ABSTRACT

Title of Dissertation: DEVELOPMENT OF MACHINE LEARNING AND ADVANCED DATA ANALYTICAL TECHNIQUES TO INCORPORATE GENOMIC DATA IN PREDICTIVE MODELING FOR *SALMONELLA ENTERICA*

Shraddha Karanth, Doctor of Philosophy, 2021

Dissertation directed by: Dr. Abani K. Pradhan, Associate Professor, Department of Nutrition and Food Science

The past few decades have seen a renaissance in the field of food safety, with the increasing usage of genomic data (e.g., whole genome sequencing (WGS)) in determining the cause of microbial foodborne illness, particularly for multi-serovar

agents such as *Salmonella enterica*. However, utilizing such data in a preventative framework, specifically in the field of quantitative microbial risk assessment (QMRA) remains in its infancy, because incorporating such large-scale datasets in statistical models is hindered by the sheer number of variables/features introduced. Thus, the goal of this research is to introduce machine learning (ML)-based approaches to potentially incorporate WGS data in various stages of a risk assessment for *Salmonella enterica*.

Specifically, we developed a machine learning-based workflow to obtain an association between gene presence/absence data from microbial whole genome sequences and severity of *Salmonella*-related health outcomes in host systems. A key contribution of this dissertation is assessing the applicability of Elastic Net model, a recursive feature selection technique, which resolves a well-known issue concerning WGS-based data analysis: variables/features outnumber the count of observations. Building on this finding, we developed a gene weighted Poisson regression method to incorporate genes into a dose-response framework for *Salmonella enterica*, thereby incorporating genetic variability directly into a risk assessment framework. Finally, we combined machine learning with count-based models to determine how significant genes interact with meteorological factors in impacting the severity of salmonellosis outbreaks.

This dissertation uncovers some interesting findings. First, although commonly used classifiers (such as random forest) performed well in predicting disease severity, logistic regression, in conjunction with Elastic Net, performed significantly better. This finding is important, as the result of a logistic regression is generally more interpretable than that of other classifiers, easing its incorporation into predictive microbial

modeling. Next, machine learning-supported count-based models, such as Poisson regression also proved to be a good fit for gene-informed dose-response modeling and determination of outbreak severity when combined with extrinsic factors such as atmospheric temperature and precipitation. Overall, this dissertation identified areas within a QMRA framework that could benefit from incorporating genetic information, and introduced ML models to incorporate such information.

DEVELOPMENT OF MACHINE LEARNING AND ADVANCED DATA ANALYTICAL TECHNIQUES TO INCORPORATE GENOMIC DATA IN PREDICTIVE MODELING FOR *SALMONELLA ENTERICA*

by

Shraddha Karanth

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:

Dr. Abani K. Pradhan, Chair

Dr. Jianghong Meng

Dr. Jitu Patel

Dr. Seong-Ho Lee

Dr. Adel Shirmohammadi

Dedication

This dissertation is dedicated to my grandmother, P. Yashoda, and my uncle, Prakash Rao, for their immeasurable love and support, and constant motivation. I'm sorry you never got to see what I have achieved. You are both dearly missed.

Acknowledgements

This work would not have been possible without the support of many incredible people. First and foremost, I would like to express my deep sense of gratitude to my advisor and mentor, Dr. Abani K. Pradhan. Thank you for accepting me into this program, guiding me through this impossible-at-times research, all of the opportunities, and your infinite patience with me. I cannot imagine a more helpful and resourceful advisor.

I would like to extend special thanks to Dr. Jianghong Meng and Dr. Jitu Patel, for providing me with many opportunities for lab-based learning and professional development. Many thanks also to the members of my dissertation committee – Drs. Meng and Patel, and Dr. Adel Shirmohammadi and Dr. Seong-Ho Lee – for their many insightful comments and suggestions on my research, scientific advice, encouragement, and support.

I am extremely grateful to all current and past members of Dr. Pradhan's lab – Dr. Collins Tanui, Dr. Abhinav Mishra, Taryn Horr, Edmund Benefo, Dr. Hao Pang, Aishwarya Rao, Shuyi Feng, and Yinzhi Qu – for their continuous support, help, and guidance over the past 5 years. Special thanks to Dr. Surabhi Rani for helping me through so many tough times. I would also like to thank faculty and staff at the Department of Nutrition and Food Science for their support, help, and patience. Thank you also to my fellow students and friends, both current and former, at the Department of Nutrition and Food Science – Abby, Andrea, Amy, Bidisha, Bobby, Jihye, Jinglin, Kathy, Lei, Paul, Qiao, Rishov, Ruth, Stratton, and many others.

I could not have achieved all that I have without the support, love, and shoulder-to-cry-on provided by my wonderful family and friends. To my best friend Roshni, all that you have done for me has kept me sane and afloat for too many years. Vaishali, Sabari, and Ahaana – you have been steadfast in supporting me and motivating me through good times and bad, and I could not be more grateful. To my TC family – Neenu, Bhavana, Sruthi, Emmy, Vivek, Shamini, Kailash, Nida, Naresh, Sriman, and Sid – thank you for always being there (and the wild adventures at Norra Fäladen). Kristi, you are a true friend and a rare find. And finally, my family – father Umesh M., mother Veena Rao, sister Shravya, husband Raveesh, and my in-laws – thank you for your belief in me, and your constant encouragement and endless love. Raveesh, you have stood by me and my choices, pushed me to my limits while always providing a helping hand, and simply been a constant source of love, support and strength all these years – words cannot describe my gratitude for having you in my life. Last but not least, thank you Shiva for lighting up our every day – everything we do, we do for you.

Table of Contents

List of Tables

along with the standardized covariates monthly mean average daily temperature (tavg, in °F), monthly mean precipitation (prcp, in inches), and monthly mean snow cover (snow, in inches).

List of Figures

List of Abbreviations

ABC – ATP-binding cassette

AdaBoost – Adaptive Boosting

AIC – Akaike Information Criterion

AMR – Antimicrobial resistance

ANOVA – Analysis of Variance

AUC-ROC – Area under the Curve – Receiver Operating Characteristic curve (also known as 'Area under the receiver operating characteristic curve')

BIC – Bayesian Information Criterion

BLASTP – Protein Basic Local Alignment Search Tool

CFSAN – Center for Food Safety and Applied Nutrition

CFU – Colony Forming Unit

COMPARE – Collaborative Management Platform for Detection and Analyses of (Re)-Emerging and Foodborne Outbreaks in Europe

CRISPR – Clustered Regularly Interspaced Short Palindromic Repeats

CSD – Collaborative Study Design

DDBJ – DNA Data Bank of Japan

EMBL – European Molecular Biology Laboratory

FDOSS – Foodborne Disease Outbreak Surveillance System

FSIS – Food Safety Inspection Service

GMI – Global Microbial Identifier

GTP – Guanosine Triphosphate

GWAS – Genome-Wide Association Studies

LR – Logistic Regression

MAF – Minor Allele Frequency

MLST – Multi-locus Sequence Typing

NARMS – National Antimicrobial Resistance Monitoring System for enteric bacteria

NCBI – National Centers for Biotechnology Information

NCDC – National Climactic Data Center

NCEI – National Centers for Environmental Information

NGS – Next-generation (Gen) Sequencing

NOAA – National Oceanographic and Atmospheric Administration

NORS – National Outbreak Reporting System of the U.S. CDC

PATRIC - Pathosystems Resource Integration Center

PCA – Principal Component Analysis

PFGE – Pulsed Field Gel Electrophoresis

QMRA – Quantitative Microbial Risk Assessment

RASTk – Rapid Annotation using Subsystem Technology

RF – Random Forest

SCM – Set Covering Machine

SNP(s) – Single Nucleotide Polymorphism(s)

SPAdes – St. Petersburg genome assembler

SRA – Sequence Read Archive

SSU – Single strand unit

STEC – Shiga-toxigenic *Escherichia coli*

SVM – Support Vector Machine

U.S. CDC – United States Centers for Disease Control and Prevention

U.S. FDA – United States Food and Drug Administration

USDA – United States Department of Agriculture

WGS – Whole Genome Sequencing

Chapter 1: Introduction

1.1 Predictive microbiology and quantitative microbial risk assessment in food safety

Food safety is a major public health issue in the U.S. and worldwide. This is of particular relevance in modern times when the food being consumed is very diverse and being sourced from across the globe; ensuring that food is "safe" requires a very proactive approach towards identifying the sources of, and factors affecting, contamination, and minimizing them across the farm-to-fork paradigm. Despite measures being taken by regulatory agencies, the industry, and the educational sector, incidences of foodborne disease outbreaks and food recalls in the U.S. and across the globe continue on a steady scale. This clearly shows the need for improvement in the safety and security of our food supply. Foodborne illnesses affect an estimated 1 in 6 Americans every year, causing 128,000 hospitalizations and 3,000 deaths (Scallan et al., 2011; U.S. CDC, 2021a). Food safety risk assessment has gained momentum in recent decades, as it can be used to provide a scientifically sound basis for informed management and policy decisions.

QMRA is a systematic approach to evaluate the likelihood of adverse health effects in humans as a result of exposure to a pathogenic microorganism, which was developed to understand and manage microbial risks to inform risk management practices and policy making (Rantsiou, Mataragas, Jespersen, & Cocolin, 2011). It uses statistical and mathematical models to understand, predict, and prevent the risks posed by pathogenic

microorganisms (Whiting & Buchanan, 1997; Pradhan et al., 2009; Guo et al., 2016 a, b). Briefly, QMRA can be used to predict the behavior and transmission of pathogens across the food production, processing, and supply chain, identify areas in the chain that could lead to contamination, and estimate the probability and consequence of adverse public health effects upon consumption of potentially contaminated products (FAO, 2001; Pradhan et al., 2009; Guo et al., 2016b, 2017; Pang, Lambertini, Buchanan, Schaffner, & Pradhan, 2017). QMRA consists of four steps: (i) hazard identification – which involves gaining knowledge about the microorganism and its association with adverse health effects in the host; (ii) hazard characterization – the likelihood of infection given the dose and the consequences of infection; (iii) exposure assessment – wherein the numbers of microorganisms, and the impact of various processing techniques on these numbers, are assessed to estimate the microbial quantities in the final food product; and (iv) risk characterization – wherein the level of risk to the exposed individual is estimated. This can then be employed to make informed management decisions based on the risk of microbial contamination to human health.

1.2. Whole genome sequencing

1.2.1. Background and applications

Whole genome sequencing is a method that can be used to provide a detailed characterization of an organism. Recently, we have seen an upswing in the use of WGS to describe what makes up a foodborne pathogen, primarily due to its rapid turnover time and cost effectiveness. Since WGS enables the extraction of the complete genetic information

of an organism, it can facilitate the identification and *in silico* prediction of genes that can be indicative of clinically important phenotypic traits, such as serotype, survival, increased virulence in human infections, and antimicrobial resistance, as well as identify the pathogen's genealogy (Deng, den Bakker, & Hendriksen, 2016).

WGS of foodborne pathogens has played a major role in identifying the key mechanisms behind pathogen virulence and survival, for improved understanding and, ultimately, control of pathogen in food (Gilmour et al., 2010). Moreover, WGS data can help identify factors that promote microbial proliferation in food and disease outbreak, such as the virulence or adaptability of specific subtypes to ecological niches in foods and their processing environments (Chen et al., 2006).

1.2.2. Applicability in quantitative microbial risk assessment

While its use in QMRA and predictive microbial modeling is not as prevalent as that in outbreak investigations, source attribution, and epidemiological investigations, WGS data has nevertheless shown great promise in this field. Recently, Njage and colleagues attempted to use next generation sequencing data to predict possible clinical outcomes resulting from exposure to shiga-toxigenic *Escherichia coli* and *Listeria* monocytogenes (Njage, Leekitcharoenphon, & Hald, 2019; Njage, Henri, Leekitcharoenphon, Mistou, & Hald, 2019). Similarly, molecular data has been used successfully to confirm source of pathogens associated with foodborne illness in pet food (Jones et al., 2019). Such models are believed to show great promise in revising existing

microbial dose-response models (Njage, Leekicharoenphon, & Hald, 2019; Njage, Henri, Leekitcharoenphon, Mistou, & Hald, 2019), as well as in improving the accuracy of models predicting pathogen growth and survival in the farm-to-fork environment (Collineau et al., 2019). Overall, the development and application of such modeling approaches has been predicted to improve the accuracy of current risk estimates for a number of foodborne pathogens, by reducing the uncertainty and variability that are inherent in such models.

In conclusion, the large molecular data set can offer the opportunity for increased insight and better decision-making than that which can be accomplished by analyzing small data sets. However, identification of the underlying trends, correlations, and relationships from such data, which would typically be absent in one-dimensional data alone, requires new and improved analytical and data management considerations (Strawn et al., 2015). This would include developing models that can effectively analyze and derive meaningful patterns from the large, noisy WGS datasets, and identifying data with the right kind of metadata to identify significant correlations between gene expression and a definite endpoint or impacting factor, as well as drawing reasonable conclusions about the same.

Chapter 2: Literature review, project rationale, and objectives

2.1. Whole genome sequencing

Whole genome sequencing is a method that can be used to provide a detailed characterization of an organism. WGS has increasingly become routine in the surveillance of bacterial foodborne pathogens, and epidemiological and outbreak investigations due to the advances in sequencing technologies (making them more mainstream; **Figure 2.1**) and their ability to completely characterize a pathogen down to its genomic level (Chen et al., 2016; Phillips et al., 2016).

2.1.1. WGS for foodborne outbreak investigation

WGS has become a viable resort for epidemiologic investigation into, and surveillance of, bacterial foodborne pathogens; thanks to the recent advances in sequencing technologies and bioinformatics tools. In fact, the term genomic epidemiology has been increasingly used to describe the practice of utilizing WGS to access, index, and analyze DNA sequence features of epidemiologic importance (Deng, den Bakker, & Hendriksen, 2016; Inns, et al., 2016; Chen, et al., 2016; Phillips, et al., 2016). Whole genome sequencing allows for the extraction of the complete genetic information of an organism; therefore, it facilitates the identification, as well as *in silico* prediction, of genetic determinants of clinically important phenotypic traits, such as serotype and antimicrobial resistance (Deng, den Bakker, & Hendriksen, 2016). Moreover, recent studies have shown that WGS of bacterial genomes can detect superspreaders, predict the existence of

undiagnosed cases and intermediates in transmission chains, suggest the likely directionality of transmission, and identify unrecognized risk factors for onward transmission (Nubel, Strommenger, Layer, & Witte, 2011; Snitkin et al., 2012; Octavia et al., 2015). In fact, WGS of foodborne pathogens has played a major role in identifying the major mechanisms behind pathogen virulence and survival, for improved understanding and, ultimately, control of pathogen proliferation in food. For example, a model was developed to describe the chromosomal evolution of strains involved in a nation-wide foodborne outbreak of *Listeria monocytogenes* in Canada, using the distribution and segregation of genetic traits such as SNPs, indels, and prophage, identified by WGS. This was the first instance wherein a next-generation sequencing (NGS) technology was used to perform a detailed genetic comparison of isolates displaying distinct pulsed field gel electrophoresis (PGFE) bands (Gilmour, et al., 2010). On the other hand, WGS data can help identify factors, such as the virulence or adaptability of specific subtypes to ecological niches in foods and food processing environments, which could promote microbial proliferation in food and infection outbreak (Chen, et al., 2006).

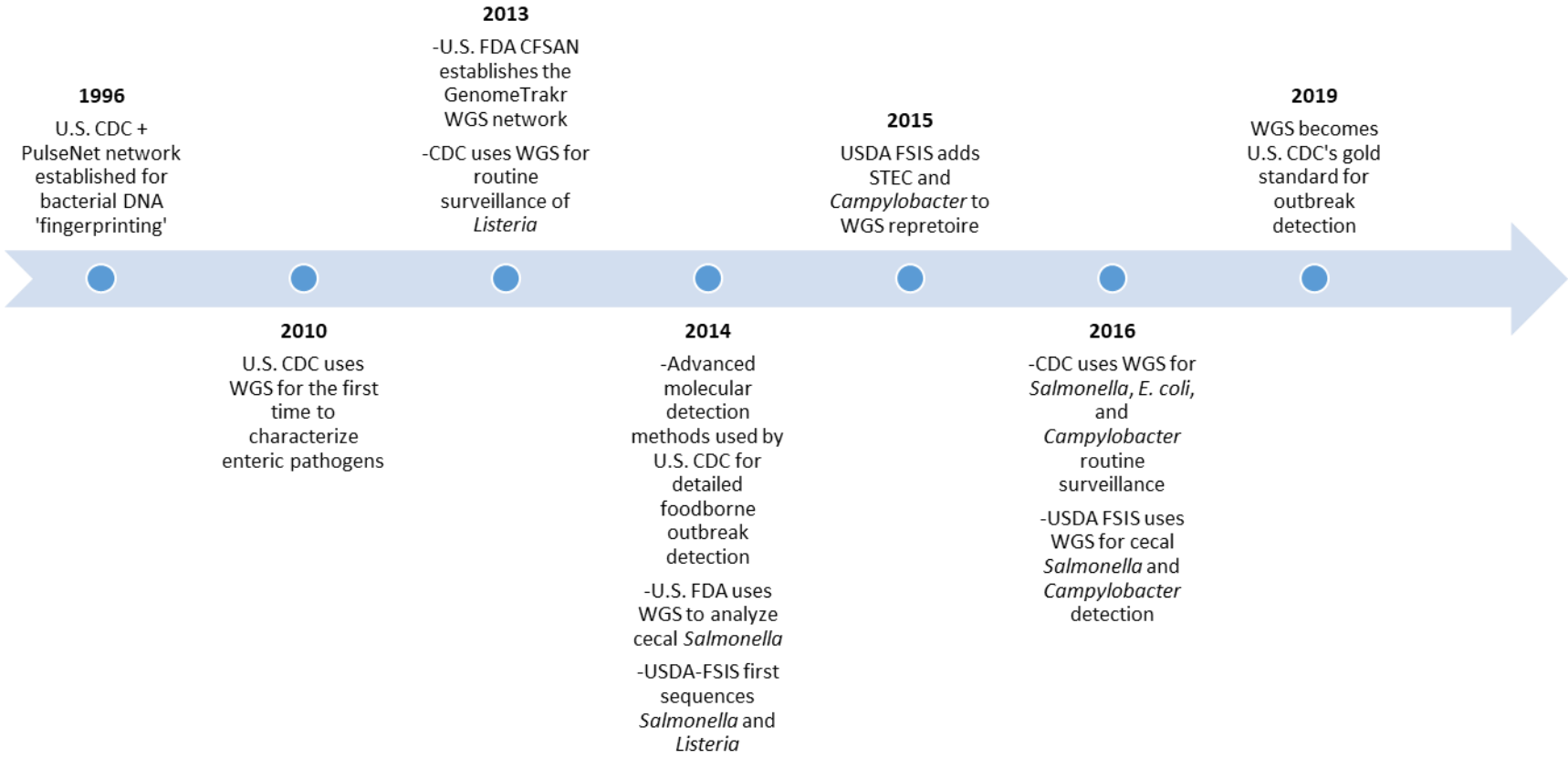

Figure 2.1. Timeline depicting usage and incorporation of WGS-based surveillance and investigations in U.S. regulatory agencies.

High-throughput sequencing technologies, or next generation sequencing (NGS) technologies, have been introduced and implemented in outbreak investigations only over the past decade, eclipsing the previously dominant automated Sanger technology-based methods (**Figure 2.1;** van Dijk, Auger, Jaszczyszyn, & Thermes, 2014). A major advantage of these technologies is that they produce massive amounts of data at greatly reduced per base-pair sequencing costs, allowing for the sequencing of complete microbial genomes at a price comparable to that of traditional subtyping methods such as PFGE or multi-locus sequence typing (MLST) (Deng, den Bakker, & Hendriksen, 2016).

2.1.2. WGS in predictive modeling and QMRA

2.1.2.1. Advantages and Limitations

Accurate risk estimation necessitates taking into account all sources of uncertainty and variability in the data used in the various steps of a risk assessment (Membre & Guillou, 2016). This could be accomplished using "omics" technologies, such as genomics, proteomics, metabolomics, and transcriptomics. Omics data could potentially be used to identify genes that encode proteins that could be involved in microbial preservation, stress response, survival, growth, and/or virulence. Therefore, these technologies have been predicted to play important roles in understanding the various functional properties of pathogens (such as survival under stress and interaction with potential hosts). Moreover, omics-based input may reduce the uncertainty involving species identity in risk assessments by providing full genome coverage, as well as new perspectives on strain diversity and physiological uncertainty. However, while sequencing data are becoming

increasingly comprehensive, they do not necessarily indicate function, nor do they always correlate to phenotypic response. Therefore, a key challenge will be to incorporate the data generated by these new technologies in the risk management decision making process (Brul, et al., 2012).

Sequencing of the entire genome helps in the identification of virulence and stress-related response. Genomic data helps identify the potential environmental stress response, survival, and virulence of microorganisms; however, these qualities may never be expressed in the microorganism, necessitating a thorough understanding of the true capabilities of microorganisms in different environments using transcriptomics (which can be used to quantify and confirm the differential expression of important genes). Transcriptional data can be analyzed by transforming the raw data into a gene expression matrix (data normalization to account for non-biological variation between samples) and subsequent data analysis (ANOVA, clustering, principle component analysis, multidimensional scaling and methods for class prediction) (Rantsiou, Mataragas, Jespersen, & Cocolin, 2011). Subsequently, bioinformatics and computational tools can be used to identify the functional elements in this data and predict the functions of genes in a genome. Furthermore, a complete metabolic profile for the microorganism can be compiled by identifying metabolic reactions that may be present due to their role in metabolic reaction cascades or pathways, which could function as indicators or interpretors of data, thereby validating the predicted phenotypic data (Alkema, Boekhorst, Wels, & van Hijum, 2016; Rantsiou, Mataragas, Jespersen, & Cocolin, 2011).

However, predicting the functions of all genes in a bacterial genome to identify its potential health and safety hazards is highly inefficient; a feasible alternative would be to selectively screen the microbial genome sequence for genes with specific functionalities, such as virulence, stress response, and persistence in unfavorable environments. The virulence potential of a bacterium can be investigated by comparing its genome sequence to data from previous epidemiological and outbreak investigations, or by comparing with a reference database containing known resistance genes and virulence factors, functionality of specific genes, and gene-function relations, such as the Salm-gene database for the serotyped isolates of *Salmonella* (Alkema, Boekhorst, Wels, & van Hijum, 2016; Deng, den Bakker, & Hendriksen, 2016). Similar approaches have been described for the identification of persistence of bacteria in food products (Vangay, Steingrimsson, Wiedmann, Stasiewicz, 2014), anaerobic spore-forming organisms in food (Doyle, et al., 2015), and potential pathogens in metagenomics data (Naccache, et al., 2014). This genomics-based method can also be applied to scenarios such as resistance to cleaning practices employed during food production (Bore & Langsrud, 2005; Fernandes, et al., 2015). However, a primary issue with employing such a manual selective screening approach is the potential loss of very informative genetic features that could significantly impact the predictive model. Therefore, there is a critical need for novel methods to analyze the available data and make impersonal decisions based on the features' contribution to the models and the model outcomes.

2.1.2.2. Potential areas for inclusion of WGS data in a modeling framework

While its use in QMRA and predictive microbial modeling is not as prevalent as that in outbreak investigations, source attribution, and epidemiological investigations, WGS data has nevertheless shown great promise in this field. Recently, Njage and colleagues attempted to use next generation sequencing data to predict possible clinical outcomes resulting from exposure to shiga-toxigenic *Escherichia coli* and *Listeria monocytogenes* (Njage, Leekitcharoenphon, & Hald, 2019; Njage, Henri, Leekitcharoenphon, Mistou, & Hald, 2019). Similarly, molecular data has been used successfully to confirm source of pathogens associated with foodborne illness in pet food (Jones et al., 2019). Such models are believed to show great promise in revising existing microbial dose-response models (Njage, Leekitcharoenphon, & Hald, 2019; Njage, Henri, Leekitcharoenphon, Mistou, & Hald, 2019), as well as in improving the accuracy of models predicting pathogen growth and survival in the farm-to-fork environment (Collineau et al., 2019). Overall, the development and application of such modeling approaches has been predicted to improve the accuracy of current risk estimates for a number of foodborne pathogens, by reducing the uncertainty and variability that are inherent in such models (**Figure 2.2**).

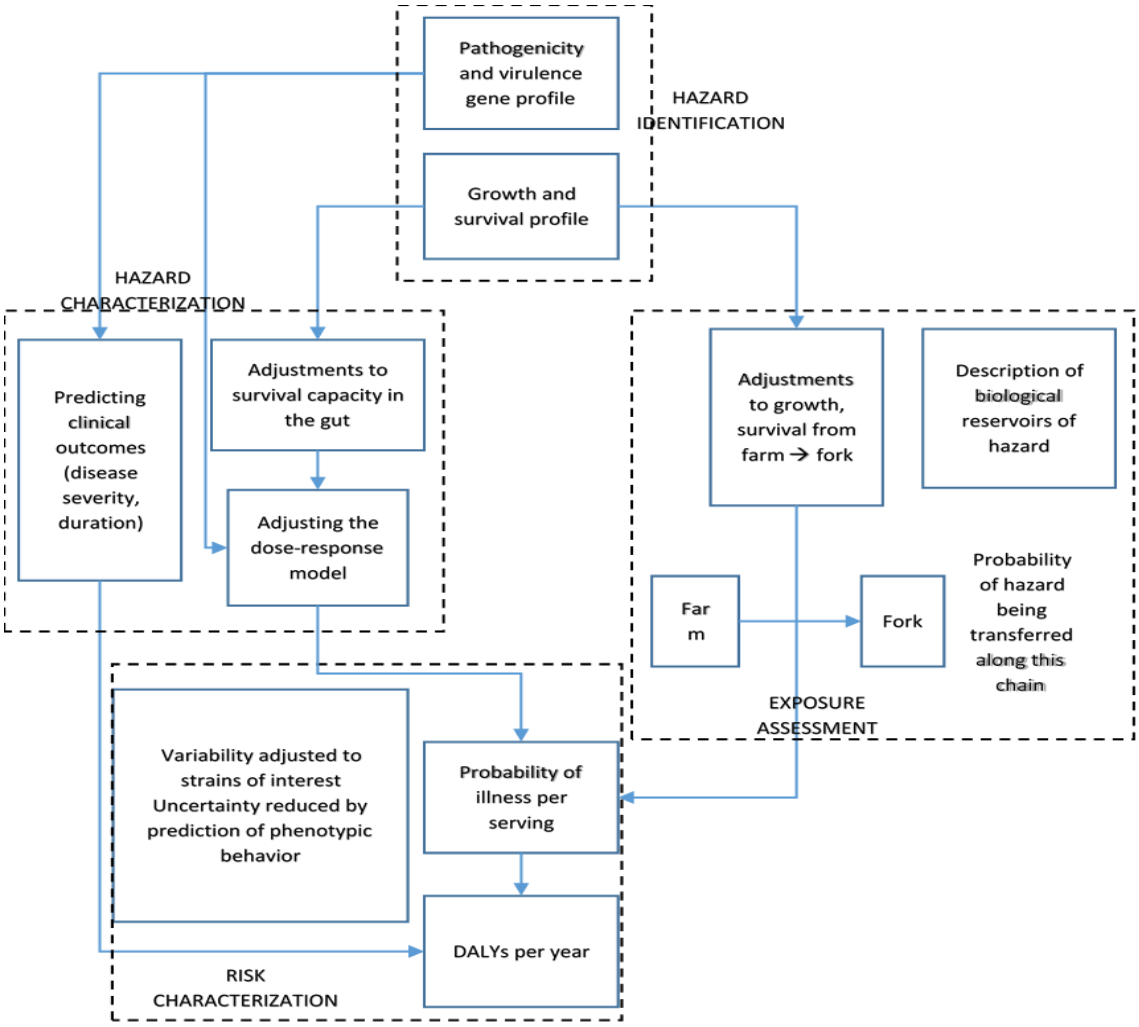

Figure 2.2. Potential areas for inclusion of WGS in a basic QMRA framework (adapted from Collineau et al., (2019)).

Recent reviews have postulated that the inclusion of WGS data would add novel dimensions to every aspect of a QMRA. For example, Collineau et al. (2019) described a basic framework for the areas in a risk assessment for antibiotic-resistant bacteria that could function as potential inclusion points for WGS data. Gaining knowledge about the differing

virulence, stress response, and antimicrobial genetic profiles of pathogens in the Hazard identification step of risk assessment would help in the development of differing pathogenicity profiles for the same. These profiles can, in turn, be used to adjust the survival capacity of the pathogen in the gut, as well as develop models to predict the adjusted clinical outcome or disease severity (hazard characterization/dose-response modeling). Simultaneously, these differing stress response profiles can be used to adjust the growth and survival capacity of the pathogen of interest in the farm-to-fork chain (farm, abattoir/processing, retail, food preparation, and consumption), which would significantly reduce uncertainty in the exposure assessment module of risk assessment. Finally, the revised models would lead to adjusting the variability in the model according to the strains of interest, reduce overall uncertainty by focusing on specific phenotypic behavior, and improve the overall accuracy of the model.

2.1.2.3. Early work in WGS-based predictive modeling

Human disease-associated sequence variation provides indirect information about the complex environmental stresses imposed on bacteria during the various steps in the food processing chain. This data can also be used to identify lineages that survive in food under stress, and subsequently infect humans after the consumption of contaminated food. Although genome-wide association studies (GWAS) have been increasingly used to identify genetic elements associated with particular phenotypes in humans, the strong population structure of bacteria resulting from clonal reproduction impedes the use of simple association mapping approaches in these microorganisms. Phenotypic differences

between complex isolates of bacteria could include differential metabolic abilities, and cell virulence, adhesiveness and invasiveness. Moreover, different complex isolates could be predominantly isolated from specific sources (while others could display a larger spread). An analysis of these observed divergences in disease-associated genetic variation between various clonal complexes should reflect different interactions with the selective conditions engineered during the food processing chain. Therefore, a thorough understanding of the functional traits associated with bacterial survival through processing plays a major role in developing models to estimate the risk of contamination and disease, as well targeted interventions to control said contamination.

In order to achieve this goal, Yahara et al. (2016) developed a GWAS approach to investigate genetic variations in *Campylobacter jejuni* isolates obtained from poultry processing and clinical infection sources, based on the capacity of bacteria to survive outside of the host through the poultry processing chain. Using a method that minimized the potential confounding effects of the strong population structure in *C. jejuni* by adjusting for the effect of relatedness between individual strains in the clonal genealogy compared to the null distribution (developed by Monte-Carlo simulation) of expected associations within each clonal complex, they identified genetic elements that were significantly over-represented among clinical *C. jejuni,* and subsequently mapped these elements to known virulence and candidate survival genes (Yahara, et al., 2016).

Alternatively, Franz et al. (2015) attempted to investigate the distribution of known virulence factors among clinical, food, and animal Shiga toxin-producing *Escherichia coli* (STEC) isolates, with the aim of identifying associations between virulence factors and phylogenetic groups, isolation sources, seropathotypes, serogroups, the presence or absence of adhesion factors such as intimin, a type of Shiga toxin, and the *rpoS* genotype in relation to the epidemiology of STEC in the Netherlands. Basically, they determined the virulence characteristics of putative pathogens from genomic information, using a method is referred to as 'predictive hazard identification.' Franz et al. (2015) performed a Chi-squared test to analyze the differences in frequencies of genetic markers and the associations between these markers. Among their most significant results, they observed that isolates expressing certain Shiga-toxin genes (*stx2a*, *stx2c*, and *stx2f*) showed higher numbers of other virulence genes, as well as a strong correlation between the expression of the adhesion gene *eae* and virulence characteristics of the bacteria.

This was expanded upon by Pielaat et al. (2015), who introduced a conceptual method for hazard identification linking genotypic information (whole-genome sequencing data) with epidemiological (subset of STEC O157:H7 isolates) and phenotypic (*in vitro* adherence to epithelial cells as a proxy for virulence) data. Assuming a homogeneous distribution of cells in the *in vitro* culture, the fractional adherence of the genotyped isolates to Caco-2 cells was calculated by dividing the number of STECs after the adhesion assay by the number of STECs added to the cells. Subsequently, isolates that were associated with an increased virulence behavior were identified by a simple linear regression model:

$$y_i = \mu + \beta x_i + \varepsilon_i$$

With y_i, μ, β, and ε indicating the fractional Caco-2 adhesion for strain i, mean response, SNP effect, and the residual error (with an independent normal distribution), respectively, and x_i indicating 0 or 1, depending on the concordance of the marker score with the reference strain. Possible errors arising from population variation, such as the effect of between-group variations (assuming a negligible within-group variation) were corrected by introducing an additional term "Gi" in the regression model (to correct for group effect):

$$y_i = \mu + \beta x_i + G_i + \varepsilon_i$$

Moreover, the authors suggested the use of the Bonferroni correction method to account for additional error terms arising from the large number of identified SNPs (compared to the number of isolates), identification of strains with an acceptable minor allele frequency (MAF) compared to the reference strain. They reasoned that despite being an informative *ex post facto* determinant of virulence potential, the dynamic nature of STEC virulence in the real world reduces the robustness of the seropathotype concept of classifying STEC serogroups into different risk classes based on the severity of disease and involvement in outbreaks as a predictive indicator of microbial risk. This issue was transcended by their approach, which offered a standardized, reproducible, serogroup-independent method for identifying potential candidate genes to be included in refined hazard identification, as SNP analysis of a broad spectrum of isolates may lead to a less biased association between genotypic and phenotypic strain characteristics. However, the authors also cautioned against blanket acceptance of their results, as (i) SNP analysis based

on comparing test strains with one reference strain would not be sufficient in identifying relevant SNPs for hazard identification, and (ii) the presence of a biomarker alone may not be the best predictor of risk, without controlling or accounting for the dependent biological factors. In conclusion, they suggested the use of a pan-STEC O157:H7 genome for a more comprehensive assessment (Pielaat et al., 2015).

2.2. A novel development – integration of machine learning strategies into a food safety modeling framework

Recent efforts toward the incorporation of WGS data into predictive modeling and risk assessment have focused on the development of advanced modeling strategies and data analytical methods to compress WGS data into a format applicable in a mathematical and statistical framework of risk. This has led to increased exploration into machine learning and deep learning methods to analyze WGS data in order to identify patterns indicative of specific pathogen behavior and derive meaningful outcomes that are relevant to a risk assessment. Simply put, machine learning is a subfield of artificial intelligence wherein a computer is trained to identify patterns based on example data or past experience to solve a given problem. Machine learning is principally different from traditional algorithmic problem-solving, as in the former, explicit instructions are not programmed, Instead, learning occurs one of two ways – (i.) via instance-based learning, i.e., from examples and generalizing to new cases based on their closeness to learned examples, or (ii) by model-based learning, i.e., training a model with data to learn parameters through optimization, and making predictions using new test data (Deng, Cao, & Horn, 2021).

Machine learning models require a certain amount of training: based on the amount of supervision provided, the primary learning models can be supervised, semi-supervised, or unsupervised. Supervised learning involves feeding the learning system with training data that are labeled with the ground truth or outcome. Examples of such models include random forest, support vector machine, Naïve Bayes, boosting (classification models) and logistic regression (regression). Unsupervised learning algorithms such as principal component analysis (PCA) involves inputting the learning system with unlabeled training data, allowing the algorithm to identify hidden patterns on its own. Semi-supervised learning occurs when few labeled or tagged instances are available among many unlabeled ones in very large datasets. Under such constraints, a semi-supervised learning algorithm weighs in on the contribution of unlabeled data on the relationship between the predictor and the outcome (Deng, Cao & Horn, 2021).

Machine learning, specifically supervised learning, has seen increasing usage in the biological and medical domain to effectively analyze the usually noisy, non-linear, high-dimensional datasets generated (Kampichler, Wieland, Calme, Weissenberger, Arriaga-Weiss, 2010). Examples include the use of machine learning to identify disease severity in patients with heart disease (Tripoliti, Papadopoulos, Karanasiou, Naka, & Fotiadis, 2017), Crohn's disease (Schuffler et al., 2013), and Parkinson's disease (Tsanas, 2012; Armañanzas, Bielza, Chaudhuri, Martinez-Martin, & Larranaga, 2013) in the biomedical field, plant disease severity indicators and genetic responses to stress in the plant sciences

domain (Mwebaze & Owomugisha, 2016), and identify ecological factors contributing to decreasing bird populations in the ecological domain (Kampichler, Wieland, Calme, Weissenberger, Arriaga-Weiss, 2010), among many others. However, their widespread usage in microbial predictive modeling especially in the food safety domain, have been hindered by the lack of available metadata and endpoints, differences in the type of data being collected and collated between different agencies (lack of consistency), and a lack of transparency in case of many of the datasets that are available.

However, despite this, strides have been made towards developing and employing various such machine learning-based modeling strategies in predictive modeling, and therefore, risk assessment. So far, researchers have made great strides in incorporating machine learning models, primarily binary classification models, into the prediction of antimicrobial susceptibility and source attribution (**Table 2.1**). However, very few studies have analyzed such large datasets with the end goal of incorporating genomic data into predictive modeling and risk assessment. Recently, Farrell, Soyer, & Quince (2018) employed a machine learning approach and a lasso logistic regression statistical model to resolve 65 functional and metabolic capacities (i.e., phenotypic traits) of 9,407 prokaryotic full-draft genomes. Wheeler, Gardner, & Barquist (2018) used a random forest approach to predict invasiveness in *Salmonella*, as well as identify a common theme of degradation of metabolic pathways in extraintestinal lineages. More recently, however, Njage, Leekitcharoenphon, & Hald (2019) & Njage, Henri, Leekitcharoenphon, Mistou, & Hald (2019) have analyzed a number of machine learning algorithms, including random forests,

support vector machine, logistic regression, and boosting to identify genetic patterns indicative of increased severity of clinical outcomes in humans infected with *Escherichia coli* and *Listeria monocytogenes.* These papers were of special relevance to food safety and QMRA, as they proposed methods to incorporate the results from such models in a risk assessment framework. A point to be noted is that, in the context of food safety, specifically in food safety risk assessment and predictive microbiology, researchers have mostly clustered around the concept of supervised machine learning. This could be due to the relative newness of the 'big data' explosion in this field (resulting in researchers not having a clear, standardized idea of the type of data to collect to assist in machine learning-based modeling). Moreover, compressing WGS data into its most important and relevant genetic features considering the outcome (stress survival, virulence in the host, disease severity, etc.) that could be informative to a risk assessment, is a task as yet unfulfilled by the research body.

Table 2.1. Some selected works in the food safety domain demonstrating the use of machine learning in analyzing and predicting patterns from WGS data. Most included works apply the supervised learning structure.

Overall goal of the study	Microorganism	Reference	Method used	Supervised (S), unsupervised (US) or semi-supervised (SS)	Basic study design (type of prediction)	Features used as predictors
Predicting antimicrobial resistance	*Mycobacterium tuberculosis*	Niehaus et al. (2014)	LR, SVM	S	Classification of AMR	SNPs
	Enteric bacteria (*E. coli*, *Enterobacter*, etc.)	Pesesky et al. (2016)	LR	S	Classifying based on antimicrobial susceptibility	AMR genes
	Klebsiella pneumoniae	Nguyen et al. (2018)	Boosting, bagging, RF, SVM, extremely random tress	S	Determining the MIC	*k*-mer
	S. enterica	Maguire et al. (2019)	LR, SCM	S	Classifying based on antimicrobial susceptibility	AMR genes, *k*-mer
	S. enterica	Nguyen et al. (2019)	Boosting	S	MIC determination	*k*-mer

Predicting host specificity	*E. coli* O157:H7	Lupolova et al. (2016)	SVM	S	Classifying into correct isolation hosts	pan-genome content
	S. enterica multiple serovars and lineages	Wheeler et al. (2018)	RF	S	Identifying invasive, host-adapted species of *Salmonella*	Pan-genome content
Source attribution	*S. enterica* serovar Typhimurium	Munck et al. (2020)	Logistic boost	S	Zoonotic attribution	Core genome MLST

LR – logistic regression; SVM – support vector machine; AMR – antimicrobial resistance; SNP – single nucleotide polymorphism; RF – random forest; MIC – minimum inhibitory concentration; SCM – set covering machine; MLST – multi-locus sequence typing.

2.3. *Salmonella* – a prime etiological agent for WGS-based predictive modeling

Salmonella enterica subsp. *enterica* is a major foodborne pathogen responsible for an estimated 1.2 million cases of foodborne illnesses per year. Despite the implementation of several preventative and control measures against *Salmonella* over the past several years, this has failed to make a significant impact on its worldwide prevalence rates. *Salmonella* is a major food-borne pathogen, with high morbidity and mortality rates and demonstrated major economic loss worldwide. This Gram negative, facultative anaerobic bacterial species consists of over 2,500 named serovars with a highly variable pathogenicity profile (CDC, 2021b). Studies have shown that, although different serovars do not actually imply pathogenicity, a limited number have been associated with a majority of the cases of human infections (U.S. CDC, 2019a, b). However, genetic evolution and horizontal gene transfer between traditionally virulent and non-virulent serovars of *Salmonella* has resulted in a significant upsurge in the incidences of foodborne salmonellosis over the past several years, with several serovars being associated with human cases of infection from consuming animal-based food sources alone (Ferrari et al., 2019; **Figure 2.3**). For example, earlier cases of chicken-borne salmonellosis were primarily attributed to the serovars Typhimuirum and Enteritidis, while recent years have seen outbreaks of salmonellosis attributed to previously unknown serovars including Schwarzengrund, Infantis, Agona, Anatum and Oranienburg (U.S. CDC, 2019a, b).


Serovar
Pork
Poultry
Beef
Seafood
Aberdeen
Agona
Albany
Amsterdam
Anatum
Arizona
Bareilly
Blockley
Bovismorbificans
Braenderup
Brancaster
Bredeney
Corvallis
Derby
Drac
Dublin
Eastbourne
Goldcoast
London
Give
Hadar
Haifa
Havana
Indiana
Kentucky
Kiambu
Kottbus
Lexington
Livingstone
London
Madelia
Mbandaka
Meleagridis
Monophasic
Montevideo
Muenchen
Muenster
Newport
Orion
Panama
Enteritidis
Infantis
Brandenburg
Heidelberg
Johannesburg
Reading
Rissen
Saintpaul
Schwarzengrund
Senftenberg
Ser 16:1
Singapore
Sinstorf
Sofia
Stanley
Paratyphi B
Ohio
Orion
Typhi
Tallahassee
Tennessee
Thompson
Typhimurium
Virchow
Weltevreden
Worthington
Zanzibar
Key
-0.5
0.5
1.5
2.5
3.5
>4.5

Figure 2.3. *Salmonella enterica* serovar prevalence in food animal matrices in North America (adapted from: Ferrari et al., 2019).

While all serovars belonging to *Salmonella enterica* share a common genome structure, and the same core genome (Anjum et al., 2005; Jacobsen, Hendriksen, Aaresturp, Ussery, & Friis, 2011; Hoffman et al., 2014), there is enormous variation in their pathogenicity (**Table 2.1**), host range (**Figure 2.3**), and epidemiology (Cheng, Eade, & Wiedmann, 2019). In terms of pathogenicity, some non-typhoidal *Salmonella* serovars have been shown to be particularly proficient in causing invasive (isolated from blood, joint fluid, etc. (Jones et al. (2008)) infections in hosts (particularly humans), similar to that shown by serovars Typhi and Paratyphi. Such serovars are of particular importance, as they are the most severe, showing a capacity to transcend the gastrointestinal tract, causing severe infection, and ultimately, hospitalization. However, it is also difficult to definitively characterize serovars as a whole as being invasive or non-invasive, as most invasive cases are associated with individuals from high-risk populations, such as very young children, the elderly, the immunodeficient (including those with human immunodeficiency virus – acquired immunodeficiency syndrome (HIV-AIDS)), and pregnant women (Scott et al., 2011; Feasey, Dougan, Kingsley, Heyderman & Gordon, 2012; Okoro et al., 2012; Ao et al., 2015; Lan et al., 2016).

In terms of epidemiology, only specific serovars (*S.* Typhi, *S.* Paratyphi A and C, and *S.* Sendai) have been shown to cause enteric fever, with a *majority* of the other serovars

causing only gastroenteritis. However, a few non-typhoidal serovars such as Choleraesuis, Dublin, Panama, and Sandiego are more likely to cause the more serious bacteremia and other forms of invasive disease, than simply diarrhea (Fierer & Guiney, 2001; Jones et al., 2008; Marzel et al., 2016). However, despite being associated with a higher incidence of invasive disease, these serovars contribute lower towards the total number of human salmonellosis cases, compared to other, less invasive serovars, such as serovar Typhimurium. Such discrepancies in invasiveness compared to overall impact on the human health index has been postulated to be due to one or more of several factors – higher level of exposure of susceptible populations to these more invasive serovars, underreporting of less severe cases, and even genetic adaptations making the more invasive serovars inherently more 'dangerous' (Jones et al., 2008). With respect to host range, some serovars are host adaptive (such as Enteritidis or Typhimurium) while others are more host-specific (Graziani et al., 2011; Capuano et al., 2013), although some serovars previously thought to be host-specific are also being implicated in human cases of infection (such as serovar Derby).

This knowledge, combined with recent findings that multiple individual virulence genes are variably distributed across the different serovars of *Salmonella*, allows us to conclude that a complicated combination of genes contribute to the overall virulence diversity (Suez et al., 2014), which presents a big challenge for virulence profiling and, by extension, for predictive modeling and risk assessment. So far, risk assessments of *Salmonella* in different food sources have focused on overall species-level count data, as

opposed to serovar- or gene-level classification, primarily due to the complexity it would introduce to the models. However, this known difference in *Salmonella* virulence, pathogenicity in the host, survival ability under various environmental and processing conditions, and antimicrobial resistance introduces significant uncertainty and variability into any predictive models for *Salmonella*. Therefore, this agent is a prime example of an etiological agent to be subjected to WGS-based distinction in terms of risk profile development.

2.4. Potential for development of advanced machine learning and data analytical models to assist in the development of WGS-based risk assessment for *Salmonella*

Although a huge volume of data in being produced in nearly all sectors of society and economy worldwide, the term ‘big data’ is rarely applied in the context of food and food safety. As a result, the utilization of advanced methods to analyze such data in this particular domain remains untapped. This represents an untapped resource for the use and application of the large amounts of data being generated by the combined agricultural, health, and environmental domains that comprise and impact the food sector. Here, we discuss the availability of such data, modeling strategies to analyze large volumes of data, and the issues with applying such methods currently being faced by researchers in the food safety and predictive microbiology domain.

2.4.1. Availability of WGS data

Several public repositories of sequencing data are currently available, with published sequences being made available for comparative genomic analysis, epidemiological investigations, and source attribution, and provides us with an unprecedented opportunity for the development of WGS-based risk assessments. Several databases and initiatives for *Salmonella* Genomic Analyses, including the National Center for Biotechnology Information's (NCBI) GenBank database, the U.S. FDA GenomeTrakr network, the NCBI Pathogen Detection database, the Collaborative Management Platform for Detection and Analyses of (Re)-Emerging and Foodborne Outbreaks in Europe (COMPARE), the Global Microbial Identifier (GMI), and the National Antimicrobial Resistance Monitoring System (NARMS), among others. The National Center for Biotechnology Information's (NCBI) GenBank database (https://www.ncbi.nlm.nih.gov/genome/genomes/152) currently lists over 10,935 complete (and annotated) genome assemblies of *Salmonella*. A majority of the labs, programs, and initiatives that are responsible for the generation of WGS data (and its related metadata) exchange data with the NCBI; these include the major worldwide WGS data repositories, the United States Food and Drug Administration's (U.S. FDA) Center for Food Safety and Applied Nutrition (CFSAN), the United States Department of Agriculture (USDA) Food Safety Inspection Service (FSIS), the U.S. Centers for Disease Control and Prevention (U.S. CDC), the CDC's National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS), the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ) (Chen et al., 2020).

The GenomeTrakr network, established by the U.S. FDA (https://www.fda.gov/food/whole-genome-sequencing-wgs-program/genometrakr-network) is the first open-source network of its kind that utilizes WGS for pathogen identification. As of October 2021, GenomeTrakr participants includes 15 federal laboratories, 36 state public health and academic laboratories, 1 U.S. hospital lab, 2 other labs located in the U.S., and 21 laboratories outside the U.S., and several other laboratories, which are authorized to collect WGS data and metadata of foodborne pathogens, including their food sources, geographic origins, and diseases manifestations, subsequently sharing them via publicly accessible databases at the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/bioproject/?term=genometrakr). This information is publicly accessible, allowing for real-time comparison and analysis of foodborne microorganisms, thereby aiding in speedy investigations of foodborne outbreaks. The bacterial sequences collected from food, the environment, and human patients collected by these international agencies during real-time, active surveillance of pathogens and foodborne disease is then uploaded to the centralized NCBI Pathogen Detection database (https://www.ncbi.nlm.nih.gov/pathogens/). This allows for easier data exchange, identification of potential sources of foodborne contamination, and rapid reporting of potential relationships between the type and source of food and human illness during traceback investigations and outbreak response (Chen et al., 2020).

The Collaborative Management Platform for Detection and Analyses of (Re-) Emerging and Foodborne Outbreaks in Europe (COMPARE; http://www.compare-europe.eu/) is a large EU project comprising a multidisciplinary research network of 29 European participants from 10 EU countries and Australia, which is funded by the European Union's Horizon 2020 research and innovation program, with the aim of speeding up the detection of, and response to, disease outbreaks among humans and animals worldwide through the use of new genome technology. Additionally, COMPARE aims to develop risk assessment models, risk-based strategies for sampling and data collection, and harmonized standards for sample processing and sequencing. The Global Microbial Identifier (GMI) (http://www.globalmicrobialidentifier.org/About-GMI) is a network of approximately 160 representatives from 32 countries working to develop a global system to aggregate, share, mine, and use microbiological genomic data to address global public health and clinical challenges. The primary goal of GMI is to employ a system that promotes equity in access and the use of current genomic technology worldwide. The GMI database gathers microbial (bacterial, fungal, viral, and parasitic) genomic information, as well as related metadata such as epidemiological information, to address the clinical challenges associated with these microbes. The GMI is an ongoing initiative, a global solution for the analysis of WGS data (including the creation of networks and regional hubs). Online bioinformatics platforms for genomic analysis include the Pathosystems Resource Integration Center (PATRIC; https://www.patricbrc.org/portal/portal/patric/Home), a genomics-centric relational database and source for the analysis of bacterial gene expression data, which can be quite

helpful for data integration. Similarly, the Center for Genomic Epidemiology (http://www.genomicepidemiology.org/) is another ongoing collaborative endeavor initiated by the National Food Institute in Denmark that aims to provide the scientific foundation for a future central database of genome data, a platform of spatio-temporal tools for the analyses of such data along with epidemiological information, as well as a web-based interface capable of facilitating the exchange of required microbial data (Chen et al., 2020).

2.4.2. Review of available modeling strategies

Currently, predictive models for microbial agents in food are restricted to the application of classification strategies to group specific isolates based on a defined endpoint, such as disease severity. Briefly put, a classification algorithm, is a function that weighs the input features so that the output separates one class into positive values and the other into negative values. Training a classification algorithm, also known as classifier training, is performed to identify the weights (and functions) that provide the most accurate and best separation of the two classes of data (Netoff, 2019). While a linear discrimination analysis is the simplest form of such an algorithm, a majority of available microbial datasets do not have a clear relative separation between the two classes. Under such conditions of complexity, novel classification methods such as random forests, support vector machine, and boosting have been developed. Random forest is an ensemble method wherein several decision trees are trained in parallel with bootstrapping, followed by aggregating the decisions of individual trees for a final decision (Misra & Wu, 2020).

Support vector machine, on the other hand, is a kernel-based method that attempts to find the optimal separating surface by projecting nonlinear separable samples onto another higher dimensional space via the use of different types of kernel functions (Pisner & Schnver, 2020). Alternatively, boosting refers to a family of classifiers that can be used to convert weak learners to strong ones. On the other hand, the application of *k*-means clustering and principal component analysis (PCA), two major unsupervised machine learning methods, in delineating patterns from datasets without any fixed ground truth (outcome variable, or unlabeled datasets), have been primarily restricted to the analysis of (i.) images obtained from various imaging technologies (viz. hyperspectral and multispectral imaging) that are currently being used or proposed for food quality and safety analysis (Qin, Burks, Kim, Chao, & Ritenour, 2008; Powell, Jacob, & Chapman, 2011; Qin, Chao, Kim, Lu & Burks,), and (ii.) irradiation damage detection for food quality and safety testing purposes (Guillén-Casla, Rosales-Conrado, León-González, Pérez-Arribas & Polo-Díez, 2011; Yang et al., 2021).

Overall, a majority of these classifiers have been found to be significantly more powerful than simple logistic regression in effectively separating classes, identifying non-linear trends, reducing the bias and variance, and reducing model overfitting. A popularly employed method to identify the most appropriate classifier for the given set of data is by training and using multiple classifiers, and subsequently making a classification decision based on the results of all the classifiers. Reporting on the success or failure of classifiers in defining the classes involves the identification of accuracy simple metrics. These include

the sensitivity (true positive rate) and specificity (true negative rate). Another metric that is commonly used as a critical measure of the classifier performance is to determine the area under the curve of a receiver operating curve (AUC-ROC), which is obtained by plotting the sensitivity against the specificity. Specifically, the latter measure is considered as a non-parametric measure of classifier performance and is very useful for comparing classifiers (Netoff, 2019). However, each of these classifiers have a number of pros and cons – while a majority of the classifiers are particularly good at gleaning patterns from noisy datasets, the results are hard to interpret, and eventually, incorporate in a predictive modeling framework. Thus, it is of utmost importance to identify a learning-based method that will accurately identify significant genes from a dataset with many more predictor variables than number of samples, while providing a simple means to interpret the results.

2.5. Project overview and objectives

Whole genome sequencing and other such molecular data is fast becoming standard in microbial epidemiology, source attribution, and pathogen tracking. However, due to the large scale of WGS datasets, its application in microbial modeling and QMRA to calculate the public health burden of pathogenic organisms remains in its early stages. In this dissertation, we aim to develop advanced statistical learning methods and data analytical techniques to overcome the challenges of incorporating large molecular datasets in different aspects of a QMRA framework. In this dissertation, we utilize a multi-pronged approach to (i.) identify areas in microbial predictive modeling and a QMRA framework that could be improved upon using WGS data, and (ii.) develop machine learning models

to analyze WGS for important expression trends that could be indicative of important bacterial phenotypes and outcomes in host systems. The following objectives have been set for this dissertation, and their relationship to each other and a QMRA framework have been highlighted in **Figure 2.3**.

1) **Development of an advanced machine learning-based workflow to identify and predict severe disease phenotype in *Salmonella enterica*.** Microbial WGS data introduces extremely large dimensions (small number of samples compared to predictor variables) that regular algorithmic statistical models cannot handle without model overfitting and introducing dimensionality issues. Currently, there is no published information on how to combat dimensionality issues in multi-collinear datasets such as WGS datasets. Moreover, a majority of available genomic data is not labeled, precluding their use in predictive models. This necessitates the development of a workflow to label, and subsequently analyze large WGS datasets for microbial modeling.

2) **Development of a weighted modeling approach to incorporate genetic heterogeneity in a dose-response modeling framework.** Current dose-response models for *Salmonella enterica* are generalized to the species level, employing generalized data from a select few serovars. Genomic data can help delineate patterns within bacteria that could be indicative of increased infectivity (and thereby, increased disease incidence). There is a need to develop an advanced

genomics-based workflow to adequately capture this genetic heterogeneity and its correlation to human disease incidence.

3) **Predicting foodborne salmonellosis outbreak severity based on genetic and meteorological trends.** Meteorological factors and innate genetic changes have been independently shown to impact salmonellosis occurrence and severity (in terms of case numbers). However, there is no published research on the combined impact of the two on outbreak severity. There is a need to quantify the interaction effects between meteorological factors (such as temperature and precipitation) and the probability of expression of various significant genes in *Salmonella enterica,* would help predict the most significant combination of genes and meteorological factors that contribute to the incidence and outbreak of food-associated salmonellosis.

It is evident that whole genome sequencing information could reduce the uncertainty due to lack of information about individual serovar behavior in a predictive modeling and QMRA framework. However, it introduces new dimensions that regular statistical models are unable to handle effectively. Machine learning can help analyze such high dimensional datasets, identifying important predictor variables in the process. However, in addition to identifying the exact areas within a QMRA framework that such information can be introduced in, the utilization of such models introduce their own set of issues and challenges. Identifying effective means to handle such data would greatly assist in futuristic, WGS-assisted risk assessments, and help risk managers take necessary action

to control and reduce the public burden of *Salmonella enterica* (and other such pathogens with multiple subspecies and serovars) in the future.

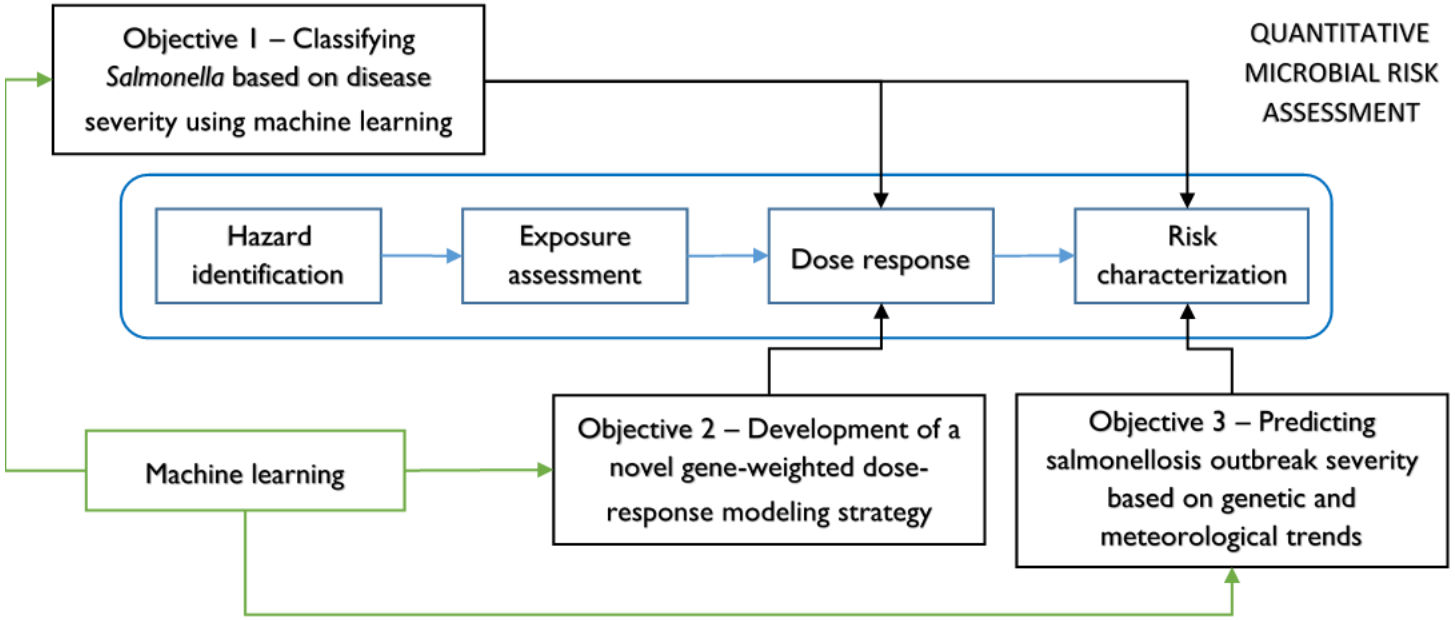

Figure 2.4. Diagram demonstrating relationships among the objectives and potential paths of integration of WGS data into a QMRA framework.

Chapter 3: Development of an advanced machine learning-based workflow to identify and predict severe disease phenotype in *Salmonella enterica*

3.1. Abstract

The increase in availability of WGS information seen in the past decades has allowed for its incorporation in predictive modeling for foodborne pathogens to account for inter- and intra-species differences in their virulence. However, this is hindered by the inability of traditional statistical methods to analyze such large amounts of data compared to the number of observations/isolates. In this study, we have explored the applicability of machine learning models to predict the disease outcome, while identifying features that exert a significant effect on the prediction. This study was conducted on *Salmonella enterica*, a major foodborne pathogen with considerable inter- and intra-serovar variation. WGS of isolates obtained from various sources (human, chicken, and swine) were used as input in four machine learning models (logistic regression with ridge, random forest, support vector machine, and AdaBoost) to classify isolates based on disease severity (extraintestinal vs. gastrointestinal) in the host. The predictive performances of all models were tested with and without Elastic Net regularization to combat dimensionality issues. Elastic Net-regularized logistic regression model showed the best area under the receiver operating characteristic curve (AUC-ROC; 0.86) and outcome prediction accuracy (0.76). Additionally, genes coding for transcriptional regulation, acidic, oxidative, and anaerobic

stress response, and antibiotic resistance were found to be significant predictors of disease severity. The genes, which predicted each outcome, could possibly be input in amended, gene-expression-specific predictive models to estimate virulence pattern-specific effect of *Salmonella* and other foodborne pathogens on human health.

3.2. Introduction

Recent advances in scientific testing methodologies has resulted in the widespread application of whole genome sequencing (WGS) for epidemiological investigations and surveillance of bacterial foodborne pathogens. Several scientific reviews and studies have concluded that the larger molecular data set afforded by WGS analysis can offer researchers with the opportunity for increased insight and better decision-making than that which can be accomplished by analyzing smaller data sets. However, integrating WGS information in a predictive microbial modeling framework has proven to be a challenging prospect due to data disaggregation caused by splitting a generalized species into its genetic content (i.e., exponential increase in the number of predictors to be considered) (Brul et al., 2012; Mughini-Gras et al., 2014; Pielaat et al., 2015; Strawn et al., 2015; Membre & Guillou, 2016).

The different serovars of the major foodborne pathogen *Salmonella enterica* subsp. *enterica* show enormous variation in pathogenicity profile, virulence, host range, disease severity, and epidemiology (Abbott, Ni, & Janda, 2012). Certain serovars of *Salmonella* have demonstrated a higher propensity for causing extraintestinal disease, including

bacteremia, systemic disease, sepsis, and infection in extraintestinal sites such as the brain, lymph, and lungs, compared to others, indicating a significant variability in cause-specific mortality (Parkhill et al., 2001; Austin, Tu, Ho, Levy, & Lee, 2013; Nuccio & Baumler, 2014; Wheeler, Gardner, & Barquist, 2018). This highlights the clear need to take into account the distinction in virulence-associated characteristics of each strain when calculating the risk of disease and disease management. This would, in turn, demonstrate the applicability of including whole genome sequencing and other genetic characterizations into predictive microbial modeling. However, the sheer number of predictor variables that this would introduce makes the modeling of genomic data a challenging prospect, while methods that can be employed to reduce the number of predictor variables may result in the loss of important predictor variables (Houle, Govindaraju, & Omholt, 2010).

Machine learning is a field of study wherein advanced computational systems are "trained" to make predictions or decisions based on inference and patterns alone, thereby simplifying complex statistical models and algorithms. A specific subset of machine learning algorithms, known as classification algorithms, have seen increasing use in the life sciences domain over the past few decades due to their particular efficacy in identifying hard-to-discern patterns from large, noisy, and complex data sets (Friedman, 1998; Bishop, 2006; Austin, Tu, Ho, Levy, & Lee, 2013). In the food safety domain, Pielaat et al. (2015) and Njage, Leekitcharoenphon, & Hald (2019), and Njage, Henri, Leekitcharoenphon, Mistou, & Hald (2019) proposed the use of ensemble classification algorithms to predict bacterial host disease characteristics using WGS data with an aim to ultimately incorporate

WGS into microbial risk assessment. However, to our knowledge, this has not been attempted in *Salmonella* so far. Additionally, making sense of genotypic data using machine learning is difficult due to the lack of availability of associated useful biological information (Austin, Tu, Ho, Levy, & Lee, 2013; Njage, Leekitcharoenphon, & Hald, 2019). This is because, in the context of exploratory sampling or epidemiological analyses, the emphasis has not so far been on associated genetic signatures to specific endpoints, including bacterial growth, survival, and host reaction. This is particularly true for bacteria such as *Salmonella* with considerable intra-species variation in virulence, survival, and host characteristics. Therefore, a major step towards the application of WGS to a modeling framework is the identification of appropriate associated endpoints.

In our study, we have attempted to obtain an association between WGS data and severity of *Salmonella*-related health outcomes in the host, indicative of virulence capacities of different strains of *Salmonella,* based on their genetic makeup. The objectives of this study were (i) to compare the accuracy of different statistical learning/machine learning algorithms in predicting gastrointestinal (disease severity = low) or extraintestinal (disease severity = high) outcomes in *Salmonella* isolates from human, poultry, and swine, based on genome sequencing data, (ii) to determine the applicability of a powerful recursive feature selection tool, which is useful in reducing data dimensionality issues, in increasing model accuracy, and (iii) to identify genetic signatures significantly associated with each outcome to possibly be input in amended microbial risk models.

3.3. Material and Methods

3.3.1. Bacterial genomes – Sample selection

Isolates of *Salmonella enterica* subsp. *enterica* were obtained from literature. Studies were selected using stringent inclusion criteria – availability of sequence accession numbers, availability of data regarding the source of isolation (host system, place of isolation within the host), and availability of disease severity information (or, alternately, information to deduce the same). Based on this, two studies were identified (Pornsukarom, van Vliet, & Thakur (2018); Rakov, Mastriani, Liu, & Schifferli (2019)) and WGS accession numbers were obtained. These studies were unique in that the studies specifically distinguished the site of isolation of the *Salmonella* isolates, the clinical endpoints in the host, and/or specified the invasive or non-invasive nature of the isolates in the host, allowing the use of this information to derive our outcome variables. The isolates included in our study were curated from among human cases of *Salmonella* infection (n = 73), and animal hosts (such as swine (n = 25) and poultry (n = 52)). Metadata associated with the included samples, including source of isolation, invasive or non-invasive subtypes, and availability of sequencing data are outlined in the referenced manuscripts.

3.3.2. Classification criteria

Several serovars of *Salmonella* are associated with a range of illnesses, from localized gut infections to bacteremia, systemic infections and sepsis, and it would be extremely useful to identify genetic signatures that are associated with an increased risk of severe disease. Prior epidemiological records have led to the classification of serovars into

the gastrointestinal and extraintestinal pathovars. A majority of serovars in *Salmonella enterica subsp. enterica* belong to the gastrointestinal pathovar and are most often associated with gastrointestinal infections. On the other hand, a small percentage of these serovars are believed to have evolved beyond this level, allowing them to disseminate beyond the intestinal mucosa and colonize systemic sites within the host (Metris et al., 2017). In this study, we employed this understanding of gastrointestinal and extraintestinal serovars, as well as the classifications detailed by Abbott, Ni, & Janda (2012) and Nuccio and Baumler (2014), wherein the site of isolation and clinical characteristics of the infected host played a key role in determining the ability of *Salmonella* serovars and subspecies to invade the intestinal epithelium to infect the blood, therein causing severe disease, to putatively divide the samples into extraintestinal and gastrointestinal serovars (**Figure 3.1**). This allows the use of potential disease severity as dependent variables in model development.

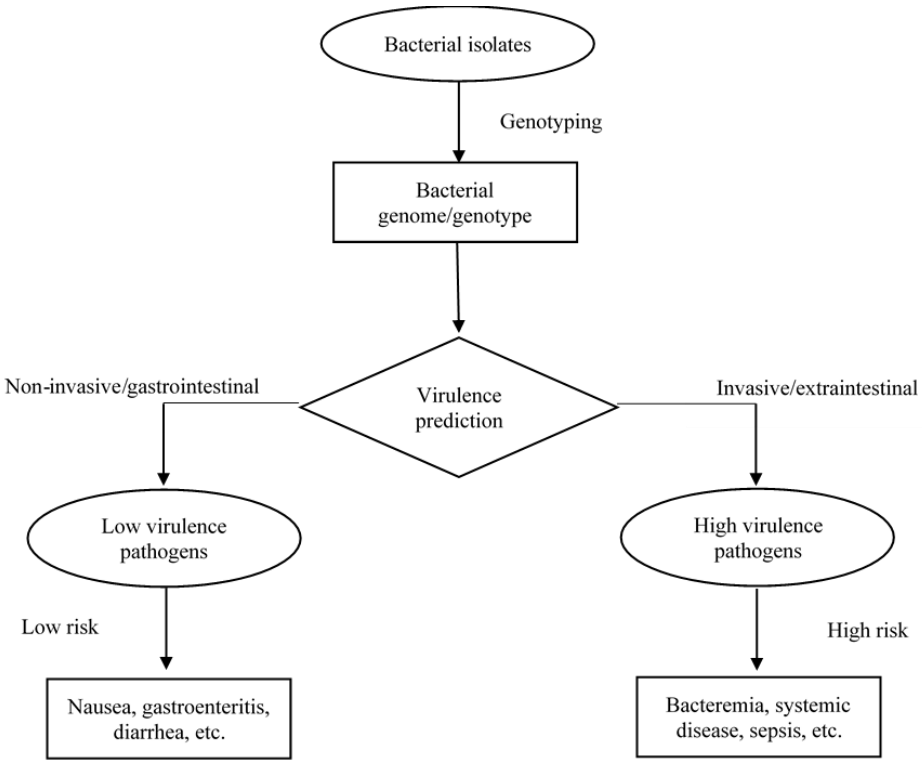

Figure 3.1. Extraintestinal and gastrointestinal virulence classification strategy for endpoint determination.

3.3.3. Bioinformatics analyses and development of a *Salmonella* pan genome

Short read sequences of all isolates from the included studies were obtained from the National Center for Biotechnology Information's (NCBI) BioProject and Sequence read archive (SRA) repositories, as well as the European Molecular Biological Laboratory database, and assembled using the PATRIC (v. 3.6.3) Bacterial Bioinformatics Resource Center, a freely available web-based platform for comprehensive comparative genomics and analyses, using the Rapid Annotation using Subsystem Technology (RASTk)-enabled genome annotation service (Brettin et al., 2015). The in-built SPAdes strategy was

employed for assembly, and assembly quality was assessed using the Quality Assessment Tool for Genome Assemblies statistics (Bankevich et al., 2012). The genomes were annotated on the same platform for uniformity of data using the in-built RASTk toolkit. Among the 850 sequenced *Salmonella* genomes obtained from the studies, 150 genomes that fit the parameters of our research and assembly and annotation quality (good sequence quality, sequence completeness score of >80%, and a contamination score <10%) were chosen in our final analysis.

The baseline machine learning model input included a *Salmonella* pan genome (in machine learning terminology, a "dictionary" of genes and gene homologs from the annotated sequences). This was created by aligning nucleotide sequences all-against-all using the *pairwise2* module in Python (Pedregosa et al., 2011). Briefly, each new annotated gene encountered by the program was locally aligned at 95% sequence similarity against the genes present in the "dictionary," with any sequences not showing a sufficient match being added as a new gene to the dictionary. This generated a dictionary of 33,185 unique genes, including potential gene homologs, which were nevertheless assumed to be heterologous, and thereby included as predictors in the initial model.

3.3.4. Predictive modeling by supervised machine learning

The genes identified in the gene "dictionary" were input in a range of ensemble classification and boosting models, including logistic regression, random forest, linear Support Vector Machine, and a gradient boosted tree model (AdaBoost), in order to find

the best prediction accuracy in distinguishing isolates according to clinical outcomes using genotype data. The overall machine learning schema is provided in **Figure 2.2**.

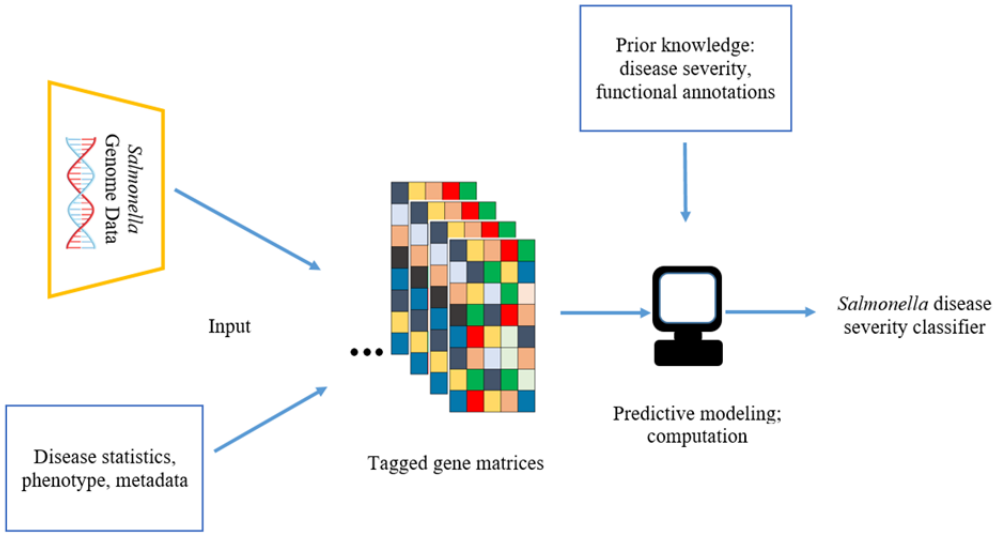

Figure 3.2. Machine learning-based disease outcome severity model schematic

3.3.4.1. Machine learning models

1. Logistic regression (*LR*) is a simple method to model the probability that a single or combination of genes can predict disease status. The simplest approach to model this probability is by fitting a pre-selected set of parameters (genes) to a linear logistic model:

$$logit\big(P(Y_i = 1|x_i)\big) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \text{.....(3.1)}$$

Where Y_i is a binary indicator for the disease severity of isolates $i = 1,\ldots, n$, X_{ij} denotes the values of individual genes $j = 1,\ldots, p$, coded as 0, 1, or m for absence, presence, and m multiple copy numbers, respectively. This approach will be carried

out on data sets of the *p* most significant genes/predictors. Since logistic regression does not perform satisfactorily when $p>>n$, the predictors are being pre-selected using the ridge penalty (which averages highly correlated predictors) on the training data alone (Bielzaa, Robles, & Larrangaa, 2011; Austin, Tu, Ho, Levy, & Lee, 2013).

2. Random forest (*RF*) is a generalized decision tree method wherein the data is repeatedly partitioned based on the values of the predictor variables into multiple decision trees using random samples of observations bootstrapped for each tree and random samples of the predictors. The resulting "forest" of these trees provides fitted values, which are more accurate than those of a single tree. The RF method is particularly useful in the analysis of data with $n<<p$, protecting against data reduction, overfitting, and multi-collinearity (Breiman, 2001; Liaw & Wiener, 2002; Matsuki, Kuperman, & van Dyke, 2016).
3. Support vector machine (*SVM*) with linear kernel is a kernel-based modeling algorithm that discriminates data points into specific categories by separating them with a hyperplane. In our analysis, a binary SVM was trained for binary classification using the sequential minimal optimization (SMO) algorithm. Maximum separation between the two classes is achieved by selecting the hyperplane with the largest "margin" (summation of the shortest distance from the separating hyperplane to the nearest data point of both categories) (Friedman, 1998; Hastie & Tibshirani, 1998; Yu & Kim, 2012).

4. Ada-boost (*AB*) or adaptive boosting is an ensemble method that has shown a great degree of accuracy in several fields of study. Boosting, a machine learning method proposed by Schapire (1990) and Freund (1995), is an ensemble method that postulates several hypotheses on different distributions, combining them to obtain a final hypothesis, yielding a final model by aggregating a large number of weighted models, which performs better than the individual constituent models (Ren, Zhang, & Suganthan, 2016). In essence, boosting creates a strong learner by reducing the bias and error in a weak learner algorithm (that proposes hypotheses with a precision that is intuitively better than random chance (>50%)), allowing for indefinite improvement in model precision and confidence (Collet, Fonlupt, Hao, Lutton, & Schoenauer, 2001).

3.3.4.2. Training the classifiers

Classifiers to identify significant features among the list of several features were trained on the *scikit learn* package in Python (Pedregosa et al., 2011), using a variety of parameters to assess model accuracy. Data was explored for class imbalance, which greatly impacts model accuracy and class-specific model performance (Velez et al., 2007; Lu et al., 2019) if the dataset is not perfectly class-balanced. However, as our combined and pre-processed dataset was not skewed towards any single clinical outcome class, we did not employ any additional methods for class-based bias reduction. Models were built by randomly dividing our dataset into training and testing sets ($2/3^{rd}$–$1/3^{rd}$ split), described to be optimal based on the sample size (n>100) (Dobbin & Simon, 2011). Hyperparameter

tuning was performed to determine the best model parameters using a 10-fold cross validation.

3.3.4.3. Model evaluation

Since our data fits the typical parameters of a high-dimensional, low sample size dataset, with the total number of predictors/features far outnumbering the total number of isolates/observations ($p>>n$), classifier accuracy could be very high for the training set, but low for the separate test set, known as model overfitting (Simon, Radmacher, Dobbin, & McShane, 2003; Subramanian & Simon, 2013). This was tested by analyzing the results of a confusion matrix, a tabular visualization of model performance, comprising two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (**Figure 3.3**).

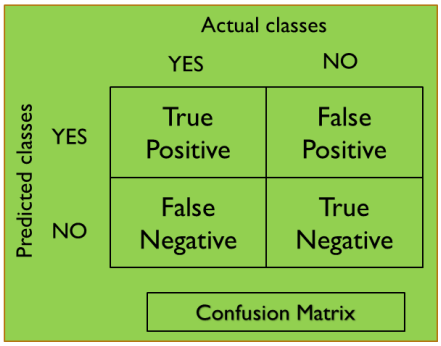

Figure 3.3. Confusion matrix setup to calculate model sensitivity, specificity, and accuracy.

A confusion matrix is an indicator of the number of correct and incorrect predictions made by the classifier compared to the actual outcomes (target value) in the

data, which allows us to evaluate model performance. Since our models analyze a binary classification problem (extraintestinal/positive vs. gastrointestinal, negative), a 2×2 confusion matrix was constructed, with the calculated accuracy depicting the agreement between the observed and predicted classes (Lasko, Bhagwat, Zou, & Ohno-Machado, 2005). We also analyzed model performance by determining the area under the receiver operating characteristic curve (AUC-ROC). The AUC measures the probability of differentiation between outcomes from a randomly collected sample independent of prior probabilities or test threshold, with AUC=0.5 indicating random or chance discrimination of the clinical outcome of isolates, and AUC=1 denoting perfect discrimination (Hastie, Tibshirani, & Friedman, 2001; Guyon, Elisseefi, & Kaelbling, 2003; Lin, Sintchenko, Kong, Gilbert, & Coiera, 2009). Analysis of variance at $\alpha = 0.05$ was used to analyze the differences in mean accuracy between the models.

3.3.4.4. Feature selection

Since the predictor variables to be used in the model are individual genes, and $n<<p$, it could lead to model overfitting. Reducing the number of non-discriminative features in genetic data with high dimensionality may improve the performance of machine learning algorithms, in a process known as regularization (Kooperberg, LeBlanc, & Obenchain, 2010; Subramanian & Simon, 2013), which shrinks the coefficient estimates towards zero, and discourages learning a more complex or flexible model to avoid the risk of overfitting. Elastic net is a penalized regression method that combines lasso and ridge to minimize data overfitting (StataCorp, 2021). When $l(\beta; Y_i,$

x_i, $i = 1,\ldots,n$) is the logistic log-likelihood, the elastic net estimate of β is the maximizer of

$$l(\beta;\ Y_i, x_i, i = 1, \ldots, n) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \beta_j^2 \quad \text{.....(3.2)}$$

where λ_1 and λ_2 are selected manually. This approach will effectively perform model selection, as the l_1 penalty $\lambda_1 \sum_{j=1}^{p} |\beta_j|$ effectively sets many coefficients β_j to 0, and the l_2 penalty $\lambda_2 \sum_{j=1}^{p} \beta_j^2$ encourages averaging of highly correlated predictors (Saabos, 2014).

3.3.5. Identification of significant predictors

In addition to determining prediction accuracy, it is important to identify features that make a relevant and informative impact on the phenotype of interest. Here, we rank features based on their importance to the phenotype of interest. In the case of a regularized logistic regression, this can simply be achieved by analyzing the coefficients for the variables. Simply put, the magnitude of the coefficient determines its importance to the model, that is, when all features are on the same scale, features adding substantially to the model have the highest coefficients, with uncorrelated features expressing coefficient values close to zero (Bielzaa, Robles, & Larrangaa, 2011). Proteins coded by the important genes were predicted by conducting a Protein Basic Local Alignment Search Tool (BLASTP) search on the NCBI web server (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) and compared against the Gammaproteobacteria sequences from Uniprot (http://www.uniprot.org) (Uniprot, 2017).

3.4. Results

3.4.1. Strain characteristics

Our dataset included *Salmonella* genomes obtained from human clinical specimens, and chicken and swine specimens (**Table 3.1**), and incorporated a mix of serovars, including Enteritidis, Typhimurium, Kentucky, Choleraesuis, Johannesburg, Senftenberg, Rissen, Derby, Newport, Gallinarum, and Pullorum, among others. The genomes were characterized as "extraintestinal" or "gastrointestinal," according to classifications made in previous epidemiological studies (Abbott, Ni, & Janda, 2012; Nuccio & Baumler, 2014; Wheeler, Gardner, & Barquist, 2018), based on the site of isolation within the host (blood or extraintestinal site such as the liver, brain, and kidneys), patterns of host restriction, and clinical characteristics observed in hosts.

Table 3.1: Isolate characteristics.

Isolate characteristics	**Extraintestinal**		**Gastrointestinal**	
	(n = 85)		(n = 65)	
Human	41	(27.3%)	32	(21.3%)
Poultry	3	(2%)	22	(14.6%)
Swine	41	(27.3%)	11	(7.3%)

3.4.2. Predictive analyses using machine learning

Here, we employed an approach for classification of *Salmonella enterica* into extraintestinal or gastrointestinal isolates based on clinical characteristics of infection in the host.

3.4.2.1. Baseline predictive modeling and model evaluation

The predictive performances of the different machine learning models were evaluated based on hold-out accuracy. The AUC **(Figure 3.4)** and balanced accuracy and predictive power obtained from the 2×2 confusion matrix **(Table 3.2)** were evaluated.

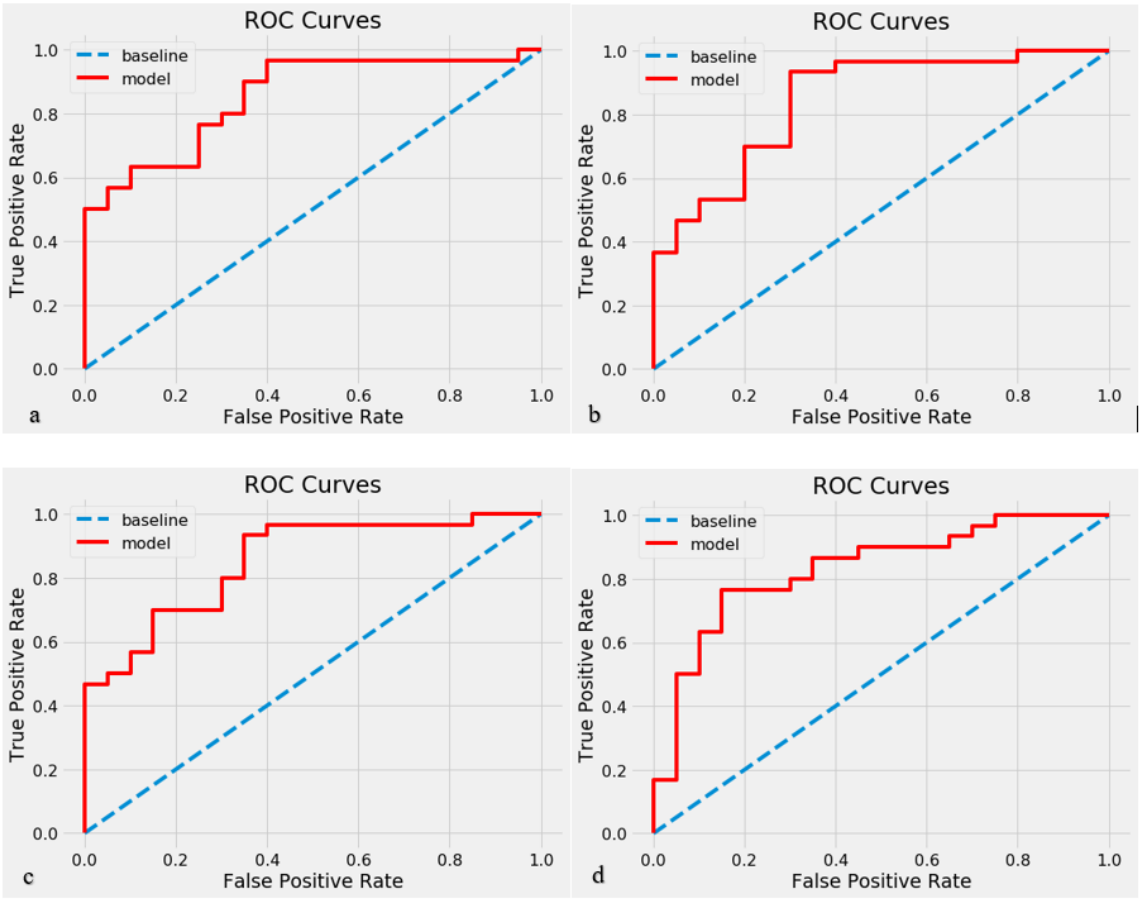

Figure 3.4. Receiver operator characteristic (ROC) curve depicting model accuracies of different machine learning classifiers without feature selection. Using BioPython, 33185 unique genes were identified from our sample isolates, which were analyzed by machine learning. Receiver operating characteristic (ROC) curves for (a) Random Forest (area under the curve, AUC = 0.85), (b) Support Vector Classifier with

linear kernel (AUC = 0.84), (c) Logistic regression (AUC = 0.85), and (d) AdaBoost (AUC = 0.83) classifier performance.

Although RF and LR resulted in a high AUC value (0.85), the models showed very low specificity (true negative rate = 0.23 for both) and positive predictive value (0.68), indicative of model overfitting. On the other hand, the boosting model AdaBoost, which has been previously shown to perform remarkably in filtering gene-gene interaction effects (Hoffmann, 2001; Zou & Hastie, 2005) (thereby excluding parameters that do not contribute to the predictive capacity of the model), showed comparatively better model accuracy, with an AUC value of 0.83, and higher sensitivity (0.8). However, the specificity remained dismal at 0.23 **(Figure 3.4; Table 3.2).**

Table 3.2: Sensitivity and accuracy of the machine learning (ML) classifiers with (w) and without (w/o) feature selection using Elastic Net.

		RF		SVM		LR		AB	
		w	w/o	w	w/o	w	w/o	w	w/o
Sensitivity	TP/TP+FN	0.7	0.75	0.6	0.7	0.75	0.7	0.75	0.8
Specificity	TN/TN+FP	0.73	0.23	0.76	0.23	0.76	0.23	0.7	0.23
Positive Predictive value	TP/TP+FP	0.63	0.68	0.63	0.67	0.68	0.67	0.62	0.69
Negative Predictive value	TN/TN+FN	0.78	0.82	0.74	0.79	0.82	0.79	0.80	0.85
Class balance accuracy	TP+TN/total	0.72	0.76	0.7	0.74	0.76	0.74	0.72	0.78

[1]Sensitivity, specificity, positive and negative predictive values, and model accuracy were computed using the confusion matrix for the test dataset (33% of included observations).

[2]RF – Random Forests, SVM – Support Vector Machine, LR – Logistic regression with ridge correction, AB – AdaBoost, TP – true positive, TN – true negative.

3.4.2.2. Application of feature selection algorithms

The accuracy of models developed with a large number of predictors is generally low due to dimensionality issues and model overfitting. However, the application of sparse regression methods such as Elastic Net (Baker & Dougan, 2007), can greatly reduce this issue by estimating parameters (genes) that are significant to the model, and can also help in selection of the final model. Application of the Elastic Net correction to our four machine learning models ($2/3^{rd}$–$1/3^{rd}$ train-test split) using an approach detailed by Kooperberg, LeBlanc, & Obenchain (2010), led to the following test-sample-validated predictions. The Elastic Net-corrected LR model with ridge penalty showed the best AUC (0.86; **Figure 3.5**) and balanced accuracy (0.76), which was significantly different from the remaining models, both with and without feature selection ($p < 0.05$). The sensitivity, specificity, negative predictive values were all ≥ 0.7, and the positive predictive value was > 0.65, showing a significant improvement compared to the non-feature selected models (**Table 3.2**).

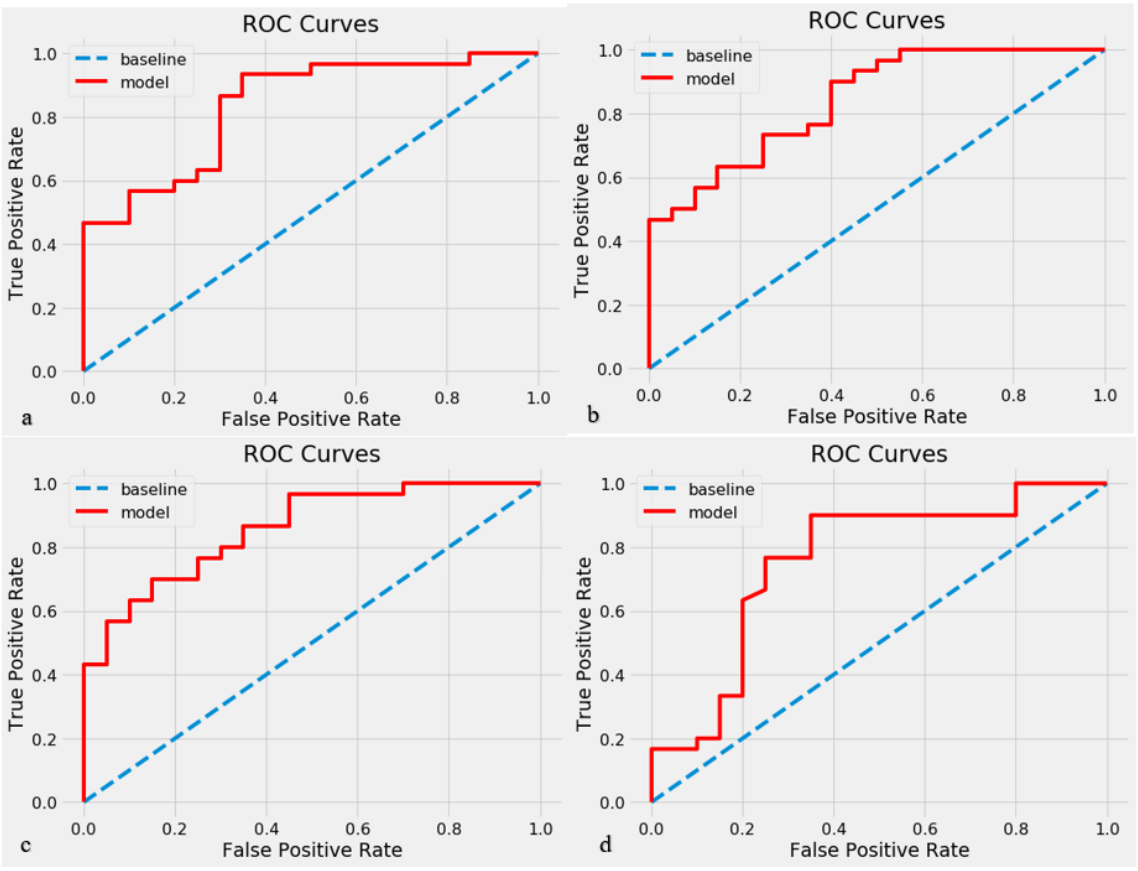

Figure 3.5. Receiver operator characteristic (ROC) curve depicting model accuracies of different machine learning classifiers with feature selection. Elastic Net regularization selected 176 predictor genes and eliminated the other 33009 as not being significant to the model. Receiver operating characteristic (ROC) curves for (a) Random Forest (area under the curve, AUC = 0.84), (b) Support Vector Classifier with linear kernel (AUC = 0.84), (c) Logistic regression (AUC = 0.86), and (d) AdaBoost (AUC = 0.76) classifier performance.

3.4.3. *Salmonella* molecular markers associated with extraintestinal vs. gastrointestinal disease

Analyzing for feature importance identified a number of genes that were important towards predicting severity of clinical outcome (**Figure 3.6**). These included genes putatively coded for integrases, lipoproteins, virulence proteins, phage proteins, transcriptional regulators, fimbrial and mobile element proteins, metabolite transporters, type III secretion system (T3SS) effectors, and heat shock proteins, among others. We observed that genes coding for proteins associated with bacterial membrane stability, DNA replication, and transcription were significantly associated with prediction of extraintestinal isolates ($p < 0.05$), while those coding for proteins required for survival in host tissue (transcriptional regulators, RNA transporters, etc.) appeared to play a role in predicting isolates capable of gastrointestinal infection. This appears to be in agreement with the results observed for gastrointestinal and extraintestinal serovar prediction shown by Wheeler, Gardner, & Barquist (2018) and Nuccio and Baumler (2014). Interestingly, a number of genes coding for hypothetical proteins were shown to be significant in the prediction of both endpoints, indicating the need for additional studies to identify their specific functional properties.

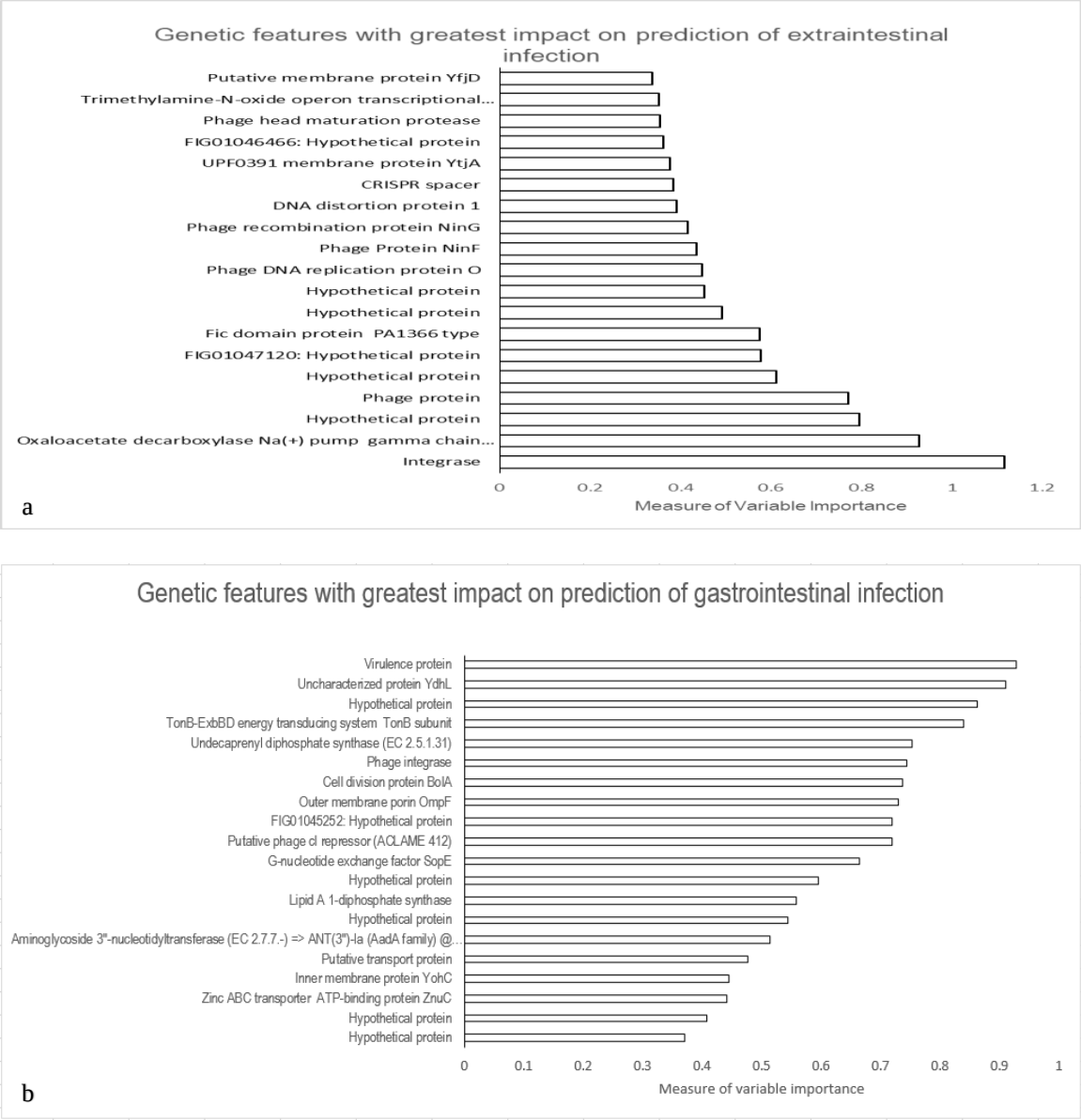

Figure 3.6. Genes identified as important predictors of disease outcome by the regularized logistic regression model.

Subsample analyses were additionally performed on individual host systems to identify any host-specific gene presence/absence patterns of importance. Although the results did not differ significantly from those observed for the generalized model, notable

differences were disregarded to preserve the generalizability of the model, as well as to not compromise on sample size.

3.5. Discussion

The aim of this study was to identify the best machine learning classifier to predict *Salmonella* clinical outcomes, highlight the need for feature selection in genomic dataset analysis, as well as identify genetic signatures significantly associated with each clinical outcome. Foodborne pathogens such as *Salmonella* exhibit many important phenotypic traits such as resistance to stress, host adaptation, survival and growth potential (Franz, Mughini-Gras, & Dallman, 2016), traits that could increase the specificity of microbial modeling by focusing on subtypes of pathogens that could pose the greatest risk (Deng, den Bakker, & Hendriksen, 2016). For the purpose of this research, a simple discriminative approach was selected to accurately model the boundary between two classes (extraintestinal vs. gastrointestinal), in order to obtain a quantitative metric for disease severity.

Machine learning algorithms predict the outcomes of complex mechanisms by isolating the most relevant input from large datasets for a specific output, while circumventing the need to understand the underlying mechanisms (Baker, Pena, Jayamohan, & Jerusalem, 2018; Vilne, Meistere, Grantiņa-Ieviņa, & Ķibilds, 2019; Xu & Jackson, 2019). Typically, disease gene prediction in biological systems, wherein the most significant disease genes are identified, can be formulated as machine learning

classification problems (Speiser, Miller, Tooze, & Ip, 2019). The past few years have seen an emergence in popularity of various ensemble methods, including tree-based RF, the kernel-based SVM, and boosting algorithms, for classification-based disease or disease-severity prediction (Austin, Tu, Ho, Levy, & Lee, 2013; Huang et al., 2018; Kegerreis et al., 2019; Njage, Leekitcharoenphon, & Hald, 2019). However, logistic regression, the simplest form of machine learning, is not commonly used for classification problems using genetic data since they perform poorly under conditions where $p>>n$, even though they are generally easier to interpret compared to other methods such as RF (Hanley & McNeil, 1983; Subramanian & Simon, 2013). Therefore, for a logistic regression to be effectively used in WGS data-based disease prediction, there is a need to incorporate a method to include only the most relevant features in the model. This is generally accomplished by the inclusion of a pre-selection method such as feature selection, which in this case means the identification and selection of the smallest possible set of relevant genes that can help achieve good predictive performance in sample classification. In our study, we propose the use of Elastic Net to identify features significant to the prediction, because it has been previously shown to outperform other feature selection methods in identifying true positives with high accuracy (Koller & Sahami, 2006).

A typical measure of classifier performance is classification accuracy, percentage of correctly classified observations. However, this metric is generally not suitable when dealing with two-class problems with skewed classes. In such cases, computing the AUC-ROC, which ranges from 0 (random classification) to 1 (perfect discrimination between

both classes) is an effective alternative (Díaz-Uriarte & Alvarez de Andrés, 2006). In our study, classifier performance was computed both by analyzing the classification accuracy and the AUC. We observed that, although the AUC was consistently high for all algorithms, the classifiers resulted in dismal specificity in identifying true negatives. Since the baseline models were developed using all genes in the developed dictionary, we attribute this to model overfitting. On the other hand, we saw a marked improvement in classification accuracy and prediction of true positives and true negatives in Elastic Net-regularized models, with LR showing the best AUC and accuracy values.

The final LR model identified a number of *Salmonella* genes that were significant in predicting host disease severity. Our analysis showed that a number of genes coding for *Salmonella* phage proteins were significant to predicting severe disease in the host. These genes coded for proteins that ranged in functionality from virulence and systemic infection acceleration, type III secretion system effectors, and putative/hypothetical proteins associated with bacterial survival under conditions observable in extraintestinal regions (Worley, Nieman, Geddes, & Heffron, 2006; Thornbrough & Worley, 2012). Bacteriophages in *Salmonella* have been known to harbor a number of virulence proteins, such as SrfH, which has been previously shown to promote phagocyte motility and accelerate systemic infection spread (Worley et al., 2006). Additionally, these elements have been isolated from many serovars of *Salmonella* previously considered to be avirulent, indicating their transmissibility among strains (Thornbrough & Worley, 2012). Similarly, the phage proteins NinF and NinG have been associated with enteric infections,

bacteremia, and meningitis, among others extraintestinal isolates of *Escherichia coli* and *S.* Typhimurium (Pizza, 2006; Desai & McClelland, 2013). The putative membrane protein yfjD has been shown to be express positive fold change in *S.* Typhimurium subjected to desiccation stress (Maserati, 2017). Additionally, the transcriptional regulatory protein TorR, which encodes one half of the two component regulatory system TorR/TorT in the tor operon (Hu et al., 2019), has also been identified as a significant predictor of severe extraintestinal disease in the host. Finally, the putative protein YfjD, while undefined in *Salmonella*, has been associated with hemolysis in *Bacillus subtilis*. This could be another example of horizontal transmission of virulent genes, causing an increase in disease severity in the hosts (Liu, Fang, Jiang, & Yan, 2009). On the other hand, genes predicting gastrointestinal disease worked towards enabling *Salmonella* survival, growth, and expansion under severe and harsh conditions, such as anaerobic, acidic, and oxidative stress. For example, the stress response protein BolA has been shown to help *S.* Typhimurium overcome host defense conditions (host cellular response), allow bacterial proliferation during the latter stages of bacterial growth, and survive acidic and oxidative stress (Mil-Homens et al., 2018). Similarly, ydhL, belonging to the ydh protein family, is an oxidoreductase that helps *Salmonella* survive oxidative stress caused by host phagocyte activity (Kim, Liu, Husain, M., & Vasquez-Torres, 2016). Alternatively, the *Salmonella* pathogenicity island 1 (SPI-1) effector protein sopE has been previously shown to elicit intestinal inflammation in the *in vivo* murine host model (Hapfelmeier et al., 2004). Additionally, proteins responsible for cellular maintenance, including znuC, an ATPase that provides energy to the bacteria for zinc uptake (Liu, Yan, Liu, & Chen, 2013), and the

membrane protein yohC were found to be significant. Finally, the aminoglycoside adenylyltransferase (aadA) protein, responsible for conferring streptomycin resistance in *Salmonella* (Singh, Drolia, Bai, & Bhunia, 2015), and the outer membrane protein ompF, responsible for conferring cephalosporin antibiotic resistance in *Salmonella* (Choi et al., 2018), were also found to be significant in predicting gastrointestinal disease.

These results provide us with a clearer idea of genes that play a role in determining the severity of host disease. This could be attributed to niche adaptation caused by parallel evolution undergone by the various *Salmonella* isolates, which allows the different isolates to survive in different host environments despite belonging to the same genus (Wheeler, Gardner, & Barquist, 2018). *Salmonella* infections and disease in the host have a wide range, from infections isolated to the gastrointestinal system (including nausea, vomiting, and diarrhea, among others), to those that transcend the intestinal barrier to colonize other parts of the body (i.e., extraintestinal infections, including bacteremia, systemic disease, and sepsis). This distinction is important to identify strains of *Salmonella* posing a relatively higher risk of severe infection in the host, and can potentially help in the development of virulence-subtype-specific dose response models, as proposed by Fritsch et al. (2018).

3.6. Conclusion

In conclusion, machine learning is a powerful tool that can be trained to identify patterns from WGS and other such high dimensional datasets that are indicative of a

specific outcome. However, the widespread use of this method, especially in predictive microbiology and WGS-informed risk assessment, is hindered by the lack of availability of enough number of isolates with associated metadata to draw meaningful inferences. Metadata, including isolation date, time, and health outcome information, are the need of the hour. However, data with such granularity is currently unavailable, since this may not fit the parameters for sample collection set by the collection agency. Therefore, an important step towards developing and validating these models is to update these metadata parameters during the collection phase of bacterial isolates. Our model helped identify genes that were significant predictors of disease severity in the host. These ranged from transcriptional regulators and stress response genes (gastrointestinal) to virulence and survival under anaerobic conditions (extraintestinal), allowing us to identify patterns associated with each form of disease. However, a number of significant predictors remained undefined, and warranted further investigation for homology with other virulence-associated genes or signals of horizontal gene transfer. We envision this as the first step towards the widespread incorporation of WGS in a predictive modeling framework, in delineating risk patterns based on the predicted disease severity in the host.

Chapter 4: Development of a weighted modeling approach to incorporate genetic heterogeneity in a dose-response modeling framework

4.1. Abstract

Estimating microbial dose-response is an important aspect of a food safety risk assessment. In recent years, there are considerable interests to advance these models with potential incorporation of gene expression data. The aim of this study was to develop a novel machine learning model that considers the weights of expression of *Salmonella* genes that could be associated with illness, given exposure, in hosts. Herein, an Elastic Net-based weighted Poisson regression method was proposed to identify *Salmonella enterica* genes that could be significantly associated with the illness response, irrespective of serovar. The best-fit Elastic Net model was obtained by 10-fold cross validation. The best-fit Elastic Net model identified 33 gene expression-dose interaction terms that added to the predictability of the model. Of these, 9 genes associated with *Salmonella* metabolism and virulence were found to be significant by the best-fit Poisson regression model ($p < 0.05$). This method could improve or redefine dose-response relationships for illness from relative proportions of significant genes from a microbial genetic dataset, which would help in refining endpoint and risk estimations.

4.2. Introduction

Salmonella enterica is a major cause of foodborne illness and significant economic burden worldwide, and has a high morbidity and mortality rate. Recent years have seen the emergence of new pathogenic serovars of *Salmonella* in cases and outbreaks of foodborne salmonellosis, necessitating the identification and development of novel methods and models to estimate the human health risk posed by these emerging variants of this bacteria (CDC, 2019). Quantitative microbial risk assessment (QMRA) is a modeling approach for estimating the risk of infection and illness as a result of exposure to microorganisms in the environment. Estimating and predicting the dose-response relationship is one of the most important aspects of a food safety risk assessment. Dose-response models provide us with the probability estimate of a specific response (such as infection, illness, or death) as a result of consuming/ingesting a specific dose of the pathogen (USDA-FSIS, 2005; QMRAWiki, 2019). A major challenge in food safety microbial risk analysis is to identify and predict relationships between low-level pathogen exposure and the potential public health outcomes (Buchanan, Havelaar, Smith, Whiting & Julien, 2009). Some of the major factors that affect the probability estimates of host illness or infection include (i) the dose, or the number of ingested organisms, over a defined period of time, (ii) virulence factors, gene expression, and other factors describing the nature of the pathogen, and (iii) the pathogenic strain (due to there being substantial strain-to-strain differences) (USDA-FSIS, 2005).

A current major limitation of existing QMRAs is that dose-response models for pathogens with multiple subtypes and serovars that can successfully infect human hosts (such as *Salmonella*) do not always take into account differences in survivability and virulence among strains. Such information is best gleaned from bacterial "omics" information, which could assist researchers in exploring bacterial growth, survival, and virulence dynamics, as well as various pathogen-host interactions leading to variations in susceptibility in the host (Brul et al., 2012; Njage, Henri, Leekitcharoenphon, Mistou, & Hald, 2019). With this in mind, efforts have been made to account for this variability in *Salmonella* virulence in developing dose-response models by defining its serotypes or serovars as strains (Oscar, 2004; Teunis et al., 2010). However, these models do not account for intra-serovar differences, or the potential for horizontal gene transfer between pathogenic and non-pathogenic variants of the same bacteria, necessitating the incorporation of higher-resolution data (such as WGS data) to improve their precision in capturing the variability in *Salmonella* virulence and its resulting effects in the host.

The inclusion of WGS data would introduce a substantial number of variables that current predictive microbial models are not equipped to handle. This can be overcome by the application of machine learning techniques to identify genetic variables that could add significance to these models. Prior research studies have (Wheeler, Gardner, & Barquist, 2018; Njage, Henri, Leekitcharoenphon, Mistou, & Hald, 2019; Njage, Leekitcharoenphon, & Hald, 2019; Munck, Njage, Leekitcharoenphon, Litrup, & Hald, 2020) proposed incorporating genetic data into dose-response models by assuming each

bacterial unit to be composed of individual genetic units. Such a method would remove the need for individual subspecies or serovar count data to determine the individual subtype-based response in human cases of foodborne disease, by instead including the probability of expression of individual genes as potential "weights" to a regular dose-response model. However, the development of such a model is hindered by the lack of studies that analyze bacterial gene expression in the context of host response to an ordered exposure to microbes.

In this study, our aim was to propose a novel machine learning-based approach to incorporate genes into a dose-response framework for *Salmonella enterica*. Herein, we use Elastic Net to identify genes from a multi-serovar pan-genome that could significantly impact dose-based host response to *Salmonella enterica* exposure irrespective of serovar. For this purpose, we used a weighting method that considers the weights of expression of genes that could be associated with a host response, obtained from pooled data from prior *Salmonella* dose-response modeling studies (Oscar et al., 2004; Teunis et al., 2010). The method proposed in this study could provide us with a new means to incorporate WGS data into a QMRA framework for *Salmonella*, as well as other pathogenic microorganisms.

4.3. Material and Methods

4.3.1. Data collection and preliminary analyses

4.3.1.1. Curation of prior dose-response models

A combination of human-feeding trial and outbreak-based dose-response model data were included in our study. Dose-response data were obtained from 1) prior human feeding trials conducted by McCullough & Eisele (1951 a, b), and described in Oscar (2004), and 2) outbreak data previously reported by George (1976), Fontaine et al. (1978), and Kasuga et al. (2004). Of the outbreak data, only those studies with relatively complete counts were employed for model building (**Table 4.1**).

4.3.1.2. Isolate selection for creation of **Salmonella** *gene dictionary (pan genome)*

Isolates of *Salmonella* belonging to the serovars Anatum, Bareilly, Derby, Enteritidis, Heidelberg, Oranienberg, Newport, Schwarzengrund, and Typhimurium were sampled from the National Center for Biotechnology Information's (NCBI) Pathogen Detection database. The isolates were selected from among those isolated from food sources (such as meat and poultry in the various stages of processing), farm animals, the environment, and from human clinical cases, in order to incorporate the genetic variations observable in various isolate environments. Additionally, isolates were randomly picked between years 1998 and 2019 to ensure that any potential gene mutations or horizontal gene transmission over time would be captured while computing the probability of gene expression. Approximately 50 isolates were selected per serovar, a number which was set based on data availability, in order to be on par with the serovar with the lowest number of

sequenced isolates, and not introduce imbalance in our final model. Metadata, including the year of isolation, place of isolation, and collection agency, among others, were collected for all included isolates.

4.3.1.3. Bioinformatics analysis and creation of **Salmonella** *pan genome*

Sequence Read Archive (SRA) Run Accession numbers for all included isolates were obtained from the NCBI SRA repository. The isolates were *de novo* assembled and annotated using the PATRIC (v.3.6.3) Bacterial Bioinformatics Resource Center, a freely available web-based platform for comprehensive comparative genomics and analyses. All bioinformatics analyses were performed in accordance with the method described in Section 3.3.3. A total of 414 isolates across the 9 included serovars which fit the quality parameters detailed in 3.3.3. were included for the creation of our gene dictionary/pan-genome and to compute the probability of gene expression per serovar.

A *Salmonella* pan genome (or dictionary) was created from the 414 annotated sequences, with each individual gene being input as a variable in our baseline model. In simple terms, this comprises a set of features that represent the input data, providing some form of parametrization of the input space used to represent the prediction function (de Mol, de Vito, & Rosascode, 2009). This was created by aligning nucleotide sequences all-against-all using the *pairwise2* module in Python, as described in Section 3.3.3. The dictionary comprised 31,030 unique genes, including potential gene homologs, which were included as predictors in the initial model. Ninety-six CRISPR repeats and an equal number

of CRISPR spacers (and their homologs) were removed from the final gene dictionary due to being repetitive, despite potentially contributing to the virulence and pathogenicity potential of *Salmonella* (Louwen, Staals, Endtz, van Baarlen, & van der Oost, 2014), similar to the Roary platform.

4.3.2. Calculating the probability of gene expression

The empirical probability of gene expression was calculated to add weights to our dose-response model as:

$$p(gene\ expression) = \frac{Number\ of\ times\ the\ gene\ is\ expressed}{Number\ of\ samples\ included\ per\ serovar} \quad \text{.....(4.1)}$$

For example, the gene Arsenic metallochaperone ArsD (transfers trivalent metalloids to ArsAB pump) showed a *p(gene expression)* = 0.98 in serovar Typhimurium, whereas the same gene had a *p(gene expression)* = 0 in serovar Newport.

The empirical probability was computed, as opposed to the theoretical probability (which reflects the number of times an event is expected to occur relative to the number of times it could potentially occur), which is calculated based on the circumstances that result in the expression of the gene, such as the specific environmental, stress, or host-interaction conditions that *Salmonella* is exposed to, that are as yet to be quantified thoroughly. Briefly, in an ideal experiment for gene-based dose-response modeling, we would be presented with the genetic data from isolates obtained from test subjects fed with a specific

dose of *Salmonella*. However, since this is not possible, we determine which genes are differentially present within each serovar (with the understanding that genes present in *all* serovars would not provide us with a statistically differential effect in the final model). Computing the probability allows us to estimate the potential for a gene to be expressed in a random isolate of the same serovar, and provide us with a differential metric to incorporate into our dose-response model.

Finally, the dose-gene expression interaction terms were created from the dose from the original dose-response models interacted with the probability of gene expression. For example, if a gene is expressed 60% of the time, our interaction term used in the model would multiply the dose with 0.6. We input this interaction term into the elastic net regularization model to identify the most significant genes/gene interaction terms adding value to the dose-response model.

4.3.3. Identification of important dose-gene interaction terms by Elastic Net

All statistical analyses and modeling were performed on STATA 16 (StataCorp, 2019). The predictability of the outcome and simplicity of a model suffers in cases where the number of features *p* in the input space is very large (or over-complete) compared to the number of samples *n*, which is simply referred to as the "large p, small n" problem (Candes & Tao, 2007). Additionally, genes sharing the same biological pathway would show a high number of correlations (Zou & Hastie, 2005). Therefore, based on suggestions from prior research, we employ Elastic Net, a powerful penalization technique, to shrink

the β of unimportant variables towards zero. Elastic Net automatically selects important features and apply continuous shrinkage to the feature dictionary, while selecting groups of correlated variables that add significantly to the model (instead of just retaining one in the group and discarding the others; Zou & Hastie, 2015). Elastic net is a penalized regression method that combines lasso (wherein many of the coefficient estimates are exactly zero) and ridge (all coefficients are nonzero, although many are small) to minimize data overfitting. The penalized objective function for Elastic Net is

$$Q = \frac{1}{N}\sum_{i=1}^{N} w_i f(y_i, \beta_0 + x_i\beta') + \lambda \sum_{j=1}^{p} \kappa_j (\frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta_j| \quad \text{.....(4.2)}$$

where N indicates the number of observations, w_i denotes the observation level weights, $f()$ denotes the likelihood contribution for the Poisson model, β_0 denotes the intercept, x_i is the 1 x p vector of covariates, β is the p-dimensional vector of coefficients on covariates x, λ is the lasso penalty parameter that must be greater than or equal to 0, K_j are coefficient level weights, and α is the Elastic Net penalty parameter that can only take on values in the [0, 1] dimension. Estimated β are those that minimize Q for given values of α and λ (penalty coefficient). Here, when $\alpha = 1$, Elastic Net reduces to lasso, and when $\alpha = 0$, it reduces to ridge regression. The functional form for the function $f()$ used when the model is Poisson is

$$f(\beta_0 + x_i\beta) = -y_i(\beta_0 + x_i\beta') + e^{(\beta_0 + x_i\beta')} \quad \text{.....(4.3)}$$

Elastic Net regularization was performed on STATA using the *elasticnet* function. In order to fit the model with Elastic Net, a set of candidate α values and a fine grid of λ values was selected. This followed the rationale set by Hastie, Tibshirani & Wainwright

(2015) that only a few points in the space between ridge regression and lasso (i.e., α value) need to be reviewed, but a finer grid over λ is needed to identify non-zero coefficients. In addition to the default candidate α values of 1, 0.75, and 0.5, lower and upper bounds of α (0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 0.8) were tested. The λ grid was set automatically. 10-fold cross validation was performed on the combined set of (α, λ) values, and the (α, λ) pair that minimized the value of the cross validation function was ultimately selected. The non-zero coefficients identified by this (α, λ) pair were then deemed as being significant and employed in further models (Stata, 2021).

4.3.4. Dose-response model development

We propose a weighted approach to the incorporation of genes of importance in a dose-response paradigm. The response variable was set as the %incidence, obtained from dose-response data from prior human-feeding and outbreak studies. We analyzed our pooled dataset (outbreak and human-feeding trial), as well as the individual subsamples (outbreak-associated dataset and human-feeding trial-associated dataset), against important genes from corresponding *Salmonella* whole genome sequences obtained via machine learning (predictor variables).

4.3.4.1. Elastic Net-based weighted Poisson regression model

In this study, we propose a weighted Poisson regression model for gene-based dose-response. This is in agreement with prior *Salmonella enterica* dose-response, which has typically followed the beta-Poisson format (McCullough & Eisele, 1951a, b; Meynell &

Meynell, 1958). This also applies to our dataset since the outcome variable is expressed in terms of *rate* data, i.e., illness (*Y*) given exposure (*t*) to a *Salmonella* dose (Anderson, 2019; Dataquest, 2019). Hence, we can interpret our model outcome as a percentage change in our predictor variable triggers a unit change in the response variable (which in itself is a percentage value). Therefore, our model is structured as:

$$y_i/_t = e^{\alpha} e^{\sum_{k=1}^{n} \beta_{ik} X_{ik}} \quad \text{.....(4.4)}$$

Where, the response variable *y/t* denotes the probability of illness given exposure, *i* is the individual dose-response observation included in the model, X_{ik} denotes the probability of a gene *k* being expressed in observation *I*, and β_{ik} denotes the weights of genes $k = 1 \ldots n$ (n = 31,030). While the total number of included genes from the initial dictionary is very large, the β_{ik} values for only those genes that are deemed significant by the Elastic Net model were included in the final regression model (with all genes with zero values being automatically eliminated from the model).

We also performed separate subsample analyses for outbreak and human feeding trial data to determine the effect on data type on the identification of genes informative to dose-response modeling.

4.4. Results

Here, we propose a machine learning-based method to incorporate whole genome sequencing data in a dose-response modeling framework that transcends serovar-level heterogeneity. In order to achieve this, (i) whole genome sequences of *Salmonella enterica*

serovars corresponding to available dose-response data were pre-processed to create a gene dictionary, (ii) gene expression weights were assigned based on serovar-level expression data, (iii) important genes were identified using Elastic Net regularization, and (iv) significant genes were incorporated in a log-linear dose-response regression model.

4.4.1. Dose-response and WGS data collection and pre-processing

Dose-response data from prior human feeding trials and *Salmonella enterica* outbreaks was obtained from literature. Datasets were pooled based on availability of complete dose-response data as well as bacterial isolates corresponding to the individual serovars. We employ a pooling design similar to the Collaborative Study Design (CSD) described by Lesko et al. (2018). In essence, the scientific commonality between the studies (i.e., dose-response) was the driver for dataset selection, irrespective of study design heterogeneity (trial vs. observational studies). This was done in order to increase our sample size, as well as investigate effect heterogeneity due to diversity in data types (Lesko et al., 2018). The response variable (probability of illness given exposure) was then standardized across the studies to remove a potential source of heterogeneity across the data. The final dataset comprised dose-response data across nine serovars (**Table 4.1**).

In order to identify the predictor variables, whole genome sequences across nine *Salmonella enterica* serovars (Anatum, Bareilly, Derby, Enteritidis, Heidelberg, Oranienberg, Newport, Schwarzengrund, and Typhimurium) were sampled from the NCBI Pathogen Detection web server. Isolates were selected from across a number of human,

animal, and environmental isolation sources to account for genetic recombination, and directionality and timing of evolutionary changes within and among serovars (Grad & Lipsitch, 2014). Short reads for each isolate were assembled and annotated on the PATRIC web server for homogeneity. The *Salmonella* pan-genome was created from the annotated sequences using settings similar to that employed by Roary (Page et al., 2015), resulting in a gene dictionary composed of 31,030 genes.

Table 4.1. Dose-response data from human feeding trials and salmonellosis outbreaks used in model building.

Study	Serovar	Food source (for outbreak-related cases only)	Dose (log10CFU)	Number Ill	Number fed	Incidence (%)
Human feeding trial-associated data points						
Eisele & McCullough 1951a	Anatum A1	*Salmonella*-spiked eggnog	4.08	0	5	0
			4.38	0	6	0
			4.82	0	6	0
			4.97	0	6	0
			5.15	0	6	0
			5.41	0	6	0
			5.77	2	6	33
			5.93	3	6	50
	Anatum A2		4.95	0	6	0
			5.65	0	6	0
			6.02	0	6	0
			6.59	0	6	0
			7	0	6	0
			7.38	0	6	0
			7.65	1	6	17
			7.38	4	8	50
	Anatum A3		5.2	0	6	0
			6.1	2	6	33
			6.67	4	6	67
	Bareilly		5.1	1	6	17

Eisele & McCullough 1951b			5.84	2	6	33
			6.23	4	6	67
	Derby		5.14	0	6	0
			5.85	0	6	0
			6.22	0	6	0
			6.81	0	6	0
			7.18	3	6	50
	Newport		5.18	1	6	17
			5.59	1	8	13
			6.13	3	6	50
Outbreak-associated data points						
Kasuga et al. 2004	Bareilly	Sauce for octopus pancake	7.14	34	68	50
Kasuga et al. 2004	Enteritidis	Tartar sauce	3.55	36	126	28.6
		Omelet	5.17	10	11	90.9
		Seared beef	5.38	3	5	60
		Natto with raw eggs	5.87	9	9	100
		Scallop cream sauce	6	30	38	78.9
		Grated yam diluted with soup	6.27	113	123	91.8
		Spaghetti salad	7.14	73	78	93.5
		Natto with raw eggs	7.78	45	191	23.56
Fontaine et al., 1978	Heidelberg	Cheese	2.3	339	1211	28
			2	1	1	100

Kasuga et al. 2004	Oranienburg	Grated yam diluted with soup	9.87	11	11	100
George, 1976	Schwarzengrund	Pancreatic extract	1.64	1	1	100
Kasuga et al., 2004	Typhimurium	Grated yam diluted with soup	5.14	40	99	40.4
		Grated yam diluted with soup	6.38	39	79	49.37

4.4.2. Machine learning-based identification of features informative to a *Salmonella* dose-response

Of the 31,030 p(gene)-dose interaction terms from our *Salmonella* pan-genome, the Elastic Net model dropped 25945 variables due to collinearity. The α value and λ penalty for the model comprising the remaining 5085 covariates were selected by 10-fold cross validation. Cross-validation helps in choosing the model that minimizes the cross-validation function (**Figure 4.1**). The overall best-fit Elastic Net model, with an α value of 0.300 and λ penalty of 28.1061 identified 33 non-zero *p(gene expression)*-dose interaction terms that were most informative to the model (**Table 4.2**). The functionality of these genes ranged from adhesion and invasion to stress response and bacterial metabolism, and also included eleven hypothetical proteins.

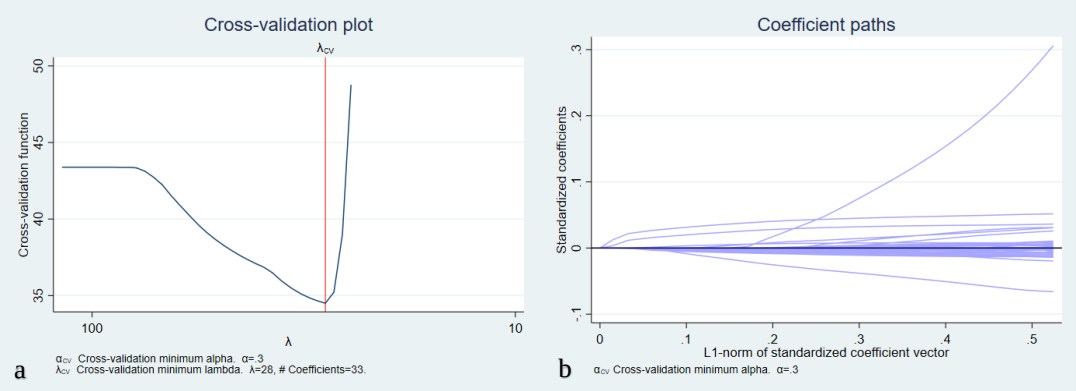

Figure 4.1. Elastic Net (a.) Cross-validation plot and (b.) Coefficient path plot indicating the best-fit α value and λ penalty that minimizes the cross validation function.

Table 4.2. Important genes associated with the gene-dose interaction terms identified by the best-fit Elastic Net model.

	Gene name	**RefSeq ID**
1	Zinc-binding GTPase YeiR	NP_416678.1
2	ABC transporter involved in cytochrome c biogenesis CcmB subunit	NZ_CP009516.1
3	FIG01046564 hypothetical protein	NA
4	SSU ribosomal protein S20p	CP007542
5	Polysaccharide export lipoprotein Wza	WP_014907221.1
6	Hypothetical protein	NA
7	Biofilm regulator BssS	NP_415578.3
8	NADH dehydrogenase (EC 1.6.99.3)	NP_460181.1
9	Putative oxidoreductase YdjL	NP_416290.1
10	Hypothetical zinc-type alcohol dehydrogenase-like protein YdjJ	NP_416288.1
11	Putative aldolase YdjI	NP_416287.1
12	Uncharacterized transcriptional regulator (DeoR family) YdfJ	WP_000347482.1
13	Secretion system chaperone SscB	NA
14	Hypothetical protein	NA
15	Secreted effector protein SteA	NP_460542.1
16	FIG01048335 hypothetical protein	NA
17	Putative invasin	NA

18	FIG01045615 hypothetical protein	NA
19	Putative membrane protein	NA
20	Cell division-associated ATP-dependent zinc metalloprotease FtsH	NC_011916.1
21	tRNA (guanine(46)-N(7))-methyltransferase (EC 2.1.1.33) trmB	NC_000913.3
22	Murein hydrolase activator NlpD	NP_417222.1
23	Hypothetical protein	NA
24	Cyclic di-GMP-binding protein BcsB	NP_312439.1
25	Phosphoglycerate transport regulatory protein PgtC	NP_461339.1
26	Hypothetical protein	NA
27	FIG01048353: hypothetical protein	NA
28	Hypothetical protein	NA
29	Hypothetical protein	NA
30	Deoxyribose-phosphate aldolase (EC4.1.2.4)	NP_418798.1
31	Putative periplasmic protein	NA
32	LysR-family transcriptional regulator STM3834	WP_000687412.1
33	Hypothetical protein	NA

4.4.3. Elastic Net-based Poisson regression model outcome

The Poisson regression dose-response model was fit on the gene expression probability-weighted log10 CFU doses. The Poisson model was selected primarily because our outcome variable (probability of illness given exposure) is a numeric count with a limited range compared to a continuous variable (Chesaniuk, 2021). The Elastic Net-based Poisson regression model identified 9 gene-dose interaction terms that significantly impacted the probability of illness ($p < 0.05$) when exposed to *Salmonella enterica* (**Table 4.2**). The model containing these 9 predictors showed a significant improvement and fit over the null model (Likelihood ratio chi-square statistic = 869.62; McFadden's $R^2 = 0.423$; probability > chi-square = 0.0000). The weighted genes varied in functionality from bacterial virulence, metabolism, and stress response.

The regression coefficients from this model can be interpreted as the predicted change in the log count of the response variable for every one unit increase in the predictor variable, i.e., p(gene)-dose interaction (controlling for the remaining predictors). A simple means of explaining the results of such a model would be that, a positive coefficient indicates an increase in the predicted value of the response variable (probability of illness) with an increase in value of the predictor variable, whereas a negative coefficient implies a decrease in the predicted response variable with an increase in the value of the predictor variable. In general, we found that genes coding for bacterial metabolism had the greatest impact on the probability of illness given exposure to *Salmonella enterica* (**Table 4.3, 4.4; Figure 4.2)**.

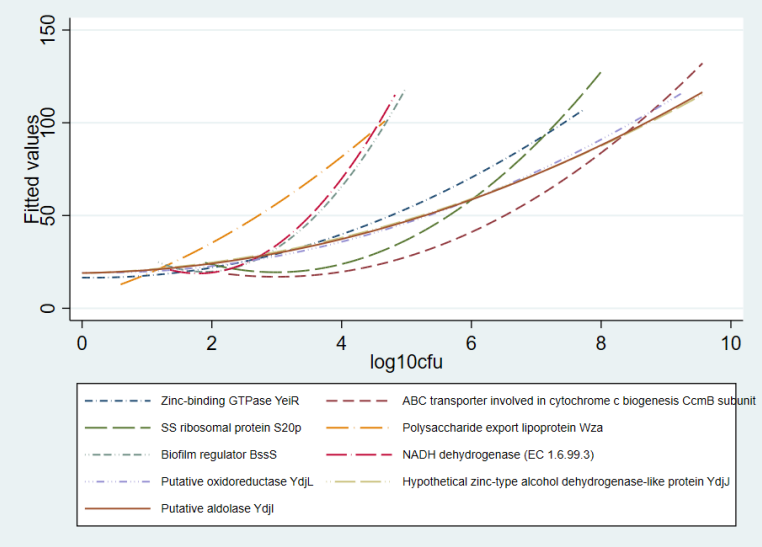

Figure 4.2. Predicted plot of the impact of the predicted values of significant *p(gene)-dose* interaction terms (predictor variables) on the probability of illness given exposure (response variable) to *Salmonella enterica.*

Finally, the subsample analysis using human feeding trial data or outbreak data alone yielded insignificant results, with the best-fit model (α value = 0.800; γ-penalty = 27.9430) identified by 10-fold cross validation selecting zero significant covariates. Therefore, for the purposes of this study, we have retained the model fit and parameters identified by the Elastic Net-based Poisson regression model developed using the pooled dataset.

Table 4.3. Significant predictor terms identified by the final Poisson model.

Significant predictor terms	**Model coefficient**	**Standard error**	**P > |z|**
Zinc-binding GTPase YeiR	-14.4135	1.1398	0.000
ABC transporter involved in cytochrome c biogenesis CcmB subunit	-13.9821	1.0935	0.000
SSU ribosomal protein S20p	-26.4459	2.0598	0.000
Polysaccharide export lipoprotein Wza	-27.9688	2.2478	0.000
Biofilm regulator BssS	-97.2884	7.5971	0.000
NADH dehydrogenase (EC 1.6.99.3)	178.1859	13.9816	0.000
Putative oxidoreductase YdjL	-235.0121	18.2942	0.000
Hypothetical zinc-type alcohol dehydrogenase-like protein YdjJ	53.3194	4.0035	0.000
Putative aldolase YdjI	203.2633	16.0184	0.000

4.5. Discussion

Traditionally, the hazard characterization step of QMRA has relied on strains of pathogens from selected cases to define and characterize hazards. However, a major issue with this approach has been the assumption that the pathogen is a single unit, thereby neglecting intra-species variation in pathogen virulence and virulence-associated functions. However, the recent advances and widespread application of sequencing technologies provides us with an unprecedented opportunity to attempt to account for these variations in predictive modeling and risk assessment approaches to potentially reduce the uncertainty and variability in the models. This is because the larger molecular data set can offer the opportunity for increased insight and better decision-making than that which can be accomplished by analyzing smaller data sets. However, the use of WGS in predictive models comprising a microbial QMRA is largely untapped and faces valid challenges related to disaggregation since the number of hazards to be considered increases exponentially when zooming in from a previous generalized species or serovars into specific genotypes (Pielaat et al., 2015).

In classical dose-response modeling of pathogens such as *Salmonella*, one must consider the high variability of the pathogen at low doses, and the potential for infection from survivors of the innate host defense systems. Infection and subsequent illness from a pathogen occurs from the proportion of ingested microorganisms that survive the human host barriers; for example, surviving the nutrient-starvation conditions typically seen in the host gastrointestinal system, adhering to, and transcending, the intestine, biofilm

formation, and virulence signaling. While current dose-response models function under the assumption that each ingested microorganism is a taxonomic unit with equal probability of inducing illness, the heterogeneity rendered as a result of horizontal gene transmission between the various taxonomic units is completely excluded or discounted (Njage, Leekitcharoenphon, & Hald, 2019).

However, the application of WGS data in a dose-response modeling framework is hindered by a number of issues. These include (i) the lack of a standardized dose-response dataset from which to make genetic inferences (e.g. separate dose-response profiles per serovar, differences in data collection), (ii) the lack of molecular data specifically associated with illness response to a given pathogen dose, necessitating the development of association models using comparable genome sequences, (iii) unavailability of sufficient molecular data from a single host (e.g. humans) to fit the requirements of a power analysis, necessitating the use of genome sequences from a myriad of hosts, and (iv) the large number of genes from such a compilation of isolates compared to the number of isolates themselves ($p >> n$). Therefore, the method proposed in this study does away with the concept of pathogenicity of a specific subgroup or serovar of a pathogen, as well as potential heterogeneity introduced due to gene transmission and differences in the pathogenic host, by directly taking into account the expression probabilities of genes identified as significant by a learning model. This is principally similar to the inverse-variance weighted methods proposed earlier for summarized data from multiple genetic variants (Ritchie et al., 2006; Burgess & Bowden, 2015; Reifeis, Hudgens, Civelek,

Mohlke, Love, 2020). Herein, the probability of expression of each genetic unit is computed and input as weights in a predictive model to estimate the dose-based response, taken here as the probability of illness (salmonellosis) as a result of exposure to the pathogen for simplicity. This all-in regression approach, similar to the whole genome random regression methods proposed for genome-wide association studies (Janss, los Campos, Sheehan, & Sorensen, 2012), towards simultaneously fitting all markers from the sample set was done to account for unaccounted population stratification. This would contribute to redefining dose-response relationships for initial infection from the relative proportions of each significant gene from a WGS dataset, which would help in refining endpoint and risk estimations.

A major issue contributing to substantial uncertainty in dose-response models is the type of data used to generate and assess the models. Dose-response relationships for pathogenic microorganisms are generally developed on data obtained from foodborne outbreaks, human trial studies, or experimental model studies. Each of these approaches have some limitations due to the inherent variability of the pathogen, host, and food source (Buchanan, Havelaar, Smith, Whiting & Julien, 2009). Moreover, the models that fit such varied types of data are different. For example, the generally agreed-upon model used in the development of dose-response relationships based on outbreak data for salmonellosis is the beta-Poisson model, with an emphasis on the difference in infection and illness. Alternatively, human feeding trial data generally fit more linear models, as evidenced in Oscar (2004). This distinction was not made in our study, since a majority of the dose-

response data points included in our study were from human feeding trials, as opposed to outbreak data, primarily due to lack of information. Moreover, we have only considered outbreak data where the pool of potentially exposed subjects have been clearly reported. Since the probability of illness (response variable) data is treated as a continuous variable between 0 and 1, but with a limited range, Poisson regression perfectly fits the requirements for this model. Hence, we can interpret our model outcome as the predicted change in the log count of the probability of illness given exposure for every one unit increase in the gene-dose interaction.

An important consideration of utilizing WGS data in any modeling studies is to work around the $p >> n$ problem described in Section 3.5. In our study, we work around this issue using Elastic Net feature selection to identify features significant to the prediction. The α value = 0.300 and λ penalty = 28.1061 were chosen by a 10-fold cross validation to minimize the bias-variance trade-off. The bias measures the accuracy of the estimates, by describing the difference between the true population parameter and the expected estimator, and the variance measures the uncertainty of the estimates. While traditional statistical modeling strategies involve the use of information criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), to determine the penalty terms to obtain a best-fit model, cross-validation is a more popular machine learning-based approach to obtain the best fit model (Oleszak, 2019). It is important to note that the McFadden's pseudo R^2 value was acceptable (0.42), despite eliminating 99.9% of the genes from the initial gene dictionary, which indicates that the

handful of genes (n = 33) selected for model building could be strongly indicative of the response metric, and gives us confidence in the elimination strategy employed by the Elastic Net model. However, it is important to note that Elastic Net is a predictor variable selection tool, and therefore does not give inferential results, necessitating further modeling (in our case, the weighted regression method) to identify significant predictor variables.

A number of genes identified by the Elastic Net-based Poisson regression model as being significant ($p < 0.05$) have been implicated in bacterial functions ranging from bacterial metabolism to virulence (**Table 4.4**). Interestingly, three of the genes (*YdjL, YdjJ,* and *YdjI*) identified as significant currently coding for hypothetical proteins or putative proteins. This demonstrates the need for further analyses to determine the functionality and effect of these proteins to the overall virulence and pathogenicity of *Salmonella*. Moreover, the large number of genes coding for metabolic functions being found as significant by our model is telling, since bacterial metabolism has been previously found to be key towards determining bacterial persistence (Amato et al., 2014), as well as defining the host-pathogen interface (Passalacqua, Charbonneau, & O'Riordan, 2016).

Our study has a few limitations. The biological basis for microbial dose-response models is a function of interactions between the pathogen, host and the matrix. This complexity was not captured in our proposed weighted gene-based modeling approach due to a number of factors, including (i) the lack of standardization of available data regarding *Salmonella* dose-response, (ii) unavailability of *Salmonella* molecular data collected in the

context of dose-response determination, and (iii) non-standardization of *Salmonella* genome sequencing data. Additionally, the non-linear beta-Poisson model is recommended for dose-response of *Salmonella* (QMRAWiki, 2021). However, our model employs a more generalized Poisson regression approach because of the limitations of Elastic Net regression. Moreover, we chose an arbitrary number of isolates per serovar (n ~ 50) to ensure that even the lowest sampled serovar is equally represented. However, choosing only 50 isolates for a serovar whose incidence is very high may not capture the overall spread; therefore, it may be better in future studies to weigh the genes based on the global probability of outbreaks involving the serovar in question to identify the trends.

Table 4.4. Functionality and significance of genes identified as 'significant' by gene-weighted Poisson regression.

Significant gene	Associated protein	Function	Relationship to dose-response	References
YeiR	Zinc-binding GTPase	Zinc homeostasis in *Escherichia coli*	Mismetallation of bacterial proteins can erode their functionality, which in turn has broad implications in bacterial and host metabolism during infection	(Blaby-Haas, Flood, Crecy-Lagard, & Zamble, 2012; Palmer & Skaar, 2016)
CcmB	ABC transporter involved in cytochrome c biogenesis CcmB subunit	Exports heme to periplasm for biogenesis of c-type cytochromes (transfers electrons between complexes III and IV of the respiratory chain), and also shows transmembrane transporter activity	Metabolism and mitochondrial function	(Stevens et al., 2011; Hough, Silkstone, Worrall, & Wilson, 2014)
SSU S20p	SSU ribosomal protein S20p	Structural constituents of ribosome	Reduction in mRNA binding, which in turn impacts rate of protein synthesis	(Tobin, Mandava, Ehrenberg, Andersson, & Sanyal, 2010)
Wza	Polysaccharide export lipoprotein Wza	Part of Wzy-dependent pathway responsible for assembly and export of capsular	Main virulence factors of bacterial cell	Cuthbertson, Mainprize, Naismith, & Whitfiled, 2009; Morais, Dee, & Suarez, 2018

		polysaccharides, constituting the outermost layer of the bacterial cell		
BssS	Biofilm regulator protein BssS	Cellular maintenance	Global regulator of genes involved in catabolite repression and stress response	Domka, Lee, & Wood, 2006
ndh	NADH dehydrogenase	-	Respiratory metabolism	Heikal et al., 2014
YdjL	Putative oxidoreductase	Catalyzes certain elements of glycolysis pathway	-	Huddleston et al., 2019
YdjJ	Putative zinc-type alcohol dehydrogenase-like protein	Metabolism	Bacterial physiology and pathogenicity	
YdjI,	Aldolase of unknown specificity	Metabolism	Bacterial physiology and pathogenicity	

4.6. Conclusion

While current microbial dose-response models are effective in providing us with the probability estimate for a response such as illness, illness given exposure, and illness given infection, they are unable to sufficiently capture heterogeneity arising due to genetic differences, which can be highlighted by whole genome sequencing data. Current models are unable to handle the large amount of data introduced by such data, requiring the use of advanced machine learning methods. Here, we propose a machine learning-supported weighted regression method to model the dose-augmented effect of significant genes on the illness response due to exposure to *Salmonella enterica*. Our method is unique in that it attempts to transcend inter-serovar and inter-host environment genetic heterogeneity using a weighted approach. This method could redefine dose-response relationships for initial infection from relative proportions of significant genes from a microbial genetic dataset, which would help in refining endpoint and risk estimations in the future.

Chapter 5: Predicting foodborne salmonellosis outbreak severity based on genetic and meteorological trends

5.1. Abstract

Several studies have shown a correlation between outbreaks of *Salmonella enterica* and climatological and meteorological trends, especially related to temperature and precipitation. Additionally, current outbreak-related studies are performed on data pooled by *Salmonella* species without taking into account its intra-species and genetic heterogeneity. In this study, we analyzed the effect of differential gene expression and a suite of meteorological factors on salmonellosis outbreak severity (typified by case numbers) using a combination of machine learning and data analytical methods. Elastic Net regularization was used to identify significant genes from a *Salmonella* pan-genome, and a multi-variable Poisson regression developed to fit the individual and mixed effects data. The best-fit Elastic Net model ($\alpha = 0.5000$; $\lambda = 2.18399$) identified 53 significant gene expression variables. The final multi-variable Poisson regression model (χ^2 = 5748.22; psuedo R^2 = 0.6688; probability > χ^2 = 0.0000) identified 127 significant predictor terms ($p < 0.10$), comprising 45 gene-only predictors, average temperature, average precipitation, and snow cover, and 79 gene-meteorological interaction terms. The significant genes ranged in functionality from bacterial metabolism, cell survival, ion transport, to stress response and bacterial virulence, and included gene variables not considered as important or significant by the baseline model. The results of this study indicate the need to co-

evaluate novel data with environmental data to develop a more holistic model to predict disease outcome severity, and thereby calculate the risk to human health.

5.2. Introduction

Despite ongoing efforts to curb the spread and proliferation of *Salmonella enterica*, its ubiquitous nature and considerable subtype diversity has contributed to a significant increase in the number of salmonellosis cases being reported both in the U.S. and globally (Scallan et al., 2011; CDC, 2021a). Moreover, *Salmonella* covers a diverse genetic landscape, with *Salmonella enterica* subsp. *enterica* (which can infect humans) alone comprising >2500 named serovars. Currently, models predicting bacterial outcome and severity do not account for intra-species variability in microbial behavior because, for the most part, the variabilities existing at the gene-level are too large in scale to be incorporated in basic statistical models. Genetic analyses of isolates could provide us with information regarding the expression of genes associated with stress tolerance, virulence, and antibiotic resistance. This could help in developing a differential virulence profile to aid in re-evaluating the existing infectivity and outbreak predictive estimates for *Salmonella enterica*.

Several studies have investigated the impact of environmental factors, specifically temperature, and precipitation, on the incidence of *Salmonella*-associated foodborne outbreaks. The impact of environmental factors on the genetic profiles of *Salmonella* plays a particularly important role in its pathogenicity; conditions unfavorable to pathogen

growth could induce a variety of survival mechanisms in the cells, which could impact the overall rate of *Salmonella* infection, modulating outbreak and illness risk estimates. Studies have shown how varied combinations of ambient temperatures, precipitation levels, and the resultant changes in food habits contributes to salmonellosis numbers in the population (Munnoch et al., 2009; Sidhu et al., 2013; Mun, 2020). This is particularly the case with higher temperatures and *Salmonella* proliferation and notifications of salmonellosis (McMichael, 2015). In general, studies have reported that the risk of *Salmonella* contamination, and subsequently, infection, increases under higher ambient temperatures, as it supports the growth of *Salmonella*. Similarly, increasing precipitation levels are also believed to increase the risk of salmonellosis incidence, as run off can increase pathogen loads in water sources. Therefore, it is important to take the impact of these variations into account when estimating the overall human risk due to *Salmonella*.

Recent studies have shown the applicability of novel approaches to re-quantify the risk of disease and outbreaks based on differences in gene expression. Chief among them is the application of novel modeling or machine learning to predict the severity or endpoint of diseases caused by pathogenic agents such as *Listeria* (Njage, Leekitcharoenphon, & Hald, 2019), *Escherichia coli* (Pielaat et al., 2015; Njage et al., 2019), and *Salmonella*. However, one of the most important contributions of this new wave of the use of advanced data analytical methods in bacterial predictive modeling is the incorporation of feature selection algorithms to reduce whole genome sequencing data into a format that can be

employed in predictive models without resulting in model overfitting or introducing bias to the same.

The objective of this project was to develop a machine learning-based regression approach was developed to attempt to quantify the interaction effects between meteorological factors (such as temperature and precipitation) and the probability of expression of various significant genes in *Salmonella enterica.* This, in turn, would help us predict the most significant combination of genes and meteorological factors that contribute to the incidence and outbreak of food-associated salmonellosis.

5.3. Material and Methods

5.3.1. Data collection

5.3.1.1. Salmonella *outbreak data*

Data regarding foodborne outbreaks of *Salmonella* was obtained from the U.S. Centers for Disease Control and Prevention's (CDC) National Outbreak Reporting System (NORS) database, which receives such data from the CDC's Foodborne Disease Outbreak Surveillance System (FDOSS). For the purpose of this study, we have only included data on outbreaks of *Salmonella* definitively associated with a food source (i.e., foodborne salmonellosis). The NORS toolkit contains a comprehensive list of outbreaks attributed to different etiological agents occurring between 1998 and 2017, and includes metadata such as the month and year of the outbreak, food source, and resultant number of illnesses,

hospitalizations, and deaths. We employed inclusion criteria, such as the availability of serovar and state data to identify relevant and complete information.

5.3.1.2. Meteorological data

Meteorological data was obtained from the National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Information (NCEI; previously the National Climactic Data Center) database (https://www.ncdc.noaa.gov/cdo-web/). Collected data included monthly climatological measures of temperature, precipitation, and snow-related statistics from the suite of climatological statistics collectively referred to as "U.S. Global Summary of the Month," measured at stations operated by the NOAA (Heim 1996; Owen & Whitehurst 2002; Arguez et al. 2012; Durre et al., 2013). In this study, data was obtained in the form of monthly averages. Specifically, we obtained the monthly average mean daily temperature, monthly average precipitation, and monthly average snowfall from all weather stations within each state of interest from NCEI. In the NCEI website (Index of /data/global-summary-of-the-month/access (noaa.gov)), temperature data is reported in °F and precipitation data is reported in inches, and the same metric is utilized in our models.

NOAA measures temperatures with the aid of numerous weather stations spread across the U.S. Incorporating data from these stations allows us to incorporate the variation in weather conditions seen across the state, specifically those with a significantly larger land mass. However, since we are taking average values across states, we standardize the

included meteorological data based on the state-specific mean and standard deviation for each month and year of interest.

5.3.1.3. Salmonella *isolates for development of pan-genome*

Salmonella isolates obtained by U.S. regulatory agencies during routine surveillance corresponding to salmonellosis outbreak occurrence were sampled from the National Center for Biotechnology Information's (NCBI) Pathogen Detection database. We applied the following inclusion criteria for isolate selection using available metadata: 'serovar' and 'state' corresponding to an outbreak, 'availability of short reads data,' and 'month and year corresponding to an outbreak' or 'month and year up to two months before an outbreak.' The latter criteria was included to account for lag time between infection in animals (or contamination of food) and actual consumption. Based on these inclusion criteria, 541 isolates spread across serovars Dublin, Enteritidis, Heidelberg, Infantis, Javiana, Montevideo, Munchen, Muenster, Newport, Reading, Saintpaul, Senftenberg, and Typhimurium were used to create our pan genome (gene dictionary).

5.3.2. WGS pre-processing and creation of pan genome

Sequence Read Archive (SRA) Run Accession numbers for all included isolates were obtained from the NCBI SRA repository. The isolates were *de novo* assembled and annotated on the PATRIC (v.3.6.3) Bacterial Bioinformatics Resource Center, a web-based platform for genomic analyses, as previously described in Section 3.3.3. The WGS of isolates that fit our quality parameters (n = 497) were used in creating an environmental

Salmonella enterica gene dictionary (pan genome) using *pairwise2* in Python, as described in Section 3.3.3. Genes annotated as coding for 'hypothetical proteins,' 'hypothetical *xyz*,' 'putative *xyz*,' CRISPR repeats, and CRISPR spacers (and their homologs) were removed for ease of use (similar to Roary), despite potentially contributing to the virulence and pathogenicity potential of *Salmonella* (Louwen, Staals, Endtz, van Baarlen, & van der Oost, 2014). This generated a dictionary/pan-genome of 18,520 unique genes.

5.3.3. Model development and statistical analysis

Here, we modeled the individual and combined effects of the predictor variables gene expression (categorical; 1 or 0), mean daily average temperature (in °F) (continuous), precipitation (in inches), and snow attributes (in inches) on the response variable (number of illnesses, or case numbers). All models were run with standardized meteorological variables recorded during the month of an outbreak (no lag), and two months before an outbreak (two-month lag) to determine the effect of sustained weather factors on illness outcome. All statistical analyses and modeling were performed on STATA 16 (StataCorp, 2019). Model significance was set at $\alpha = 0.05$, and predictor significance was tested at both $\alpha = 0.10$ and $\alpha = 0.05$.

5.3.3.1. Feature selection

Since our gene predictor matrix is very large (n = 18,520), it could result in dimensionality issues and model overfitting. Additionally, a gene-based dataset is bound to comprise a large number of variables that are highly correlated. Thus, we combat this

using the Elastic Net feature selection method, as described in Sections 3.3.4.4. and 4.3.3. Elastic Net regularization was performed on STATA using the *elasticnet* function. The penalized objective function for Elastic Net is provided in equation 4.1. The functional form for the function *f()* from equation 4.1. used in a linear (ordinary least squares) and Poisson (or another count model, such as a negative binomial) model are provided in equations (5.1) and (5.2), respectively.

$$f(y_i, \beta_0 + x_i\beta') = \frac{1}{2}(y_i - \beta_0 - x_i\beta')^2 \quad \text{.....(5.1)}$$

$$f(y_i, \beta_0 + x_i\beta) = -y_i(\beta_0 + x_i\beta') + e^{(\beta_0 + x_i\beta')} \quad \text{.....(5.2)}$$

In this study, we tested the default α values (1, 0.75, and 0.5) and a fine grid of λ values, according to Hastie, Tibshirani & Wainwright (2015). The λ grid is set automatically during the run. The (α, λ) pair that minimized the value of the cross validation function during 10-fold cross validation was selected, and the significant non-zero coefficients identified by this (α, λ) pair were employed in further models as independent predictor variables (Stata, 2021).

5.3.3.2. Poisson regression

A Poisson regression model was developed to explain the outcome of case numbers (count data; response variable), with gene presence/absence data functioning as the primary predictor, and meteorological factors as the covariates. Simply put, our model is structured as:

$$\Pr(Y_i = y_i|\mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!} \quad \text{.....(5.3)}$$

Where,

$$\mu_i = t_i e^{(\beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \ldots + \beta_k X_{ki})} \quad \ldots\ldots(5.4)$$

Where, the response variable denotes case numbers over the included time period, i the outbreak observation included in the model, and X_i denotes a vector of independent variables – significant genes identified by Elastic Net, monthly average mean daily temperature, monthly average precipitation, and monthly average snowfall – and their interaction terms, $\beta = 1 \ldots k$ indicates the regression coefficients, and μ the risk of a new occurrence of the event during a specified exposure event t (if no exposure is given, t is assumed to be 1). Hence, the model outcome can be interpreted as, for a one unit change in the predictor variable, the difference in log of expected counts (response) is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant. While the total number of included genes from the initial dictionary is very large, the values for only those genes that are deemed significant by the Elastic Net model were included in the final regression model (all genes with zero values being automatically eliminated from the model).

5.3.3.3. Negative binomial regression

Since a Poisson regression makes a restrictive assumption that the mean is equal to the variance, we also fit the data to a negative binomial regression, which is a generalization of the former model that loosens this restrictive assumption, as shown in another study (Shirriff, 2019). The fundamental negative binomial regression equation is written as:

$$\Pr(Y_i = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i+\alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+1)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \quad \text{.....(5.5)}$$

Where μ_i, or the mean incidence rate per unit exposure *t* (if no exposure is given, *t* is assumed to be 1) is:

$$\mu_i = e^{(lnt_i+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots\beta_k X_{ki})} \quad \text{.....(5.6)}$$

Here $\beta = 1\ldots k$ denote the regression coefficients, $\alpha = 1/v$, where *v* denotes the scale parameter of the gamma (or negative binomial) noise parameter, and $X=1\ldots k$ indicates the matrix of predictor variables. As in the Poisson regression, important genes and meteorological factors recorded during the outbreak period were used as independent variables.

5.4. Results

Here, a machine learning-based method to identify important genes, meteorological factors (and their combinations) that impact salmonellosis outbreak numbers irrespective of *Salmonella enterica* serovar-level heterogeneity is developed using whole genome sequencing information. In order to achieve this, (i) whole genome sequences of *Salmonella enterica* serovars isolated from varied environmental sources, corresponding to human outbreaks of salmonellosis, were pre-processed to create a *Salmonella* pan genome, (ii) meteorological data corresponding to human outbreaks of salmonellosis were obtained and processed, (iii) important genes were identified using Elastic Net regularization, and (iv) significant genes were incorporated in count-based models, along

with meteorological factors as predictor variables to identify their individual or combined impact on illness numbers.

5.4.1. Outbreak and WGS data collection and preprocessing

Data from human outbreaks of *Salmonella enterica* was obtained from the NORS dashboard. Relevant outbreaks were selected for further analyses based on our inclusion criteria. Two hundred and eighty five outbreaks without serovar information and 338 multi-state occurrences were dropped, leaving us with 2844 outbreaks definitively attributed to different serovars of *Salmonella* that were included for further analyses. Subsequently, we matched the outbreaks to *Salmonella enterica* isolates obtained from food sources and the environment based on the date and time of the outbreak and matching serovar, in order to build our *Salmonella* pan genome. This provided us with 249 data points (outbreaks), with a salmonellosis case (illness) number of 7385 (**Figure 5.1**). Of these, the public health burden of a large number of these outbreaks was comparatively lower, with a majority of outbreaks having salmonellosis case (illness) numbers ≤60. As is common with most count-based datasets, our dataset has a large number of data points for a few values (i.e., case numbers ≤60), resulting in a skewed frequency distribution of data points (**Figure 5.2**).

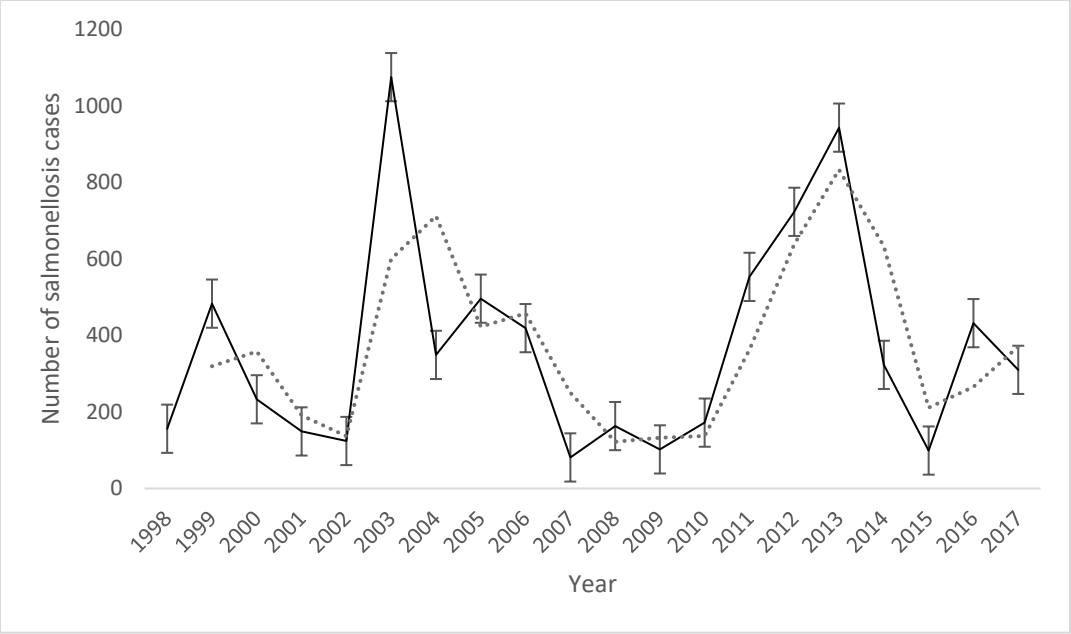

Figure 5.1. Data trends - yearly trend in foodborne salmonellosis case numbers (1998–2017) included in our study. Error bars indicate standard error and the trend line is calculated by simple two-period moving averages method.

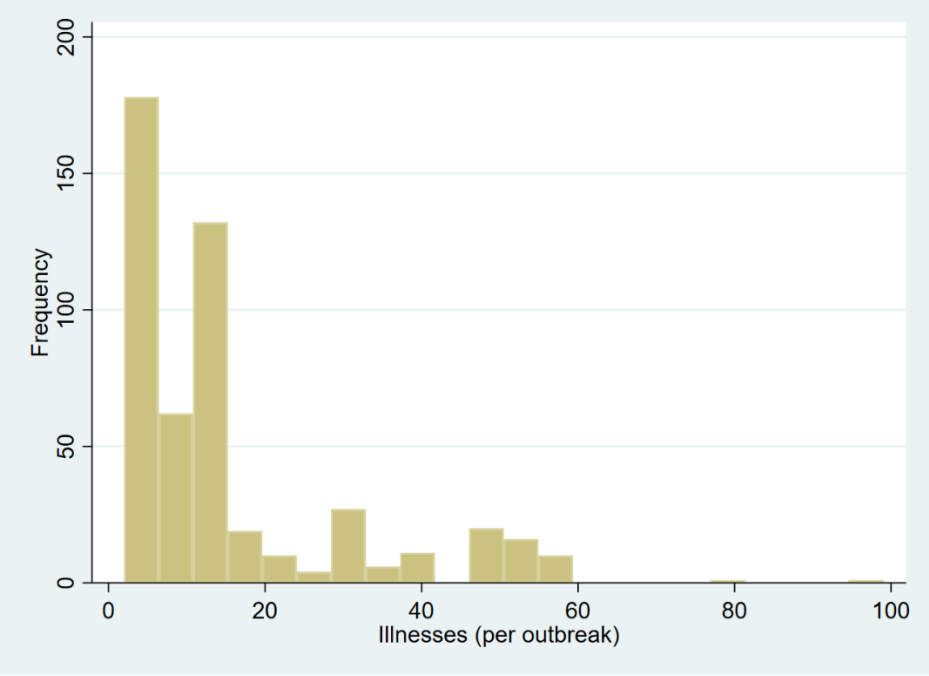

Figure 5.2. Histogram depicting trends in foodborne salmonellosis illness cases per outbreak included in our study. A majority of the outbreaks included in our study had a small number of overall reported case numbers ($n < 20$), with some outlier outbreaks with > 60 reported illnesses, representing a skewed distribution of discrete values that are generally handled using count models.

The *Salmonella* pan genome serves as the raw dataset to identify important gene predictor variables. Whole genome sequences across the included *Salmonella enterica* serovars (and matching the time frame of salmonellosis outbreaks) were sampled from the NCBI Pathogen Detection web server. Isolates were selected from across a number of human, animal, and environmental isolation sources to account for genetic recombination, and directionality and timing of evolutionary changes within and among serovars (Grad & Lipsitch, 2014). Short reads for each isolate were assembled and annotated on the PATRIC web server for homogeneity. The final *Salmonella* pan-genome, composed of 18,520 annotated genes, was created from the annotated sequences using settings similar to that employed by Roary (Page et al., 2015), with further restrictions (as described in section 2.2.2) to obtain interpretable model results.

A preliminary comparison of month-wise outbreak trends with the most impactful meteorological data (temperature and precipitation) revealed that the highest temperatures (summer months – June–August; **Figure 5.3.**) and months with highest rainfall (and

consequently, highest precipitation in spring and early fall; **Figure 5.4.**) were correlated with increased salmonellosis case numbers.

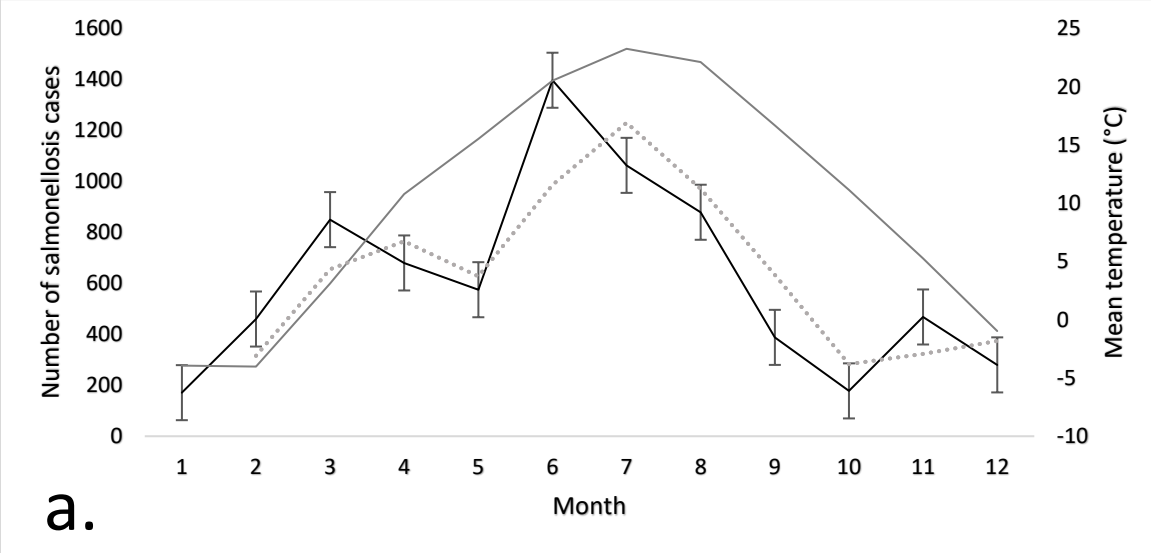

Figure 5.3. Monthly trend in salmonellosis cases (included in our study) and mean temperature. Solid black line indicates the monthly trend in salmonellosis cases, dotted black line indicates a simple 2-point moving averages trend line, error bars indicate standard error, and the grey line indicates the monthly average temperature.

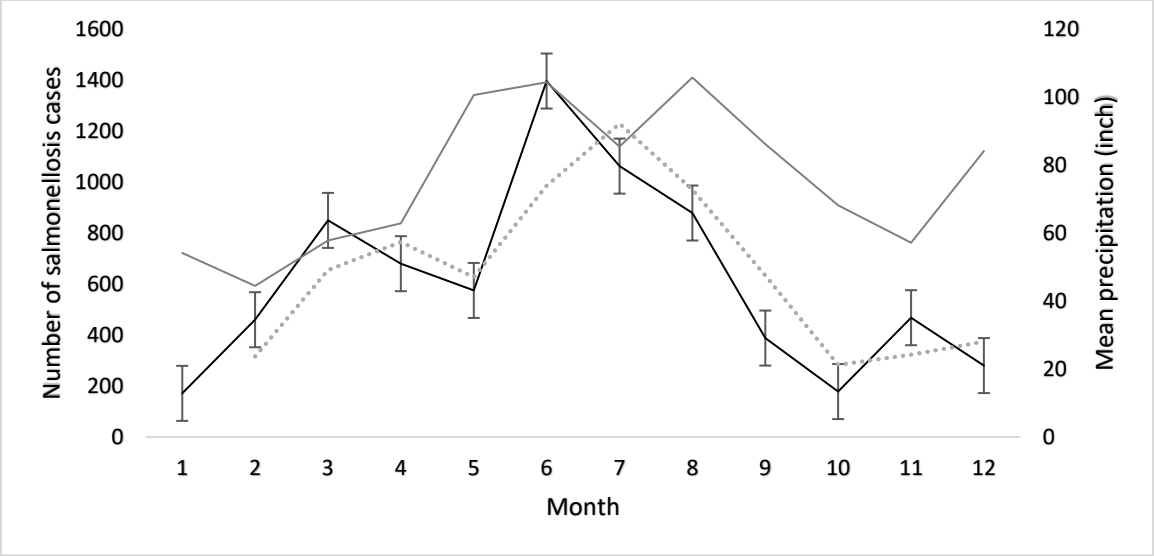

Figure 5.4. Monthly trend in salmonellosis cases (included in our study) and mean precipitation. Solid black line indicates the monthly trend in salmonellosis cases, dotted black line indicates a simple 2-point moving averages trend line, error bars indicate standard error, and the grey line indicates the monthly average precipitation.

5.4.2. Machine learning-based identification of genes informative to *Salmonella* outbreak prediction model

Of the 18,520 genes comprising our *Salmonella* pan-genome, the best-fit Elastic Net model (α value = 0.500 and λ penalty = 2.18399) selected by 10-fold cross validation, which helps in choosing the model that minimizes the cross-validation function (**Figure 5.5**), identified 53 distinct, non-zero genes that were most informative to the model (**Table 5.1**). The functionality of these genes ranged from adhesion and invasion to temperature stress response and bacterial metabolism. Of note, 13 of the selected significant genes coded for bacterial phage proteins.

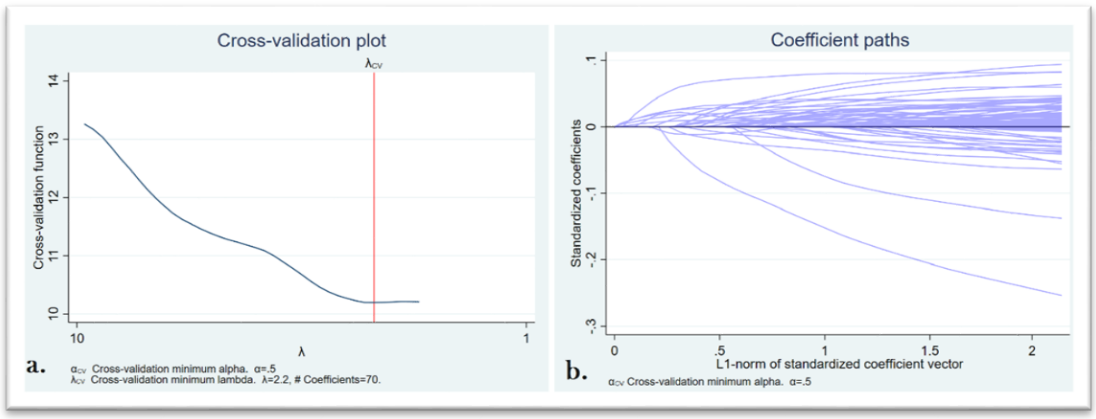

Figure 5.5. Best-fit Elastic Net cross-validation plot and coefficient path.

Table 5.1. Important genes identified by the best-fit Elastic Net model.

Gene name	Function	Reference
TnpA transposase	Sets threshold for activation of *Salmonella* pathogenicity island (SPI1) effector proteins, thereby impacting bacterial virulence.	Ellis, Trussler, Charles, & Maniford (2017)
Iron-sulfur cluster assembly protein SufD	Oxidative stress response through the formation and protection of iron-sulfur clusters.	Saini, Mapolelo, Chahal, Johnson, & Outten (2010)
Conjugative transfer protein 123	Spreading of mobile genetic elements (transposons, plasmids, etc.) among bacteria – indicative of horizontal gene transmission.	Zatyka & Thomas (1998)
Replication protein	Involved in biochemical pathway allowing for interaction between *Salmonella* and *E. coli*.	Maurer, Osmond, Shekhtman, Wong, & Botstein (1984)
Phototransferase system (PTS), D-glucosaminate-specific IIA component (EC 2.7.1.203)	Found only in bacteria, catalyzes transport and phosphorylation of many monosaccharides, disaccharides, and sugar derivatives such as amino sugars.	Deutscher, Francke, & Postma (2006); Zhi et al. (2020)
PTS – IIB component (EC 2.7.1.203)	Deletion of this gene could result in inefficient utilization of glucose and glycerol during invasion of host systems.	
Oxaloacetate decarboxylase Na(+) pump, alpha chain (EC 4.1.1.3)	Involved in bacterial growth and survival	Liu et al. (2013)
Beta chain (EC 4.1.1.3)		
Secreted effector protein SteA	*Salmonella* virulence – suppression of host inflammatory response.	Gulati, Shukla, & Mukhopadhyaya (2019)
Phage lysozyme R (EC 3.2.1.17)	Lytic transglycosidases	Holtje, Mirelman, Sharon, & Schwarz (1975)

Two-component transcriptional response regulator BtsR	Together with YpdA/YpdB, balances physiological state of cells within the population via nutrient uptake.	Vilhena et al. (2018)
Uncharacterized protein YciW	L-cysteine and L-methionine metabolism.	Kawane et al. (2014)
Phage tail fiber protein GpH	Potential horizontal transmission from *Yersinia pestis*, with putative receptive binding function.	Born, Braun, Scholz, & Grass (2020)
Large repetitive protein	Putative adhesin during initial cell surface interactions, highly specific to *Salmonella*.	Danckert, Hoppe, Bier, & von Nickisch-Rosenegk (2014)
Copper/silver efflux RND transporter, transmembrane protein CusA	Protein in the cus operon of the Copper Homeostasis and Silver Resistance Island (CHASRI) mobile element, which allows for protection under copper-mediated stress under aerobic, anaerobic, and facultative anaerobic conditions.	Haendiges, Brown, Tikekar, & Hoffmann,
Fructokinase (EC 2.7.1.4)	Potentially assists in *Salmonella* growth in presence of fructose and absence of phosphorylation.	Postma & Stock (1980)
Phage protein Ogr	Late gene expression function in SopEΦ phage that encodes virulence protein SopE.	Pelludat, Mirold, & Hardt (2003)
Phage tail tip, host specificity protein J	Potential presence of adhesion proteins that help in host recognition.	Dunne et al. (2018)
Cobalamin synthase (EC 2.7.8.26)	Required for synthesis of cobalamin, which in turn is used by *Salmonella* to uptake B12, required for anaerobiosis, especially during infected host systems.	Paiva, Penha Filho, Junior & Lemos (2011)
Aminoglycoside 3''-nucleotidyltransferase, AadA family (EC 2.7.7.-)	Involved in streptomycin stress response	Singh, Drolia, Bai, & Bhunia (2015)
RNA polymerase sigma factor RpoS	Stress sigma factor, required for *Salmonella* survival under starvation and stress conditions.	Nickerson & Curtiss (1997)

IncF plasmid conjugative transfer protein TraD	Pilus extension, pilus retraction, formation of stable mating pairs, and formation of lumen through which plasmid transfer occurs.	Frost, Ippen-Ihler, & Skurray (1994)
TolA protein	Virulence, membrane integrity, lipopolysaccharide production, and bile and serum resistance.	Paterson et al. (2009)
***Salmonella enterica* serovar Choleraesuis 50k virulence plasmid DNA**	N/A	-
Mobile element protein	N/A	-
SbmA protein	Inner membrane protein in Gram-negative bacteria involved in internalization of glycopeptides and prokaryotic and eukaryotic antimicrobial peptides.	Runti et al. (2013)
Phage tail protein GpU	Potential Mg(II)-mediated oligomerization and biological function.	Edmonds et al. (2007)
Phage integrase	Helps delineate *Salmonella* diversity; Mediate unidirectional site-specific recombination between two DNA recognition sequences.	Groth & Calos (2004); Colavecchio et al. (2017)
Rep protein	N/A	
Phage baseplate assembly protein GpV	Phage spike protein.	Buttner et al. (2016)
HTH-type transcriptional regulator MlrA	Regulatory protein that binds to large intergenic region upstream of the csgD promoter to modulate gene expression in response to changing environmental stimuli.	Gerstel, Park, & Römling (2003); Shen & Fang (2012)
L-carnitine/gamma-butyrobetaine antiporter	Exchanger for l-carnitine and γ-butyrobetaine.	Jung et al. (2002)
Type III secretion spans bacterial envelope protein (YscO)	Required for high-level expression and secretion of V antigen and Yops. Virulence type III export chaperone ortholog of InvI, with effector delivery (to host) function	Payne & Straley (1998); Evans & Hughes (2009)

Type III secretion and flagellar regulator RtsA	Belonging to *Salmonella* pathogenicity island 1 (SPI1), encoding a type III secretion system (T3SS) required for invasion of epithelial cells.	Ellermeier & Slauch (2003)
Uncharacterized J domain-containing protein YbeV	N/A	-
ADP-ribose pyrophosphatase of COG1058 family (EC 3.6.1.13) / Nicotinamide-nucleotide amidase paralog YfaY	N/A	-
Transcriptional regulator STM2275, GntR family	Transcriptional regulator that controls a variety of cellular processes such as cell motility, glucose metabolism, bacterial resistance, pathogenesis and virulence.	Li, Wang, Su & Lu (2021)
Uncharacterized major facilitator superfamily (MFS)-type transporter	Multidrug efflux pump, which can extrude compounds like metabolites, quorum-sensing molecules, and virulence factors, with a large spectrum of substrate specificities.	Pasqua et al. (2019)
Methionine ABC transporter permease protein	Biological transport	Interpro (2021)
Multidrug efflux system MdtABC-TolC, inner-membrane proton/drug antiporter MdtC (RND type)	Extrusion of substrates such as novobiocin, bile salts, quinolones, fosfomycin, detergents, zinc, and myricetin.	Anes, McCusker, Fanning, & Martins (2015)
Phage activator protein cII	Transcriptional activation.	Obuchowski et al. (1997)
DNA-damage-inducible protein I	Cold adaption.	Smith, Arany, Orrego & Eisenstadt (1997)
tRNA-(ms[2]io[6]A)-hydroxylase (MiaE)	Di-iron binding domain (MiaE)	-

Tn21 protein of unknown function Urf2	N/A	
ABC transporter, permease protein STM1634 (cluster 3, basic aa/glutamine/opines)	Indispensable for transport of solutes across biological membrane and ATP hydrolysis function.	Schneider & Hunke (1998)
Phage tail fiber, side tail fiber protein Stf	Host recognition and important effectors during the infection process.	Andres, Baxa, Hanke, Seckler, & Barbirz (2010)
Phage tail fiber assembly protein GpG		
Phage protein		
Resolvase	Nucleases involved in genetic recombination.	Massey, Boew, Sheehan, Dougan & Dorman (2000)
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA	Virulence plasmid.	Rychlik, Gregorova, & Hradecka (2006)
FIL protein	Filamentation.	Uniprot (2017)
Phage replication protein GpB	DNA replication and packaging.	Fane et al. (2006)
Phage replication protein GpA, endonuclease		

N/A – not available

5.4.3. Poisson regression model outcome

Poisson regression models were developed using a matrix of predictor variables comprising genes and meteorological data (and a combinations of these factors). The Poisson model was selected primarily because our outcome variable (illness case numbers) is a numeric count with a limited range compared to a continuous variable (Chesaniuk, 2021). Here, we can interpret the model coefficients as follows: for a one unit change in the predictor variable (x_1), the difference in log of expected case numbers changes by the corresponding regression coefficient (β_1). A simple means of explaining the results of such a model would be that, a positive coefficient indicates an increase in the predicted value of the response variable (salmonellosis illness/case numbers) with an increase in value of the predictor variable, whereas a negative coefficient implies a decrease in the predicted response variable with an increase in the value of the predictor variable.

The baseline Poisson regression model identified 28 *Salmonella* genes that were significant in predicting salmonellosis case numbers at $p < 0.05$ and 5 that were significant at $p < 0.10$ (**Table 5.2**). The model containing these 33 predictors showed a significant improvement and fit over the null model (Likelihood ratio χ^2 statistic = 4604.21; pseudo $R^2 = 0.5357$; probability > $\chi^2 = 0.0000$). The weighted genes varied in functionality from metabolism (antiporters, efflux pump-related, ion transport-related, transcriptional regulators), survival (e.g. replication protein), virulence (phage proteins), and stress response (iron sulfur cluster assembly protein, for example). Of note is that a majority of predictor variables that negatively impacted the outcome were associated with bacterial metabolism.

Table 5.2. Baseline Poisson regression model coefficients.

Predictor variable	Coefficient	Standard error	z	P>z	[95% confidence interval]	
Iron-sulfur cluster assembly protein SufD	28.01664	6.891953	4.07	0	14.47036	41.56293
Replication protein	27.246	7.898414	3.45	0.001	11.72149	42.77051
PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)	33.38321	6.588831	5.07	0	20.43271	46.3337
Oxaloacetate decarboxylase Na(+) pump alpha chain (EC 4.1.1.3)	-10.6636	4.13916	-2.58	0.01	-18.7992	-2.528
Secreted effector protein SteA	48.02776	5.872315	8.18	0	36.4856	59.56992
Phage lysozyme R (EC 3.2.1.17)	-10.9388	4.016991	-2.72	0.007	-18.8342	-3.04327
Two-component transcriptional response regulator BtsR	15.31843	4.756487	3.22	0.001	5.969454	24.66741
Uncharacterized protein YciW	25.77064	5.295008	4.87	0	15.36319	36.1781
Large repetitive protein	-5.58476	2.869412	-1.95	0.052*	-11.2247	0.05513
Copper/silver efflux RND transporter transmembrane protein CusA	4.673564	2.556099	1.83	0.068*	-0.3505	9.697633
PTS system ascorbate-specific IIC component	10.75809	3.805999	2.83	0.005	3.277311	18.23886
Phage tail tip host specificity protein J	8.521956	3.662542	2.33	0.02	1.323149	15.72076
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)	11.03298	4.310288	2.56	0.011	2.561012	19.50495
Phage tail fiber protein GpH	-13.5206	4.676838	-2.89	0.004	-22.713	-4.32816
IncF plasmid conjugative transfer protein TraP	18.70792	9.034205	2.07	0.039	0.95099	36.46485
TolA protein	10.50026	4.994796	2.1	0.036	0.682878	20.31764
SbmA protein	40.63866	8.477824	4.79	0	23.97531	57.30201
Rep protein	12.78836	6.803667	1.88	0.061*	-0.58439	26.16112

IncF plasmid conjugative transfer protein TraD	15.53458	6.889965	2.25	0.025	1.992205	29.07696
HTH-type transcriptional regulator MlrA	22.73903	5.480819	4.15	0	11.96636	33.5117
L-carnitine/gamma-butyrobetaine antiporter	29.93934	6.410017	4.67	0	17.34031	42.53837
Type III secretion spans bacterial envelope protein (YscO)	-3.41751	1.29988	-2.63	0.009	-5.97245	-0.86257
ADP-ribose pyrophosphatase of COG1058 family (EC 3.6.1.13)	-19.3482	6.497515	-2.98	0.003	-32.1192	-6.5772
Methionine ABC transporter permease protein	-15.5589	7.862062	-1.98	0.048	-31.0119	-0.10579
PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)	-4.09485	2.194291	-1.87	0.063*	-8.40777	0.218082
Multidrug efflux system MdtABC-TolC inner-membrane proton/drug antiporter MdtC (RND type)	-20.5966	9.031413	-2.28	0.023	-38.348	-2.84515
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)	2.031516	0.986939	2.06	0.04	0.091667	3.971366
inhibits host DNA replication	6.708811	2.776113	2.42	0.016	1.2523	12.16532
Phage tail fiber side tail fiber protein Stf	7.215132	4.295992	1.68	0.094*	-1.22874	15.659
Phage tail fiber assembly protein GpG	20.62922	8.136763	2.54	0.012	4.636236	36.62221
Phage replication protein GpB	29.05992	12.73747	2.28	0.023	4.024143	54.0957
Phage replication protein GpA endonuclease	-38.1249	13.51386	-2.82	0.005	-64.6867	-11.5631
Mobile element protein	9.907769	2.68062	3.7	0	4.63895	15.17659
Constant	72.14902	15.7775	4.57	0	41.13801	103.16

Data for variables that did not pass the threshold for significance, or which were automatically omitted by the model, not included herein.

P values indicated with * are significant at $\alpha = 90\%$. All other significant observations are significant at $\alpha = 95\%$.

The final Poisson regression model (no lag) with monthly average mean daily temperature, monthly average mean precipitation, and monthly average snowfall identified 125 predictor terms that were significant (at $p < 0.10$ (n = 8) or 0.05 (n = 117)) in predicting the outcome variable. These terms included 45 gene only predictors, each of the 3 meteorological predictor covariates, and 79 gene-meteorological interaction terms (**Table 5.3**). The model containing these 127 predictors showed a significant improvement and fit over the null and baseline models (Likelihood ratio χ^2 statistic = 5748.22; pseudo R^2 = 0.6688; probability > χ^2 = 0.0000). We observed that a number of gene predictors that were dropped by the baseline model as not significantly associated with outcome prediction were included in this model, indicating the significance of the joint impact of meteorological stressors and bacterial gene composition on the severity of outbreaks (as typified by case numbers).

Although the two-month lag model also showed a significant improvement in fit compared to the null and baseline models, it only identified 63 significant individual and mixed effect predictors, dropping important covariates like mean average daily temperature and mean average daily precipitation. Moreover, the coefficients (and their relationship to the outcome) of the remaining covariates more or less corresponded to the no-lag model (with the notable exception of *SteA*; data not included). Thus, the results of this model were dropped from further consideration.

Table 5.3. Poisson regression coefficients for multi-variable Poisson regression. Presence/absence of genes identified as significant by the Elastic Net model were input along with the standardized covariates monthly mean average daily temperature (tavg, in °F), monthly mean precipitation (prcp, in inches), and monthly mean snow cover (snow, in inches).

Significant predictor variables	**Coefficient**	**Standard error**	**z**	**P>z**	**[95% confidence interval]**	
Individual effect (gene or meteorological factor only)						
TnpA transposase	-4.24205	0.774781	-5.48	0	-5.76059	-2.7235
Iron-sulfur cluster assembly protein SufD	1.681351	0.281515	5.97	0	1.129592	2.23311
Conjugative transfer protein 123	2.26942	1.234303	1.84	0.066*	-0.14977	4.68861
PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)	0.472935	0.200257	2.36	0.018	0.08044	0.865431
Secreted effector protein SteA	1178.586	422.1747	2.79	0.005	351.1385	2006.033
Phage lysozyme R (EC 3.2.1.17)	-1.96364	0.330304	-5.94	0	-2.61103	-1.31626
Two-component transcriptional response regulator BtsR	1.401105	0.181993	7.7	0	1.044405	1.757805
Uncharacterized protein YciW	0.342621	0.169996	2.02	0.044	0.009435	0.675807
Fructokinase (EC 2.7.1.4)	-20.6446	5.038915	-4.1	0	-30.5207	-10.7685
Phage protein Ogr	1.076102	0.692009	1.56	0.12	-0.28021	2.432414
PTS system ascorbate-specific IIC component	1.45159	0.864899	1.68	0.093*	-0.24358	3.146761
Phage tail tip host specificity protein J	1.624397	0.658393	2.47	0.014	0.333971	2.914824
RNA polymerase sigma factor RpoS	1.493614	0.501178	2.98	0.003	0.511322	2.475905
IncF plasmid conjugative transfer protein TraP	0.741612	0.220333	3.37	0.001	0.309768	1.173456

TolA protein	2.054254	0.428836	4.79	0	1.213751	2.894757
SbmA protein	5.159942	0.737322	7	0	3.714817	6.605067
Phage tail protein GpU	9.279699	2.714573	3.42	0.001	3.959234	14.60016
Phage tail fiber protein GpH	140.9729	18.1362	7.77	0	105.4266	176.5192
Rep protein	-7.31411	1.011574	-7.23	0	-9.29676	-5.33146
IncF plasmid conjugative transfer protein TraD	1.908935	0.637275	3	0.003	0.6599	3.15797
Phage baseplate assembly protein GpV	-8.76387	2.61576	-3.35	0.001	-13.8907	-3.63707
HTH-type transcriptional regulator MlrA	1.343493	0.383769	3.5	0	0.591319	2.095667
L-carnitine/gamma-butyrobetaine antiporter	25.93318	15.80572	1.64	0.101*	-5.04548	56.91183
Type III secretion spans bacterial envelope protein (YscO)	-0.74477	0.202041	-3.69	0	-1.14077	-0.34878
Type III secretion and flagellar regulator RtsA	-0.87805	0.497011	-1.77	0.077*	-1.85218	0.096069
Uncharacterized J domain-containing protein YbeV	268.0571	95.46974	2.81	0.005	80.93987	455.1744
ADP-ribose pyrophosphatase of COG1058 family (EC 3.6.1.13)	-35.2471	5.016747	-7.03	0	-45.0798	-25.4145
Transcriptional regulator STM2275 GntR family	4.146038	1.159361	3.58	0	1.873731	6.418344
Aminoglycoside 3"-nucleotidyltransferase (EC 2.7.7.-) => ANT(3")-Ia (AadA family); ANT(9)-I	136.1866	49.16544	2.77	0.006	39.82414	232.5491
Multidrug efflux system MdtABC-TolC inner-membrane proton/drug antiporter MdtC (RND type)	-1.44521	0.230433	-6.27	0	-1.89685	-0.99357
Phage integrase	0.474894	0.179422	2.65	0.008	0.123233	0.826554
Phage activator protein cII	1.050271	0.333611	3.15	0.002	0.396407	1.704136
DNA-damage-inducible protein I	-0.71886	0.292764	-2.46	0.014	-1.29267	-0.14505
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)	-0.96144	0.178365	-5.39	0	-1.31103	-0.61185
Phage tail fiber side tail fiber protein Stf	14.19948	1.549145	9.17	0	11.16321	17.23575

Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)	-0.54206	0.287737	-1.88	0.06	-1.10602	0.021891
Phage activator protein cII	-1.50948	0.486527	-3.1	0.002	-2.46306	-0.5559
Phage protein	1.440505	0.450688	3.2	0.001	0.557172	2.323837
Resolvase	1.599016	0.802224	1.99	0.046	0.026687	3.171346
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA	9.330222	0.949352	9.83	0	7.469527	11.19092
Oxaloacetate decarboxylase Na(+) pump alpha chain (EC 4.1.1.3)	1.930903	0.590861	3.27	0.001	0.772838	3.088968
FIL protein	-8.39159	3.500809	-2.4	0.017	-15.2531	-1.53013
Phage replication protein GpB	-114.636	46.17484	-2.48	0.013	-205.137	-24.1348
Phage replication protein GpA endonuclease	122.9997	45.98945	2.67	0.007	32.86203	213.1374
Mobile element protein	-1.15607	0.523836	-2.21	0.027	-2.18277	-0.12937
Tavg	-14.5395	4.47902	-3.25	0.001	-23.3182	-5.76077
Prcp	4.655756	1.668996	2.79	0.005	1.384584	7.926927
Snow	0.124827	0.035906	3.48	0.001	0.054452	0.195202

Gene-meteorological factor interaction effects

Iron-sulfur cluster assembly protein SufD*tavg	-0.05214	0.020777	-2.51	0.012	-0.09286	-0.01142
Conjugative transfer protein 123*tavg	-1.33849	0.5176	-2.59	0.01	-2.35297	-0.32401
PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)*tavg	0.116912	0.037725	3.1	0.002	0.042973	0.190851
Secreted effector protein SteA*tavg	-52.5854	18.76711	-2.8	0.005	-89.3682	-15.8025
Uncharacterized protein YciW*tavg	0.062993	0.011097	5.68	0	0.041243	0.084742
Large repetitive protein*tavg	-0.08067	0.023613	-3.42	0.001	-0.12695	-0.03439

Copper/silver efflux RND transporter transmembrane protein CusA*tavg	-0.13928	0.032813	-4.24	0	-0.2036	-0.07497
Oxaloacetate decarboxylase Na(+) pump alpha chain (EC 4.1.1.3)*tavg	0.06511	0.026838	2.43	0.015	0.012509	0.117711
Fructokinase (EC 2.7.1.4)*tavg	17.74434	6.26332	2.83	0.005	5.468461	30.02022
Phage protein Ogr*tavg	-0.34707	0.177876	-1.95	0.051*	-0.6957	0.001565
PTS system ascorbate-specific IIC component*tavg	-0.32484	0.100093	-3.25	0.001	-0.52102	-0.12867
Phage tail tip host specificity protein J*tavg	-0.13404	0.042579	-3.15	0.002	-0.21749	-0.05058
Cobalamin synthase (EC 2.7.8.26)*tavg	0.183903	0.067219	2.74	0.006	0.052156	0.31565
TolA protein*tavg	0.074754	0.024308	3.08	0.002	0.027111	0.122398
Phage tail fiber protein GpH*tavg	-0.05204	0.023564	-2.21	0.027	-0.09822	-0.00586
Phage tail protein GpU*tavg	-0.88777	0.199956	-4.44	0	-1.27968	-0.49587
Mobile element protein*tavg	-0.55117	0.112759	-4.89	0	-0.77217	-0.33017
IncF plasmid conjugative transfer protein TraD*tavg	0.628332	0.223032	2.82	0.005	0.191196	1.065467
Phage baseplate assembly protein GpV*tavg	1.188794	0.360594	3.3	0.001	0.482042	1.895545
Uncharacterized J domain-containing protein YbeV*tavg	4.236758	1.532786	2.76	0.006	1.232552	7.240964
ADP-ribose pyrophosphatase of COG1058 family (EC 3.6.1.13)*tavg	2.558907	0.390711	6.55	0	1.793127	3.324687
Transcriptional regulator STM2275 GntR family*tavg	-0.16095	0.044063	-3.65	0	-0.24731	-0.07459
Aminoglycoside 3"-nucleotidyltransferase (EC 2.7.7.-) => ANT(3")-Ia (AadA family); ANT(9)-I*tavg	8.074298	2.95607	2.73	0.006	2.280507	13.86809
Phage integrase*tavg	-0.02751	0.011759	-2.34	0.019	-0.05056	-0.00446
Phage activator protein cII*tavg	-0.07942	0.020528	-3.87	0	-0.11965	-0.03918
DNA-damage-inducible protein I*tavg	0.073154	0.016927	4.32	0	0.039977	0.106331
Tn21 protein of unknown function Urf2*tavg	0.066194	0.018991	3.49	0	0.028971	0.103416

ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)*tavg	0.078788	0.009355	8.42	0	0.060453	0.097122
Phage tail fiber side tail fiber protein Stf*tavg	-0.99623	0.108582	-9.17	0	-1.20905	-0.78341
Phage activator protein cII*tavg	0.069115	0.039959	1.73	0.084*	-0.0092	0.147432
Resolvase*tavg	0.58649	0.111544	5.26	0	0.367867	0.805113
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA*tavg	-0.6919	0.073709	-9.39	0	-0.83636	-0.54743
Conjugative transfer protein 123*prcp	0.146314	0.057701	2.54	0.011	0.033222	0.259406
Secreted effector protein SteA*prcp	-18.0371	6.515214	-2.77	0.006	-30.8067	-5.26754
Copper/silver efflux RND transporter transmembrane protein CusA*prcp	0.020353	0.003022	6.73	0	0.014429	0.026276
Oxaloacetate decarboxylase Na(+) pump alpha chain (EC 4.1.1.3)*prcp	-0.01043	0.004412	-2.36	0.018	-0.01908	-0.00178
Fructokinase (EC 2.7.1.4)*prcp	-2.72804	0.978749	-2.79	0.005	-4.64635	-0.80973
PTS system ascorbate-specific IIC component*prcp	0.032766	0.010782	3.04	0.002	0.011633	0.053898
Cobalamin synthase (EC 2.7.8.26)*prcp	-0.02309	0.005223	-4.42	0	-0.03333	-0.01286
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)*prcp	0.042143	0.005954	7.08	0	0.030474	0.053812
Aminoglycoside 3"-nucleotidyltransferase (EC 2.7.7.-) => ANT(3")-Ia (AadA family); ANT(9)-I*prcp	0.009526	0.004561	2.09	0.037	0.000587	0.018465
TolA protein*prcp	-0.02852	0.005031	-5.67	0	-0.03838	-0.01866
Phage integrase*prcp	-0.02598	0.011609	-2.24	0.025	-0.04874	-0.00323
Type III secretion and flagellar regulator RtsA*prcp	0.006916	0.00275	2.51	0.012	0.001526	0.012307
Uncharacterized J domain-containing protein YbeV*prcp	-3.10552	1.107264	-2.8	0.005	-5.27571	-0.93532
Transcriptional regulator STM2275 GntR family*prcp	-0.01364	0.005098	-2.67	0.007	-0.02363	-0.00364
Uncharacterized MFS-type transporter*prcp	0.011674	0.004931	2.37	0.018	0.002009	0.021339

PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)*prcp	0.005396	0.002515	2.15	0.032	0.000466	0.010326
tRNA-(ms[2]io[6]A)-hydroxylase (MiaE)*prcp	-0.00374	0.001347	-2.78	0.006	-0.00638	-0.0011
Tn21 protein of unknown function Urf2*prcp	-0.00756	0.002201	-3.44	0.001	-0.01188	-0.00325
Phage tail fiber side tail fiber protein Stf*prcp	-0.10854	0.012485	-8.69	0	-0.13301	-0.08407
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA*prcp	0.040433	0.005659	7.14	0	0.029341	0.051525
FIL protein*prcp	0.057762	0.015416	3.75	0	0.027547	0.087977
Phage replication protein GpB*prcp	-0.04512	0.015057	-3	0.003	-0.07464	-0.01561
Conjugative transfer protein 123*snow	-0.05152	0.020964	-2.46	0.014	-0.09261	-0.01043
Phage tail fiber protein GpH*snow	-0.00324	0.001211	-2.68	0.007	-0.00561	-0.00087
Large repetitive protein*snow	-0.00601	0.001517	-3.96	0	-0.00898	-0.00304
Copper/silver efflux RND transporter transmembrane protein CusA*snow	-0.01487	0.003705	-4.01	0	-0.02213	-0.00761
Fructokinase (EC 2.7.1.4)*snow	0.976185	0.336255	2.9	0.004	0.317137	1.635233
Phage protein Ogr*snow	-0.00667	0.002224	-3	0.003	-0.01103	-0.00231
PTS system ascorbate-specific IIC component*snow	-0.04662	0.015005	-3.11	0.002	-0.07603	-0.01721
Phage tail tip host specificity protein J*snow	0.003887	0.002257	1.72	0.085*	-0.00054	0.00831
Phage integrase*snow	-0.02156	0.00385	-5.6	0	-0.0291	-0.01401
Type III secretion spans bacterial envelope protein (YscO)*snow	0.004681	0.0012	3.9	0	0.002328	0.007034
Uncharacterized J domain-containing protein YbeV*snow	0.010743	0.003605	2.98	0.003	0.003677	0.01781
Transcriptional regulator STM2275 GntR family*snow	-0.01245	0.003713	-3.35	0.001	-0.01972	-0.00517
Uncharacterized MFS-type transporter*snow	-0.01828	0.007975	-2.29	0.022	-0.03391	-0.00265
PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)*snow	-0.01652	0.00356	-4.64	0	-0.02349	-0.00954

Aminoglycoside 3"-nucleotidyltransferase (EC 2.7.7.-) => ANT(3")-Ia (AadA family); ANT(9)-I*snow	-0.09406	0.035315	-2.66	0.008	-0.16327	-0.02484
DNA-damage-inducible protein I*snow	0.006397	0.001231	5.2	0	0.003985	0.008809
Tn21 protein of unknown function Urf2*snow	0.004971	0.001707	2.91	0.004	0.001626	0.008317
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)*snow	0.00568	0.000893	6.36	0	0.003929	0.007431
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)*snow	0.003524	0.001564	2.25	0.024	0.000458	0.006589
Phage activator protein cII*snow	0.010357	0.003427	3.02	0.003	0.00364	0.017074
Phage protein*snow	-0.00533	0.003119	-1.71	0.088*	-0.01144	0.000785
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA*snow	-0.06416	0.007295	-8.79	0	-0.07845	-0.04986
FIL protein*snow	0.049875	0.03015	1.65	0.098*	-0.00922	0.108969
Phage replication protein GpB*snow	-0.06796	0.029887	-2.27	0.023	-0.12654	-0.00938
Mobile element protein*snow	0.017178	0.006452	2.66	0.008	0.004532	0.029824
Constant	-371.006	144.9256	-2.56	0.01	-655.055	-86.9568

Data for variables that did not pass the threshold for significance, or which were automatically omitted by the model, not included herein.

P values indicated with * are significant at $\alpha = 90\%$. All other significant observations are significant at $\alpha = 95\%$.

5.4.4. Negative binomial regression model outcome

The negative binomial model, a variant of a Poisson regression model was developed similar to the Poisson model, using significant gene presence/absence and meteorological factors as covariates. The negative binomial regression model was used to loosen the restrictions set by a Poisson model (mean = variance). We found that the negative binomial model did not perform as well as the Poisson regression in fitting the data. The mixed negative binomial model, using all covariates and interaction terms did not show a good improvement over the null model (χ^2 statistic = 912.44; Pseudo R^2 = 0.2442; probability > χ^2 = 0.0000). Moreover, we observed that the covariates 'mean average daily temperature' and 'mean precipitation' did not significantly impact the model (**Table 5.4**). Thus, the results of this model were dropped from further consideration.

Table 5.4. Negative binomial regression coefficients for multi-variable regression**.** Presence/absence of genes identified as significant by the Elastic Net model were input along with the standardized covariates monthly mean average daily temperature (tavg, in °F), monthly mean precipitation (prcp, in inches), and monthly mean snow cover (snow, in inches).

Predictor variables	**Coefficient**	**Standard error**	**z**	**P > z**	**[95% confidence interval]**	
Individual effect (gene or meteorological factor only)						
TnpA transposase	6.350179	3.259748	1.95	0.051*	-0.03881	12.73917
Iron-sulfur cluster assembly protein SufD	0.906526	0.367715	2.47	0.014	0.185818	1.627233
Replication protein	6.724287	3.396197	1.98	0.048	0.067864	13.38071
PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)	0.847904	0.301911	2.81	0.005	0.256168	1.439639
Phage lysozyme R (EC 3.2.1.17)	-1.29812	0.436763	-2.97	0.003	-2.15416	-0.44208
Two-component transcriptional response regulator BtsR	1.239956	0.186895	6.63	0	0.873648	1.606264
Phage tail fiber protein GpH	-0.61126	0.283359	-2.16	0.031	-1.16664	-0.05589
Large repetitive protein	-1.48541	0.580664	-2.56	0.011	-2.62349	-0.34733
Phage protein Ogr	-3.77682	1.003906	-3.76	0	-5.74444	-1.8092
Phage tail tip host specificity protein J	3.969457	0.890443	4.46	0	2.224221	5.714693
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)	10.99693	4.53113	2.43	0.015	2.116079	19.87778
IncF plasmid conjugative transfer protein TraP	1.55751	0.305619	5.1	0	0.958509	2.156512
Mobile element protein	-9.02398	3.478404	-2.59	0.009	-15.8415	-2.20643

SbmA protein	2.568552	0.522142	4.92	0	1.545173	3.591932
Rep protein	10.23198	4.57048	2.24	0.025	1.274002	19.18995
Mobile element protein	26.07243	10.95793	2.38	0.017	4.595289	47.54957
IncF plasmid conjugative transfer protein TraD	-845.222	372.908	-2.27	0.023	-1576.11	-114.336
HTH-type transcriptional regulator MlrA	2.109655	0.55162	3.82	0	1.0285	3.190811
L-carnitine/gamma-butyrobetaine antiporter	2.014045	0.600934	3.35	0.001	0.836236	3.191854
ADP-ribose pyrophosphatase of COG1058 family (EC 3.6.1.13)	12.58691	2.259701	5.57	0	8.157979	17.01584
Multidrug efflux system MdtABC-TolC inner-membrane proton/drug antiporter MdtC (RND type)	-1.6171	0.324067	-4.99	0	-2.25226	-0.98194
DNA-damage-inducible protein I	0.796208	0.348809	2.28	0.022	0.112556	1.47986
Tn21 protein of unknown function Urf2 Type III secretion spans bacterial envelope protein (YscO)	0.715662	0.304443	2.35	0.019	0.118966	1.312358
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)	0.744915	0.177554	4.2	0	0.396915	1.092915
Phage tail fiber side tail fiber protein Stf	32.33057	16.73855	1.93	0.053*	-0.47639	65.13753
Resolvase	-21.2964	11.64346	-1.83	0.067*	-44.1172	1.524357
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA Phage P4-associated	8.789888	2.868127	3.06	0.002	3.168463	14.41131
tavg	0.466526	0.30291	1.54	0.124	-0.12717	1.060219
prcp	0.074717	0.164465	0.45	0.65	-0.24763	0.397062
snow	0.03985	0.0206	1.93	0.053*	-0.00053	0.080225
Interaction effects						
TnpA transposase*tavg	-5.17343	2.416302	-2.14	0.032	-9.9093	-0.43757

Phage tail fiber protein GpH*tavg	-0.06262	0.022853	-2.74	0.006	-0.10741	-0.01783
Large repetitive protein*tavg	0.080944	0.049002	1.65	0.099*	-0.0151	0.176987
Phage protein Ogr*tavg	0.86395	0.280612	3.08	0.002	0.313961	1.413939
Phage tail tip host specificity protein J *tavg	-0.24449	0.060643	-4.03	0	-0.36335	-0.12563
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)*tavg	2.037308	0.989519	2.06	0.04	0.097887	3.976729
Aminoglycoside 3"-nucleotidyltransferase (EC 2.7.7.-) => ANT(3")-Ia (AadA family)*tavg	-0.06018	0.036491	-1.65	0.099*	-0.1317	0.011343
TolA protein*tavg	0.158033	0.050522	3.13	0.002	0.059012	0.257055
Mobile element protein*tavg	-6.48205	2.728607	-2.38	0.018	-11.83	-1.13408
IncF plasmid conjugative transfer protein TraD*tavg	46.00818	20.17627	2.28	0.023	6.463419	85.55293
Phage baseplate assembly protein GpV*tavg	-0.57812	0.315755	-1.83	0.067*	-1.19698	0.040751
HTH-type transcriptional regulator MlrA*tavg	0.087311	0.034408	2.54	0.011	0.019874	0.154749
Type III secretion and flagellar regulator RtsA*tavg	-0.05393	0.031872	-1.69	0.091*	-0.1164	0.00854
Phage integrase*tavg	-0.02937	0.017225	-1.7	0.088*	-0.06313	0.004394
DNA-damage-inducible protein I*tavg	-0.06352	0.020778	-3.06	0.002	-0.10425	-0.0228
tRNA-(ms[2]io[6]A)-hydroxylase (MiaE)*tavg	-0.03361	0.016187	-2.08	0.038	-0.06534	-0.00189
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)*tavg	-0.03902	0.012758	-3.06	0.002	-0.06402	-0.01401
Phage tail fiber side tail fiber protein Stf*tavg	-1.00217	0.522355	-1.92	0.055*	-2.02597	0.021625
Resolvase*tavg	6.021522	2.869371	2.1	0.036	0.397659	11.64539
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA Phage P4-associated*tavg	-0.8177	0.382217	-2.14	0.032	-1.56683	-0.06857
Uncharacterized J domain-containing protein YbeV*tavg	-0.103	0.044536	-2.31	0.021	-0.19029	-0.01572
Secreted effector protein SteA*prcp	0.094434	0.036348	2.6	0.009	0.023193	0.165675

Phage tail fiber protein GpH*prcp	0.009745	0.003258	2.99	0.003	0.00336	0.01613
Large repetitive protein*prcp	-0.00813	0.003052	-2.66	0.008	-0.01411	-0.00215
Oxaloacetate decarboxylase Na(+) pump alpha chain (EC 4.1.1.3)*prcp	-0.01337	0.004625	-2.89	0.004	-0.02243	-0.0043
Cobalamin synthase (EC 2.7.8.26)*prcp	-0.01207	0.003514	-3.44	0.001	-0.01896	-0.00519
Phage tail fiber protein GpH*prcp	0.106022	0.064268	1.65	0.099*	-0.01994	0.231985
Phage integrase*prcp	-0.03064	0.016849	-1.82	0.069*	-0.06367	0.002381
HTH-type transcriptional regulator MlrA*prcp	-0.01746	0.006582	-2.65	0.008	-0.03036	-0.00455
Type III secretion and flagellar regulator RtsA*prcp	0.003708	0.002238	1.66	0.098*	-0.00068	0.008095
ADP-ribose pyrophosphatase of COG1058 family (EC 3.6.1.13)*prcp	-0.21986	0.033612	-6.54	0	-0.28574	-0.15398
PTS system IIA component - PTS system D-glucosaminate-specific IIA component (EC 2.7.1.203)*prcp	0.008889	0.002989	2.97	0.003	0.00303	0.014747
Phage activator protein cII*prcp	-0.0083	0.002331	-3.56	0	-0.01287	-0.00373
DNA-damage-inducible protein I*prcp	0.007461	0.002108	3.54	0	0.00333	0.011592
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)*prcp	0.002627	0.000744	3.53	0	0.001169	0.004085
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)*prcp	0.012045	0.002537	4.75	0	0.007071	0.017018
Resolvase*prcp	-0.02429	0.007435	-3.27	0.001	-0.03886	-0.00972
FIL protein*prcp	0.013396	0.006395	2.09	0.036	0.000862	0.02593
Mobile element protein*prcp	0.043004	0.007958	5.4	0	0.027407	0.058601
Large repetitive protein*snow	0.012401	0.003546	3.5	0	0.005452	0.01935
Oxaloacetate decarboxylase Na(+) pump alpha chain (EC 4.1.1.3)*snow	0.015105	0.005222	2.89	0.004	0.004869	0.02534
Phage protein Ogr*snow	0.01511	0.003915	3.86	0	0.007438	0.022783

Phage tail tip host specificity protein J*snow	-0.01283	0.003155	-4.07	0	-0.01901	-0.00665
Cobalamin synthase (EC 2.7.8.26)*snow	0.010069	0.004247	2.37	0.018	0.001744	0.018393
Oxaloacetate decarboxylase Na(+) pump beta chain (EC 4.1.1.3)*snow	0.057655	0.024671	2.34	0.019	0.009302	0.106009
Type III secretion and flagellar regulator RtsA*snow	-0.00582	0.002826	-2.06	0.039	-0.01136	-0.00028
Uncharacterized J domain-containing protein YbeV*snow	0.008446	0.002693	3.14	0.002	0.003167	0.013725
Phage activator protein cII*snow	0.006826	0.001782	3.83	0	0.003332	0.010319
DNA-damage-inducible protein I*snow	-0.00746	0.001757	-4.25	0	-0.01091	-0.00402
ABC transporter permease protein STM1634 (cluster 3 basic aa/glutamine/opines)*snow	-0.00467	0.000966	-4.83	0	-0.00656	-0.00277
Zinc binding domain / DNA primase (EC 2.7.7.-) Phage P4-associated / Replicative helicase RepA Phage P4-associated*snow	-0.06552	0.038122	-1.72	0.086*	-0.14023	0.0092
FIL protein*snow	-0.01141	0.002534	-4.5	0	-0.01638	-0.00644
Constant	-6.82743	5.862327	-1.16	0.244	-18.3174	4.66252

Data for variables that did not pass the threshold for significance, or which were automatically omitted by the model, not included herein

(excluding individual meteorological factors, as applicable).

P values indicated with * are significant at $\alpha = 90\%$. All other significant observations are significant at $\alpha = 95\%$.

5.5. Discussion

Climatological and meteorological factors have been repeatedly implicated in the rise in incidence and impact (in terms of number of illnesses, hospitalizations, etc.) of illnesses caused by bacterial agents such as *Salmonella enterica* (Rose et al., 2001; Simental & Martinez-Urtaza, 2008; McMichael, 2015). Particularly, a positive association has been reported between diarrheal disease numbers and temperature increase (Singh et al., 2001). Moreover, a number of studies have indicated that factors such as increased temperatures and precipitation (as well as relative humidity) in the environment lead to an increase in environmental *Salmonella* presence and persistence (Akil, Ahmed & Reddy, 2014).

The bacterial genetic code holds the key to unlocking the many secrets governing bacterial pathogen growth, survival, proliferation, and pathogenicity. However, its potential is only now being realized with the advent of whole genome sequencing. Currently, WGS is being applied to surveillance and disease outbreak investigation, and identifying the key mechanisms behind pathogen virulence and survival to understand and control pathogens in food (Fritsch, Guillier, & Augustin, 2018; Pornsukarom, van Vliet, & Thakur, 2018). However, identifying the underlying trends, correlations, and relationships from such data adds multiple dimensions to even simple survival kinetics, necessitating multi-dimensional analytical considerations (Strawn et al., 2015). Thus, a primary consideration of researchers is to develop methods to analyze and obtain meaningful data from WGS, specifically in the case of preventative modeling of pathogen growth, survival,

and overall human health risk. Researchers are increasingly looking towards machine learning and advanced data analysis to overcome these issues. However, so far, the joint impact of a pathogen's genetic expression and meteorological factors such as temperature and precipitation on the pathogen's infectivity and outbreak severity (in terms of case numbers) has not been analyzed. In this project, we utilize gene presence-absence data, which is more readily obtainable from whole-genome sequencing, compared to gene abundance data, which would be a more ideal metric for effect estimation.

Here, we used outbreak case numbers consolidated by outbreak area (state), month and year as the outcome variable, and gene expression data and meteorological variables, specifically state-wise mean daily temperature, mean daily precipitation, and mean snowfall, as predictor variables in a Poisson regression model (since the outcome variable is in counts). While only the most important among the large number of genetic predictors were selection by Elastic Net feature selection, meteorological data for each observation was averaged from data obtained from all weather stations in the respective state during the month of the outbreak, similar to the approach used by Akil, Ahmed, & Reddy (2014). The best-fit Poisson regression model identified a number of genes and gene-meteorological factor interaction terms (n = 127) that significantly contributed to salmonellosis outbreak numbers (**Table 5.3**). In general, we observed that a majority of significant gene-only variables were positively correlated with salmonellosis case numbers. Among those that were negatively correlated, the gene functionality ranged from phage-related virulence, bacterial metabolism, and membrane transport. In sharp contrast,

interaction effects of a large number of phage proteins with environmental temperature were negatively correlated with outbreak severity, indicating that the combined effect of a unit increase in temperature and gene expression led to corresponding decrease in the log counts of outbreaks. Concurrently, we observed that the temperature-interaction effects of a large percentage of metabolism and cell maintenance-related proteins were positively correlated with outbreak severity. This is in agreement with the conclusions of Pin et al. (2012) and Dawoud et al. (2017), who reported an upregulation in stress-, energy metabolism-, and cellular mechanism-related genes in *Salmonella enterica* under thermal and other stress conditions. We also observed a positive correlation between the mean precipitation effect and outbreak numbers, which is in line with a prior report by Soneja et al. (2016). The precipitation-gene expression interaction patterns were similar to those observed for the temperature-gene expression effects. Interestingly, we also observed a positive correlation between average snowfall and outbreak numbers, which in turn could be correlated with the increased precipitation (Holley et al., 2008; Piekarska, 2010).

We obtained some confounding results regarding the effect of temperatures on outbreak severity (as defined by case numbers). We observed that, for a one °F increase in average temperature, the difference in log of expected case numbers would be expected to decrease by 14.5395. While this relationship is contrary to published literature and our own research into the relationship between meteorological factors and outbreak trends (**Figure 4.3**), that have repeatedly found a positive association between increasing temperature and salmonellosis incidence rates, these results are in agreement with those of Semenov, van

Bruggen, van Overbeek, Termorshuizen, & Semenov (2007), who reported similar inconsistent conclusions about temperature levels contributing to *Salmonella* survival. In essence, they found that *Salmonella* survival significantly declined with increasing mean temperatures, indicating that fixed measures of parameters such as temperature and precipitation need not necessarily capture the impact of fluctuating temperatures (as is commonly seen under natural conditions, captured by meteorological measurements) on the characteristics of *Salmonella*. We attribute these results to the discrete nature of our data, and to the secondary nature of our study, which make it impossible to pinpoint exact causal relations between the predictor and response variables.

Our study has a few limitations. As in the case of most analyses pertaining to foodborne outbreaks, our dataset is limited by underreporting of illnesses. For example, a majority of illnesses may not be serious enough to warrant testing, let alone hospitalization. Second, since our WGS dataset is built from among isolates obtained to correlate with salmonellosis outbreaks, the initial pan genome dataset is not wholly representative of all *Salmonella* serovars that have caused foodborne diseases in humans. Moreover, while WGS can determine if a microbe is the root cause of a foodborne outbreak, a lack of defined thresholds regarding genetic differences and the dependency of similarity (to other isolates) identification on prior knowledge (from previous outbreaks, etc.) makes it difficult to conclusively determine the level of mutation needed to identify an isolate as truly being 'different.' Finally, due to the small number of data points, meteorological factors have been pooled across all sampled states, since the effects of these factors taken from

individual state level data were not significant. Such issues necessitate field- and laboratory-level analyses of the changes observed in pathogens under specific conditions that can be observed in the environment to truly capture the genome-level effect of factors (such as meteorological factors) on *Salmonella* persistence and virulence, and subsequently, its effect on outbreak numbers.

5.6. Conclusion

Meteorological factors such as ambient temperature, precipitation, and humidity have been shown to significantly impact the incidence and severity of bacterial foodborne outbreaks. The increase in availability of WGS data has allowed for the rapid detection of bacteria to help with disease epidemiology, as well as predict disease severity based on presence or absence of significant genes or gene groups. Studies have also shown how meteorological factors specifically upregulate or downregulate the expression of specific bacterial genes, such as those coding for stress tolerance and bacterial metabolism. However, so far, there have been no studies attempting to identify the combined impact of bacterial gene expression and meteorological factors on outbreak severity, primarily due to the lack of controlled datasets and models to efficiently analyze such large datasets. Machine learning is a powerful tool that can be trained to identify patterns from large datasets that are indicative of a specific outcome. In this project, we developed multi-variable Poisson regression models to determine the impact of *Salmonella enterica* genes, pooled (by month and year) meteorological factors, and combinations of the two, on *Salmonella* outbreak severity. We identified a large number of genes that significantly

impacted the outcome, specifically those coding for metabolism, cellular function, and stress response. Ambient temperature and precipitation also played a role (individually and in combination with significant genes) in predicting outcome severity. We envision this as the first step towards incorporating the effect of bacterial gene expression in models predicting bacterial foodborne outbreak severity, which are traditionally based on environmental and processing-related factors.

Chapter 6: Summary and future studies

6.1. Summary

Although WGS and other molecular data help delineate the specific characteristics of microbial and host systems under conditions encountered in the food system and under infection conditions, their utilization in the field of QMRA remains in its infancy. This is due to the vast number of variables/features introduced by such data, which traditional algorithmic models are unable to process effectively. Machine learning and advanced data analytics are novel methods that have recently entered the food safety domain. The applicability of these methods in predictive microbial modeling, explanatory modeling, and in quantitative microbial risk assessment is increasingly being explored. This project has identified and successfully evaluated the applicability of these novel techniques in microbial genomic modeling. These in turn, would be eventually incorporated in various stages of a molecular data-informed QMRA framework to better inform

In Chapter 3, we focused on the development of machine learning-based methods to analyze whole genome sequences of *Salmonella enterica,* and subsequently classify isolates based on the severity of infection in host systems. A primary obstacle towards utilizing gene expression datasets in microbial modeling frameworks is the non-availability of associated metadata and endpoint data, as well the very large number of predictors compared to the number of available samples. In this project, we developed a workflow to identify genes that significantly contributed to the model outcome using a combination of

Elastic Net feature selection and machine learning classification modeling. Among the four classification models tested, Elastic Net-regularized logistic regression with ridge proved to be the most accurate, with an AUC-ROC of 0.86 and high sensitivity and specificity values, compared to the other tested models. This is especially important, as it is easy to interpret the coefficients of a logistic regression, compared to other machine learning-based models. The best-fit logistic regression model identified a number of genes, varying in functionality from virulence, stress response, to bacterial metabolism, that contributed to different illness outcomes in the host, and could therefore be the focus of further outbreak severity studies.

Chapter 4 focused on the development of a machine learning-based workflow to modulate *Salmonella enterica* dose response based on gene expression. In this project, we developed a method to directly incorporate gene expression data and machine learning modeling into a risk assessment framework. Multi-serovar bacterial species like *Salmonella* tend towards differing dose-response profiles in host systems. Studies have postulated that the genetic makeup, influenced by recombination and horizontal gene transmission, of each serovar could impact this differential response. In our study, we developed a machine learning-based weighted regression method to incorporate genes into a dose-response framework, using the weights of expression of genes that could be associated with a host response. A weighted Poisson regression was employed for this purpose in order to effectively deal with count data (%incidence). The cross-validated model identified 9 predictor variables – a majority of which coded for metabolism-related

functions – as contributing significantly towards model performance. Tellingly, only dehydrogenases and aldolases, which are responsible for bacterial survival under anaerobic host conditions, and contribute significantly to bacterial pathogenicity, were found to positively impact the dose-response. This study provides a potential means to identify and directly incorporate significant genetic entities (as opposed to serovar-level information) in refined dose-response modeling frameworks.

Chapter 5 focused on developing a machine learning-based method to identify and correlate genetic and meteorological factors to salmonellosis outbreak incidence and severity. Although previous studies have reported on a relationship between meteorological events (extrinsic factors) and genetic factors (intrinsic factors) on salmonellosis outbreak severity individually, none have analyzed or reported on the joint effect of these extrinsic and (microbial) intrinsic effects on outbreak severity. In our study, we employed a machine learning-supported weighted count-based regression approach for this purpose. We found that a weighted-Poisson approach provided the best model fit, with a pseudo R^2 value > 0.68, compared to the less restrictive negative binomial model. The final model identified a number of genes, mostly coding for bacterial metabolism, stress response, and (less frequently) virulence, that independently, and interacting with meteorological factors, contributed to salmonellosis outbreak severity. These models are envisioned as the first step towards incorporating the effect of bacterial gene expression in models predicting bacterial foodborne outbreak severity, which are traditionally based on environmental and processing-related factors.

6.2. Future studies

This dissertation represents our current best knowledge of the applicability of machine learning and advanced data analytical methods to incorporate molecular data into a predictive modeling and risk assessment framework. Several data gaps were identified and elaborated on in each chapter. Potential areas of research and methods required for, and pertaining to, the incorporation of molecular data in a QMRA framework are proposed as follows:

1. A majority of molecular data that is currently being collected for pathogenic surveillance and outbreak investigations do not have, collect or report on, associated metadata. Such information is very useful to establish the outcome variable (confirmed, as opposed to inferential) in machine learning-based modeling. This could help in the development of more accurate, representative models to predict bacterial characteristics and behavior under various food processing and infective conditions to be employed in a risk assessment framework.
2. Currently, a majority of the available WGS and other molecular data does not support the determination of causal relationships – i.e., machine learning can at best identify correlations between the expression or non-expression of certain genes or gene subsets and relevant phenotypic functions. However, in order to determine causal relationships from WGS-informed predictive microbial growth, survival, and death models, experimental data is needed to determine gene expression patterns under various processing and stress-related conditions.

3. *Salmonella enterica*, in particular, is composed of a number of serovars that infect humans at various levels of infectivity, and interacts differently under various processing-related stress conditions. Our current knowledge on *Salmonella* infectivity, pathogenicity, dose-response, and survival, is dependent on experimental data generated from a select few serovars. In order to obtain a more representative, labeled dataset, with which to train learning models, more experimental data from serovars that are under-represented, but are highly infective/invasive, or demonstrating specific stress-response profiles are needed.
4. Although whole genome sequences alone, and the resultant pan genome, provide a wealth of information that could be very informative to predictive microbial models, studies have repeatedly indicated that the presence of a gene in and of itself does not guarantee its expression and functionality. Thus, analyzing the expression characteristics of bacteria under the various food processing- (such as temperature stress, acid stress, desiccation stress, etc.) and infectivity-related (such as the impact of host immune system) stressors, in the form of stress-specific proteomics and transcriptomics could help in further refining existing growth and survival models using laboratory-informed data.

Bibliography

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., Jones, J. L., & Griffin, P. M. (2011). Foodborne illness acquired in the United States: Major pathogens. *Emerging Infectious Diseases, 17(1)*, pp. 7–15.

United States Centers for Disease Control and Prevention – U.S. CDC. (2021a). *Foodborne burden - CDC*. Retrieved October 1, 2021, from http://www.cdc.gov/foodborneburden/index.html

Rantsiou, K., Mataragas, M., Jespersen, L., & Cocolin, L. (2011). Understanding the behavior of foodborne pathogens in the food chain: New information for risk assessment analysis. *Trends in Food Science and Technology, 22*, S21-S29.

Whiting, R. C., & Buchanan, R. L. (1997). Development of a quantitative risk assessment model for *Salmonella* Enteritidis in pasteurized liquid eggs. *International Journal of Food Microbiology, 36,* pp.111–125.

Pradhan, A. K., Ivanek, R., Gröhn, Y. T., Geornaras, I., Sofos, J. N., & Wiedmann, M. (2009). Quantitative risk assessment for *Listeria monocytogenes* in selected categories of deli meats: impact of lactate and diacetate on listeriosis cases and deaths. *Journal of Food Protection, 72,* pp.978–989.

Guo, M., Mishra, A., Buchanan, R. L., Dubey, J. P., Hill, D. E., Gamble, & Pradhan, A. K. (2016a). Quantifying the risk of human *Toxoplasma gondii* infection due to consumption of domestically produced lamb in the United States. *Journal of Food Protection, 79,* pp.1181–1187.

Guo, M., Mishra, A., Buchanan, R. L., Dubey, J. P., Hill, D. E., Gamble, H. R., Jones, J. L., Du, X., & Pradhan, A. K. (2016b). Development of dose-response models to predict the relationship for human *Toxoplasma gondii* infection associated with meat consumption. *Risk Analysis, 36,* pp. 926–938.

Food and Agricultural Organization of the United Nations (FAO)/World Health Organization (WHO). (2001). Risk characterization of *Salmonella* spp. in eggs and broiler chickens and *Listeria monocytogenes* in ready-to-eat foods. Joint FAO/WHO expert consultation on risk assessment of microbiological hazards in foods, 2001. FAO headquarters, Rome.

Guo, M., Lambertini, E., Buchanan, R. L., Dubey, J. P., Hill, D. E., Gamble, H. R., Jones, J. L., & Pradhan, A. K. (2017). Quantifying the risk of human *Toxoplasma gondii* infection due to consumption of fresh pork in the United States. *Food Control, 73,* pp.1210–1222.

Pang, H., Lambertini, E., Buchanan, R. L., Schaffner, D. W., & Pradhan A. K. (2017). Quantitative microbial risk assessment for *Escherichia coli* O157:H7 in fresh-cut lettuce. *Journal of Food Protection, 80,* pp.302–311.

Deng, X., den Bakker, H. C., & Hendriksen, R. S. (2016). Genomic epidemiology: whole-genome-sequencing–powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annual Review of Food Science and Technology*, *7,* pp.353–374.

Chen, Y., Burall, L. S., Luo, Y., Timme, R., Melka, D., Muruvanda, T., Payne, J., Wang, C., Kastanis, G., Maounounen-Laasri, A., de Jesus, A. J., Curry, P. E., Stones, R.,

K'Aluoch, O., Liu, E., Salter, M., Hammack, T. S., Evans, P. S., Parish, M., Allard, M. W., Datta, A., Strain, E. A., & Brown, E. W. (2016). Isolation, enumeration and whole genome sequencing of *Listeria monocytogenes* in stone fruits linked to a multistate outbreak. *Applied and Environmental Microbiology, 82,* pp.7030–7040..

Njage, P. M. K., Leekitcharoenphon, P., & Hald, T. (2019). Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: Predicting clinical outcomes in shigatoxigenic *Escherichia coli*. *International Journal of Food Microbiology, 292*, 72–82.

Njage, P. M. K., Henri, C., Leekitcharoenphon, P., Mistou, M. Y., Hendriksen, R. S., & Hald, T. (2019). Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. *Risk Analysis, 39(6),* 1397–1413. doi: 10.1111/risa.13239.

Collineau, L., Boerlin, P., Carson, C. A., Chapman, B., Fazil, A., Hetman, B., McEwen, S. A., Parmley, E. J., Reid-Smith, R. J., Taboada, E. N., & Smith, B. A. (2019). Integrating Whole-Genome Sequencing Data Into Quantitative Risk Assessment of Foodborne Antimicrobial Resistance: A Review of Opportunities and Challenges. *Frontiers in Microbiology* 10: 1107.

Strawn, L. K., Brown, E. W., David, J. R. D., Den Bakker, H. C., Vangay, P., Yiannas, F., & Wiedmann, M. (2015). Big data in food. *Food Technology, 69,* pp.42–49.

Phillips, A., Sotomayor, C., Wang, Q., Holmes, N., Furlong, C., Ward, K., Howard, P., Octavia, S., Lan, R., & Sintchenko, V. (2016). Whole genome sequencing of Salmonella Typhimurium illuminates distinct outbreaks caused by an endemic

multi-locus variable number tandem repeat analysis type in Australia, 2014. *BMC Microbiology*, *16*(1), pp.1–9.

Gilmour, M. W., Graham, M., Van Domselaar, G., Tyler, S., Kent, H., Trout-Yakel, K. M., Larios, O., Allen, V., Lee, B., & Nadon, C. (2010). High-throughput genome sequencing of two Listeria monocytogenes clinical isolates during a large foodborne outbreak. *BMC Genomics*, *11*(1), pp.1–15.

Chen, Y., Ross, W. H., Gray, M. J., Wiedmann, M., Whiting, R. C., & Scott, V. N. (2006). Attributing risk to *Listeria monocytogenes* subgroups: dose response in relation to genetic lineages. *Journal of Food Protection, 69*(2), pp.335–344.

Inns, T., Ashton, P., Herrera-Leon, S., Lighthill, J., Foulkes, S., Jombart, T., Rehman, Y., Fox, A., Dallmann, T., de Pinna, E., Browning, L., Coia, J. E., Edeghere, O., & Vivancos, R. (2016). Prospective use of whole-genome sequencing (WGS) detected a multi-country outberak of *Salmonella* Enteritidis. *Epidemiology and Infection, 145(2),* pp.1–10.

Nubel, U., Strommenger, B., Layer, F., & Witte, W. (2011). From types to trees: reconstructing the spatial spread of Staphylococcus aureus based on DNA variation. *International Journal of Medical Microbiology, 301,* 614-618.

Snitkin, E., Zelasny, A., Thomas, P., Stock, F., Henderson, T., Palmore, T., & Segre, J. (2012). Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine, 4,* 148ra116.

Octavia, S., Wang, Q., Tanaka, M., Kaur, S., Sintchenko, V., & Lan, R. (2015). Delineating community outbreaks of Salmonella enterica serovar Typhimurium by use of whole-genome sequencing: Insights into genomic variability within an outbreak. *Journal of Clinical Microbiology, 53*, 1063-1071.

van Dijk, E., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics, 30*(9), 418-426.

Membre, J.-M., & Guillou, S. (2016). Latest developments in foodborne pathogen risk assessment. *Current Opinion in Food Science, 8*, 120-126.

Brul, S., Bassett, J., Cook, P. Kathariou, S., McClure, P., Jasti, P. R., & Betts, R. (2012). 'Omics' technologies in quantitative microbial risk assessment. *Trends in Food Science and Technology, 27*, 12–24.

Alkema, W., Boekhorst, J., Wels, M., & van Hijum, S. A. (2016). Microbial bioinformatics for food safety and production. *Briefings in Bioinformatics, 17*(2), 283-292.

Vangay, P., Steingrimsson, J., Weidmann, M., & Stasiewicz, M. (2014). Classification of Listeria monocytogenes persistence in retail delicatessen environments using expert elicitation and machine learning. *Risk Analysis, 30*, 1830-1845.

Doyle, C., Gleeson, D., Jordan, K., Beresford, T., Ross, R., Fitzgerad, G., & Cotter, P. (2015). Anaerobic sporeformers and their significance with respect to milk and dairy products. *International Journal of Food Microbiology, 197*, 77-87.

Naccache, S., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A. L., Luk, K. C., E., B., Wadford, D. A., Messenger, S. L., Genrich, G. L., Pellegrino, K., Grard, G., Leroy, E., Schneider, B. S., Fair, J. N.,

Martinez, M. A., Isa, P., Crump, J. A., DeRisi, J. L., Sittler, T., Hackett, J., Miller, S., & Chiu, C. Y. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Research, 24*, 1180-1192.

Bore, E., & Langsrud, S. (2005). Characterization of micro-organisms isolated from dairy industry after cleaning and fogging disinfection with alkyl amine and peracetic acid. *Journal of Applied Microbiology, 98*, 96-105.

Fernandes, M. D., Kabuki, D., & Kuaye, A. (2015). Biofilms of Enterococcus faecalis and Enterococcus faecium isolated from the processing of ricotta and the control of these pathogens through cleaning and sanitization procedures. *International Journal of Food Microbiology, 200*, 97-103.

Jones, J. L., Wang, L., Ceric, O., Nemser, S. M., Rotstein, D. S., Jurkovic, D. A., Rosa, Y., Byrum, B., Cui, J., Zhang, Y., Brown, C. A., Burnum, A. L., Sanchez, S., & Reimschuessel, R. (2019). Whole genome sequencing confirms source of pathogens associated with bacterial foodborne illness in pets fed raw pet food. *Journal of Veterinary Diagnostic Investigation, 31*(2), 235–240.

Yahara, K., Méric, G., Taylor, A., de Vries, S., Murray, S., Pascoe, B., Mageiros, L., Torralbo, A., Vidal, A., Ridley, A., Komukai, S., Wimalrathna, H., Cody, A. J., Colles, F. M., McCarthy, N., Harris, D., Bray, J. E., Jolley, K. A., Maiden, M. C. J., Bentley, S. D., Parkhill, J., Bayliss, C. D., Grant, A., Maskell, D., Didelot, X., Kelly, D. J., & Sheppard, S. K. (2016). Genome-wide association of functional

traits linked with Campylobacter jejuni survival from farm to fork. *Environmental Microbiology*. *19(1),* pp.361–380.

Franz, E., van Hoek, A. H., Wuite, M., van der Wal, F. J., de Boer, A. G., Bouw, E. I., & Aarts, H. J. (2015). Molecular hazard identification of non-O157 Shiga toxin-producing *Escherichia coli* (STEC). *PLoS One*, *10*(3), e0120353.

Pielaat, A., Boer, M. P., Wijnands, L. M., van Hoek, A. H., Bouw, E., Barker, G. C., Teunis, P. F., Aarts, H. J., & Franz, E. (2015). First step in using molecular data for microbial food safety risk assessment; hazard identification of *Escherichia coli* O157:H7 by coupling genomic data with in vitro adherence to human epithelial cells. *International Journal of Food Microbiology, 213*, 130–138.

Deng, X., Cao, S., & Horn, A. L. (2021). Emerging applications of machine learning in food safety. *Annual Reviews of Food Science and Technology, 12,* pp.513–538.

Kampichler, C., Wieland, R., Calme, S., Weissenberger, H., & Arriaga-Weiss, S. (2010). Classification in conservation biology: A comparison of five machine-learning methods. *Ecological Information, 5*(6): pp.441–450.

Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques. *Computational and Structural Biotechnology Journal*, *15*, pp.26-47.

Schüffler, P. J., Mahapatra, D., Tielbeek, J. A., Vos, F. M., Makanyanga, J., Pendsé, D. A., Nio, C. Y., Stoker, J., Taylor, S. A., & Buhmann, J. M. (2013). A model development pipeline for Crohn's disease severity assessment from magnetic

resonance images. In International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging (pp. 1–10). Springer, Berlin, Heidelberg.

Tsanas, A. (2012). Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning (Doctoral dissertation, Oxford University, UK).

Armañanzas, R., Bielza, C., Chaudhuri, K. R., Martinez-Martin, P., & Larranaga, P. (2013). Unveiling relevant non-motor Parkinson's disease severity symptoms using a machine learning approach. *Artificial Intelligence in Medicine*, 58(3), 195-202.

Mwebaze, E., & Owomugisha, G. (2016, December). Machine learning for plant disease incidence and severity measurements from leaf images. In 2016 15th IEEE international conference on machine learning and applications (ICMLA) (pp. 158–163). IEEE.

Farrell, F., O. S. Soyer, and C. Quince. 2018. Machine learning based prediction of functional capabilities in metagenomically assembled microbial genomes. *bioRxiv* 307157 [pre-print].

Wheeler, N. E., Gardner, P. P., & Barquist, L. (2018). Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS Genetics.* https://doi.org/10.1371/journal.pgen.1007333.

Niehaus, K. E., Walker, T. M., Crook, D. W., Peto, T. E., & Clifton, D. A. (2014, June). Machine learning for the prediction of antibacterial susceptibility in

Mycobacterium tuberculosis. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 618-621). IEEE.

Pesesky, M. W., Hussain, T., Wallace, M., Patel, S., Andleeb, S., Burnham, C. A. D., & Dantas, G. (2016). Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data. *Frontiers in Microbiology*, *7*, pp.1887.

Nguyen, M., Brettin, T., Long, S. W., Musser, J. M., Olsen, R. J., Olson, R., ... & Davis, J. J. (2018). Developing an in silico minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Scientific Reports*, *8*(1), pp.1–11.

Maguire, F., Rehman, M. A., Carrillo, C., Diarra, M. S., & Beiko, R. G. (2019). Identification of primary antimicrobial resistance drivers in agricultural nontyphoidal *Salmonella enterica* serovars by using machine learning. *Msystems*, *4*(4), pp.e00211-19.

Nguyen, M., Long, S. W., McDermott, P. F., Olsen, R. J., Olson, R., Stevens, R. L., ... & Davis, J. J. (2019). Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *Journal of Clinical Microbiology*, *57*(2), pp.e01260–18.

Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L., & Gally, D. L. (2016). Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proceedings of the National Academy of Sciences*, *113*(40), pp.11312-11317.

Munck, N., Njage, P. M. K., Leekitcharoenphon, P., Litrup, E., & Hald, T. (2020). Application of whole-genome sequences and machine learning in source attribution of *Salmonella* Typhimurium. *Risk Analysis, 40(9),* 1693–1705.

US Centers for Disease Control and Prevention (2021b). *Salmonella* – Questions and answers. Available at: Questions and Answers | Salmonella | CDC. Accessed October 1, 2021.

US Centers for Disease Control and Prevention. (2019a). Reports of selected *Salmonella* outbreak investigations. Available at: https://www.cdc.gov/salmonella/outbreaks.html. Accessed 22 June 2019.

US Centers for Disease Control and Prevention. (2019b). Estimates of foodborne illness in the United States. Available at: http://www.cdc.gov/foodborneburden/index.html. Accessed 20 June 2019.

Ferrari, R. C., Rosario, D. K. A., Cunha-Neto, A., Mano, S. B., Figueiredo, E. E. S., & Conte-Junior, C. A. (2019). Worldwide epidemiology of *Salmonella* serovars in animal-based foods. *Applied and Environmental Microbiology, 85*(14).

Anjum, M. F., Marooney, C., Fookes, M., Baker, S., Dougan, G., Ivens, A., & Woodward, M. J. (2005). Identification of core and variable components of the *Salmonella enterica* subspecies I genome by microarray. *Infection and Immunity*, *73*(12), pp.7894–7905.

Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W., & Friis, C. (2011). The *Salmonella enterica* pan-genome. *Microbial Ecology*, *62*(3), pp.487–504.

Cheng, R. A., Eade, C. R., & Wiedmann, M. (2019). Embracing diversity: differences in

virulence mechanisms, disease severity, and host adaptations contribute to the success of nontyphoidal *Salmonella* as a foodborne pathogen. *Frontiers in Microbiology, 10,* 1368.

Jones, T. F., Ingram, L. A., Cieslak, P. R., Vugia, D. J., Tobin-D'angelo, M., Hurd, S., et al. (2008). Salmonellosis outcomes differ substantially by serotype. *Journal of Infectious Diseases, 198,* pp.109–114.

Scott, J. A. G., Berkley, J. A., Mwangi, I., Ochola, L., Uyoga, S., Macharia, A., et al. (2011). Relation between falciparum malaria and bacteraemia in Kenyan children: a population-based, case-control study and a longitudinal study. *Lancet, 378,* pp.1316–1323.

Feasey, N. A., Dougan, G., Kingsley, R. A., Heyderman, R. S., and Gordon, M. A. (2012). Invasive nontyphoidal *salmonella* disease: an emerging and neglected tropical disease in Africa. *Lancet, 379,* pp.2489–2499.

Okoro, C. K., Kingsley, R. A., Connor, T. R., Harris, S. R., Parry, C. M., Al-Mashhadani, M. N., et al. (2012). Intracontinental spread of human invasive *Salmonella Typhimurium* pathovariants in sub-Saharan Africa. *Nature Genetics, 44,* pp.1215–1221.

Ao, T. T., Feasey, N. A., Gordon, M. A., Keddy, K. H., Angulo, F. J., and Crump, J. A. (2015). Global burden of invasive nontyphoidal *Salmonella* disease, 2010. *Emerging Infectious Diseases, 21*, pp.941.

Lan, N. P. H., Phuong, T. L. T., Huu, H. N., Thuy, L., Mather, A. E., Park, S. E., et al. (2016). Invasive nontyphoidal *Salmonella* infections in Asia: clinical observations,

disease outcome and dominant serovars from an infectious disease hospital in Vietnam. *PLoS Neglected Tropical Diseases, 10,* e0004857.

Hoffmann, M., Zhao, S., Pettengill, J., Luo, Y., Monday, S. R., Abbott, J., Ayers, S. L., Cinar, H. L., Muruvanda, T., Li, C., Allard, M. W., Whichard, J., Meng, J., Brown, E. W., & McDermott, P. F. (2014). Comparative genomic analysis and virulence differences in closely related *Salmonella enterica* serotype Heidelberg isolates from humans, retail meats, and animals. *Genome Biology and Evolution*, *6*(5), pp.1046–1068.

Fierer, J. & Guiney, D. G. (2001). Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection. *Journal of Clinical Investigation 107*(7), pp.775–780.

Marzel, A., Desai, P. T., Goren, A., Schorr, Y. I., Nissan, I., Porwollik, S., Valinsky, L., McClelland, M., Rahav, G., & Gal-Mor, O. (2016). Persistent infections by nontyphoidal *Salmonella* in humans: epidemiology and genetics. *Clinical Infectious Diseases*, *62*(7), pp.879–886.

Graziani, C., L. Busani, A. M. Dionisi, A. Caprioli, S. Ivarsson, I. Hendenstrom, et al. 2011. "Virulotyping of *Salmonella Enterica* serovar Napoli strains isolated in Italy from human and nonhuman sources." *Foodborne Pathog. Dis.* 8(9): 997–1003.

Capuano, F., A. Mancusi, R. Capparelli, S. Esposito, and Y. T. R. Proroga. 2013. "Characterization of drug resistance and virulotypes of *Salmonella* strains isolated from food and humans." *Foodborne Pathog. Dis.* 10(11): 963–68.

Suez, J., Porwollik, S., Dagan, A., Marzel, A., Schorr, Y. I., Desai, P. T., Agmon, V.,

McClelland, M., Rahav, G. & Gal-Mor, O. (2014). Virulence gene profiling and pathogenicity characterization of non-typhoidal *Salmonella* accounted for invasive disease in humans. *PLoS One, 8(3),* e58449.

Chen, J., Karanth, S., & Pradhan, A. K. (2020). Quantitative microbial risk assessment for *Salmonella*: Inclusion of whole genome sequencing and genomic epidemiological studies, and advances in the bioinformatics pipeline. *Journal of Agriculture and Food Research, 2,* e100045.

Netoff, T. I. (2019). The ability to predict seizure onset. In *Engineering in Medicine* (pp.365–378). Academic Press.

Misra, S., & Wu, Y. (2019). Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. *Machine Learning for Subsurface Characterization, 289.*

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine Learning* (pp.101–121). Academic Press.

Qin, J., Burks, T. F., Kim, M. S., Chao, K., & Ritenour, M. A. (2008). Citrus canker detection using hyperspectral reflectance imaging and PCA-based image classification method. *Sensing and Instrumentation for Food Quality and Safety*, *2*(3), 168–177.

Powell, D. A., Jacob, C. J., & Chapman, B. J. (2011). Enhancing food safety culture to reduce rates of foodborne illness. *Food Control*, *22*(6), pp.817–822.

Qin, J., Chao, K., Kim, M. S., Lu, R., & Burks, T. F. (2013). Hyperspectral and multispectral imaging for evaluating food safety and quality. *Journal of Food Engineering*, *118*(2), pp.157–171.

Guillén-Casla, V., Rosales-Conrado, N., León-González, M. E., Pérez-Arribas, L. V., & Polo-Díez, L. M. (2011). Principal component analysis (PCA) and multiple linear regression (MLR) statistical tools to evaluate the effect of E-beam irradiation on ready-to-eat food. *Journal of Food Composition and Analysis*, *24*(3), pp. 456–464.

Yang, M., Liu, X., Luo, Y., Pearlstein, A. J., Wang, S., Dillow, H., ... & Zhang, B. (2021). Machine learning-enabled non-destructive paper chromogenic array detection of multiplexed viable pathogens on food. *Nature Food*, *2*(2), pp.110–117.

Mughini-Gras, L., Smid, J., Enserink, R., Franz, E., Schouls, L., Heck, M., & van Pelt, W. (2014). Tracing the sources of human salmonellosis: A multi-model comparison of phenotyping and genotyping methods. *Infection, Genetics, Evolution*, *28*, 251–260.

Abbott, S. L., Ni, F. C., & Janda, J. M. (2012). Increase in extraintestinal infections caused by *Salmonella enterica* subspecies II-IV. *Emerging Infectious Diseases*, *18*(*4*), 637–639. https://doi.org/10.3(201/eid1804.111386.

Parkhill, J., Dougan, G., James, K. D., Thomson, N. R., Pickard, D., Wain, J., Churcher, C., Mungall, K. L., Bentley, S. D., Holden, M. T., Sebaihia, M., Baker, S., Basham, D., Brooks, K., Chillingworth, T., Connerton, P., Cronin, A., Davis, P., Davies, R. M., Dowd, L., White, N., Farrar, J., Feltwell, T., Hamlin, N., Haque, A., Hien, T. T., Holroyd, S., Jagels, K., Krogh, A., Moule, S., O'Gaora, P., Parry, C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., & Barrell,

B. G. (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature, 413*, 848–852.

Austin P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology 44*(*6)*, 398–407.

Nuccio, S.-P. & Baumler, A. J. (2014). Comparative analysis of *Salmonella* genomes identifies a metabolic network for escalating growth in the inflamed gut. *mBio, 5(2)*, e00929-14. https://doi.org/10.1128/mBio.00929-14.

Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics 11*, 855–866.

Friedman, J. H. (1998). Data mining and statistics: What's the connection? *Computer Science and Statistics, 29(1)*, 3–9.

Bishop, C. M. (2006). Pattern recognition and machine learning (Information science and statistics). Berlin, Heidelberg: Springer-Verlag. ISBN 978-0-387-31073-2.

Pornsukarom, S., van Vliet, A., & Thakur, S. (2018). Whole genome sequencing analysis of multiple *Salmonella* serovars provides insights into phylogenetic relatedness, antimicrobial resistance, and virulence markers across humans, food animals and agriculture environmental sources. *BMC Genomics, 19(1)*, 801. DOI: https://doi.org/10.1186/s12864-018-5137-4.

Rakov, A. V., Mastriani, E., Liu, S. L., & Schifferli, D. M. (2019). Association of *Salmonella* virulence factor alleles with intestinal and invasive serovars. *BMC Genomics, 20(1)*, 429. doi: 10.1186/s12864-019-5809-8.

Metris, A., Sudhakar, P., Fazekas, D., Demeter, A., Ari, E., Olbei, M., Branchu, P., Kingsley, R.A., Baranyi, J., Korcsmaros, T. (2017). SalmoNet, an integrated network of ten Salmonella enterica strains reveals common and distinct pathways to host adaptation. *System Biology Application* 3(31).

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Shukla, M., Thomason, 3rd, J. A., Stevens, R., Vonstein, V., Wattam, A. R., & Xia, F. (2015). RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports, 5*, 8365.

Bankevich, A., Nurk, S., Antipov, D., Gurevich,A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19(5)*, 455–477.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Bielzaa, C., Robles, V., & Larranagaa, P. (2011). Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Systems with Applications, 38(5),* 5110–5118.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science, 16,* 199–215.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2,* 18–22.

Matsuki, K., Kuperman, V., & van Dyke, J. A. (2016). The Random Forests statistical technique: An examination of its value for the study of reading. *Scientific Studies of Reading, 20(1),* 20–33.

Friedman, H. (1998). Another approach to polychotomous classification. *Tech. Rep.* Stanford University, Department of Statistics, Stanford, CA 10: 1895–1924.

Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. In: Advances in neural information processing systems. MIT Press, Cambridge, MA.

Yu, H., & Kim, S. (2012). SVM Tutorial — Classification, Regression and Ranking. In: Handbook of Natural Computing (Rozenberg, G., Bäck, T., & Kok, J. N., Eds.). Springer, Berlin, Heidelberg.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5(2),* 197–227.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation,* 256–285.

Ren, Y., Zhang, L., & Suganthan, P. N. (2016). Ensemble classification and regression: recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, *11*, 41–53.

Collet, P., Fonlupt, C., Hao, J. K., Lutton, E., & Schoenauer, M. (Eds.). (2001). Artificial Evolution. 5th International Conference, Evolution Artificielle, EA (2001 Le Creusot, France, October 29-31, 2001.

Velez, D. R., White, B. C., Motsinger, A. A., Bush, W. S., Ritchie, M. D., Williams, S. M., & Moore, J. H. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetics and Epidemiology, 31*, 306–315.

Lu, H., Xu, Y., Ye, M., Yan, K., Gao, Z., & Jin, Q. (2019). Learning misclassification costs for imbalanced classification on gene expression data. *BMC Bioinformatics, 20(25)*, 681.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics, 4*, 31.

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. M. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, *95*, 14–18.

Subramanian, J., & Simon, R. (2013). Overfitting in prediction models – is it a problem only in high dimensions? *Contemporary Clinical Trials, 36*, 636–641.

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics, 38(5),* 404–415.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The elements of statistical learning: Data mining, inference, and prediction. Springer.

Guyon, I., Elisseefi, A., & Kaelbling, L. P. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3(7-8),* 1157–1182.

Lin, F., Sintchenko, V., Kong, F., Gilbert, G. L., & Coiera, E. (2009). Commonly used molecular epidemiology markers of *Streptococcus agalactiae* do not appear to predict virulence. *Pathology, 41(6),* 576–581.

Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetics and Epidemiology, 34(7),* 643–652.

StataCorp (2021). ElasticNet – ElasticNet for prediction and model selection. Retrieved from https://www.stata.com/manuals/lassoelasticnet.pdf. Accessed March 1, 2021.

Saabos, A. (2014). Selecting good features – Part II: linear models and regularization [Blog post]. Retrieved from https://blog.datadive.net/selecting-good-features-part-ii-linear-models-and-regularization/

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research, 45(D1),* D158–D169.

Hoffmann, F. (2001). Boosting: a genetic fuzzy classifier, presented at: IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th. Volume 3.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society series B, 67,* 301–320.

Baker, S., and Dougan, G. (2007). The genome of *Salmonella enterica* serovar Typhi. *Clinical Infectious Diseases, 45*, S29–S33.

Franz, E., Gras, L., & Dallman, T. (2016). Significance of whole genome sequencing for surveillance, source attribution, and microbial risk assessment of foodborne pathogens. *Current Opinion in Food Science, 8*, 74-79.

Baker, R. E., Pena, J. M., Jayamohan, J., & Jerusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters, 14(5).*

Vilne, B., Meistere, I., Grantiņa-Ieviņa, L., & Ķibilds, J. (2019). Machine learning approaches for epidemiological investigations of food-borne disease outbreaks. *Frontiers in Microbiology, 10*, 1722.

Xu, C. M., & Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biology, 20*, 76.

Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems and Applications, 134*, 93–101.

Huang, S., Cai, N., Pacheco, P. P., Nanandres, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) learning in cancer genomics. *Cancer Genomics and Proteomics, 15(1),* 41–51.

Kegerreis, B., Catalina, M. D., Bachali, P., Geraci, N. S., Labonte, A. C., Zeng, C., Stearrett, N., Crandall, K. A., Lipsky, P. E., & Grammer, A. C. (2019). Machine learning approaches to predict lupus disease activity from gene expression data. *Scientific Reports,* 9, 9617.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148(3),* 839–843.

Koller, D., & Sahami, M. (1996). Toward optimal feature selection. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, pp. 284–292.

Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics, 7(3).*

Worley, M. J., Nieman, G. S., Geddes, K., & Heffron, F. (2006). *Salmonella* Typhimurium disseminates within its host by manipulating the motility of infected cells. *Proceedings of the National Academy of Sciences U.S.A., 103(47)*: 17915–17920.

Thornbrough, J. M., & Worley, M. J. (2012). A naturally occurring single nucleotide polymorphism in the *Salmonella* SPI-2 Type III effector srfH/sseI controls early extraintestinal dissemination. *PLoS One, 7(9),* e45245.

Pizza, M. (2006). Proteins and nucleic acids from meningitis/sepsis-associated [patent]. Accessed on: April 5, 2020. Available at: https://patents.justia.com/patent/8758764

Desai, P.T., & McClelland, M. (2013). Integrative analysis of salmonellosis cases in Israel, 1995-(2012 reveals under-reporting of outbreaks, association of serovar i9,12:l,v:-

with extraintestinal salmonellosis and dissemination of endemic *S.* Typhimurium DT104 [unpublished]. Accessed on: April 01, 2020. Available at: https://matrix.bio.anl.gov/pub/CSGID/genomes/90371/GCA_000636135.1_ASM63613v1_genomic.gbff

Maserati, A. (2017). *Salmonella's* desiccation survival and thermal tolerance: genetic, physiological, and metabolic factors [Doctoral dissertation, University of Minnesota]. Accessed on: April 4, 2020. Available at: file:///C:/Users/akplab2/Downloads/Maserati_umn_0130E_18401.pdf.

Hu, Y., Wang, Z., Qiang, B., Xu, Y., Chen, X., Li, Q., & Jiao, X. (2019). Loss and gain in the evolution of the *Salmonella enterica* serovar Gallinarum biovar Pullorum genome. *American Society of Microbiology, 4(2), e00627-18*

Liu, J., Fang, C., Jiang, Y., & Yan, R. (2009). Characterization of a hemolysin gene *ytjA* from *Bacillus subtilis*. *Current Microbiology, 58(6),* 642–647.

Mil-Homens, D., Barahona, S., Moreira, R. N., Silva, I. J., Pinto, S. N., Fialho, A. M., & Arraiano, C. M. (2018). Stress response protein BolA influences fitness and promotes *Salmonella enterica* serovar Typhimurium virulence. *Applied and Environmental Microbiology, 84(8),* e02850-17.

Kim, S., Liu, L., Husain, M., & Vasquez-Torres, A. (2016). Antioxidant defense by thioredoxin can occur independently of canonical thiol-disulfide oxidoreductase enzymatic activity. *Cell Reports, 14(12),* 2901–2911.

Hapfelmeier, S., Ehrbar, K., Stecher, B., Barthel, M., Kremer, M., & Hardt, W. D. (2004). Role of the *Salmonella* pathogenicity island 1 effector proteins SipA, SopB, SopE,

and SopE2 in *Salmonella enterica* subspecies 1 serovar Typhimurium colitis in streptomycin-pretreated mice. *Infection and Immunity, 72(2),* 795–809.

Liu, M., Yan, M., Liu, L., & Chen, S. (2013). Characterization of a novel zinc transporter ZnuA acquired by *Vibrio parahaemolyticus* through horizontal gene transfer. *Frontiers in Cellular and Infection Microbiology, 3(61).*

Singh, A. K., Drolia, R., Bai, X. & Bhunia, A. K. (2015). Streptomycin induced stress response in *Salmonella enterica* serovar Typhimurium shows distinct colony scatter signature. *PLoS One*, e0135035.

Choi, K. M., Kim, M. H., Cai, H., Lee, Y. J., Hong, Y., & Ryu, P. Y. (2018). Salicylic acid reduces OmpF expression, rendering *Salmonella enterica* serovar Typhimurium more resistant to cephalosporin antibiotics. *Chonnam Medical Journal, 54(1),* 17–23.

Fritsch, L., Felten, A., Palma, F., Mariet, J., Radomski, N., Mistou, M., Augustin, J., & Guillier, L. (2018). Insights from genome-wide approaches to identify variants associated to phenotypes at pan-genome scale: application to *L. monocytogenes*' ability to grow in cold conditions. *International Journal of Food Microbiology, 291,* 181–188.

C.D.C. United States Centers for Disease Control and Prevention. (2019c). Reports of *Salmonella* Outbreak Investigations from 2019. Retrieved from U.S. Centers for Disease Control and Prevention. https://www.cdc.gov/salmonella/outbreaks-2019.html. Accessed December 20, 2019.

United States Department of Agriculture – Food Safety Inspection Service (2005). Risk assessments of *Salmonella* Enteritidis in shell eggs and salmonella spp. in egg products. Retrieved from https://www.fsis.usda.gov/wps/wcm/connect/16abcb5f-21fc-4731-9346-fd8401582ba4/SE_Risk_Assess_Oct2005.pdf?MOD=AJPERES. Accessed June 20, 2020.

QMRAWiki (2019). Dose response assessment. Retrieved from http://qmrawiki.org/content/dose-response-assessment. Accessed July 5, 2020.

Buchanan, R. L., Havelaar, A. H., Smith, M. A., Whiting, R.C., & Julien, E. (2009). The key events dose-response framework: its potential for application to foodborne pathogenic microorganisms. *Critical Reviews in Food Science Nutrition 49(8)*, 718–728.

Oscar, T. (2004). Dose-response model for 13 strains of *Salmonella*. *Risk Analysis, 24(1)*, 41–49.

Teunis, P. F. M., Kasuga, F., Fazil, A., Ogden, I. D., Rotariu, O., & Strachen, N. J. C. (2010). Dose–response modeling of *Salmonella* using outbreak data. *International Journal of Food Microbiology 144(2)*, 243–249.

McCullough, N. B., & Eisele, C. W. (1951a). Experimental human salmonellosis. I. Pathogenicity of strains of *Salmonella* meleagridis and Salmonella anatum obtained from spray-dried whole egg. *Journal of Infectious Disease*, 88, 278 – 279.

McCullough, N. B., & Eisele, C. W. (1951b). Experimental human salmonellosis. III. Pathogenicity of strains of Salmonella newport , Salmonella derby , and Salmonella

Bareilly obtained from spray-dried whole egg. *Journal of Infectious Disease*, 89, 209 – 213.

George, R. H. (1976). Small infectious doses of *Salmonella*. *The Lancet, 1130.*

Fontaine, R. E., Arnon, S., Martin, W. T., Vernon Jr., T. M., Gangarosa, E. J., Farmer 3rd, J. J., Moran, A. B., Silliker, J. H., & Decker, D. L. (1978). Raw hamburger: an interstate common source of human salmonellosis. *American Journal of Epidemiology, 107,* 36–45.

Kasuga, F., Hirota, M., Wada, M., Yunokawa, T., Toyofuku, H., Shibatsuji, M., Michino, H., Kuwasaki, T., Yamamoto, S., & Kumagai, S. (2004). Archiving of food samples from restaurants and caterers—quantitative profiling of outbreaks of foodborne salmonellosis in Japan. *Journal of Food Protection, 67,* 2024–2032.

De Mol, C., de Vito, E., Rosascode, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity, 25(2),* 201–230.

Louwen, R., Staals, R. H. J., Endtz, H. P., van Baarlen, P., & van der Oost, J. (2014). The Role of CRISPR-Cas systems in virulence of pathogenic bacteria. *Microbiology and Molecular Biological Reviews, 78(1),* 74–88.

Candes, E., & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist. 35(6),* 2313–2351.

Hastie, T. J., Tibshirani, R. J. & Wainwright, M. (2015). Statistical learning with sparsity: the lasso and generalizations. Boca Raton, FL: CRC Press.

Meynell, G. G. & Meynell, E. W. (1958). The growth of micro-organisms in vivo with particular reference to the relation between dose and latent period. *Epidemiology & Infection*, 56.

Anderson, C. J. (2019). Poisson regression for regression of counts and rates. Retrieved from https://education.illinois.edu/docs/default-source/carolyn-anderson/edpsy589/lectures/4_glm/4glm_3_beamer_post.pdf. Accessed February 25, 2021.

Dataquest. (2019). Tutorial: Poisson regression in R. Retrieved from https://www.dataquest.io/blog/tutorial-poisson-regression-in-r/. Accessed February 25, 2021.

Lesko, C. R., Jacobson, L. P., Althoff, K. N., Abraham, A. G., Gange, S. J., Moore, R. D., Modur, S., & Lau, B. (2018). Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. *International Journal of Epidemiology, 47(2),* 654–668.

Grad, Y. H. & Lipsitch, M. (2014). Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. *Genome Biology, 15(11),* 538.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. J., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics, 31(22),* 3691–3693.

Chesaniuk, M. (2021). Chapter 19: Logistic and Poisson regression. Retrieved from https://ademos.people.uic.edu/Chapter19.html. Accessed February 28, 2021.

Ritchie, M.E., Diyagama, D., Neilson, J., van Laar, R., Dobrovic, A., Holloway, A., & Smyth, G. K. (2006). Empirical array quality weights in the analysis of microarray data. *BMC Bioinformatics, 7,* 261.

Burgess, S., & Bowden, J. (2015). Integrating summarized data from multiple genetic variants in Mendelian randomization: Bias and coverage properties of inverse-variance weighted methods. *Annals of Applied Statistics* [submitted].

Reifeis, S. A., Hudgens, M. G., Civelek, M., Mohlke, K. L., Love, M. I. (2020). Assessing exposure effects on gene expression. *Genetic Epidemiology, 44(6),* 601–610.

Janss, L., los Campos, G., Sheehan, N., & Sorensen, D. (2012). Inferences from genomic models in stratified populations. *Genetics, 192(2),* 693–704.

Worley, M. J., Nieman, G. S., Geddes, K., & Heffron, F. (2006). *Salmonella* Typhimurium disseminates within its host by manipulating the motility of infected cells. *Proceedings of the National Academy of Sciences U.S.A., 103(47),* 17915–17920.

Oleszak, M. (2019). Regularization: Ridge, Lasso, and Elastic Net. Retrieved from https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net. Accessed on July 10, 2020.

Blaby-Haas, C. E., Flood, J. A., Crecy-Lagard, V., & Zamble, D. B. (2012). YeiR: a metal binding GTPase from *Escherichia coli* involved in metal homeostasis. *Metallomics, 4(5),* 488–497.

Palmer, L. D. & Skaar, E. P. (2016). Transition metals and virulence in bacteria. *Annual Reviews in Genetics, 50,* 67–91.

Stevens, J. M., Mavridou, D. A. I., Hamer, R., Kritsiligou, P., Goddard A. D., & Ferguson, S. J. (2011). Cytochrome c biogenesis system I. *The FEBS Journal, 278(22),* 4170–4178.

Hough, M. A., Silkstone, G., Worrall, J. A. R., & Wilson, M. T. (2014). Chapter 8 – NO binding to the proapoptotic cytochrome-c cardiolipin complex. *Vitamins and Hormones, 96,* 193–209.

Tobin, C., Mandava, C. S., Ehrenberg, M., Andersson, D. I., & Sanyal, S. (2010). Ribosomes lacking protein S20 are defective in mRNA binding and subunit association. *Journal of Molecular Biology, 397(3),* 767–776.

Cuthbertson, L., Mainprize, I. L., Naismith, J. H., & Whitfield, C. (2010). Pivotal roles of the outer membrane polysaccharide export and polysaccharide copolymerase protein families in export of extracellular polysaccharides in gram-negative bacteria. *Microbiology and Molecular Biology Reviews, 73(1),* 155–177.

Morais, V., Dee, V., & Suarez, N. (2018). Purification of capsular polysaccharides of *Streptococcus pneumoniae*: traditional and new methods. *Frontiers in Bioengineering and Biotechnology, 6,* 145.

Domka, J., Lee, J., & Wood, T. K. (2006). YliH (BssR) and YceP (BssS) regulate *Escherichia coli* K-12 biofilm formation by influencing cell signaling. *Applied and Environmental Microbiology, 72,* 2449–2459.

Heikal, A., Nakatani, Y., Dunn, E.,Weimar, M. R., Day, C. L., Baker, E. N., Lott, S., Sazanov, L. A., & Cook, G. M. (2014). Structure of the bacterial type II NADH dehydrogenase: a monotopic membrane protein with an essential role in energy generation. *Molecular Microbiology, 9(5),* 950–964.

Huddleston, J. P., Thoden, J. B., Dopkins, B. J., Narindoshvili, T., Fose, B. J., Holden, H. M., & Raushel, F. M. (2019). Structural and functional characterization of YdjI, an aldolase of unknown specificity in *Escherichia coli* K12. *Biochemistry, 58(31),* 3340–3353.

Amato, S. M., Fazen, C. H., Henry, T. C., Mok, W. W. K., Orman, M. A., Sandvik, E. L., Volzing, K. G., & Brynildsen, M. P. (2014). The role of metabolism in bacterial persistence. *Frontiers in Microbiology, 5,* 70.

Passalacqua, K. D., Charbonneau, M.-E., & O'Riordan, M. X. D. (2016). Bacterial metabolism shapes the host:pathogen interface. *Microbial Spectrum, 4(3),* 10.

QMRAWiki. (2021). Dose response – Recommended best-fit parameters. Retrieved from http://qmrawiki.org/content/recommended-best-fit-parameters?page=1. Accessed February 21, 2021.

Munnoch, S. A., Ward, K., Sheridan, S., Fitzsimmons, G. J., Shadbolt, C. T., Piispanen, J. P., Wang, Q., Ward, J. T., Worgan, T. L. M., Oxenford, C., Musto, J. A., McAnulty, J., & Durrheim, D. N. (2009). A multi-state outbreak of *Salmonella Saintpaul* in

Australia associated with cantaloupe consumption. *Epidemiology and Infection, 137*, pp.367–374.

Sidhu, J. P. S., Ahmed, W., Gernjak, W., Aryal , R., McCarthy, D., Palmer, A., Kolotelo, P., & Toze, S. (2013). Sewage pollution in urban stormwater runoff as evident from the widespread presence of multiple microbial and chemical source tracking markers. *Science of the Total Environment, 463–464*, pp.488–96.

Mun, S. G. (2020). The effects of ambient temperature changes on foodborne illness outbreaks associated with the restaurant industry. *International Journal of Hospitality Management, 85*, 102432.

McMichael, A. (2015). Extreme weather events and infectious disease outbreaks. *Virulence, 6(6)*, pp. 543–547.

Heim, R. R., Jr. (1996). An overview of the 1961–90 climate normals products available from NOAA's National Climatic Data Center. Preprints, *22nd Conf. on Agricultural and Forest Meteorology*, Atlanta, GA, Amer. Meteor. Soc., 193–196.

Owen, T. W., & Whitehurst, T. (2002). United States climate normals for the 1971–2000 period: Product descriptions and applications. Preprints, *Third Symp. on Environmental Applications: Facilitating the Use of Environmental Information*, Orlando, FL, Amer. Meteor. Soc., J4.3. [Available online at https://ams.confex.com/ams/annual2002/webprogram/Paper26747.html.]

Arguez, A., Durre, I., Applequist, S., Vose, R. S., Squires, M. F., Yin, X., Heim, R., Jr., & Owen, T. W. (2012). NOAA's 1981–2010 U.S. Climate Normals: An overview. *Bulletin of the American Meteorological Society, **93***, 1687–1697.

Durre, I., Squires, M. F., Vose, R. S., Yin, X., Arguez, A., & Applequist, S. (2013). NOAA's 1981–2010 U.S. Climate Normals: Monthly Precipitation, Snowfall, and Snow Depth. *Journal of Applied Meteorology and Climatology, 52(11),* 2377–2395.

Shirriff, V. E. (2019). Impacts Of Ambient Temperature On Foodborne Salmonella Infection. Public Health Theses. Retrieved from https://elischolar.library.yale.edu/ysphtdl/1845. Accessed September 6, 2021.

Ellis, M. J., Trussler, R. S., Charles, O., & Haniford, D. B. (2017). A transposon-derived small RNA regulates gene expression in *Salmonella* Typhimurium. *Nucleic acids research, 45*(9), 5470–5486.

Saini, A., Mapolelo, D. T., Chahal, H. K., Johnson, M. K., & Outten, F. W. (2010). SufD and SufC ATPase activity are required for iron acquisition during in vivo Fe-S cluster formation on SufB. *Biochemistry, 49,* pp. 9402–9412.

Zatyka M. & Thomas, C. M. (1998). Control of genes for conjugative transfer of plasmids and other mobile elements, *FEMS Microbiology Reviews, 21(4),* 291–319.

Maurer, R., Osmond, B. C., Shekhtman, E., Wong, A., & Botstein, D. (1984). Functional interchangeability of DNA replication genes in Salmonella typhimurium and Escherichia coli demonstrated by a general complementation procedure. *Genetics, 108*(1), pp. 1–23.

Deutscher, J., Francke, C., & Postma, P. W. (2006). How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. *Microbiology and Molecular Biology Reviews*:*MMBR*, *70(4)*, pp. 939–1031.

Zhi, Y., Lin, S. M., Ahn, K. B., Ji, H. J., Guo, H. C., Ryu, S., Seo, H. S., & Lim, S. (2020). ptsI gene in the phosphotransfer system is a potential target for developing a live attenuated Salmonella vaccine. *International Journal of Molecular Medicine*, *45(5)*, pp. 1327–1340.

Liu, Y., Ho, K. K., Su, J., Gong, H., Chang, A. C., & Lu, S. (2013). Potassium transport of Salmonella is important for type III secretion and pathogenesis. *Microbiology (Reading, England)*, *159(Pt 8)*, pp. 1705–1719.

Gulati, A., Shukla, R., & Mukhopadhyaya, A. (2019). *Salmonella* effector SteA suppresses proinflammatory responses of the host by interfering with IκB degradation. *Frontiers in Immunology*.

Holtje, J.V., Mirelman, D., Sharon, N., and Schwarz, U. (1975). Novel type of murein transglycosylase in *Escherichia coli*. *Journal of Bacteriology*, *124*, pp.1067–1076.

Vilhena, C., Kaganovitch, E., Shin, J. Y., Grünberger, A., Behr, S., Kristoficova, I., Brameyer, S., Kohlheyer, D., & Jung, K. (2018). A single-cell view of the BtsSR/YpdAB pyruvate sensing network in *Escherichia coli* and its biological relevance. *Journal of Bacteriology*, *200*(1).

Kawano, Y., Ohtsu, I., Tamakoshi, A., Shiroyama, M., Tsuruoka, A., Saiki, K., Takumi, K., Nonaka, G., Nakanishi, T., Hishiki, T., Suematsu, M., & Takagi, H. (2014).

Involvement of the yciW gene in L-cysteine and L-methionine metabolism in *Escherichia coli*. *Journal of Bioscience and Bioengineering, 119*.

Born, F., Braun, P., Scholz, H. C., & Grass, G. (2020). Specific detection of Yersinia *pestis* based on receptor binding proteins of phages. *Pathogens, 9(8),* 611.

Danckert, L., Hoppe, S., Bier, F. F., & von Nickisch-Rosenegk, M. (2014). Rapid identification of novel antigens of *Salmonella* Enteritidis by microarray-based immunoscreening. *Mikrochimica Acta, 181(13-14),* pp.1707–1714.

Haendiges, J., Brown, E. W., Tikekar, R., & Hoffmann, M. (2021). A comprehensive genomic analysis of the CHASRI in *Salmonella enterica* isolated from food and environmental sources. Retrieved from download (fda.gov). Accessed 27 September, 2021.

Postma, P. W., & Stock, J. B. (1980). Enzymes II of the phosphotransferase system do not catalyze sugar transport in the absence of phosphorylation. *Journal of bacteriology, 141(2),* pp.476–484.

Pelludat, C., Mirold, S., & Hardt, W.D. (2003).The SopEΦ Phage integrates into the ssrA gene of *Salmonella enterica* serovar Typhimurium A36 and is closely related to the Fels-2 prophage. *Journal of Bacteriology, 185(17),* pp.5182-5191.

Dunne, M., Denyes, J. M., Arndt, H., Loessner, M. J., Leman, P. G., & Klumpp, J. (2018). *Salmonella* phage S16 tail fiber adhesin features a rare poly-glycine rich domain for host recognition. *Structure, 26(12),* pp.1573–1582.

Paiva, J. B., Penha Filho, R. A., Junior, A. B., & Lemos, M. V. (2011). Requirement for cobalamin by Salmonella enterica serovars Typhimurium, Pullorum, Gallinarum

and Enteritidis during infection in chickens. *Brazilian journal of microbiology : [publication of the Brazilian Society for Microbiology]*, *42*(4), pp.1409–1418.

Singh, A. K., Drolia, R., Bai, X., & Bhunia, A. K. (2015). Streptomycin induced stress response in *Salmonella enterica* serovar Typhimurium shows distinct colony scatter signature. *PLoS One*

Nickerson, C. A., & Curtiss, R., 3rd (1997). Role of sigma factor RpoS in initial stages of *Salmonella* Typhimurium infection. *Infection and Immunity*, *65(5)*, pp.1814–1823.

Frost, L. S., Ippen-Ihler, K., & Skurray, R. A. (1994). Analysis of the sequence and gene products of the transfer region of the F sex factor. *Microbiological Reviews*, *58*, pp.162–210.

Paterson, G. K., Northen, H., Cone, D. B., Willers, C., Peters, S. E., & Maskell, D. J. (2009). Deletion of TolA in *Salmonella* Typhimurium generates an attenuated strain with vaccine potential. *Microbiology (Reading, England)*, *155(Pt 1)*, pp.220–228.

Runti, G., Lopez Ruiz, M., Stoilova, T., Hussain, R., Jennions, M., Choudhury, H. G., Benincasa, M., Gennaro, R., Beis, K., & Scocchi, M. (2013). Functional characterization of SbmA, a bacterial inner membrane transporter required for importing the antimicrobial peptide Bac7(1-35). *Journal of Bacteriology*, *195(23)*, pp.5343–5351.

Edmonds, L., Liu, A., Kwan, J. J., Avanessy, A., Caracoglia, M., Yang, I., ... & Donaldson, L. W. (2007). The NMR structure of the gpU tail-terminator protein from bacteriophage lambda: identification of sites contributing to Mg (II)-mediated

oligomerization and biological function. *Journal of molecular biology*, *365*(1), pp.175-186.

Groth, A. C., & Calos, M. P. (2004). Phage integrases: biology and applications. *Journal of molecular biology, 335(3),* pp.667–678.

Colavecchio, A., D'Souza, Y., Tompkins, E., Jeukens, J., Freschi, L., Emond-Rheault, J. G., Kukavica-Ibrulj, I., Boyle, B., Bekal, S., Tamber, S., Levesque, R. C., & Goodridge, L. D. (2017). Prophage integrase typing is a useful indicator of genomic diversity in *Salmonella enterica*. *Frontiers in Microbiology*, *8*, 1283.

Büttner, C. R., Wu, Y., Maxwell, K. L., & Davidson, A. R. (2016). Baseplate assembly of phage Mu: Defining the conserved core components of contractile-tailed phages and related bacterial systems. *Proceedings of the National Academy of Sciences of the United States of America, 113(36),* pp.10174–10179.

Gerstel U, Park C, & Römling, U. (2003). Complex regulation of csgD promoter activity by global regulatory proteins. *Molecular Microbiology, 49,* pp.639–654.

Shen, S., & Fang, F. C. (2012). Integrated stress responses in *Salmonella*. *International journal of food microbiology*, *152(3),* pp.75–81.

Jung, H., Buchholz, M., Clausen, J., Nietschke, M., Revermann, A., Schmid, R., Jung, K. (2002). CaiT of *Escherichia coli*, a new transporter catalyzing l-carnitine/γ-butyrobetaine exchange. *Membrane Transport Structure Function and Biogenesis, 277(42),* pp.39251–39258

Payne, P. L. & Straley, S. C. (1998). YscO of *Yersinia pestis* is a mobile core component of the Yop secretion system. *Journal of Bacteriology, 180(15).*

Evans, L. D., & Hughes, C. (2009). Selective binding of virulence type III export chaperones by FliJ escort orthologues InvI and YscO. *FEMS microbiology letters*, *293*(2), pp.292–297.

Ellermeier, C. D., & Slauch, J. M. (2003). RtsA and RtsB coordinately regulate expression of the invasion and flagellar genes in *Salmonella enterica* serovar Typhimurium. *Journal of Bacteriology*, *185(17)*, pp.5096–5108.

Liu, G. F., Wang, X. X., Su, H. Z., & Lu, G. T. (2021). Progress on the GntR family transcription regulators in bacteria. *Yi chuan Hereditas*, *43(1)*, pp.66–73.

Pasqua, M., Grossi, M., Zennaro, A., Fanelli, G., Micheli, G., Barras, F., Colonna, B., & Prosseda, G. (2019). The varied role of efflux pumps of the MFS family in the interplay of bacteria with animal and plant cells. *Microorganisms*, *7(9)*, 285.

Interpro (2021). ABC transporter type 1, transmembrane domain MetI-like [IPR000515]. ABC transporter type 1, transmembrane domain MetI-like (IPR000515) - InterPro entry - InterPro (ebi.ac.uk).

Anes, J., McCusker, M. P., Fanning, S., & Martins, M. (2015). The ins and outs of RND efflux pumps in *Escherichia coli*. *Frontiers in microbiology*, *6*, 587.

Obuchowski, M., Giladi, H., Koby, S., Szalewska-Pałasz, A., We, A., Oppenheim, A. B., ... & We, G. (1997). Impaired lysogenisation of the Escherichia coli rpoA341 mutant by bacteriophage λ is due to the inability of CII to act as a transcriptional activator. *Molecular and General Genetics MGG*, *254*(3), pp.304–311.

Smith, C. M., Arany, Z., Orrego, C., & Eisenstadt, E. (1997). DNA damage-inducible loci in *Salmonella* Typhimurium. *Journal of Bacteriology*, *173(11)*.

Schneider, E. & Hunke, S. ATP-binding-cassette (ABC) transport systems: Functional and structural aspects of the ATP-hydrolyzing subunits/domains. *FEMS Microbiology Reviews, 22(1),* pp.1–20.

Andres, D., Baxa, U., Hanke, C., Seckler, R., & Barbirz, S. (2010). Carbohydrate binding of *Salmonella* phage P22 tailspike protein and its role during host cell infection. *Biochemical Society Transactions, 38(5),* pp.1386–1389.

Massey, R. C., Bowe, F., Sheehan, B. J., Dougan, G., & Dorman, C. J. (2000). The virulence plasmid of *Salmonella* typhimurium contains an autoregulated gene, rlgA, that codes for a resolvase-like DNA binding protein. *Plasmid, 44*(1), pp.24–33.

Rychlik, I., Gregorova, D., & Hradecka, H. (2006). Distribution and function of plasmids in *Salmonella enterica*. *Veterinary microbiology*, *112*(1), pp.1–10.

Fane, B. A., Brentlinger, K. L., Burch, A. D., Chen, M., Hafenstein, S. U. S. A. N., Moore, E. R. I. C. A., ... & Uchiyama, A. S. A. K. O. (2006). wX174 et al., the Microviridae. In *The Bacteriophages* (pp. 129-145). Oxford: Oxford University Press.

Rose, J. B., Epstein, P. R., Lipp, E. K., Sherman, B. H., Bernard, S. M., Patz, J. A. (2001). Climate variability and change in the United States: potential impacts on water- and foodborne diseases caused by microbiologic agents. *Environmental Health Perspectives, 109,* pp.211.

Simental, L., & Martines-Urtaza, J. (2008). Climate patterns governing the presence and permanence of salmonellae in coastal areas of Bahia de Todos Santos, Mexico. *Applied and Environmental Microbiology, 74,* pp. 5918–5924.

Singh, R. B., Hales, S., de Wet, N., Raj, R., Hearnden, M., & Weinstein, P. (2001). The influence of climate variation and change on diarrheal disease in the Pacific Islands. *Environmental Health Perspectives, 109,* pp. 155–159.

Akil, L., Ahmad, H. A., & Reddy, R. S. (2014). Effects of climate change on Salmonella infections. *Foodborne Pathogens and Disease*, *11*(*12*), 974–980. https://doi.org/10.1089/fpd.2014.1802

Pin, C., Hansen, T., Muñoz-Cuevas, M., de Jonge, R., Rosenkrantz, J. T., Löfström, C., Aarts, H., & Olsen, J. E. (2012). The transcriptional heat shock response of *Salmonella* Typhimurium shows hysteresis and heated cells show increased resistance to heat and acid stress. *PLoSOne, 7(12),* pp.e51196.

Dawoud, T. M., Davis, M. L., Park, S. H., Kim, S. A., Kwon, Y. M., Jarvis, N., O'Bryan, C. A., Shi, Z., Crandall, P. G., Ricke, S. C. (2017). The potential link between thermal resistance and virulence in *Salmonella*: a review. *Frontiers in Veterinary Science, 4*.

Soneja, S., Jiang, C., Upperman, C. R., Murtugudde, R., Mitchell, C. S., Blythe, D., Sapkota, A. R., & Sapkota, A. (2016). Extreme precipitation events and increased risk of campylobacteriosis in Maryland, U.S.A. *Environmental Research, 149*, pp. 216–221. doi: 10.1016/j.envres.2016.05.021.

Holley, R., Walkty, J., Blank, G., Tenuta, M., Ominski, K., Krause, D., Ng, L. K. (2008). Examination of *Salmonella* and *Escherichia coli* translocation from hog manure to forage, soil, and cattle grazed on the hog manure-treated pasture. *Journal of Environmental Quality, 37(6),* 2083–2092.

Piekarska, K. (2010). Mutagenicity of airborne particulates assessed by *Salmonella* assay and the SOS Chromotest in Wrocław, Poland. *Journal of the Air & Waste Management Association, 60(8),* pp. 993–1001.

Semenov, A. V., van Bruggen, A. H. C., van Overbeek, L., Termorshuizen, A. J., & A. M. Semenov. (2007). Influence of temperature fluctuations on Escherichia coli O157:H7 and Salmonella enterica serovar Typhimurium in cow manure. *FEMS Microbiology Ecology, 60(3),* 419–428.