads	HRSD	19-Jan-17	12690702	1920
USA1-P2-IN1	Influent	USA	37.201889, -76.447378	Hampton Roads	HRSD	19-Jan-17	13460770	11776

P1&P2 represents two wastewater treatment plants sampled in each country

Sample Collection

WWTP Influent samples were collected after the grit removal and screening process, whereas effluent samples were collected after the disinfection process. Grab samples were collected from each site and transported to the lab on ice. Biomass from the liquid samples was filter-concentrated onto three separate 0.45 µm filters after homogenizing each sample by shaking. Each membrane filter was then preserved in 50% ethanol at -20 °C and subsequently shipped to the Molecular Biology Lab at Virginia Tech for DNA extraction and further analyses.

DNA Extraction

DNA was extracted from the filter concentrated samples using a FastDNA Spin Kit (MP Biomedicals) for soil according to the prescribed protocol. The total DNA was eluted in 100 µL of water and stored at -20 °C until further analysis. The concentration and quality of extracted DNA was analyzed using spectrophotometric analysis by NanoPearl and Qubit Spectrometer, and via gel electrophoresis.

Shotgun Metagenomic Sequencing

Library preparation via TrueSeq library prep kit and shotgun metagenomics sequencing on the Illumina HiSeq2500 platform with 2x100 paired-end reads were performed by the Biocomplexity Institute Genomic Sequencing Center, Blacksburg, VA, USA. Two of the samples were duplicated to verify the sequencing reproducibility. Further the FastQ files obtained from Shotgun Metagenomic sequencing and public databases were uploaded onto the MetaStorm server for downstream analyses [41].

Bioinformatics Analysis

The read matching pipeline was used for the functional annotation of metagenomics data by comparing the raw reads with a reference database using the approach called marker gene analysis. This approach uses the Diamond BLASTX aligner with the representative hit approach having E-value <1e-10, identity>90%, and minimum length of 25 amino acids for the annotations. Sequences were annotated to antibiotic resistance function using the comprehensive antibiotic resistance database CARD (1.0.6). The database version was consistent throughout the analyses. Further, the samples were compared based on the relative abundance of annotated genes where each sample was normalized based on the total number of 16S rRNA genes present in the sample.

Also, it should be pointed out that CARD contains various efflux proteins that could be found in both antibiotic resistant bacteria as well as antibiotic susceptible bacteria, and may not classify as a valid markers of resistance phenotype. In previous studies, they were related to efflux of antibiotic and classified as ARGs. Hence, in this study, efflux proteins were also considered to characterize the ARG profiles.

Statistical Analysis

A non-parametric multivariate statistical test, PERMANOVA was performed to compare whether the distribution and abundances of ARGs among various environments or defined groups, is statistically different.

NMDS was done on relative abundance matrix of resistant genes obtained from MetaStorm to visualize the level of similarity between the samples in the metadata using Bray Curtis Similarity method. Firstly, the similarity analysis was done with all the annotated genes obtained from MetaStorm server and then compared to the NMDS plot generated based on the relative abundance metric of ARGs selected upon the application of the ET algorithm. The analysis represents the effectiveness of ET algorithm in selecting the genes specific to the environment and enhancing the classification of different environments. All of the statistical analyses were performed using PAleontological STastics software (version 3.18)

Data Analysis using Extremely Randomized Trees Algorithm

Data Preprocessing & Labeling

After retrieving the annotated data from MetaStorm, the samples were labeled into groups. The labels are based on the kind of classification needed and are hypothesis driven. This could change based upon the study and type of information one wants to retrieve from the data. Some potential labeling examples are: labeling the samples based on the geographical location (country), type of environment (soil, wastewater, surface water, river, air, etc.) or sampling season (winter or summer). Figure 2.3 shows an example of how the labeling was done for influent wastewater samples collected from different countries. Similar strategy was used to label/group the data depending on the basis of classification.

(a) Raw Data

Annotated ARGs from Database

Sample Names

Relative Abundance

samples	AAC(6')-Ie-APH(2")-Ia	aadA6/aa	aadA25	aadA5	aadA13	DHA-15	OXA-333	OXA-256	TLA-1
IND1-P1-IN1	0.00888	0.00363	0.005	0.02008	0.00963	0.00061	0.00028	0.06079	0.00049
PHL1-P2-IN1	0.00059	0.01665	0.0093	0.00395	0.00124	0.00261	0.00069	0.04228	0.02068
USA1-P2-IN1	0.00041	0.00375	0.00148	0.00099	0	0	0.00284	0.06078	0.00414
CHE1-P2-IN1	0.00079	0.0108	0.00032	0.00595	0.00191	0	0.01922	0.00459	0
HKG1-P2-IN1	0.00693	0.00364	0.01014	0.00475	0.00201	0	0.00043	0.01537	0.00057
PHL1-P1-IN1	0.00082	0.01532	0.00764	0.00448	0.00316	0.00236	0.0002	0.02561	0.00676
HKG2-P2-IN1	0.01041	0.00188	0.00661	0.00487	0.00066	0.00015	0.00063	0.01744	0
SWE1-P1-IN1	0.00012	0.00123	0.00022	0.0013	0.00064	0	0.00641	0.00362	0.00018
USA1-P1-IN1	0.00013	0.00326	0.00183	0.00115	0	0	0.00989	0.0342	0.0046
HKG2-P1-IN1	0.01176	0.00062	0.00649	0.00261	0	0	0.00062	0.0077	0
IND1-P2-IN1	0.00602	0.00121	0.00714	0.02354	0.00809	0.00106	0	0.04662	0
SWE1-P2-IN1	0.0001	0.00303	0.00053	0.00196	0.00053	0	0.00902	0.00386	0
CHE1-P1-IN1	0.00065	0.01053	0.00047	0.00426	0.00187	0	0.0222	0.00677	0
HKG1-P1-IN1	0.00796	0.00238	0.01255	0.00728	0.00249	0.00019	0.00054	0.01268	0.00023

(b) Labeled Data

samples	Label	AAC(6')-Ie-APH(2")-Ia	aadA6/aa	aadA25	aadA5	aadA13	DHA-15	OXA-333	OXA-256	TLA-1
IND1-P1-IN1	IND-IN	0.00888	0.00363	0.005	0.02008	0.00963	0.00061	0.00028	0.06079	0.00049
PHL1-P2-IN1	PHL-IN	0.00059	0.01665	0.0093	0.00395	0.00124	0.00261	0.00069	0.04228	0.02068
USA1-P2-IN1	USA-IN	0.00041	0.00375	0.00148	0.00099	0	0	0.00284	0.06078	0.00414
CHE1-P2-IN1	CHE-IN	0.00079	0.0108	0.00032	0.00595	0.00191	0	0.01922	0.00459	0
HKG1-P2-IN1	HKG-IN	0.00693	0.00364	0.01014	0.00475	0.00201	0	0.00043	0.01537	0.00057
PHL1-P1-IN1	PHL-IN	0.00082	0.01532	0.00764	0.00448	0.00316	0.00236	0.0002	0.02561	0.00676
HKG2-P2-IN1	HKG-IN	0.01041	0.00188	0.00661	0.00487	0.00066	0.00015	0.00063	0.01744	0
SWE1-P1-IN1	SWE-IN	0.00012	0.00123	0.00022	0.0013	0.00064	0	0.00641	0.00362	0.00018
USA1-P1-IN1	USA-IN	0.00013	0.00326	0.00183	0.00115	0	0	0.00989	0.0342	0.0046
HKG2-P1-IN1	HKG-IN	0.01176	0.00062	0.00649	0.00261	0	0	0.00062	0.0077	0
IND1-P2-IN1	IND-IN	0.00602	0.00121	0.00714	0.02354	0.00809	0.00106	0	0.04662	0
SWE1-P2-IN1	SWE-IN	0.0001	0.00303	0.00053	0.00196	0.00053	0	0.00902	0.00386	0
CHE1-P1-IN1	CHE-IN	0.00065	0.01053	0.00047	0.00426	0.00187	0	0.0222	0.00677	0
HKG1-P1-IN1	HKG-IN	0.00796	0.00238	0.01255	0.00728	0.00249	0.00019	0.00054	0.01268	0.00023

Labels/Groups

Figure 2.3: This figure represents the methodology of Data Labeling. The Raw data is a sample subset of WWTP influent samples. Further, this raw data was labeled according to their sampling locations (in this case countries), the highlighted column in the dataset.

<u>Step-wise execution of ET Algorithm</u>

1. ExtraTrees Classifier

The ET algorithm was executed on the labeled dataset using Python (3.2.5). The Scikit learn pre-built classifier ExtraTreesClassifier was used to build the ensemble and to calculate the variable importance. The number of estimators was set at a default value of 1000. The algorithm gives a list of features (genes) best suited for discriminating classes.

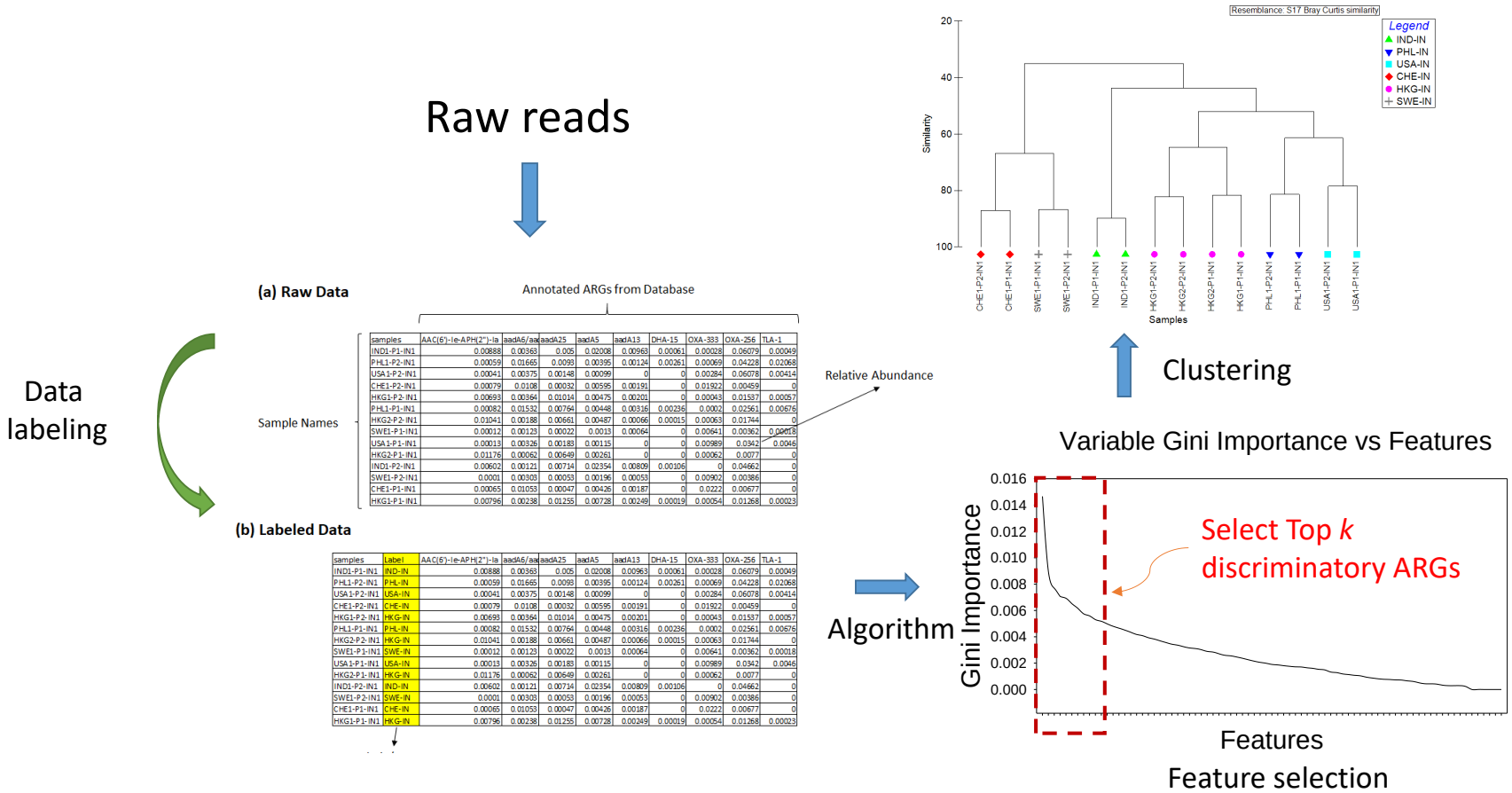
2. Identification of Discriminatory ARGs using ExtraTrees Classifier

The input dataset consists of a matrix X*nxm* where rows represent the samples and columns the features (ARGs) and a vector input Y containing the labels of each sample. The system, will take

the X and Y as the input and return a list of k ARGs defined as the most discriminatory ARGs. Top *k* genes were chosen in the descending order of Gini importance and were then extracted from the main dataset. The new dataset contains only the classes and the *k* attributes chosen using the classifier.

