CHALMERS
UNIVERSITY OF TECHNOLOGY

The Genetic Compatibility between Antibiotic Resistance Genes and Bacterial Hosts

Evaluated by measuring differences in k-mer distributions and comparing gene codon usage with tRNA availability

Master's thesis in Engineering Mathematics and Computational Science

ENYA ARVIDSSON & JOHANNA LUNDSTRÖM

DEPARTMENT OF MATHEMATICAL SCIENCES

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

Master's thesis 2025

The Genetic Compatibility between Antibiotic Resistance Genes and Bacterial Hosts

Evaluated by measuring differences in k-mer distributions and comparing gene codon usage with tRNA availability

ENYA ARVIDSSON & JOHANNA LUNDSTRÖM

CHALMERS
UNIVERSITY OF TECHNOLOGY

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

The Genetic Compatibility between Antibiotic Resistance Genes and Bacterial Hosts
Evaluated by measuring differences in k-mer distributions and comparing gene codon usage with tRNA availability
ENYA ARVIDSSON & JOHANNA LUNDSTRÖM

Supervisor: Erik Kristiansson, Full Professor in Applied Mathematics and Statistics
Examiner: Erik Kristiansson, Full Professor in Applied Mathematics and Statistics

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

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

The Genetic Compatibility between Antibiotic Resistance Genes and Bacterial Hosts
Evaluated by measuring differences in k-mer distributions and comparing gene codon usage with tRNA availability

ENYA ARVIDSSON & JOHANNA LUNDSTRÖM
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Antibiotic resistance is a growing global health concern, driven by bacteria exchanging antibiotic resistance genes (ARGs) through horizontal gene transfer. The factors influencing the spread of ARGs across bacteria are not entirely understood, though genetic compatibility has been proposed as a contributing factor. This project aimed to explore genetic compatibility between ARGs and bacterial genomes by creating two metrics. The first metric, the 5mer score, was created by looking at nucleotide composition, specifically comparing 5-mer distributions using Euclidean distance. The second metric, the tRNA score, was created by comparing codon usage in the genes with the tRNA availability in the bacterial hosts. The results showed that both scores capture certain aspects of genetic compatibility and that higher compatibility correlates with increased likelihood of horizontal gene transfer. Although some transfers have occurred with poor scores, this suggests that transfers can still take place despite lower genetic compatibility, for example under evolutionary pressure. Gene length was identified as an important factor to take into account when working with 5-mers. Further studies include implementing the 5mer score in machine learning to determine the spread of ARGs, and refining the tRNA score due to its limitations, including how the scores were determined.

Keywords: genetic compatibility, antibiotic resistance genes, codon usage, nucleotide composition, tRNA availability, horizontal gene transfer

Acknowledgements

We would like to thank our examiner and supervisor Erik Kristiansson for his helpful guidance, relevant input, and support during this thesis work. Additionally, we want to thank Sophia Axillus, David Lund, and Laleh Varghaei, for their valuable insights throughout the project. Finally, we are grateful for our opponent Lovisa Rosin, whose thoughtful critique and constructive feedback helped us improve the quality of our thesis.

Enya Arvidsson & Johanna Lundström, Gothenburg, June 2025

Contents

1

Introduction

Antibiotics are essential in modern medicine for treating bacterial infections. The use of antibiotics has given rise to antibiotic resistance, which have led to pathogenic bacteria surviving antibiotic treatment. As a result, antibiotic resistance has become a global health concern. The rise of antibiotic resistance is largely driven by bacteria's ability to exchange genetic material. Mobile antibiotic resistance genes (ARGs) can therefore be transferred from harmless bacteria into pathogens, accelerating the dissemination of antibiotic resistance.

The factors that influence a bacterium's ability to successfully acquire and maintain a specific gene are still not fully understood. Some studies suggest that genetic compatibility between genes and bacterial hosts may play a significant role in this process [1][2]. Elements contributing to this compatibility could include the gene's DNA sequence, its encoded resistance mechanism, and transcriptional regulation in the cell. For instance, the resistance mechanism must function effectively within the host to confer selective advantage, while the sequence composition and transcriptional regulation may impact expression levels and associated fitness cost. Although, exactly what influences the genetic compatibility and its significance for the spread of mobile genes remains unclear.

1.1 Aims

The aim of this project is to explore genetic compatibility between ARGs and their bacterial hosts. This will be achieved by developing two metrics: one based on nucleotide composition, and another that assesses the gene's codon usage in relation to the host cell's transfer RNA (tRNA) availability. These metrics will be used to evaluate if genetic compatibility is of importance for the dissemination of ARGs. In addition, this project will examine differences in compatibility across bacterial phyla and between mobile and non-mobile ARGs. The two metrics will also be compared to each other, with a focus on identifying the underlying factors they capture.

This is a computational project, with data sourced from the Comprehensive Antibiotic Resistance Database (CARD) and National Center for Biotechnology Information (NCBI).

2

Theory

This chapter provides an overview on how antibiotic resistance genes (ARGs) are transferred between bacteria, through horizontal gene transfer. It presents information about genetic compatibility between ARGs and bacterial genomes, and it addresses the role of transfer RNA (tRNA) in translation and its availability in bacteria.

2.1 Antibiotic resistance

Antibiotics are important in treatment of many fatal diseases, and resistance to these antibiotics arises when bacteria evolve to survive the effects of them [3]. As a result, infections can become difficult or impossible to treat [4]. Antibiotic resistance genes use different mechanisms to survive the use of antibiotics [5]. The resistance mechanisms can be divided into four main groups: drug inactivation, limitation of drug uptake, active drug efflux, and modification of drug target [5].

These ways of resisting antibiotics help explain how resistance can spread, and the use of antibiotics further increases this resistance [6], posing a significant risk to global public health. Antibiotic resistance can arise in different ways. One way is through changes in genetic material caused by mutations, but it can also occur when bacteria transfer resistance genes between each other [3].

2.2 Horizontal gene transfer

The evolution and diversity of bacteria are largely driven by their ability to transfer genes between them, a process called horizontal gene transfer [7]. This process can occur through three main mechanisms: transformation, transduction, and conjugation [8]. Transformation involves a bacterium taking up free DNA from its environment. Transduction occurs when a virus transfers DNA from one bacterium to another, and conjugation involves the direct transfer of plasmid DNA through physical contact between bacteria.

Horizontal gene transfer is important for the dissemination of antibiotic resistance, as mobile ARGs are exchanged between bacteria. The most common method for ARGs being transferred between bacteria is conjugation [9]. Before these ARGs

reach pathogens, they are likely transferred between diverse bacteria. In a study examining ARG transfers between evolutionary distant bacteria, specifically across different phyla, a substantial number of inter-phyla transfers were identified [10]. The highest number of transfers was observed for aminoglycoside resistance gene AAC(3), while transfer frequency was generally lower for beta-lactamases. The study also found that conjugative systems, direct cell to cell contact, are rarely shared between bacterial phyla. Instead, the dissemination of ARGs between distant bacteria is most likely due to other mechanisms [10].

2.3 Genetic compatibility

Genetic compatibility between genes and bacterial genomes influence how likely it is that a gene will be successfully taken up by a bacterial host. One way to measure this compatibility is by analysing codon usage. This was investigated in a study where researchers found that codon usage compatibility increases the probability that a gene is taken up by a bacterial host through horizontal gene transfer [1]. Codon usage (CU) was categorized into three groups; poor, typical and rich CU, depending on how many of the host's abundant codons were present in the gene. The results showed that most genes involved in horizontal gene transfer have typical CU, some have rich CU, and very few have poor CU. This suggests that having sufficient CU compatibility is enough for a gene to be taken up, while poor CU makes successful uptake unlikely.

While codon usage provides one aspect of genetic compatibility, another approach involves analysing sequences of k nucleotides in length, known as k-mers. A recent study investigated this by comparing differences in 5-mer distributions of gene-genome and genome-genome pairs, using random forest models to predict the spread of ARGs [2]. The results showed that genetic incompatibility affected the performance of the models. For ARGs encoding tetracycline efflux pumps, incompatibility had a large influence on the performance, whereas class B beta-lactamases showed a lower influence. The study also found that environmental co-occurrence also plays a role in the spread of ARGs, it facilitates horizontal transfer of ARGs, while high genetic incompatibility reduces the likelihood of a successful transfer [2].

Beyond codon usage and nucleotide composition, other aspects of genetic compatibility have been studied. One study found that codon adaptability index (CAI), GC-content and mRNA-folding energy in *E.coli* are factors that are less important to the compatibility [11]. Instead, they concluded that the fitness and functionality of resistance genes are more affected by the resistance mechanism and the phylogenetic origin of the gene.

2.4 Translation and tRNA availability

In protein synthesis, translation is the process where proteins are produced from the template of messenger RNA (mRNA). During translation, tRNAs recognise the

codons on the mRNA molecule and bind via an anticodon on one end, while carrying the corresponding amino acid on the other [12]. This process occurs on the ribosome, which has space for three tRNA molecules. When a third tRNA enters, the first one, which is now uncharged, is released and can be recharged for later use in the process [13].

There are many different tRNA molecules in a cell, though not necessarily one for each codon. The number of tRNA types differs between species, in bacteria the minimum number of tRNA molecules required to translate all amino acids is 31 [12]. This is due to wobble base pairing, where the third position of the mRNA codon, known as the wobble position, can mismatch with the anticodon base on the tRNA [12]. In Table 2.1, the possible base pairings at the wobble position and their corresponding anticodon bases are shown.

Table 2.1: The nucleotide bases adenine (A), cytosine (C), guanine (G), and uracil (U) with their possible pairings at the wobble position in translation. The anticodon can also carry the nucleoside inosine (I) due to deamination of the nucleoside adenosine [12].

Wobble position base	Possible anticodon base
A	U or I
C	G or I
G	C or U
U	A, G, or I

Understanding wobble base pairing helps explain how a limited set of tRNAs can decode multiple codons. The actual abundance of tRNA molecules in a cell differs between bacteria [12], and the factors influencing this availability has been investigated in various studies. In *E. coli*, the abundance of tRNA generally corresponds to the codon usage in the bacterial genome, although this correlation is less significant for phage and transposon genes [14]. One study estimated the tRNA abundance in certain bacteria by determining translationally optimal codons from RNA sequencing data [15]. This approach gave a more accurate result than methods that estimate the tRNA abundance by gene copy number.

3

Methods

In this project, the genetic compatibility has been investigated between antibiotic resistance genes (ARGs) and bacteria, using different methods. Firstly, the structure of the genes and bacterial genomes were looked at, to get a metric of how similar they are. This was done by using the distribution of 5-mers, as well as looking at the GC-content. Secondly, the tRNA availability in the bacteria was compared to the codons in the genes, to identify the compatibility in terms of translational efficiency. Reference genes have also been looked at to find reference values for both metrics.

