2024

Genome-Based Pathogenicity Potential of Salmonella Isolated from Diverse Sources

Jared MR Crocco
Wilfrid Laurier University, croc0500@mylaurier.ca

Genome-Based Pathogenicity Potential of *Salmonella* Isolated from Diverse Sources

by

Jared Mitchell Rong Crocco

Honours Bachelor of Science in Biology and Chemistry, Wilfrid Laurier University, 2020

THESIS

Submitted to the Department of Biology

Faculty of Science

in partial fulfilment of the requirements for the

Master of Science in Integrative Biology

Wilfrid Laurier University

Abstract

Bacterial human pathogens are among the leading causes of death around the world, especially in low income and developing countries. One important element in a bacterium's ability to cause disease are genes that directly contribute to pathogenicity called virulence factors. A second significant aspect are antimicrobial resistance genes which allow microorganisms to persist in the presence of antimicrobial agents. In this project I aimed to determine if *Salmonella* isolated from different sources differed in pathogenicity profiles based on the complement of genes identified through genomic analysis. Accordingly, *Salmonella* genomes were organized into 8 groups: animal, clinical, human, environmental, food, water source, plant, and nut. A negative control, consisting primarily of non-pathogenic *E. coli*, was also included. To determine disease-causing potential, the proteins encoded by these genomes were compared against the Virulence Factor Database (VFDB), the PathFam database, and the Comprehensive Antimicrobial Database (CARD). The negative controls coded for significantly fewer proteins matching the VFDB and PathFams, but significantly more matching the CARD, than all other groups. Though visibly overlapping, most isolation sources were found to be significantly different to each other (p value < 0.05), aside from the very small nut and plant groups. When clustered by their specific matches to the VFDB and CARD, genomes from the same environmental groups did not cluster together. Therefore, while the groups were statistically different from each other in number of matches, those differences were not due to group-specific virulence factors. Though most isolation source groups were found to be significantly different in VFDB, CARD and PathFam matches, further analyses are needed to determine if that difference is large enough to influence

Salmonella’s disease-causing potential. Methods and results from this analysis can be built upon in the future to better identify potential pathogens isolated from different environments.

Acknowledgements

I would like to extend my deepest gratitude to my supervisor Dr. Gabriel Moreno-Hagelsieb for his support throughout the completion of my thesis. He was always willing to help me on problems I was stuck on, and passionate to talk about ideas and questions I had. He always had a smile on his face during every meeting and was always enthusiastic to talk about new results or ideas we wanted to test. I am thankful for all the time and energy you had devoted to me in the pursuit of my MSc under your supervision. I would like to thank my committee members, Dr. Robin Slawson and Dr. Joel Weadge for their support and feedback throughout each stage of my thesis.

Table of contents

List of Tables

Lists of Figures

Ch. 1 - Introduction

1. 1 The Information Age

With the rise of the information age, large amounts of data are being collected and need to be analyzed. Now more than ever, humans are generating large amounts of data, along with technological advancements that make collecting and analyzing enormous amounts of data possible. This is also evident in the scientific community, most prevalent in the -informatics fields, where the amount of raw data from experiments can pile up. Next generation sequencing techniques make it possible to sequence millions of base pairs at low error rates, while still being cost effective (Slatko *et al*., 2018). This sequencing revolution has resulted in large amounts of genome sequences from bacteria that cannot efficiently be experimentally tested. Bacteria and pathogens themselves are constantly evolving in new ways to cause disease, a better understanding of them can lead to new preventative measures and remedies. To determine the pathogenicity potential of isolates by sequence analysis, it might be necessary to manage the large number of sequences when experimental analyses might not be feasible. Pathogenic potential inferred from sequence comparison may be done by uncovering the presence of genes whose products directly participate in an organism's ability to cause disease. However, the lack of experimental confirmation of the pathogenic phenotype should be considered. Therefore, in this thesis, the term pathogenicity is more appropriate when referring to a *Salmonella*'s potential to cause disease.

1.2 Horizontal Gene Transfer

Evolution of bacteria can happen in two main ways: divergence of vertically inherited genetic material and horizontal gene transfer. Vertically inherited genetic material is the transfer of genetic information from parents to offspring, while horizontal gene transfer is the lateral transfer of genetic information between organisms (Daubin *et al*., 2016). Horizontal gene transfer can occur by three main mechanisms: transformation, conjugation, and transduction. Transformation is the uptake of genetic information from its surrounding environment, and conjugation is the transmission of genetic information to another cell via a pilus (Daubin *et al*., 2016). Plasmids can carry a wide variety of genes, such as those conferring antimicrobial resistance that may be beneficial for survival in a given environment (Daubin *et al*., 2016). Transduction is caused when a bacteriophage transfers genetic information from one cell to another (Daubin *et al*., 2016). Horizontal gene transfer is the most common way new genomic material is introduced to a host organism and may lead to the acquisition of beneficial genes. Horizontal gene transfer can introduce large segments of foreign genetic information between different species within a single generation, which is a far faster process for the acquisition of new characteristics than divergence of vertically inherited genes (Daubin *et al*., 2016).

1.3 Pathogenicity islands

A bacterium can acquire pathogenicity-related genomic regions, known as pathogenicity islands, through horizontal gene transfer. Genomic islands are long segments of DNA that are tightly associated together and flanked by mobile genetic elements (ex transposons) (Messerer *et al*., 2017). This movement of large segments of closely associated DNA may lead to immediate fitness advantages in the long and short term. Pathogenicity islands contain genes coding for several virulence factors, which include proteins that contribute to the overall virulence, or disease-causing ability, of an organism (Desvaux *et al*., 2020). These pathogenicity islands can code for a seemingly endless variety of pathogenic effector proteins or large structures useful in the transport of effectors to the host's cells. Some products of pathogenicity island genes are toxins that can be injected into a target cell, adherence proteins that allow for continual close contact with a target cell, or invasion proteins that allow for cell penetration through larger structures such as epithelial linings (Kombade and Kaur, 2021).

Toxins are substances produced by a pathogenic organism that directly or indirectly affect the target host in a negative way. Toxins can be broken down into two major categories: endotoxins and exotoxins. Endotoxins are typically found in the membranes of Gram-negative bacteria and can be released into an environment upon death or cell lysis of a bacterium (Popoff, 2018). For example, Lipid A, an endotoxin, acts as an anchor for lipopolysaccharides (LPS) that form the outer membrane in Gram-negative bacteria (Śmiechowicz, 2022). Upon cell death, endotoxins are released into the surrounding environment and trigger the innate immune response in humans. Endotoxins can cause an exaggerated immune response by binding to surface cell receptors such as TLR4/CD14/MD2, leading to the secretion of proinflammatory cytokines, nitric oxide, and eicosanoids (Farhana and Khan, 2023). This causes an intense inflammatory response from the host immune system, if enough endotoxin accumulates, septic shock can occur leading to damaged organs (Gyawali *et al*., 2019).

Exotoxins are toxins that are secreted by an organism or can be released upon cell lysis, if they had been accumulating in a cell. For example, *Clostridium tetani* is a spore forming bacterium. After entering the body, the spores germinate and form vegetative cells after an incubation period of 3-12 days (Sheehan *et al*., 2023). *Clostridium tetani* releases the exotoxin tetanospasmin, which travels via retrograde axonal transport to the spinal cord and stops inhibitory neurotransmitter release from inhibitory neurons (Sheehan *et al*., 2023). This exotoxin causes sustained muscle contractions, since the inhibitory neurotransmitters are needed to stop muscle contractions, which are being blocked by the tetanospasmin (Sheehan *et al*., 2023).

Another large group of virulence factors found in pathogenicity islands are proteins related to the adherence and invasion of target cells. Protein structures protruding extracellularly, such as fimbriae and pili, allow for adherence and association of pathogens to target cells. Pathogens, such as *Pseudomonas aeruginosa*, use type IV pili to latch onto target cells, leading to the upregulation of exotoxin virulence factors, resulting in Pneumonia infections (Persat *et al*., 2015). *Yersinia enterocolitica* is a common bacterial pathogen that can be ingested from improperly cooked or prepared food, causing severe abdominal pains and diarrhea (Uliczka *et al*., 2011). *Yersinia enterocolitica* produces extracellular surface invasin proteins YadA (*Yersinia* adhesin A) and Ail (attachment invasion locus) (Uliczka *et al*., 2011). Invasin can be found encoded chromosomally or in plasmids from horizontal gene transfer in pathogenic strains of *Y. enterocolitica*. Invasin allows for tight attachment to mammalian epithelial cells of the gastrointestinal tract, leading to the manifestation of yersinosis (Uliczka *et al*., 2011). After penetration through the epithelial wall, *Y. enterocolitica* colonizes lymphoid tissue, then makes its way to other organs in the infected host (Aziz *et al*., 2023). Yersinosis manifests as a typical gastrointestinal bacteria-based disease, with common symptoms of diarrhea, nausea, vomiting and fever (Aziz *et al*., 2023).

1.4 *Salmonella enterica*

Salmonella infections are among the most common food-borne pathogen illnesses around the world with *Salmonella enterica* being the most pathogenic species (Jajere *et al*., 2019 and Ferrari *et al*., 2019). *S. enterica* causes salmonellosis; an infection caused by the ingestion of improperly cooked/prepared food, with common symptoms being diarrhea and vomiting. The Centre for Disease Control and Prevention (CDC) estimates that 1.35 million people in the United States of America alone are infected by *Salmonella* each year (CDC, 2020). Salmonellosis is typically a mild infection that can be easily prevented but becomes a serious problem in underdeveloped countries with poor sanitation and food quality. Contaminated water with human or animal fecal matter harbours numerous amounts of pathogenic bacteria. Over 2000 serovars of *S. enterica* have been characterized along with pathogenicity islands SPI-1 (*Salmonella* pathogenicity island 1), SPI-2 and other SPIs. Serovars/Serotypes are strains found within a species that can be classified into distinct groups, depending on surface structures found on the bacterial lipopolysaccharides (LPS), flagella and polysaccharides (Ferrari *et al*., 2019). *Salmonella enterica* Typhi, is a serotype of *S. enterica* that causes typhoid fever. Surface polysaccharides, such as Vi antigen, produced by *S. enterica* Typhi, inhibit signalling pathways in a host's immune response by targeting membrane signaling molecules and signify a unique group of *Salmonella* (Parween *et al*., 2019). Furthermore, pathogenicity islands, such as SP-1 and SP-2 contain other virulence factors that encode for capsule formation, adhesion systems, and type 3 secretion systems that can provide unique phenotypes during infection (Jajere, 2019).

1.5 Average Nucleotide Identity (ANI) and Mash

Average nucleotide identity (ANI) is a metric used to determine if a set of genomes belong to the same species, by using a threshold of more than 95 percent similarity (Ciufo *et al*., 2018). ANI is a determinant of how similar two genomes are to one another and can be used to confirm the taxonomy of an unidentified bacterium. ANI is an accurate tool that can be almost universally applied to several genomic analyses. ANI can also be used to reclassify previously misclassified genomes (Ciufo *et al*., 2018).

An alternative to calculating ANI is using the Mash software for estimating how similar genomes are to one another. The Mash software uses a MinHash approach (Ondov *et al*., 2016) where a sample of k-mers, which are DNA segments 20-40 base pairs long, is taken from a genome (Ondov *et al*., 2016). These k-mers are transformed into hashes, and these hashes are stored into a "sketch" of the genome. These sketches can then be efficiently compared to one another for matching k-mers (Ondov *et al*., 2016). The proportion of matching k-mers is then used to compute an estimate of the similarity of the genomes. Mash can produce similar results as ANI but is orders of magnitude faster (Hernández-Salmerón *et al*., 2023). Mash achieves this increased speed because it only takes k-mers from a portion of the genome and extrapolates for the rest of the genome. ANI on the other hand, uses the entire genome, resulting in much longer runtimes than those of mash (Hernández-Salmerón & Moreno-Hagelsieb, 2022).

1.6 Hidden Markov models

Hidden Markov models (HMM) are statistical models that are used to describe the evolution of some observable variable that depend on internal factors (hidden states) that cannot be directly observed (Yoon, 2009). Each hidden state has transition probabilities between each state that are affected by the hidden state which came before it. Each hidden state also influences the observed variable, and the hidden variable can be determined based on the probabilities of the observed variables (Yoon, 2009). For example, we can determine the weather (hidden state) of a distant town that we cannot directly observe, by talking to a citizen of that town over the phone (observed variable) (Jurafsky and Martin 2020). The person's mood changes depending on the weather of that day, he/she is more likely to be happy if the day was sunny, and more likely to be sad if it was raining. HMMs are used in a large variety of applications such as finance, signal processing and pattern recognition. They have been extensively used in the field of bioinformatics for modeling eukaryotic genes and in the construction of protein family profiles (Yoon, 2009). Protein family profiles, built on the basis of a multiple alignment of homologous proteins, consider the probability of each amino acid residue occurring at each position in the alignment, as well as its influence on what residue comes next, while also accounting for insertions and deletions. Once constructed, HMM protein profiles can be used to classify unknown proteins into corresponding protein families and thus determine their potential functions.

HMMs can be made from all known sequences for certain protein families that contribute to the virulence of a bacterium, such as the proteins that constitute type III secretion systems. These HMM can then be compared to newly discovered or unknown protein sequences to help identify the protein and determine its function. An HMM can do a better job than a standard pairwise alignment because of the increased sensitivity and specificity that it provides. HMMs can accurately identify pathogenicity-related proteins from a newly isolated bacteria and help determine treatment options if it is found to be pathogenic.

1.7 Pairwise vs Hidden Markov models (HMM) alignments

Pairwise alignment such as blast alignments are the standard in computational biology for comparing DNA or protein sequences to one another. An alignment compares each nucleotide base or amino acid residues of the two input sequences to assess their degree of similarity. Pairwise alignments are fast and simple to run, but also have difficulty in identifying distant homologs (Park *et al*., 1998). An alternative to the standard pairwise alignment is using Hidden Markov models (HMM) described above.

1.8 Hierarchical clustering

Hierarchical clustering is a data analysis technique that is used in diverse disciplines for grouping data into clusters based on a shared metric (Zhang *et al*., 2017). These clusters can be visualized as dendrograms, branching tree-like structures that have data points as “leaves”, and “branches” which link data that share the clustering metric. The clustering metric could be a genome similarity measure such as ANI or Mash, grouping together genomes that are genetically similar to one another. Other metrics could be based on shared matches with databases such as the Pfam, to assess whether the same proteins from the databases are being matched with proteins in different genomes.

1.8 Antimicrobial resistance

Bacterial infections can be aggravated when the pathogenic bacteria also harbour genes that confer antimicrobial resistance. Microorganisms can typically be killed by a variety of antimicrobial agents that impede their growth and/or reproduction. Microorganisms can also develop mechanisms to protect themselves from the effects of antimicrobial agents, allowing them to persist and survive in the presence of an antimicrobial compound. Resistance to antimicrobials can come in many forms, such as reduced permeability of the outer membrane that limits the amount of antimicrobial that can accumulate inside a cell. Gram-negative bacteria have this added potential for innate resistance because of their additional lipopolysaccharide (LPS) outer membrane, as opposed to a Gram-positive bacterium with no LPS outer membrane (Darby *et al*., 2023). Another form of innate resistance to antimicrobials are efflux pumps, they can be found encoded on the chromosomes of organisms. Efflux pumps are transmembrane proteins that allow for the transport of molecules out of a cell. They are essential for transporting waste out of a cell and creating ion gradients for more energy demanding processes. Efflux pumps can be used to rapidly transport incoming antimicrobials out of the cell, not allowing them to accumulate and bind to their target (Reygaert, 2018).

Biofilms are formed by the colonization of bacterial communities and can also provide a physical barrier against antimicrobials and prevent them from entering the cell (Vestby *et al*., 2020). Microorganisms can also modify or protect the target of the antimicrobial, reducing the ability for the antimicrobial to bind and perform its function. Antimicrobials typically target essential processes for a cell, limiting their ability for growth and leading to their eventual death. For example, fluoroquinolones are a class of drugs that target DNA gyrase and topoisomerase

IV, two essential proteins in the DNA replication process (Reygaert, 2018). Biofilms can prevent fluoroquinolone class drugs from reaching its intended target, resulting in the drug being ineffective because it cannot bind to its target protein inside the target cell.

Instead of modifying the target, microorganisms can also modify an incoming antimicrobial to inhibit its effects. Two common strategies for modification of an antimicrobial drug are drug inactivation and degradation. A drug can become inactivated by the transfer of a chemical group such as an acetyl or adenyl group, reducing the binding potential of the drug to its target (Reygaert, 2018). An antimicrobial can be degraded using enzymes that cleave off its functional groups via hydrolysis (De Pascale *et al.*, 2010). These functional groups are essential for the antimicrobial to bind to its target, without any binding occurring, it is rendered useless (De Pascale *et al.*, 2010). Antimicrobial resistance mechanisms can allow pathogenic bacteria to persist in an environment, allowing them to survive and propagate disease. Antimicrobial resistance may not directly be related to virulence, but it can be a tool used by pathogenic bacteria to help them survive in a host.