3. Clustering

Hierarchical clustering was obtained using PRIMER-E Software (v6). The clustering technique used was *"group average"* and the similarity measure was estimated using Bray Curtis method. The clustering was used to validate the classification of samples into their respective groups or labels. The new dataset was then executed using an R code to generate a heatmap projecting the relative abundances of genes. The library used for heatmap was *Complex Heatmap* [42]. The obtained heatmap has class variables (i.e., labeled samples) in the columns and the attributes (ARGs) in the rows.



samples	AAC(6')-Ie-APH(2")-Ia	aadA6/aa	aadA25	aadA5	aadA13	DHA-15	OXA-333	OXA-256	TLA-1
IND1-P1-IN1	0.00888	0.00363	0.005	0.02008	0.00963	0.00061	0.00028	0.06079	0.00049
PHL1-P2-IN1	0.00059	0.01665	0.0093	0.00395	0.00124	0.00261	0.00069	0.04228	0.02068
USA1-P2-IN1	0.00041	0.00375	0.00148	0.00099	0	0	0.00284	0.06078	0.00414
CHE1-P2-IN1	0.00079	0.0108	0.00032	0.00595	0.00191	0	0.01922	0.00459	0
HKG1-P2-IN1	0.00693	0.00364	0.01014	0.00475	0.00201	0	0.00043	0.01537	0.00057
PHL1-P1-IN1	0.00082	0.01532	0.00764	0.00448	0.00316	0.00236	0.0002	0.02561	0.00676
HKG2-P2-IN1	0.01041	0.00188	0.00661	0.00487	0.00066	0.00015	0.00063	0.01744	0
SWE1-P1-IN1	0.00012	0.00123	0.00022	0.0013	0.00064	0	0.00641	0.00362	0.00018
USA1-P1-IN1	0.00013	0.00326	0.00183	0.00115	0	0	0.00989	0.0342	0.0046
HKG2-P1-IN1	0.01176	0.00062	0.00649	0.00261	0	0	0.00062	0.0077	0
IND1-P2-IN1	0.00602	0.00121	0.00714	0.02354	0.00809	0.00106	0	0.04662	0
SWE1-P2-IN1	0.0001	0.00303	0.00053	0.00196	0.00053	0	0.00902	0.00386	0
CHE1-P1-IN1	0.00065	0.01053	0.00047	0.00426	0.00187	0	0.0222	0.00677	0
HKG1-P1-IN1	0.00796	0.00238	0.01255	0.00728	0.00249	0.00019	0.00054	0.01268	0.00023

samples	Label	AAC(6')-Ie-APH(2")-Ia	aadA6/aa	aadA25	aadA5	aadA13	DHA-15	OXA-333	OXA-256	TLA-1
IND1-P1-IN1	IND-IN	0.00888	0.00363	0.005	0.02008	0.00963	0.00061	0.00028	0.06079	0.00049
PHL1-P2-IN1	PHL-IN	0.00059	0.01665	0.0093	0.00395	0.00124	0.00261	0.00069	0.04228	0.02068
USA1-P2-IN1	USA-IN	0.00041	0.00375	0.00148	0.00099	0	0	0.00284	0.06078	0.00414
CHE1-P2-IN1	CHE-IN	0.00079	0.0108	0.00032	0.00595	0.00191	0	0.01922	0.00459	0
HKG1-P2-IN1	HKG-IN	0.00693	0.00364	0.01014	0.00475	0.00201	0	0.00043	0.01537	0.00057
PHL1-P1-IN1	PHL-IN	0.00082	0.01532	0.00764	0.00448	0.00316	0.00236	0.0002	0.02561	0.00676
HKG2-P2-IN1	HKG-IN	0.01041	0.00188	0.00661	0.00487	0.00066	0.00015	0.00063	0.01744	0
SWE1-P1-IN1	SWE-IN	0.00012	0.00123	0.00022	0.0013	0.00064	0	0.00641	0.00362	0.00018
USA1-P1-IN1	USA-IN	0.00013	0.00326	0.00183	0.00115	0	0	0.00989	0.0342	0.0046
HKG2-P1-IN1	HKG-IN	0.01176	0.00062	0.00649	0.00261	0	0	0.00062	0.0077	0
IND1-P2-IN1	IND-IN	0.00602	0.00121	0.00714	0.02354	0.00809	0.00106	0	0.04662	0
SWE1-P2-IN1	SWE-IN	0.0001	0.00303	0.00053	0.00196	0.00053	0	0.00902	0.00386	0
CHE1-P1-IN1	CHE-IN	0.00065	0.01053	0.00047	0.00426	0.00187	0	0.0222	0.00677	0
HKG1-P1-IN1	HKG-IN	0.00796	0.00238	0.01255	0.00728	0.00249	0.00019	0.00054	0.01268	0.00023

Figure 2.4: Methodology for Feature selection and Clustering

Results and Discussion

ARG Abundance in WWTPs Influent and Effluent Samples

The core resistome or core ARGs were defined as those ARGs that were detected in all the samples of the same type. Among 14 wastewater influent samples, 189 ARGs were shared among all the samples and this is hereafter referred as the wastewater influent core (Supplementary Table S1). In 12 wastewater effluent samples, 48 ARGs were common to all the samples. The shared ARGs accounted for 88.4±2.4% and 61.5±11.6% of the Total Relative Abundance in influent and effluent samples, respectively. Based on the annotations, every sample had various distinct ARGs detected in them. Supplementary Table S2 presents the number of distinct ARGs detected in each influent and effluent sample. It was interesting to note that the relative abundance of distinct ARGs in these samples only contributed 11.6±2.4% and 38.5±11.6% towards the total relative abundance of influent and effluent samples, respectively.

As observed, more than half of the annotated ARGs in each of the wastewater influent samples were distinct, but their contribution towards the total relative abundance was small. Existing methods of analysis only consider similarity and dissimilarity of the samples, often biased towards the strong variables, in this case highly abundant genes. Shared ARGs in influent samples accounts for almost 90% of total relative abundance, so the distinct ARGs may get over-shadowed and important information would possibly get missed when applying typical methodologies. However, each distinct ARG is important in the grand scheme of combating antibiotic resistance. Hence, it is also essential to explore the low-level variations and associated ARGs.

Firstly, to capture the geospatial variation of ARG composition in wastewater collected from different locations the WWTP influent samples were grouped according to their sampling location (countries). The ARG abundance in influent samples varied from 1.4 to 3 ARG copies per 16S rRNA copy with samples from India and Hong Kong having the highest relative abundance of ARGs as compared with samples from Sweden, USA, and Switzerland (Figure 2.5). Secondly, the effect of biological treatment was analyzed by grouping the samples based on their type (i.e., influent vs. effluent). The analysis showed a decrease in the average number of ARG copies per 16S rRNA from 2.4 for the influent samples to 1.4 in the effluent samples (Figure 2.6). The PERMANOVA test indicated that total ARG abundances were significantly different (P-value <0.005) among all the groups taken into consideration.

With the metagenomics analysis, it should be noted that the annotations are based on the database used and could be biased toward these databases. There are several databases available for antibiotic resistance (such as DeepARG-DB, SNC-ARDB, CARD). This study used CARD as it is a well-curated database and extensively documented in the literature.

In consequence, the annotations obtained may not be representative of the actual/or full-profile of these samples as these depend on the information available in the literature. Considering the complexity of microbial environments as well as the numerous niches and corresponding anthropogenic pressures, the potential for many novel or unidentified ARGs should not come as a surprise. Importantly, their absence from the available databases should be considered when attempting to characterize any environment. Furthermore, the annotations are based on similarity search, which infers that there could be many ARGs that were missed during the annotations owing to the limited knowledge, computational abilities and available technologies.

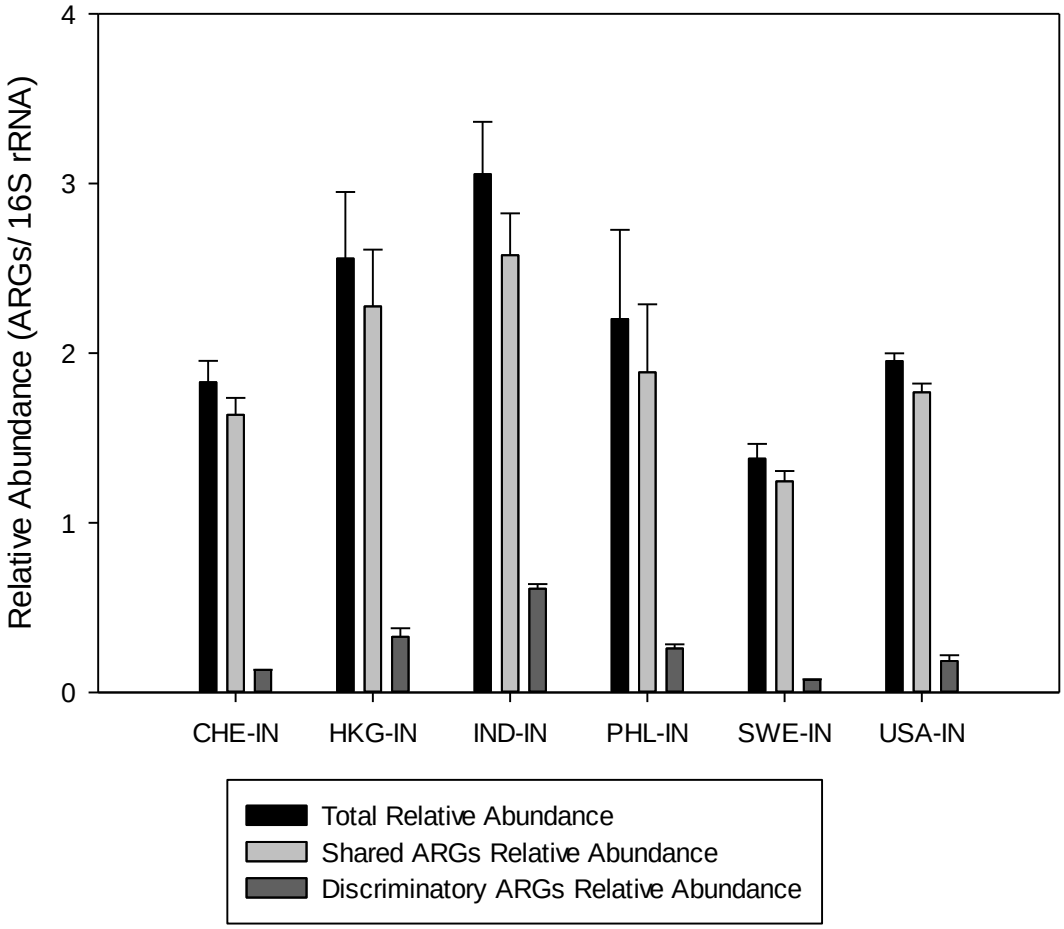


Figure 2.5: Mean relative abundance of influent samples segregated according to the sampling location.