The bacterial genomes considered in this project were downloaded from National Center for Biotechnology Information (NCBI) [16], as well as their taxonomy. Only genomes that were considered not to be contaminated were kept, which was approximately 1.6 million genomes. The ARGs studied were collected from CARD [17], which contained approximately 6 000 genes.

All code for this project was written in Python and can be found in the following GitHub repository. Some analysis and tables were created using Excel.

3.1 Data processing

A filtering was applied to the bacterial data, where a maximum of 10 bacterial genomes were kept from each species. This was done to reduce bias caused by overrepresented species, and resulted in approximately 77 000 genomes remaining. An additional filtering was performed to keep only the bacteria from the six largest phyla, for easier analysis later in the project. This resulted in approximately 73 000 genomes.

A BLAST [18] database was created using the nucleotide sequences of the 73 000 bacterial genomes. A BLASTn search was then performed to find where the genes are present in the genomes. The following parameters were used: -perc_identity 95, -max_target_seqs 100000, -evalue 1e-5, -qcov_hsp_perc 90 and -best_hit_score_edge 0.1.

After the BLASTn search, the results were filtered to avoid retaining multiple gene matches at the same location in one genome. For overlapping hits, the match with the highest bit score was kept. Overlaps of up to 20% were allowed, but any match

overlapping more than that was considered to target the same region in the genome.

3.2 Genetic compatibility by 5-mers

The initial approach to measure genetic compatibility was by comparing k-mer distributions, specifically 5-mers. A 5-mer is a sequence of five nucleotides from a gene or genome. 5-mers were used since they not only capture codons, but also how the codons are arranged in the genome. The 5-mers were counted using the program KMC [19] with parameters -ci1, and -cs500000 for genomes and -cs10000 for genes. Some genes and genomes contained other letters than A, C, G and T which KMC handles by excluding the 5-mers containing these letters [20]. The 5-mers were counted canonically, meaning that a 5-mer and its reverse complement were treated as the same. The 5-mer counts were then normalised to obtain each gene and genomes' respective distribution. These distributions were compared by calculating the Euclidean distance between each gene and genome. This metric will be referred to as the 5mer score.

3.2.1 Adjusting for gene length

After calculating the 5mer score between each gene and genome, a correlation between the gene length and the 5mer score was observed. To determine whether this correlation depended on a biological or technical factor, it was further investigated by the following method. Only genes that were at least 500 nucleotides long were included. For the genes longer than 500 nucleotides, a random start and end position 500 nucleotides apart was selected, keeping a segment of 500 nucleotides from each gene. The 5-mers were then counted for these segments, the counts were normalised and the Euclidean distance was calculated between each gene segment and genome. This metric will be referred to as the length-adjusted 5mer score.

3.2.2 Comparing GC-content

The GC-content, percentage of the nucleotides G and C, was calculated for each ARG and bacterial genome. Then, the GC-content ratio and difference were computed for each gene-genome pair. The ratio was determined by dividing the GC-content of the ARG by the GC-content of the genome, while the difference was obtained by subtracting the GC-content of the genome from the GC-content of the ARG.

3.2.3 Defining a worst case score

To further compare the 5-mer distributions, the maximum absolute difference between distributions was calculated, which represents the value for the 5-mer with the largest difference between the gene and the genome. The relative difference was also taken into account, which was calculated by dividing the maximum difference with the average between the values causing the maximum difference. Both metrics were calculated for each gene-genome pair.

These values, the maximum difference and relative difference, were then scaled to the range $[0, 1]$. For a value x, the scaled value x_{scaled} was calculated by

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}. \tag{3.1}$$

Then, for each gene-genome pair, the values were combined by multiplying them with 0.5 and adding them together. This score will be referred to as 5mer worst case.

3.3 Genetic compatibility by tRNA availability

Another way to measure compatibility was considered by comparing the tRNA availability in the host cell to the codon usage of the gene.

To quantify the tRNA availability, the program tRNAscan-SE [21] was run on the filtered set of approximately 77 000 genomes. This provided a list of tRNAs found in the genome, along with the specific anticodons used in translation. For some genomes only a few, or none, tRNAs were returned by the program. The results were therefore filtered by removing genomes with fewer than 35 rows, corresponding to a minimum of 31 tRNAs. After filtering, approximately 71 000 genomes remained. Anticodons containing letters other than A, C, G, and T were excluded.

The output from tRNAscan-SE was then compared to the codon usage of each gene in the dataset by comparing their distributions and calculating a one-sided score. The reversed complement of the anticodons were compared to the codons of the genes. The method searched for available translations and assigned them a weight based on the effectiveness of the match. An exact match received a weight of 1, codons with wobble position T and G, see Table 2.1, received a weight of 2, codons with a possible I at wobble position received 4, and codons with no available translation received a weight of 5.

Each codon received a score by multiplying the assigned weight with the squared difference between codon frequencies of the gene and genome. The difference was only considered if the value for the gene was higher than the genome. The total tRNA score for each gene-genome pair was obtained by summing the scores of all individual codons.

3.4 Reference genes

Some genes were selected as reference genes to represent high and low compatibility. The genes chosen as highly compatible references were chromosomal genes in their respective host cells. These included hp1181 in *Helicobacter pylori* [22], vatF in *Yersinia enterocolitica* [23], APH(9)-Ia in *Legionella pneumophila* [24], among others. The total number of compatible reference genes was 94, and 518 data points

were included.

The incompatible reference genes included class B beta-lactamases in gram-positive bacteria and vancomycin resistance genes in gram-negative bacteria. The antibiotic vancomycin affects the cell wall of gram-positive bacteria, however, gram-negative bacteria have an outer membrane that blocks the antibiotic [25], making vancomycin resistance genes unnecessary in gram-negative bacteria. Similarly, beta-lactamases are less prevalent in gram-positive bacteria due to its absence of periplasmic space, which is where these enzymes typically function [26]. The exact genes considered were *NDM*, *IMP*, and *VIM*, which were compared to bacteria of the phylum Bacillota, and *van* genes in bacteria of the phylum Pseudomonadota. The total number of incompatible reference genes was 98, and 500 data points were included.

4

Results and Discussion

This chapter presents the key results from the analysis of genetic compatibility between antibiotic resistance genes (ARGs) and bacterial genomes. The analysis is divided into two metrics; 5mer score, which is a comparison between 5-mer distributions, and tRNA score, which compares codon usage in the genes with tRNA availability in host cells. In relation to these scores, gene length, GC-content, and worst case are considered, as well as specific reference genes. In addition to presenting the results, this chapter also provides discussions around the results and their biological interpretation.

In the following analysis, we will distinguish between gene-genome pairs where the gene is found in the genome during a BLAST search, and gene-genome pairs where the gene is not found in the genome. These are called 'matches' and 'non-matches' respectively.

4.1 5mer score analysis

The 5mer score was first analysed by creating histograms for each gene. An example is presented in Figure 4.1, which displays the histograms for the ARG *tet(Q)*.

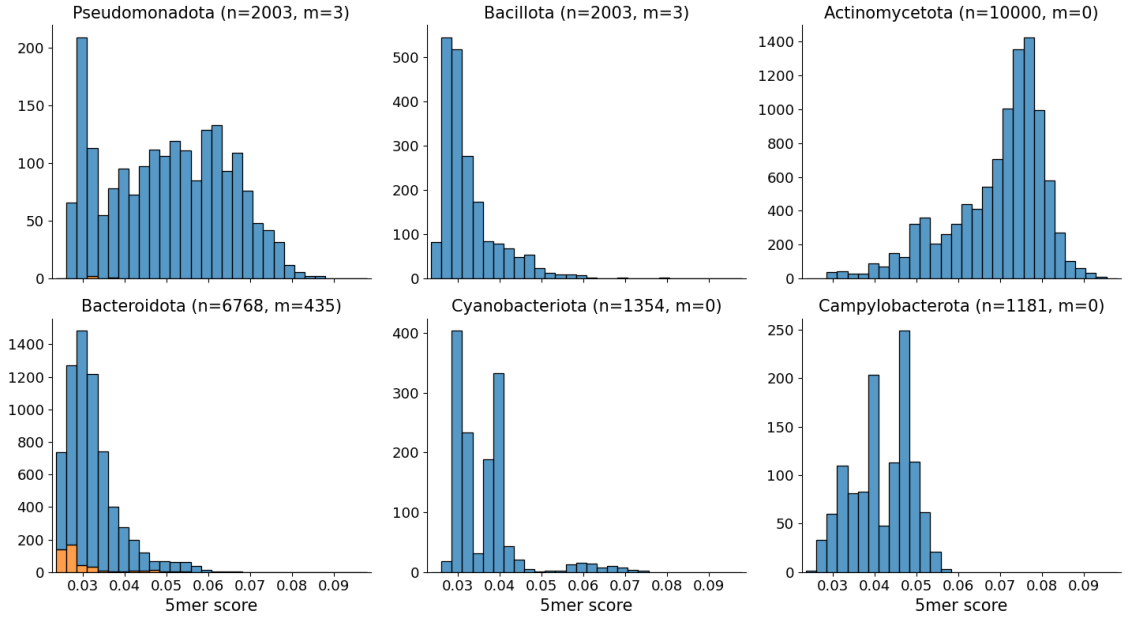


Figure 4.1: For the gene *tet(Q)*, a histogram from each of the six largest phyla in our dataset is shown. The number of bacterial genomes is plotted against the 5mer score, where a low score indicates a more genetically similar pair of *tet(Q)* and genome. Orange indicates the matches, and blue indicates the non-matches. In each phylum, n represents the total number of genomes in the histogram and m the number of matches.

Notably, most matches of *tet(Q)* are found in Bacteroidota with a low 5mer score (< 0.04). This observation aligns with the presumed evolutionary origin of *tet(Q)* in Bacteroidota [27]. Even though *tet(Q)* has not been detected in Cyanobacteriota or Campylobacterota, the figure indicates low values of the 5mer score for these phyla. This suggests that the ARG could potentially spread to them as well. In contrast, the higher values observed for Actinomycetota may indicate a lower likelihood of transfer to that phylum. Interestingly however, the gene *TEM-116* has matches in Actinomycetota despite 5mer scores reaching values as high as 0.065, as presented in Figure A.1. The gene *tet(W)* also has matches in Actinomycetota although with lower 5mer scores, as seen in Figure A.2. These observations suggest that, based on the 5mer score, there is no apparent barrier to horizontal gene transfer between phyla.

When looking at histograms for different genes, it was observed that the matches generally get low scores. This was verified with a Wilcoxon signed rank test that showed significance, $p = 0.0$ ($p < 10^{-308}$), when testing whether, for each gene, the mean 5mer score for matches is lower than the gene's overall mean 5mer score.

To further illustrate this trend, Table 4.1 presents the five genes with the lowest mean 5mer scores, representing the overall most compatible genes. Their mean 5mer scores for the matches are shown in Table 4.2.

Table 4.1: Genes with the lowest overall mean 5mer score, including their respective minimum and maximum scores and gene length.