1.9 Environmental and Clinical bacteria

Clinical bacteria isolates are taken from a host that has been infected by that bacterium, typically taken from hospital patients that had been showing symptoms from a disease. Environmental bacteria isolates would be taken from the outside environment, typically from soil samples where large populations of bacteria would reside. Clinical isolates can be expected to be pathogenic because they are taken from a host already infected by a known disease-causing bacterium, while an environmental bacterium with the same genetic classification might be less pathogenic, or have less pathogenic potential. Pathogenic and non-pathogenic strains of bacteria have been observed within the same species. For example, *Escherichia coli* K-12 is a non-pathogenic strain of *E. coli*, while *E. coli* O157:H7 is a pathogenic strain of *E. coli* that can cause intestinal hemorrhages (Stromberg *et al*., 2018). *Salmonella* isolates are all typically considered pathogenic, but non-pathogenic strains might still be undiscovered. Comparing the genomes of environmental *Salmonella* to clinical isolates may uncover differences between the two groups in their capacity to cause disease. Specifically, clinical genomes may have key virulence factors that may not be present in the genomes of environmentally isolated pathogens.

Hypothesis and Objectives

Hypothesis

Salmonella genomes isolated from the environment will show less pathogenicity potential than genomes isolated from pathogenicity-related sources, such as clinical or human samples.

Objectives

Overarching Objective - To determine if there is a relationship between the isolation source of *Salmonella enterica* and their pathogenic potential.

Objective 1 - To determine the complement of virulence factor genes and antimicrobial resistance potential of *Salmonella* genomes to give insight into their overall pathogenic potential.

Objective 2 - To determine if there is a relationship between the isolation sources of *Salmonella* genomes and their virulence factor complements and/or antimicrobial resistance-related genes.

Ch. 2 - Methods

2.1 *Salmonella enterica* genomes

Seventy-five environmental *Salmonella enterica* bacteria were isolated from Clair and Silver Lakes within the Laurel Creek sub-watershed of the Grand River in the Kitchener/Waterloo region by Dr. Robin Slawson's lab at Wilfrid Laurier University (Thomas *et al*., 2013). These genomes were downloaded from the Salfos database (https://salfos.ibis.ulaval.ca). Another 165 clinical *S. enterica* genomes were downloaded from the National Center for Biotechnology Information (NCBI). The genome dataset was later expanded to include approximately 6000 *S. enterica* genomes taken from the Bacterial and Viral Bioinformatics research center (BV-BRC), formally known as the Pathosystems Resource Integration Center (PATRIC) (Olson *et al*., 2023).

2.2 Negative control genomes

Bacteria closely related to *Salmonella*, such as *E. coli*, that were classified as non-pathogenic by Cosentino S *et al*., (2013) were used to construct a negative control dataset. This set of genomes was used as a non-disease-causing negative control dataset to compare against all *Salmonella* (Table 1).

Table 1. Bacteria name and strain of non-pathogenic negative control genomes

Organism	**Strain**
Escherichia coli	K-12 substr. MG1655
Escherichia coli	B str.REL606
Escherichia coli	ATCC 8739
Escherichia coli	APEC 078
Escherichia coli	SE15
Escherichia coli	KO11FL
Escherichia coli	IAI1
Escherichia coli	BL21(DE3)
Escherichia coli	BL21-Gold(DE3)pLysS AG
Escherichia coli	SE11
Escherichia coli	ABU 83972
Escherichia coli	SMS-3-5
Escherichia coli	DH1
Escherichia coli	HS
Salmonella enterica	subsp. enterica serovar Choleraesuis str. SC-B67

2.3 Isolation source classification

Information about the isolation sources of *Salmonella* genomes was taken from metadata available at the BV-BRC database (Olson *et al*., 2023). The genomes were categorized and grouped depending on where the bacteria were isolated from. Isolation source groups consisted of human, animal, plant, environmental, water source, food and nuts. An *ad hoc* python script was written to assign each genome into an isolation source group based on their isolation source provided in the metadata table. Lists of all unique isolation sources and groups can be found in Figure 21 in the Appendix. Any fecal or tissue sample taken from a human patient was classified into the "human" grouping. Uncharacterized isolation sources, such as just "humans", were also added to the human grouping. Different varieties of animals, such as chickens, cows and pigs, were put into the "animal" category. Environmental and water source groups were separated to see if the two would group separately when comparing their pathogenicity. Environmental genomes consisted of samples taken from farms or non-specific locations that were designated as just "environmental". Seeds were categorized into the plant grouping and could potentially be merged with the nut grouping if no clear distinction can be observed between the two. Most of the clinical genomes were labelled as just "clinical" in the metadata, but potentially could be reclassified into the human grouping depending on results. Unspecified clinical genomes could be reclassified into the human grouping depending on the results, since they were most likely taken from a human subject. A large majority of the genomes did not have an isolation source listed in the metadata table, they were omitted from the grouping process and further analysis. Isolation sources that had no specific designation of where it came from such as "liver", were omitted from the grouping process. With such a large number of unique isolation sources and inaccurate wording in the metadata, many genomes were not given a grouping because the python script used keywords in order to add a grouping, and sorting the genomes one-by-one by hand would be too lengthy of a task.

2.4 Non-redundant *Salmonella* protein database

A non-redundant *Salmonella* protein database was compiled using all the proteins annotated in the *Salmonella* genomes in this project. Since the very same proteins are likely to be annotated in thousands of the genomes analysed, a non-redundant dataset was produced using an *ad-hoc* program written by Dr Moreno-Hagelsieb, to avoid working with the same proteins thousands of times.

2.5 Pairwise protein comparisons

To infer virulence factors, the proteins annotated in all genomes used in this study were aligned against the Virulence Factor Database (VFDB) (Liu *et al*., 2022), using the DIAMOND alignment software (Buchfink *et al*., 2015). To infer antimicrobial resistance, initial comparisons against the Comprehensive Antibiotic Resistance Database (CARD) (Alcock *et al*., 2023) were also performed using DIAMOND (Buchfink *et al*., 2015), but later produced and classified using the Resistance Gene Identifier (RGI) software pipeline produced by the same group that compiled the CARD database (Alcock *et al*., 2023).

2.6 Comparison against Hidden Markov Models (HMM)

All genomes taken from the BV-BRC database, environmental samples isolated from the Kitchener/Waterloo region, and negative controls were compared against Hidden Markov

Models (HMMs) from the Pfam database (Mistry *et al*., 2021), and from the VFDB produced by de Nies *et al.* (de Nies *et al*., 2021). The alignments were performed with hmmscan, from the HMMER software (Finn *et al*., 2011).

2.7 Hierarchical clusters and tanglegrams

A custom python script was made to produce matrices that outline the number of occurrences of each individual match against the VFDB, Pfam and CARD for all genomes. These matrices were used to produce hierarchical clusters to visualize the similarities and differences between clustering metrics. Hierarchical clusters can help determine if different metrics produce groups that are related to one another by different characteristics, and whether different metrics produce similar hierarchies. These hierarchical clusters were produced using the divisive clustering method implemented as “diana” in the cluster package available in R (Ihaka and Gentleman, 1996). Hierarchical clusters were used to produce tanglegrams to calculate entanglement values, to better determine how similar the clustering metrics are. Tanglegrams and Basker’s correlations were produced using the package dendextend in R (Galili, 2015).

Ch. 3 - Results

3.1 Initial results: Clinical versus Environmental *Salmonella*

Initially, this research was aimed at comparing 75 genomes of *Salmonella* isolated from the environment, specifically, those isolated primarily from Clair and Silver Lakes in the Waterloo region by Dr. Slawson's lab at Wilfrid Laurier University, against 165 clinical isolates. The initial hypothesis was that the genome analyses would show that the lake samples had a lower pathogenicity potential.

3.1.1 Average Nucleotide Identity

To verify that the *S. enterica* genomes were correctly classified and thus belonged to a single species, their average nucleotide identity was calculated. Most genomes had an average nucleotide identity above the species threshold of 95% (Fig. 1) (Jain *et al*., 2018).

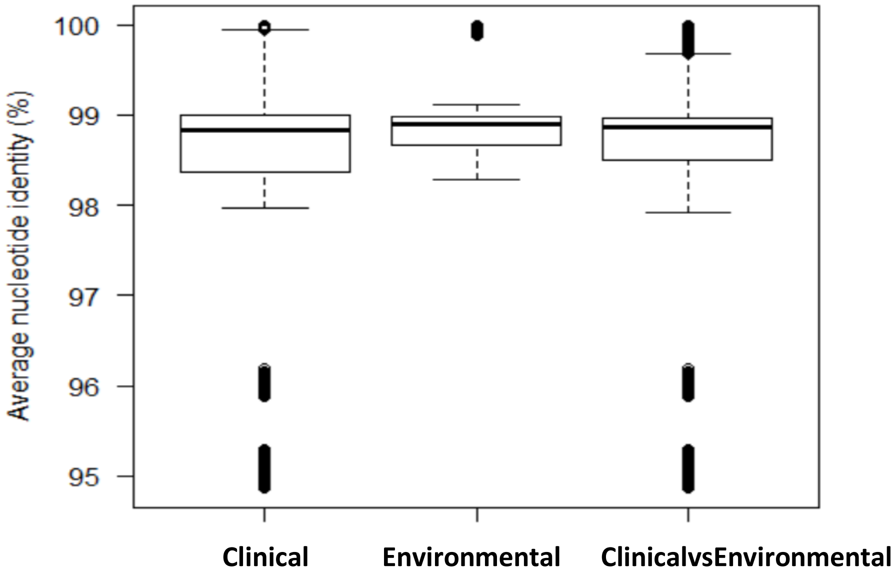


Figure 1. Average nucleotide identity (ANI) scores of clinical and environmental genomes. The ANI between almost all genomes were above the 95% threshold, within each group and against each group, confirming that all genomes belong to isolates of a single species. The bolded horizontal line represents the median of the data, and the box represents the middle 50% of all data from the sample. The whiskers that extend from the box are the maximum and minimum values that are not outliers. The white circles beyond the whiskers are values that are outliers, outliers are defined as data that is 1.5 times larger or smaller than the interquartile range.

3.1.2 Alignment with the Virulence Factor Database (VFDB)

The clinical and environmental genomes were compared against the VFDB to identify potential pathogenicity-related proteins (Fig. 2). The clinical samples of *S. enterica* were found to have a median value of 961, minimum of 923 and a maximum of 1022 (Fig. 2). The environmental samples of *S. enterica* were found to have a median value of 968, a minimum value of 936 and a maximum value of 1015 (Fig. 2). The clinical samples also had 3 outlier values that were 1.5 times larger than the interquartile range, while the environmental samples had none. The clinical samples were found to be non-normally distributed after a Shapiro-Wilks test was conducted resulting in a test statistic *W* = 0.977 and a p-value of 8.40x10^{-3}. This value is lower than the alpha value of 0.05, so the null hypothesis that the data was normally distributed was rejected. The environmental data set was also found to be non-normally distributed after a Shapiro-Wilks test was conducted resulting in a test statistic *W* = 0.898 and a p-value of 1.93x10^{-5}.

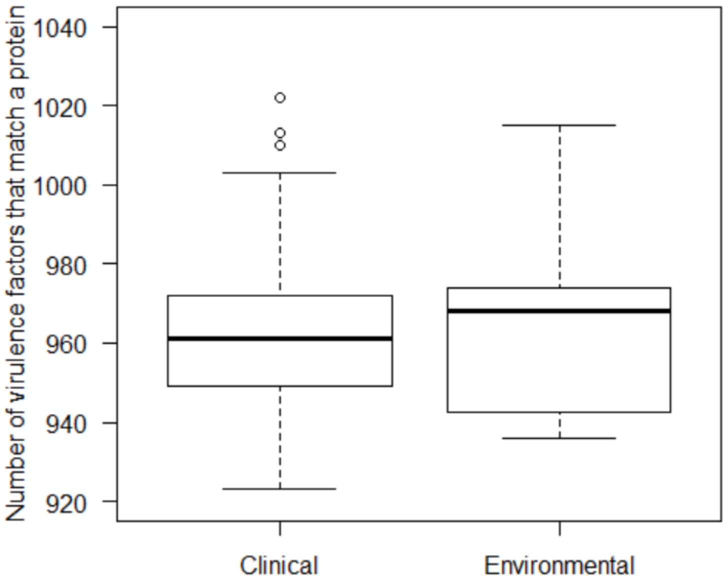


Figure 2. Number of virulence factors in the VFDB that match a protein in the genomes of clinical and environmental *S. enterica*. The clinical and environmental genomes have similar amounts of matches with the VFDB. The clinical *S. enterica* were found to have a median of 961 matches, while the environmental ones had a median of 968 matches.

3.1.3 Completeness of Virulence Factor Sets

Larger virulence factor proteins or structures, such as efflux pumps, would require multiple genes to produce a functional protein. All genes that are required to produce an efflux pump make up a virulence factor set. The percent completeness of virulence factor sets was calculated for the clinical and environmental genomes to give insight into their pathogenic potential. The clinical and environmental genomes were found to have almost the same amount of virulence factor sets completed, around 41% (Fig. 3).

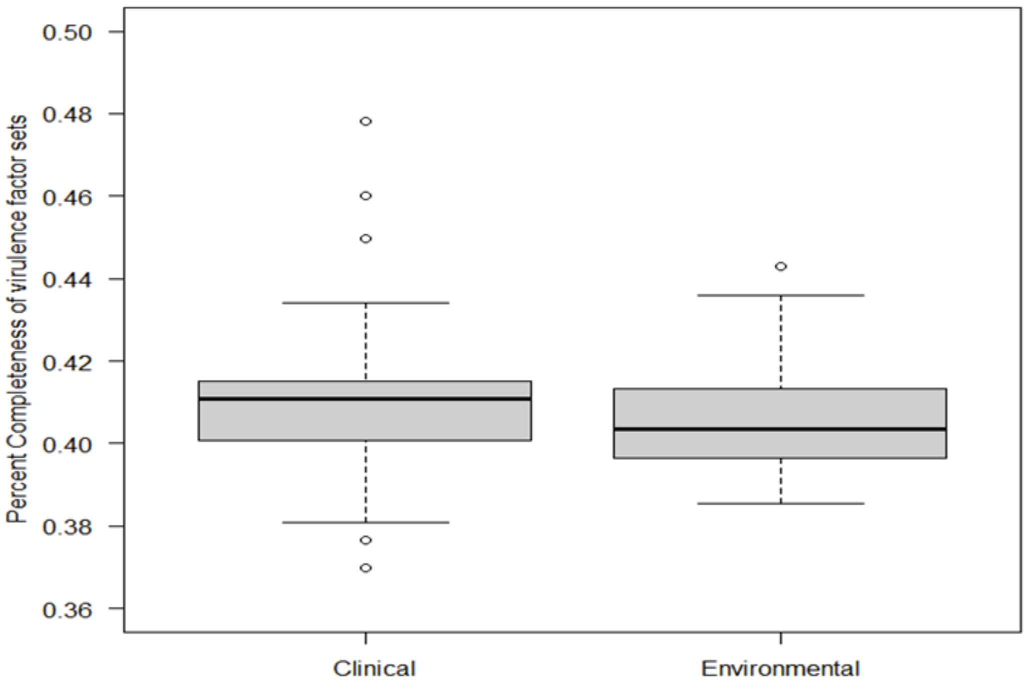


Figure 3. Percent completeness of virulence factor gene sets in the clinical and environmental *S. enterica* genomes. The clinical and environmental genomes have a similar amount of complete virulence factor sets, with around 41% being complete.

3.1.4 Alignment with the Comprehensive Antibiotic Resistance Database (CARD)

The clinical and environmental genomes were aligned with the CARD database, the clinical genomes were found to have a median of 27.41 matches, while the environmental genomes had a median of 13.81 matches (Fig. 4). The clinical genomes had a total number of matches of 4522, and the environmental genomes had 1036 matches total. It should be noted that there are approximately double the number of environmental genomes than there are clinical, on average the clinical genomes had double the number of matches than environmental when sample size is considered.