Legend: CHE: Switzerland, HKG: Hong Kong, IND: India, PHL: Philippines, SWE: Sweden, USA: United States of America. IN: Influent

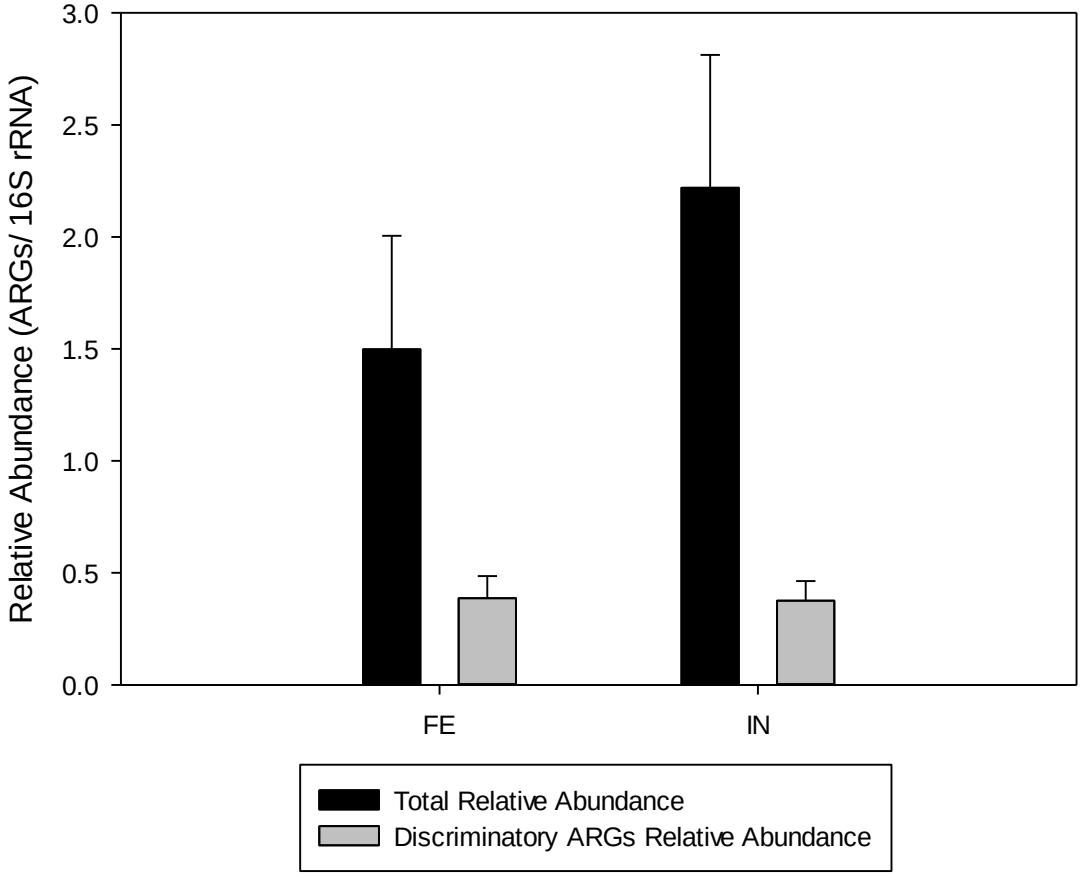


Figure 2.6: Mean relative abundance of wastewater samples segregated according to sample type.

Legend: CHE: Switzerland, HKG: Hong Kong, IND: India, PHL: Philippines, SWE: Sweden, USA: United States of America. IN: Influent. FE: Final Effluent

Identification of discriminative ARGs by the user defined labels

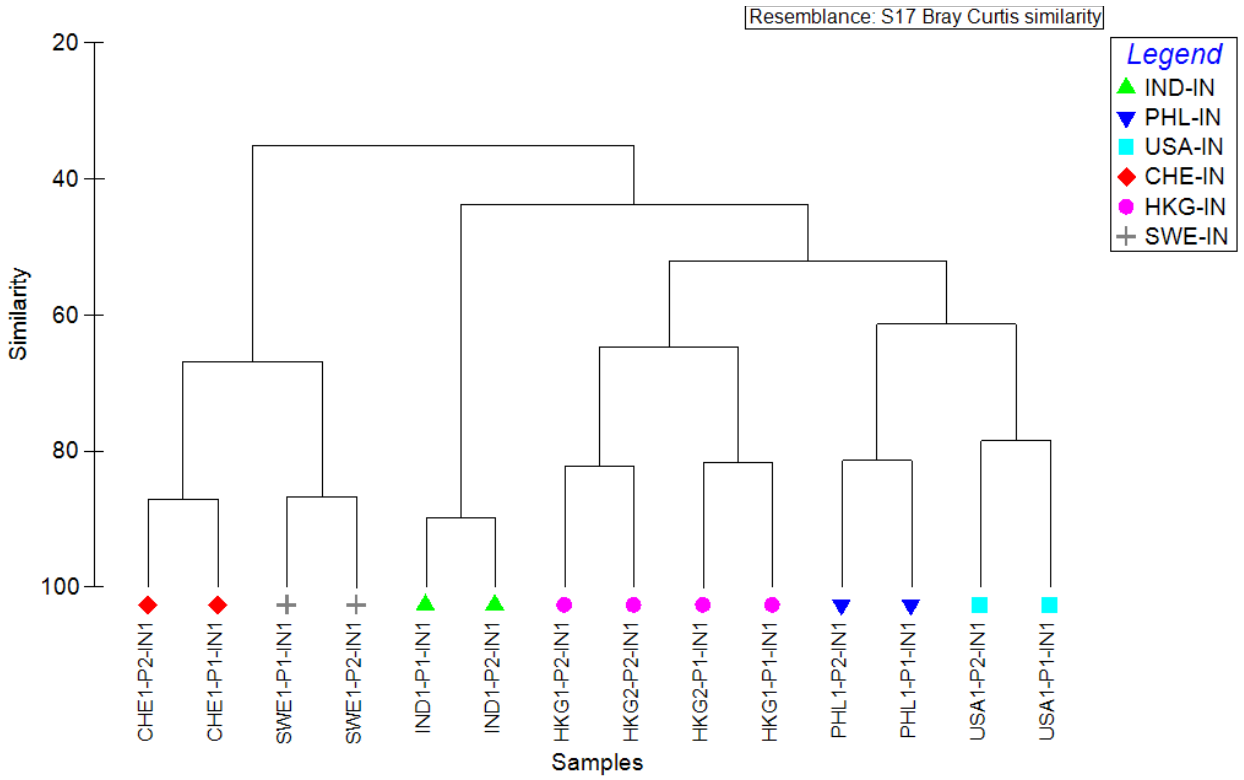
Classification based on sampling location

As discussed in the previous sections, influent samples from different locations were grouped according to countries where they were collected in order to find the discriminative ARGs between these samples. ExtraTrees classifier was used to identify the discriminative ARGs among influent samples. The top 50 discriminative ARGs according to the Gini importance score were selected to cluster these samples. The similarity in ARG compositions among the samples was evaluated using NMDS. As shown in Figure 2.8, the distances among the same group samples decreased and the samples clustered more closely when similarity was evaluated using the discriminatory genes identified by the classifier. This is also evident from the hierarchical clustering in Figure 2.7(a) as all the samples clustered in their respective groups. This shows that ET algorithm was able to select the discriminative ARGs.

The heatmap shown in Figure 2.7(b) was generated using the relative abundances of the 50 discriminatory genes, which were further sorted into 10 antibiotic classes. Based on the observation, the genes conferring resistance to aminoglycoside antibiotics were found to be more prevalent in samples from India and Hong Kong when compared with other locations. The beta-lactam resistance genes *VEB-3* and *VEB-5* were detected in all locations except Sweden and Switzerland. In contrast, relative abundance of *OXA-333* showed five-fold increase in samples from Sweden, USA and Switzerland when compared with others. Interestingly, the Fosfomycin resistance gene *FosC2* was only detected in the samples from USA. MLS resistance genes showed higher abundance in India, Hong Kong, the Philippines and the USA when compared with Sweden and Switzerland. Multidrug category genes were found to be present everywhere but with highest abundance in India and Hong Kong. Quinolone resistance gene qnrb65 was only detected in samples from Sweden. The highest relative abundance was observed for the clinically important sulfonamide resistance gene *sul1* in samples from India followed by Hong Kong, Philippines, USA, Sweden and Switzerland. Sulfonamide resistance genes were among the highest abundant genes in all the influent samples. Another noticeable observation was that tetracycline resistance gene (*tetD*) was only detected in samples from Hong Kong and Philippines. Though at relatively low levels in influent samples, trimethoprim resistance genes (i.e., dfrB3 and dfrA16) showed stronger presence in samples from India. In descending order, the ratio of total relative abundance

of discriminatory genes to the total relative abundance is highest for India (19-20%) followed by Hong Kong (10-16%), Philippines (10-13%), USA (8-10%), Switzerland (7-7.5%) and Sweden (5-5.3%). This ratio suggests that of the 50 discriminatory genes, India has the highest percentage in its influent samples. In general, feature projection methods have a bias towards genes with high relative abundance which could overshadow the ARGs with low relative abundance and may miss their variations. However, such genes could be important in associating the environmental impacts of such genes and should be included in the study. This algorithm could identify such ARGs and their variations, despite their lower relative abundance. In essence, the ET algorithm effectively captures low level variances with a better resolution, which are difficult to detect in the broad spectrum studies. The analysis provides insight into the antibiotic resistance presence at different locations geographically.

(a)



(b)

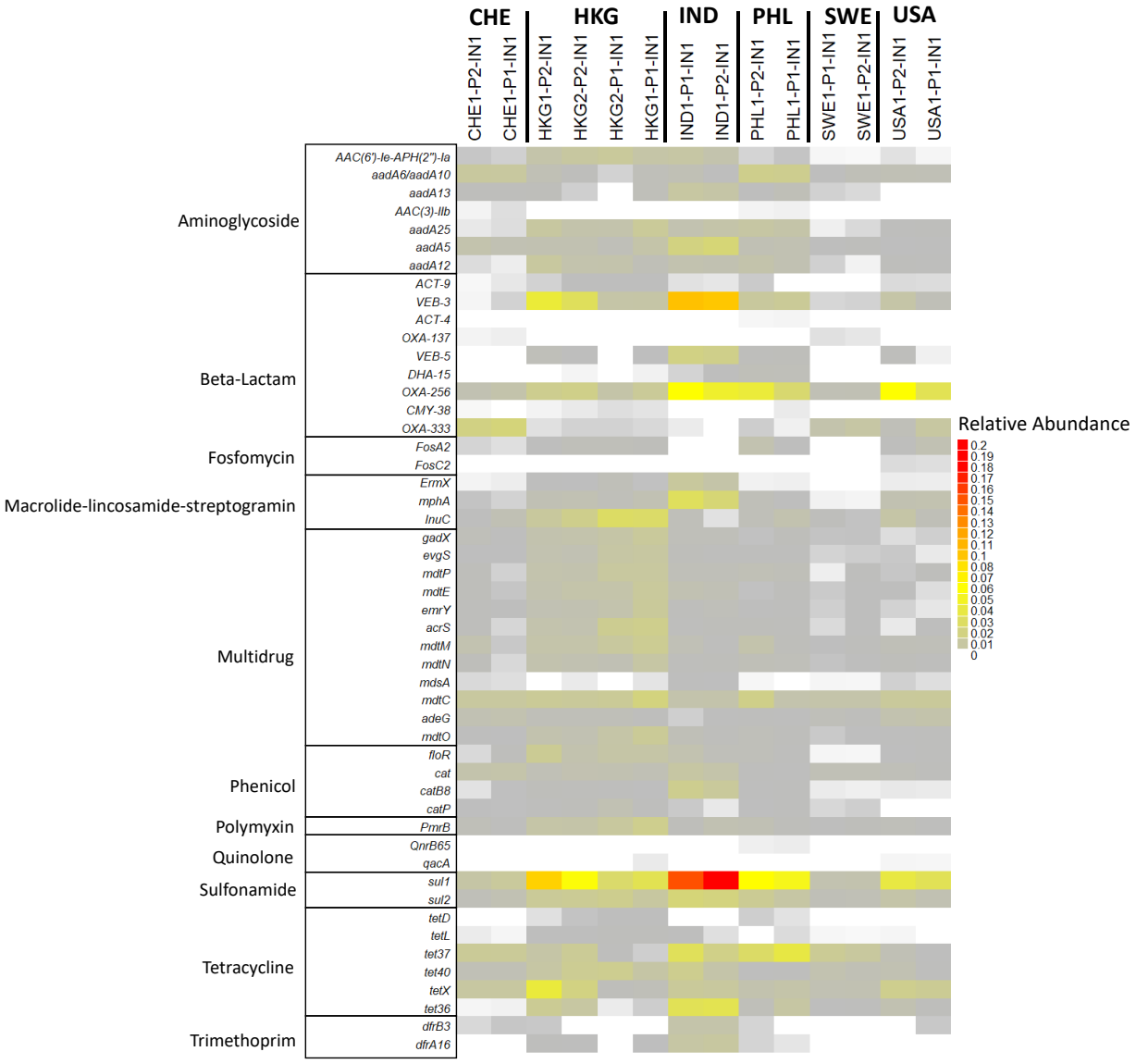
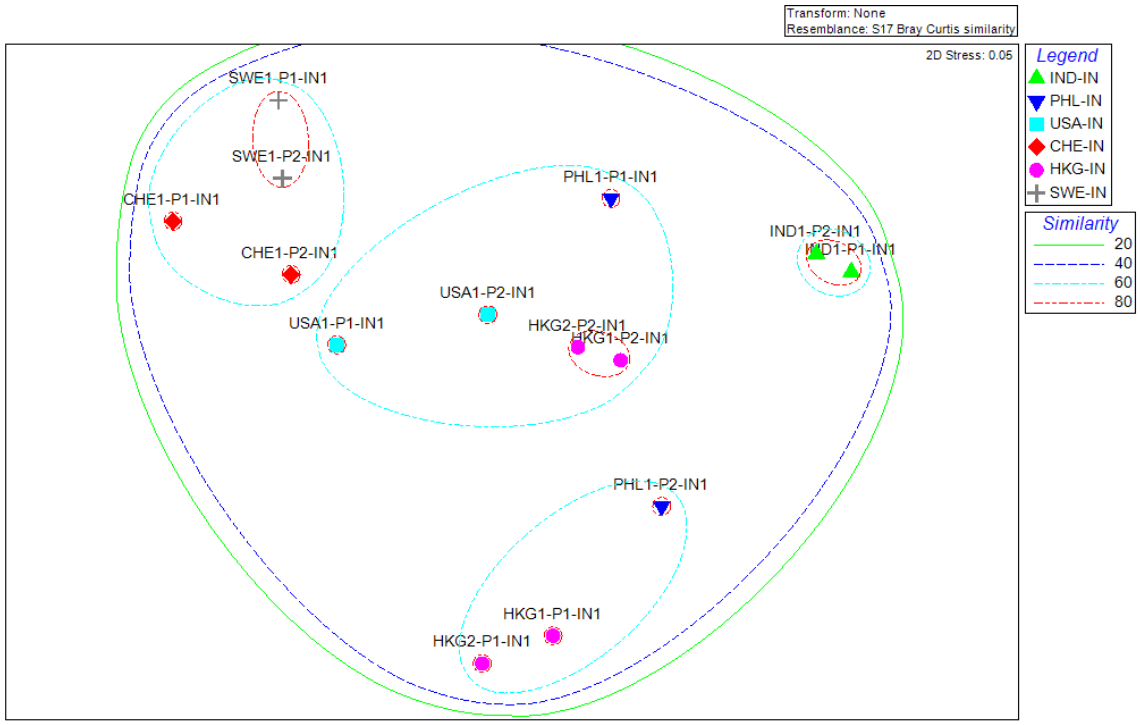


