

Wilfrid Laurier University

## Scholars Commons @ Laurier

---

Theses and Dissertations (Comprehensive)


---

2021

# Composition and Homology in the Taxonomic Classification of Escherichia coli

Tanya Irani  
[iran1960@mylaurier.ca](mailto:iran1960@mylaurier.ca)

Follow this and additional works at: <https://scholars.wlu.ca/etd>

 Part of the [Bioinformatics Commons](#), [Computational Biology Commons](#), [Genetics Commons](#), [Genomics Commons](#), [Integrative Biology Commons](#), [Molecular Genetics Commons](#), and the [Other Genetics and Genomics Commons](#)

---

### Recommended Citation

Irani, Tanya, "Composition and Homology in the Taxonomic Classification of Escherichia coli" (2021). *Theses and Dissertations (Comprehensive)*. 2381.  
<https://scholars.wlu.ca/etd/2381>

This Thesis is brought to you for free and open access by Scholars Commons @ Laurier. It has been accepted for inclusion in Theses and Dissertations (Comprehensive) by an authorized administrator of Scholars Commons @ Laurier. For more information, please contact [scholarscommons@wlu.ca](mailto:scholarscommons@wlu.ca).

Composition and Homology in the  
Taxonomic Classification of *Escherichia*  
*coli*

By

Tanya Irani

MSc Integrative Biology, Wilfrid Laurier University, 2021

THESIS

Submitted to the Department of Biology

Faculty of Science

In partial fulfillment of the requirements for the

Masters of Science in Integrative Biology

Wilfrid Laurier University

© Tanya Irani 2021

# Abstract

As new techniques have been introduced, specifically the possibility of complete genome sequencing, better methods of defining bacterial species have also been proposed. One of the most recently proposed methods, using bioinformatic techniques, is to calculate the average nucleotide identity (ANI) between the homologous genome segments of different isolates. Another method for species discrimination that has been tested successfully is the similarity of DNA compositional signatures. However, in a recent update, DNA signatures split the available *Escherichia coli* complete genomes into three groups. To check if this result was consistent with such genomes belonging to different species, we tested methods based on genomic composition and compared them to classic homology methods. The five methods used were ANI, DNA signatures, 16s rRNA, 23s rRNA, and genomic similarity score. All species discrimination methods grouped genomes of *E. coli* slightly differently. However, the DNA signatures and ANI split the groups similarly, suggesting that methods of delimitation based on genetic composition are just as effective as methods based on homology.

# Acknowledgements

I would like to express the deepest level of gratitude to my supervisor, Dr. Gabriel Moreno-Hagelsieb for the overwhelming level of support offered throughout the duration of my undergraduate and graduate thesis. I am truly thankful for all the energy and time you invested by providing me with insightful feedback throughout this process. I would also like to thank Julie Hernandez for the selfless support you offered, I am eternally grateful. As well as Dr. Andrew Doxey Dr. Michael Suits for their support and guidance as members of my thesis committee. Finally, I would like to thank my family and friends for their unconditional support throughout the duration of this thesis, and throughout the entirety of my graduate degree. I am truly thankful for all of you.

# Table of Contents

<b>Abstract</b>	<b>2</b>
<b>Acknowledgements</b>	<b>3</b>
<b>Table of Contents</b>	<b>4</b>
<b>List of Tables</b>	<b>5</b>
<b>List of Figures &amp; Illustrations</b>	<b>6</b>
<b>Introduction</b>	<b>9</b>
<b>Methods</b>	<b>18</b>
<b>Results</b>	<b>21</b>
<b>Discussion</b>	<b>35</b>
<b>Conclusions</b>	<b>41</b>
<b>Integrative Biology statement</b>	<b>41</b>
<b>References</b>	<b>42</b>
<b>Appendix</b>	<b>46</b>

# List of Tables

## Introduction

**Table 1.1.** Four common tests performed in order to classify bacteria as either *Escherichia coli* or *Shigella*. The similarity in the results of these tests provides an explanation as to why there are genomes of *Shigella* that are found in databases under the species of *E. coli*.

## Results

**Table 3.1.** Species found in the three groups obtained using DNA signatures at a 0.03 threshold. The first two clusters mainly contained genomes labeled as *Escherichia coli*, with the second containing most of the *Shigella* genomes. However, the third group contained only one *E. coli* genome, shown in red. The third cluster was left out of the analyses as it was suspected to be a misclassified genome of *Klebsiella*, which was later confirmed as it was later reclassified as *Klebsiella pneumoniae* in the original database.

**Table 3.3.** Amount of genomes of *Shigella* and *E.coli* found in a single group based on different methods. These thresholds were determined by comparing same-species genome pairs as positive datasets against same-genus genome pairs as negative datasets as indicated by the ROC curves.

## Appendix

**Table A.1.** ROC curve data based on all the different methods of discrimination used, when comparing Family versus species.

**Table A.2.** ROC curve data based on all the different methods of discrimination used when comparing Genus versus species.

**Table A.3.** Number of Groups at different distances for di, tri and tetra nucleotide DNA signatures.

# List of Figures & Illustrations

## Introduction

**Figure 1.1.** Comparison of homology based methods of discrimination to methods based on composition. Homology based methods compare the portions of the genomes that are significantly similar to each other, the intersection, as shown in the top Venn diagram. Genomic composition methods compare the compositions of whole genomes, without alignment.

**Figure 1.2.** Comparison of two methods used in species classification. The first image is the ANI method, which has to find matching regions between genomes to determine how similar they are to one another. The second image is the method of DNA signatures. The signature calculation is exemplified with the dinucleotide AA, where the predicted AA value is derived from the genomic proportion of AT and the observed is learned from the genome itself. The observed value is then divided by the predicted value to give a ratio of 1.17 for the dinucleotide AA for this given genome. The dinucleotide DNA signature consists of the vector of these observed/expected ratios for each dinucleotide. The resulting vectors, DNA signatures, are compared to each other using Manhattan distances.

**Figure 1.3.** Comparison of number of clusters and genomes using three methods of species classification. The results show that based on the DNA signatures method there were three clusters found, with the largest containing 324 genomes. The ANI method gave 10 different clusters, with the largest containing 560 genomes and the 16S rRNA method gave 28 different clusters with the largest containing 540 genomes.

## Results

**Figure 3.1.** Hierarchical clusters based on homology methods of classification, including ANI 16S rRNA and 23S rRNA. It is interesting to note that ANI clustered most *Shigella* genomes together into a single group. The *Shigella* 16S and 23S rRNA genes did not group together as clearly.

**Figure 3.2.** Hierarchical clusters based on compositional methods of classification, including di, tri and tetranucleotide DNA signatures. All three methods grouped most *Shigella* genomes close together.

**Figure 3.3.** Tanglegram representing similarities between hierarchical clusters of ANI and tri-nucleotide DNA signatures. On the right is the hierarchical cluster for ANI and on the left is the hierarchical cluster for DNA signatures. The number on top is an entanglement coefficient which corresponds to how well the two clusters align to each other. The low entanglement of 0.08 indicates that the results of the two methods are very similar.

**Figure 3.4.** ROC curves for DNA signatures and ANI, with same-family genome pairs used as negative datasets. A ROC curve is a performance measurement that is often used for classification problems testing different thresholds. These graphs are plotted with the true positive rate (sensitivity) on the y-axis, against the false positive rate (1 - specificity) on the x-axis. The area under the curve (AUC) represents the accuracy of the method for classification. The AUC suggested that all methods of classification, DNA signatures and ANI, were able to differentiate between species within the same taxonomic family.

**Figure 3.5.** ROC curves for DNA signatures and ANI, with same-genus genome pairs used as a negative dataset. The AUCs were lower than those displayed in Figure 3.4. The AUCs for DNA signatures were not as high as the AUC for ANI, meaning that DNA signatures were not as good at differentiating between species of the same genus as ANI.



**Figure 3.6.** Hierarchical clusters based on MASH and Dashing. Both methods kept most of the *Shigella* genomes into the same group.

**Figure 3.7.** Tanglegram of ANI vs MASH, where the hierarchical cluster of ANI is seen on the right side and the hierarchical cluster of MASH is seen on the left. The entanglement coefficient seen here is 0.01, indicating that the results of the hierarchical clusters for both MASH and ANI are very similar to one another, and almost identical.

**Figure 3.8.** Tanglegram of ANI vs Dashing-MASH, where the hierarchical cluster of ANI is seen on the right side and the hierarchical cluster of Dashing-MASH is seen on the left. The entanglement coefficient seen here is 0.6, indicating that the results for the hierarchical clusters of Dashing and ANI are not well aligned, meaning the methods group them very differently from one another.

**Figure 3.9.** ROC curves for MASH and Dashing-Mash, with same-genus genome pairs used as a negative dataset. What is seen is that both had an AUC of 0.96, indicating that both methods are effective in differentiating between genus and species, where genus is the negative dataset.

# Introduction

The definition and identification of species have been a challenge in biological sciences for centuries. Many species concepts have been proposed in order to classify plants, animals, and bacteria. As new methods of bacterial classification are introduced, the classification of species continues to change as well. Early methods of bacterial classification started by looking at the morphology of bacteria and conducting biochemical tests to group genomes into the same species (Scheutz and Strockbine 2015). The problem with lab techniques is that they are laborious and sometimes unable to be used as certain bacteria cannot be cultured in the lab. Although these methods are still used today, more and more, they are used in conjunction with sequence analysis methods to overcome some of the difficulties of lab techniques. Most sequence analyses compare homologous genome segments to determine which organisms belong together. These methods are accurate but can be very time-consuming. Due to the time inefficiency and computer processing that these methods take, it is possible that new, heuristic approaches need to be considered. An alternative to comparing homologous segments is to compare the composition of genomes to discriminate between genomes of the same species (Moreno-Hagelsieb et al. 2013).

Although species classification has come a long way, with different methods incorporated, some errors may arise due to inappropriate classification methods for some genomes of bacteria. A classic example of mislabeled species includes genomes of *Shigella*, which often group with organisms classified as *Escherichia coli* because of their genomic similarity (Lan and Reeves 2002). So the question lies in whether their genomes are similar enough that they should be considered the same species or whether they should be separate species altogether.

