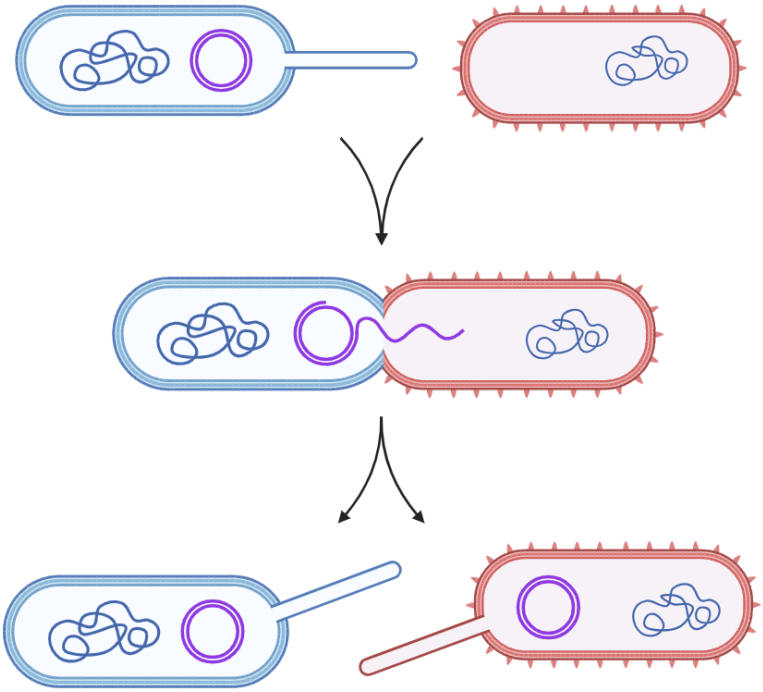
CHALMERS
UNIVERSITY OF TECHNOLOGY



Identification of New Mobile Antibiotic Resistance Genes using Metagenomic Abundances

A Correlation-Based Approach Using Zero-Inflation Models to Detect Early Spread of Antibiotic Resistance Genes

Master's thesis in Engineering Mathematics and Computational Science

IDA BERG & ELLEN KONINGEN

DEPARTMENT OF MATHEMATICAL SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Identification of New Mobile Antibiotic Resistance Genes using Metagenomic Abundances

A Correlation-Based Approach Using Zero-Inflation Models to Detect Early Spread of Antibiotic Resistance Genes

IDA BERG
ELLEN KONINGEN

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
Gothenburg, Sweden 2025

Identification of New Mobile Antibiotic Resistance Genes using Metagenomic Abundances
A Correlation-Based Approach Using Zero-Inflation Models to Detect Early Spread of Antibiotic Resistance Genes
IDA BERG
ELLEN KONINGEN

Supervisor: David Lund, Department of Mathematical Sciences
Supervisor: Helga Ólafsdóttir, Department of Mathematical Sciences
Supervisor: Erik Kristiansson, Department of Mathematical Sciences
Examiner: Erik Kristiansson, Department of Mathematical Sciences

Master's Thesis 2025
Department of Mathematical Sciences
Division of Applied Mathematics and Statistics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of gene transfer from a non-pathogenic bacterium to a pathogen. Created in BioRender.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Identification of New Mobile Antibiotic Resistance Genes using Metagenomic Abundances
A Correlation-Based Approach Using Zero-Inflation Models to Detect Early Spread of Antibiotic Resistance Genes
IDA BERG, ELLEN KONINGEN
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Antibiotic resistance is an increasing concern worldwide, as it compromises the treatments of infectious diseases and their use in many medical procedures. This project aims to identify antibiotic resistance genes (ARGs) that are spreading among bacterial populations but have not yet been detected in known pathogens.

To achieve this, ARG sequences were aligned against a large bacterial database of genomes. The resulting matches were filtered for relevance and analysed using correlation-based methods. Several strategies for handling zero counts in the data were tested and different correlation metrics were evaluated. Ultimately, a zero-inflated negative binomial model combined with Pearson's weighted correlation metrics was selected as it was the most effective approach.

The method found matches between genes and bacterial hosts, that were then compared to already known hosts found in the literature. A small group of genes was selected based on the weak taxonomic relation of their identified hosts to the known hosts. Detailed analysis of these genes showed that the method had been successful in identifying several potentially mobile ARGs that appear to be in the early stages of spreading into non-pathogenic bacterial hosts.

Keywords: Antibiotic Resistance Genes, Pearson Correlation, Zero-Inflated Models, Metagenomics, Horizontal Gene Transfer, Pathogens, Taxonomy.

Acknowledgements

We would like to express our deepest gratitude to our examiner Erik Kristiansson for making this project possible. Thank you for your enthusiasm, support and creative ideas.

We would also like to give a special thanks to our supervisors David Lund and Helga Ólafsdóttir for your support and for always taking the time to discuss solutions and answer our questions. Thank you David for helping us make sense of all data, and thank you Helga for supporting and helping us with coding issues.

Ida Berg & Ellen Koningen, Gothenburg, May 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

ARG	Antibiotic Resistant Genes
AMR	Antimicrobial Resistance
BLAST	Basic Local Alignment Search Tool
HGT	Horizontal Genes Transfer
VGT	Vertical Gene Transfer
Tax-ID	Taxonomic Identification Number
ZINB	Zero-Inflated Negative Binomial
ZIP	Zero-Inflated Poisson

Contents

List of Figures

List of Tables

1

Introduction

1.1 Background

Infectious diseases remain one of the leading causes of death globally, and the growing resistance to treatments against these microbes is becoming a threat to public health [1]. As antimicrobial resistance (AMR) increases, the effectiveness of current treatments and therapies is decreasing [1, 2]. This is making treatable diseases such as pneumonia and gonorrhoea a global health challenge again [3], and AMR has become one of the global health crises of the 21st century [4].

The implications of AMR extend far beyond treatments of infectious diseases. Routine medical procedures, such as surgeries, organ transplants, and chemotherapy, rely heavily on effective antibiotics to prevent and treat infections. Without reliable antimicrobial agents, the safety of these methods is severely compromised [2].

A bacterium that is resistant to antibiotics has a antibiotic resistant gene (ARG) that can for instance alter the target site of, or even destroy, the antibiotic [5]. The gene can occur naturally in a host, or it can be acquired through either mutation or through gene transfer from other bacteria [5]. Genes that can spread between hosts like this are called mobile genes [6].

1.2 Aim

The aim of the project is to examine the possibility of identifying mobile ARGs that have not yet spread to pathogenic bacteria. To achieve this, the goal is to develop a method that utilizes gene and organism abundances to detect ARGs that may have recently transferred to new, non-pathogenic hosts. The analysis will be based on the assumption that there is a weak relationship between the abundance of mobilized genes and their original hosts.

The project is limited to only study bacteria from samples of waste water and the human gut. Additionally, the project will not include any deeper interpretation of the raw data related to sampling location or donor identity.

2

Theory

2.1 BLAST

Basic Local Alignment Search Tool (BLAST) is a tool that can be used to find local similarities between nucleotide or protein sequences by comparing input sequences to a database [7]. Matching sequences can give an understanding of functionalities and evolutionary relationships between organisms [7]. BLAST has built-in databases, but it is also possible to produce a personalized database from FASTA files using *makeblastdb*, which is a program in the *BLAST+* package [8, 9]. The FASTA file format contains a header for each sequence in the dataset [10], and is a common format used in bioinformatics. Then *BLASTn* [8] can be used to compare the query against the database [9].

2.2 Taxonomy

Taxonomy is the science of how living organisms are named and classified [11]. It organises organisms into categories that start of broad and become increasingly specific. Taxonomy is a standardised way of naming organisms, making data easier to manage and share. It is also important when comparing organisms, studying how they are related and at what level [11]. Figure 2.1 shows the levels of biological taxonomic classification.

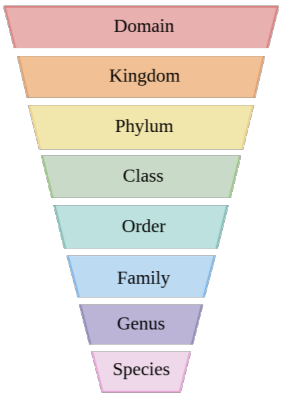


Figure 2.1: Taxonomic classification levels.

2.3 Metagenomics

Metagenomics is a study that is used to analyse metagenomes, which refers to the collection of all present genomes in an environmental sample [12]. This sample could for instance be collected from soil or a lake and the sample would then consist of a selection of microbes and their biological sequences [12]. This information can be analysed to find patterns, similarities and relationships between genes and organisms [12].

2.4 Antibiotic Resistance

One of the major global health crises of the 21st century is the spread of antibiotic resistance among pathogens. This section will introduce how this spreading happens and how resistance works.

2.4.1 Antibiotic Resistance Genes

An antibiotic can either inhibit the growth of bacteria or kill bacteria [13]. How this is done depends on the antibiotic itself, but usually they inhibit some pathway in the bacteria, making the cell not function as it should [13]. Some examples of this are inhibition of cell wall synthesis, protein synthesis and DNA synthesis. For a bacterium to become resistant, its genes have been modified to withstand the antibiotics characteristic [13]. This could for instance mean that the bacterium develops a way to modify the target where the antibiotic binds to the cell or change the efflux pumps expressions to get rid of antibiotics that invade the cell. These mechanisms are often results of ARGs [14].

Common places for ARGs to be found in bacteria is in, for instance, waste waters and the human gut [15]. Most of the antibiotics that are given to humans are used at home and thus end up in the sewage [16]. The sewage handled by the waste water treatment plants contain bacteria from many different sources, making it possible for them to interact and exchange genes. In addition to that, this kind of environment is filled with biocides, pharmaceuticals and heavy metals which causes stress to the bacteria living there [16]. This can promote horizontal gene transfer between the bacteria, and thus hasten the antibiotic resistance growth.

2.4.2 Gene Transfer in Bacteria

Bacteria are able to use both vertical gene transfer (VGT) and horizontal gene transfer (HGT) in order to spread their genetic material [17]. HGT is regarded as the major pathway in the spread of antibiotic resistance but there is also significant involvement of VGT [17].