Figure 2.7: (a) Hierarchical clustering & (b) Heatmap of WWTP-Influent samples from different countries based on the relative abundance of discriminatory ARGs.

The list of ARGs presented in the heatmap were extracted by applying ET algorithm for the classification of different samples.

Legend: CHE: Switzerland, HKG: Hong Kong, IND: India, PHL: Philippines, SWE: Sweden, USA: United States of America. P1 & P2 stands for the different WWTP sampled in each country. IN: Influent

(a)



(b)

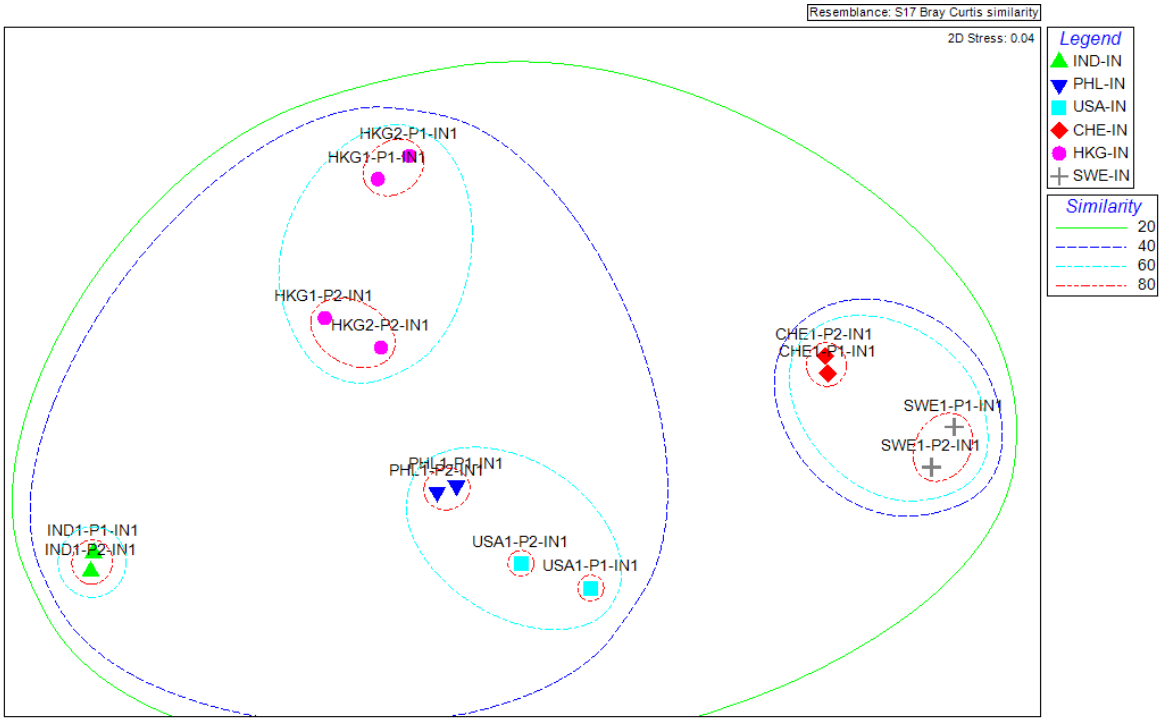


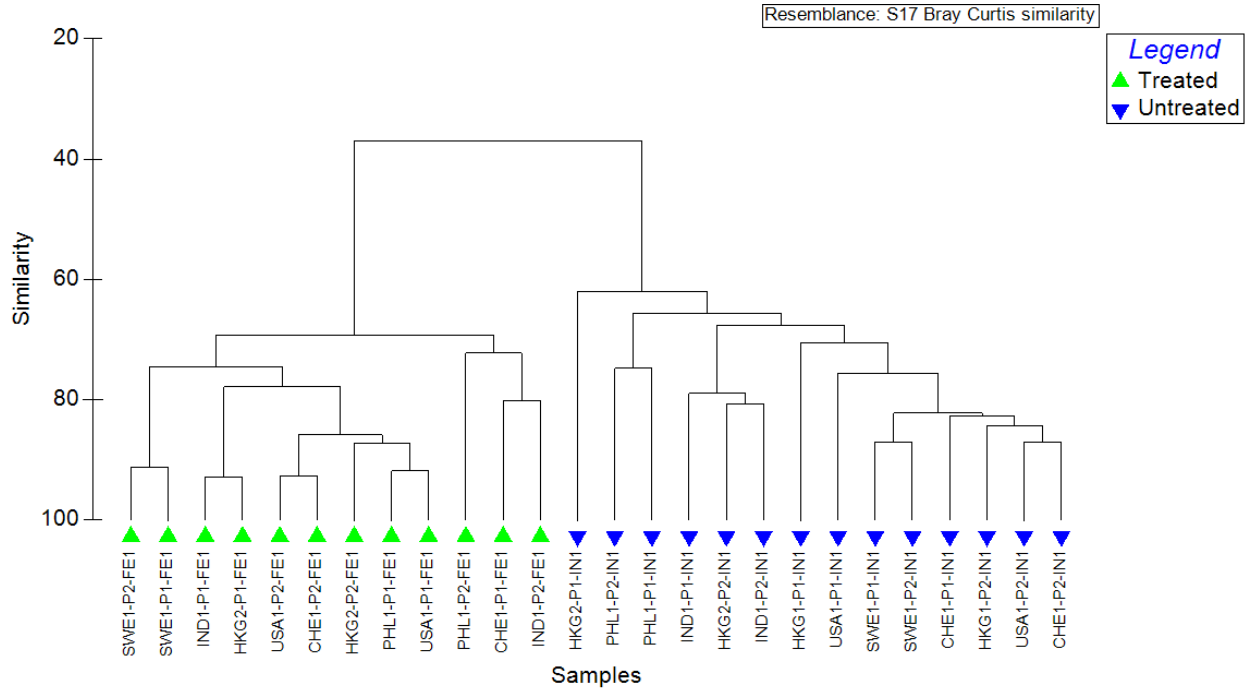
Figure 2.8: (a) NMDS plot for influent samples using all the annotated ARGs (b) NMDS Plot for influent samples using the discriminatory ARGs

Classification between WWTP influent and effluent samples

In this analysis, an attempt was made to capture the differences between influent and effluent samples in the context of antibiotic resistance at the gene level. Top 50 discriminatory genes were selected using ET algorithm in the descending order of their Gini importance. The clustering shown in Figure 2.9(a) represents the classification of wastewater influent (untreated) and final effluent (treated) samples using the genes selected using the ET algorithm. The relative abundance of discriminatory ARGs was visualized using a heatmap (Figure 2.9(b)). The discriminatory genes were found to be associated with 11 different antibiotic classes. The NMDS based similarity analysis showed increased similarity and better clustering among the samples using discriminatory ARGs (see Figure 2.10). The analysis was found to be effective in capturing and visualizing the effect of biological treatment in a wastewater plant. It successfully classified genes where there was variation pre and post treatment as important genes. For example, all the genes explored in the aminoglycoside, beta-lactam, glycopeptide, phenicol, tetracycline resistance categories showed a decline in final effluent. There were several instances where an increase in gene abundance was observed post treatment such as in the *amrB* multidrug resistance gene, trimethoprim resistance genes *dfrE* and *dfrB6*, and rifamycin resistance gene *rph*. The largest variation was observed in trimethoprim resistance gene *dfrE* which increased two folds on average in effluent samples. This proves that this methodology was beneficial in identifying the ARGs that were varying in untreated and treated wastewater samples.

With the identification and inclusion of novel ARGs in the databases, there is an increase in the size and dimensionality of metagenomics data which presents subsequent challenges in analyzing such data. The proposed methodology can simplify such analyses and provide valuable insights as discussed above.