Gene name	Mean	Min	Max	Gene length
AcrF	0.0389	0.0182	0.0790	3106
acrD	0.0390	0.0186	0.0825	3115
mdsB	0.0398	0.0198	0.0861	3169
mdtF	0.0398	0.0202	0.0812	3115
mdtB	0.0399	0.0189	0.0873	3124

It can be noted that all of the most compatible genes have a gene length over 3000 nucleotides. They also have a low minimum 5mer score around 0.02.

Table 4.2: Information of the matches for the genes listed in Table 4.1, including the mean 5mer score for the matches, minimum and maximum scores, number of phyla they are found in and number of matches.

Gene name	Mean match	Min match	Max match	Phyla	Matches
AcrF	0.0185	0.0182	0.0189	1	158
acrD	0.0200	0.0190	0.0207	1	251
mdsB	0.0215	0.0211	0.0225	1	432
mdtF	0.0208	0.0203	0.0212	1	256
mdtB	0.0240	0.0226	0.0257	1	321

Table 4.2 shows that the genes have lower mean 5mer score for the matches than their respective overall mean 5mer scores shown in Table 4.1. This aligns with the result of the Wilcoxon signed rank test. The lower mean scores for the matches may indicate that the genes are chromosomal in the genomes where a match has been found. Indeed, all genes presented in the table have been found on chromosomes of the bacteria that we have found them in [28]–[32].

In comparison to the most compatible genes, Table 4.3, show the least compatible mobile genes. These are defined as the genes which have spread to at least 3 phyla and have the highest mean 5mer score.

Table 4.3: Genes with the highest overall mean 5mer score that have spread to at least 3 phyla, including their respective minimum and maximum scores and gene length.

Gene name	Mean	Min	Max	Gene length
qacG	0.0884	0.0551	0.127	325
APH(2")-If	0.0819	0.0381	0.125	895
ErmT	0.0766	0.0367	0.120	736
lnuA	0.0763	0.0458	0.119	487
ErmC	0.0759	0.0362	0.120	736

It can be noted that all the least compatible mobile genes have a gene length under 900 nucleotides. The match information for these genes are presented in Table 4.4.

Table 4.4: Information of the matches for the genes listed in Table 4.3, including the mean 5mer score for the matches, minimum and maximum scores, number of phyla they are found in and number of matches.

Gene name	Mean match	Min match	Max match	Matches	Phyla
qacG	0.0674	0.0610	0.106	28	3
APH(2")-If	0.0613	0.0453	0.0951	18	4
ErmT	0.0555	0.0395	0.0937	52	3
lnuA	0.0518	0.0473	0.0841	69	4
ErmC	0.0649	0.0409	0.106	169	3

All genes in Table 4.4 have a quite high 5mer score for max match, which indicates that bacteria can take up genes even if they are not highly compatible according to the 5mer score. This can be due to evolutionary pressure, where the gene is beneficial or even essential for survival, despite not being optimal for growth efficiency.

Notably, the gene lengths for the top 5 most compatible genes are greater than 3000 nucleotides compared to the least compatible where all genes are less than 900 nucleotides. This suggests that the length of the gene has an effect on the 5mer score.

4.1.1 Gene length correlation

Previous results showed an indication that the gene length might influence the 5mer score. Figure 4.2 presents the relationship between the mean 5mer score and the gene length for all genes.

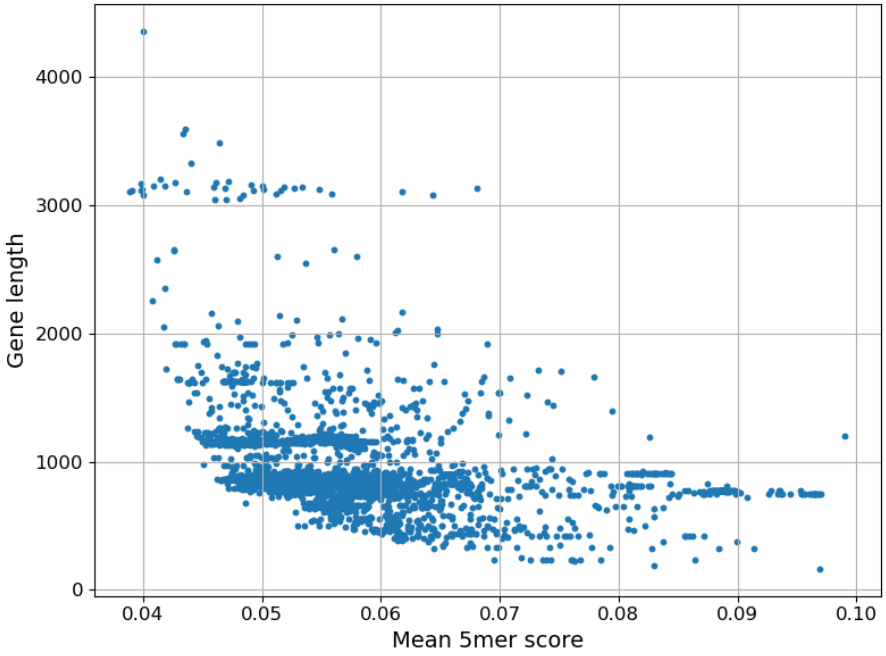


Figure 4.2: The relationship between mean 5mer score and gene length plotted for each gene. A Spearman correlation coefficient of -0.286 ($p = 1.70 \cdot 10^{-114}$) and a Pearson correlation coefficient -0.304 ($p = 3.09 \cdot 10^{-129}$) indicate a weak or moderate negative correlation with high significance.

The Spearman and Pearson correlation tests from Figure 4.2, indicate a significant, but weak or moderate negative effect. This means that genes with larger gene length are associated with slightly lower mean 5mer scores. The figure shows that short genes cannot attain low mean 5mer scores. Although, it is unclear whether this effect is technical or biological.

To investigate this further, only 500 nucleotides of each gene were analysed, see Section 3.2.1 for a more detailed explanation. Figure 4.3 presents the relationship between the mean length-adjusted 5mer score and the original gene lengths, for all genes.

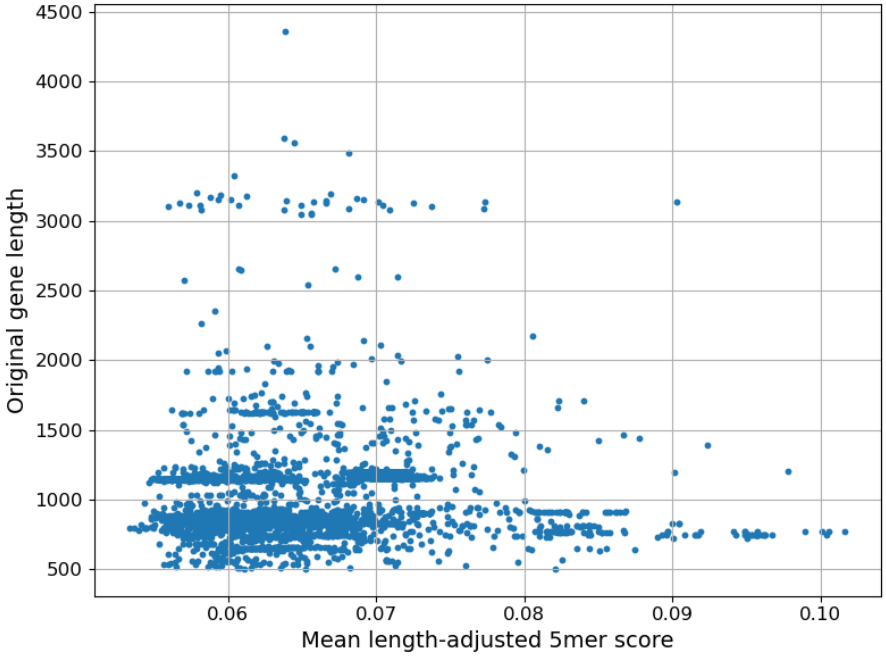


Figure 4.3: The relationship between the original gene length and the mean length-adjusted 5mer score. A Spearman correlation coefficient of 0.237 ($p = 2.93 \cdot 10^{-76}$) and a Pearson correlation coefficient 0.145 ($p = 5.28 \cdot 10^{-29}$) indicate a weak positive correlation.

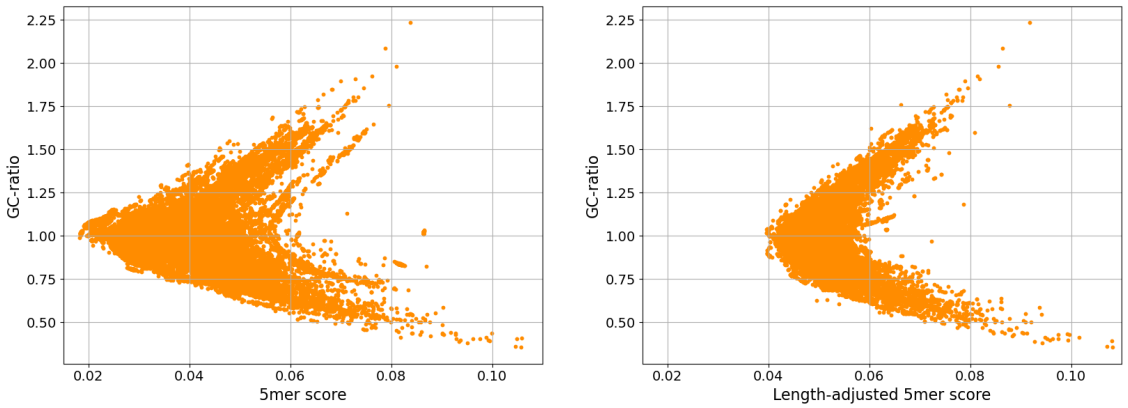
In Figure 4.3, the Spearman correlation shows a weak positive correlation, compared to the negative correlation for the original data in Figure 4.2. This implies that the effect seen on the original data is technical since the same correlation is not seen after adjusting the score for the gene length. This technical bias arises because the score is based on counting 5-mers in each gene. In shorter genes, each additional 5-mer has a proportionally greater impact compared to longer genes. As a result, shorter genes cannot achieve scores as low as those of longer genes.

To further explore how length adjustment affects the results, a histogram for the gene *tet(Q)* using the length-adjusted 5mer score is shown in Figure A.3, corresponding to the original 5mer score histogram in Figure 4.1. While the overall distribution shapes are similar, the values of the scores differ, and there is a sligt shift in the location of the matches. However, the matches are still low compared to the non-matches.

To assess whether the matches have a lower mean than the overall mean, for the length-adjusted 5mer score, a Wilcoxon signed rank test was also performed. The result was significant, $p = 0.0$ ($p < 10^{-308}$).

4.1.2 GC-content comparison

In Figure 4.4, the GC-ratio between ARGs and bacterial genomes is plotted against either the 5mer score or the length-adjusted 5mer score, for each gene-genome pair in which the ARG is present.



(a) 5mer score.

(b) Length-adjusted 5mer score.