VGT refers to the passing on of genetic material from mother cell to daughter cells during cell division [18]. HGT is a term that refers to the exchanging of genetic material between cells and it plays a significant role in the evolution of bacteria [19]. Mechanisms behind gene transfer continue to be discovered and specified, however, HGT generally occurs through three main mechanisms: transformation, conjugation and transduction [20]. Through transformation, the bacteria take up extracellular material from their surroundings and conjugation refers to the transferring of DNA through direct contact [20]. Transduction is a process where DNA is transferred from one bacterial cell to another using a bacteriophage [21]. The DNA is packaged into the bacteriophage which is sent out of the donor cell [21]. The bacteriophage is then adsorbed to the surface of the host cell and the DNA is ejected into its cytoplasm.

2.5 Data handling

This section describes the theory of data preprocessing methods such as transformation and normalisation. It also explains the principles behind correlation calculations and the strategies for handling zero-inflated data.

2.5.1 Transformation

Metagenomic data is in many cases skewed, meaning that the distribution is uneven [22]. This makes applying statistical analysis difficult since many methods assume normally distributed data. A common approach to dealing with skewed data is applying transformations, such as logarithmic or square root transformations, which help normalise the data and make it more suitable for analysis [23].

2.5.2 Normalisation

In order to be able to make a fair comparison of datasets, they must be on a common scale. If that is not the case, a normalisation can be applied to adjust the different scales to match each other [24]. One way of doing this is to divide the raw individual counts in a dataset by their scaling factor, which for instance could be the total amount of reads [25].

2.6 Correlation

To examine biological relationships between, for instance, two organisms, one can use correlation. Correlation is measured through a correlation coefficient. Normalised to be a number between -1 and 1, the coefficient describes the extent to which a set of points supports a relation between two variables [26]. A value close to 1 indicates a strong direct relationship, while a value approaching -1 indicates a strong inverse relationship [27]. Values near zero indicate little, if any, relationship between the variables.

There are several methods that can be used, all with their benefits and disadvantages. Some methods that are commonly used in bioinformatics are Pearson's correlation and Spearman's rank correlation. Both these methods measure similarity of expression between the two variables [28], for example gene expression or the occurrence of two organisms in an ecosystem. The difference in the methods lies in how they measure this similarity, and thus they will each perform better for different types of data [28].

Pearson's correlation method can only find linear relationships and is also very sensitive to outliers [28]. The Spearman correlation method is a more robust alternative to the Pearson method. It is better suited for data that has a non-normal distribution, and it finds non-linear relationships [28]. It is also not as sensitive to outliers as Pearson's correlation method [28]. However, Spearman's method converts the data

into rank data, which comes with some loss of information. For continuous data, it therefore has lower statistical accuracy [28].

2.6.1 Weighted Correlation

A correlation coefficient can also be calculated with the use of weights that are assigned to the two correlating variables x and y. The equation for the weighted and normalised Pearson coefficient, r, is [29]:

$$r = \frac{cov}{\sqrt{var_x \cdot var_y}}, \tag{2.1}$$

where the weighted covariance, cov, is calculated as

$$cov = \sum_i w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w), \tag{2.2}$$

where w_i is the weight for entry i and $\bar{x}_w$, $\bar{y}_w$ are weighted means. The weighted variances for the two variables, var_x and var_y, are calculated as

$$var_x = \sum_i w_i(x_i - \bar{x}_w))^2, \tag{2.3}$$

$$var_y = \sum_i w_i(y_i - \bar{y}_w))^2. \tag{2.4}$$

That the function is normalised means that the sum of the weights are one [29], and thus the weighted means, $\bar{x}_w$ and $\bar{y}_w$, are calculated:

$$\bar{x}_w = \sum_i w_i x_i, \tag{2.5}$$

$$\bar{y}_w = \sum_i w_i y_i. \tag{2.6}$$

For the purpose of this project, a non-normalised version needed to be used. The variables x and y can also have different weights, w_{xi} and w_{xi}, and thus the weighted means were instead calculated [30]:

$$\bar{x}_w = \frac{\sum_i w_{xi} x_i}{\sum_i w_{xi}}, \tag{2.7}$$

$$\bar{y}_w = \frac{\sum_i w_{yi} y_i}{\sum_i w_{yi}}. \tag{2.8}$$

The formula for weighted covariance was then modified to

$$cov = \sum_i w_{xi} w_{yi} (x_i - \bar{x}_w)(y_i - \bar{y}_w), \tag{2.9}$$

and the weighted variances to

$$var_x = \sum_i (w_{xi}(x_i - \bar{x}_w))^2, \tag{2.10}$$

$$var_y = \sum_i (w_{yi}(y_i - \bar{y}_w))^2. \tag{2.11}$$

Weighted correlation is done when, rather than treating all counts equally, one wants to give more importance to certain data based on, for example, their reliability. This can help minimise the risk of the analysis being skewed or in other ways not accurately represent the reality.

2.7 Zero Handling

It is not uncommon when working with sequencing data that the data will be high in sparsity, meaning it has a large number of zeros [31]. This can be an issue when modelling the data since operations such as logarithms or division will not be possible. Another issue with zeros is the fact that they can arise for several reasons, and therefore also have different meanings [31].

The zeros can either be biological or non-biological zeros [31]. A biological zero means that there is nothing to detect. For instance, an organism was not found because it was absent, which would make it a true zero. A non-biological zero can have different causes, for example low abundances where it is interpreted as zero because the count is too low to be detected. Another cause could be data handling, where zeros emerge from, for example, preparation of a sample or data processing. It is therefore of as much importance to analyse the zero counts as the non-zero counts.

There are several ways to approach the issue of zero counts [31]. If the goal is to simply remove zeros in order to be able to perform certain operations, one could use a pseudo-count in the form of a small number [32].

A way to distinguish between different types of zeros is to use zero-inflated models [31]. Two examples of such models are the Zero-Inflated Poisson (ZIP) model and the Zero-Inflated Negative Binomial (ZINB) model. They are both built on the assumption that a zero-inflated dataset can be separated into two parts, one for the actual counts and one for the excess zeros [33, 34]. The most defining difference between them is that the ZIP model works better for data that has the same mean and variance, while the ZINB model is better suited for overdispersed data, where the variance is greater than the mean [34].

3

Methods

3.1 Files and Datasets

This chapter will describe the methods used to produce the results of the project. A link to the GitHub page with the code can be found in Appendix C along with descriptions of the files and pipelines.

Eight files were given for the project:

1. **ARG sequences**: A FASTA file with ARGs.

2. **Genome**: A text file with paths to roughly 1.6 million genomes of bacteria that are possible hosts to ARGs.

3. **ARG counts**: A file containing a count matrix, describing the abundances of ARGs in about 6000 metagenomic samples from waste water and the human gut. See table 3.1 for the structure of the file.

4. **Taxonomy**: A text file with the taxonomy for all relevant bacterial genomes.

5. **Organism counts**: Three files each containing a count matrix describing the abundances of organisms in about 6000 metagenomic samples in total. Two are from waste water and one from the human gut. See table 3.1 for the structure of the file.

6. **Conversion table**: A text file with a conversion table for two kinds of identification names for genomes, Assembly ID and Contig ID.

7. **Gene status**: A text file with all given genes and if they are established or latent, meaning previously known in the context or not.

8. **Normalisation counts**: A text file with all organisms given in the project, and their summed counts over all samples.

Table 3.1: Structure for count matrices

	Samples $\rightarrow$		
Genes/Organisms	Count	$\cdots$	Count
$\downarrow$	$\vdots$	$\ddots$	$\vdots$
	Count	$\cdots$	Count

3.2 Preparation of Data

This section describes the preprocessing steps applied to the data in order to make it fit the goals of the analysis.

3.2.1 Matching of ARGs in Organism Genomes

The dataset *ARG sequences* was made into a database using *makeblastdb*. Alignment of the dataset *Genome* was then done against the database using *blastn*. When BLAST had been run, the results could be filtered to only keep the ARGs that had an alignment where at least 70% of the ARG sequence was covered, and had a minimum of 90% sequence similarity.

3.2.2 Refining Count Matrices

In *Organism counts*, the organisms were identifiable by taxonomic identification number (tax-ID). The taxonomic IDs were linked to taxonomic names through the provided file *Conversion table* and the file *Taxonomy* was used to filter for organisms that had been classified at a specific taxonomic level. This process was performed separately for both species and genus levels, resulting in two versions of the organism count matrix.

When preparing the data for the correlation of ARGs to hosts, only organisms and genes that had been matched in the BLAST results were of interest. The count matrices were therefore filtered using the BLAST results table. After this filtering, the result was a list of the organisms' proper names (genus or species) and their corresponding taxonomic IDs. The provided file *Taxonomy* was then again used to filter out irrelevant organisms that had been found in the metagenome, such as viruses, and finally a new count matrix was formed of only the counts of relevant organisms. The matrices were also filtered based on their samples, only samples that existed in both matrices were kept.

Lastly, the count matrices were normalised. This was done using the *Normalisation counts* file for the organism counts. To normalise the gene counts, a file was created with the sum of all counts for each gene across all samples, before refining the matrix. The file could then be used to normalise the gene counts as well.

3.3 Histograms

To choose a method suitable for finding correlations, the distribution of the data was examined using histograms. These were created for the full genes and organism datasets, as well as for the three sets of organisms separately. They were also made for the genes and organisms that had the highest and lowest count, organised in three different ways. Firstly, the most extreme individual count was found and whichever gene or organism this belonged to was saved. Secondly, all counts for each gene or organism were added together, and the one with the most extreme value of all sums was saved. Lastly, the average count was calculated for each gene or organism, and the one with the most extreme value was saved. These histograms were also created with transformed data, testing if logarithmic or square root transformation would give the best distribution to continue with. Some of these histograms can be seen in the results chapter, in section 4.1.1.