(a)


Resemblance: S17 Bray Curtis similarity
Legend
Treated
Untreated
Similarity
20
40
60
80
100
SWE1-P2-FE1
SWE1-P1-FE1
IND1-P1-FE1
HKG2-P1-FE1
USA1-P2-FE1
CHE1-P2-FE1
HKG2-P2-FE1
PHL1-P1-FE1
USA1-P1-FE1
PHL1-P2-FE1
CHE1-P1-FE1
IND1-P2-FE1
HKG2-P1-IN1
PHL1-P2-IN1
PHL1-P1-IN1
IND1-P1-IN1
HKG2-P2-IN1
IND1-P2-IN1
HKG1-P1-IN1
USA1-P1-IN1
SWE1-P1-IN1
SWE1-P2-IN1
CHE1-P1-IN1
HKG1-P2-IN1
USA1-P2-IN1
CHE1-P2-IN1
Samples

(b)

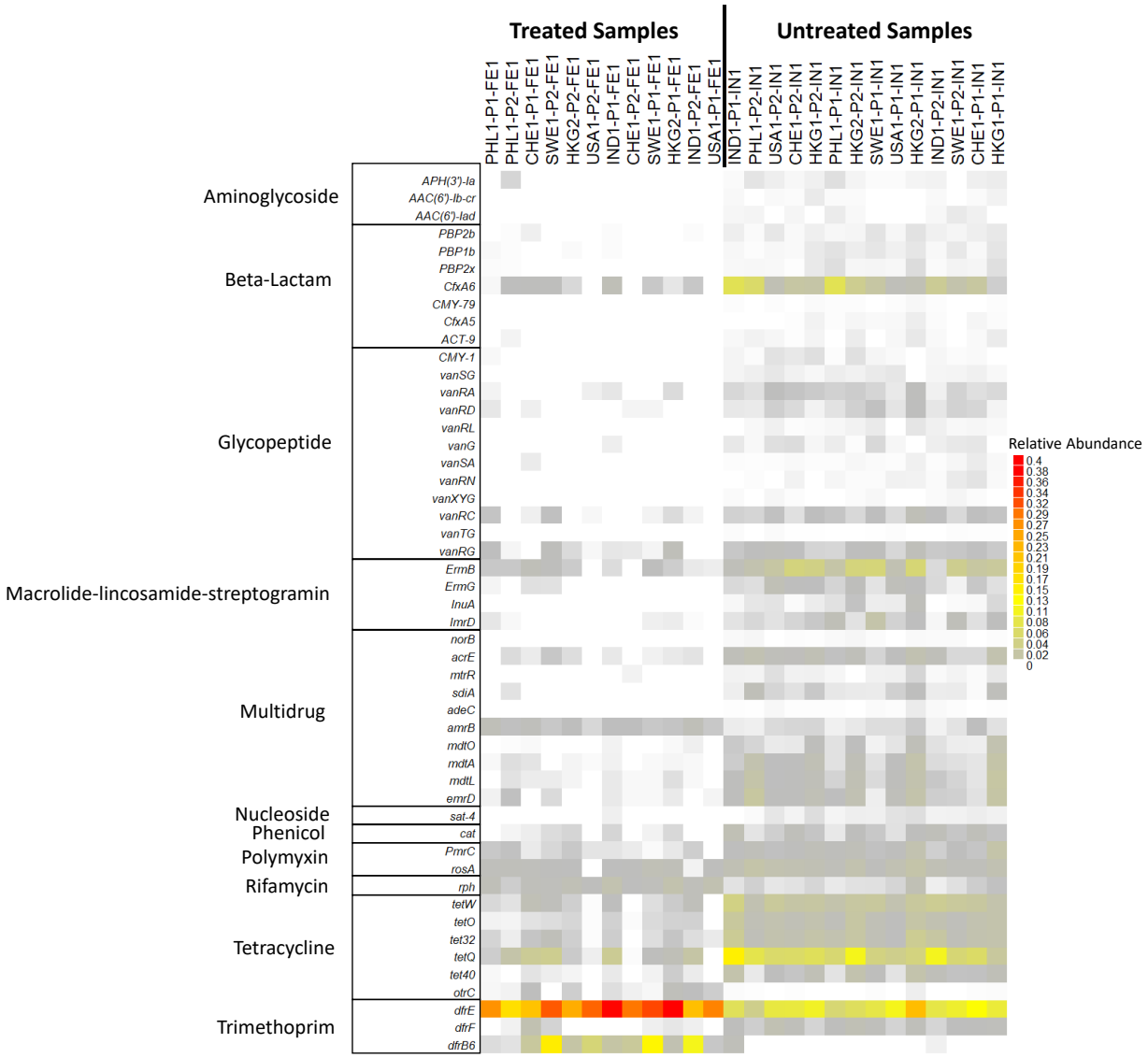
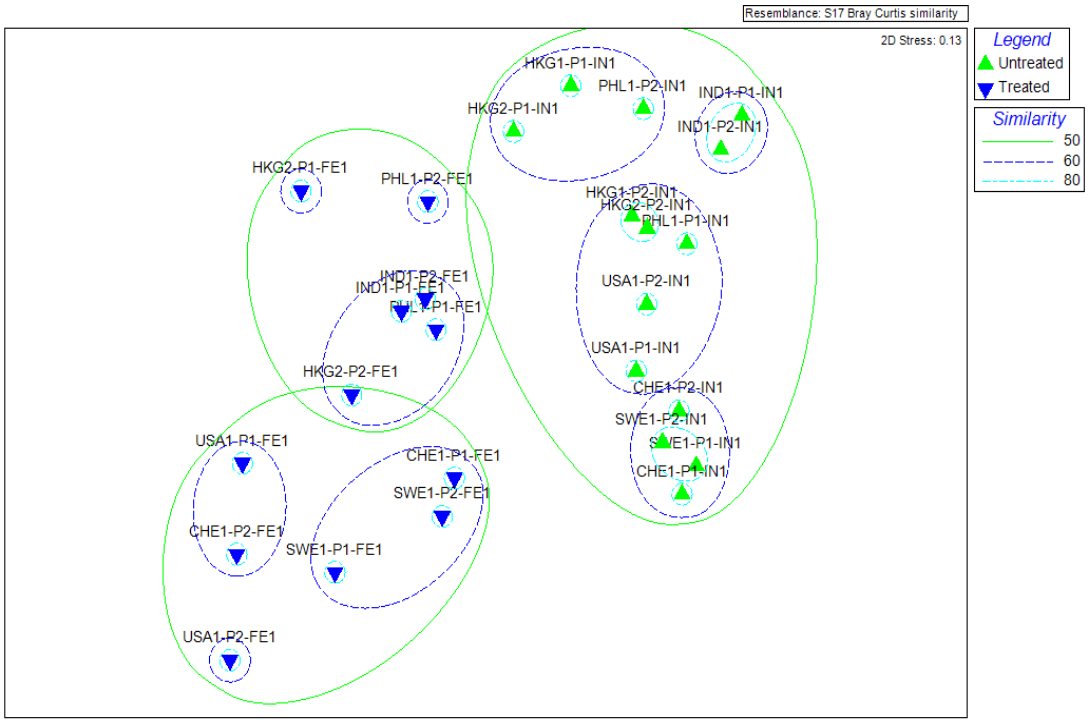
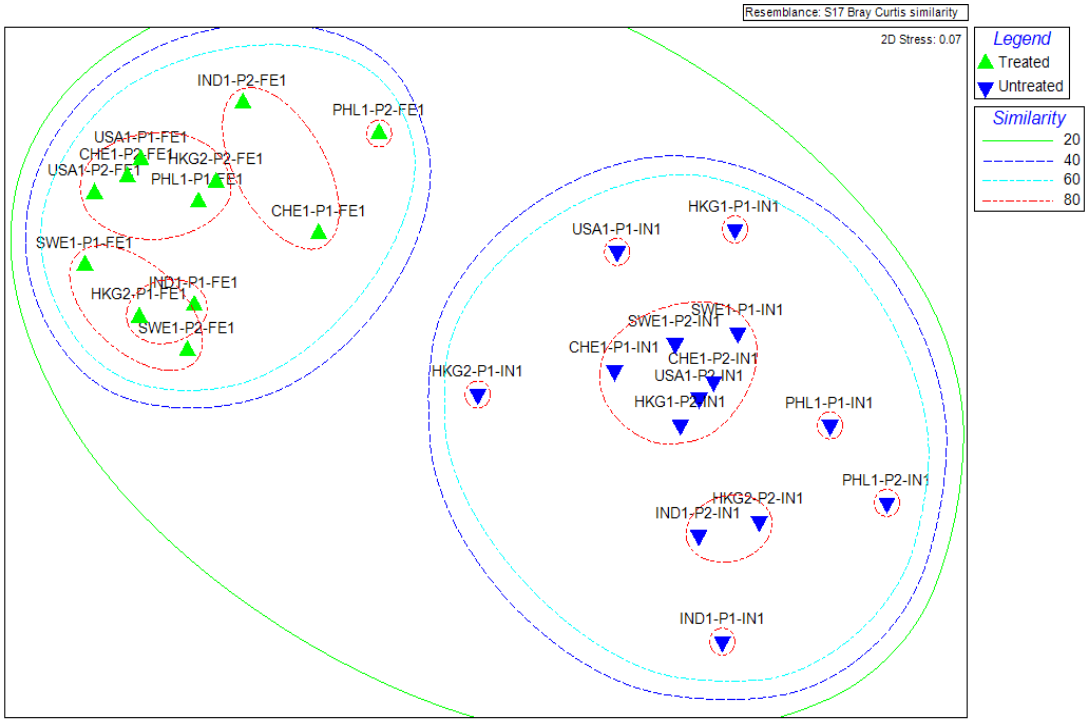


Figure 2.9: (a) Hierarchical clustering & (b) Heatmap between Influent and Effluent samples of WWTP based on the relative abundance of discriminatory ARGs. The list of ARGs presented in the heatmap were extracted by applying ET algorithm for the classification of different samples.

Legend: CHE: Switzerland, HKG: Hong Kong, IND: India, PHL: Philippines, SWE: Sweden, USA: United States of America. P1 & P2 stands for the different WWTP sampled in each country. IN: Influent. FE: Final Effluent



(a)



(b)

Figure 2.10: (a) NMDS plot for influent and effluent samples using all the annotated ARGs (b) NMDS Plot for influent and effluent samples using the discriminatory ARGs

Comparison with different aquatic environment samples

In this analysis, wastewater influent and effluent samples from the present study were compared with metagenomes retrieved from public databases to compare different aquatic environments (Table 1 & Table 2). The purpose of this comparison was to assess the performance of the ET algorithm on public database samples, validating that the algorithm is effective on samples outside of the current study.

The samples included in the public database analysis were selected based on two key criteria (1) the effluent was sampled from an aquatic environment (2) the samples selected would represent a spectrum of different environments, including river, farm and hospital effluents. This study aimed to assess different aquatic environments to identify whether there are any ARGs specific to certain environments, for example, if there was an ARG only detected in hospital effluent. Additionally, showing a range of aquatic environments could elucidate any patterns among environments and demonstrate interconnectivity between different aquatic environments.

Based on these desired qualifications, two papers describing aquatic environments of interest were selected and their associated samples were retrieved from public databases. The R.Farm effluent, R.Hospital effluent and R.River water metagenomes were selected from the study conducted by Rowe et al in Cambridge and the N.Hospital effluent samples were taken from study conducted in Singapore by Ng. Charmaine et al.

Wastewater influent sample core ARGs were compared with other environments. The core ARG from these different samples were compared with influent sample's core (Figure 2.11). It was found that all the 48 core ARGs found in final effluent samples were shared with influent samples core. Similarly, influent core was compared with farm effluent and river water and it was found that 147 and 124 ARGs, respectively, were in common. Further, the two hospital core ARGs were compared with influent core ARGs. The result showed about 157 ARGs were common across all samples showing that the influent samples core is very similar to the core of wastewater samples from the hospitals, which is reasonable. Interestingly, Hospital-2 had 150 ARGs which were unique to those metagenomes indicating that hospital wastewater samples serve as a great reservoir for ARGs.

Further to capture the distinctness among these samples, ET algorithm was used to generate the list of discriminatory ARGs which could classify these samples. The similarity/dissimilarity analysis using NMDS (Figure 2.13) and hierarchical clustering (Figure 2.12(a)) shows that discriminatory genes were able to further classify the samples in their respective groups. As observed in Figure 2.12(a), three major clusters were identified i.e. hospital effluents (two clusters), WWTP influent samples (one cluster), and farm effluent, river source water and effluent source water (one cluster). The heatmap (Figure 2.12(b)) represents the relative abundances of discriminatory genes categorized according to their antibiotic classes in rows and clustering of environmental samples (influent, effluent, farm effluent, river source water, and hospital effluents) in columns.

The heatmap offers insight to the dispersion of certain genes (Figure 2.12(b)). For example, glycopeptide resistance genes had relatively lower abundance in hospital discharges tested in this study. Also, ARGs pertaining to aminoglycoside (*AAC(6')-IB*, *APH(3')-IB, APH(6)-ID, APH(3')-IA),* beta-Lactams *(OXA-226,OXA-135,* and *TEM-126*), and MLS (*msrE)* were abundant across all the hospital discharges. The identification of a few, very specific ARGs in the hospital discharges provides evidence that the use of certain specific drugs in the hospitals potentially increases abundance of specific ARGs in discharges.