Figure 4.4: Comparison of GC-ratio and the two metrics (a) 5mer score and (b) length-adjusted 5mer score, for matches. The GC-ratio was calculated by dividing the GC-content of the ARG by the GC-content of the corresponding bacterial genome.

Figure 4.4 indicates that gene-genome pairs with large differences in GC-content cannot attain low values of the scores. This seems reasonable since a large difference in GC-content will correlate with the compositions of the 5-mers. The figure furthermore suggests that the two scores captures not only the nucleotide composition, but also the sequential order of the nucleotides. When comparing the two scores, it becomes evident that the length-adjusted 5mer score does not reach values as low as the 5mer score, as it accounts for gene length.

The two scores were also compared with the GC-difference in Figure A.4, where the GC-content for the bacterial genome is subtracted from the GC-content of the corresponding ARG. The figure shows no additional pattern beyond those previously observed with the GC-ratio.

4.1.3 Worst case comparison

The 5mer worst case metric was compared to the 5mer score for the gene *tet(Q)* in Figure 4.5.

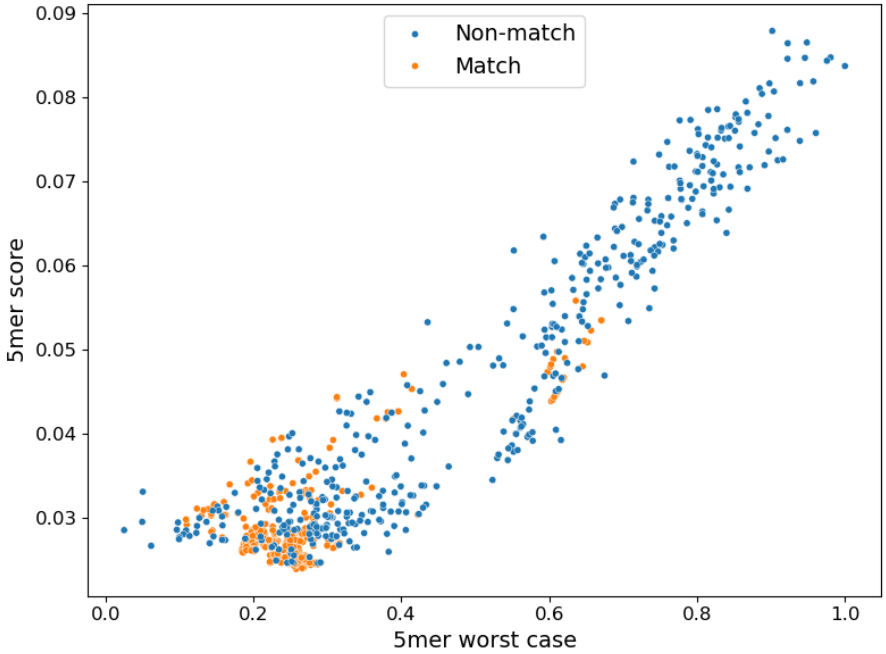
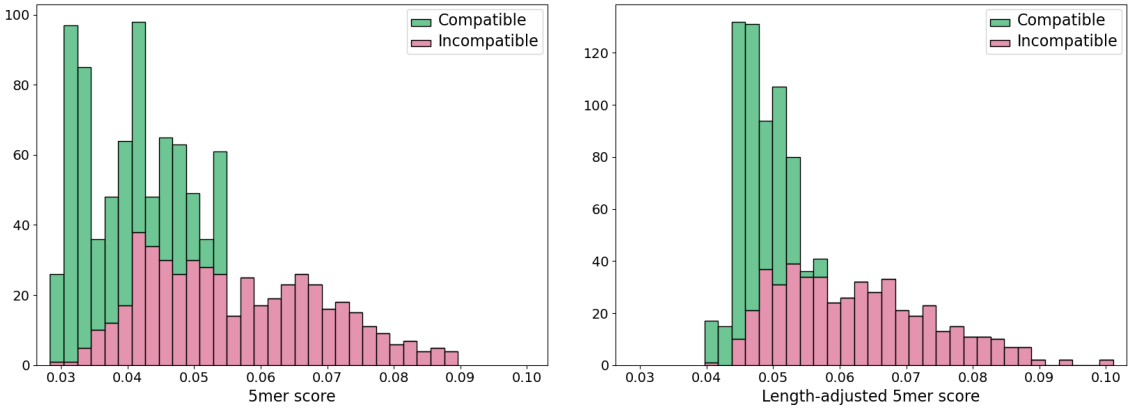


Figure 4.5: For the gene *tet(Q)*, the 5mer score is plotted against the metric 5mer worst case. Orange indicates genomes where the gene is present, and the blue indicates genomes where the gene is not present.

Figure 4.5 shows that the relationship between the two metrics 5mer score and 5mer worst case is linear. It can also be seen that the matches get low values in both measurements. It seems reasonable that these measurements correlates since they are both built from 5-mers.

4.1.4 Reference genes

The reference genes chosen in Section 3.4 are visualised in Figure 4.6. The x-axis represents either the 5mer score, or the length-adjusted 5mer score, while the y-axis indicates the number of gene-genome pairs. The compatible reference genes are chromosomal genes in bacterial hosts that they are present in, and the incompatible reference genes are class B beta-lactamases in gram-positive bacteria, as well as vancomycin resistance genes in gram-negative bacteria.



(a) 5mer score.

(b) Length-adjusted 5mer score.

Figure 4.6: The 5mer score and the length-adjusted 5mer score are plotted for the compatible and incompatible reference genes. Both show significance when comparing if the compatible references have a lower distribution than the incompatible references using a Wilcoxon rank sum test ((a) $p = 1.48 \cdot 10^{-94}$, (b) $p = 5.56 \cdot 10^{-108}$).

The significant p-values support the validity of the approach used to select reference genes, since it aligns with our scores of genetic compatibility. Notably, the values of the 5mer score for the incompatible reference genes show a greater variability compared to the compatible genes. This may reflect the difficulty in defining what a suitable incompatible reference gene would be. The compatible reference reveals a lower variation for the length-adjusted 5mer score than the 5mer score. This reduction in variation could be due to a wide range in gene length for the compatible reference genes.

4.2 tRNA score analysis

To analyse the tRNA score, histograms were created for all genes. In Figure 4.7 an example is presented for the gene *tet(Q)*.

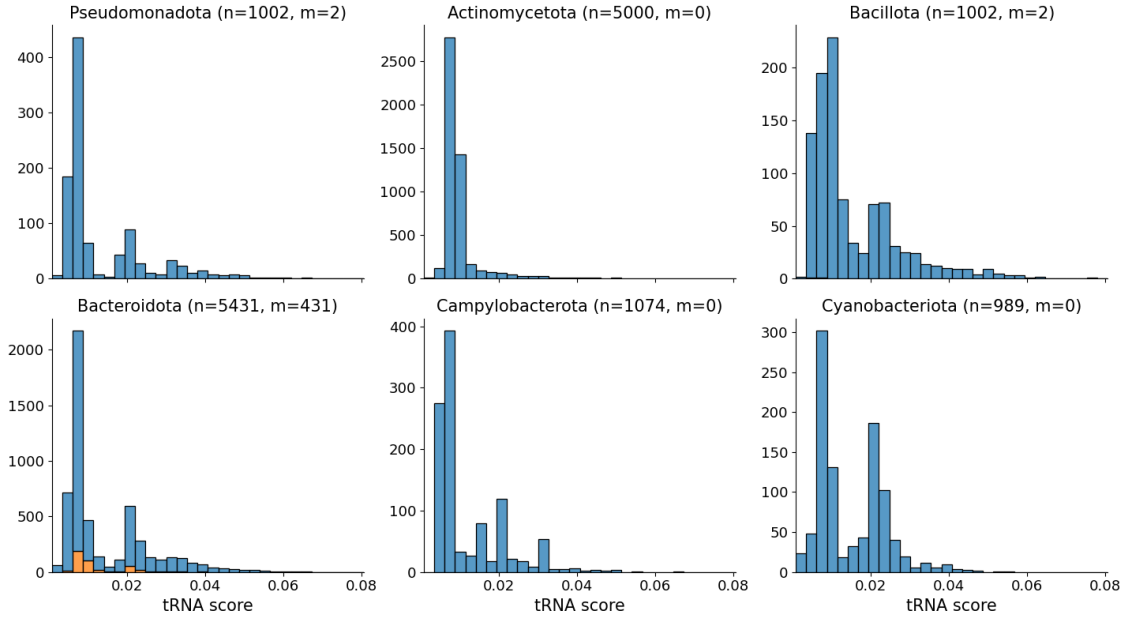
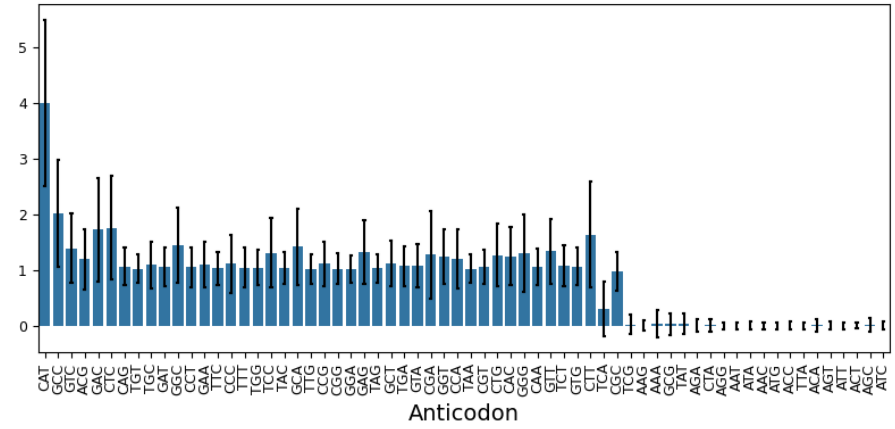
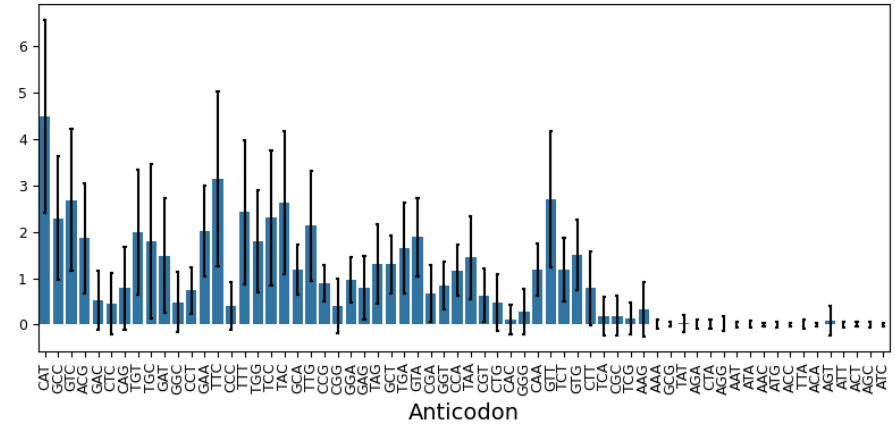


Figure 4.7: Histograms for *tet(Q)* representing the six largest phyla in our dataset, where the y-axis is the number of bacteria, and the x-axis is the tRNA score. Orange indicates the matches, and blue indicates the non-matches. In each phylum, n represents the total number of bacteria in the histogram and m the number of matches.