3.4 Zero Handling

Both the count matrix for the genes and the ones for the organisms included counts in which the gene or organism was not present in any samples, or present but to a very low degree, i.e. both biological and technical zeros. In order to carry out a well-rounded statistical analysis, all zeros should not be handled in the same way. In this project, the genes and organisms that had more than 90% zeros were removed and then five approaches were used to manage the zero abundance before calculation of correlation.

3.4.1 Simple Approaches

The first approach was to not apply any additional restrictions and to calculate Pearson's correlation coefficient directly from the filtered matrices.

If the correlating gene and organism have zero counts in the same sample, this was classified as a double zero and the counts were not included in the correlation calculation. This way of filtering ensured that if a sample had one zero count this would still be compared to non-zero counts, but if both counts were zero they would not be compared since it is not clear if they are both biological or non-biological. The second approach was to remove these double zeros in addition to filtering the data.

3.4.2 Zero-Inflation Modelling Approaches

The last three approaches were based on using zero-inflation models to separate biological zeros from non-biological zeros. To decide which model to use, a scatter plot of randomly sampled counts was created to check the relation between the mean and variance of the data. Since the data was shown to be overdispersed, see figure 4.3 in chapter 4, the ZINB model could be chosen.

All zeros in the filtered data were given a probability of being zero-inflated using the ZINB model. These values were inserted into a probability matrix with the same layout and dimensions as the count matrices. For entries that had non-zero values in the count matrix, the probabilities were set to zero. If a row contained all zeros, each of their probabilities were set to one. A zero that had a low probability of being zero-inflated meant that the organism or gene most likely was not entirely absent, but that it was discarded during screening because the abundance was very low. These zeros were thus still relevant in the analysis.

The third approach to zero handling was to set a threshold for the highest probability that would still indicate a zero-inflated zero. In this project, the threshold was set to 0.6, which was the lowest without removing too much data from the analysis. For each double zero, their probability of being false zeros were compared to the threshold. If at least one of them was classified as a false zero, they were kept for the correlation calculation.

In the fourth approach, when a double zero was encountered, a random number between zero and one was generated for each zero. The probabilities of those zeros not being zero-inflated are compared to their respective random numbers, and the pair is kept if at least one of them is below their respective threshold.

For the fifth approach, the probabilities were used to calculate the weighted correlation between the rows. A weight matrix was created based on the probability matrix, where the weights were calculated using the function,

$$w = (1 - p) \cdot 0.01, \tag{3.1}$$

where w is the weight and p is the probability. The weighted correlation coefficient was then calculated for each pair of rows using the formula described in section 2.6.1.

3.5 Correlation

After analysing the histograms and the distribution of the data, Pearson's correlation was decided to be the best alternative for finding correlations. Correlation using all five methods was done for gene-to-gene correlation, and for organism-to-organism correlation. The resulting coefficients were stored in a list with all matches. A heatmap was then created for each correlation to visualise the result. The built-in function for clustering the genes and organisms in the heatmap was used to give a better structure of the results. After analysing the heatmaps, one method was chosen to continue with when moving on to gene-to-organism correlation, which was the weighted correlation. This correlation was then done on gene-to-organism relationships on both the species level and the genus level of the organism taxonomy. Since there were no significant differences between the genus and species results, it was decided to only continue with species. A final heatmap for this correlation was

then made.

A column to describe if the gene was well known and established, or if it was a latent gene, was added to the correlation list. Another addition to the list was how many times the gene and the host in the match were present in the entire results, respectively. These facts were also added to the heatmap for easier interpretation.

3.6 Analysis of the Results

Once the final correlation list and heatmap were completed, they were analysed for matches showing unexpected patterns - specifically, cases where both the gene and host appeared only once in the results. These genes were submitted to BLAST on the NCBI website to identify known hosts, which were then compared to the ones found in this study. The hosts that only shared taxonomy from the kingdom level or up of the organism's taxonomy, were considered different enough to be kept in the analysis. These hosts were then researched together with the newly found hosts and the genes they have in common.

4

Results

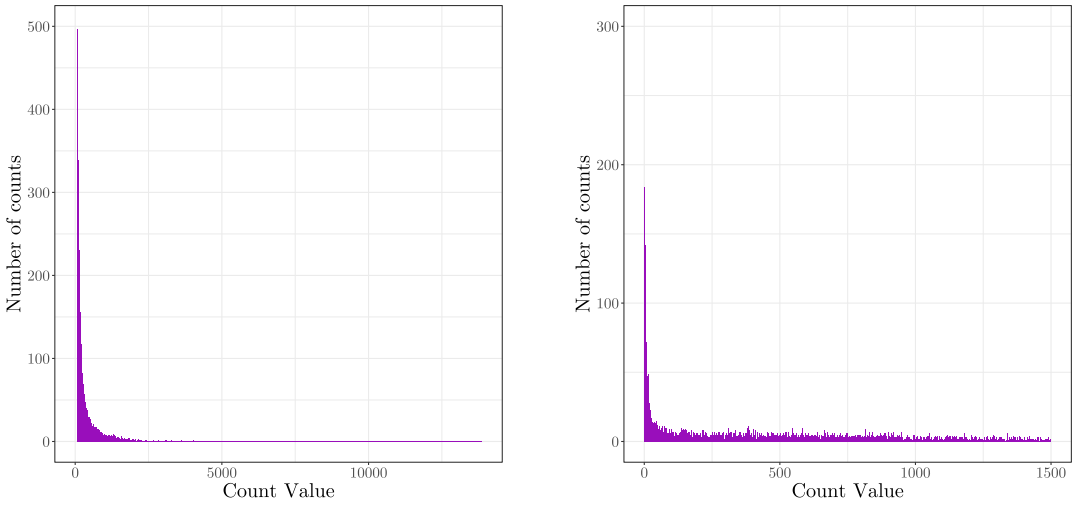
4.1 Characteristics of the Data

This section presents the characteristics of the data, which informed decisions made when selecting appropriate analysis methods.

4.1.1 Histograms

Some of the histograms that were created to examine the distributions of the data can be seen in figure 4.1, where the data studied is genes. As can be observed in the figures, the data follows an exponential distribution for both the full dataset and for individuals. This trend could be observed for both genes and organisms.

Some histograms in this section will have a limit to the y-axis to better visualise the distribution. This is because the number of zeros was occasionally so high that the y-axis increased to an extreme extent, making it difficult to interpret the figure.

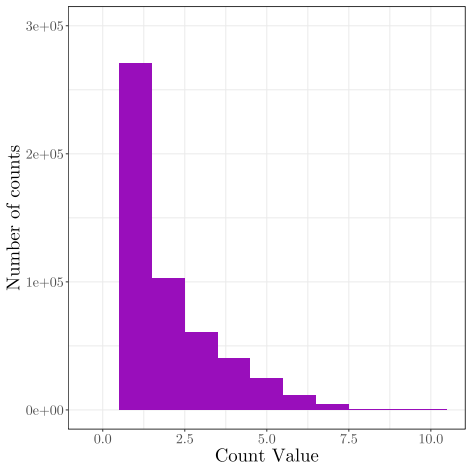


(a) All genes.

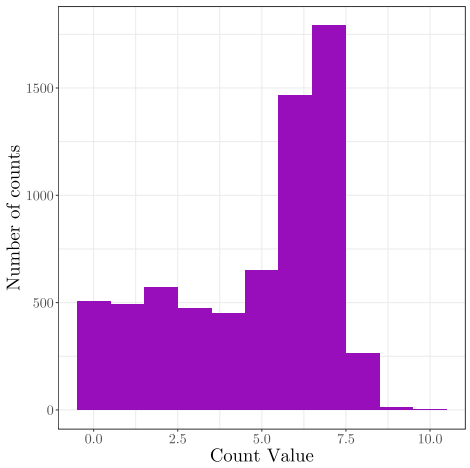
(b) The gene with highest mean count across all samples.

Figure 4.1: Histograms of gene counts with the count value on the x-axis and the y-axis shows the frequency of each value.

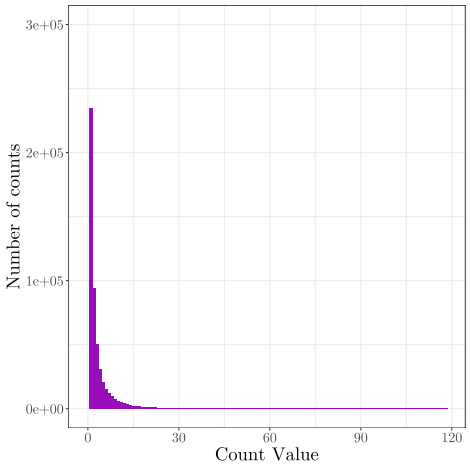
Two data transformations were evaluated to improve the normality of the distribution: logarithmic and square-root transformations. Histograms where the transformations have been applied can be seen in figure 4.2. When examining the images for all genes (subfigures 4.2a and 4.2c), it is clear that the logarithmic transformation resulted in a distribution with a much shorter tail than the square root transformation. The histograms in the subfigures 4.2b and 4.2d show even more clearly that the logarithmic transformation gives a more normal distribution. The square root transformed data also showed some improvement in distribution of the counts. However, there is still a pronounced peak at the lowest values and the tail is longer.



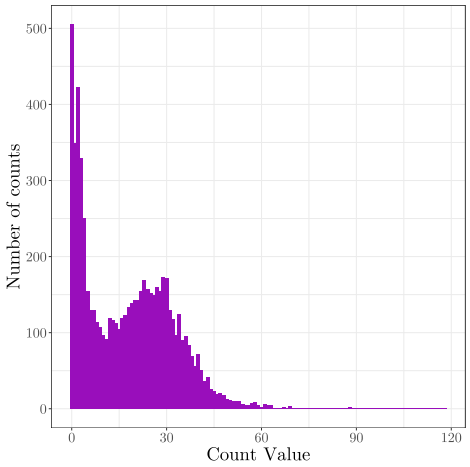
(a) Log-transformed: All genes.



(b) Log-transformed: The gene with highest mean count.



(c) Square-root-transformed: All genes.