Similarly, aminoglycosides (*AAC(6')-Iad, aadA* 14) and glycopeptide (*van*XYN) resistance genes were specifically detected only in farm effluent samples, and there is potentially a farm-specific characteristic that alludes to an increased loading of this gene type. Multidrug resistance gene *arl*S and MLS resistance gene *mph*B were only detected in farm, hospital, and river samples from study done by Rowe et al in Cambridge, UK. Hence, it could be possible that these samples might be influenced by each other. Multidrug resistance gene *mgr*A had an increased abundance in river samples. Bacitracin, tetracycline (*tet39 and tetQ),* and trimethoprim resistance genes were abundant across all the groups. Such specific observations and analysis can improve our overall understanding about antibiotic resistance and the causation of where certain gene types are abundant. Additionally, it can help us generate potential hypotheses to identify the stressors that exacerbate ARG prevalence. Results from this heatmap suggest that influences and variations in ARG abundance on a fine scale can be well captured and visualized by the ET algorithm.

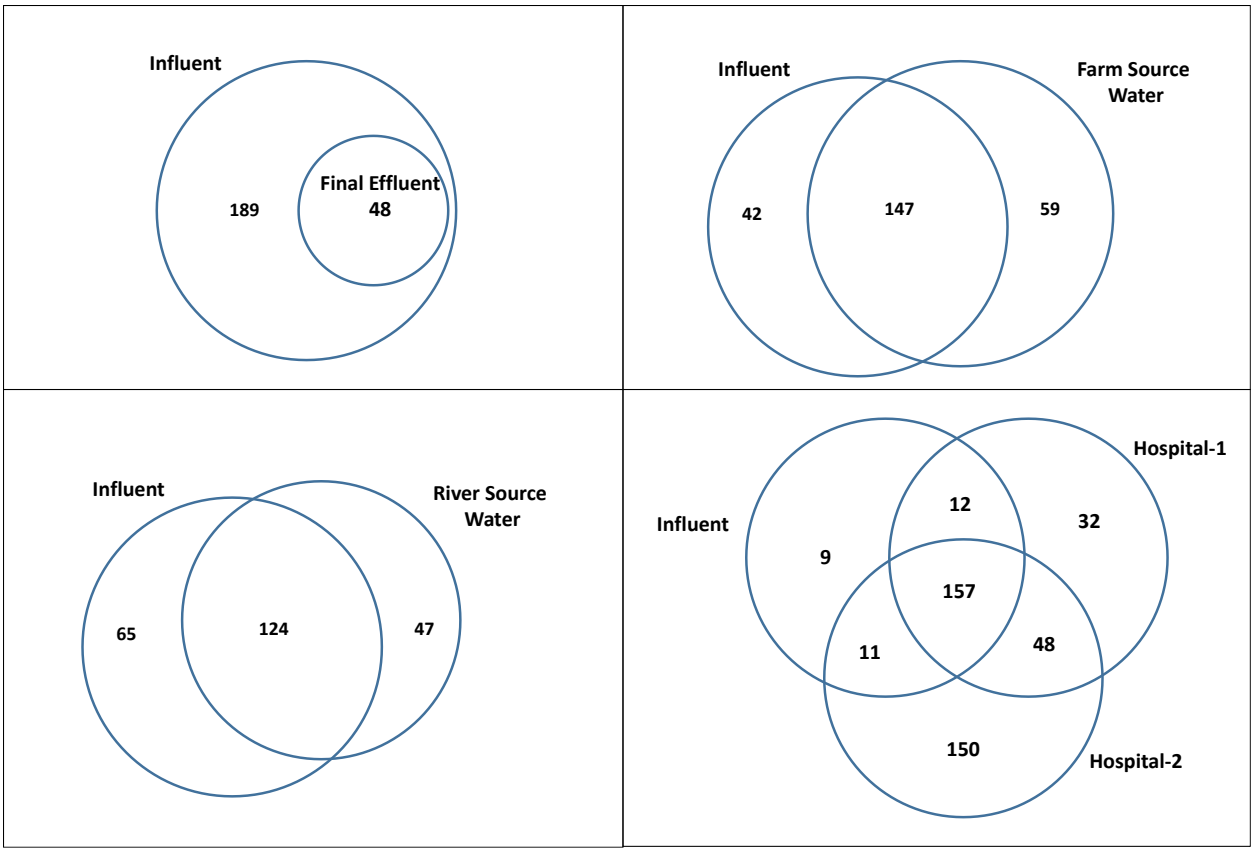
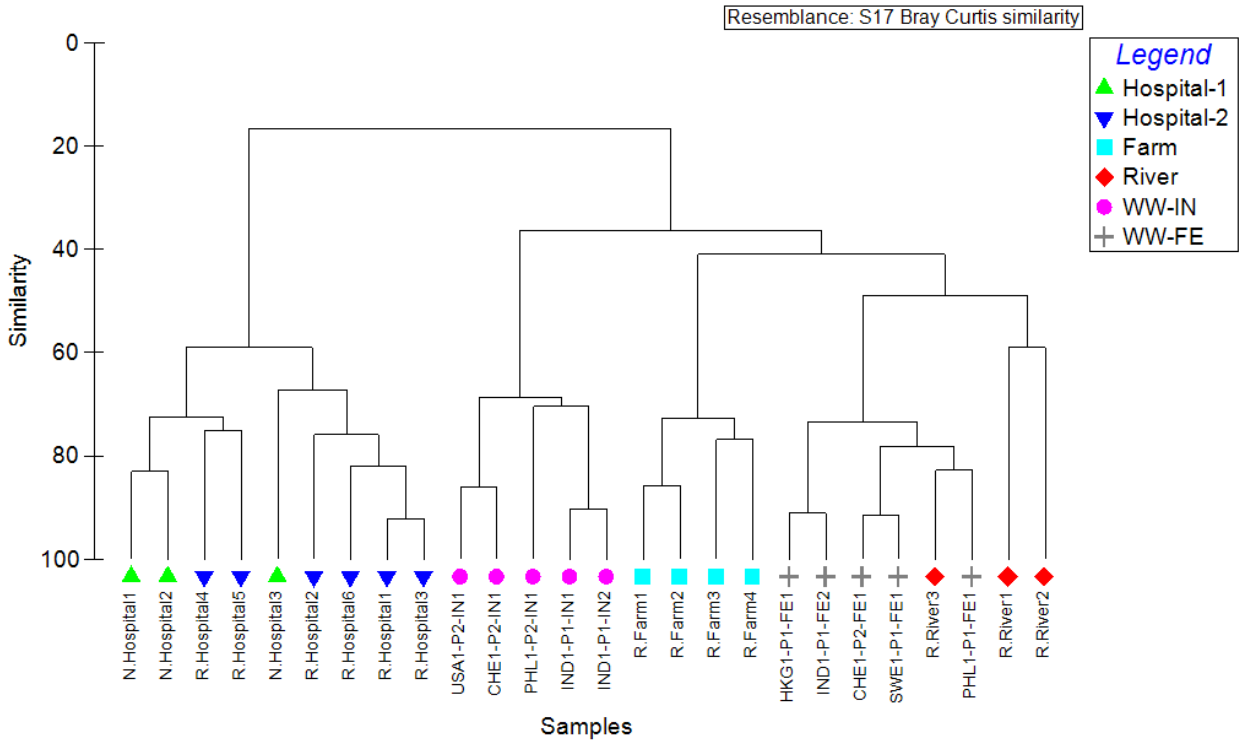


Figure 2.11: Comparison of different environmental samples with Wastewater-Influent Sample using the Core ARGs.

(a)


Resemblance: S17 Bray Curtis similarity
Legend
Hospital-1
Hospital-2
Farm
River
WW-IN
WW-FE
Similarity
0
20
40
60
80
100
N.Hospital1
N.Hospital2
R.Hospital4
R.Hospital5
N.Hospital3
R.Hospital2
R.Hospital6
R.Hospital1
R.Hospital3
USA1-P2-IN1
CHE1-P2-IN1
PHL1-P2-IN1
IND1-P1-IN1
IND1-P1-IN2
R.Farm1
R.Farm2
R.Farm3
R.Farm4
HKG1-P1-FE1
IND1-P1-FE2
CHE1-P2-FE1
SWE1-P1-FE1
R.River3
PHL1-P1-FE1
R.River1
R.River2
Samples