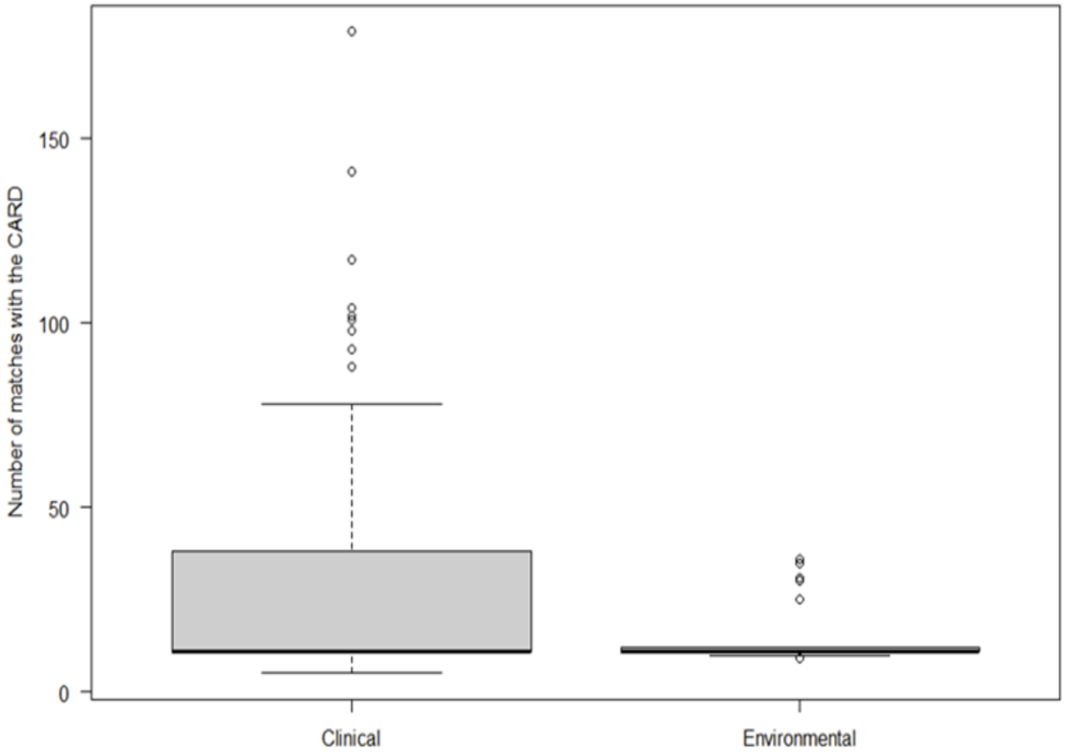


Figure 4. Number of proteins that matched the Comprehensive Antibiotic Resistance database (CARD) in both sets of *S. enterica*. The clinical genomes were found to have more matches than the environmental genomes. The clinical genomes had an average of 27.41 matches, and the environment an average of 13.81 matches.

3.1.5 Cluster analysis

The clinical and environmental genomes were clustered by their ANI scores to determine if phylogeny played a part in their pathogenicity (Fig. 5). The two genome sets did not produce large distinct clusters from one another, instead they had smaller “clusters of clusters” mixed throughout the hierarchical cluster. When the genomes were clustered by their matches with the CARD, the environmental genomes had two large clusters at the bottom of the hierarchical clusters, with smaller clusters scattered throughout (Fig. 5). The CARD cluster also had one very large branch containing almost half the genomes in the entire cluster at the bottom of the cluster (Fig. 5). Even when the genome sets were clustered by their matches with the VFDB, two distinct clusters were not produced, and the genomes were mixed throughout the hierarchical cluster (Fig. 6). Tanglegrams were produced to compare how similar two hierarchical clusters were to one another as this can help determine if two clustering metrics are related. Tanglegrams have an entanglement value that ranges from 0.00-1.00, the larger the value, the less similar the two hierarchical clusters are to one another. An entanglement value of 0.00 would mean that the two hierarchical clusters are identical, and the two-clustering metrics would produce the same results. The comparison between the clustering metrics of matches with the CARD and VFDB had an entanglement score of 0.59 (Fig. 7), a fairly high level of entanglement. The comparison of the CARD and ANI hierarchical clusters produced a similar entanglement score of 0.64 (Fig. 8), and ANI compared with the VFDB had a much lower entanglement score of 0.32 (Fig. 9).

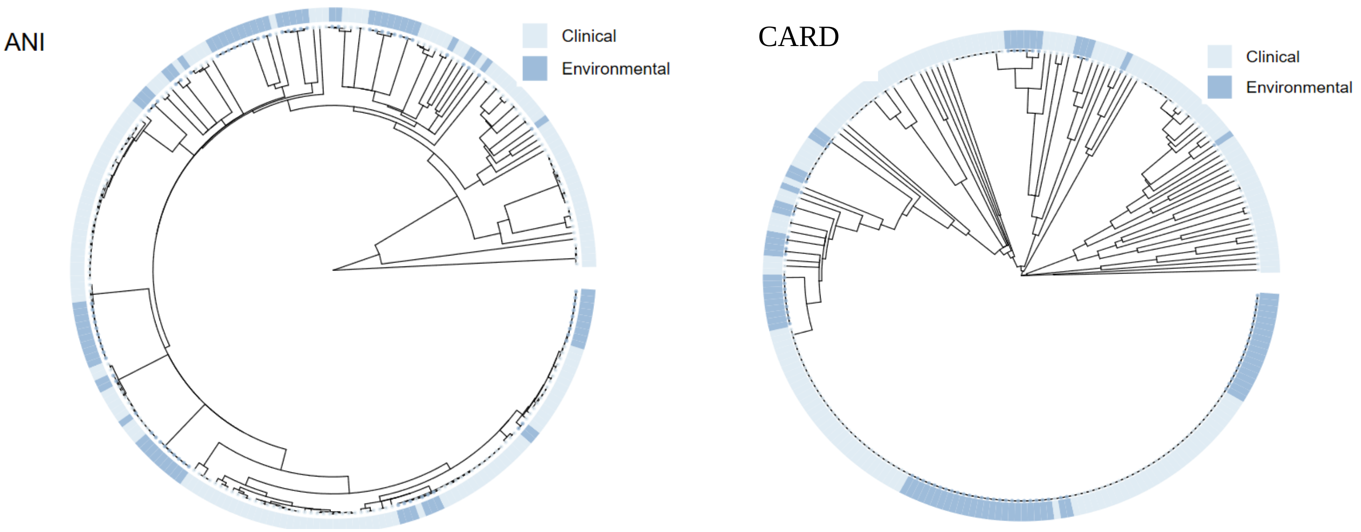


Figure 5. Hierarchical clusters of clinical and environmental *S. enterica* genomes. Cluster based on Average Nucleotide Identity (ANI, left) and on matches with the Comprehensive Antibiotic Database (CARD, right). The two genome sets when clustered by ANI did not have a distinct separation between the two groups, while clustering by CARD matches produced slightly larger distinct clusters. The CARD cluster also has one very large branch at the bottom of the cluster, almost half the size of the cluster.

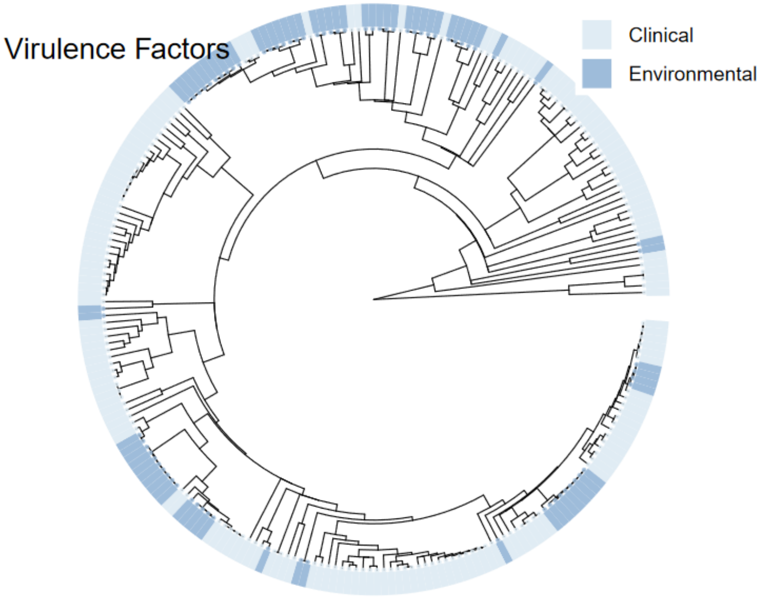


Figure 6. Clinical and environmental *S. enterica* genomes clustered by genes that produce a match with the Virulence Factor Database (VFDB). The genome sets did not produce two very distinct groups, but smaller clustered groups can be seen.

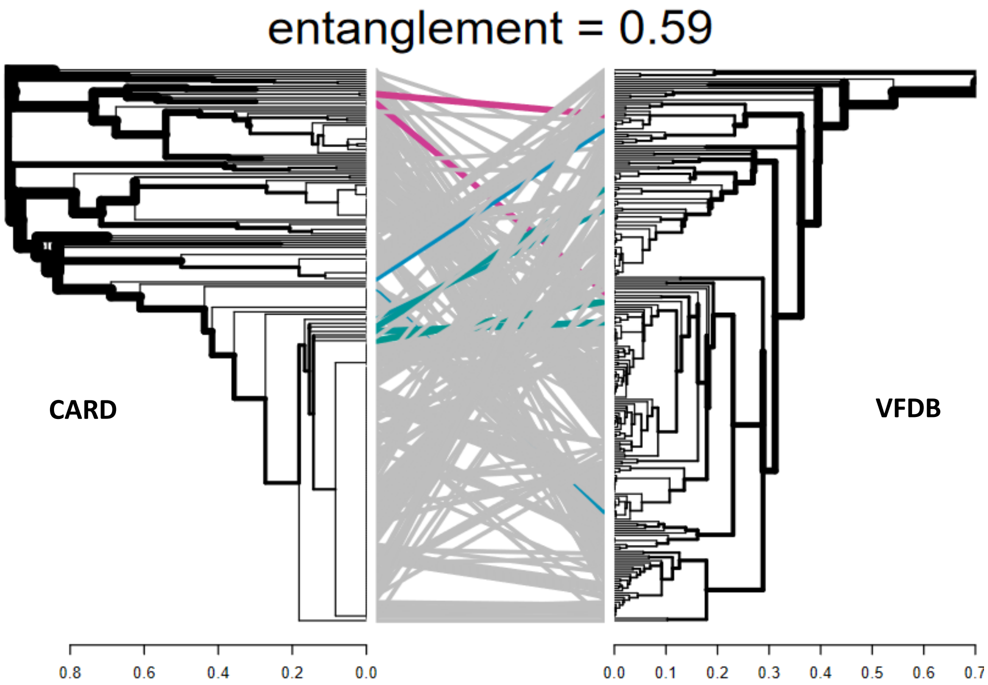


Figure 7. Comparison between hierarchical clusters by the CARD and VFDB matches in both sets of *S. enterica* genomes. With entanglement values ranging from 0.00-1.00, a value of 0.59 has a high level of entanglement between the two clusters. The colored lines represent regions of the clusters that are shared between one another.

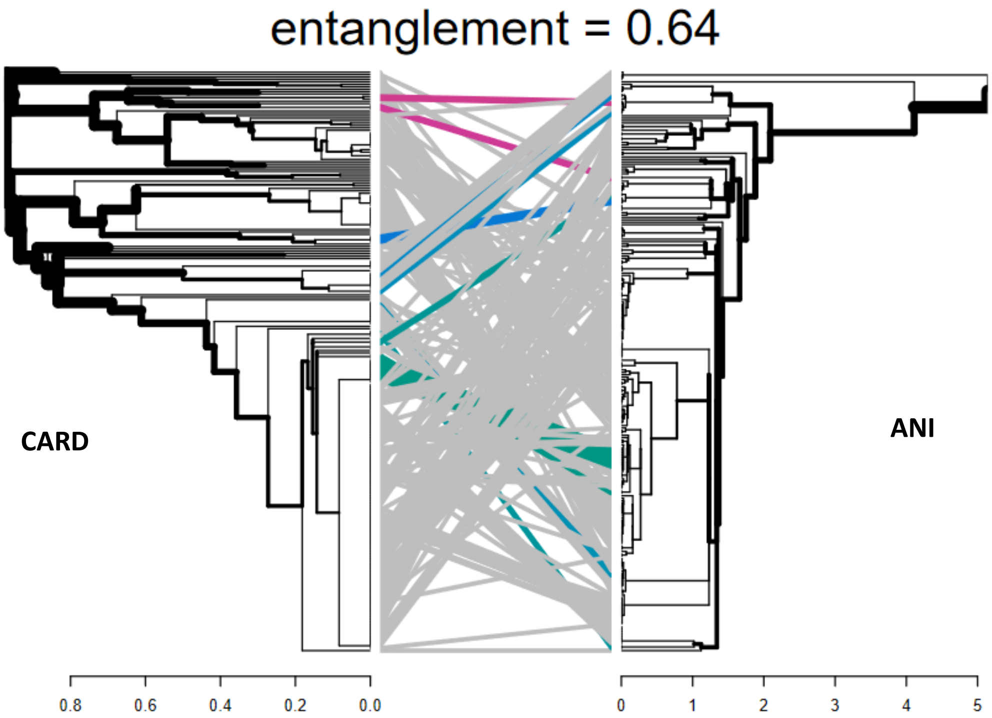


Figure 8. Comparison between hierarchical clusters by the CARD matches and ANI scores in both sets of *S. enterica* genomes. With entanglement values ranging from 0.00-1.00, a value of 0.64 has a high level of entanglement between the two clusters.

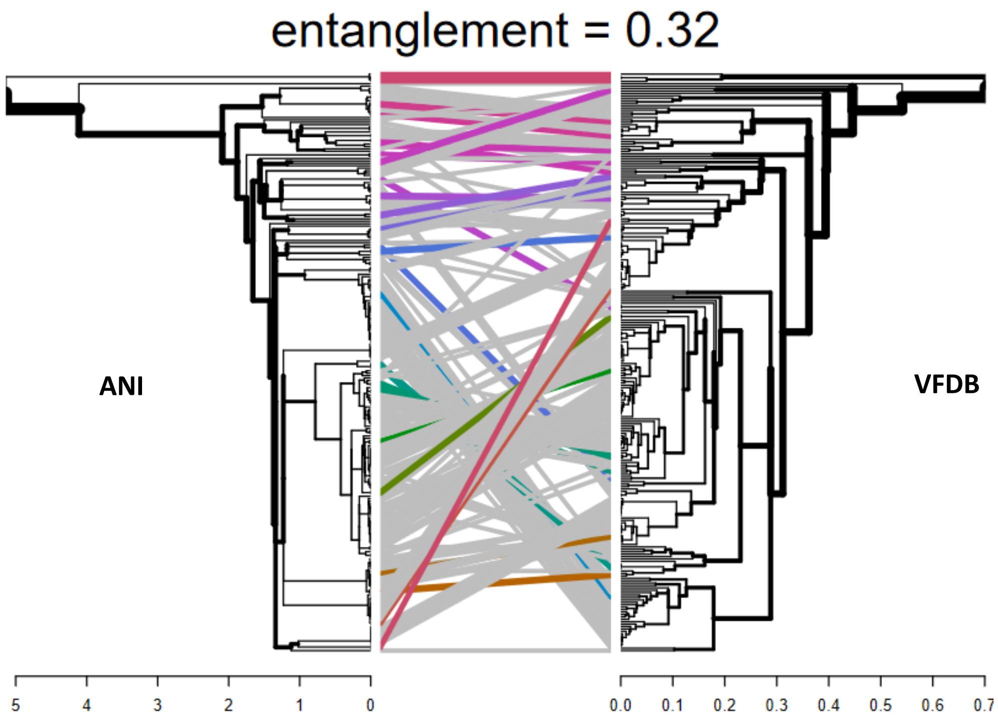


Figure 9. Comparison between the hierarchical clusters by ANI scores and the VFDB matches in both sets of *S. enterica* genomes. Entanglement values range from 0.00-1.00, a value of 0.32 has a moderate level of entanglement between the two clusters.

3.2 Expanded genome dataset

The genome dataset was expanded to include approximately 6,000 *Salmonella* genomes from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC), specifically those containing enough information to be classified into different isolation source categories.

3.2.1 Isolation Source groups of *Salmonella* taken from the BV-BRC

The dataset of *Salmonella* genomes was categorized based on their isolation source (Fig. 10). The largest genome groups were genomes isolated from animal and human sources, such as organ tissue and feces. The next largest groups were genomes isolated from water sources such as lakes and riverbeds, or environmental samples from farms and forests. The nut, plant, food, and negative control genomes were the smallest of the groups, but they were included because they are a vastly different environment than an animal or human source. A *Salmonella* found on a nut, or a plant would be exposed to different environmental conditions than a *Salmonella* found in an animal. These different environmental conditions may give rise to phenotypic differences as well, resulting in genomic differences that may be detectable.

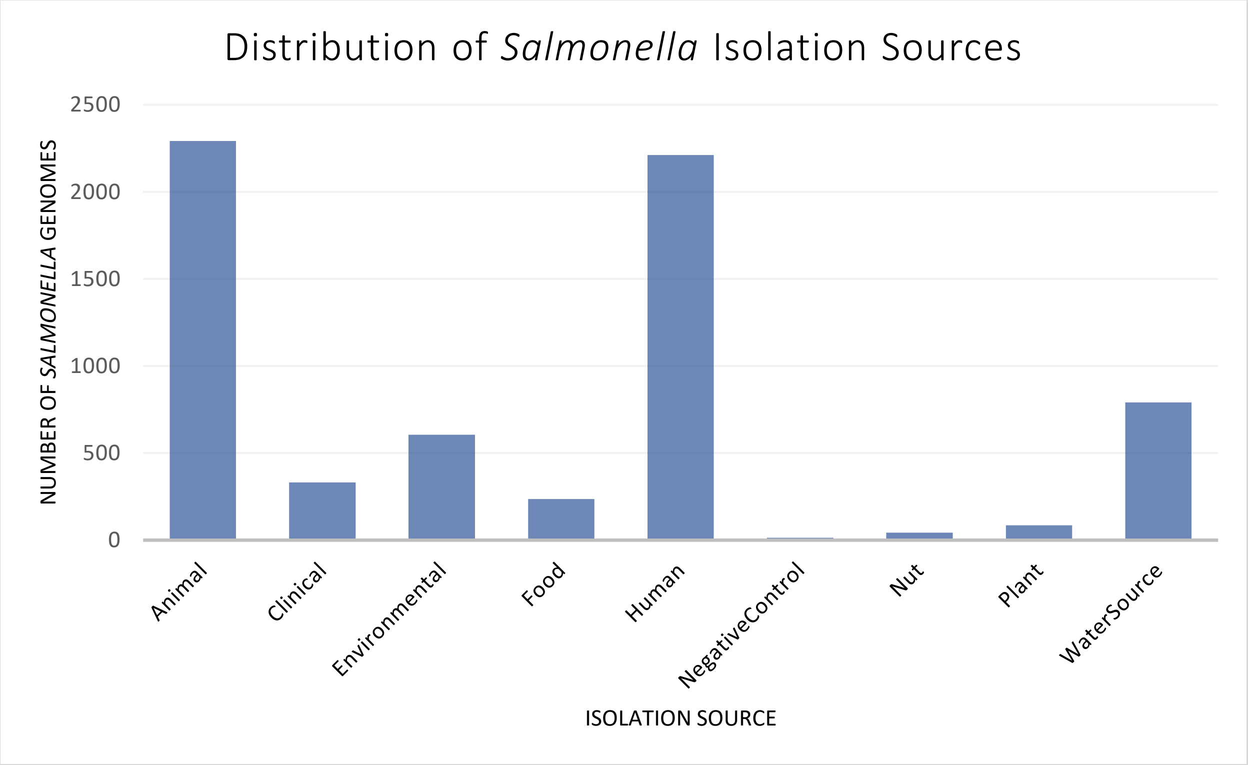


Figure 10. Isolation sources of *Salmonella* genomes. Isolation data was taken from metadata available at the Bacterial and Viral Bioinformatics Resource Center (BV-BRC). *Salmonella* isolated from human and animal sources make up the majority of the genomes, while water source and environment comprised the next largest datasets. The food, nut, plant, and negative control groups were the smallest of the groups.

3.2.2 Inference of pathogenicity potential

The number of proteins that produced a match with the VFDB was determined to give insight into the pathogenic potential of the *Salmonella* genomes (Fig. 11). The negative control genomes were found to have fewer matches than all other genome datasets, with a mean of 1014 matches (Fig. 11). This difference in matches was found to be statistically significant because the p-value from an un-paired two sample Wilcoxon test was less than the threshold value of 0.05 (Table 2). A P-value of <0.05 shows that the difference in means is very unlikely due to random chance, while a value of >0.05 would mean that the difference in means is most likely due to random chance. The nut and plant groups were consistently found to be not significantly different than the other groups (Table 2). The clinical and environmental groups were also found to be not significantly different from one another (Table 2). The number of proteins that produced a match with the HMM-VFDB was determined for *Salmonella* genomes isolated from different sources (Fig. 12). The negative controls had fewer matches with the HMM-VFDB, but over 2800 genes related to pathogenicity is too many when compared to the total number of genes a *Salmonella* has in its genome.

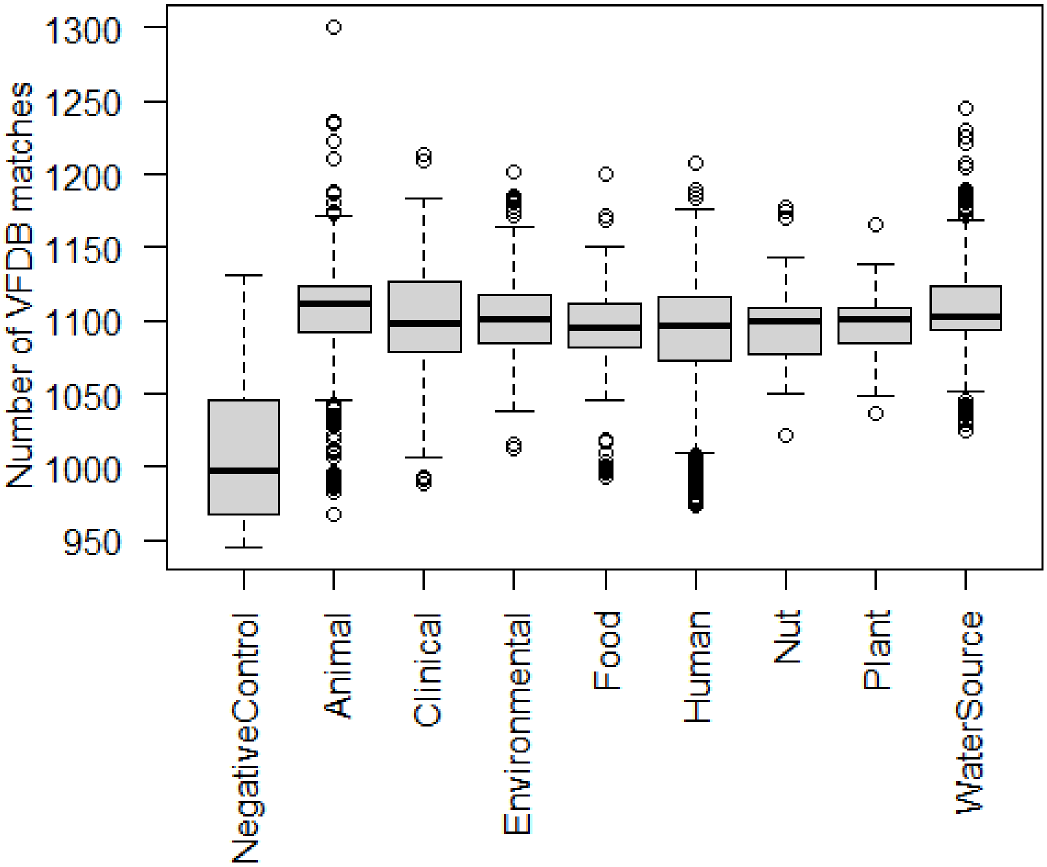


Figure 11. Number of proteins that produced a match with the VFDB by sequence alignment. The negative control *Salmonella* genomes were found to have less proteins that matched the VFDB than other isolation sources.

Table 2. P-Values from an un-paired two sample Wilcoxon test comparing VFDB matches.

	Animal	**Clinical**	**Environmental**	**Food**	**Human**	**Nut**	**Plant**	**WaterSource**
Negative Control	4.7e-05	1.96e-05	1.81e-06	1.48e-05	1.96e-05	4.96e-05	1.20e-05	9.64e-07
Animal		1.42e0-5	9.32e-11	6.82e-13	<2.2e-16	7.00e-04	2.48e-06	0.01
Clinical			0.84	0.0475	6.07e-4	0.29	0.20	2.64e-03
Environmental				6.05e-03	1.93e-06	0.15	0.01	5.43e-04
Food					0.82	0.98	0.66	3.75e-08
Human						0.94	0.71	<2.2e-16
Nut							0.76	0.01
Plant								1.69e-03

Legend: 🟪 = Significant difference 🟨 = No significant difference

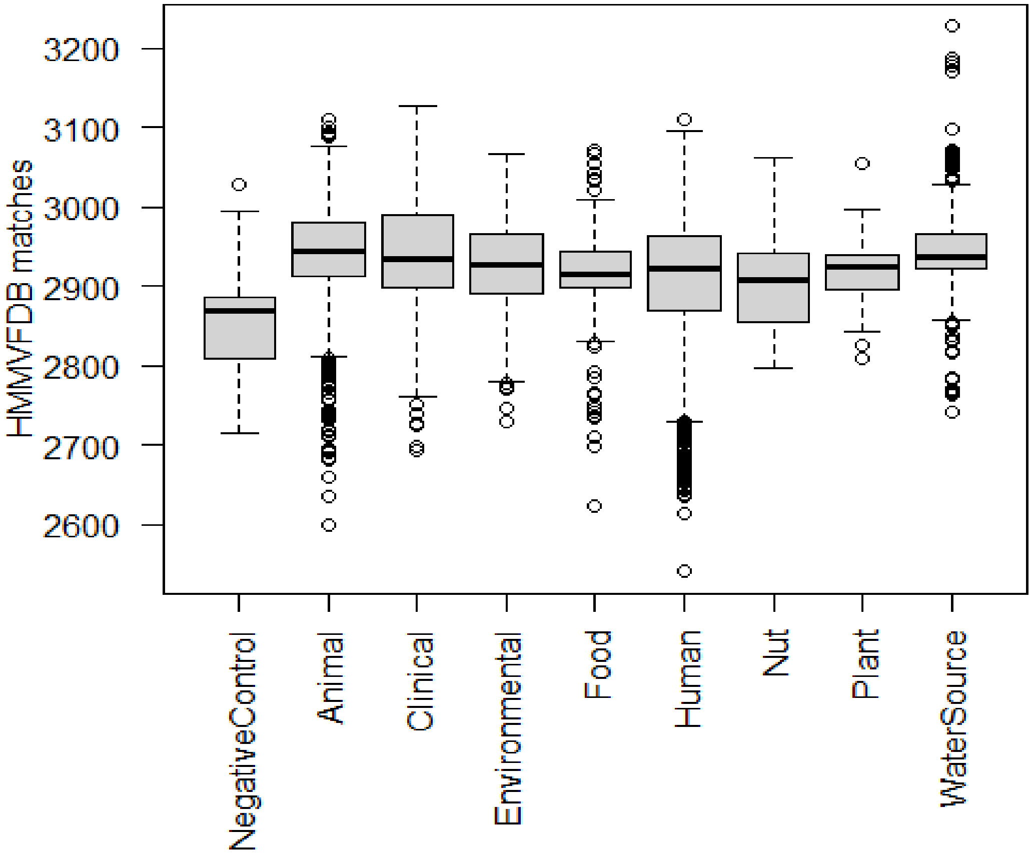


Figure 12. Number of proteins that produced a match with the HMM-VFDB. The negative control was found to have a higher number of matches with the HMM-VFDB than all other groups.

The PathFam database was produced by Lobb *et al*. in 2021 and ranked all Pfam protein domains from most pathogenic associated, to least pathogenic associated (Lobb *et al*., 2021). The protein domains were ranked by how often they were found in the genomes of pathogenic bacteria compared to non-pathogenic bacteria. The *Salmonella* proteins were compared against the highest-ranking domains to determine the amounts of pathogenicity associated domains of the different environmental groups. The negative controls were found to have significantly less pathogenicity associated domains than all other groups (Fig. 13 and Table 3). The nut and plant groupings were, once again, not found to be significantly different to some of the other groups, as well as the human group when compared to the clinical and environmental groups (Table 3). The human group was also found to be most similar to the clinical and environmental groups (Table 3).

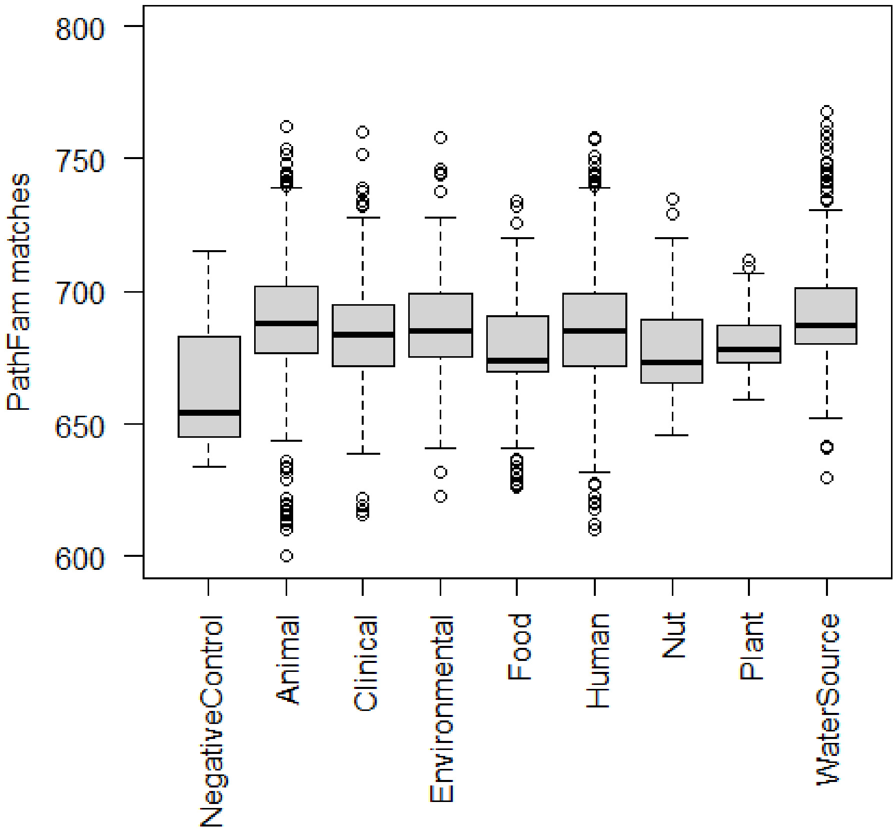


Figure 13. Number of Pfam domains that produced a match with the most pathogenicity associated domains from the PathFams database. The negative control genomes were found to have less matches with pathogenicity associated domains from the PathFam database than all other groups.

Table 3. P-Values from an un-paired two sample Wilcoxon test comparing PathFam matches.

	Animal	Clinical	Environmental	Food	Human	Nut	Plant	WaterSource
Negative Control	1.60e-04	1.63e-03	5.4e-04	0.01	4.9e-04	0.02	4.4e-03	1.04e-04
Animal		1.42e-05	2.32e-04	<2.2e-16	3.91e-08	3.37e-05	1.65e-07	0.18
Clinical			0.23	7.50e-04	0.14	9.10e-04	8.10e-03	1.12e-06
Environmental				1.09e-07	0.86	1.26e-03	5.99e-04	<2.2e-16
Food					9.27e-08	0.35	0.43	<2.2e-16
Human						4.47e-03	9.76e-04	1.34e-07
Nut							0.09	3.20e-06
Plant								0.09

Legend: 🟪 = Significant difference 🟨 = No significant difference

3.2.3 Antimicrobial resistance genes

The number of proteins that produced a match with the CARD database was determined to give insight into the antibiotic resistant capabilities of the *Salmonella* genomes. The negative control dataset was found to have more matches with the CARD than the other datasets with a mean of 56.5 matches (Fig. 14). This difference was found to be statistically significant based on the results of an unpaired two sample Wilcoxon test between the negative controls and each individual grouping (Table 4). All matches were classified as either strict or perfect matches to ensure no false positives were present. Perfect matches are identical matches to the protein in the database, while a strict is a variant of that protein, that falls within the inputted blast cut-off threshold (Alcock *et al*., 2023). Even when filtered for only perfect matches, the negative control genomes had significantly more matches than all other groups (Fig. 15).

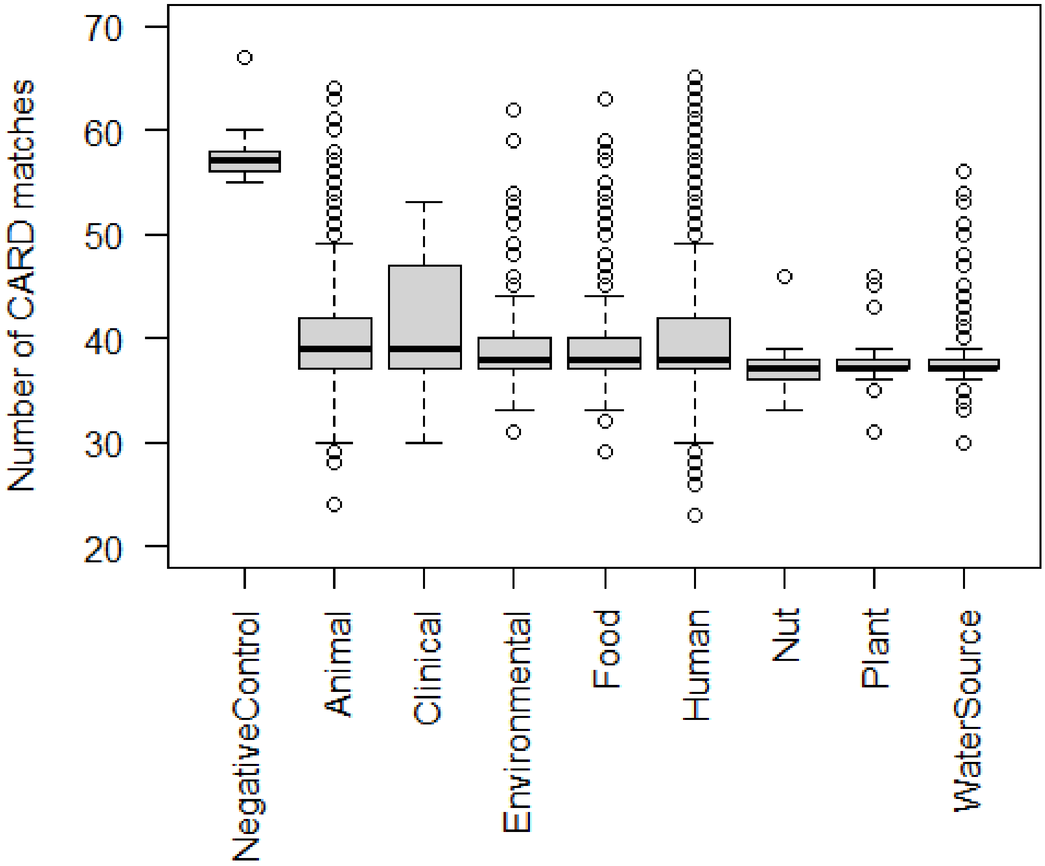


Figure 14. Number of proteins that produced a match with the CARD. The Negative controls were found to have a higher amount of CARD matches than all other datasets.

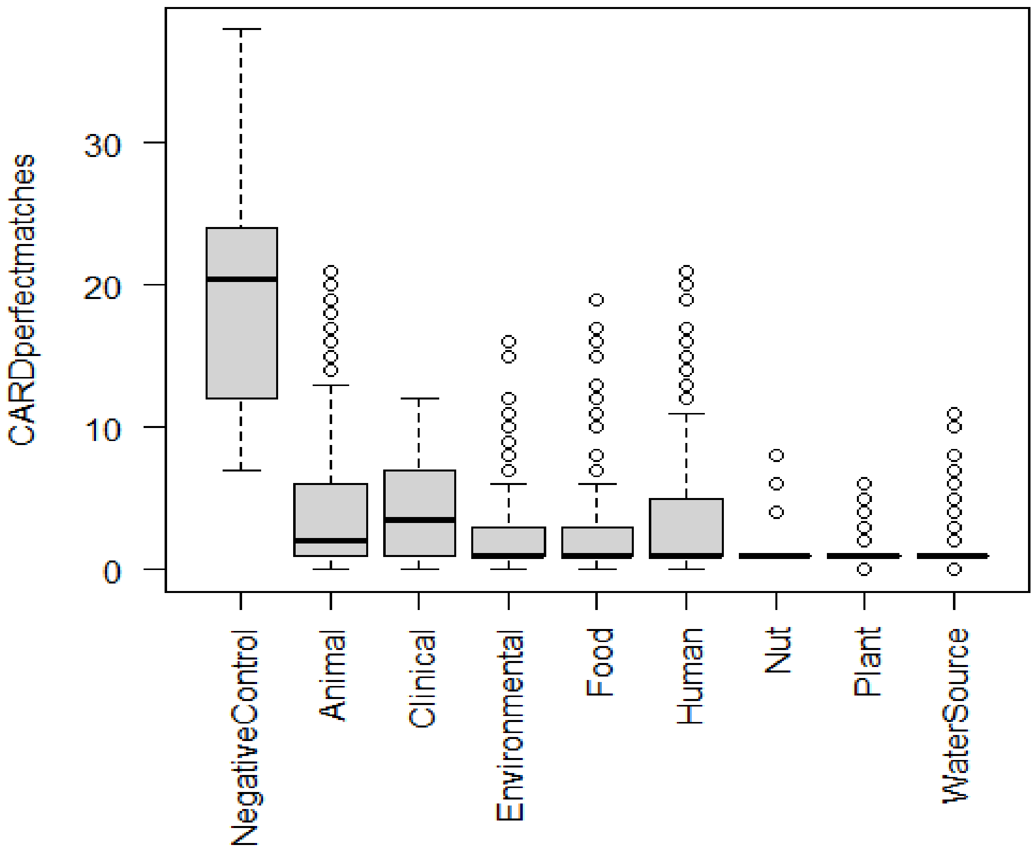


Figure 15. Number of perfect matches with the CARD. The negative control genomes were found to have significantly more matches than other isolation sources.

Table 4. P-Values from an un-paired two sample Wilcoxon test comparing CARD matches.

	Animal	Clinical	Environmental	Food	Human	Nut	Plant	WaterSource
Negative Control	2.59e-10	1.8e-10	7.14e-11	4.21e-10	1.97e-10	1.74e-08	4.82e-10	1.00e-11
Animal		0.14	<2.2e-16	5.18e-07	<2.2e-16	9.63e-09	3.00e-14	<2.2e-16
Clinical			4.42e-13	9.99e-05	1.16e-08	6.93e-07	1.01e-09	<2.2e-16
Environmental				5.67e-03	9.33e-03	9.44e-03	8.26e-03	2.82e-04
Food					0.467	3.17e-04	7.28e-06	1.53e-09
Human						4.07e-03	5.90e-04	1.98e-12
Nut							0.47	0.13
Plant								0.25

Legend: 🟪 = Significant difference 🟨 = No significant difference

3.2.4 Cluster analyses

The *Salmonella* genomes were clustered together by their matches with the Pfam, only the human grouping produced a large distinct cluster on the right side of the hierarchical cluster (Fig. 16). The human grouping also had medium sized clusters grouped together on the left side of the hierarchical cluster, but still mingling with other groups. The water source grouping also produced some small clusters that were grouped together at the far right of the dendrogram, as well as the animal grouping at the bottom of the hierarchical cluster. Overall, all groups were mixed and not distinct from one another, only the human group produced one large distinct cluster (Fig. 16). When the *Salmonella* genomes were clustered by which proteins they matched in the VFDB, the human grouping once again produced a large distinct cluster on the right side of the hierarchical cluster (Fig. 17). Genomes from the water source grouping clustered together near the bottom of the hierarchical cluster, but not very distinct from other groups. The animal grouping produced a small distinct cluster near the bottom right of the hierarchical cluster. Besides the one large distinct human cluster, no groupings were found in a distinct large cluster separated from the other groups. When the *Salmonella* genomes were clustered together by their Mash scores, the human grouping produced a large distinct cluster at the top right of the hierarchical cluster, as well as smaller clusters spread out near the bottom left (Fig. 18). The animal grouping produced a large distinct cluster on the left side of the hierarchical cluster, but also has a lot of smaller clusters scattered all around. The water source grouping had some medium sized clusters at the bottom of the hierarchical cluster, but they were still mixed with other groups. None of the groups produced large clusters that were distinct from all other groups, besides the large human cluster on the right side. The hierarchical clusters from clustering by VFDB and Mash were compared using a tanglegram to determine if there was a relationship between the two (Fig. 19). A tanglegram lines up two hierarchical clusters and determines if the same genomes are found in the same spots in each of the clusters, producing an entanglement value. Entanglement values range from 0.00 to 1.00, with 0.00 being no entanglement, and 1.00 have full entanglement between the two clusters. No entanglement between the two clusters means that the genomes line up identically, and the same genomes are found in the same spots in the clusters. An entanglement value of 0.00 can still be achieved with different clusters, if the genomes line up the same when compared, the branching within the hierarchical cluster can still differ. The VFBD and Mash had an entanglement value of 0.17, a moderate level of entanglement, and the two clusters are fairly similar to one another (Fig. 19). The Baker Gamma correction correlation is a measure of similarity between two hierarchical clusters that ranges from -1 to 1, with values closer to 0 mean the two trees are not similar. A Baker's Gamma correction correlation has an inverse relationship with entanglement values, a lower entanglement value will result in a higher Baker's Gamma value. The VFDB and mash clusters had a Baker's Gamma correction correlation of 0.73, meaning the two clusters are very similar. The hierarchical clusters for the VFDB and Pfam were similar to one another because of their low entanglement value of 0.09 (Fig. 20). The VFDB and Pfam clusters were also very similar to one another with a Baker's gamma of 0.66.

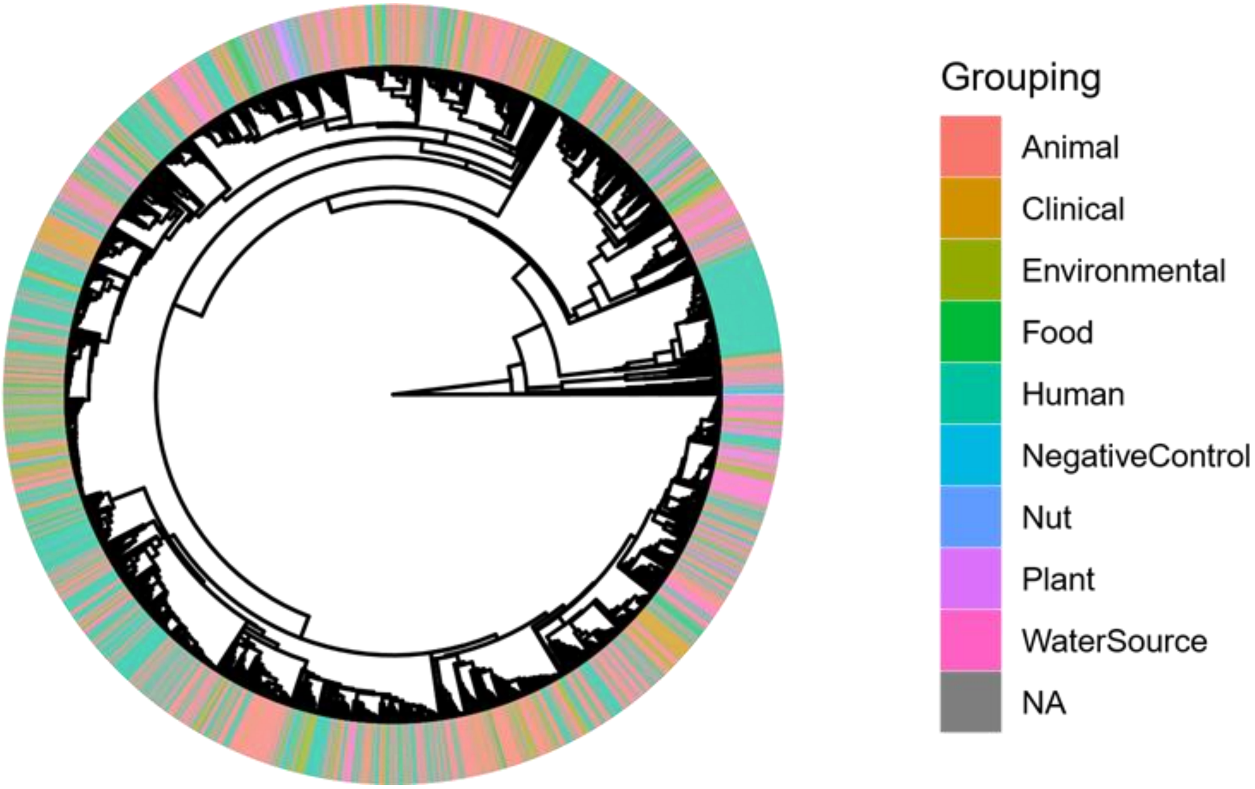


Figure 16. Pfam-based Hierarchical cluster of *Salmonella* genomes. The human isolation group was the only one with a large distinct cluster, as well as a lot of smaller clusters. All groups mixed with one another, and none were distinctly separated from the rest.

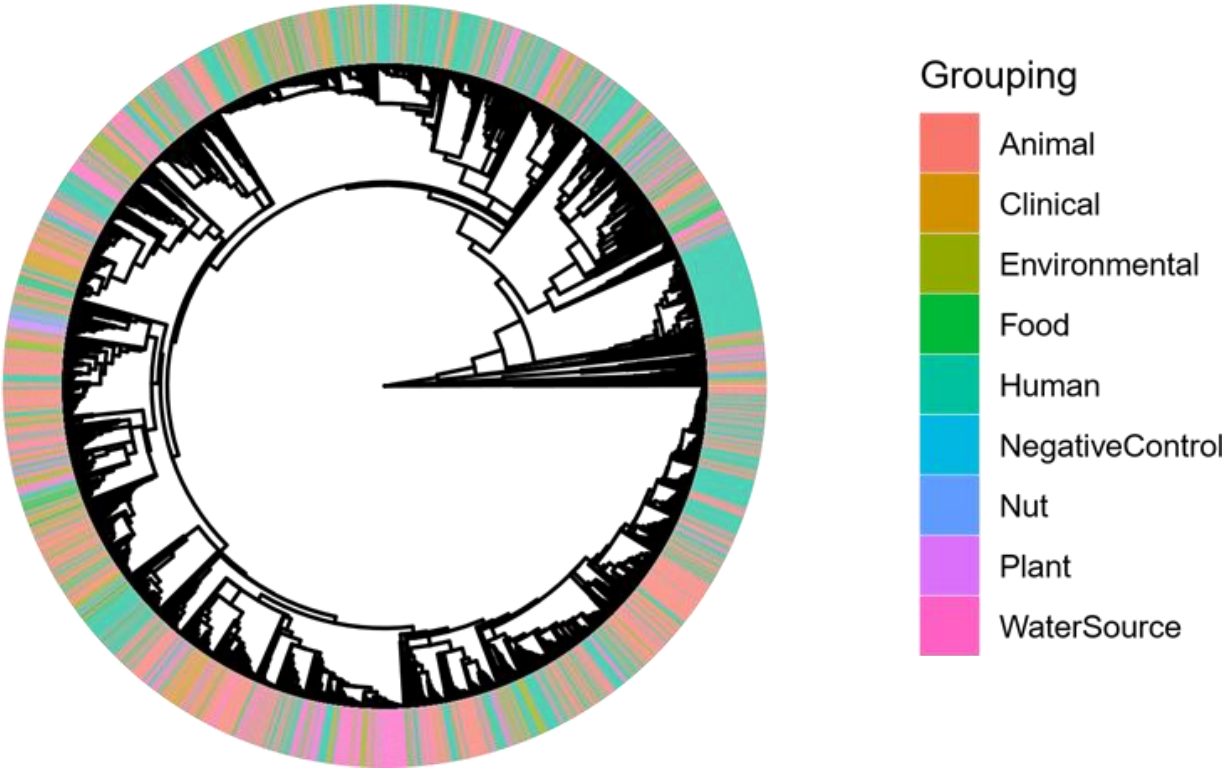


Figure 17. VFDB-based hierarchical cluster of *Salmonella* genomes. As in the Pfam-based cluster, the human group produced the largest and most abundant clusters, followed by the animal group. Each group did not separate into distinct clusters, but were found mixed throughout.

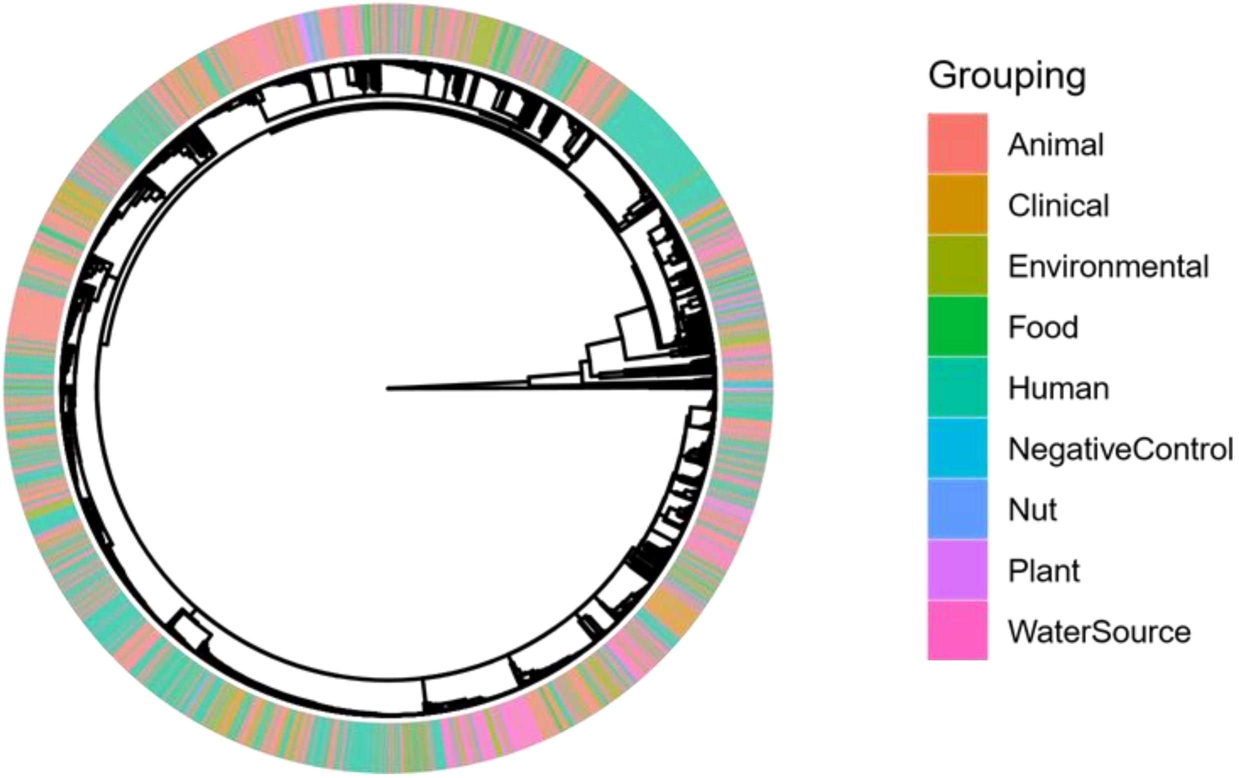


Figure 18. Mash distance-based hierarchical cluster of *Salmonella* genomes. The human group had one large distinct cluster (top right), with many smaller clusters spread throughout. The water source group produced a large relatively distinct cluster (bottom right), as well as the animal group (left side).

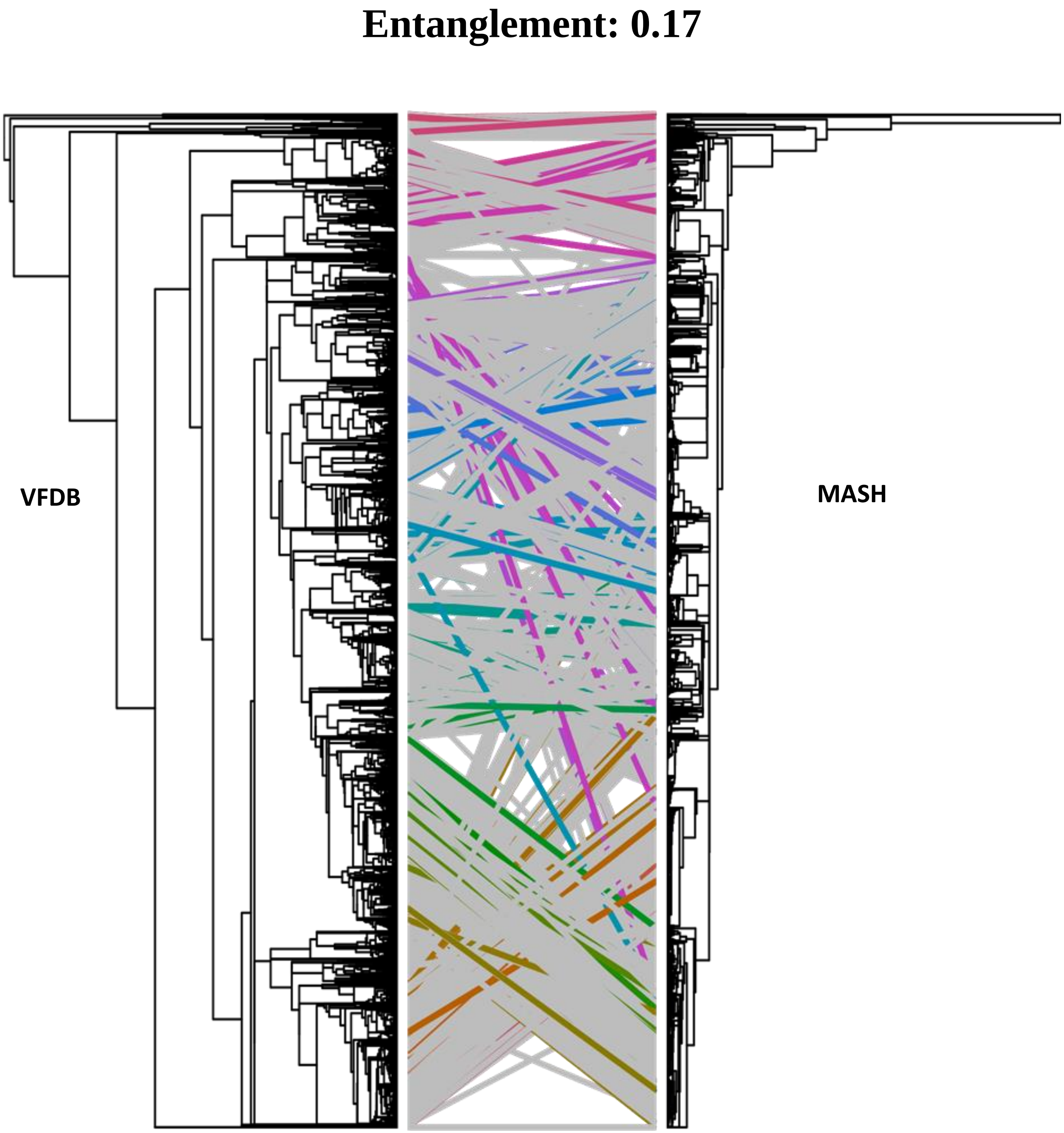


Figure 19. Comparison between hierarchical clusters of VFDB matches (left) and Mash scores (right) for all *Salmonella* genome groups. With an entanglement score of 0.17, the two clusters have little entanglement, and the clustering metrics produce similar results. The clusters had a corresponding Baker's Gamma correlation coefficient of 0.73.

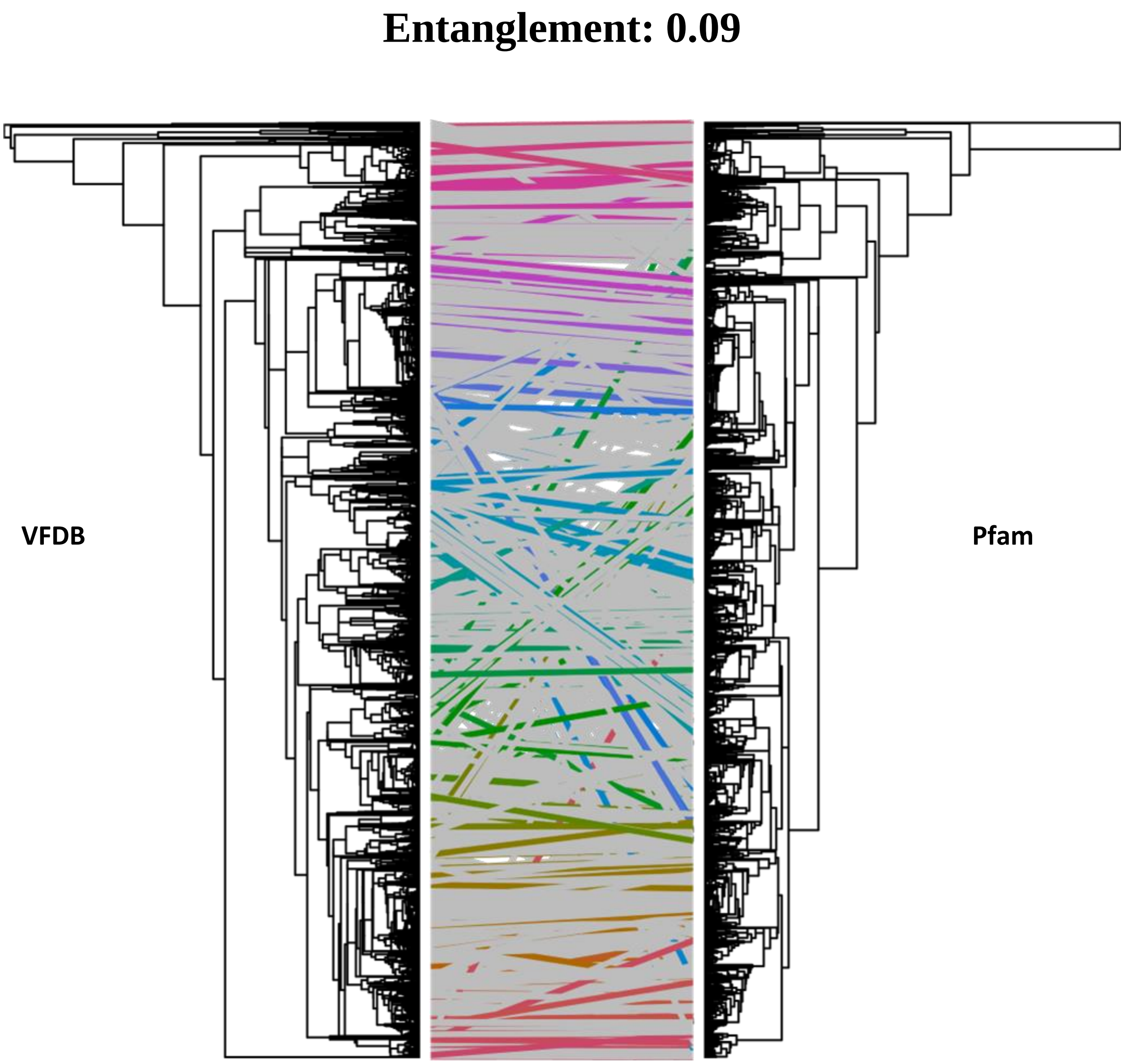


Figure 20. Comparison between hierarchical clusters of VFDB and Pfam matches for all *Salmonella* genome groups. With an entanglement score of 0.09, the two clusters have very little entanglement, and produce similar results. The clusters had a corresponding Baker's Gamma correlation coefficient of 0.66.

Ch. 4 - Discussion

4.1 Analysis of the initial clinical and environmental *Salmonella*

The initial genome dataset consisted of the environmental samples isolated primarily from Clair and Silver Lakes in the Kitchener/Waterloo region (Thomas *et al*., 2013) and clinical samples of known virulence taken from the Pathosystems Resource Integration Center (PATRIC) database, now known as the Bacterial and Viral Bioinformatics Resource Center (BV-BRC). The initial goal was to compare the two genome datasets and determine the pathogenic potential of the newly isolated environmental samples to those of the known disease-causing clinical samples. Both genome sets were annotated with Prokka and aligned against the VFDB and CARD using DIAMOND. The clinical genomes had a median of 961 proteins that matched proteins in the VFDB, while the environmental genomes had a median of 968 (Fig. 1). This was an unexpected result, I initially hypothesized that the clinical genomes would have a lot more matches with the VFDB than the environmental genomes. The clinical genomes were taken from a database of known disease-causing bacteria, so I initially thought that would also be reflected in their alignment with the VFDB.

The clinical genomes had a median 27.41 matches with the CARD, while the environmental genomes had a lower median of 13.81 matches (Fig. 4). This was a somewhat predicted result because I hypothesized that the clinical samples would have more antibiotic resistance related genes because they are taken from a database of known disease-causing pathogens. Patients infected with these pathogenic bacteria would likely have been exposed to antimicrobial drugs during their treatment process. Exposer to antimicrobial drugs may give rise to antimicrobial resistance, as well as genomic differences in their amount of antimicrobial resistance genes.

The genome datasets did not group separately from each other when clustered by ANI, VFDB and CARD matches (Figs. 5 and 6). Instead, the hierarchical clusters showed the clinical and environmental genomes interspersed with each other. Thus, the pathogenicity potential differences were not related to the genome similarity, or to similarity in specific protein family content. The CARD cluster had a very large branch that contained almost half of the genomes in the cluster (Fig. 5). All the genomes in that branch matched the same proteins in the CARD, but there was insufficient data to distinguish them based on the number of matches with CARD. Alignment with the CARD on average produced 13.81 matches for the environmental genomes, and 27.41 for the clinical genomes (Fig. 4). With so few matches, there might not be enough information to properly differentiate genomes from one another. The VFDB and ANI had little relation because of the high entanglement value produced when both hierarchical clusters were compared to one another (Fig. 9). Since ANI is a metric of how similar genomes are to one another (Ciufo *et al*., 2018) genomes that were similar to one another did not match the same proteins in the VFDB (Fig. 9). Comparing clusters produced with CARD and ANI showed a similar result, with an entanglement value of 0.64. Therefore, the overall genome similarity had no relationship on matches with the CARD and antibiotic resistance potential (Fig. 8). The hierarchical clusters made from the CARD and VFDB matches had a large entanglement value of 0.59, meaning that both clusters were very different from each other (Fig. 7). The proteins from the CARD and VFDB matched different proteins in the *Salmonella* genomes, and there was no relationship between the CARD and VFDB matches, as suggested by the entanglement value

(Fig. 7). Though these results are apparently contradictory, ANI is based only on the parts of the genomes that can be aligned with each other (Ciufo *et al*., 2018), thus shared, while CARD and VFDB matches can happen against any proteins (Camacho *et al*, 2009 and Newell *et al*., 2013), whether the genes coding for them are shared or not.

Since both genome sets could not be differentiated by the pairwise alignment with the VFDB, Hidden Markov models (HMM) were employed to potentially produce a distinction between the two. With the initial environmental and clinical datasets, HMMs were going to be used to reduce the number of false positives that may have arisen from a standard pairwise alignment. HMMs tend to have higher specificity than pairwise alignments because they can take into account the amino acid (or nucleotide) that came before and after an amino acid (or nucleotide) during an alignment (Yoon, 2009).

HMMs would reduce false positives by ensuring that the protein of interest is not matching any unwanted proteins that might flag a non-pathogenic protein as a pathogenic one. This can happen when a pathogenicity-related protein is related by homology with a non-pathogenic protein homolog **(**Diepol *et al*., 2015). For example, Type III Secretion Systems (T3SS) share numerous closely related proteins with flagella proteins (Diepol *et al*., 2015). Flagella proteins could potentially be falsely flagged as a pathogenicity-related protein. A program using HMMs was going to be developed by me, to address this problem and help to determine a bacterium’s pathogenic potential more accurately. However, an article was published producing both HMM for VFDB and software to address this exact issue (Nies *et al*., 2021), while I was still working on mine. Thus, instead of producing my own HMMs and software pipeline, I worked on trying to implement the other group's software and use their HMM. Before proceeding, I decided to expand the dataset of *Salmonella* genomes.

4.2 Expanding the dataset to include ~6000 *Salmonella* genomes from different isolation sources.

The dataset of *Salmonella* genomes was expanded to try and incorporate ~11,000 genomes taken from the PATRIC (Now BV-BRC) database as of December 2021, as well as their metadata information. The genomes were then sorted and grouped by their isolation source listed in the metadata table (Figure 21 in the Appendix). Some genomes had either no isolation source information or there was not a sufficient amount to be placed into a group, resulting in the number of genomes being trimmed down to ~6000. Some overlap between the different isolation sources should be noted due to the limited information that was available in the metadata table. Genomes listed with an unspecified "Clinical" isolation source could have also potentially categorized into the "Human" group since they were most likely taken from a human in a clinical setting. Genomes listed with an unspecified "Environmental" isolation source could have been grouped together with those with the water source isolation. "Water source" was made its own distinct group because of the abundant number of genomes that had a more specific isolation source than just "environmental". The negative control dataset was constructed with genomes that were closely related to *Salmonella* and were classified as non-pathogenic (Cosentino *et al*., 2013). A total of 14 genomes were used to create the negative control dataset and were compared with the *Salmonella* isolated from multiple sources.

A Hidden Markov Model (HMM) made from the Virulence Factor Database (VFDB) was constructed by de Nies *et al* (2021). Unfortunately, I was unable to install the full software pipeline developed by the authors, despite trying to install it under different operating systems (Darwin and Linux). Still, I aligned this HMM dataset with all sets of *Salmonella* proteins. With the added sensitivity and specificity that HMMs can provide, isolation sources could be better distinguished from one another. This was not the case and the alignment produced far too many matches when compared to the total number of genes in a *Salmonella's* genome. *Salmonella* genomes typically have ~5000 coding genes (Stevens *et al*., 2018) and the alignment with the HMM VFDB produced ~3000 matches (Fig. 12), this would mean that over half the genome would be related to the bacterium's virulence. The setup and construction of the HMM may play a large role in its efficacy and should be fine tuned to a project's desired needs. The HMM used also contained HMMs from multiple different databases and not just the VFDB, which could have played a part in the number of matches that were produced. It is also possible that the software pipeline I could not install would have cleaned the HMM results and reduced the total number of matches.

The pairwise alignment with the Virulence Factor Database (Non-HMM) resulted in the negative control genomes producing less matches than all other genome groups (Fig. 11). This was an expected result since the negative control genomes were chosen because they should be less pathogenic than regular *Salmonella* genomes. The negative controls having fewer matches than the other genome groups reinforced the concept that an alignment with the VFDB can differentiate non-pathogenic from pathogenic bacteria (Liu *et al*., 2022). There was a clear visual decrease in the number of matches with the VFDB (Fig. 11), but this was reaffirmed after an unpaired two-sample Wilcoxon test showed that there was also a statistically significant difference between the negative controls and all other groups (Table 2). A surprising result was that even though the other groups had a similar number of matches to the VFDB, most were found to be significantly different from one another (Table 2). The only groups that were consistently not significantly different from one another were the plant and nut groups, which may have been because those groups had very few genomes compared to the others. Even with the nut and plant grouping not being as large as the other groups, they were included because they are a vastly different environment than an animal or human, this difference in conditions may have given rise to genomic differences as well. The genomes from the nut and plant groups could be reclassified into the environmental grouping, or more genomes added to those groups to have a more accurate representation.

All genome groups had a large number of total matches with the VFDB, approximately 1000-1100, which represents between one fifth and one quarter of the total number of genes in a *Salmonella's* genome. Similar genome analysis projects with *S. enterica* saw 193 genes that matched the VFDB (Cui *et al*., 2021), but the matches were also filtered by sequence identity, as well as by coverage of the VFDB protein. Sequence identity is how many matches are shared between two sequences as a percentage, while sequence coverage is the percentage of the protein length covered by the alignment (Newell *et al*., 2013). A coverage of at least 70% of the VFDB matched protein was used in the alignments. A sequence identity threshold was not used because it may exclude genes and protein sequences that have diverged a lot, but the protein still retained its original function. Different percent identity thresholds could be tested in the future to try and filter out some non-pathogenicity related sequences that may have matched with a virulence factor, while still maintaining a high percent coverage.

The Comprehensive Antibiotic Resistance Database (CARD) has frequently been used to help determine the pathogenic potential of a bacterium by providing information on antimicrobial resistance genes (Alcock *et al*., 2023). Antimicrobial resistance genes are key features for pathogens that help them persist in an environment and propagate diseases. The negative controls had more proteins that produced a match with the CARD, with all matches being either strict or perfect (Fig. 14). Proteins that produced a "perfect" match were identical to the sequence in the CARD, while a "strict" match is a protein that is similar to a functional variant in the CARD (Alcock *et al*., 2023). Even when filtered for just perfect matches, the negative controls produced more matches with the CARD than that of the other genome groups (Fig. 15). A large majority of the matches were to efflux pumps in the CARD database, a common antimicrobial resistance mechanism that will rapidly pump out antimicrobials from a cell. This was also consistent with the CARD:Live project, where 60% of genomes submitted were found to have an antimicrobial resistant gene (ARG) from the major facilitator superfamily antibiotic efflux pumps (Alcock *et al*., 2023).

Efflux pumps are an intrinsic gene found chromosomally on all bacteria and are used to transport molecules in and out of a cell. High-level multidrug resistance is often caused by many different resistance mechanisms working synergistically, with the efflux pump being a gateway to the activation of other mechanisms in some instances (Saw *et al*., 2016; Nolivos *et al*., 2019; Buckner *et al*., 2017). For example, if the *acrAB* gene is deleted or inhibited in *Enterobacterales,* it had downstream effects and can lead to decreased expression of *OmpF*, a gene that encodes for outer membrane protein F (Porins) (Saw *et al*., 2016). Porins are outer membrane proteins that allow for the diffusion of molecules into a cell, such as antimicrobials. With decreased expression of the *OmpF* gene, less channels would be available for an antimicrobial to enter a cell, reducing its effectiveness. Efflux pumps were also found to be instrumental in acquiring new plasmids through horizontal gene transfer in *E. coli* (Nolivos *et al*., 2019). For example, in the presence of the antibiotic tetracycline, which inhibits translation in bacteria, the AcrAB-TolC efflux pump was needed for the intake of plasmids that carry the gene for tetracycline resistance; TetA (Nolivos *et al*., 2019). The AcrAB-TolC efflux pump would rapidly pump out incoming tetracycline, so the newly acquired plasmid with *TetA* can be successfully translated. The same can be seen in *Klebsiella pneumoniae*, where the acquisition of multi-drug-resistant plasmids caused an increase in the transcription of efflux genes (Buckner *et al*., 2017). This would help explain why most of the matches with the CARD in all groups were efflux pumps. The efflux pump proteins themselves can be used in antimicrobial resistance mechanisms and can also have downstream regulatory effects, as well as being essential in the horizontal gene transfer of plasmids (Saw *et al*., 2016; Nolivos *et al*., 2019; Buckner *et al*., 2017).

The negative controls having more matches with the CARD might be explained by the relationship between antimicrobial resistance and virulence. Antimicrobial resistance has been suggested to come at a fitness cost in an antimicrobial free environment because of the high genetic burden needed for resistance (Beceiro *et al*., 2013). *S. enterica* expressing the *AmpC* gene that codes for resistance to beta-lactam targeting antimicrobials, were found to have decreased invasion rates of target cells and decreased intracellular replication inside the invaded cells (Morosini *et al*., 2000). The *Salmonella* colonies also appeared flattened and rough (normally smooth and raised) when producing AmpC beta-lactamase, but the inclusion of the regulatory *AmpR* gene reversed these characteristics (Morosini *et al*., 2000). Ciprofloxacin-resistant *S. enterica* were found to have lower growth rates and formed smaller colonies when compared to their non-resistant counterparts (O'Regan *et al*., 2010). Resistant *S. enterica* were also more susceptible to environmental stresses such as pH increases and osmotic susceptibility by the addition of salts (O'Regan *et al*., 2010). Ciprofloxacin-resistant *S.enterica* were also found to be more susceptible to other types of antibiotics such as; ampicillin, chloramphenicol and tetracycline (O'Regan *et al*., 2010). These changes also compounded into decreased virulence in the form of significantly decreased swim motility, swarm motility, and invasion of target cells (O'Regan *et al*., 2010). In contrast, other studies showed that antimicrobial resistant *Salmonella* had increased virulence (Tamayo *et al*., 2002 and Eswarappa *et al*., 2008). Other studies have also found that the addition of antimicrobial resistance genes had no associated cost for *Salmonella* due to compensatory mutations (Nilsson *et al*., 2006 and Andersson *et al*., 2010). Since the negative control dataset was primarily composed of *E. coli*, the same three outcomes from the addition of antimicrobial resistance have been observed (Beceiro *et al*., 2013). The negative controls having fewer matches with the VFDB and more matches with the CARD, which suggests that increased antimicrobial resistance decreases virulence.

The PathFam matches displayed a similar result as the pairwise alignment with the VFDB. Matches with the PathFam database helped determine how much pathogenicity-associated protein families there were in each of the groups. The negative controls were found to have less pathogenicity-associated protein families than all other groups (Fig. 13). This difference was found to be statistically significant as well using an unpaired two sample

Wilcoxon test (Table 4). Similar to the matches with the VFDB, the negative controls should have less pathogenicity-associated protein families than the other groups because they should be non-disease-causing bacteria. The PathFam database is shown to be a good tool in differentiating non-pathogenic from pathogenic bacteria and can be used in the future for pathogen identification of an unknown sample. The PathFam database was also better at differentiating the nut and plant groups from the other groups. The nut and plant groups were significantly different from more groups when comparing their PathFam matches than when comparing their VFDB matches (Tables 2 and 3).

When clustered by the VFDB, Pfam and Mash scores, the isolation groups did not separate into corresponding distinct clusters from one another, instead they mixed throughout the hierarchical clusters (Figs. 16,17,18). Aside from one large animal cluster in each dendrogram, no group was found to be distinctly clustered on its own (Figs. 16,17,18). The CARD match clustered because the original dataset clustered (Fig. 5) and showed that there was not sufficient information to properly group genomes, resulting in a large single branch. VFDB matches and Mash scores produced similar results because of their low entanglement score of 0.17 (Fig. 19). With the two clusters having a low level of entanglement, genomes that were similar to one another also matched the same proteins in the VFDB. This relationship was not due to a *Salmeonlla*'s isolation source, if it were, genomes from the same grouping would be clustered together in the hierarchy (Fig. 17). Genomes that are similar to one another were matching the same proteins in the VFDB, but not because they were from the same isolation group. If *Salmonella* from the same groups were matching the same proteins in the VFDB, that would be reflected in large distinct clusters in the hierarchical cluster.

With an entanglement value of 0.09, the VFDB and Pfam clusters were very similar, sharing a lot of the same branches (Fig. 20). Since the Pfam is a protein domain database, and virulence factors are a subset of protein families, the similar clusters were expected. Virulence factors themselves being a small subset of proteins inside larger protein families and domains, they should be similar to matches from the Pfam. This is also a similar case as the relationship between VFDB and Mash (Fig. 19), the VFDB and Pfam matches (Fig. 20) in the *Salmonella* genomes were similar, because the genomes are so similar. If the VFDB and Pfam relationship was due to a *Salmonella*'s isolation source, the groups would have been clustered together in the hierarchical clusters (Figs. 16 and 17). To reiterate, even though these clustering metrics are found to be similar, the causation of this similarity is not due to a *Salmonella*'s isolation source.

With all these results in mind, all isolation groups were found to have similar levels of virulence due to their similar number of matches with the VFDB. Even though most groups were found to be significantly different in this metric, more analysis is needed to further highlight key differences in the groups to provide a definitive answer. The environmental genomes being potentially just as pathogenic as the clinical genomes contrasted my initial hypothesis. *Salmonella* from diverse isolation sources all around the world seem to have the same potential to be pathogenic, regardless of isolation source.

Ch. 5 Future directions

Different percent identity thresholds can be experimented with, to further filter down the total number of matches with the VFDB. HMMs and pairwise alignments could be used in tandem to make use of the benefits of each, while minimizing each other's weakness. HMM models would need to be tailored for specific projects since "one size fits all" models did not give good results. I would also expand the negative control dataset to include more *Salmonella* genomes and increase the size in general. The reclassification of the groupings might be needed to merge some smaller groups with larger ones that consistently did not have a significant difference with other groups. The genome groupings were also be constructed to contain the same number of genomes as each other, to limit the effect of sample size differences in further analysis. Lastly, I would create a more updated workflow for identifying virulence and antimicrobial resistance related genes. More post alignment steps could be utilized to further reduce the total number of matches with the VFDB, or looking at specific genes in the VFDB that are found in *Salmonella.*

Ch. 6 Integrative Biology Statement

To me, integrative biology is the idea of using different fields of biology to solve problems through different perspectives. Biology projects might span several disciplines within biology such as: ecology, pathology, molecular genetics, physiology etc. With the invention of the internet, it has become easier than ever to communicate and share information with anyone around the world. This network has allowed for the coordination between people that have diverse backgrounds and expertise. These collaborations can add new perspectives to problems and questions that may arise and help us to understand large biological systems that may span multiple fields of biology.

My thesis project was integrative in nature because I was able to work with *Salmonella* genomes isolated by Dr. Slawson's lab. They were able to isolate the *Salmonella* from the environment and used molecular genetics techniques to sequence the genomes for me to analyze. My analysis used predominantly computational biology techniques, such as blast alignments, to better understand the *Salmonella*'s potential to cause disease, investigating the pathology of them. I also used statistical analysis techniques such as Hidden Markov Models and hierarchical clusters to better visualize and understand trends within the genomes. This is why I believe that my thesis was integrative, I collaborated with people from different fields and did not strictly use techniques and ideas from my discipline alone.

Ch. 7 Summary

The initial environmental genomes isolated from the Waterloo region and clinical lab isolates were found to have a similar number of matches with the VFDB, suggesting a similar amount of pathogenic potential. The environmental genomes were found to have less matches with the CARD than the clinical lab isolates, suggesting less antibiotic resistant potential than the clinical genomes. The dataset was then expanded to include ~6000 genomes from the BV-BRC database. These new genomes were sorted into different groups based on their isolation source, in order to determine if there is a relationship between isolation source and pathogenic potential. The negative control dataset was found to have significantly less matches with the VFDB, as well as significantly more matches with the CARD than other groups. Most groups were found to be significantly different from others besides the very small nut and plant groups. The PathFam database was able to differentiate between the negative control group because it had less pathogenic associated domains. When clustered by Mash scores, matches with the VFDB, and matches with the CARD, the groups did not produce distinct clusters and were found dispersed throughout. The difference in VFDB matches was not due to different isolation sources but was attributed to the genomes being similar. Isolation source did not seem to have an effect on a *Salmonella's* disease-causing potential, but further analysis is needed to confirm this conclusion.

References

Alcock BP, Huynh W, Chalil R, *et al*. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res*. 2023;51(D1):D690-D699. doi:10.1093/nar/gkac920

Andersson, D. I., & Hughes, D. (2010). Antibiotic resistance and its cost: is it possible to reverse resistance?. Nature reviews. Microbiology, 8(4), 260–271. https://doi.org/10.1038/nrmicro2319

Aziz M, Yelamanchili VS. Yersinia Enterocolitica. [Updated 2023 Jul 3]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK499837/

Beceiro, A., Tomás, M., & Bou, G. (2013). Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world?. Clinical microbiology reviews, 26(2), 185–230. https://doi.org/10.1128/CMR.00059-12

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods. Volume 12: 59-60.

Buckner, M. M. C., Saw, H. T. H., Osagie, R. N., McNally, A., Ricci, V., Wand, M. E., Woodford, N., Ivens, A., Webber, M. A., & Piddock, L. J. V. (2018). Clinically Relevant Plasmid-Host Interactions Indicate that Transcriptional and Not Genomic Modifications Ameliorate Fitness Costs of Klebsiella pneumoniae Carbapenemase-Carrying Plasmids. mBio, 9(2), e02303-17. https://doi.org/10.1128/mBio.02303-17

Carrol AC, Wong A. 2018. Plasmid persistence: costs, benefits, and the plasmid paradox. Canadian Journal of Microbiology. Volume 65(5): 293-304. Center for Disease Control and Prevention. 2020. *Salmonella*

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC bioinformatics*, *10*, 421. https://doi.org/10.1186/1471-2105-10-421

Ciufo S, Kannan S, Sharma S, Badretdin *et al*. 2018. Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. Internal Journal of Systematic and Evolutionary Microbiology. Volume 68(7): 2386-2392.

Cosentino S, Voldby Larsen M, Møller Aarestrup F, Lund O (2013) PathogenFinder - Distinguishing Friend from Foe Using Bacterial Whole Genome Sequence Data. PLoS ONE 8(10): e77302. https://doi.org/10.1371/journal.pone.0077302

Cui, L., Wang, X., Zhao, Y., Peng, Z., Gao, P., Cao, Z., Feng, J., Zhang, F., Guo, K., Wu, M., Chen, H., & Dai, M. (2021). Virulence Comparison of *Salmonella* enterica Subsp. enterica Isolates from Chicken and Whole Genome Analysis of the High Virulent Strain S. Enteritidis 211. Microorganisms, 9(11), 2239. https://doi.org/10.3390/microorganisms9112239

Darby, E. M., Trampari, E., Siasat, P., Gaya, M. S., Alav, I., Webber, M. A., & Blair, J. M. A. (2023). Molecular mechanisms of antibiotic resistance revisited. *Nature reviews. Microbiology*, *21*(5), 280–295. https://doi.org/10.1038/s41579-022-00820-y

Daubin, V., & Szöllősi, G. J. (2016). Horizontal Gene Transfer and the History of Life. *Cold Spring Harbor perspectives in biology*, *8*(4), a018036. https://doi.org/10.1101/cshperspect.a018036

de Nies, L., Lopes, S., Busi, S.B. *et al.* PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* **9**, 49 (2021). https://doi.org/10.1186/s40168-020-00993-9

De Pascale, G., & Wright, G. D. (2010). Antibiotic resistance by enzyme inactivation: from mechanisms to solutions. *Chembiochem : a European journal of chemical biology*, *11*(10), 1325–1334. https://doi.org/10.1002/cbic.201000067

Desvaux, M., Dalmasso, G., Beyrouthy, R., Barnich, N., Delmas, J., & Bonnet, R. (2020). Pathogenicity Factors of Genomic Islands in Intestinal and Extraintestinal *Escherichia coli*. *Frontiers in microbiology*, *11*, 2065. https://doi.org/10.3389/fmicb.2020.02065

Diepold, A., & Armitage, J. P. (2015). Type III secretion systems: the bacterial flagellum and the injectisome. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 370(1679), 20150020. https://doi.org/10.1098/rstb.2015.0020

Eswarappa, S. M., Panguluri, K. K., Hensel, M., & Chakravortty, D. (2008). The yejABEF operon of *Salmonella* confers resistance to antimicrobial peptides and contributes to its virulence. Microbiology (Reading, England), 154(Pt 2), 666–678. https://doi.org/10.1099/mic.0.2007/011114-0

Farhana A, Khan YS. Biochemistry, Lipopolysaccharide. [Updated 2023 Apr 17]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554414/

Ferrari, R. G., Rosario, D., Cunha-Neto, A., Mano, S. B., Figueiredo, E., & Conte-Junior, C. A. (2019). Worldwide Epidemiology of *Salmonella* Serovars in Animal-Based Foods: a Meta-analysis. *Applied and environmental microbiology*, *85*(14), e00591-19. https://doi.org/10.1128/AEM.00591-19

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. *Nucleic acids research*, *42*(Database issue), D222–D230. https://doi.org/10.1093/nar/gkt1223

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research, 39(Web Server issue), W29–W37. https://doi.org/10.1093/nar/gkr367

Galili T (2015). “dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering.” Bioinformatics. doi:10.1093/bioinformatics/btv428

Galli, T., Chilcote, T., Mundigl, O., Binz, T., Niemann, H., & De Camilli, P. (1994). Tetanus toxin-mediated cleavage of cellubrevin impairs exocytosis of transferrin receptor-containing vesicles in CHO cells. *The Journal of cell biology*, *125*(5), 1015–1024. https://doi.org/10.1083/jcb.125.5.1015

Gyawali, B., Ramakrishna, K., & Dhamoon, A. S. (2019). Sepsis: The evolution in definition, pathophysiology, and management. *SAGE open medicine*, *7*, 2050312119835043. https://doi.org/10.1177/2050312119835043

Hauser M, Steinegger M, Soding J. 2016. MMSeqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics. Volume 32(9): 1323-1330.

Hernández-Salmerón, J. E., & Moreno-Hagelsieb, G. (2022). FastANI, Mash and Dashing equally differentiate between *Klebsiella* species. *PeerJ*, *10*, e13784. https://doi.org/10.7717/peerj.13784

Hernández-Salmerón JE, Irani T, Moreno-Hagelsieb G. Fast genome-based species delimitation: Enterobacterales and beyond. bioRxiv; 2023. DOI: 10.1101/2023.04.05.535762.

Higgins, D., Mukherjee, N., Pal, C., Sulaiman, I. M., Jiang, Y., Hanna, S., Dunn, J. R., Karmaus, W., & Banerjee, P. (2020). Association of Virulence and Antibiotic Resistance in *Salmonella*-Statistical and Computational Insights into a Selected Set of Clinical Isolates. *Microorganisms*, *8*(10), 1465. https://doi.org/10.3390/microorganisms8101465

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. Journal of computational and graphical statistics, 5(3), 299-314.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature communications, 9(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Jajere S. M. (2019). A review of *Salmonella enterica* with particular focus on the pathogenicity and virulence factors, host specificity and antimicrobial resistance including multidrug resistance. *Veterinary world*, *12*(4), 504–521. https://doi.org/10.14202/vetworld.2019.504-521

Jaina Mistry and others, Pfam: The protein families database in 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D412–D419, https://doi.org/10.1093/nar/gkaa913

Jurafsky D and Martin JH. (2020). Speech and Language Processing: Hidden Markov models. *Stanford University*. Chapter A: p1-17.

Kombade, S., & Kaur, N. (2021). Pathogenicity Island in *Salmonella*. IntechOpen. doi: 10.5772/intechopen.96443

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). VFDB 2022: a general classification scheme for bacterial virulence factors. Nucleic acids research, 50(D1), D912–D917. https://doi.org/10.1093/nar/gkab1107

Lobb, B., Tremblay, B.JM., Moreno-Hagelsieb, G. *et al.* PathFams: statistical detection of pathogen-associated protein domains. *BMC Genomics* 22, 663 (2021). https://doi.org/10.1186/s12864-021-07982-8

McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, Bhullar K, Canova MJ, De Pascale G, Ejim L, Kalan L, King AM, Koteva K, Morar M, Mulvey MR, O'Brien JS, Pawlowski AC, Piddock LJ, Spanogiannopoulos P, Sutherland AD, Tang I, Taylor PL, Thaker M, Wang W, Yan M, Yu T, Wright GD. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother. 2013 Jul;57(7):3348-57. doi: 10.1128/AAC.00419-13. Epub 2013 May 6. PMID: 23650175; PMCID: PMC3697360.

Messerer, M., Fischer, W., & Schubert, S. (2017). Investigation of horizontal gene transfer of pathogenicity islands in Escherichia coli using next-generation sequencing. *PloS one, 12*(7), e0179880. https://doi.org/10.1371/journal.pone.0179880

Morosini, M. I., Ayala, J. A., Baquero, F., Martínez, J. L., & Blázquez, J. (2000). Biological cost of AmpC production for *Salmonella* enterica serotype Typhimurium. Antimicrobial agents and chemotherapy, 44(11), 3137–3143. https://doi.org/10.1128/AAC.44.11.3137-3143.2000

Newell, P. D., Fricker, A. D., Roco, C. A., Chandrangsu, P., & Merkel, S. M. (2013). A Small-Group Activity Introducing the Use and Interpretation of BLAST. Journal of microbiology & biology education, 14(2), 238–243. https://doi.org/10.1128/jmbe.v14i2.637

Nilsson, A. I., Zorzet, A., Kanth, A., Dahlström, S., Berg, O. G., & Andersson, D. I. (2006). Reducing the fitness cost of antibiotic resistance by amplification of initiator tRNA genes. Proceedings of the National Academy of Sciences of the United States of America, 103(18), 6976–6981. https://doi.org/10.1073/pnas.0602171103

Nolivos, S., Cayron, J., Dedieu, A., Page, A., Delolme, F., & Lesterlin, C. (2019). Role of AcrAB-TolC multidrug efflux pump in drug-resistance acquisition by plasmid transfer. Science (New York, N.Y.), 364(6442), 778–782. https://doi.org/10.1126/science.aav6390

Olson, R. D., Assaf, R., Brettin, T., Conrad, N., Cucinell, C., Davis, J. J., Dempsey, D. M., Dickerman, A., Dietrich, E. M., Kenyon, R. W., Kuscuoglu, M., Lefkowitz, E. J., Lu, J., Machi, D., Macken, C., Mao, C., Niewiadomska, A., Nguyen, M., Olsen, G. J., Overbeek, J. C., … Stevens, R. L. (2023). Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. Nucleic acids research, 51(D1), D678–D689. https://doi.org/10.1093/nar/gkac1003

Ondov BD, Treangen TJ, Melsted P, *et al*. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17(1):132. Published 2016 Jun 20. doi:10.1186/s13059-016-0997-x

O'Regan, E., Quinn, T., Frye, J. G., Pagès, J. M., Porwollik, S., Fedorka-Cray, P. J., McClelland, M., & Fanning, S. (2010). Fitness costs and stability of a high-level ciprofloxacin resistance phenotype in *Salmonella* enterica serotype enteritidis: reduced infectivity associated with decreased expression of *Salmonella* pathogenicity island 1 genes. Antimicrobial agents and chemotherapy, 54(1), 367–374. https://doi.org/10.1128/AAC.00801-09

Park J, Karplus K, Barrett C, Hughey R, Haussler D, *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol 284: 1201–1210

Parween, F., Yadav, J., & Qadri, A. (2019). The Virulence Polysaccharide of *Salmonella* Typhi Suppresses Activation of Rho Family GTPases to Limit Inflammatory Responses From

Epithelial Cells. Frontiers in cellular and infection microbiology, 9, 141. https://doi.org/10.3389/fcimb.2019.00141

Persat, A., Inclan, Y. F., Engel, J. N., Stone, H. A., & Gitai, Z. (2015). Type IV pili mechanochemically regulate virulence factors in Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(24), 7563–7568. https://doi.org/10.1073/pnas.1502025112

Popoff M. R. (2018). "Bacterial Toxins" Section in the Journal Toxins: A Fantastic Multidisciplinary Interplay between Bacterial Pathogenicity Mechanisms, Physiological Processes, Genomic Evolution, and Subsequent Development of Identification Methods, Efficient Treatment, and Prevention of Toxigenic Bacteria. *Toxins*, *10*(1), 44. https://doi.org/10.3390/toxins10010044

Raetz, C. R., & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annual review of biochemistry*, *71*, 635–700. https://doi.org/10.1146/annurev.biochem.71.110601.135414

Reygaert W. C. (2018). An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS microbiology*, *4*(3), 482–501. https://doi.org/10.3934/microbiol.2018.3.482

Saw, H. T., Webber, M. A., Mushtaq, S., Woodford, N., & Piddock, L. J. (2016). Inactivation or inhibition of AcrAB-TolC increases resistance of carbapenemase-producing Enterobacteriaceae to carbapenems. The Journal of antimicrobial chemotherapy, 71(6), 1510–1519. https://doi.org/10.1093/jac/dkw028

Seeman T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics. Volume 30(14): 2068-2069.

Sheehan, J. R., Sadlier, C., & O'Brien, B. (2022). Bacterial endotoxins and exotoxins in intensive care medicine. *BJA education*, *22*(6), 224–230. https://doi.org/10.1016/j.bjae.2022.01.003

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current protocols in molecular biology*, *122*(1), e59. https://doi.org/10.1002/cpmb.59

Śmiechowicz J. (2022). The Rationale and Current Status of Endotoxin Adsorption in the Treatment of Septic Shock. *Journal of clinical medicine*, *11*(3), 619. https://doi.org/10.3390/jcm11030619

Stevens, M. J. A., Zurfluh, K., Althaus, D., Corti, S., Lehner, A., & Stephan, R. (2018). Complete and Assembled Genome Sequence of *Salmonella* enterica subsp. enterica Serotype Senftenberg N17-509, a Strain Lacking *Salmonella* Pathogen Island 1. Genome announcements, 6(12), e00156-18. https://doi.org/10.1128/genomeA.00156-18

Stromberg, Z. R., Van Goor, A., Redweik, G. A. J., Wymore Brand, M. J., Wannemuehler, M. J., & Mellata, M. (2018). Pathogenic and non-pathogenic *Escherichia coli* colonization and host inflammatory response in a defined microbiota mouse model. *Disease models & mechanisms*, *11*(11), dmm035063. https://doi.org/10.1242/dmm.035063

Tamayo, R., Ryan, S. S., McCoy, A. J., & Gunn, J. S. (2002). Identification and genetic characterization of PmrA-regulated genes and genes involved in polymyxin B resistance in *Salmonella* enterica serovar typhimurium. Infection and immunity, 70(12), 6770–6778. https://doi.org/10.1128/IAI.70.12.6770-6778.2002

Thomas, J. L., Slawson, R. M., & Taylor, W. D. (2013). Salmonella serotype diversity and seasonality in urban and rural streams. *Journal of applied microbiology*, *114*(3), 907–922. https://doi.org/10.1111/jam.12079

Uliczka, F., Pisano, F., Schaake, J., Stolz, T., Rohde, M., Fruth, A., Strauch, E., Skurnik, M., Batzilla, J., Rakin, A., Heesemann, J., & Dersch, P. (2011). Unique cell adhesion and invasion properties of Yersinia enterocolitica O:3, the most frequent cause of human Yersiniosis. *PLoS pathogens*, *7*(7), e1002117. https://doi.org/10.1371/journal.ppat.1002117

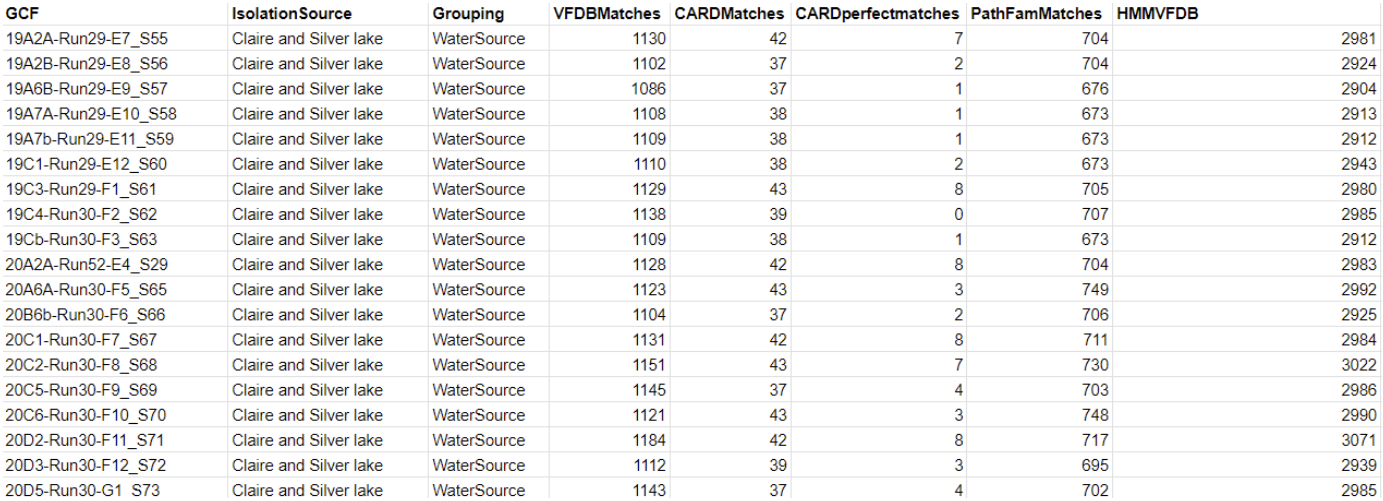
Ursell, L. K., Metcalf, J. L., Parfrey, L. W., & Knight, R. (2012). Defining the human microbiome. *Nutrition reviews*, *70 Suppl 1*(Suppl 1), S38–S44. https://doi.org/10.1111/j.1753-4887.2012.00493.x

Vestby, L. K., Grønseth, T., Simm, R., & Nesse, L. L. (2020). Bacterial Biofilm and its Role in the Pathogenesis of Disease. *Antibiotics (Basel, Switzerland)*, *9*(2), 59. https://doi.org/10.3390/antibiotics9020059

Yoon B. J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current genomics*, *10*(6), 402–415. https://doi.org/10.2174/138920209789177575

Zhang, Z., Murtagh, F., Van Poucke, S., Lin, S., & Lan, P. (2017). Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. Annals of translational medicine, 5(4), 75. https://doi.org/10.21037/atm.2017.02.05

Appendix



GCF	IsolationSource	Grouping	VFDBMatches	CARDMatches	CARDperfectmatches	PathFamMatches	HMMVFDB
19A2A-Run29-E7_S55	Claire and Silver lake	WaterSource	1130	42	7	704	2981
19A2B-Run29-E8_S56	Claire and Silver lake	WaterSource	1102	37	2	704	2924
19A6B-Run29-E9_S57	Claire and Silver lake	WaterSource	1086	37	1	676	2904
19A7A-Run29-E10_S58	Claire and Silver lake	WaterSource	1108	38	1	673	2913
19A7b-Run29-E11_S59	Claire and Silver lake	WaterSource	1109	38	1	673	2912
19C1-Run29-E12_S60	Claire and Silver lake	WaterSource	1110	38	2	673	2943
19C3-Run29-F1_S61	Claire and Silver lake	WaterSource	1129	43	8	705	2980
19C4-Run30-F2_S62	Claire and Silver lake	WaterSource	1138	39	0	707	2985
19Cb-Run30-F3_S63	Claire and Silver lake	WaterSource	1109	38	1	673	2912
20A2A-Run52-E4_S29	Claire and Silver lake	WaterSource	1128	42	8	704	2983
20A6A-Run30-F5_S65	Claire and Silver lake	WaterSource	1123	43	3	749	2992
20B6b-Run30-F6_S66	Claire and Silver lake	WaterSource	1104	37	2	706	2925
20C1-Run30-F7_S67	Claire and Silver lake	WaterSource	1131	42	8	711	2984
20C2-Run30-F8_S68	Claire and Silver lake	WaterSource	1151	43	7	730	3022
20C5-Run30-F9_S69	Claire and Silver lake	WaterSource	1145	37	4	703	2986
20C6-Run30-F10_S70	Claire and Silver lake	WaterSource	1121	43	3	748	2990
20D2-Run30-F11_S71	Claire and Silver lake	WaterSource	1184	42	8	717	3071
20D3-Run30-F12_S72	Claire and Silver lake	WaterSource	1112	39	3	695	2939
20D5-Run30-G1 S73	Claire and Silver lake	WaterSource	1143	37	4	702	2985

Figure 21. Snapshot of information for all genomes used in this study, contains isolation source information and grouping information, as well as counts from all alignments. Full table available upon request.