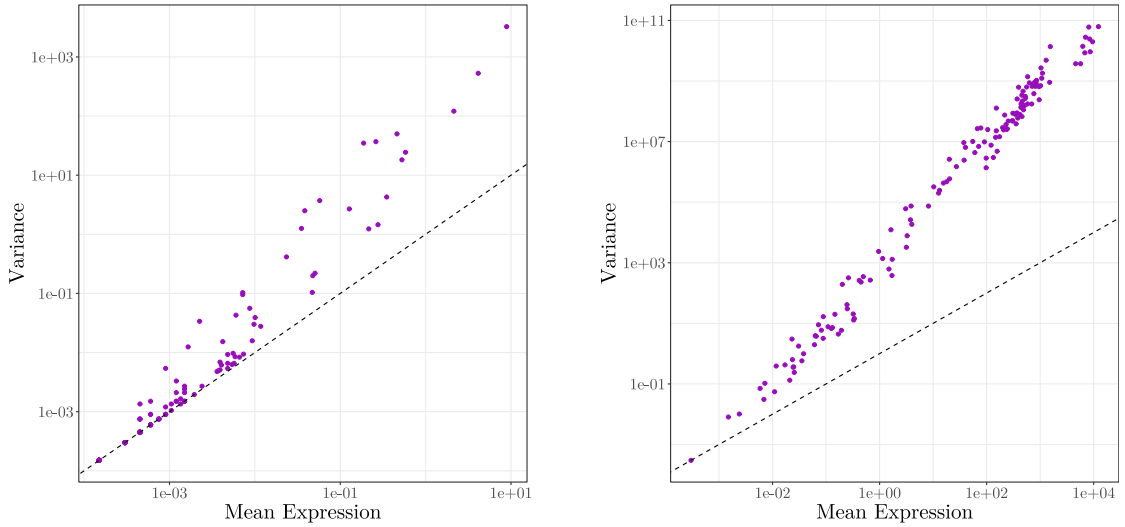


(d) Square-root-transformed: The gene with highest mean count.

Figure 4.2: Histograms of transformed gene counts. Subfigures 4.2a and 4.2b show log-transformed values, while 4.2c and 4.2d show square root-transformed values. The left figures includes counts from all genes, while the right figures have counts from the gene with the highest mean count across all samples. The count values are on the x-axis, and the y-axis shows the frequency of each value.

4.1.2 Dispersion

The scatter plots that were used to decide which zero-inflation model to use can be studied in figure 4.3. It is clear that the data is overdispersed in both datasets as the data points generally have higher variance than mean.



(a) Genes. **(b)** Organisms.

Figure 4.3: Scatter plots over the mean and variance of randomly sampled counts from the gene data.

4.2 ARG to Host Correlations

Table 4.1 shows twelve hand-selected gene-host matches that were found to be of particular interest. For each match, both gene and host were only found to have a correlation with one counterpart in the dataset. All genes were classified as latent, except *Gene2* and *Gene4*, which are a well-known and established ARGs. The "Gene Type" column in the table was derived from the final portion of each gene name. The full original names of the genes can be found in table A.1 in Appendix A.

Table 4.1: Correlation results

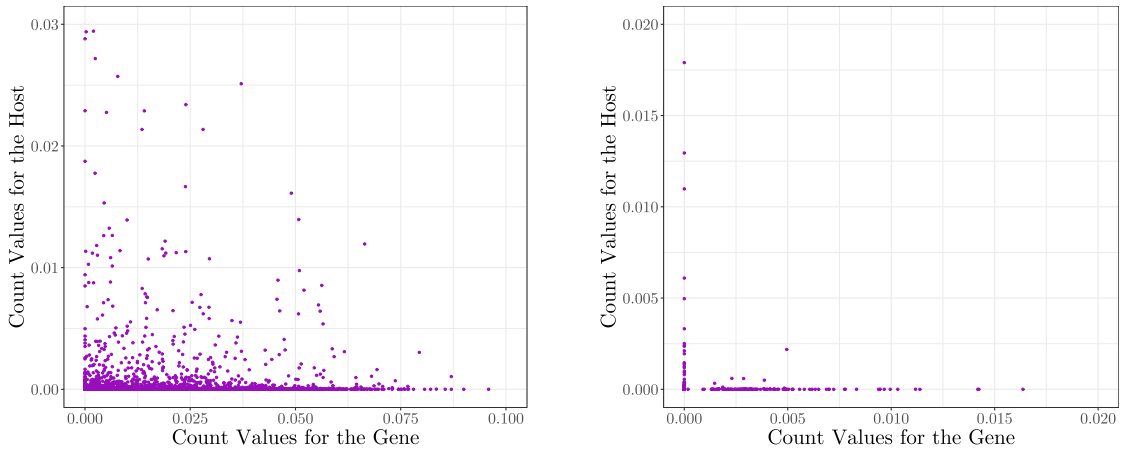
Gene	Gene Type	Organism	Correlation coefficient
Gene1	class_A	Croceicoccus marinus	-0.0096838
Gene2	tet.37.	Vibrio campbellii	-0.0091839
Gene3	class_C	Sinorhizobium americanum	-0.0067380
Gene4	GES.23	Francisella tularensis	-0.0065228
Gene5	class_A	Mycolicibacterium hassiacum	-0.0029663
Gene6	aac6p_class2	Sphingopyxis alaskensis	-0.0021758
Gene7	class_A	Paraburkholderia sprentiae	-0.0021110
Gene8	tet_rpg	Paenibacillus pabuli	-0.0016013
Gene9	aac6p_class2	Novosphingobium resinovorum	-0.0005616
Gene10	tet_rpg	Vagococcus carniphilus	-0.0005580
Gene11	erm_typeF	Planktothrix agardhii	-0.0000192
Gene12	class_A	Ornithobacterium rhinotracheale	0.0008685

None of the matches above were present in the NCBI database. The taxonomic similarity between these hosts and those found in NCBI ranged from the family level to the kingdom level, indicating that some findings were more distantly related than others. The relevant matches that were retrieved from the NCBI database for comparison can be seen in table 4.2.

Table 4.2: Matches from NCBI database

Gene	Organism
Gene1	Riemerella anatipestifer
Gene2	Prevotella histicola
Gene3	Acinetobacter johnsonii
Gene4	Pseudomonas aeruginosa
Gene5	Acidovorax carolinensis Acidovorax delafieldii Acidovorax facilis Acidovorax radicis Pseudomonas aeruginosa Pseudomonas guariconensis
Gene6	Agathobacter rectalis Lachnospiraceae bacterium Roseburia intestinalis
Gene7	Coprobacter fastidiosus
Gene8	Clostridium beijerinckii Clostridium sp. Cetobacterium somerae Clostridium saccharoperbutylacetonicum Clostridium acetobutylicum Clostridium estertheticum
Gene9	Ruthenibacterium lactatiformans
Gene10	Lachnospiraceae bacterium uncultured Candidatus Saccharibacteria bacterium
Gene11	No match found
Gene12	Bacteroides difficilis Bacteroides faecis Bacteroides faecium Bacteroides finegoldii Bacteroides ovatus Bacteroides zhangwenhongii

The match with the highest correlation was between the gene *GCA_000203195.1_ASM20319v1_FR824044.1_seq1...tet_rpg* and the host *Faecalibacillus intestinalis*. This match had the correlation coefficient 0.1811726, which is about one hundred times larger than the largest one listed in table 4.1, *Gene12*. The gene appeared in a total of 40 matches and the host appeared in three. Figure 4.4 illustrates the correlation structure for this match alongside that of *Gene1* and *Croceicoccus marinus*. These scatter plots visualise the difference of a strong and a weak correlation between matches. A stronger correlation has more evenly distributed points across the plot area, while a weak correlation shows clustering along the axes. There is a restriction on the axis lengths in the plots to remove outliers and make the image easier to interpret. Because of this, the axis ranges differ between plots, which is important to consider when analysing the plots.



(a) Strong correlation.

(b) Weak correlation.

Figure 4.4: Scatter plots that visualises the correlation between one gene on the x-axis and one organism on the y-axis. The axis-lengths differ depending on the data. **(a)** shows an example of strong correlation between the gene *GCA_000203195.1_ASM20319v1_FR824044.1_seq1...tet_rpg* and the host *Faecalibacillus intestinalis*, and **(b)** shows an example of weak correlation between the gene *Gene1* and the host *Croceicoccus marinus*.

A segment of the final heatmap of the correlations can be seen in figure 4.5. Some genes of interest and some other genes with distinct patterns or multiple host associations are included to reflect the overall structure of the full heatmap.

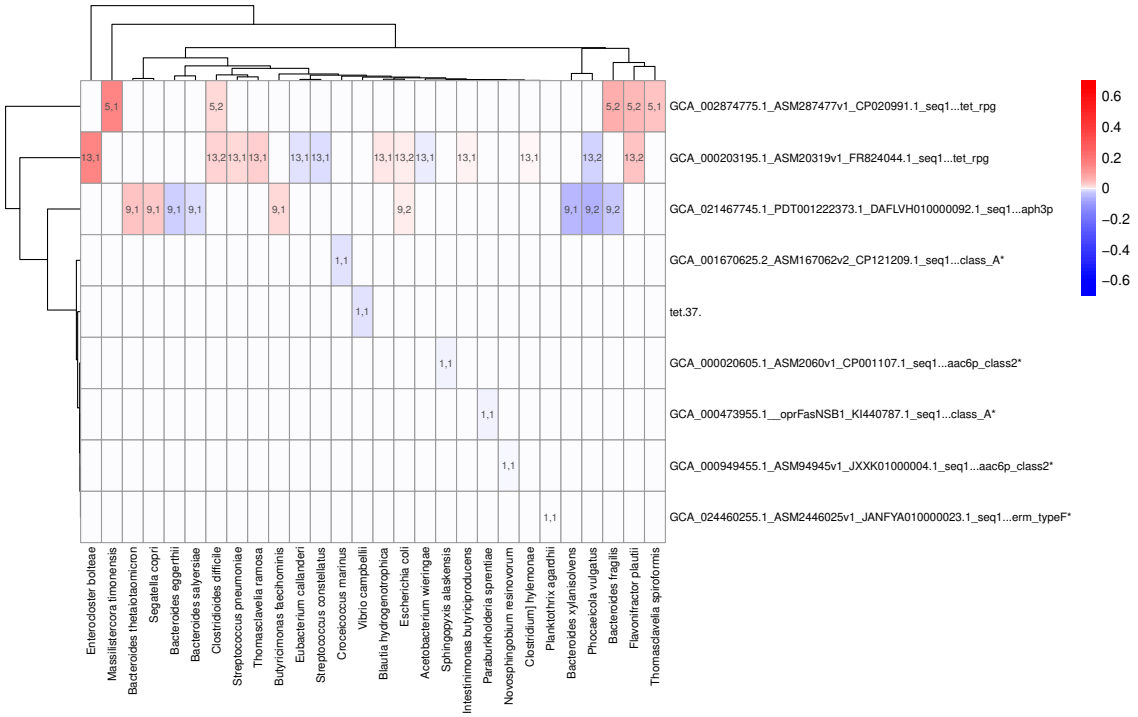


Figure 4.5: A heatmap showing selected gene–host correlation matches. Red indicates a positive correlation, and blue indicates a negative correlation. An asterisk (*) next to a gene name denotes a latent gene. The numbers in each box represent the total number of times the gene and host appear in the map, presented as "gene count, host count". Values are only shown for pairs where there has been a match between a gene and a host.