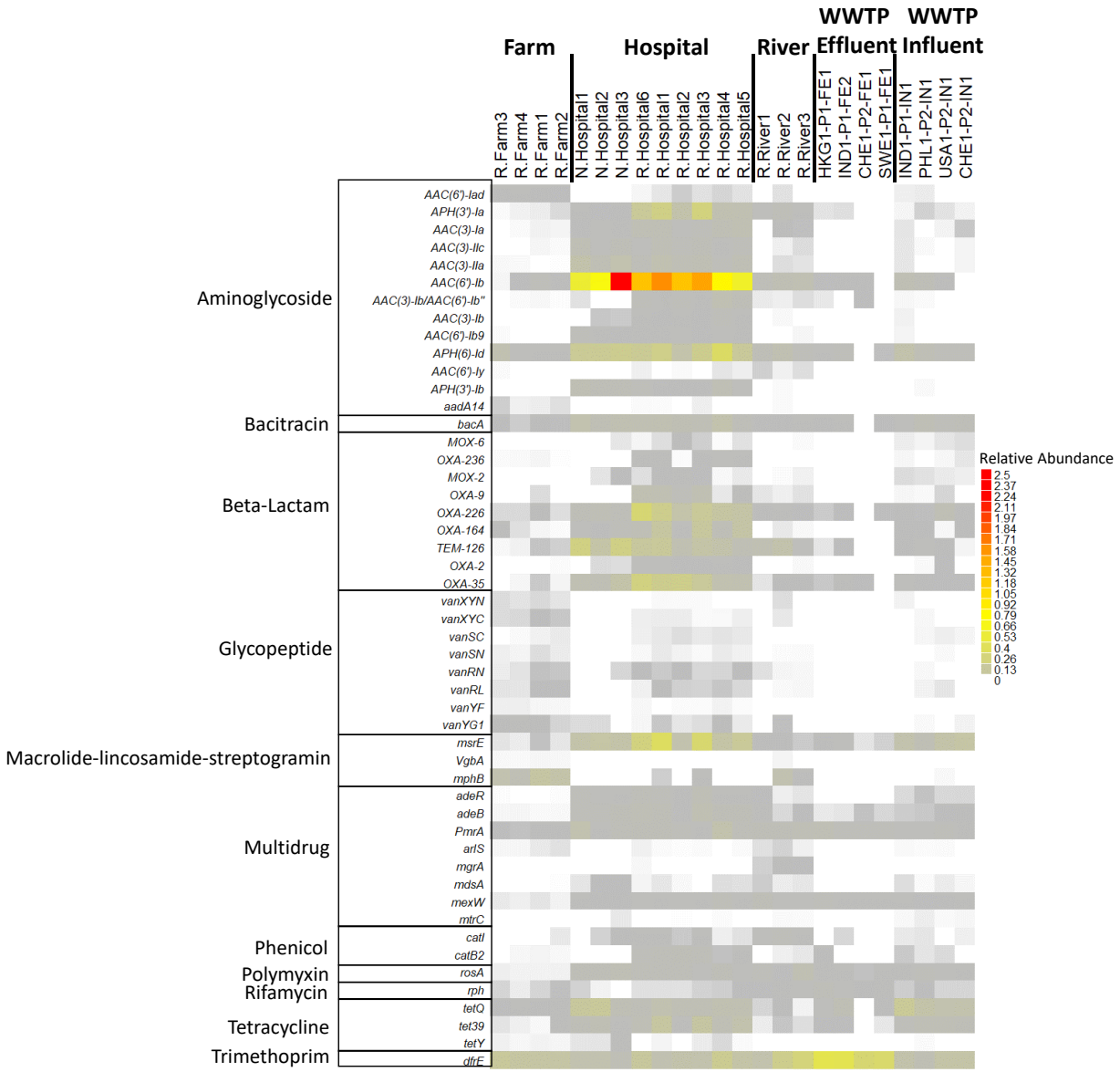
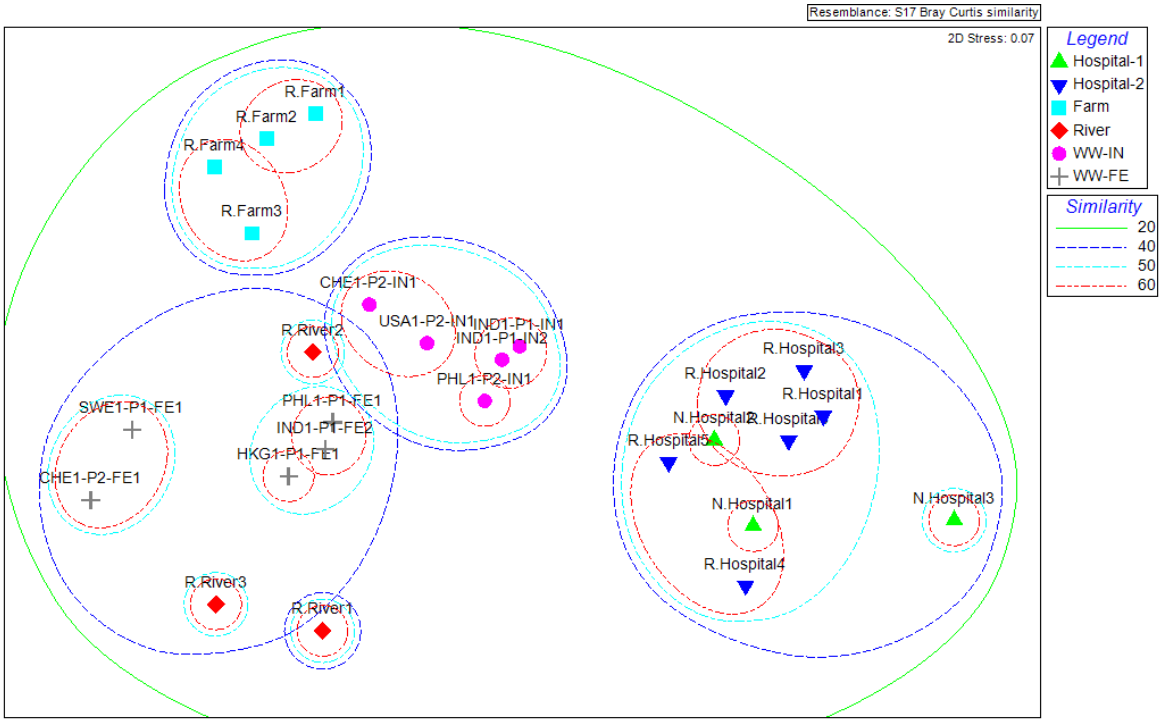
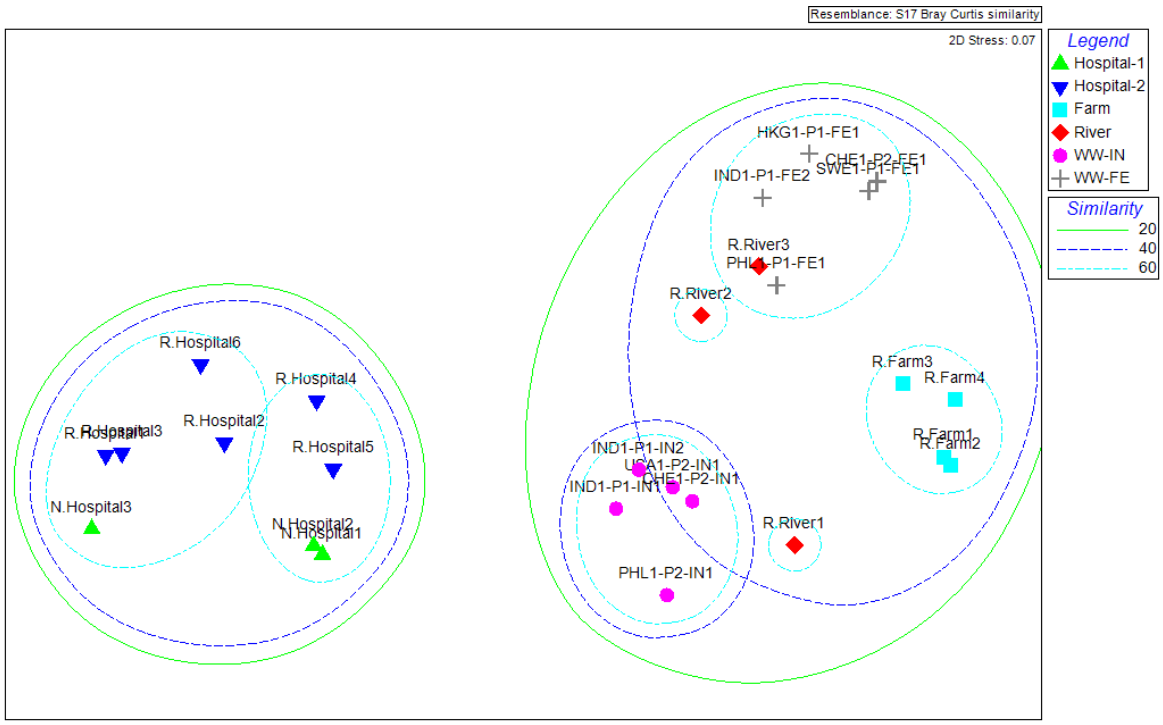


Figure 2.12: (a) Hierarchical clustering & (b) Heatmap of different aquatic environment samples based on the relative abundance of discriminatory ARGs. The list of ARGs presented in the heatmap were extracted by applying ET algorithm for the classification of different samples.



(a)



(b)

Figure 2.13: (a) NMDS plot for environmental samples using all the annotated ARGs (b) NMDS Plot for environmental samples using the discriminatory ARGs

Conclusions

In this study, a new methodology was formulated to capture the variances in ARG profile among a group of similar/dissimilar environmentally derived metagenomic data. The metagenomes were grouped based on various groups such as sampling location, untreated & treated wastewater samples and type of aquatic environment. The methodology was developed using ET algorithm which identifies discriminatory ARGs among a group of samples categorized in their group. Further, this methodology was used in identifying discriminatory ARGs in wastewater samples from different geographical location, between untreated and treated wastewater, and water/wastewater samples from different environments. It was observed that there were some ARGs which were specific to some wastewater samples from certain locations. Also, ARG variations were observed in the metagenomic data based on the geographical location encompassing a broad-range of site-specific variables. This suggests that site-specific factors (for example anthropogenic stressors) may have a certain effect in shaping the ARG profiles. Comparing untreated and treated wastewater revealed that biological treatments have a definite impact on shaping the ARG profile. While there were several ARGs which got removed after the treatment, there were some ARGs which showed an increase in abundance irrespective of location and treatment plant specific variables. Moreover, it was interesting that ET algorithm was able to capture such information. Different aquatic environments were compared and environment specific ARGs were effectively identified using ET algorithm. Also, it was interesting to see that the hospital metagenomes which were retrieved from two different studies showed high similarity. Even after the identification of discriminatory ARGs, the hospital metagenomes tend to cluster together. This raises a speculation that maybe hospital wastewater always comprises of some specific ARGs. While this still needs more validation, it was interesting to see such patterns as this sets some basic guideline for further research.

While the developed methodology demonstrates a great potential to analyze the metagenomic data, it should be noted that there are further certain factors which could play a major role in data interpretation. As the novel ARGs are continuously being added to the databases, the ARG profiles obtained from different databases could also be different. Since, the presented methodology uses the relative abundance metrics of ARGs to identify the discriminatory ARGs, it is expected that using different databases could generate a different set of discriminatory ARGs. Hence, to be consistent with the results, database version for ARG annotation was kept consistent throughout

the study and this is also advised for further studies. Also, as possible with most of the metagenomic data analysis, the prior contaminations occurred during the sample processing could also create some biases in the results. Another point to be noted is that in this study, selection of discriminatory ARGs was done empirically by the users. In further studies, optimization algorithms could be devised to help in selecting the discriminatory ARGs.

The main contribution of this methodology is that it provides an improved visualization of the ARG profile by targeting low-level variations that get overshadowed in typical similar/dissimilar type analyses. Further validation of the approach could be achieved by expanding the sample size. The specific observation validations could be made using qPCR and other bio-molecular techniques.

Future studies can be formulated around this ET algorithm for different research objectives and desired hypotheses, for example, one could frame a study to see the effectiveness of each stage of WWTPs in removing ARGs or one can study river samples from across the globe to see the geospatial variation in ARG profile in natural water bodies. The proposed methodology presents an effective way to analyze and visualize the metagenomic data and retrieve useful information from it. Also, this would aid in targeted studies in the field of antibiotic resistance. These analyses may offer a path to frame best management practices to mitigate antibiotic resistance proliferation.

References

1. Lushniak, B.D., *Antibiotic resistance: a public health crisis.* Public Health Reports, 2014. **129**(4): p. 314-316.
2. Organization, W.H., *Overcoming antimicrobial resistance.* Overcoming antimicrobial resistance., 2000.
3. Control, C.f.D. and Prevention, *Antibiotic resistance threats in the United States, 2013*. 2013: Centres for Disease Control and Prevention, US Department of Health and Human Services.
4. Martínez, J.L., *Antibiotics and antibiotic resistance genes in natural environments.* Science, 2008. **321**(5887): p. 365-367.
5. Novo, A., et al., *Antibiotic resistance, antimicrobial residues and bacterial community composition in urban wastewater.* Water research, 2013. **47**(5): p. 1875-1887.
6. Smillie, C.S., et al., *Ecology drives a global network of gene exchange connecting the human microbiome.* Nature, 2011. **480**(7376): p. 241.
7. Holmes, A.H., et al., *Understanding the mechanisms and drivers of antimicrobial resistance.* The Lancet, 2016. **387**(10014): p. 176-187.
8. Schmieder, R. and R. Edwards, *Insights into antibiotic resistance through metagenomic approaches.* Future microbiology, 2012. **7**(1): p. 73-89.
9. Gaze, W.H., et al., *Influence of humans on evolution and mobilization of environmental antibiotic resistome.* Emerging infectious diseases, 2013. **19**(7).
10. Wright, G.D., *Antibiotic resistance in the environment: a link to the clinic?* Current opinion in microbiology, 2010. **13**(5): p. 589-594.
11. Martinez, J.L., *The role of natural environments in the evolution of resistance traits in pathogenic bacteria.* Proceedings of the Royal Society of London B: Biological Sciences, 2009. **276**(1667): p. 2521-2530.
12. Alonso, A., P. Sanchez, and J.L. Martinez, *Environmental selection of antibiotic resistance genes.* Environmental microbiology, 2001. **3**(1): p. 1-9.
13. D'costa, V.M., et al., *Sampling the antibiotic resistome.* Science, 2006. **311**(5759): p. 374-377.
14. Aminov, R.I., *The role of antibiotics and antibiotic resistance in nature.* Environmental microbiology, 2009. **11**(12): p. 2970-2988.

15. McEneff, G., et al., *A year-long study of the spatial occurrence and relative distribution of pharmaceutical residues in sewage effluent, receiving marine waters and marine bivalves.* Science of the Total Environment, 2014. **476**: p. 317-326.

16. Rowe, W.P., et al., *Overexpression of antibiotic resistance genes in hospital effluents over time.* Journal of Antimicrobial Chemotherapy, 2017. **72**(6): p. 1617-1623.

17. Bruchmann, J., S. Kirchen, and T. Schwartz, *Sub-inhibitory concentrations of antibiotics and wastewater influencing biofilm formation and gene expression of multi-resistant Pseudomonas aeruginosa wastewater isolates.* Environmental Science and Pollution Research, 2013. **20**(6): p. 3539-3549.

18. Fick, J., et al., *Contamination of surface, ground, and drinking water from pharmaceutical production.* Environmental Toxicology and Chemistry, 2009. **28**(12): p. 2522-2527.

19. Kristiansson, E., et al., *Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements.* PloS one, 2011. **6**(2): p. e17038.

20. Pruden, A., M. Arabi, and H.N. Storteboom, *Correlation between upstream human activities and riverine antibiotic resistance genes.* Environmental science & technology, 2012. **46**(21): p. 11541-11549.

21. LaPara, T.M., et al., *Tertiary-treated municipal wastewater is a significant point source of antibiotic resistance genes into Duluth-Superior Harbor.* Environmental science & technology, 2011. **45**(22): p. 9543-9549.