Figure 4.7 shows a wide range for the tRNA score, indicating a high variation in compatibility. This could be due to how the tRNA score is calculated, where large differences in distributions can disproportionally influence the overall score. As shown in Figure 4.7, most bacteria across all phyla get low tRNA scores, indicating better compatibility. However, it is observed that the distributions vary between phyla, especially for the two phyla Actinomycetota and Bacillota. To analyse this further, their tRNA anticodon distributions are presented in Figure 4.8.



(a) Actinomycetota.



(b) Bacillota.

Figure 4.8: The distributions of anticodons for the two phyla (a) Actinomycetota and (b) Bacillota. For each anticodon, the mean number of tRNA molecules across the bacteria within each phylum is plotted, together with the standard deviations.

The tRNA anticodon distributions differ notably between the two phyla shown in Figure 4.8, which may help explain the variation in their histograms in Figure 4.7. It can also be noted that the phylum Actinomycetota contains 21 % of the genus Streptomyces in our dataset, making the data skewed and potentially influencing the results. For tRNA anticodon distributions of the remaining phyla, see Figure A.5.

After examining histograms for several genes, it was observed that the matches tend to have low tRNA scores, just like the trend seen with the 5mer scores. To further support this observation, a Wilcoxons signed rank test was performed, which confirmed that the mean tRNA score for matches is significantly lower than the overall mean tRNA score ($p = 3.23 \cdot 10^{-76}$).

4.2.1 GC-content comparison

Presented in Figure 4.9 is the GC-ratio between ARGs and bacterial genomes where the ARGs are present, plotted against the tRNA score.

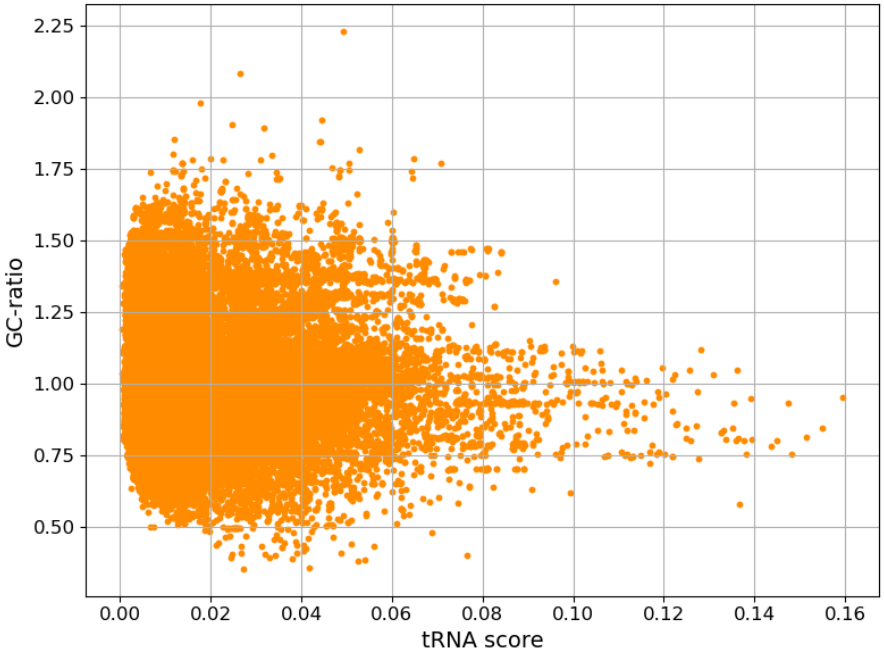


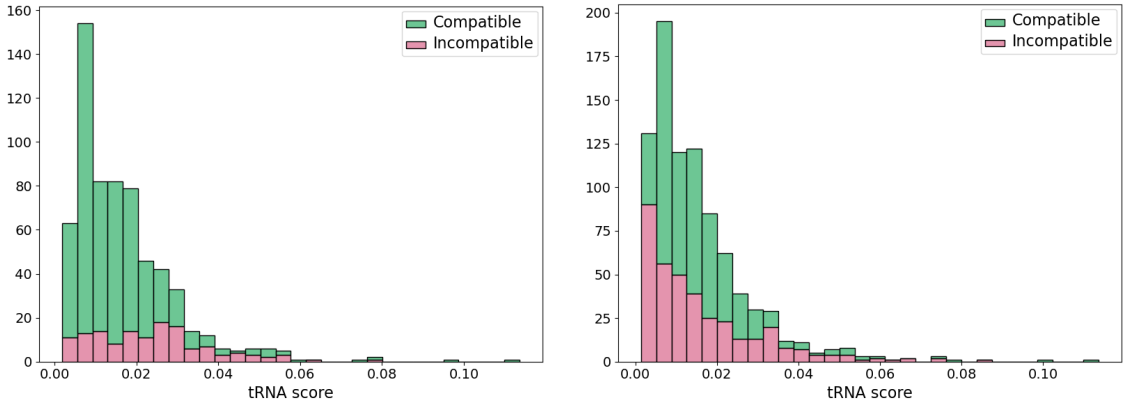
Figure 4.9: Comparison of GC-ratio and tRNA score, for matches. The GC-ratio was calculated by dividing the GC-content of the ARG by the GC-content of the corresponding bacterial genome.

Both Figures 4.9 and A.6, which shows the GC-difference, suggests that low values of the tRNA score are not affected by the GC-content. The reason for this could be that the tRNA anticodons does not reflect the GC-content in the bacterial genome, although one could also argue that they should reflect the GC-content since the tRNA anticodons should be altered to translate the genome effectively. However, the GC-content is calculated on the whole genome, not just the coding regions, which could affect this.

For high values of the tRNA score, both figures show that the GC-content is similar between the genes and genomes. However, this might be due to more gene-genome pairs within the GC-ratio range of 0.75-1.2. To further investigate this possibility, an additional plot was created in Figure A.7, which showed that the correlation between GC-ratio and high tRNA score values remained. The bacteria or genes associated with higher tRNA scores may be closely related bacteria, or many of the same genes. To confirm these observations, additional analyses are required.

4.2.2 Reference genes

The same reference genes that were used in Section 4.1.4 for the 5mer scores, were also used for the tRNA score. These reference genes are shown in Figure 4.10, where the number of gene-genome pairs is plotted against the tRNA score. The figure consists of two plots; in both, the compatible genes are the same, while the incompatible genes are divided between class B beta-lactamases in one plot and vancomycin resistance genes in the other.



(a) Class B beta-lactamases.

(b) Vancomycin resistance genes.

Figure 4.10: The tRNA scores are plotted for compatible (green) and incompatible (pink) reference genes. A Wilcoxon rank sum test was found to be significant for the class B beta-lactamases ((a) $p = 1.04 \cdot 10^{-12}$), but not for the vancomycin resistance genes ((b) $p = 0.937$)), when comparing if the compatible reference has a lower distribution than the incompatible reference.

Compared to the strong significance observed for the reference genes with the 5mer score in Figure 4.6 ($p = 1.48 \cdot 10^{-94}$ and $p = 5.56 \cdot 10^{-108}$), the tRNA score for the same reference genes only imply significance for the class B beta-lactamases in Bacillota ($p = 1.04 \cdot 10^{-12}$). This suggests that the tRNA score could be gene-specific and therefore depend on the particular gene being analysed.

4.3 Comparing 5mer score and tRNA score

A comparison between the metrics length-adjusted 5mer score and tRNA score will be performed. The length-adjusted 5mer score is chosen for this analysis since it accounts for gene length, making it suitable for direct comparison between genes.

In Figure 4.11, the mean length-adjusted 5mer score is plotted against the mean tRNA score, for each gene.

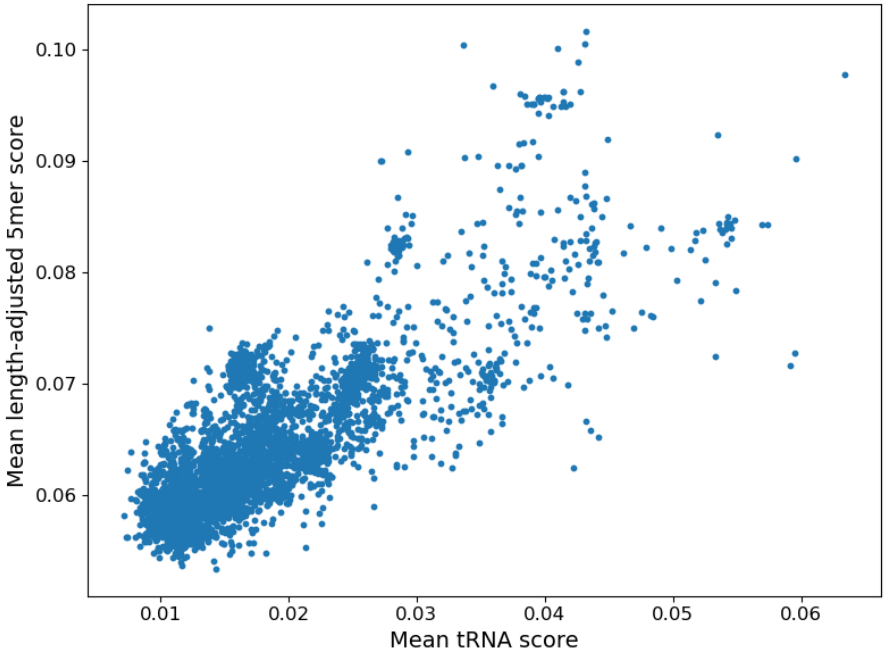


Figure 4.11: For each gene, the mean length-adjusted 5mer score is plotted against the mean tRNA score. A Spearman correlation coefficient of 0.779 ($p = 0.0$) indicate a strong positive correlation.

The Spearman correlation test from Figure 4.11 indicate a significant positive effect. This means that there is a correlation between the two metrics length-adjusted 5mer score and tRNA score. The genes in the bottom left corner can be considered the overall most compatible genes. Excluding genes with no matches, five genes were identified as being among the 300 lowest-scoring genes in both metrics. These five genes are shown in Table 4.5.

Table 4.5: Most compatible ARGs with respect to their mean length-adjusted 5mer score and mean tRNA score, including the number of matches, number of phyla they are found in and their gene length.

Gene name	Mean length-adjusted 5mer score	Mean tRNA score	Matches	Phyla	Gene length
OXA-9	0.0563	0.00731	47	1	826
CARB-23	0.0563	0.00904	4	1	940
OXA-919	0.0562	0.00808	14	1	829
erm(56)	0.0548	0.00942	8	1	802
AcrF	0.0560	0.00961	157	1	3106

The genes listed in Table 4.5 are the most compatible ones based on both metrics, suggesting a high potential for horizontal transfer across multiple phyla. However, each gene is only found in a single phylum. This could be due to ecological separation or because there has been no evolutionary pressure for other bacteria to acquire them.

In contrast, the least compatible genes that are found in at least three phyla are shown in Table 4.6. These genes were among the six highest-scoring genes in both the length-adjusted 5mer score and the tRNA score, considering the genes that had spread to at least three phyla.

Table 4.6: Least compatible ARGs that have spread to at least 3 phyla with respect to their mean length-adjusted 5mer score and tRNA score, including their number of matches, number of phyla they are found in, and their gene length.

Gene name	Mean length-adjusted 5mer score	Mean tRNA score	Matches	Phyla	Gene length
AAC6_Ie_APH2_Ia	0.0878	0.0431	309	4	1441
APH(2")-If	0.0831	0.0545	18	4	895
tet(K)	0.0810	0.0416	91	3	1381
ErmC	0.0793	0.0406	164	3	736

While these genes may appear as outliers in terms of compatibility, their persistence and spread across phyla suggest some evolutionary advantage in certain conditions. It is also interesting to note that these four genes were found among the six genes that scored the highest values in both metrics, considering the ones that had spread to at least three phyla. This overlap indicates that both scores captures similar aspects of gene compatibility, particularly at the extreme end representing the least compatible genes.

Overall, we have explored genetic compatibility using two metrics. The 5mer score compares similarities in nucleotide composition by capturing both the distribution of codons and how they are arranged within the sequences. This metric of genetic compatibility revealed that matched gene-genome pairs exhibited a lower mean 5mer score compared to the overall mean. This suggests that for an ARG to be successfully acquired by a bacterium, the gene and host must exhibit similarities in their nucleotide sequences. This is consistent with the results reported in two previous studies, where it was shown that high genetic incompatibility reduces the probability of a transfer [2] and that compatibility of codon usage between genes and bacterial genomes increases the probability that a gene gets taken up by a bacterial host [1]. The 5mer score is sensitive to large differences in GC-content, as discussed in Section 4.1.2, and showed a moderate correlation with the gene length. Given that this metric is based on methods established in earlier research, it is likely to be both robust and reliable.

In contrast, the tRNA score represents a more explorative approach, focusing on translational efficiency by comparing codon usage in the genes and the tRNA availability in the bacteria. Unlike the 5mer score, the tRNA score did not exhibit a clear correlation with either GC-content or gene length, see Figure A.8. One might expect the two metrics to show similar results, given that tRNA availability in cells has been shown to correlate with the codon usage in genomes [14]. However, the 5mer score accounts for the entire genome, including non-coding regions, which may not be of importance. The distribution of tRNA availability could therefore be more

closely correlated to the specific coding regions of the bacterial genome. Additionally, biological factors such as growth rate may influence the interpretation of the tRNA score, since slow-growing bacteria can have only a single copy of many of their tRNA genes, in contrast to fast-growing bacteria which have a larger variability in tRNA gene copy number [15]. This could result in a higher score for slow-growing bacteria despite effective translation. Nevertheless, there are instances in which both metrics yield consistent results. As shown in Table 4.5, both scores identified the same genes as the least compatible mobile genes, indicating alignment at the extremes of the scores.

5

Conclusion

In conclusion, both metrics, the 5mer score and the tRNA score, capture certain aspects of genetic compatibility. This interpretation is supported by the significance of the reference genes, see Section 4.1.4. Although tRNA score did not show significance for the vancomycin resistance genes, it still gave a significant result for the beta-lactamases which is statistically meaningful, see Section 4.2.2. The results also show that overall, the matching gene-genome pairs tend to have a lower score compared to the average for the gene, suggesting that a certain degree of compatibility is required for horizontal gene transfer.

Nevertheless, the relationship between compatibility and successful gene transfer is not absolute. Some matches exhibit relatively high scores, indicating that transfers can still occur between genetically less compatible genes and genomes. This could reflect evolutionary pressures, where bacteria acquire genes that offer a selective advantage, even at the cost of reduced growth efficiency. The results of the 5mer worst case indicate that a high worst case value reduces the likelihood of gene transfer. This aligns with previous studies, which have concluded that while high similarity is not essential for successful transfer, high dissimilarity significantly lowers the probability. Additionally, low compatibility scores were observed for gene-genome pairs with no detected matches. ARGs tend to spread in environments that are favourable, thus, certain bacteria may not have acquired specific ARGs simply because they have not been exposed to such environments. This underlines a limitation of the database where not all genes have disseminated to their full potential range.

The 5mer score was found to depend on gene length, and further analysis indicated that it was a technical bias. Therefore, gene length is an important factor to take into account when working with nucleotide composition. To further analyse this metric, it would be interesting to implement it in a machine learning algorithm to see if it can predict the spread of ARGs based on genetic compatibility.

While the 5mer score shows potential for predictive modelling, the tRNA score has several limitations that need to be further studied to improve its reliability. The limitations include how the tRNA availability was quantified, how the weighted scores were chosen, and that large differences in distributions heavily impacted the score. Analysis of tRNA availability as a metric of genetic compatibility remains relevant, given that it is important for the translational efficiency in bacteria.

Bibliography

[1] A. Medrano-Soto, G. Moreno-Hagelsieb, P. Vinuesa, J. A. Christen, and J. Collado-Vides, "Successful lateral transfer requires codon usage compatibility between foreign genes and recipient genomes," *Molecular Biology and Evolution*, vol. 21, no. 10, pp. 1884–1894, Jul. 2004. DOI: 10.1093/molbev/msh202.

[2] D. Lund, M. Parras-Moltó, J. S. Inda-Díaz, *et al.*, "Genetic compatibility and ecological connectivity drive the dissemination of antibiotic resistance genes," *bioRxiv*, 2024. DOI: 10.1101/2024.10.15.617735.

[3] H. Y and G. N., *Antibiotic Resistance*, https://www.ncbi.nlm.nih.gov/books/NBK513277/, Accessed: 2025-05-08.

[4] WHO, *Antimicrobial resistance*, https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance, Accessed: 2025-05-08.

[5] R. WC, "An overview of the antimicrobial resistance mechanisms of bacteria," *AIMS Microbiol*, vol. 4, no. 3, 2018. DOI: https://doi.org/10.3934/microbiol.2018.3.482.

[6] *Antibiotika och antibiotikaresistens*, https://www.folkhalsomyndigheten.se/smittskydd-beredskap/antibiotika-och-antibiotikaresistens/, Accessed: 2025-05-08.

[7] S. Srivastava, "Bacteria and science of genetics," in *Genetics of Bacteria*, Springer, 2013, ch. 1.

[8] B. Alberts, R. Heald, A. Johnson, *et al.*, "Pathogens and infection," in *Molecular Biology of the Cell*, 7th ed., W. W. Norton & Company, 2022, ch. 23.

[9] S. Tao, H. Chen, N. Li, T. Wang, and W. Liang, "The Spread of Antibiotic Resistance Genes In Vivo Model," *Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 2022, Jul. 2022. DOI: 10.1155/2022/3348695.

[10] M. Parras-Moltó, D. Lund, S. Ebmeyer, D. J. Larsson, A. Johnning, and E. Kristiansson, "The transfer of antibiotic resistance genes between evolutionary distant bacteria," *bioRxiv*, 2024. DOI: 10.1101/2024.10.22.619579.

[11] A. Porse, T. S. Schou, C. Munck, M. M. H. Ellabaan, and M. O. A. Sommer, "Biochemical mechanisms determine the functional compatibility of heterologous genes," *Nature Communications*, vol. 9, no. 522, 2018. DOI: https://doi.org/10.1038/s41467-018-02944-3.

[12] B. Alberts, R. Heald, A. Johnson, *et al.*, "How cells read the genome: From dna to protein," in *Molecular Biology of the Cell*, 7th ed., W. W. Norton & Company, 2022, ch. 6.

[13] D. P. Clark, "Protein synthesis," in *Molecular biology*. Elsevier Science & Technology Books, 2005, ch. 8.

[14] T. Ikemura, "Correlation between the abundance of escherichia coli transfer rnas and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the e. coli translational system," *Journal of Molecular Biology*, vol. 151, no. 3, pp. 389–409, 1981. DOI: https://doi.org/10.1016/0022-2836(81)90003-6.

[15] Y. Wei, J. Silke, and X. Xia, "An improved estimation of trna expression to better elucidate the coevolution between trna abundance and codon usage in bacteria," *Scientific Reports*, vol. 9, p. 3184, Feb. 2019. DOI: 10.1038/s41598-019-39369-x.

[16] P. Kitts, D. Church, F. Thibaud-Nissen, J. Choi, V. Hem, and S. V, "Assembly: a resource for assembled genomes at NCBI," *Nucleic acids research*, vol. 44, no. D1, 2016.

[17] B. P. Alcock, W. Huynh, R. Chalil, *et al.*, "CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database," *Nucleic Acids Research*, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36263822/.

[18] C. Camacho, T. Madden, T. Tao, R. Agarwala, and A. Morgulis, *BLAST® command line applications user manual*, Bethesda (MD): National Center for Biotechnology Information (US), 2008. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279690/pdf/Bookshelf_NBK279690.pdf.

[19] M. Kokot, M. Długosz, and S. Deorowicz, "Kmc 3: Counting and manipulating k-mer statistics," *Bioinformatics*, vol. 33, no. 17, pp. 2759–2761, May 2017, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btx304.

[20] S. Deorowicz, A. Debudaj-Grabysz, and S. Grabowski, "Disk-based -mer counting on a pc," *BMC bioinformatics*, vol. 14, p. 160, May 2013. DOI: 10.1186/1471-2105-14-160.

[21] T. Lowe, *Trnascan-se: A program for improved transfer rna detection in genomic sequence*, 2001.

[22] CARD, *hp1181*, https://card.mcmaster.ca/ontology/40735, Accessed: 2025-04-25.

[23] CARD, *vatF*, https://card.mcmaster.ca/ontology/40399, Accessed: 2025-04-25.

[24] CARD, *APH(9)-Ia*, https://card.mcmaster.ca/ontology/39062, Accessed: 2025-04-25.

[25] G. D. Wright, "Mechanisms of resistance to antibiotics," *Elsivier*, 2003.

[26] M. Toth, N. T. Antunes, N. K. Stewart, *et al.*, "Class D β-lactamases do exist in Gram-positive bacteria," *Nature chemical biology*, vol. 12, no. 1, 2016. DOI: https://doi.org/10.1038/nchembio.1950.

[27] CARD, *tet(Q)*, https://card.mcmaster.ca/ontology/36330, Accessed: 2025-05-05.

[28] CARD, *AcrF*, https://card.mcmaster.ca/ontology/36641, Accessed: 2025-04-29.