5

Discussion

5.1 Choice of Methods

All methods used in the study were chosen based on their benefits over other methods. NCBI is the most used organisation both for alignments and research, due to it being up to date and having a substantial amount of information compared to other sites.

The next decision was to choose which kind of correlation to use. After researching different correlation methods, it was found that the most common ones in the field of bioinformatics were Pearson's correlation and Spearman's rank correlation. Pearson's correlation was chosen since it has higher accuracy and assumes linearity, which aligns with this study. Pearson is, however, more sensitive to outliers and non-normal distributed data. These disadvantages were handled since outliers belonged mostly to organisms that were not kept in the study, and the data was transformed to have more of a normal distribution.

Zero handling was also a big part of the study due to the data being high in sparsity. The decision to remove genes and organisms that had more than 90% zeros in their count matrices was made because such a low level of non-zero values will likely result in unreliable correlations. This will be an issue even after the initial filtration, which is why research was done to find additional approaches to handle high sparsity, where zero-inflation models were found to be the most suitable and standard choice for this purpose. The probability of being zero-inflated gave the specific zeros a measure of their relevance in the study, which is why this information could be used to decide if the zero was worth keeping or not in the correlation calculation. The counts that were non-zero values were given a probability of zero in the probability matrix, because the probability describes the likelihood of the count being zero-inflated and thus it only applies to zero-counts. On the other hand, genes or organisms with only zeros were assigned a probability of one as all their counts can be presumed to be zero-inflated. Such all-zero genes or organisms were possible since the zero-inflation probability matrices were generated from the unfiltered matrices to retain as much information as possible for the calculations. In the end, the zeros that were kept after the analysis were likely to not be true zeros. Such zeros indicate the presence of low-abundance signals rather than a true absence of signal, and two

of them should theoretically have the same connection as two low, non-zero counts. In contrast, two zeros that are interpreted as true absences cannot be considered related since co-absence does not always imply a connection, both may simply be missing independently.

The results of the five correlation approaches were used to decide which one should be used in the final analysis. The approach to use only the 90% filtered data, with no other zero handling, gave high correlations since all zeros could contribute as much as every other count. Thus, low-abundance genes and organisms were given a higher correlation coefficient than they should have. The removing of double zeros eliminated of some of these high values, but it did not affect the strong negative correlations that appeared from when zeros were compared to other count values, which also gives an unreliable result. Based on this, it was then decided to use one of the zero-inflation models. The threshold and the probability model were very similar and allowed more zeros than was wanted in the correlation calculation. The weighted correlation could be more strict on the use of zeros while still using the information from the zero-inflation probabilities. This approach, therefore gave the most reliable results.

5.2 Interpretation of Single Matches of Genes and Hosts

The overall low correlation coefficients, with the highest at 0.1811726, is expected as the large datasets result in that genes usually exist in several hosts, meaning they cannot correlate strongly with only one host. The gene with the highest observed correlation was found in 40 matches in total, indicating its presence in at least 40 different species. Its counts will therefore be affected by all these hosts, meaning the strength of its correlation with any one of them will be diluted. However, for a single-appearance match between a gene and a host, a strong correlation is expected. This is because both the gene and the host appear only once in the results and therefore rely solely on each other. Unexpectedly, this was not the case when examining the matches that only had single appearances.

To further narrow down the number of matches selected for detailed analysis, only single-appearance matches involving genes with fewer than thirty known associations in the NCBI database were considered. These matches can all be seen in table B.1 in Appendix B and some of the genes can be seen in the heatmap in figure 4.5. In each case, the identified host differed taxonomically from the known hosts found in the NCBI database at least at the family level, meaning that the relation is weak and thus the match found in this project is unexpected. Despite the expectation of a strong correlation for these single matches, they all had a correlation coefficient very close to zero. This suggests that the selected genes do not correlate with their hosts, indicating that the gene is likely to be mobile and thus present in other hosts as well, but at levels that are too low to detect. It is therefore reasonable to say that these genes have only recently started to spread among new hosts.

5.2.1 Detailed Analysis of Single Gene–Host Matches

A detailed analysis was done on some of the matches in order to verify that the hosts found in the study had not been seen before for their respective genes.

Gene1, *Gene3*, *Gene5* and *Gene7* are all latent genes. They all belong to subgroups of genes that code for beta-lactamases, a type of enzymes [35–37]. *Gene1*, *Gene5* and *Gene7* are Class A genes [35, 36] while *Gene3* belongs to the subgroup Class C [35, 37]. There is a lot of diversity within the groups of beta-lactamases, but they all confer resistance through inactivation of the antibiotic by hydrolysing peptide bonds [38]. They are known to hydrolyse penicillins [35, 36]. The fact that the genes are latent makes them extra interesting to this study, since they have not been reported on in the literature in connection to antibiotic resistance in pathogens.

This study matched *Gene1* to *Croceicoccus marinus*. This is an aerobe and gram-negative bacterium that has been found in marine sediment [39, 40]. The NCBI database did not find the same results but rather matched the gene to *Riemerella anatipestifer*. This bacterium has the same kingdom as *Croceicoccus marinus* but is not related any more closely. *Riemerella anatipestifer* is also gram-negative but has only been found to affect ducklings, gosling, turkeys, and other fowl [41].

Gene5 was matched to *Mycolicibacterium hassiacum*, an aerobe, rod-shaped bacterium that was isolated from urine [39, 42]. It is tolerant towards high temperatures and grows very quickly, it has occasionally been associated with infections and disease in humans [43]. The NCBI database found six matches in total, all except two with the genus Acidovorax. Acidovorax is found predominantly in activated sludge environments [44, 45]. Some strains are capable of growing in anaerobic environments through the use of nitrate as a terminal electron acceptor, for example *Acidovorax delafieldii*. The other two matches found in the NCBI database instead had the genus Pseudomonas. *Pseudomonas aeruginosa* are known to be multiresistant and can colonize airways, wounds, and catheters in patients with a weakened immune system [46]. The resistance of *Pseudomonas aeruginosa* can be because of several mechanisms, including modified targets, active efflux, reduced permeability and degrading enzymes [47]. Acidovorax and Pseudomonas have the same phylum but they both only share domain with *Mycolicibacterium hassiacum* and are thus very distantly related to the host found through this study.

Gene2 and *Gene4* are both established genes. *Gene2* is of the type tet(37), which is an oxidoreductase that makes bacteria resistant to tetracycline [35, 48]. This type of resistance gene is the most widespread and dominant ARG in humans [49]. *Gene4* is another type of beta-lactamase, just as *Gene1*, *Gene5* and *Gene7* [35, 50]. The same type of gene has been found in multi-resistant isolates in hospitals [51]. However, since the aim of this study is to identify new hosts for genes that are undocumented in pathogens, these established genes are not of as much interest as the latent ones.

Gene6 and *Gene9* are latent genes and are a part of the gene family AAC(6'). They are aminoglycoside N6'-acetyltransferase enzymes which inactivate aminogly-

coside antibiotics through acetylation of the 6-amino group of the antibiotic [35, 52]. Aminoglycoside antibiotics are a group of broad-spectrum antibiotics used to treat infections caused by a large variety of gram-negative bacteria, and are commonly prescribed to children [53]. Together, all types of aminoglycoside N-acetyltransferases (AACs) comprise the largest group of genes causing this type of resistance and have been found in about 70% of gram-negative bacterial strains isolated from infected patients [54].

The match found for *Gene6* was *Sphingopyxis alaskensis*. This is an aerobe, gram-negative bacterium found in sea water [39, 55]. It has been isolated in large quantities from waters around Alaska as well as from the North Sea and the North Pacific over a long period of time. Thus, it seems to be one of the most common bacteria in these environments [56, 57]. In the NCBI database several matches were found, all sharing the same family: Lachnispiradeae. These are anaerobic, spore forming bacteria that ferment plant polysaccharides into short fatty acids and alcohols [58]. They are among the most abundant taxa in the human gut microbiota [59].

Gene8 and *Gene10* are both genes that are resistant to the antibiotic tetracyline, similarly to *Gene2*. They produce tetracycline-resistant ribosomal protection protein which binds to the 30S ribosomal subunit and prevents the antibiotic from interacting with the organism [35, 60].

The common denominator among all these findings is that none of the genes were found to have a host that was also documented in the literature, which is also stated in table B.1 in appendix B. In each example above, the hosts identified in this study differed from the known hosts in the NCBI database at least at kingdom level, meaning that there was no clear relation. However, the hosts that were found for a gene in the literature were all taxonomically similar, shared functions or had been found in similar environments. Out of all ARGs listed in the results table, table 4.1, there were only two that were matched with already established pathogens. This suggests that this project has successfully identified mobile ARGs that have not yet spread to pathogenic hosts.