22. Walsh, T.R., et al., *Dissemination of NDM-1 positive bacteria in the New Delhi environment and its implications for human health: an environmental point prevalence study.* The Lancet infectious diseases, 2011. **11**(5): p. 355-362.

23. Gao, P., M. Munir, and I. Xagoraraki, *Correlation of tetracycline and sulfonamide antibiotics with corresponding resistance genes and resistant bacteria in a conventional municipal wastewater treatment plant.* Science of the Total Environment, 2012. **421**: p. 173-183.

24. Goldstein, R.E.R., et al., *Methicillin-resistant Staphylococcus aureus (MRSA) detected at four US wastewater treatment plants.* Environmental health perspectives, 2012. **120**(11): p. 1551.

25. Nagulapally, S.R., et al., *Occurrence of ciprofloxacin-, trimethoprim-sulfamethoxazole-, and vancomycin-resistant bacteria in a municipal wastewater treatment plant.* Water Environment Research, 2009. **81**(1): p. 82-90.
26. Munir, M., K. Wong, and I. Xagoraraki, *Release of antibiotic resistant bacteria and genes in the effluent and biosolids of five wastewater utilities in Michigan.* Water research, 2011. **45**(2): p. 681-693.
27. Koike, S., et al., *Monitoring and source tracking of tetracycline resistance genes in lagoons and groundwater adjacent to swine production facilities over a 3-year period.* Applied and environmental microbiology, 2007. **73**(15): p. 4813-4823.
28. Smith, C.J. and A.M. Osborn, *Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology.* FEMS microbiology ecology, 2009. **67**(1): p. 6-20.
29. Lindgreen, S., K.L. Adair, and P.P. Gardner, *An evaluation of the accuracy and speed of metagenome analysis tools.* Scientific reports, 2016. **6**: p. 19233.
30. Simon, C. and R. Daniel, *Metagenomic analyses: past and future trends.* Applied and environmental microbiology, 2011. **77**(4): p. 1153-1161.
31. Li, B., et al., *Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes.* The ISME journal, 2015. **9**(11): p. 2490.
32. Ju, F., et al., *Metagenomic analysis on seasonal microbial variations of activated sludge from a full-scale wastewater treatment plant over 4 years.* Environmental microbiology reports, 2014. **6**(1): p. 80-89.
33. Rizzo, L., et al., *Urban wastewater treatment plants as hotspots for antibiotic resistant bacteria and genes spread into the environment: a review.* Science of the total environment, 2013. **447**: p. 345-360.
34. Ramette, A., *Multivariate analyses in microbial ecology.* FEMS microbiology ecology, 2007. **62**(2): p. 142-160.
35. Teeling, H. and F.O. Glöckner, *Current opportunities and challenges in microbial metagenome analysis—a bioinformatic perspective.* Briefings in bioinformatics, 2012. **13**(6): p. 728-742.

36. Geurts, P., D. Ernst, and L. Wehenkel, *Extremely randomized trees.* Machine learning, 2006. **63**(1): p. 3-42.

37. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

38. Breiman, L., *Classification and regression trees*. 2017: Routledge.

39. Breiman, L., *Some properties of splitting criteria.* Machine Learning, 1996. **24**(1): p. 41-47.

40. Ng, C., et al., *Characterization of metagenomes in urban aquatic compartments reveals high prevalence of clinically relevant antibiotic resistance genes in wastewaters.* Frontiers in microbiology, 2017. **8**.

41. Arango-Argoty, G., et al., *MetaStorm: A Public Resource for Customizable Metagenomics Annotation.* PloS one, 2016. **11**(9): p. e0162442.

42. Gu, Z., R. Eils, and M. Schlesner, *Complex heatmaps reveal patterns and correlations in multidimensional genomic data.* Bioinformatics, 2016. **32**(18): p. 2847-2849.

43. Rokach, L. and O. Maimon, *Clustering methods*, in *Data mining and knowledge discovery handbook*. 2005, Springer. p. 321-352.

Chapter 3

Conclusions

In this study, environmentally derived metagenomic data categorized in different groups such as sampling location, aquatic environments, and untreated and treated wastewaters were analyzed to understand ARG variations. A new methodology using the ET algorithm was developed to categorize environmentally derived metagenomic data based on relative ARG abundance. This methodology was applied in identifying discriminatory ARGs in environmental samples based on the above-mentioned groups. ARG variations were observed in the metagenomic data based on their location which encompasses a broad-range of site-specific variables. This suggests that site-specific factors (for example anthropogenic stressors) could augment certain changes in ARG abundance profile. Moreover, specific ARGs were identified with higher relative abundance in certain environments, but their relative abundance was lower in other samples. For example, data from hospital discharges were clustered with a high relative abundance of specific ARGs. These specific ARGs, however, were in lower abundance in other sample types like wastewater and farm discharges. Thus, the source of discharge into water body could impact the ARG abundance profile. This is possibly due to the presence of different stressors in the different sources that may preferentially select to increase the abundance of specific ARGs. Untreated and treated wastewater samples had varying ARG abundance. This could be due to the high microbial activity and presence of co-selecting factors in the biological treatment processes of WWTP. Such a nature of biological treatment can influence ARG abundance both positively and negatively. Although this is a known possibility, the ever-increasing ARG database could make analyses of these effects difficult. Despite the addition of novel ARGs to the current databases, ET can effectively handle multi-dimensional metagenomic data. It is observed that ET can holistically and simplistically study the groups and can provide a visual understanding of similarities and dissimilarities in the metagenomic data.

While the proposed methodology demonstrates great potential for the analysis of metagenomic data, it should be noted that there are certain factors which could play a major role in data interpretation. As novel ARGs are continuously being added to the databases, the ARG profiles obtained from different databases could also be different. Since, the presented methodology uses the relative abundance metrics of ARGs to identify discriminatory ARGs, it is expected that using

different databases could generate a different set of discriminatory ARGs. Hence, to be consistent with the results, database version for ARG annotation was kept consistent throughout the study and this is also advised for further studies. Also, as possible with most of the metagenomic data analyses, any sample contamination that occurred during the sample processing could also create some biases in the results. Another point to be noted is that in this study, selection of discriminatory ARGs was done empirically by the users. In future studies, optimization algorithms could be devised to help in selecting the discriminatory ARGs.

The methodology developed here can be used by other researchers as they attempt to analyze their metagenomic data and get better insights into the ARG profiles. The primary contribution of this approach is that it provides an improved visualization of the ARG profile by targeting low-level variations that are not demonstrated in typical similar/dissimilar type analyses. Further validation of the approach could be achieved by expanding the sample size and sample replicates in the study. Also, the specific observations could be validated using qPCR and other bio-molecular techniques.

Future studies can be formulated around the proposed methodology for different research objectives and desired groupings, for example, one could frame a study characterizing other class of samples, like runoff instead of wastewater or each treatment process in a WWTP to track the effectiveness of each process in removing the ARGs. This method offers an effective way to analyze and visualize the metagenomics data, and frame targeted studies with the obtained information. Ultimately, these analyses may offer a path to frame best management practices to mitigate antibiotic resistance proliferation.

Appendix: Supplementary Information for Chapter 2

Table S1: Core ARGs list using WWTP influent samples

AAC(6')-Ib	*cat*	*lmrC*	*mexN*	*sav1866*
AAC(6')-Ie-APH(2'')-Ia	*catB8*	*lmrD*	*mexQ*	*sdiA*
aadA	*CblA-1*	*lnuB*	*mexT*	*smeB*
aadA11	*ceoB*	*lnuC*	*mexW*	*smeD*
aadA12	*cepA beta-lactamase*	*lnuD*	*mphA*	*smeE*
aadA17	*CfxA2*	*lsaB*	*msrE*	*smeR*
aadA23	*CfxA3*	*lsaE*	*mtrA*	*sul1*
aadA25	*CfxA6*	*macB*	*mtrD*	*sul2*
aadA5	*cmeB*	*marA*	*nalD*	*TEM-126*
aadA6/aadA10	*CMY-114*	*MCR-1*	*novA*	*tet32*
aadA7	*cpxA*	*mdfA*	*OpmH*	*tet34*
abeM	*cpxR*	*mdsB*	*oprJ*	*tet35*
acrA	*CRP*	*mdtA*	*oprM*	*tet36*
acrB	*dfrA3*	*mdtB*	*oprN*	*tet37*
acrD	*dfrE*	*mdtC*	*OXA-119*	*tet39*
acrE	*dfrF*	*mdtD*	*OXA-12*	*tet40*
acrF	*emrA*	*mdtE*	*OXA-129*	*tet44*
acrS	*emrB*	*mdtF*	*OXA-226*	*tetA(P)*
adeB	*emrD*	*mdtG*	*OXA-256*	*tetB(P)*
adeG	*emrE*	*mdtH*	*OXA-31*	*tetC*
adeI	*emrK*	*mdtK*	*OXA-347*	*tetG*
adeJ	*emrR*	*mdtL*	*PBP1a*	*tetM*
adeK	*emrY*	*mdtM*	*PBP1b*	*tetO*
adeN	*EreA*	*mdtN*	*PBP2b*	*tetQ*

AER-1	*EreA2*	*mdtO*	*PBP2x*	*tetS*
amrB	*Erm(35)*	*mdtP*	*phoP*	*tetW*
ANT(2'')-Ia	*ErmB*	*mefA*	*phoQ*	*tetX*
ANT(6)-Ia	*ErmF*	*mefB*	*PmrA*	*tolC*
ANT(6)-Ib	*ErmG*	*mel*	*PmrB*	*TriC*
APH(3'')-Ib	*evgA*	*mexA*	*PmrC*	*vanRA*
APH(3')-IIIa	*evgS*	*mexB*	*PmrE*	*vanRC*
APH(6)-Id	*floR*	*mexC*	*PmrF*	*vanRD*
arlR	*FosA5*	*mexD*	*QnrS6*	*vanRG*
arnA	*gadX*	*mexF*	*ramA*	*vanRI*
bacA	*GES-21*	*mexI*	*rifampin phosphotransferase*	*vanSA*
baeR	*golS*	*mexJ*	*robA*	*vatB*
baeS	*H-NS*	*mexK*	*rosA*	*VEB-3*
bcrA	*LCR-1*	*mexL*	*rosB*	

Table S2: Number of distinct ARGs in WWTP influent and effluent samples

WWTP Samples	Number of ARGs annotated	No. of Distinct ARGs
CHE1-P1-IN1	382	193
CHE1-P2-IN1	415	226
HKG1-P1-IN1	430	241
HKG1-P2-IN1	424	235
HKG2-P1-IN1	343	154
HKG2-P2-IN1	420	231
IND1-P1-IN1	386	197
IND1-P2-IN1	389	200
PHL1-P1-IN1	385	196
PHL1-P2-IN1	443	254
SWE1-P1-IN1	363	174
SWE1-P2-IN1	384	195
USA1-P1-IN1	398	209
USA1-P2-IN1	409	220
PHL1-P1-FE1	271	223
PHL1-P2-FE1	299	251
CHE1-P1-FE1	245	197
SWE1-P2-FE1	197	149
HKG2-P2-FE1	232	184
USA1-P2-FE1	144	96
IND1-P1-FE1	265	217
CHE1-P2-FE1	190	142
SWE1-P1-FE1	199	151
HKG2-P1-FE1	251	203
IND1-P2-FE1	241	193
USA1-P1-FE1	180	132

S3: Sample Code for identification of discriminatory ARGs using ExtraTrees Classifier in Python:

```
import numpy as np

import matplotlib.pyplot as plt

import pylab

import pandas as pd

from sklearn.datasets import make_classification

from sklearn.ensemble import ExtraTreesClassifier

from sklearn.feature_selection import SelectFromModel

from numpy import random


X=pd.read_excel(file path)

z=X.pop('samples')

y=X.pop('Label')

print(X.shape)

df=pd.DataFrame(X)

forest=ExtraTreesClassifier(n_estimators=1000,random_state=0)

forest=forest.fit(X, y)

importances =
pd.DataFrame({'feature':X.columns,'importance':np.round(forest.feature_importances_,3)})

importances = importances.sort_values('importance',ascending=False).set_index('feature')

print(importances)

importances.plot.bar()

z=importances.index[0:50]

Y=X[z]

Y.index=y
```