In order to understand the discrepancies between the labeling and the genomes of *E.coli* and *Shigella*, it may be important to look at the specific biochemical tests used to classify

genomes into those species. **Table 1.1** lists a few tests used to classify these bacteria (Strockbine et al. 2015). With bacteria, the definitive tests started with the determination of the morphology of individuals. These bacterial species have the same shape. Although few tests and results are present in the table below, it still exemplifies similarities between the different bacteria. If researchers found a strain of *E. coli* that had the same biochemical results as *Shigella*, it could easily be misclassified. The similarity of these results can give an inaccurate classification of individual genomes into a species. This highlights why it is so important to use new methods of species classification for bacterial species, where the genomes are analyzed and compared. This can give researchers more insight into the similarity of individuals and allow for an adequate classification of bacterial species.

<i>Species</i>	<i>shape</i>	<i>Gram staining</i>	<i>Catalase</i>	<i>Oxidative/fermentative</i>
<i>Escherichia coli</i>	rods	Gram negative	positive	Fermentative
<i>Shigella</i>	rods	Gram negative	positive	Fermentative

**Table 1.1.** Four common tests performed in order to classify bacteria as either *Escherichia coli* or *Shigella*. The similarity in the results of these tests provides an explanation as to why there are genomes of *Shigella* that are found in databases under the species of *E. coli*.

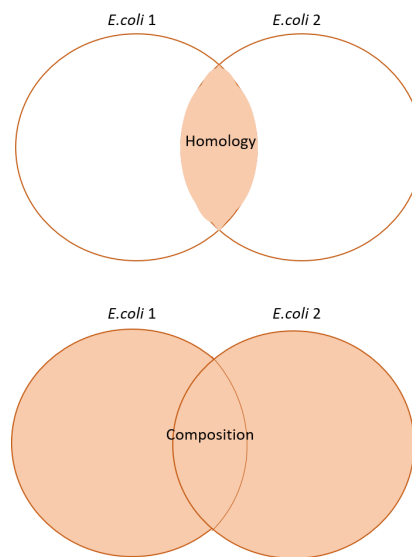
However, in order to understand why these genomes may be classified as the same species, it is crucial to consider their differences. When reading literature about what some of the differences are between the genomes of *E.coli* and *Shigella*, what was found was that one of the main differences between *E.coli* and *Shigella* is that, unlike *E.coli*, *Shigella* cannot ferment

lactose (Devanga Ragupathi et al. 2018). There are four different kinds of *Shigella* found in the database labeled as *E.coli*, including *S.flexneri* and *S. boydii*, which do not contain any *Lac* genes. *S.dysenteriae*, which contains *LacA* and *LacB* but does not contain the *LacZ* gene. The fourth *Shigella* in the database is *S.sonnei* which has all three *Lac* genes but can still not ferment lactose as there is no permease activity (Devanga Ragupathi et al. 2018).

A current method of species classification, based on genome comparison techniques, is to calculate the average nucleotide identity (ANI). ANI is a method based on homology. As the name indicates, ANI is the measure of nucleotide-level genomic similarity between complete genomes; It is a similarity index given to genomes based on the homology. Researchers at Michigan State University and Gent University (Goris et al. 2007)), suggested that an ANI of 95.5% better grouped genomes of the same species together. This threshold is comparable in discriminating power to the 70 percent threshold suggested before for the DNA-DNA hybridization method (Goris et al. 2007). Thus, ANI has been shown to be effective in classifying species. However, while much faster and cheaper than DNA-DNA hybridization, it can still become a bottleneck when working with large databases because it can take months to compare many genomes.

When talking about methods of classification based on composition, the most basic level of composition, GC content, would not be expected to contain enough information for species delimitation. For this reason, higher levels of compositional analysis might be necessary. In 1999, a group of researchers calculated and analysed genomic composition of species. They discovered that each genome has a characteristic “signature” defined as the ratios between the observed dinucleotide frequencies and their expected frequencies given the genomic GC content. The authors found that a comparison of the signatures of different genomes provided a measure of similarity that grouped genomes similarly to what would be expected from a phylogenetic analysis (Campbell et al. 1996).

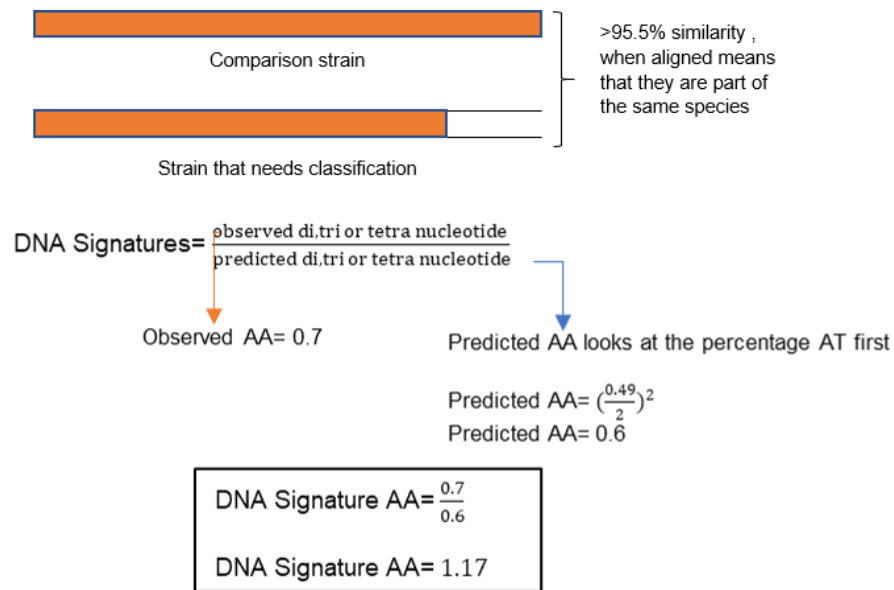
When comparing methods of classification based on homology *versus* genomic composition, it is important to understand what portion of the genomes are being compared (**Figure 1.1**). Methods that are based on homology compare the portions of the genomes that align well, whereas methods based on composition look at the entire genome and compare them. Looking at the whole genome means that there would be more differences found between genomes even if they belonged to the same species (Welch et al. 2002), which is why it is appropriate to be skeptical about this method in terms of bacterial classification.



**Figure 1.1.** Comparison of homology based methods of discrimination to methods based on composition. Homology based methods compare the portions of the genomes that are significantly similar to each other, the intersection, as shown in the top Venn diagram. Genomic composition methods compare the compositions of whole genomes, without alignment.

Based on the results given by Campbell et al. (1996) when using DNA signatures, a measure to group organisms could be suggested by comparing the DNA compositional signature between genomes of the same species. In a paper by Moreno-Hagelsieb et al. (2013), DNA signatures of genomes of the same species were analyzed. The results found that

genomes of the same species condense at a tri-nucleotide signature distance of 0.03 when using Manhattan distances (Moreno-Hagelsieb et al. 2013). The reason that this method was tested for species discrimination is because of its efficiency. Determining differences in DNA signatures is a lot quicker than calculating ANI, because there is no alignment required for calculating and comparing DNA signatures, which is a requirement to determine the ANI between two genomes (**Figure 1.2**).



**Figure 1.2.** Comparison of two methods used in species classification. The first image is the ANI method, which has to find matching regions between genomes to determine how similar they are to one another. The second image is the method of DNA signatures. The signature calculation is exemplified with the dinucleotide AA, where the predicted AA value is derived from the genomic proportion of AT and the observed is learned from the genome itself. The observed value is then divided by the predicted value to give a ratio of 1.17 for the dinucleotide AA for this given genome. The dinucleotide DNA signature consists of the vector of these observed/expected ratios for each dinucleotide. The resulting vectors, DNA signatures, are compared to each other using Manhattan distances.

Michael Richter and Ramon Rosello-Mora compared ANI and tetra-nucleotide composition for genomes of the species *Methanococcus maripaludis* (Richter et al. 2009). They found that tetra-nucleotides could group genomes of the same species, with some fuzziness as this method is, as explained above, alignment-free. In contrast with the tetranucleotide composition that Richter used, the research presented in this thesis focused mainly on using tri-nucleotide DNA signatures to achieve a more heuristic approach to species discrimination. Tri-nucleotide DNA signatures results were also compared to ANI to ensure that accuracy was also kept while improving efficiency.

Up until 2019 DNA signatures were shown to organize all genomes of *Escherichia coli* into a single group, along *Shigella* genomes. This suggested that DNA signatures were an appropriate method of classification for bacterial species. However, in a 2019 genome database update, the tri-nucleotide DNA signatures split the genomes of *E. coli* into three groups at the same, previously established, distance threshold of 0.03. This suggested that either these genomes of *E. coli* should be grouped into three different species, or that DNA signatures might not be as effective in grouping genomes of the same species as previously thought. These differences highlight the issue that different species discrimination methods can give rise to different results and, therefore, inaccurate species classifications. The research conducted for this thesis focused on reclassifying *E.coli*, using different species classification methods, to determine how these genomes should be differentiated and whether methods of classification based on composition do as well as methods of classification based in homology. Doing so will also aid in evaluating the adequacy of DNA signatures for species delimitation.

My masters' thesis aimed at comparing methods of classification based on the composition of genomes versus the homology of genome segments. The comparisons involved 1072 genomes of *Escherichia coli* as classified by the NCBI database combined with the tri-nucleotide signatures, which are part of 2882 genomes belonging to the Enterobacteriaceae

family in the NCBI database. The method used for classification based on composition were di-, tri- and tetra-nucleotide DNA signatures. The methods used for classification based on homology were the average nucleotide identity, 16S rRNA, and 23S rRNA.

## Previous work

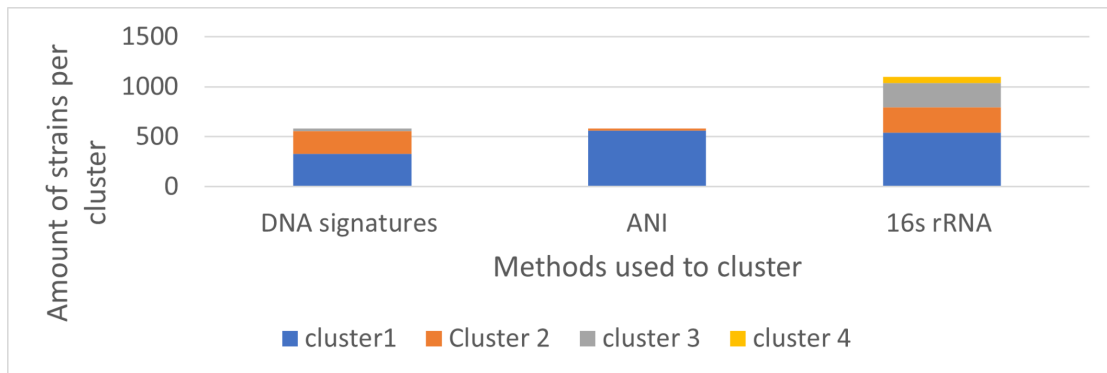
For my undergraduate thesis, classification was done based on three methods; ANI, 16S rRNA, and DNA signatures of all the genomes of *E.coli* available in the NCBI database. All three methods were used on 560 complete genomes, which grouped together with genomes labeled as *Escherichia coli* in the NCBI database in 2018. The threshold for ANI was based on a study by Johan Goris and his colleagues (Goris et al. 2007), where they determined the ANI between several species. The program used to find the ANI of all the genomes was fastANI (Jain et al. 2018). The results for all the ANI were then put into a table and clustered at a similarity of 95.5%. The 16S rRNA sequences for *E.coli* were selected from a collection of all 16S rRNA sequences of all of the bacterial genomes in the RefSeq database (O'Leary et al. 2016). To extract these sequences from the collection, an *ad hoc* program was written. This program was written in python and put all the 16S rRNA in a file in order to cluster them later. The threshold used to determine groups based on the 16S rRNAs was 98.5% (Kim et al. 2014). The third method used to determine the similarity between *E. coli* genomes was distances based on DNA compositional signatures (Campbell et al.1999). A program, written by Dr. Moreno-Hagelsieb, was used to compare the DNA signatures of all *E. coli* genomes to one another. This program grouped genomes based on a distance of 0.03. The number of groups and genomes per group were saved into their respective files for all methods to compare results later.

What was seen based on these comparisons was that each method gave different results (**Figure 1.3**). ANI resulted in 10 different groups, with the largest containing 560



genomes. DNA signatures gave three different groups, with the first containing 324 genomes. Lastly, the 16S rRNA method gave 28 different groups, with the largest having 540 genomes.

Interestingly, in all three methods, there were 24 genomes that grouped out and all labeled as genomes of *E. coli*; This was an unexpected result because genomes labeled as *Shigella* and *Citrobacter* were mixing with those labeled *E.coli* species in the NCBI database. We expected that genomes named differently would be the ones that would group themselves out. The results thus suggested that these 24 genomes were different enough to be considered a different species.



**Figure 1.3.** Comparison of number of clusters and genomes using three methods of species classification. The results show that based on the DNA signatures method there were three clusters found, with the largest containing 324 genomes. The ANI method gave 10 different clusters, with the largest containing 560 genomes and the 16S rRNA method gave 28 different clusters with the largest containing 540 genomes.

In an update of the NCBI database, we found that the 24 consistently problematic genomes had been reclassified as *Citrobacter*, which is consistent with the suggestion that DNA signatures were, like the other methods tested, currently rejecting them from the main *E. coli* groups.

## Objectives

The objectives of my MSc thesis research are as follows:

1. To build hierarchical clusters for all DNA signature groups, containing genomes labeled as *E.coli*, based on the following measures for classification: ANI, DNA signature similarity, as well as 16S rRNA and 23S rRNA gene similarities testifying for more traditional approaches.
2. To determine whether methods of classification based on composition (DNA signatures) are comparable to methods of classification based on homology (ANI).
3. To test the accuracy of these methods for species delimitation against the whole Enterobacteriaceae family.

# Methods

## Selecting genomes

We downloaded complete genomes from NCBI's RefSeq database (O'Leary et al. 2016). Trinucleotide DNA signatures were calculated for all genomes and grouped based on a cutoff of 0.03, previously found to correspond to a species threshold (Moreno-Hagelsieb et al. 2013). The genomes were selected by bringing in all genomes found in a group with at least one genome of *Escherichia coli* in it.

## Average Nucleotide Identity

Average nucleotide identity was calculated using fastANI v. 1.2 (Jain et al 2018). We used a fragment size option of 1020 nucleotides. The same method was also done for all genomes that were labeled as *Escherichia*.

## DNA signatures

DNA signatures are vectors containing the ratio of observed and predicted proportions of each of di, tri, or tetranucleotides (Campbell et al. 1996, Moreno-Hagelsieb et al. 2013). The observed were calculated with a single nucleotide sliding window as published previously (Moreno-Hagelsieb et al. 2013).

## 16S and 23S rRNA

The 16S and 23S rRNA gene sequences were found using infernal v. 1.1.3 (Nawrocki and Eddy, 2013) against the RF00177 (16S rRNA) and RF02541 (23S rRNA) covariance models. The percent identity between all rRNA sequences found were calculated using the vsearch program

v. 2.16 (Rognes et al. 2016). This program works by using a fast heuristic word searching method.

## MASH

Another method used for comparing genomic distances was MASH (Ondov et al. 2019). This method compares genomes based on long-nucleotide composition, which makes it something of an intermediate composition/homology method. MASH uses two primary functions for sequence comparisons, known as sketch and dist. The sketch function converted the collection of sequences into a MinHash sketch—the dist function then compared the sketches and returned an estimate of the Jaccard index. MASH defaults at 1000 sketches per genome. We selected 5000 sketches instead (`mash sketch -s 5000`).

## Dashing

DASHING works very similar to MASH, where it creates sketches in order to compare distances (Baker and Langmead 2019). The difference, however, is that Dashing uses HyperLogLog sketches rather than MinHash sketches. Options other than the defaults for `dashing cmp`, were selected to produce a mash-like result (`--mash-dist`), with a k-mer size of 21 to make it more comparable to MASH (`-k 21`), and the sketch size option that produced the best jaccard-index estimates in the publication ( $2^{14}$ , selected using `-S 14`).

## Selecting cutoffs

After the groups were created, the cutoffs for the family and genus were double-checked using the `cutpointr` R library (Christian Thiele, 2021: <https://CRAN.R-project.org/package=cutpointr>). The library found the optimal cutpoint for each method, including the di, tri, and tetra nucleotides

in DNA signatures; ANI, MASH, and DASH. The program also calculated the different prediction statistics for all methods of classification.

## Hierarchical Clusters

Hierarchical clusters were built using a program written by Dr. Moreno-Hagelsieb. The program is a wrap up that takes the appropriate distance/similarity files for each method as appropriate and produces an R script to produce hierarchical clusters. The program can also cut the hierarchical clusters at selected thresholds. Hierarchical clusters were compared using cluster, MCMCpack, ape and reshape2 packages in R.

# Results

## Selecting *Escherichia coli* genomes

We first selected genomes from the NCBI database, and trinucleotide DNA signatures were calculated for all genomes and grouped based on a cutoff of 0.03. The *E. coli* groups were selected by bringing in all genomes found in a group with at least one genome of *Escherichia coli* in it. The genomes of *E. coli* were found in 3 different groups based on their DNA signatures (**Table 3.1**). The first group contained 527 genomes in total, with 523 of these genomes labeled as *Escherichia coli*. The group also contained three genomes of *Shigella* and one genome of *Salmonella* with no species-level designation. In the second group, there were 544 genomes in total, with 446 of them labeled as *Escherichia* and 101 of them labeled as *Shigella*. In the third group, there were 450 genomes in total, with 448 of the genomes labeled as *Klebsiella*, one of them labeled as *Escherichia coli*, and one labeled as *Enterobacteriaceae bacterium*. We assumed that this group contained a *Klebsiella* genome mislabeled as *E. coli*, thus deciding to ignore this group. Later on, we found that the genome had been reclassified as *Klebsiella pneumoniae* in NCBI.

Cluster 1	Cluster 2	Cluster 3
523 <i>Escherichia coli</i> 1 <i>Salmonella</i> HNK130 1 <i>Shigella boydii</i> 1 <i>Shigella flexneri</i>	444 <i>Escherichia coli</i> 39 <i>Shigella flexneri</i> 23 <i>Shigella dysenteriae</i> 22 <i>Shigella sonnei</i> 14 <i>Shigella boydii</i> 1 <i>Shigella Marmoate</i> 1 <i>Escherichia</i> E4742	399 <i>Klebsiella pneumoniae</i> 22 <i>Klebsiella quasipneumoniae</i> 21 <i>Klebsiella variicola</i> 2 <i>Klebsiella aerogenes</i> 1 <i>Klebsiella quasivariicola</i> 1 <i>Klebsiella</i> PO552 1 <i>Klebsiella</i> P1CD1 1 <i>Klebsiella</i> LY 1 <i>Escherichia coli</i> 1 Enterobacteriaceae unclassified

**Table 3.1.** Species found in the three groups obtained using DNA signatures at a 0.03 threshold. The first two clusters mainly contained genomes labeled as *Escherichia coli*, with the

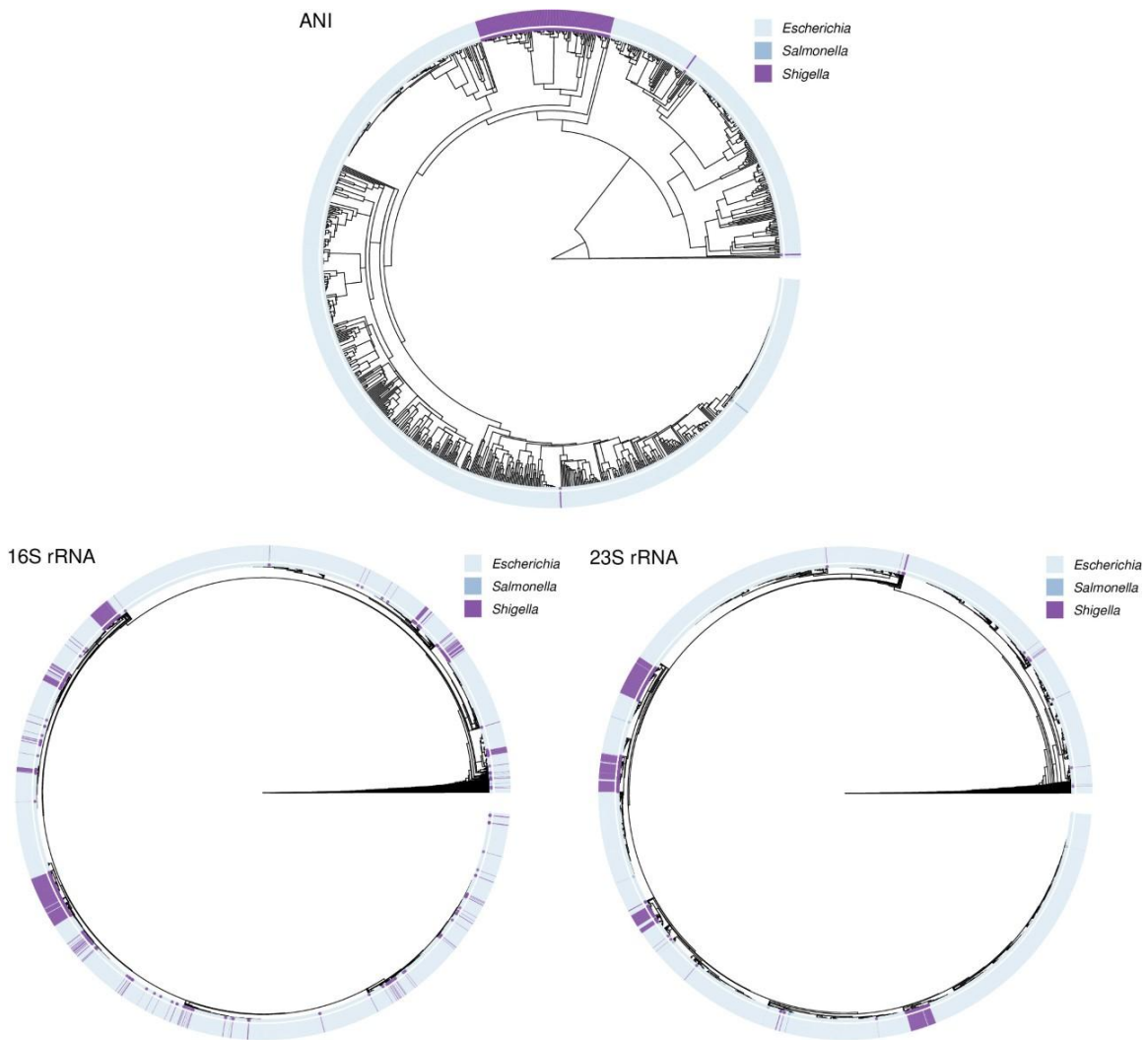
second containing most of the *Shigella* genomes. However, the third group contained only one *E. coli* genome, shown in red. The third cluster was left out of the analyses as it was suspected to be a misclassified genome of *Klebsiella*, which was later confirmed as it was later reclassified as *Klebsiella pneumoniae* in the original database.

## Hierarchical clusters

Once selecting the two proper *E. coli* groups, I produced hierarchical clusters for the remaining 1071 genomes using all the methods of classification (**Figures 3.1, 3.2**).

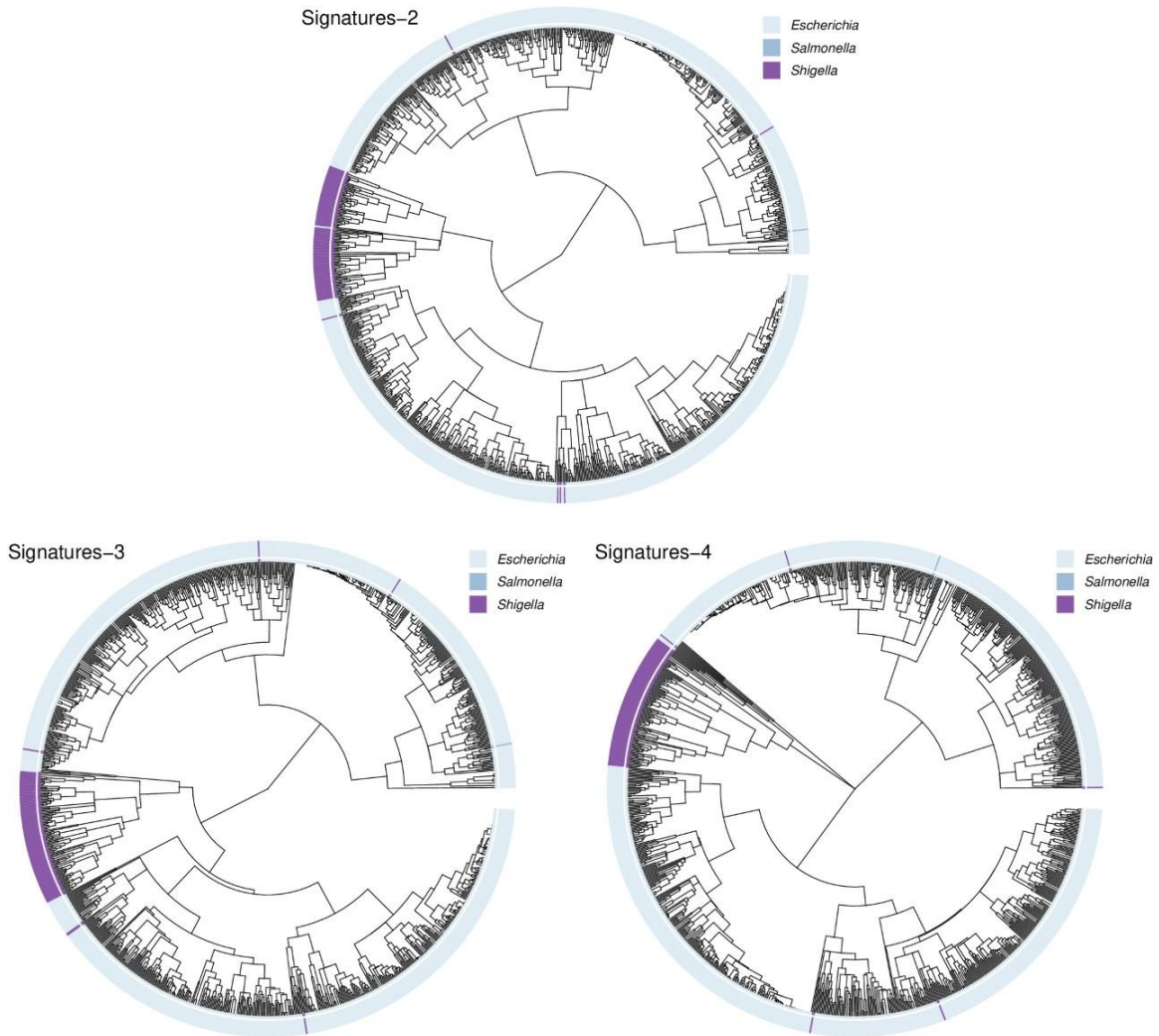
The 16S and 23S rRNA gene clusters displayed many more nodes than any other cluster, because each genome contained several rRNA genes. While identical copies of the rRNA genes, within the same genome, were ignored, that was not enough to keep a single representative sequence per genome (**Figure 3.1**). The abundance of non-identical 16S and 23S rRNA genes makes them a difficult choice for bacterial species classification or delimitation.

After the hierarchical clusters were produced, I noticed that there seemed to be many similarities between the ANI cluster (**Figure 3.1**) and DNA signatures cluster (**Figure 3.2**). In both cases, almost all the genomes of *Shigella* are found together. A numerical analysis, using entanglement, was then done to compare the alignment of the hierarchical cluster, based on ANI to that based on trinucleotide DNA signatures using an entanglement (**Figure 3.3**). The entanglement score is a measure between one and zero, where one represents full entanglement, and 0 represents no entanglement; A lower entanglement coefficient corresponds to a good correspondence between the clusters. The entanglement of 0.08 further shows that both classification methods, ANI and DNA signatures, produce very similar results.

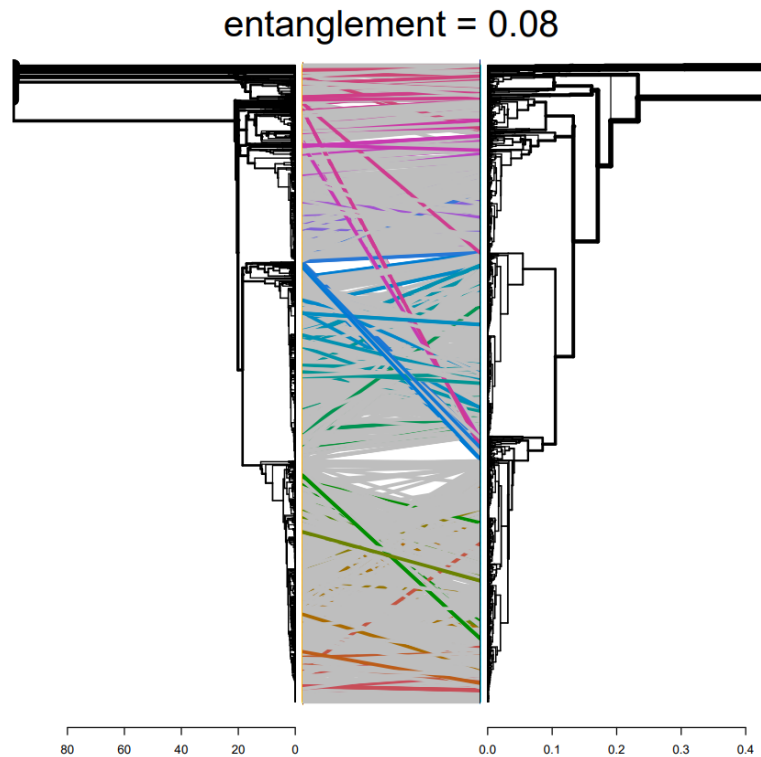


**Figure 3.1.** Hierarchical clusters based on homology methods of classification, including ANI 16S rRNA and 23S rRNA. It is interesting to note that ANI clustered most *Shigella* genomes together into a single group. The *Shigella* 16S and 23S rRNA genes did not group together as clearly.





**Figure 3.2.** Hierarchical clusters based on compositional methods of classification, including di, tri and tetranucleotide DNA signatures. All three methods grouped most *Shigella* genomes close together.



**Figure 3.3.** Tanglegram representing similarities between hierarchical clusters of ANI and tri-nucleotide DNA signatures. On the right is the hierarchical cluster for ANI and on the left is the hierarchical cluster for DNA signatures. The number on top is an entanglement coefficient which corresponds to how well the two clusters align to each other. The low entanglement of 0.08 indicates that the results of the two methods are very similar.

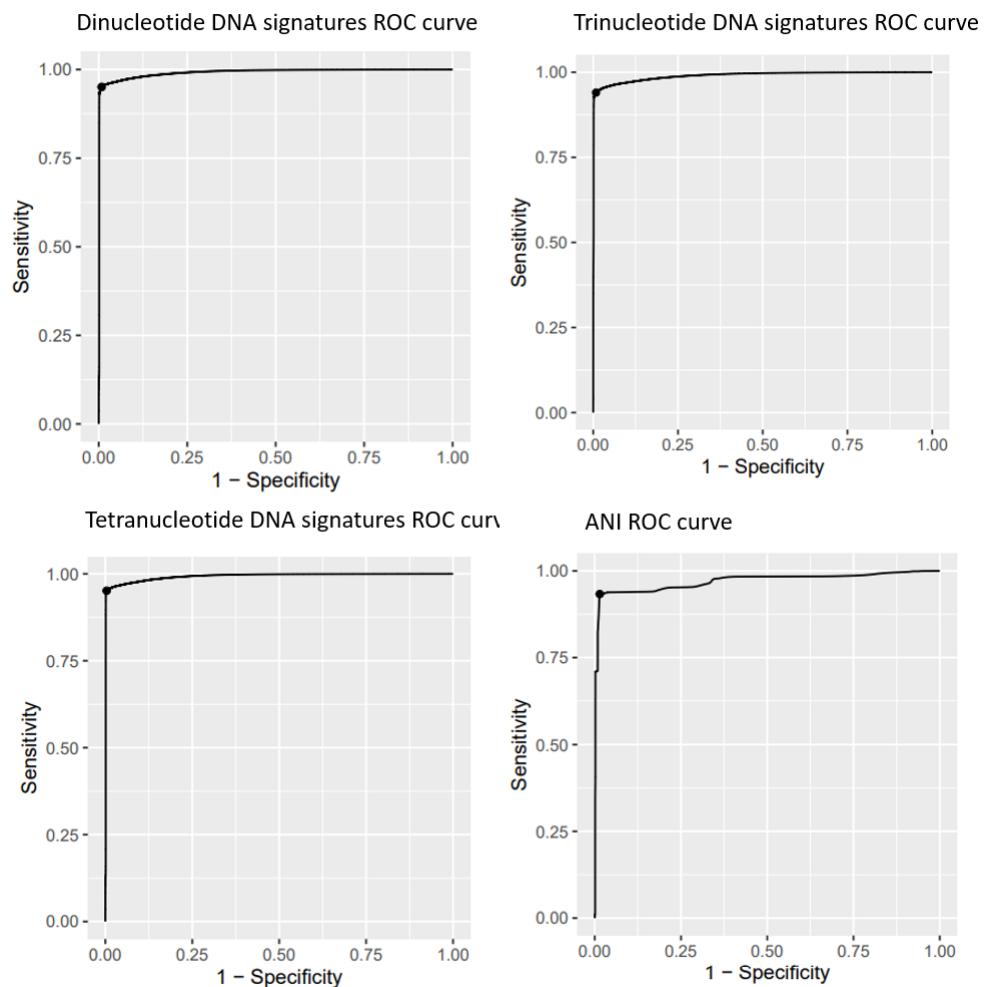
## Classification quality

Receiver operating characteristic (ROC) curves were produced where the positive data set consisted of pairs of genomes classified into the same species, while negatives consisted of, either pairs of genomes classified in the same family, but different species; or pairs of genomes classified in the same genus, but different species. A ROC curve is a performance measurement that is often used for classification problems testing different thresholds. These graphs are plotted with the true positive rate (sensitivity) on the y- axis, against the false positive

rate (1 - specificity) on the x-axis. The area under the curve (AUC) represents the accuracy of the method for classification.

The graphs in **figure 3.4** highlight that DNA signatures can accurately differentiate between different groups when using same-family pairs as negatives. This is seen due to the high value for the area under the curve (AUC), which ranges from 0.9899 for dinucleotide signatures all the way to 0.9929 for the tetranucleotide signatures. The AUC for DNA signatures are comparable to the AUC for ANI, which showed an AUC of 0.9903.

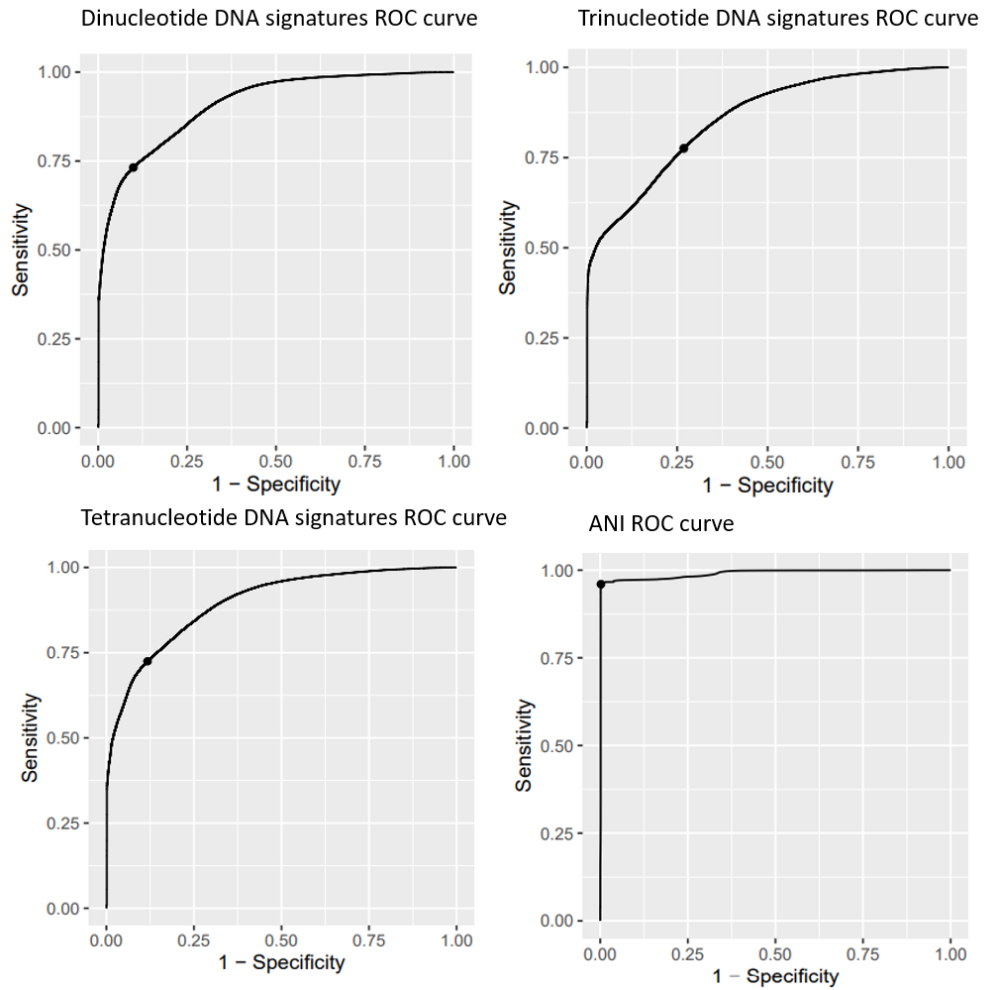
ANI can accurately differentiate species when tested using same-genus genomes as a negative dataset, as seen on the ROC curve with an AUC of 0.97 (**Figure 3.4**). ROC curves for di, tri, and tetranucleotide DNA signatures, with same-genus as negative dataset (**Figure 3.5**) shows AUC of 0.8555 for dinucleotide signatures, 0.8962 for tri-nucleotide signatures, and 0.9081 for tetra-nucleotide signatures.



Method	AUC	Opt. cutpoint	sensitivity	specificity	accuracy
<b>DNA sig - 2</b>	0.9899	0.012	0.9402	0.9914	0.9517
<b>DNA sig - 3</b>	0.9919	0.0206	0.9508	0.9917	0.9616
<b>DNA sig - 4</b>	0.9929	0.0325	0.9516	0.9961	0.9616
<b>ANI</b>	0.9903	7.955	0.9601	0.9979	0.9686

**Figure 3.4.** ROC curves for DNA signatures and ANI, with same-family genome pairs used as negative datasets. A ROC curve is a performance measurement that is often used for

classification problems testing different thresholds. These graphs are plotted with the true positive rate (sensitivity) on the y- axis, against the false positive rate (1 - specificity) on the x-axis. The area under the curve (AUC) represents the accuracy of the method for classification. The AUC suggested that all methods of classification, DNA signatures and ANI, were able to differentiate between species within the same taxonomic family.



Method	AUC	Opt. cutpoint	sensitivity	specificity	accuracy
DNA sig - 2	0.8555	0.0056	0.7755	0.7315	0.7353

<b>DNA sig - 3</b>	0.8962	0.0123	0.7247	0.8823	0.8689
<b>DNA sig - 4</b>	0.9081	0.0189	0.7316	0.9011	0.8867
<b>ANI</b>	0.97	5.12	0.9332	0.9853	0.9808

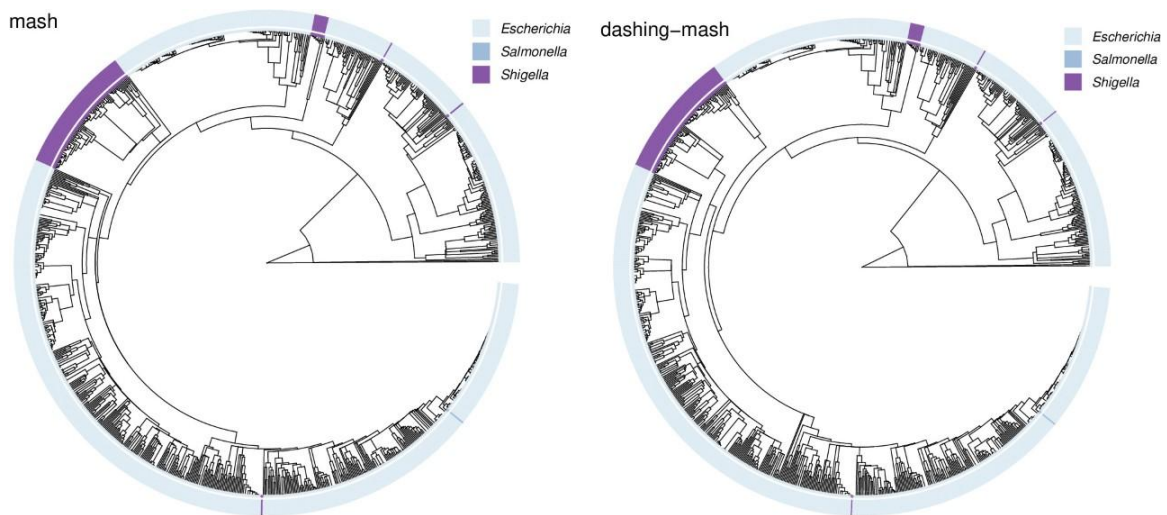
**Figure 3.5.** ROC curves for DNA signatures and ANI, with same-genus genome pairs used as a negative dataset. The AUCs were lower than those displayed in **Figure 3.4**. The AUCs for DNA signatures were not as high as the AUC for ANI, meaning that DNA signatures were not as good at differentiating between species of the same genus as ANI.

## Oligonucleotide methods for species delimitation

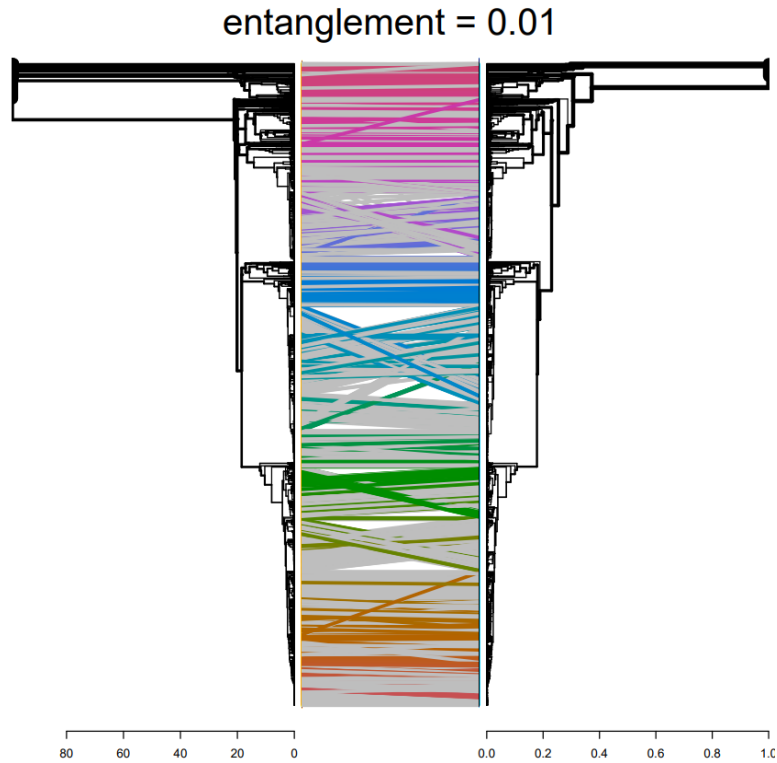
Initially, this work did not contemplate testing what we call “intermediate” classification methods, like MASH (Ondov et al. 2019) and DASHING (Baker, D.N. & Langmead, B. 2019). We call them intermediate because they compare oligonucleotide content, where the oligonucleotides, the k-mers, are much longer than four (defaults of 21 and 31, respectively), and thus approach a homology-based comparison. Given the results with the ROC analyses above, where DNA signatures gave results of lesser quality compared to ANI, I thought it worth testing these other methods. **Figure 3.6** highlights the hierarchical clusters for both methods.

As observed, both methods cluster *Shigella* genomes in a similar way as ANI. The hierarchical clusters were also compared against the ANI cluster using tanglegrams. **Figure 3.7** shows the tanglegram for ANI on the right against MASH on the left. This figure highlights how similar the hierarchical clusters for both these methods are due to the entanglement score of 0.01. **Figure 3.8** shows the tanglegram of ANI on the right against Dashing MASH on the left. This figure highlights the difference between these two programs, seen by an entanglement score of 0.6.

The ROC curves for MASH and DASHING, using same-genus genome pairs as negative datasets, are shown in **Figure 3.9** highlight the accuracy of both MASH and Dashing, based on the high values for the AUC.

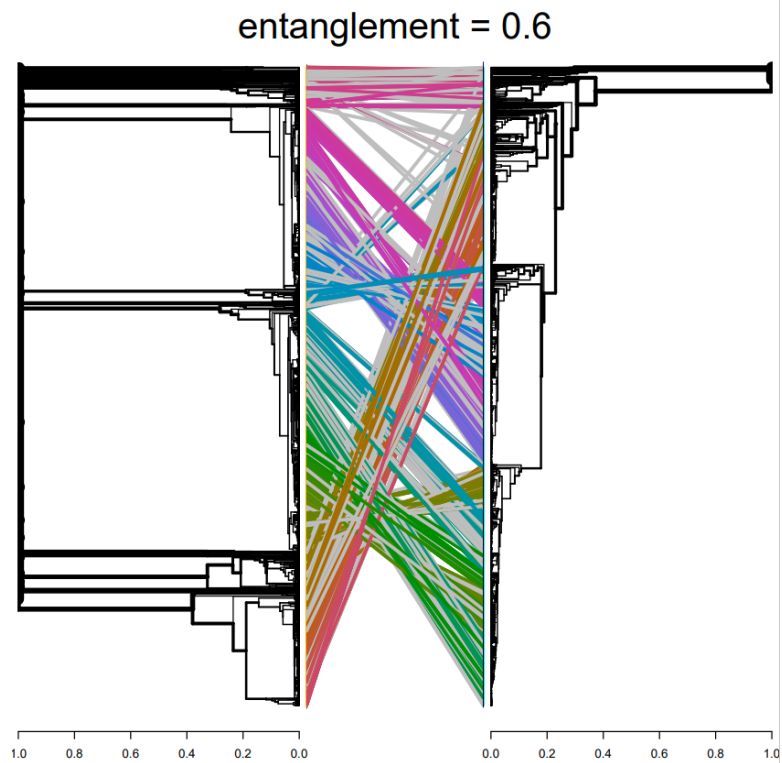


**Figure 3.6.** Hierarchical clusters based on MASH and Dashing. Both methods kept most of the *Shigella* genomes into the same group.

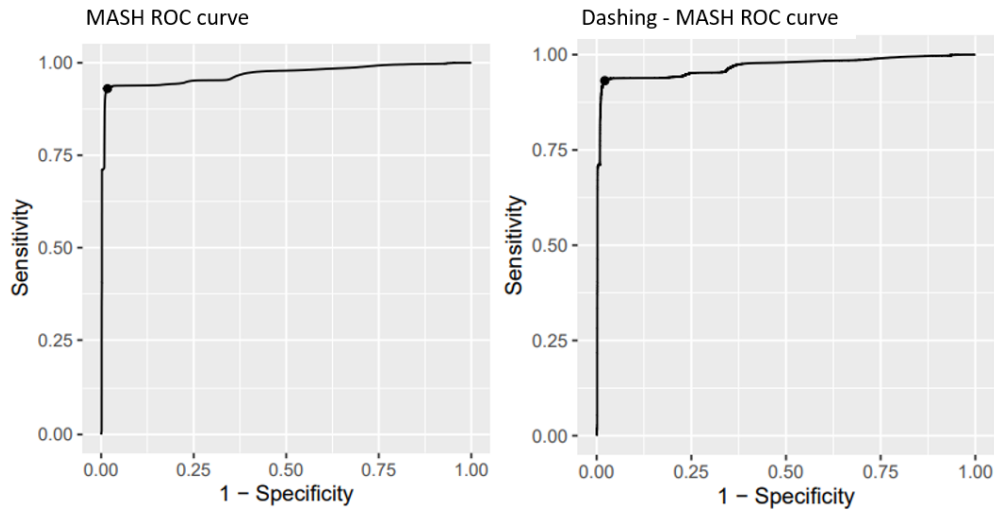


**Figure 3.7.** Tanglegram of ANI vs MASH, where the hierarchical cluster of ANI is seen on the right side and the hierarchical cluster of MASH is seen on the left. The entanglement coefficient seen here is 0.01, indicating that the results of the hierarchical clusters for both MASH and ANI are very similar to one another, and almost identical.





**Figure 3.8.** Tanglegram of ANI vs Dashing-MASH, where the hierarchical cluster of ANI is seen on the right side and the hierarchical cluster of Dashing-MASH is seen on the left. The entanglement coefficient seen here is 0.6, indicating that the results for the hierarchical clusters of Dashing and ANI are not well aligned, meaning the methods group them very differently from one another.



Method	AUC	Opt. cutpoint	sensitivity	specificity	accuracy
MASH	0.9683	0.0477	0.9301	0.9827	0.9782
Dashing - MASH	0.9682	0.0465	0.9317	0.9783	0.9744

**Figure 3.9.** ROC curves for MASH and Dashing-Mash, with same-genus genome pairs used as a negative dataset. What is seen is that both had an AUC of 0.96, indicating that both methods are effective in differentiating between genus and species, where genus is the negative dataset.

Since all methods produced very similar AUC values with the same-family negative datasets, I further analysed the performance of these methods using the same-genus negative dataset, where more differences could be expected. Thus, the hierarchical clusters containing all of the *Enterobacteriaceae* genomes were cut into groups after determining the optimal cutoffs for each method using the same-genus negative dataset. An analysis was done of the resulting groups to determine how the *Shigella* and *E.coli* grouped (**Table 3.3**). What I noticed is that, based on these cutoffs, the signature-based thresholds divided the *Enterobacteriaceae* into

more groups than other methods, with ANI, mash and dashing keeping all of the genomes of *E.coli* and *Shigella* into a single group. This table highlights the discrepancies between the method of DNA signatures and ANI when tested with organisms of the same genus as negative datasets.

<b>Method: threshold</b>	<b>Number of groups</b>	<b>Largest number of <i>Shigella</i> in a group (out of 101)</b>	<b>Largest number of <i>E.coli</i> in a group (out of 967)</b>
ANI: 5.12	155	101 (group 1)	967 (group 1)
Dinucleotide: 0.0056	190	40 (group 5)	195 (group 6)
Trinucleotide: 0.0123	159	51 (group 2)	276 (group 12)
Tetranucleotide: 0.0189	157	64 (group 2)	254 (group 11)
MASH: 0.0477	158	101 (group 1)	967 (group 1)
Dashing: 0.0465	185	101 (group 10)	967 (group 10)

**Table 3.3.** Amount of genomes of *Shigella* and *E.coli* found in a single group based on different methods. These thresholds were determined by comparing same-species genome pairs as positive datasets against same-genus genome pairs as negative datasets as indicated by the ROC curves.

# Discussion

A heuristic technique is an approach to a problem that employs a method that is not guaranteed to be perfect, but rather good enough to reach an immediate short-term goal. When considering DNA signatures as a method for bacterial classification, this is precisely the word that comes to mind. The definition and identification of species have been a challenge in Biological sciences for centuries. As new methods of bacterial classification are introduced, the classification of species continues to change as well. My thesis looked at multiple classification methods to determine whether methods based on genomic composition, specifically DNA signatures, are as effective in classifying bacterial species as methods based on homology (ANI). One of the main reasons to test methods of classification based on genomic composition is to approach bacterial classification with a heuristic model in mind.

The first step in these analyses was genomes selection from the NCBI database. There were 101 genomes of *Shigella* in the NCBI database that were grouped with *E.coli*. In order to understand why these genomes may be classified as the same species, it is crucial to consider their similarities and differences. When reading literature about what some of the differences are between the genomes of *E.coli* and *Shigella*, what was found was that one of the main differences between *E.coli* and *Shigella* is that, unlike *E.coli*, *Shigella* cannot ferment lactose (Devanga Ragupathi et al. 2018). There are four different kinds of *Shigella* found in the database labeled as *E.coli*, including *S.flexneri* and *S. boydii*, which do not contain any *Lac* genes. *S.dysenteriae*, which contains *LacA* and *LacB* but does not contain the *LacZ* gene. The fourth *Shigella* in the database is *S.sonnei* which has all three *Lac* genes but can still not ferment lactose as there is no permease activity (Devanga Ragupathi et al. 2018). Although their genomes may be similar, the differences are in parts of the genome that contain functional genes and that belong to the core genome, thus suggesting that *Shigella* strains should be classified as a different species.

The methods were all tested on genomes that were grouped with those classified as *Escherichia coli* according to the NCBI database. Interestingly, the DNA signatures and ANI hierarchical clusters were very similar, as they grouped 97 of the 101 genomes of *Shigella* together into a separate node (**Figure 3.1, Figure 3.2**). Therefore, I tested the similarity of these clusters based on a tanglegram (**Figure 3.3**). The tanglegram for this comparison produced an entanglement score of 0.08. An entanglement score, again, gives a measure to how well aligned the two clusters are to one another (An introduction to cutpoint. (n.d.)). The low entanglement score between the trinucleotide DNA signatures and ANI hierarchical clusters shows that the two are very similar. This provides evidence suggesting that DNA signatures are as effective in clustering genomes of bacterial species as ANI.

After comparing the hierarchical clusters, ROC curves were produced for di, tri, and tetranucleotide signatures (**Figure 3.4**). A ROC curve is a performance measurement that is often used for classification problems at various thresholds. The graphs are plotted with the true positive rate (sensitivity) against the false positive rate (1 - specificity) to show the area under the curve and the optimal threshold. The ROC curves for the DNA signatures highlight that this method can differentiate between genomes from organisms of the same species and other organisms in the same taxonomic family; this is seen due to the high area under the curve (AUC), ranging from 0.9899 for dinucleotides to 0.9929 for tetranucleotides. The AUC for ANI, using the same datasets, showed that it could differentiate between species accurately, with an AUC of 0.9903.

The results above are very similar. However, a more precise comparison was made producing ROC curves for DNA signatures and ANI, but this time checking the separation between genomes of organisms of the same species against organisms within the same genus. This time around, DNA signatures could not discriminate between the genus and species as accurately as ANI (**Figures 3.5**). ANI produced an AUC of 0.97, while for the DNA signatures,

the range was between 0.8555 for dinucleotides and 0.9081 for tetranucleotide DNA signatures. Although these are still high AUCs, they are not as good as the one for ANI.

After seeing that DNA signatures could not discriminate as well between genus and species, I tested two other methods. The first method was MASH; this method is somewhat in between composition and homology regarding how it classifies species. MASH uses a MinHash algorithm, effectively eliminating the resemblance of two genomes or metagenomes (Ondov et al. 2019). Initially, this method was left out of consideration when it came to picking strategies for bacterial classification. We thought it would be slower than DNA signatures as it is an in-between method based on composition and homology. What I found was that this method was quicker than DNA signatures and much quicker than ANI.

The first thing done was that a hierarchical cluster was created for MASH, as seen in **figure 3.6**. What was seen is that the hierarchical cluster grouped the majority of *Shigella* genomes together into a separate node. Based on the similarity of this cluster to the one produced by ANI, another tanglegram was made, as seen in **figure 3.7**. This time on the right, there was ANI, and on the left, there was the hierarchical cluster for MASH. What was seen in this comparison was that the two hierarchical clusters were well aligned, proven by the entanglement score of 0.01. After testing the similarity between ANI and MASH hierarchical clusters, ROC curves were created for MASH for genus versus species. What was seen in this curve is that this method worked very well when discriminating between genus and species, proven by the AUC of 0.9683.

One more tested method was Dashing; this method is also used for estimating genomes or sequence datasets. The difference is that Dashing sketches genomes more rapidly than previous MinHash-based methods, such as MASH, while still providing greater accuracy. Rather than a MinHash method for sketching, Dashing uses a HyperLogLog sketch (Baker and Langmead 2019). The first thing was that a hierarchical cluster was created for Dashing - MASH, as seen in **figure 3.6**. What was seen in this hierarchical cluster was that, yet again, the

majority of the genomes of *Shigella* were grouped together in a separate node. Based on this similarity, another tanglegram was created, as seen in **figure 3.8**. This time on the left, there was the hierarchical cluster for Dashing, and on the right was the hierarchical cluster for ANI. What was seen in this comparison was that the two hierarchical clusters were not well aligned, as seen by an entanglement score of 0.6, meaning that there was much entanglement. After testing the similarity between ANI and Dashing hierarchical clusters, ROC curves were created for Dashing-MASH for the same-genus negative dataset. What was seen in this curve is that this method worked very well when discriminating between genus and species, proven by the AUC of 0.9682 (**figure 3.9**). This method has the same accuracy as MASH does in terms of discriminating between genomes when it comes to comparing the genus against the species. The difference, however, is that Dashing is a lot faster as it uses the HyperLogLog sketches, and it does this in the same step that it separates the genome. Whereas for MASH, the sketches need to be done in a separate, initial step.

After testing all these methods, it was seen that they gave different results, which is what was expected. Hierarchical clusters gave a clearer understanding of where genomes of *Shigella* were grouped and if there was a way to separate them from all the other genomes of *E.coli*. Based on the hierarchical clusters of all the methods, it is evident that one method that gave clean results in how it grouped the genomes of *Shigella* was the method of DNA signatures. It grouped 97 of the 101 genomes of *Shigella* in one group. **Table 3.3** shows a comparison of the number of groups found based on each DNA signature method and at different thresholds and the number of genomes of *Shigella* found grouped. This table highlights the fact that one group based on trinucleotide DNA signatures captured 94% of the genomes of *Shigella* found in the NCBI database, listed as *E.coli* in one group, suggesting that based on this method, *Shigella* should be separated into different species.

Based on these results, it is evident that DNA signatures are an acceptable classification method, as it organizes genomes of *Shigella* into a separate group, just as ANI does. This also

suggests that *Shigella* should be separated from the *Escherichia* species even though their genomes are very similar. Based on ROC curves, it also suggests that DNA signatures should be used as a method of classification to discriminate between the family against the species. When discriminating between the genus and the species, DNA signatures do not seem as adequate based on an area under the curve of 89%, compared to the ANI method, which has an area under the curve of 97%. When discriminating between genus and species, it is more effective to use a classification method such as MASH or Dashing. Both these methods were seen to be more effective than ANI when discriminating between genus and species, and both were seen to be doing so with much more efficiency than the latter.

At present, *Shigella* and *Escherichia* genera are considered to be unique species based on their genotypes. Unlike *E.coli*, *Shigella* genomes are nonmotile due to a deletion in the *fliF* operon or an ISI insertion mutation in the *flhD* operon. *Shigella* also does not ferment lactose, as genomes of *S. flexneri* and *S. boydii* do not contain any *lac* genes required for fermentation. *S.dysenteriae* is known to have some *lac* genes (*lacY* and *lacZ*) but is lacking the *lacA* gene, which is required for fermentation. *S.sonnei* has all three of the *lac* genes but is still unable to ferment lactose as a result of there being no permease activity. These observations have led researchers to believe that *Shigella* originated as a result of convergent evolution (Ragupathi et al.2018). Going forward, I think it is important to check whether the *lac* gene is why DNA signatures grouped the genomes of *Shigella* together. If this is the case, it would confirm again that DNA signatures are an excellent classification method. Another thing to confirm would be whether the *Lac* genes would be found in the core genome. This is important because if they are found in genes that are a part of the core genome, this would suggest yet again that *Shigella* and *E.coli* should be part of different species rather than being considered the same.

I suggest using DNA signatures or other compositional classification methods when discriminating between the family and species to classify species as it is a lot quicker. If discrimination needs to be done between genus and species, methods that use homologies



such as ANI or some intermediate between genomic composition and homology can be used, such as MASH or Dashing. The best method of classification based on the results seen here is a combination of ones that look at just the composition and ones that also consider the homology.

Going forward, trinucleotide DNA signatures, MASH, and Dashing should be tested on all bacterial species, and these three methods should be compared to one another on a larger scale. This will ensure that these methods are consistent with one another and make sure that the produced results are accurate and efficient.

Over the years, bacterial species classification has moved away from more lab-intensive methods due to many different reasons. One of those is that it can be a prolonged approach to classifying species, and it cannot be used on non-culturable cells (Franco-Duarte, 2019). As classification methods are constantly changing, as they have throughout the years, it is essential to consider methods of classification based on genomic composition as it offers a heuristic approach to bacterial species classification. A new method of bacterial classification based on both genomic composition and homology that is being proposed is the Naïve Bayes hybrid model, which takes the intersection of the predictions produced by two classifiers to produce a high-confidence set of predictions in which classified fragments are rarely incorrect and often assigned to the most appropriate taxonomic rank given the available set of reference genomes (Parks, 2011). This method incorporates genomic composition into the classification of bacterial species to achieve a more accurate and efficient method of bacterial classification. This again proves that there should be more research conducted to determine better classification methods in bacterial species that can be done quickly; this involves looking at methods based on genomic composition or an intermediate between homology and composition.

# Conclusions

As classification methods and technologies have changed throughout the years, the focus has changed from using biochemical tests to genome-sequence methods, and then from using methods that are accurate to using methods that take a heuristic approach to bacterial classification. One of the primary shifts has been in using methods that consider the genomic composition instead of methods that focus on just the homology of genomes. My thesis focused on comparing these two different methods of bacterial discrimination on genomes of *Escherichia coli*, using DNA signatures, a compositional method, against ANI, a method based on homology. My findings showed that DNA signatures could accurately and efficiently discriminate between genomes of *Escherichia coli* and organize them into a hierarchical cluster, almost as well as ANI and at a fraction of the time. Going forward, research should focus on using methods of classification based on composition or a mix of both composition and homology; one example of this is the program MASH. This program was an intermediate between ANI and DNA signatures and worked at a fraction of the time that DNA signatures did. This further proved how genomic composition could be used as a heuristic approach to species classification. As methods of classification progress, the goals have also shifted to make methods not just accurate but also efficient; with that, it is essential to consider methods based on genomic composition.

## Integrative Biology statement

What does it mean to be integrative? The word by definition means to unify separate things. Integrative biology to me means to use multiple skills and approaches to solve biological problems. My thesis was very representative of integrative biology, as it was using programming skills in order to answer questions pertaining to the biological sciences, in terms of species definition.

# References

An introduction to cutpointr. (n.d.). Retrieved June 2, 2021, from <https://cran.r-project.org/web/packages/cutpointr/vignettes/cutpointr.html>

Baker DN, Langmead B (2019). Dashing: Fast and accurate genomic distances with HyperLogLog. *Genome Biology*, 20(1), 1–12. <https://doi.org/10.1186/s13059-019-1875-0>

Bohlin J, Eldholm V, Pettersson JHO, Brynildsrud O, Snipen L (2017). The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. *BMC Genomics*, 18(1), 1–11. <https://doi.org/10.1186/s12864-017-3543-7>

Campbell A, Mrázek J, Karlin S (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc Natl Acad Sci USA*, 96(16), 9184–9189. <https://doi.org/10.1073/pnas.96.16.9184>

Caputo A, Merhej V, Georgiades K, Fournier PE, Croce O, Robert C, Raoult D (2015). Pan-genomic analysis to redefine species and subspecies based on quantum discontinuous variation: The *Klebsiella* paradigm. *Biology Direct*. <https://doi.org/10.1186/s13062-015-0085-2>

Devanga Ragupathi NK, Muthuirulandi Sethuvel DP, Inbanathan FY, Veeraraghavan B (2018). Accurate differentiation of *Escherichia coli* and *Shigella* serogroups: challenges and strategies. *New Microbes and New Infections*, 21, 58–62. <https://doi.org/10.1016/j.nmni.2017.09.003>

Franco-Duarte R, Cernakova L, Kadam S, Kaushik KS, Salehi B, Bevilacqua A, Corbo MR, Antolak H, Dybka-Stepien K, Leszczewicz M, Relison Tintino S, Alexandrino de Souza VC, Sharifi-Rad J, Coutinho HDM, Martins N, Rodrigues CF (2019). Advances in chemical and

biological methods to identify microorganisms—from past to present. *Microorganisms*, 7, 130. <https://doi.org/10.3390/microorganisms7050130>

Goris J, Konstantinidis KT, Klappenbach JA, Coeyne T, Vandamme P, Tidge JM (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, 57, 81-91. Retrieved from [www.microbiologyresearch.org](http://www.microbiologyresearch.org)

Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, 9, 5114. <https://doi.org/10.1038/s41467-018-07641-9>

Kim H, Oh HS, Park SC, Chun J (2014). Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *International Journal of Systematic and Evolutionary Microbiology*, 64, pp. 346-35. <https://doi.org/10.1099/ijs.0.059774-0>

Lan R, Reeves PR (2002). *Escherichia coli* in disguise: molecular origins of *Shigella*. *Microbes and Infection*, 4(11), pp. 1125–1132. [https://doi.org/10.1016/S1286-4579\(02\)01637-4](https://doi.org/10.1016/S1286-4579(02)01637-4)

Lorén JG, Farfán M, Fusté MC (2018). Species Delimitation, Phylogenetic Relationships, and Temporal Divergence Model in the Genus *Aeromonas*. *Frontiers in microbiology*, 9, 770. <https://doi.org/10.3389/fmicb.2018.00770>

Moreno-Hagelsieb G, Wang Z, Walsh S, Elsherbiny A (2013). Phylogenomic clustering for selecting non-redundant genomes for comparative genomics. *Bioinformatics*, 29(7), pp. 947-949. <https://doi.org/10.1093/bioinformatics/btt064>

Nawrocki EP, Eddy SR (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>

O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*, 44(D1), D733-4. <https://doi.org/10.1093/nar/gkv1189>

Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biology*, 20(1), 1–13. <https://doi.org/10.1186/s13059-019-1841-x>

Parks DH, MacDonald NJ, Beiko RG (2011). Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, 12(1), 1–16. <https://doi.org/10.1186/1471-2105-12-328>

Richter M, Rosello-Mora R (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 106 (45) 19126-19131. <https://doi.org/10.1073/pnas.0906412106>

Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4, e2584. <https://doi.org/10.7717/peerj.2584>

Strockbine NA, Maurelli AT (2015). Bergey's manual of systematics of archaea and bacteria. *Wiley Online library*. <https://doi-org.libproxy.wlu.ca/10.1002/9781118960608.gbm01168>

Sheutz F, Strockbine NA (2015). *Escherichia*. *Bergey's Manual of Systematics of Archaea and Bacteria*. <https://doi.org/10.1002/9781118960608.gbm01147>

Thompson CC, Chimetto L, Edwards RA, Swings J, Stackebrandt E, Thompson FL (2013). Microbial genomic taxonomy. *BMC Genomics*. <https://doi.org/10.1186/1471-2164-14-913>.

Welch RA, Burland V, Plunkett G, 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Sonnenberg MS, Blattner FR (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* 99:17020-4. Doi: <https://doi.org/10.1073/pnas.252529799>

Wiens JJ (2007). Species Delimitation: New Approaches for Discovering Diversity . *Systematic Biology*, 56 (6), pp. 875-878. <https://doi.org/10.1080/10635150701748506>

# Appendix

<b>Program</b>	<b>AUC</b>	<b>n</b>	<b>n_pos</b>	<b>n_neg</b>	<b>optimal cutpoint</b>	<b>acc</b>	<b>sensitivity</b>	<b>specificity</b>
MASH	0.9893	3943836	3057146	886689	0.0666	0.9686	0.9602	0.9977
ANI	0.9903	3943836	3057147	886689	7.955	0.9686	0.9601	0.9979
Signatures-3	0.9919	3943836	3057147	886689	0.0206	0.96	0.9508	0.9917
Signatures-2	0.9899	3943836	3057147	886689	0.012	0.9517	0.9402	0.9914
Signatures-4	0.9929	3943836	3057147	886689	0.0325	0.9616	0.9516	0.9961
Dashing- MASH	0.99	3943836	3057147	886689	0.0653	0.9684	0.96	0.9975
Dashing- Jaccard	0.99	3943836	3057147	886689	0.8622	0.9684	0.96	0.9975

**Table A.1.** ROC curve data based on all the different methods of discrimination used, when comparing Family versus species.

<b>Program</b>	<b>AUC</b>	<b>n</b>	<b>n_pos</b>	<b>n_neg</b>	<b>optimal cutpoint</b>	<b>acc</b>	<b>sensitivity</b>	<b>specificity</b>
Signature-3	0.8962	969356	82667	886689	0.0123	0.8689	0.7247	0.8823
ANI	0.97	969356	82667	886689	5.12	0.9808	0.9332	0.9853
Signature-4	0.9081	969356	82667	886689	0.0189	0.8867	0.7316	0.9011
Signature-2	0.8555	969356	82667	886689	0.0056	0.7353	0.7755	0.7315
MASH	0.9683	969356	82667	886689	0.0477	0.9782	0.9301	0.9827
Dashing- MASH	0.9682	969356	82667	886689	0.0465	0.9744	0.9317	0.9783
Dashing- Jaccard	0.9682	969356	82667	886689	0.7748	0.9744	0.9317	0.9783

**Table A.2.** ROC curve data based on all the different methods of discrimination used when comparing Genus versus species.



DNA signature distance cutoffs	Dinucleotide DNA signature	Trinucleotide DNA signature	Tetranucleotide DNA Signature
0.01	5	20	58
0.02	2	3	12
0.03	1	2	4
0.04	1	1	4
0.05	1	1	1
0.06	1	1	1
0.07	1	1	1
0.08	1	1	1
0.09	1	1	1
0.1	1	1	1

**Table A.3.** Number of Groups at different distances for di, tri and tetra nucleotide DNA signatures.