The final gene, *Gene11*, is latent and codes for Erm enzymes. These are part of the RNA methyltransferase family and have the ability to methylate part of the ribosomal RNA, resulting in resistance to several antibiotics called the MLSb phenotype [35, 61]. Erm enzymes function by preventing MLSB antibiotics from binding to their active site on the 50S ribosomal subunit [62].

Gene11 was matched to the host *Planktothrix agardhii* which is a bloom-forming cyanobacteria that is often found in lakes and can produce metabolites that are toxic to humans [63]. When looking up the gene in the NCBI database, there were no matches found for potential hosts. This means that the bacteria *Planktothrix agardhii* also could be a new host for this ARG since it was not a previously known match.

5.3 Conclusions

The results of this study have shown that the use of weighted correlations with zero-inflation probabilities is a successful method for identifying potentially mobile ARGs in new, non-pathogenetic hosts.

The genes that appear once in the results correlated only weakly with organisms that were also found only once in the results. This suggests that these genes have probably spread from their original host, but only recently, and have not yet been strongly established in bacteria. The hosts identified for these genes did not coincide with the matches found for the same genes in the literature. The known host and the host found through this study were only very distantly related, while the different known hosts had strong connections to each other. This further supports the conclusion that the method developed in this study has the potential to detect emerging mobile ARGs that have not yet spread to pathogens.

For future research, it would be interesting to analyse the metadata of the samples, for instance where they are retrieved, and if there are any trends of types of ARGs between countries. To simplify the analysis, it would be beneficial to automate the comparison between the hosts found in the study and the ones found in the NCBI database. In this way, even more genes could be investigated and the study could be expanded to find even more mobile genes.

Bibliography

[1] A. Mack, D.A. Relman, and E.R. Choffnes. *Antibiotic Resistance : Implications for Global Health and Novel Intervention Strategies: Workshop Summary*. National Academies Press, 2011.

[2] *Antimicrobial resistance — who.int*. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance. [Accessed 10-02-2025]. 2023.

[3] A. Lund. *Antibiotic resistance – the silent pandemic*. https://news.ki.se/antibiotic-resistance-the-silent-pandemic. Accessed: 2025-06-02. 2022.

[4] S. Bhardwaj et al. "Antibiotics and Antibiotic Resistance - Flipsides of the Same Coin". In: *Current Pharmaceutical Design* 28.28 (2022), pp. 2312–2329. DOI: http://dx.doi.org/10.2174/1381612828666220608120238.

[5] James P. Coleman and C. Jeffrey Smith. "Microbial Resistance". In: *xPharm: The Comprehensive Pharmacology Reference*. Ed. by S.J. Enna and David B. Bylund. New York: Elsevier, 2007, pp. 1–3. ISBN: 978-0-08-055232-3. DOI: https://doi.org/10.1016/B978-008055232-3.60228-4. URL: https://www.sciencedirect.com/science/article/pii/B9780080552323602284.

[6] S. Ebmeyer, E. Kristiansson, and J. Larsson. "Unraveling the origins of mobile antibiotic resistance genes using random forest classification of large-scale genomic data". In: *Environment International* 198 (2025). DOI: https://doi.org/10.1016/j.envint.2025.109374.

[7] NCBI. *BLAST: Basic Local Alignment Search Tool*. https://blast.ncbi.nlm.nih.gov/Blast.cgi. [Accessed 03-02-2025].

[8] S.F. Altschul et al. "Basic local alignment search tool". In: *Journal of Molecular Biology* 215.3 (1990), pp. 403–410. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2.

[9] NCBI. *Quick start*. https://www.ncbi.nlm.nih.gov/books/NBK569856/. [Accessed 03-02-2025]. 2008.

[10] S. Baek and M.H. Sung. *Statistical Genomics: Methods and Protocols*. Humana, 2016.

[11] Secretariat of the Convention on Biological Diversity. *GUIDE TO THE GLOBAL TAXONOMY INITIATIVE*. https://www.cbd.int/doc/publications/cbd-ts-30.pdf. [Accessed 05-02-2025]. 2010.

[12] J.L. Bouchot et al. "Chapter 14 - Advances in Machine Learning for Processing and Comparison of Metagenomic Data". In: *Computational Systems Biology (Second Edition)*. Ed. by A. Kriete and R. Eils. Second Edition. Oxford: Aca-

demic Press, 2014, pp. 295–329. DOI: https://doi.org/10.1016/B978-0-12-405926-9.00014-9.

[13] S.B. Singh, K. Young, and L.L. Silver. "What is an "ideal" antibiotic? Discovery challenges and path forward". In: *Biochemical Pharmacology* 133 (2017). Antibiotics - Meeting the Challenges of 21st Century Health Care: Part I, pp. 63–73. DOI: https://doi.org/10.1016/j.bcp.2017.01.003.

[14] M. Galgano et al. "Acquired Bacterial Resistance to Antibiotics and Resistance Genes: From Past to Future". In: *Antibiotics* 14.3 (2025), p. 222. DOI: https://doi.org/10.3390/antibiotics14030222.

[15] M. Zhuang et al. "Distribution of antibiotic resistance genes in the environment". In: *Environmental Pollution* 285 (2021), p. 117402. DOI: https://doi.org/10.1016/j.envpol.2021.117402.

[16] A. Karkman et al. "Antibiotic-Resistance Genes in Waste Water". In: *Trends in Microbiology* 26 (2018), pp. 220–228. DOI: https://doi.org/10.1016/j.tim.2017.09.005.

[17] B. Li et al. "Dissecting horizontal and vertical gene transfer of antibiotic resistance plasmid in bacterial community using microfluidics". In: *Environment International* 131 (2019), p. 105007. DOI: https://doi.org/10.1016/j.envint.2019.105007.

[18] J.H. Bethke et al. "Vertical and horizontal gene transfer tradeoffs direct plasmid fitness — pmc.ncbi.nlm.nih.gov". In: *Mol Syst Biol.* 19 (2022). DOI: https://doi.org/10.15252/msb.202211300.

[19] B.J. Arnold, I.T. Huang, and W.P. Hanage. "Horizontal gene transfer and adaptive evolution in bacteria". In: *Nature Reviews Microbiology* 20 (2022), pp. 206–218. DOI: https://doi.org/10.1038/s41579-021-00650-4.

[20] R.K. Holmes and M.G. Jobling. *Medical Microbiology*. 4th ed. University of Texas Medical Branch at Galveston, 1996. Chap. 5.

[21] A. Thierauf, G. Perez, and S Maloy. "Generalized Transduction". In: *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions*. Ed. by M.R.J. Clokie and A.M. Kropinski. Totowa, NJ: Humana Press, 2009, pp. 267–286. ISBN: 978-1-60327-164-6. DOI: 10.1007/978-1-60327-164-6_23.

[22] Calle M.L. "Statistical Analysis of Metagenomics Data". In: *Genomics Inform* 17.1 (2019). DOI: https://doi.org/10.5808/GI.2019.17.1.e6.

[23] K. Shedden. *Introduction to Data Science | Data transformations*. https://dept.stat.lsa.umich.edu/~kshedden/introds/topics/transformations/. Accessed: 2025-03-11. 2021.

[24] D. Freedman, R. Pisani, and R. Purves. *Statistics: Fourth International Student Edition*. W.W. Norton Company, 2007.

[25] Illumina, Inc. *Depth of Sequencing Normalization*. https://support.illumina.com/help/BS_App_Targeted_RNA_OLH_15067849/Content/Source/Informatics/DiffExpressDepthSeqNorm.htm. Accessed: 2025-04-10. n.d.

[26] J.R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements, (2nd ed.)* University Science Books, 1997.

[27] National Council on Measurement in Education. *Glossary of Important Assessment and Measurement Terms*. https://web.archive.org/web/201707

22194028/http://www.ncme.org/ncme/NCME/Resource_Center/Glossary/NCME/Resource_Center/Glossary1.aspx?hkey=4bb87415-44dc-4088-9ed9-e8515326a061#anchorC. Accessed: 2025-04-10. 2017.

[28] J. Hou et al. "Distance correlation application to gene co-expression network analysis". In: *BMC Bioinformatics* 23.81 (2022). DOI: https://doi.org/10.1186/s12859-022-04601-6.

[29] J.F.P. Costa. "Weighted Correlation". In: *International Encyclopedia of Statistical Science*. Springer, 2011. DOI: https://doi.org/10.1007/978-3-642-04898-2_612.

[30] National Institute of Standards and Technology. *WEIGHTED CORRELATION*. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/weigcorr.htm. Accessed: 2025-03-13. 2020.

[31] J.D. Silverman et al. "Naught all zeros in sequence count data are the same". In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2789–2798. DOI: https://doi.org/10.1016/j.csbj.2020.09.014.

[32] A. Kaul et al. "Analysis of Microbiome Data in the Presence of Excess Zeros". In: *Frontiers in Microbiology* 7.8 (2017). DOI: https://doi.org/10.3389/fmicb.2017.02114.

[33] Times Series Reasoning. *Zero-Inflated Poisson Regression Model*. https://timeseriesreasoning.com/contents/zero-inflated-poisson-regression-model/. Accessed: 2025-03-11. n.d.

[34] UCLA Statistical Computing. *Zero-Inflated Negative Binomial | R Data Analysis Examples*. https://stats.oarc.ucla.edu/r/dae/zinb/. Accessed: 2025-03-11. n.d.

[35] B. P. Alcock et al. "CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database." In: *Nucleic acids research* 51.1 (2023), pp. 690–699. DOI: https://doi.org/10.1093/nar/gkac920.

[36] CARD. *class A beta-lactamase*. https://card.mcmaster.ca/ontology/36217. Accessed: 2025-05-12. 2017.

[37] CARD. *class C beta-lactamase*. https://card.mcmaster.ca/ontology/36215. Accessed: 2025-05-22. 2017.