[29] CARD, *acrD*, https://card.mcmaster.ca/ontology/36630, Accessed: 2025-04-29.

[30] CARD, *mdtF*, https://card.mcmaster.ca/ontology/37176, Accessed: 2025-04-29.
[31] CARD, *mdtB*, https://card.mcmaster.ca/ontology/37173, Accessed: 2025-04-29.
[32] CARD, *mdsB*, https://card.mcmaster.ca/ontology/37170, Accessed: 2025-04-29.

A

Appendix 1

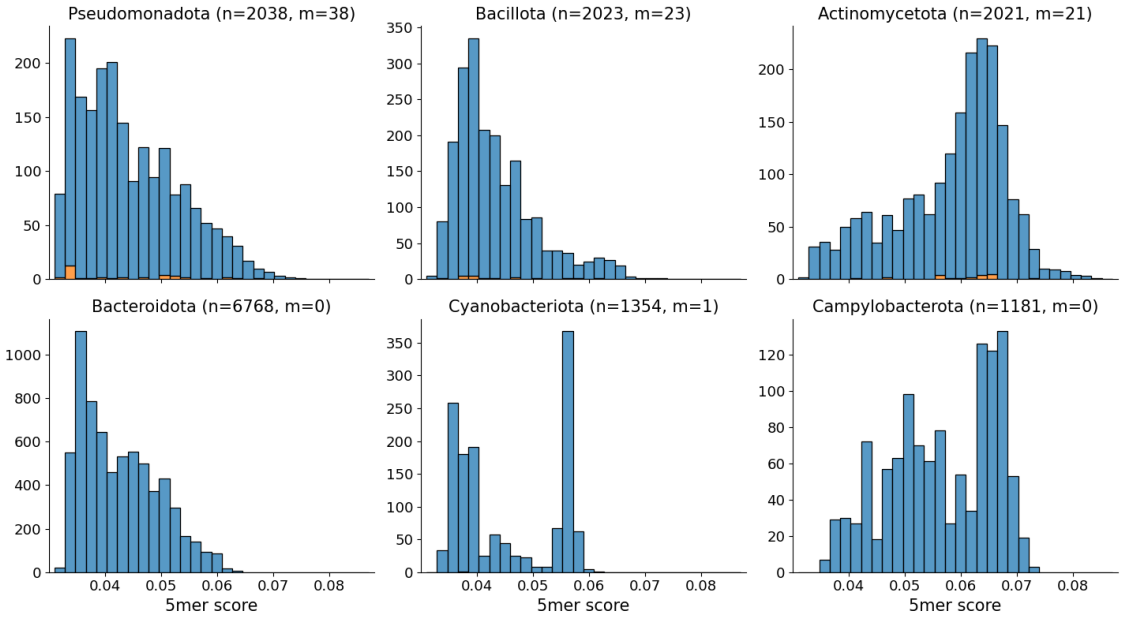


Figure A.1: Histogram for *TEM-116* using the metric 5mer score. The figure depicts the six largest phyla, where orange represents the genomes where the gene has been found and blue where it has not been found.

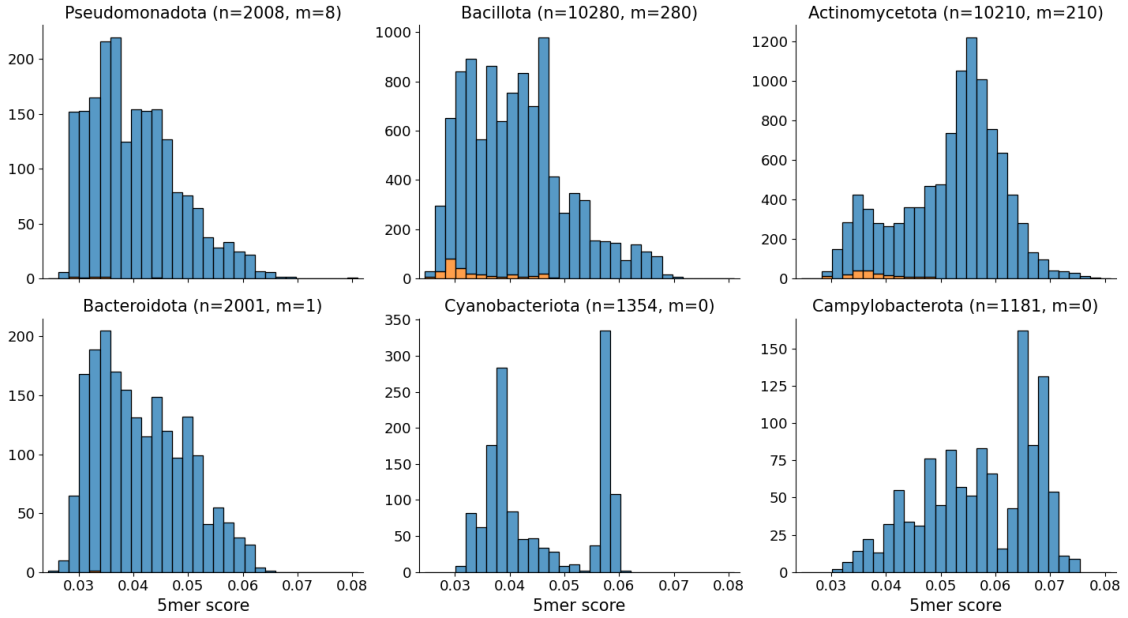


Figure A.2: Histogram for *tet(W)* using the metric 5mer score. The figure depicts the six largest phyla, where orange represents the genomes where the gene has been found and blue where it has not been found.

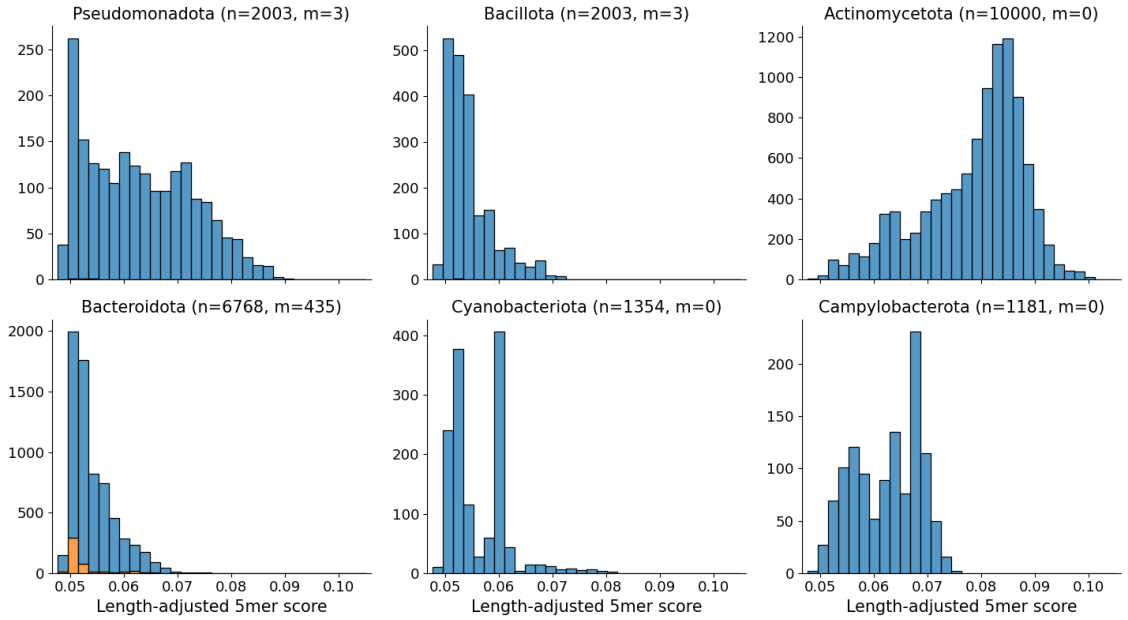
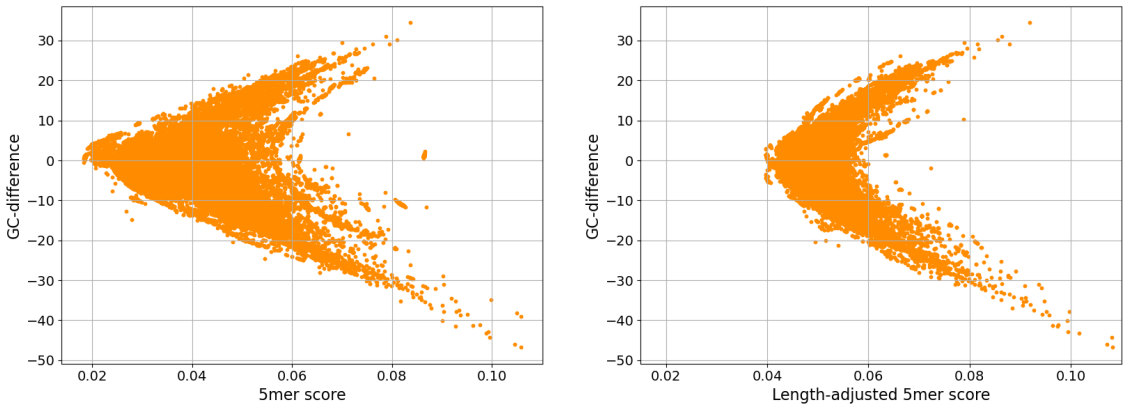


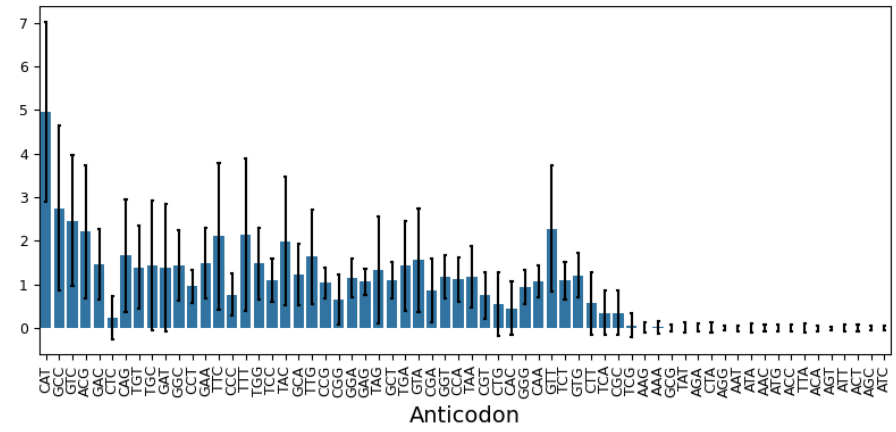
Figure A.3: Histogram for *tet(Q)* using the metric length-adjusted 5mer score. The figure depicts the 6 largest phyla, where orange represents the genomes where the gene has been found and blue where it has not been found.



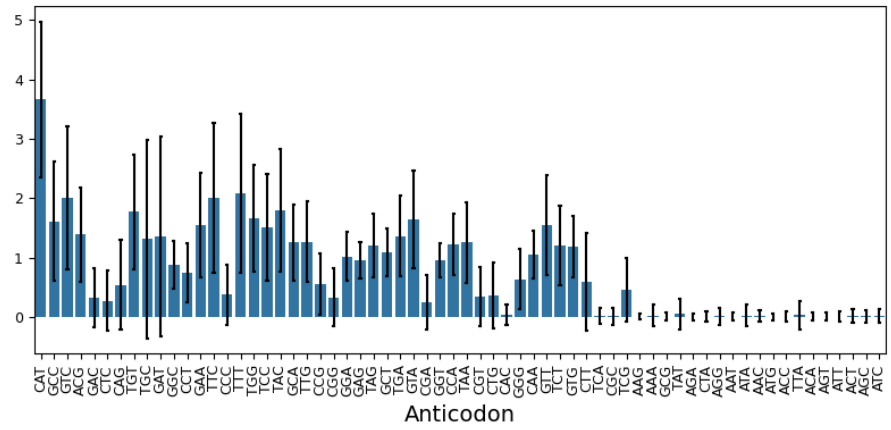
(a) 5mer score.

(b) Length-adjusted 5mer score.

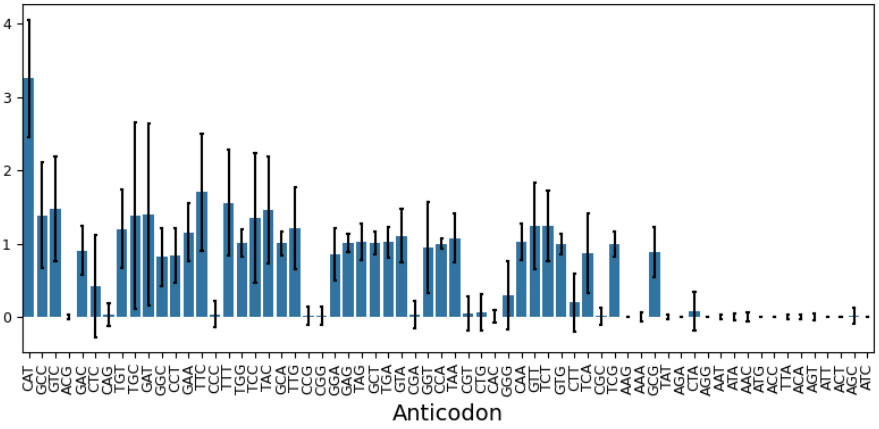
Figure A.4: Comparison of GC-difference and the two metrics (a) 5mer score and (b) length-adjusted 5mer score, for matches. The GC-difference was calculated by subtracting the GC-content for the bacterial genome from the GC-content of the corresponding ARG.



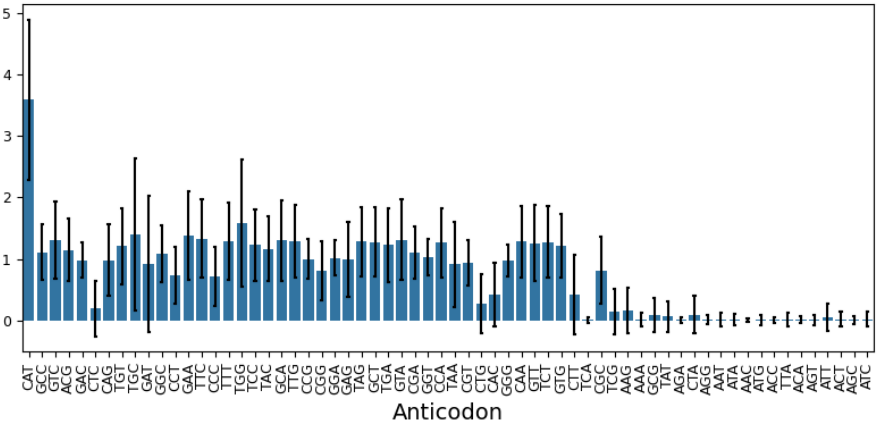
(a) Pseudomonadota



(b) Bacteroidota



(c) Campylobacterota



(d) Cyanobacteriota

Figure A.5: The distributions of anticodons for the 6 largest phyla. The mean number of tRNA molecules for each anticodon is plotted together with the standard deviations.

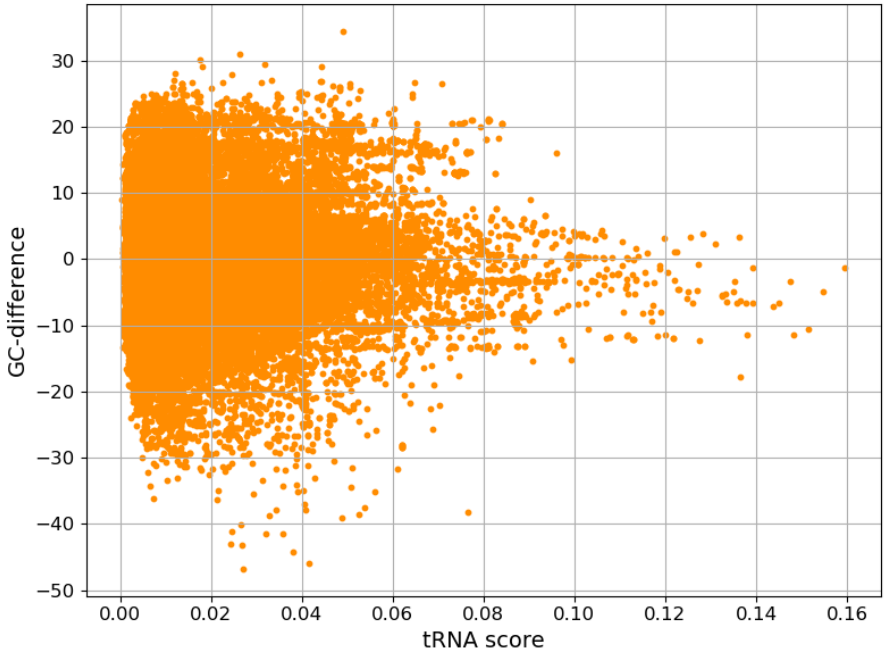


Figure A.6: Comparison of ARGs and the bacterial genomes where they are present based on GC-content and tRNA score, using the method GC-difference, where the GC-content for the bacterial genome is subtracted from the GC-content of the corresponding ARG.

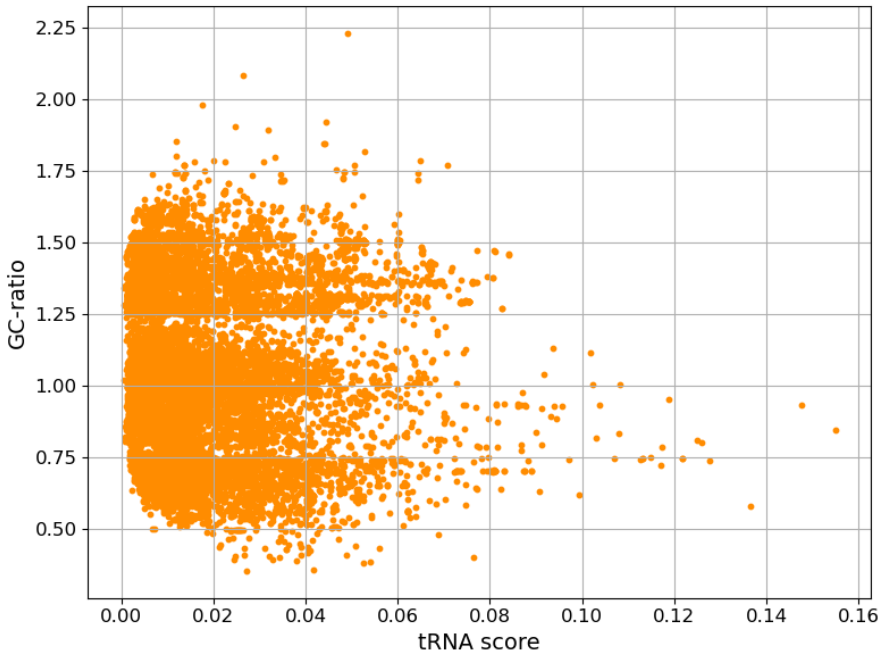


Figure A.7: A downsampled version of Figure 4.9 where the downsampling was performed in the following way. The data points were divided into GC-ratio bins (< 0.5, 0.5-0.75, 0.75-1.00, 1.00-1.25, 1.25-1.50, > 1.50). All data points were kept from the < 0.5 and > 1.5 bins. For the remaining bins, a number of data points equal to the size of the smallest bin among them was randomly selected and kept.

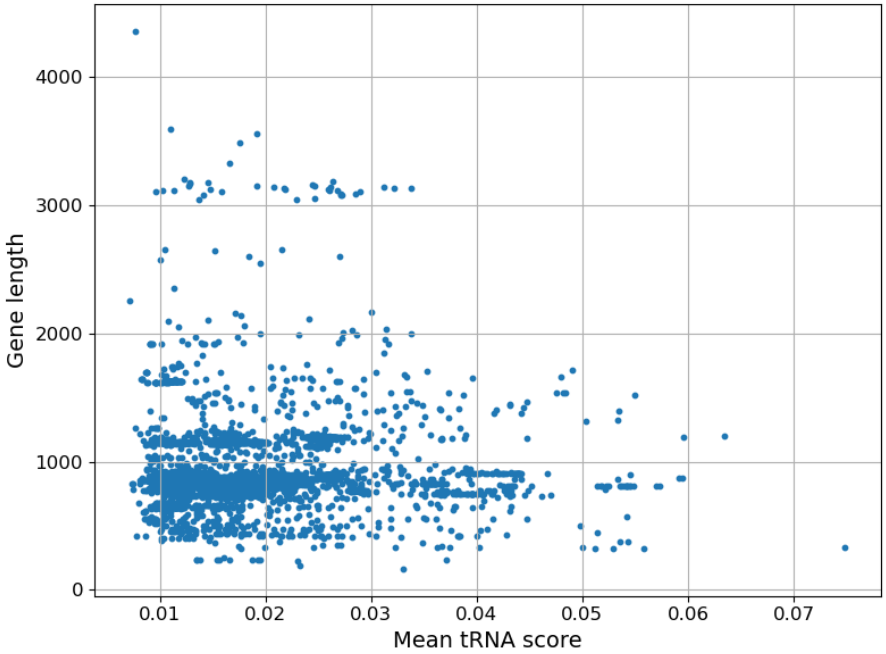


Figure A.8: The relationship between the mean tRNA score and gene length plotted for each gene. A Spearman correlation coefficient of 0.128 ($p = 1.32 \cdot 10^{-23}$) and a Pearson correlation coefficient of 0.0479 ($p = 1.94 \cdot 10^{-4}$).

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

CHALMERS
UNIVERSITY OF TECHNOLOGY