[38] F. Majiduddin, I. Materon, and T. Palzkill. "Molecular analysis of beta-lactamase structure and function". In: *Int J Med Microbiol* 292.2 (2002), pp. 127–37. DOI: https://doi.org/10.1078/1438-4221-00198.

[39] I. Schober et al. "BacDive in 2025: the core database for prokaryotic strain data". In: *Nucleic Acids Research* 53.1 (2025), pp. 748–756. DOI: https://doi.org/10.1093/nar/gkae959.

[40] BacDive. *Croceicoccus marinus*. https://bacdive.dsmz.de/strain/133445. Accessed: 2025-05-12. 2024.

[41] G. Hitchener. *Riemerella anatipestifer Infection in Poultry (New Duck Disease, Infectious Serositis)*. https://www.msdvetmanual.com/poultry/riemerella-anatipestifer-infection/riemerella-anatipestifer-infection-in-poultry. Accessed: 2025-05-12. 2024.

[42] BacDive. *Mycobacterium hassiacum*. https://bacdive.dsmz.de/strain/8463. Accessed: 2025-05-12. 2024.

[43] D. Smedile et al. “Genomics Insights into Mycolicibacterium Hassiacum Causing Infection in a Cat with Pyogranulomatous Dermatitis and Panniculitis”. In: *Pathogens* 13.9 (2024), p. 785. DOI: https://doi.org/10.3390/pathogens13090785.

[44] MiDAS. *Genus: Acidovorax*. https://www.midasfieldguide.org/guide/fieldguide/genus/acidovorax. Accessed 2025-05-12. 2024.

[45] M.K.D. Dueholm et al. “MiDAS 5: Global diversity of bacteria and archaea in anaerobic digesters”. In: *Nat Commun* (2024). DOI: https://doi.org/10.1038/s41467-024-49641-y.

[46] Region Skåne. *Multiresistenta Pseudomonas aeruginosa*. https://vardgivare.skane.se/vardriktlinjer/vardhygien/smittor/multiresistenta-pseudomonas-aeruginosa/. Accessed 2025-05-12. 2025.

[47] A.M. Spagnolo, M. Sartini, and M.L. Cristina. “Pseudomonas aeruginosa in the healthcare facility setting”. In: *Reviews in Medical Microbiology* 32.3 (2021), pp. 169–175. DOI: https://doi.org/10.1097/MRM.0000000000000271.

[48] CARD. *tet(37)*. https://card.mcmaster.ca/ontology/39305. Accessed: 2025-05-12. 2017.

[49] Y. Yan et al. “Metagenomic and network analysis revealed wide distribution of antibiotic resistance genes in monkey gut microbiota”. In: *Microbiological Research* 254 (2022), p. 126895. DOI: https://doi.org/10.1016/j.micres.2021.126895.

[50] CARD. *GES-23*. https://card.mcmaster.ca/ontology/38752. Accessed: 2025-05-22. 2023.

[51] P. Bogaerts et al. “GES extended-spectrum -lactamases in Acinetobacter baumannii isolates in Belgium”. In: *Antimicrobial agents and chemotherapy* 54.11 (2010), pp. 4872–4878. DOI: https://doi.org/10.1128/AAC.00871-10.

[52] CARD. *AAC(6')*. https://card.mcmaster.ca/ontology/36484. Accessed: 2025-05-12. 2020.

[53] E. Germovsek, C. I. Barker, and M. Sharland. “What do I need to know about aminoglycoside antibiotics?” In: *Archives of disease in childhood. Education and practice edition* 102.2 (2017), pp. 89–93. DOI: https://doi.org/10.1136/archdischild-2015-309069.

[54] S. Ahmed et al. “Retention of antibiotic activity against resistant bacteria harbouring aminoglycoside-N-acetyltransferase enzyme by adjuvants: a combination of in-silico and in-vitro study”. In: *Sci Rep* 10 (2020). DOI: https://doi.org/10.1038/s41598-020-76355-0.

[55] BacDive. *Sphingopyxis alaskensis*. https://bacdive.dsmz.de/strain/14273. Accessed: 2025-05-12. 2024.

[56] JGI Genome Portal. *Sphingopyxis alaskensis RB2256*. https://genome.jgi.doe.gov/portal/sphal/sphal.home.html. Accessed 2025-05-12. 2024.

[57] H. Nordberg et al. “The genome portal of the Department of Energy Joint Genome Institute: 2014 updates”. In: *Nucleic Acids Res.* 42.1 (2014), pp. 26–31.

[58] M. Boutard et al. “Functional Diversity of Carbohydrate-Active Enzymes Enabling a Bacterium to Ferment Plant Biomass”. In: *PLOS Genetics* 10.11 (2014). DOI: https://doi.org/10.1371/journal.pgen.1004773.

[59] P. Kanki and D.J. Grimes. *Infectious diseases selected entries from the Encyclopedia of sustainability science and technology*. Springer, 2013.

[60] CARD. *tetracycline-resistant ribosomal protection protein*. https://card.mcmaster.ca/ontology/38752. Accessed: 2025-05-22. 2019.

[61] CARD. *Erm 23S ribosomal RNA methyltransferase*. https://card.mcmaster.ca/ontology/36699. Accessed: 2025-05-12. 2019.

[62] L.M. Powell, S.J. Choi, and M.E. et al. Grund. "Regulation of erm(T) MLSB phenotype expression in the emergent emm92 type group A Streptococcus." In: *npj Antimicrob Resist* 2.44 (2024). DOI: https://doi.org/10.1038/s44259-024-00062-3.

[63] R. Kurmayerm, L. Deng, and E. Entfellner. "Role of toxic and bioactive secondary metabolites in colonization and bloom formation by filamentous cyanobacteria Planktothrix". In: *Harmful Algae* 54 (2016), pp. 69–86. DOI: https://doi.org/10.1016/j.hal.2016.01.004.

A

Appendix 1: Translation table

Below is the translation table, table A.1, for gene names that are used in the report.

Table A.1: Gene name translations

Report name	Original name
Gene1	GCA_001670625.2_ASM167062v2_CP121209.1_seq1...class_A
Gene2	tet.37.
Gene3	GCA_001707755.1_ASM170775v1_MBDL01000008.1_seq1...class_C
Gene4	GES.23
Gene5	GCA_000302535.1_ASM30253v1_CP003872.1_seq1...class_A
Gene6	GCA_000020605.1_ASM2060v1_CP001107.1_seq1...aac6p_class2
Gene7	GCA_000473955.1_oprFasNSB1_KI440787.1_seq1...class_A
Gene8	GCA_002940805.1_ASM294080v1_LRDH01000095.1_seq1...tet_rpg
Gene9	GCA_000949455.1_ASM94945v1_JXXK01000004.1_seq1...aac6p_class2
Gene10	GCA_001717125.1_ASM171712v1_MCGI01000004.1_seq1...tet_rpg
Gene11	GCA_024460255.1_ASM2446025v1_JANFYA010000023.1_seq1...erm_typeF
Gene12	GCA_000269545.1_PB_Bact_thet_CL09T03C10_V1_AKBZ01000002.1_seq1...class_A

B

Appendix 2: Gene to Host Matches

Table B.1 contains the gene to host matches that were found to be interesting in this study, along with information on the hosts that were found for the gene in the NCBI database. A red coloured gene indicates that it is latent. Only matches where less than thirty hosts were found in NCBI are listed. Only hosts that had at least 50% query coverage are considered relevant findings.

Table B.1: Gene to host matches

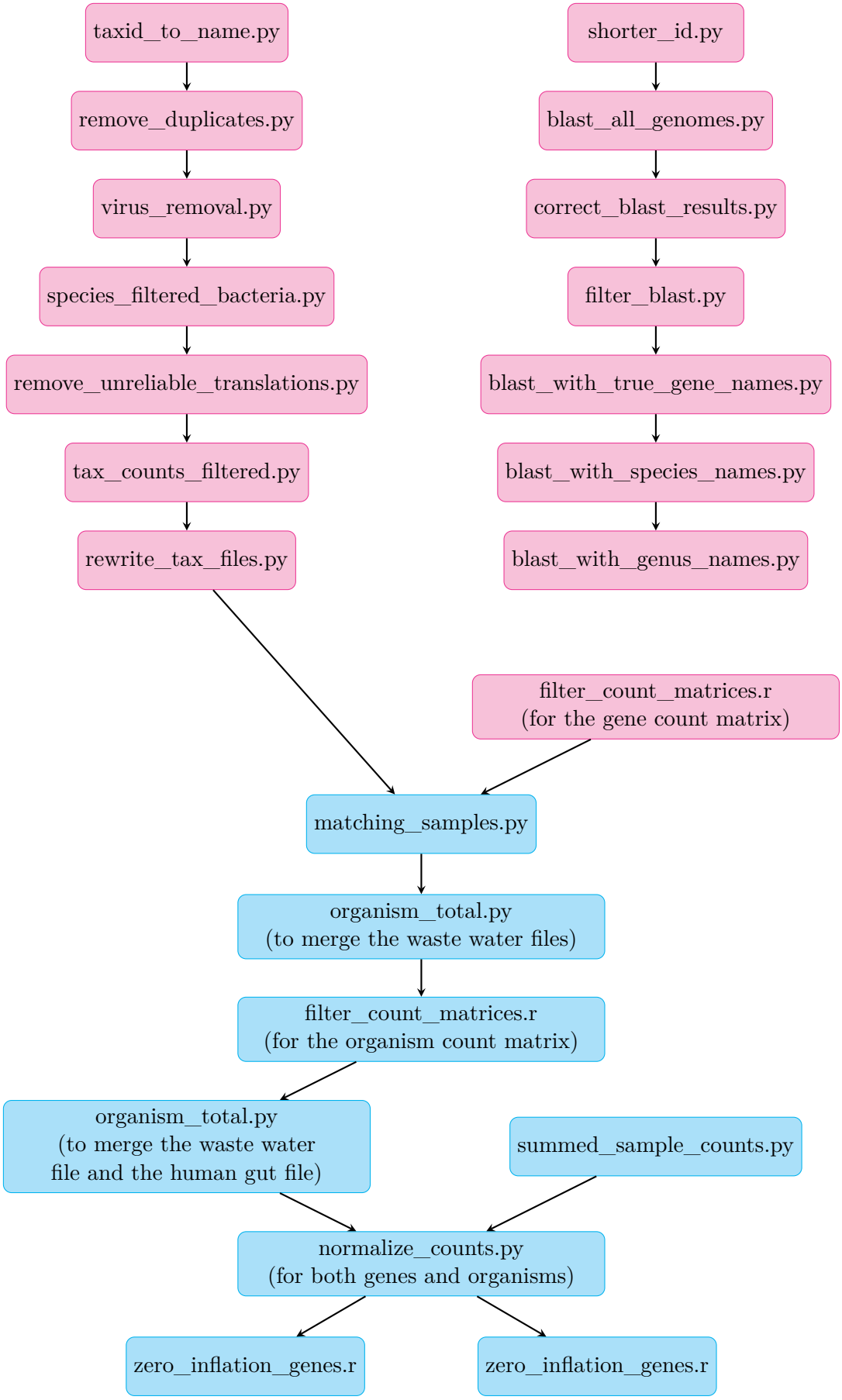
Gene	Host	Data about hosts found in the NCBI Database
Gene1	Croceicoccus marinus	One other host found with similarity to the host found in the study at kingdom level.
Gene2	Vibrio campbellii	One other host found with similarity to the host found in the study at kingdom level.
Gene3	Sinorhizobium americanum	One other host found with similarity to the host found in the study at phylum level.
Gene4	Francisella tularensis	One other host found with similarity to the host found in the study at class level.
Gene5	Mycolicibacterium hassiacum	Six other hosts found, all with similarity to the host found in the study at domain level. The hosts are similar to each other but divided between two genera.
Gene6	Sphingopyxis alaskensis	Three other hosts found, all with similarity to the host found in the study at domain level. All these hosts have the same family.
Gene7	Paraburkholderia sprentiae	One other host found with similarity to the host found in the study at kingdom level.
Gene8	Paenibacillus pabuli	Six other hosts found, all with a similarity to the host found in the study at phylum level. Five of them are similar down to genus level, and one who differs from the other at domain level.
Gene9	Novosphingobium resinovorum	One other host found with similarity to the host found in the study at domain level.
Gene10	Vagococcus carniphilus	Two other hosts found with the closest similarity to the host found in the study at phylum level. One is uncultured but they both have the same domain.
Gene11	Planktothrix agardhii	No other match found.
Gene12	Ornithobacterium rhinotracheale	Six other matches found, all with a similarity to the host found in the study at phylum level. They all have the same genus.

C

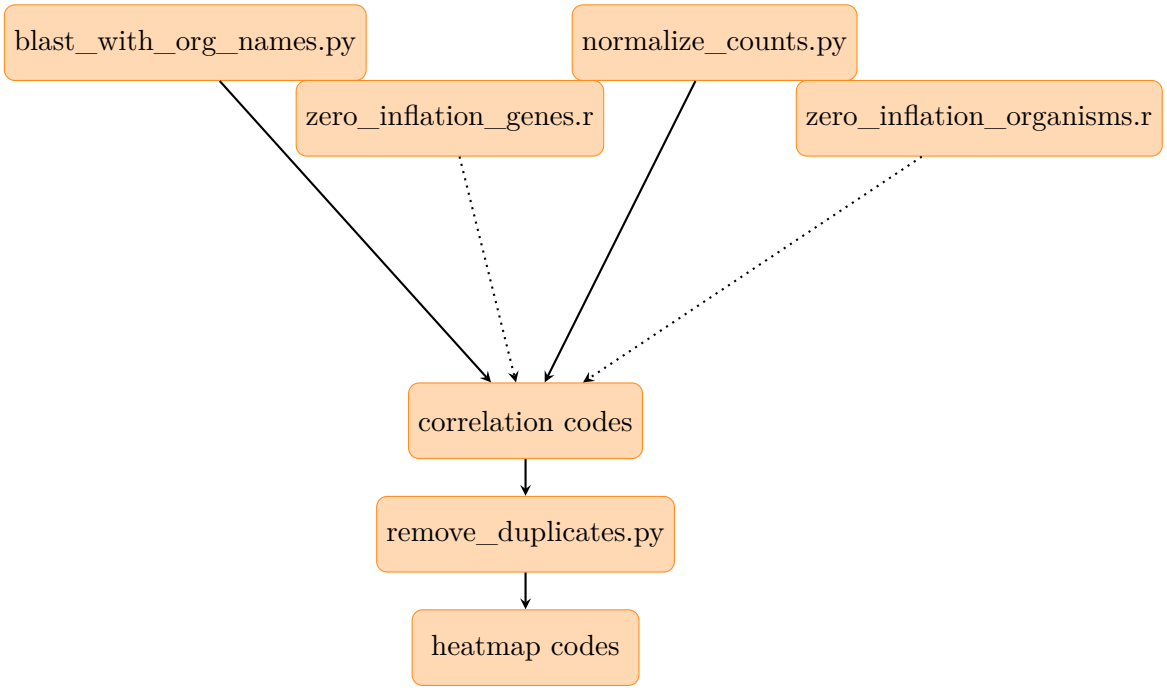
Appendix 3: Code pipelines

This chapter will describe the execution order of the scripts that were developed for the project, using flowcharts and lists. The code is available on the project's GitHub page at this link.

In the flowcharts, nodes that have the same colour indicate that the scripts can be run parallelly. Some scripts can be used for several input files. An example is the script used for filtering count matrices, which can be used for both the gene count matrix and the organism count matrix. The first flowchart illustrates the pipelines for handling the count matrices of organisms, the handling of BLAST output, and generating zero-inflation matrices.


taxid_to_name.py
remove_duplicates.py
virus_removal.py
species_filtered_bacteria.py
remove_unreliable_translations.py
tax_counts_filtered.py
rewrite_tax_files.py
shorter_id.py
blast_all_genomes.py
correct_blast_results.py
filter_blast.py
blast_with_true_gene_names.py
blast_with_species_names.py
blast_with_genus_names.py
filter_count_matrices.r
(for the gene count matrix)
matching_samples.py
organism_total.py
(to merge the waste water files)
filter_count_matrices.r
(for the organism count matrix)
organism_total.py
(to merge the waste water
file and the human gut file)
summed_sample_counts.py
normalize_counts.py
(for both genes and organisms)
zero_inflation_genes.r
zero_inflation_genes.r

The second scheme is a continuation of the first scheme and describes the steps required to perform correlation analysis and generate heatmaps. The dotted lines indicate that the previous files are only necessary for some of the scripts. For example, the script zero_inflation_genes.py will not be needed when performing correlations on only organisms, or when a correlation analysis does not take zero-inflations into account.



The scripts for the different correlation methods and heatmaps are:

- For genes:
  - correlation_genes_filter.r
  - correlation_genes_double_zeros.r
  - correlation_genes_zinb_probabilities.r
  - correlation_genes_zinb_threshold.r
  - correlation_genes_zinb_weighted.r
  - heatmap_genes.r
- For organisms:
  - correlation_organisms_filter_separate.r
  - correlation_orgs_double_zeros.r

  - correlation_orgs_zinb_probabilities.r
  - correlation_orgs_zinb_threshold.r
  - correlation_org_zinb_weighted.r
  - heatmap_orgsanisms.r
- For both genes and organisms together:
  - correlation_both_filter.r
  - correlation_both_zinb_weighted.r
  - heatmap_both.r
  - heatmap_latent.py
  - heatmap_numbers.py
- To sort the files according to the value of the correlation coefficients:
  - sort_correlations.r

Additional scripts were developed to study the nature of the data and evaluate the methods. These are not necessarily run in a specific order relative to the other scripts and can be run independently of the main pipeline. They are:

- overdispersion.r
- highest_lowest_individual.py
- highest_lowest_mean.py
- highest_lowest_sum.py
- histogram_multiple_files.r
- histogram_one_file.r
- outliers_counts.py
- outliers_orgs.py
- histogram_outlier_proportions.r
- scatterplot_gene_vs_org.r

D

Appendix 4: Contributions

For the most part, we have worked together to gather information and write the report. We both have corrected and rewritten text in all sections. The table below shows the main author for each part of the report.

Contributions

§	Title	Author
1	**Introduction**	
1.1	Background	Ida
1.2	Aim	Ellen
2	**Theory**	
2.1	BLAST	Ida
2.2	Taxonomy	Ellen
2.3	Metagenomics	Ida
2.4	Antibiotic Resistance	
2.4.2	Antibiotic Resistance Genes	Ida
2.4.1	Gene Transfer in Bacteria	Ellen
2.5	Data Handling	
2.5.2	Transformation	Ellen
2.5.3	Normalisation	Ellen
2.6	Correlation	Ellen
2.6.1	Weighted Correlation	Ellen
2.7	Zero Handling	Ida
3	**Methods**	
3.1	Files and Datasets	Ida
3.2	Preparation of Data	
3.2.1	Matching of ARGs in Organism Genomes	Ida
3.2.2	Refining Count Matrices	Ellen
3.3	Histograms	Ida
3.4	Zero Handling	Ellen
3.4.1	Filtering	Ellen
3.4.2	Zero-Inflation Models	Ellen
3.5	Correlation	Ida
4	**Results**	
4.1	Characteristics of the Data	
4.1.1	Histograms	Ida
4.1.2	Dispersion	Ida
4.2	ARG to Host Correlations	Ida
5	**Discussion**	
5.1	Choice of Methods	Ida
5.2	Interpretation of Single Matches of Genes and Hosts	Ida
5.2.1	Detailed Analysis of Single Gene-Host Matches	Ellen
5.3	Conclusions	Ellen
A	**Appendix 1: Translation table**	Ida
B	**Appendix 2: Gene to Host Matches**	Ida
C	**Appendix 3: Code pipelines**	Ida
D	**Appendix 4: Contributions**	Ida

The creation of code has been done together by discussing approaches and learning the language better together. The scripts have also been rewritten several times by both authors. It is therefore not possible to assign a specific author to different scripts.

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY