Contents lists available at ScienceDirect

Current Research in Food Science

journal homepage: www.sciencedirect.com/journal/current-research-in-food-science

Machine learning to predict foodborne salmonellosis outbreaks based on genome characteristics and meteorological trends

Shraddha Karanth [a], Jitendra Patel [b], Adel Shirmohammadi [c], Abani K. Pradhan [a,d,*]

[a] Department of Nutrition and Food Science, University of Maryland, College Park, MD, 20742, USA
[b] Environmental Microbial & Food Safety Lab, USDA-ARS, Beltsville, MD, 20705, USA
[c] Environmental Science & Technology, University of Maryland, College Park, MD, 20742, USA
[d] Center for Food Safety and Security Systems, University of Maryland, College Park, MD, 20742, USA

ARTICLE INFO

Keywords:
Salmonellosis
Meteorological variables
Whole genome sequencing
Elastic net
Outbreaks

ABSTRACT

Several studies have shown a correlation between outbreaks of *Salmonella enterica* and meteorological trends, especially related to temperature and precipitation. Additionally, current studies based on outbreaks are performed on data for the species *Salmonella enterica*, without considering 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 scale (typified by case numbers) using a combination of machine learning and count-based modeling methods. Elastic Net regularization model 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 ($\alpha = 0.50$; $\lambda = 2.18$) identified 53 significant gene features. The final multi-variable Poisson regression model ($\chi^2 = 5748.22$; pseudo $R^2 = 0.669$; probability $> \chi^2 = 0$) identified 127 significant predictor terms ($p < 0.10$), comprising 45 gene-only predictors, average temperature, average precipitation, and average snowfall, and 79 gene-meteorological interaction terms. The significant genes ranged in functionality from cellular signaling and transport, virulence, metabolism, and stress response, and included gene variables not considered as significant by the baseline model. This study presents a holistic approach towards evaluating multiple data sources (such as genomic and environmental data) to predict outbreak scale, which could help in revising the estimates for human health risk.

1. Introduction

Salmonella enterica subsp. *enterica*, a facultative anaerobic bacteria, is a leading cause of foodborne illness worldwide. Several statistics have shown that foodborne *Salmonella* exerts considerable impact on public health and mortality rates, affecting an estimated 1 million people, with 23,128 hospitalizations, and 452 deaths in the U.S. annually (Scallan et al., 2011). Despite ongoing efforts to curb the spread and proliferation of this bacteria, its ubiquitous nature, considerable within-species diversity, and horizontal gene transmission of virulence genes from traditionally more pathogenic to less- or non-pathogenic serovars has resulted in a significant increase in the number of salmonellosis cases being reported both in the U.S. and globally (CDC, 2021; Scallan et al., 2011). Among the foodborne routes, salmonellosis outbreaks are increasingly being attributed to non-meat and -poultry sources, such as produce, ready-to-eat foods, oils and grains, and bakery products. Factors such as the prevalent environmental conditions and farm practices, as well as the presence of vectors such as livestock and wildlife, untreated manure, level of crop maturity, presence of native biota that may promote or inhibit the growth of human pathogens, inadequacies in food handling practices, and water quality could contribute to the proliferation of *Salmonella* pre- and post-harvest, which, in turn, would lead to increased human foodborne exposure (Ehuwa et al., 2021). Moreover, *Salmonella* covers a diverse genetic landscape, with *Salmonella enterica* subsp. *enterica* alone comprising >2500 named serovars. Currently, models predicting bacterial infection outcome and outbreak scale do not account for intra-species variability in microbial (specifically *Salmonella*) behavior because, for the most part, variabilities existing at the gene-level are too large in scale to be incorporated in basic statistical models. Molecular analyses of isolates could provide us with information regarding the expression of genes associated with virulence and survival in isolates under various conditions of isolation (Adzitey et al.,

[*] Corresponding author. Department of Nutrition and Food Science, University of Maryland, 0112 Skinner Building, College Park, MD, 20742, USA.
E-mail address: akp@umd.edu (A.K. Pradhan).

https://doi.org/10.1016/j.crfs.2023.100525
Received 26 December 2022; Received in revised form 15 May 2023; Accepted 27 May 2023
Available online 28 May 2023

2020). Additionally, the presence/absence of genes (and its frequency) associated with stress tolerance, virulence, and antibiotic resistance in *Salmonella* could help in developing a differential virulence profile that could aid in re-evaluating the existing infectivity and outbreak predictive estimates for *Salmonella enterica*. For example, the genes associated with biofilm production (*adrA*, *bapA*), virulence (*hilA*, *invA*, *invC, invG*, *prgH*), and temperature stress (*rpoS*, *rpoE*, and *rpoH*) have been previously associated with exposure of different serovars of *Salmonella* such as Enteritidis and Tyhpimurium to stressful temperature and pH conditions (Sirsat et al., 2015; Badie et al., 2021).

Several studies have investigated the impact of environmental factors, specifically temperature and precipitation, on the incidence of *Salmonella*-associated foodborne outbreaks (Stephen and Barnett, 2016; Shirriff, 2019). 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* infections, modulating outbreak and illness risk estimates. Studies have shown how varied combinations of ambient temperatures and precipitation levels, and the resultant changes in animal (e.g. *Salmonella* shedding) and human behavior (e.g. recreational activities) and food habitats contributes to salmonellosis infections in the population (Mun, 2020; Munnoch et al., 2009; Sidhu et al., 2013). 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 (particularly at temperatures 30°C and higher), as it supports the growth of *Salmonella* (Yun et al., 2016). Similarly, increasing precipitation levels are also believed to increase the risk of salmonellosis incidence, as runoff can increase pathogen loads in water sources, which would get distributed to a wider area and create conditions (high water activity) that promote the growth of bacteria (Stephen and Barnett, 2016). 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 et al., 2019a), *Escherichia coli* (Njage et al., 2019b; Pielaat et al., 2015), and *Salmonella* (Karanth et al., 2022; Tanui et al., 2022). An important contribution of this new wave 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. This method helps with issues such as model overfitting or bias introduction due to $p >> n$ (much larger number of predictors compared to number of samples).

The objective of this study was to develop a machine learning-based regression approach to quantify the interaction effects between meteorological factors (such as temperature and precipitation) and genes that might be involved in a *Salmonella enterica* strain's response to environmental stressors within outbreak situations. This, in turn, would help us predict the most significant combination of genes and meteorological factors that contribute to the incidence of foodborne outbreaks of salmonellosis.

2. Materials and methods

2.1. Data collection

2.1.1. 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; https://www.cdc.gov/fdoss/annual-reports/index.html). For this study, only data from *Salmonella* outbreaks definitively associated with a food source (i.e., foodborne salmonellosis) were included, with sporadic cases being excluded completely. The latter was excluded since it would be difficult to find correlating *Salmonella* isolates, leading to an incomplete dataset. The NORS toolkit contains a comprehensive list of outbreaks attributed to different etiological agents that have occurred 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. Stringent inclusion criteria, such as the availability of serovar data and U.S. state wherein outbreak occurs, and complete number of illnesses per outbreak, were applied to identify relevant and complete data points.

2.1.2. Meteorological data

Meteorological data were 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 (Arguez et al., 2012; Durre et al., 2013; Heim 1996; Owen and Whitehurst 2002). In this study, data was obtained in the form of monthly averages.

NOAA measures climatological data with the aid of numerous weather stations spread across the U.S. Incorporating data from all stations within a state of interest would allow us to incorporate variations in weather conditions seen across the state, specifically those with a larger land mass. In this study, the monthly average temperature, monthly average precipitation, and monthly average snowfall was obtained from all weather stations within each state of interest from NCEI. Average values were taken across all weather stations within each state, and the values standardized using the state-specific mean and standard deviation for each month and year of interest. In essence, for each observation, meteorological data was the average of data obtained from all weather stations in the respective state during the month of the outbreak. 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. In this study, however, the results are reported in °C and cm, respectively.

2.1.3. Salmonella isolates for development of environmental Salmonella 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. Inclusion criteria set for isolate selection included the availability of 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. Multiple isolates were selected from across various sources (corresponding to each outbreak data point) in order to incorporate the genetic variations observable within and between serovars, and in various isolate environments. 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 obtained to create our pan genome (gene dictionary).

2.2. WGS pre-processing

2.2.1. WGS assembly and annotation

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 web-based Bacterial and Viral Bioinformatics Resource Center (BV-BRC; formerly known as PATRIC (v.3.6.3) Bacterial Bioinformatics Resource Center). The in-built SPAdes (Bankevich et al., 2012) assembler was used for genome assembly and the Rapid Annotation using Subsystems Technology (RASTk)-enabled genome annotation service (Brettin et al., 2015) was employed for genome annotation. Although many of the included isolates had available WGS in the NCBI Genomes repository, all WGS were assembled and annotated on PATRIC for uniformity. Isolates (n = 497) that fit the quality parameters of good sequence quality according to QUAST statistics (Gurevich et al., 2013), had complete sequence information, and returned a sufficient annotation score were included in the dataset for pan-genome creation.

2.2.2. *Salmonella pan-genome creation*

In order to obtain important predictor variables for use in our model, a “dictionary” of genes and gene homologs was developed from the annotated sequences. A dictionary is, in simple terms, 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 et al., 2009). In the case where input functions are individual, unique genes from a number of samples, this dictionary refers to the pan-genome. The pan-genome was developed by aligning nucleotide sequences all-against-all using *pairwise2* in Python as described previously (Karanth et al., 2022). All-against-all basic local alignment search tool (BLAST) is an established method to search for homologous pairs of sequences in a database. 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, despite potentially contributing to the virulence and pathogenicity potential of *Salmonella* (Louwen et al., 2014). This generated a dictionary/pan-genome of 18,520 unique genes, including potential gene homologs, which were nevertheless assumed to be heterologous and included as predictors in the initial model.

2.3. *Model development and statistical analysis*

Here, the individual and combined effects of the predictor variables gene presence/absence (categorical; 1 or 0), mean daily average temperature (in °C) (continuous), precipitation (in cm), and snowfall average (in cm) on the response variable (number of illnesses per outbreak) was modeled. All models were run with standardized meteorological variables (averaged across each individual state, as described in 2.1.2) recorded during the month of an outbreak (no lag) and two months before an outbreak (two-month lag). The latter analysis (two-month lag) was performed to determine the delayed effect of weather factors, particularly temperature, on outbreak outcome, since prior studies have described how salmonellosis risk appears to be highest 2–6 weeks after exposure to elevated temperatures (Robinson et al., 2022). All statistical analyses and modeling were performed on STATA 16 (StataCorp, 2019). Model significance was set at $p < 0.05$, and predictor significance was tested at both $p < 0.05$ and 0.10.

2.3.1. *Feature selection: identification of highly predictive genes*

The size of the predictor matrix (p = 18,520; i.e., the number of predictor variables) in our model would be much larger than the number of samples (n = 497), which is also known as the *p » n* problem. As such, our model could suffer from predictive capacity, dimensionality issues, and model overfitting (Candes and Tao, 2007). Additionally, genes coding for the same biological pathway tend to be highly correlated (Zou and Hastie, 2005), which would also impact the predictive capacity of the model. Model quality and interpretation can be improved by employing penalization techniques that would identify fewer significant, highly predictive variables, which could be employed in the model. In this study, this was performed using Elastic Net, a powerful method used to automatically reduce (or ‘shrink’) the number of non-discriminative (or non-informative) features within the dictionary, while selecting groups of correlated variables that add significantly to the model (Zou and Hastie, 2005). In our study, the penalized objective function for Elastic Net is

$$Q=\frac{1}{N}\sum_{i=1}^{N}w_if(y_i,\beta_0+x_i\beta^{'})+\lambda\sum_{j=1}^{p}\kappa_j\left(\left(\frac{1-\alpha}{2}\right)\beta_j^2+\alpha|\beta_j|\right) \tag{1}$$

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 and controls the amount of shrinkage, 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 and controls the type of shrinkage. The estimated β minimizes the penalized objective function 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 (equation and explanation adapted from: StataCorp, 2021).

The functional form for the function *f()* used in a linear (ordinary least squares) and Poisson (or other count model, such as negative binomial) model are provided in equations (2) and (3), respectively.

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

$$f(y_i,\beta_0+x_i\beta)=-y_i(\beta_0+x_i\beta^{'})+e^{(\beta_0+x_i\beta^{'})} \tag{3}$$

Elastic Net regularization was performed on the complete dataset on STATA using the *elasticnet* function. In this study, the default α values (1, 0.75, and 0.5) and a fine grid of auto-generated λ values were tested, according to Hastie et al. (2015). The λ grid is set automatically during the run. The best (α, λ) pair was selected by 10-fold cross validation; in essence, cross-validation was employed to determine the predictive performance of our model, wherein a part of the dataset with complete information was used to estimate the generalizability of the model in the absence of external data. The (α, λ) pair that minimized the value of the cross validation function was selected, and the significant non-zero coefficients identified by this (α, λ) pair were employed in further models as independent predictor variables (StataCorp, 2021).

2.3.2. *Poisson regression*

A Poisson regression model was developed to explain the outcome of outbreak case numbers (count data; response variable), with gene presence/absence as the primary predictors, and meteorological factors as the covariates. The Poisson model, a count-based regression model, was selected primarily because our outcome (or dependent) variable (illness case numbers) is a numeric count with a limited positive value range compared to a continuous variable (Chesaniuk, 2021). Count-based regression models can handle these characteristics of counts as a dependent variable and do so by using a log-link, thus modeling the log of the count (Weisburd et al., 2021). Simply put, our model is structured as:

$$\Pr(Y_i=y_i|\mu_i,t_i)=\frac{e^{-\mu_it_i}(\mu_it_i)^{y_i}}{y_i!} \tag{4}$$

Where,

$$\mu_i=t_ie^{(\beta_1X_{1i}+\beta_2X_{2i}+\ldots\ldots+\beta_kX_{ki})} \tag{5}$$

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 temperature, monthly average precipitation, and monthly average snowfall – and their interaction terms, $\beta =$ 1 … 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). 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 or shrunken values being automatically eliminated from the model). The model fit was determined by analyzing the pseudo R^2.

2.3.3. Negative binomial regression

Since a Poisson regression makes a restrictive assumption that the mean is equal to the variance, the data was also fitted to a second count-based model, 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:

$$\Pr(Y_i = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i} \tag{6}$$

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

$$\mu_i = e^{(ln\ t_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_k X_{ki})} \tag{7}$$

In our study, $\beta = 1 \ldots k$ 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.

3. Results

Here, a machine learning-based method to identify genetic and meteorological features that impact salmonellosis outbreak scale (typified by number of cases within each outbreak) irrespective of *Salmonella enterica* serovar-level heterogeneity was developed. 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, along with meteorological factors, as predictor variables in count-based models to identify their individual or combined impact on illness numbers.

3.1. Outbreak and WGS data collection and preprocessing

Relevant human outbreaks of *Salmonella* were selected from the NORS dashboard 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, the outbreaks were matched 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 the *Salmonella* pan genome (Fig. 1). The number of cases within a large number of included outbreaks was comparatively lower, with a majority of outbreaks having ≤60 cases (reported illnesses). This skewed distribution is common with datasets with discrete data points that are commonly analyzed by count-based

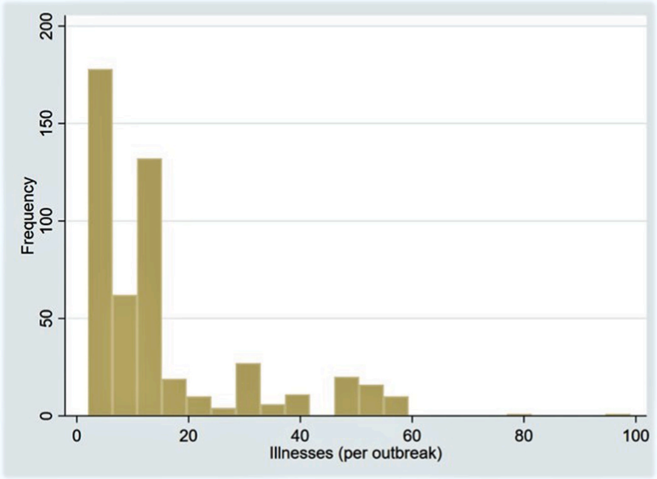

Fig. 2. Histogram depicting outbreak trends (illnesses per outbreak) observed in our study. A majority of the outbreaks included in our study had a small number of overall reported case numbers (n < 20).

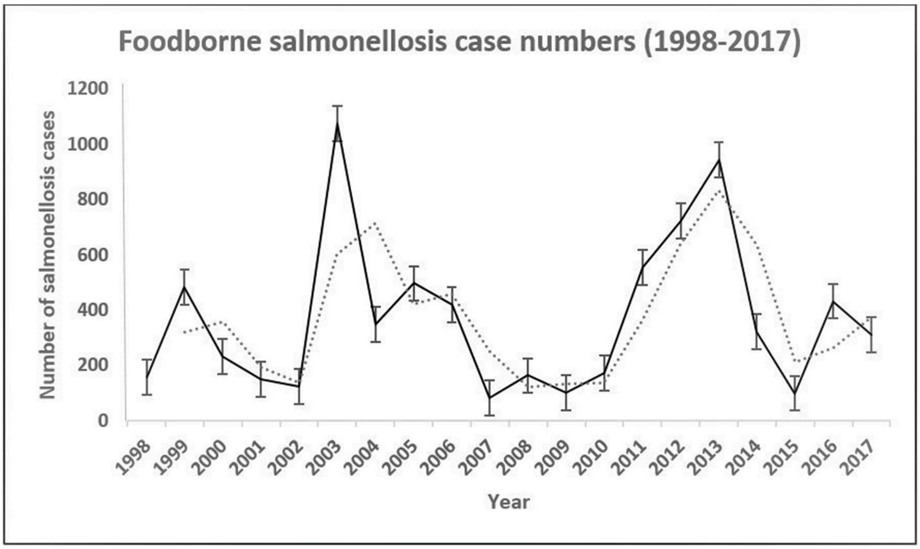

Fig. 1. Yearly trend in foodborne salmonellosis case numbers (1998–2017). Only outbreaks included in the final dataset and the number of illnesses (or case numbers) from these outbreaks are included in this model.

models (Fig. 2).

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. Multiple isolates for each data point were included to account for genetic recombination, and directionality and timing of evolutionary changes within and among serovars (Grad and Lipsitch, 2014). Short reads for each isolate were assembled and annotated on the PATRIC web server for homogeneity and the final *Salmonella* pan-genome was constructed.

3.2. Exploratory data analysis – meteorological data

Exploratory analyses of the trends in month-wise outbreak case numbers compared to meteorological factors alone (excluding the impact of genetic factors) revealed that the highest temperatures were observed between the months of June–August (Fig. 3) and the highest rainfall averages (and consequently, the highest precipitation) were observed in the month of August (Fig. 4). These, in turn, were correlated with increased salmonellosis case numbers. This corresponded well with prior knowledge regarding the correlation between meteorological factors, such as temperature, and outbreak case numbers. For example, Shirriff (2019) found that peak salmonellosis case numbers were observed during the warmest months of the year (June, July, and August) in the U.S. states of Florida, Illinois, Maryland, Minnesota, New York, Ohio, and Washington (Shirriff, 2019).

3.3. Machine learning-based identification of genes informative to Salmonella illness prediction model

Of the 18,520 genes comprising our *Salmonella* pan-genome, the best-fit Elastic Net model (α value = 0.50 and λ penalty = 2.18) was selected by 10-fold cross validation. The model minimizing the number of variables to provide a stable model fit (Fig. 5) identified 53 distinct, non-zero gene predictor variables that were most informative to the model (Fig. 6). It is important to note that the genes identified as significant by the model were accessory genes (i.e., did not belong to the *Salmonella* core genome). This is because genes that are present in all isolates would not add to the predictive capacity of the final regression model (and would only serve as noise in the model). The functionality of these genes ranged from virulence (such as the secreted effector protein coding gene *SteA*, putative adhesion large repetitive protein coding gene, virulence-associated TolA protein coding gene *tolA*, among others) to temperature-related stress response (RNA polymerase sigma factor-encoding gene *rpoS*) and bacterial metabolism (such as ABC transporters, transcriptional regulators, and fructokinase-coding gene, among others). Of note, 13 of the selected significant genes coded for bacterial phage proteins. The identified important genes and the known or presumed functionalities of the coded proteins are provided in Supplementary Table 1.

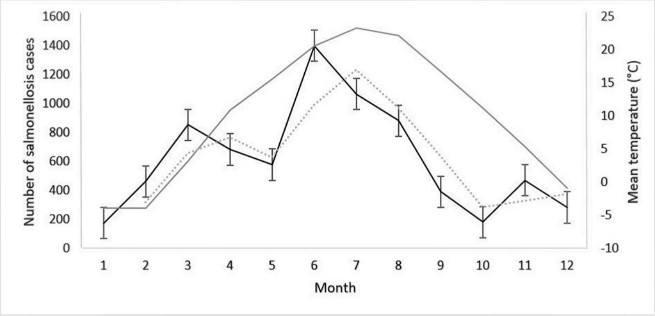

Fig. 3. Exploratory data analysis I. Monthly trend in salmonellosis cases (included in our study) viewed alongside the mean monthly temperature. Solid black line indicates the monthly trend in salmonellosis cases, dotted black line indicates a simple 2-point moving averages trend line, error bars indicate standard error, and the grey line indicates the monthly average temperature.

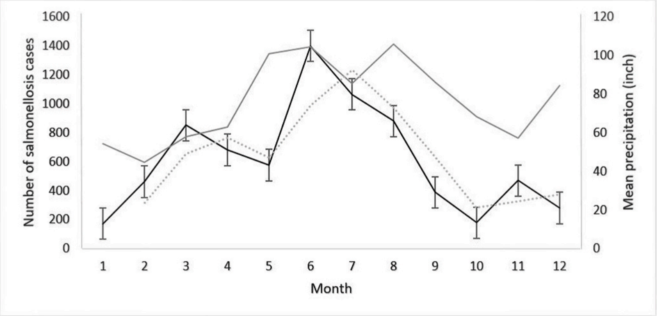

Fig. 4. Exploratory data analysis II. Monthly trend in salmonellosis cases (included in our study) viewed alongside the mean monthly precipitation. Solid black line indicates the monthly trend in salmonellosis cases, dotted black line indicates a simple 2-point moving averages trend line, error bars indicate standard error, and the grey line indicates the monthly average precipitation.

3.4. Poisson regression model outcome

Poisson regression models were developed using a matrix of gene presence/absence and meteorological data, and combinations of these. In our study, the model coefficients are interpreted 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) corresponding to the coefficient of the predictor variable, whereas a negative coefficient implies a decrease in the predicted response variable with an increase in the value of the coefficient 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 $p < 0.10$ (Fig. 7). The model containing these 33 predictors showed a significant improvement and fit over the null model (Likelihood ratio χ^2 statistic = 4604.21; McFadden's pseudo $R^2 = 0.536$; probability $> \chi^2 = 0$). 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). Notably, a majority of predictor variables that negatively influenced the outcome were associated with signaling and other cellular processes.

The final Poisson regression model (no lag) with monthly average temperature, monthly average precipitation, and monthly average snowfall identified 127 predictor terms that were significant (at $p < 0.10$ (n = 8) or 0.05 (n = 119)) 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 (Supplementary Table 2). The model containing the 119 predictors (significant at $p < 0.05$) showed a significant improvement and fit over the null and baseline models (Likelihood ratio χ^2 statistic = 5748.22; McFadden's pseudo $R^2 = 0.669$; probability $> \chi^2 = 0$). We observed that a number of gene predictors that were dropped (i.e. not significant) 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 outbreak scale (as typified by case numbers).

Although the two-month lag model also showed a significant

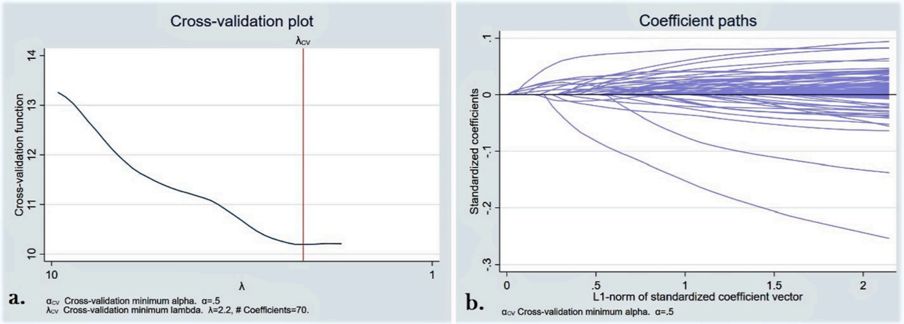

Fig. 5. Best-fit Elastic Net cross-validation plot and coefficient path. a) The CV plot indicates the best-fit α value and λ penalty that minimizes the cross validation function. The best-fit Elastic Net model netted 53 non-duplicate coefficients, at an α value = 0.50 and λ penalty = 2.18. b) The coefficient path, or solution path, for the Elastic Net model provides a compact representation of all optimal solutions for the model.

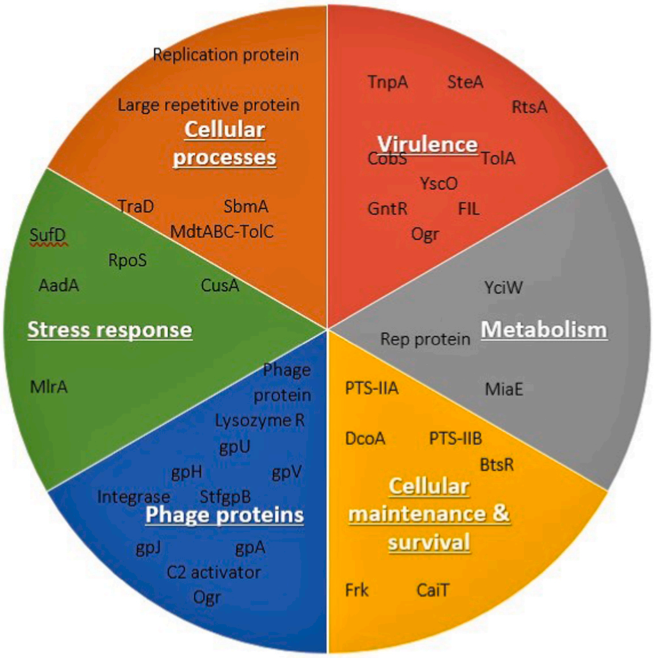

Fig. 6. Significant genes identified by Elastic Net model (n = 53) and their functional classes. The functional classes (derived from an extensive literature survey) corresponded well with the Kyoto Encyclopedia of Genes and Genomes (KEGG) Orthologous Group (OG) classification.

improvement in fit compared to the null and baseline models, important covariates like monthly average temperature and monthly average precipitation were dropped from the model (that is, they were found to be not statistically different). Moreover, the coefficients (and their relationship to the outcome) of the remaining covariates 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.

3.5. Negative binomial model outcome

The negative binomial 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. We found that the negative binomial model did not perform as well as the Poisson regression in fitting the data. The baseline (gene only) negative binomial model identified 28 predictor variables (χ^2 statistic = 517.34; Pseudo R^2 = 0.139; probability > χ^2 = 0) that mostly corresponded with those identified by the baseline Poisson model, but was not much better than the null model. However, based on the improved results observed for the multivariable Poisson model, a multivariable negative binomial model was also developed. This model showed an improvement over the null and baseline models, but performed significantly poorer compared to the multi-variable Poisson model, and was ultimately dropped from further consideration (χ^2 statistic = 912.44; McFadden's pseudo R^2 = 0.244; probability > χ^2 = 0). Moreover, the covariates 'mean average daily temperature' and 'average precipitation' were observed to not significantly impact the model (data not included). Thus, the results of this model were dropped from further consideration.

4. 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* (McMichael, 2015; Rose et al., 2001; Simental and Martinez-Urtaza, 2008). Particularly, a positive association has been reported between diarrheal disease numbers and temperature increase (Singh et al., 2001). Moreover, 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 et al., 2014).

The bacterial genetic code holds the key to unlocking the many secrets of bacterial pathogen growth, survival, proliferation, and pathogenicity. However, its potential is only now being realized, especially after the advent of whole genome sequencing. Whole genome sequencing is a relatively new technology that is increasingly being used by a number of public health laboratories to definitively identify and characterize microbial causes of foodborne illnesses. Currently, in large part thanks to reducing costs and rapid turnover time, WGS is being applied to surveillance and disease outbreak investigation, and identifying the key mechanisms behind pathogen virulence and survival to understand and minimize the occurrence of pathogens in food (Fritsch et al., 2018a; Pornsukarom et al., 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 large datasets and obtain meaningful data from WGS, specifically in the case of preventative modeling of pathogen growth, survival, and overall human health risk.

Machine learning is increasingly being applied in the food safety domain to incorporate WGS data in many aspects of predictive modeling, specifically in identifying trends in bacterial virulence (Karanth et al., 2022; Njage et al., 2019a, 2019b; Tanui et al., 2022), pathogen source attribution (Munck et al., 2020), and in developing

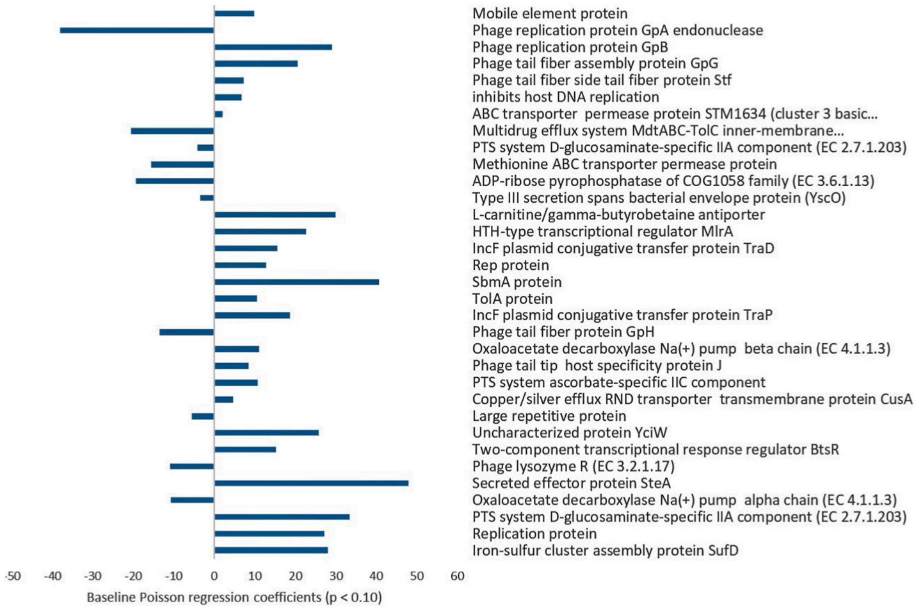

Fig. 7. Significant genes identified in the baseline Poisson regression model. The baseline model determined the impact of genes only on salmonellosis outbreak scale. The plot demonstrates the relative importance of the signficant genes - the Poisson regression coefficients (β_k) indicate the multiplicative change in the expected count of salmonellosis cases for a unit increase in the corresponding gene predictor, given all other predictors are kept constant.

gene-based risk assessments (Fritsch et al., 2018b) to predict the risk of disease given exposure. However, so far, studies have not analyzed the joint impact of a pathogen's genetic expression and meteorological factors such as temperature and precipitation on the pathogen's infectivity and outbreak scale (in terms of case or illness numbers). In this study, we have utilized 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 monthly average temperature, precipitation, and snowfall, as predictor variables in a Poisson regression model (since the outcome variable is in counts) to identify genes and environmental covariates that are highly correlated with salmonellosis case numbers. An important consideration when utilizing large gene matrices in predictive modeling is identifying and selecting the smallest possible set of relevant genes that can help achieve good predictive performance, without model overfitting or including features that are irrelevant or redundant to the prediction process (Guyon et al., 2003). In the presence of such data, it is important to employ a statistical approach to select meaningful subsets of predictors for samples with complete data, similar to the approach used by Amene et al. (2016) to predict mortality rates associated with foodborne diseases. Elastic Net has previously proven to be effective (with an accuracy of >90%) in identifying genetic features of interest related to lung cancer (Hughey and Butte, 2015). Thus, in our study, feature selection was performed by Elastic Net to pre-process the large genetic dataset to be employed in our predictive models. Of the 53 distinct non-zero gene terms identified as important by Elastic Net, 27 coded for metabolism, cellular maintenance, biological transport, virulence, and stress response-related processes, 13 were phage proteins associated mostly with bacterial virulence, and the remaining 13 did not present with a clear functional classification (Fig. 6). These also corresponded well with the Kyoto Encyclopedia for Genes and Genomes (KEGG) orthologous group classification, wherein a majority of the identified genes coded for metabolism and signaling and cellular processes, including secreted effectors, transporters, and assorted protein families. 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 et al. (2014), with the trends, particularly in relation to the average temperatures, corresponding well with those reported previously by Shirriff (2019).

The best-fit Poisson regression model, which contained 127 significant genes (n = 45), meteorological attributes (n = 3) and gene interaction (n = 79) terms, had a pseudo R^2 of 0.669. Since the pseudo R^2 value is influenced by the sample size, number of predictor variables, and number of categories of the dependent variable, setting the interpretation for model fit and stability using the pseudo R^2 must explicitly consider these characteristics (Hemmert et al., 2018). Our pseudo R^2 of 0.669 is higher than the benchmark range highlighted by Hemmert et al. (2018), who identified a pseudo R^2 of >0.4 to indicate an excellent model fit, given the sample size and number of predictor (and interaction) variables included in our model. This model identified a number of genes and gene-meteorological interaction terms that significantly contributed to salmonellosis outbreak scale (Supplementary Table 2). Among the 51 genes that impacted the Poisson model individually (n = 45) or interacting with a meteorological term, 25 coded for proteins associated with bacterial signaling and cellular processes (such as secreting virulence effectors) and metabolism (such as participating in the pentose phosphate pathway or deoxyribonuclease pathway), 9 coded for phage proteins, and others coded for proteins participating in a host of miscellaneous cellular activities, such as homologous recombination and environmental signal processing. Specifically, genes coding for *Salmonella* effector A, which is translocated by both the *Salmonella* pathogenicity island (SPI)-1 and the SPI-2 type 3 secretion systems (T3SSs) (SteA), putative resistance protein (YqiE), putative chaperone with DNA J-like domain (YbeV), phage tail fiber protein, mobile element protein, and phage replication protein GpA exerted a noticeable impact on the number of illnesses, compared to other variables. 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 with the number of illnesses, 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 scale, indicating that for every one unit increase in temperature, the probability of the interacting gene predicting the log of the illness outcome increases by the value of the coefficient. For example, when an interactive predictor has a coefficient of 0.05, for every 1 unit increase in temperature, the gene's effect on the outcome increases by 0.05. Concurrently, we observed that the temperature-interaction effects of a large percentage of metabolism and cell maintenance-related proteins were positively correlated with outbreak scale. This is in agreement with the conclusions of Dawoud et al. (2017) and Pin et al. (2012), 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 average precipitation effect and outbreak scale, 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 scale, which in turn could be correlated with the increased precipitation (Holley et al., 2008; Piekarska, 2010). Interestingly, genes coding for the effector protein SteA and the uncharacterized protein YbeV, which are associated with *Salmonella* virulence (Azimi et al., 2019) and stress response (Kobayashi et al., 2005), respectively, were significant individually and in combination with all three meteorological variables.

Our study showed some confounding results regarding the effect of temperatures on outbreak scale (as defined by number of illnesses). We observed that, for a one °C increase in average temperature, the difference in log of expected case numbers would be expected to decrease by 195.35. While this relationship is contrary to published literature, which have repeatedly found a positive association between temperature and salmonellosis incidence rates, and our own exploratory data analysis, the results are in agreement with those of Semenov et al. (2007), who reported similar inconsistent conclusions about temperature levels contributing to *Salmonella* survival. In essence, they found that *Salmonella* survival significantly declined with increasing average 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*. This also corresponded with the results of Kynčl et al. (2021) from a long-term retrospective study conducted in the Czech Republic, who found an asymptotic curve approaching the extremes of mean monthly temperatures, despite a linear relationship between air temperatures and outbreak cases between 1 and 15°C.

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 associated with *Salmonella* infection may be self-limiting, and therefore not serious enough to warrant testing, let alone hospitalization. Second, since our WGS dataset is built from among isolates obtained to correlate with salmonellosis outbreaks (based on time and location of isolation, relative to time and location of outbreaks), the initial pan genome dataset is not wholly representative of all *Salmonella* serovars specifically associated with 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.' Additionally, our machine learning-based model predictions are performed using historical monitoring data and other available information, and not experimental data, with the goal of predicting future trends in food safety-related risks. Since these models cannot be validated using experimental data, as this would require artificial manipulation of the meteorological covariates, we acknowledge that these models cannot be used to derive causality. However, the model predictions (and resultant correlations) have been validated using cross validation, and can be further validated with the availability of data in the future. Finally, due to the small number of data points, meteorological factors have been pooled within each 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 scale.

5. Conclusion

In our study, we developed multi-variable Poisson regression models to determine the impact of *Salmonella enterica* genes, pooled (by month and year) meteorological factors, and their combinations on *Salmonella* outbreak scale. 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 scale. However, our study had a few limitations: since our dataset was defined by our inclusion criteria, the total number of isolates included was limited. Moreover, since exact peer-to-peer matched isolate data was unavailable corresponding to each outbreak, our data collection was based on specific associations. Here, we attempted to overcome this by analyzing multiple sequences across sources to incorporate potential heterogeneities. Increasing data availability, as well as incorporating important metadata parameters during the collection phase of bacterial isolates, are important steps towards developing more well-rounded datasets to develop and validate these models in the future. We envision this as the first step towards incorporating the effect of bacterial gene expression in models predicting bacterial foodborne outbreak scale, which are traditionally based on environmental and processing-related factors.

CRediT authorship contribution statement

Shraddha Karanth: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Visualization, Writing – original draft. **Jitu Patel:** Writing – review & editing. **Adel Shirmohammadi:** Writing – review & editing. **Abani K. Pradhan:** Conceptualization, Methodology, Resources, Writing – review & editing, Funding acquisition, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported through a grant from the U.S. Department of Agriculture National Institute of Food and Agriculture (NIFA) Agriculture and Food Research Initiative (award number 2020-67017-30785). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the USDA-NIFA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.crfs.2023.100525.

References

Adzitey, F., Asante, J., Kumalo, H.M., Khan, R.B., Somboro, A.M., Amoako, D.G., 2020. Genomic investigation into the virulome, pathogenicity, stress response factors, clonal lineages, and phylogenetic relationship of *Escherichia coli* strains isolated from meat sources in Ghana. Genes 11 (12), 1504.

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

Amene, E., Hanson, L.A., Zahn, E.A., Wild, S.R., Döpfer, D., 2016. Variable selection and regression analysis for the prediction of mortality rates associated with foodborne diseases. Epidemiol. Infect. 144 (9), 1959–1973.

Arguez, A., Durre, I., Applequist, S., Vose, R.S., Squires, M.F., Yin, X., Heim Jr., R., Owen, T.W., 2012. NOAA's 1981–2010 U.S. climate normals: an overview.Bull. Am. Met. Soc 93, 1687–1697.

Azimi, T., Zamirnasta, M., Sani, M.A., Soltan Dallal, M.M., Nasser, A., 2019. Molecular mechanisms of *Salmonella* effector proteins: a comprehensive review. Infect. Drug Resist. 13, 11–26.

Badie, F., Saffari, M., Moniri, R., Alani, B., Atoof, F., Khorshidi, A., Shayestehpour, M., 2021. The combined effect of stressful factors (temperature and pH) on the expression of biofilm, stress, and virulence genes in *Salmonella* enterica ser. Enteritidis and Typhimurium. Arch. Microbiol. 203 (7), 4475–4484.

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. J. Comput. Biol. 19 (5), 455–477.

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. Sci. Rep. 5, 8365.

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

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

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. Front. Vet. Sci. 4.

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

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. J. Appl. Meteorol. Climatol. 52 (11), 2377–2395.

Ehuwa, O., Jaiswal, A.K., Jaiswal, S., 2021. *Salmonella,* food safety, and food handling practices. Foods 10 (5), 907.

Fritsch, L., Felten, A., Palma, F., Mariet, J., Radomski, N., Mistou, M., Augustin, J., Guillier, L., 2018a. 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. Int. J. Food Microbiol. 291, 181–188.

Fritsch, L., Guillier, L., Augustin, J.C., 2018b. Next generation quantitative microbiological risk assessment: refinement of the cold smoked salmon-related listeriosis risk model by integrating genomic data. Microb. Risk Anal. 10, 20–27.

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

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29 (8), 1072–1075.

Guyon, I., Elisseefi, A., Kaelbling, L.P., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (7–8), 1157–1182.

Hastie, T.J., Tibshirani, R.J., Wainwright, M., 2015. Statistical Learning with Sparsity: the Lasso and Generalizations. CRC Press, Boca Raton, FL.

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

Hemmert, G.A.J., Schons, L.M., Wieseke, J., Schimmelpfennig, H., 2018. Log-likelihood-based pseudo-R2 in logistic regression: deriving sample-sensitive benchmarks. Socio. Methods Res. 47 (3), 507–531.

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. J. Environ. Qual. 37 (6), 2083–2092.

Hughey, J.J., Butte, A.J., 2015. Robust meta-analysis of gene expression using the elastic net. Nucleic Acids Res. 43 (12), e79.

Karanth, S.K., Tanui, C.K., Meng, J., Pradhan, A.K., 2022. Exploring the predictive capability of advanced machine learning in identifying severe disease phenotype in *Salmonella enterica*. Food Res. Int. 151, 110817.

Kobayashi, H., Miyamoto, T., Hashimoto, Y., Kiriki, M., Motomatsu, A., Honjoh, K., Iio, M., 2005. Identification of factors involved in recovery of heat-injured Salmonella Enteritidis. J. Food Protect. 68 (5), 932–941.

Kynčl, J., Špačková, M., Fialová, A., Kyselý, J., Malý, M., 2021. Influence of air temperature and implemented veterinary measures on the incidence of human salmonellosis in the Czech Republic during 1998–2017. BMC Publ. Health 21.

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. Microbiol. Mol. Biol. Rev. 78 (1), 74–88.

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

Mun, S.G., 2020. The effects of ambient temperature changes on foodborne illness outbreaks associated with the restaurant industry. Int. J. Hospit. Manag. 85, 102432.

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 Anal. 40 (9), 1693–1705.

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. Epidemiol. Infect. 137, 367–374.

Njage, P.M.K., Leekitcharoenphon, P., Hald, T., 2019a. Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: predicting clinical outcomes in shigatoxigenic *Escherichia coli*. Int. J. Food Microbiol. 292, 72–82.

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

Owen, T.W., Whitehurst, T., 2002. United States climate normals for the 1971–2000 period: Product descriptions and applications. Preprints,Third symposium on environmental applications: Facilitating the use of environmental information. Amer. Meteor. Soc., Orlando, FL. J4.3. [Available online at: https://ams.confex.com/ams/annual2002/webprogram/Paper26747.html. J4.3. [Available online at:

Piekarska, K., 2010. Mutagenicity of airborne particulates assessed by *Salmonella* assay and the SOS Chromotest in Wrocław, Poland. J. Air Waste Manage. Assoc. 60 (8), 993–1001.

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. Int. J. Food Microbiol. 213, 130–138.

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. PLoS One 7 (12), e51196.

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 Genom. 19 (1), 801. https://doi.org/10.1186/s12864-018-5137-4.

Robinson, E.J., Gregory, J., Mulvenna, V., Segal, Y., Sullivan, S.G., 2022. Effect of temperature and rainfall on sporadic salmonellosis notifications in Melbourne, Australia 2000–2019: a time-series analysis. Foodb. Pathog. Dis. 19 (5).

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. Environ. Health Perspect. 109, 211.

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. Emerg. Infect. Dis. 17 (1), 7–15.

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

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 6 September 2021).

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. Sci. Total Environ. 463–464, 488–496.

Simental, L., Martines-Urtaza, J., 2008. Climate patterns governing the presence and permanence of salmonellae in coastal areas of Bahia de Todos Santos, Mexico. Appl. Environ. Microbiol. 74, 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. Environ. Health Perspect. 109, 155–159.

Sirsat, S.A., Baker, C.A., Park, S.H., Muthaiyan, A., Dowd, S.E., Ricke, S.E., 2015. Transcriptomic response of *Salmonella* Typhimurium heat shock gene expression under thermal stress at 48°C. J. Food Res. 4 (5).

Soneja, S., Jiang, C., Upperman, C.R., Murtugudde, R., Mitchell, C.S., Blythe, D., Sapkota, A., 2016. Extreme precipitation events and increased risk of campylobacteriosis in Maryland. U.S.A. Environ. Res. 149, 216–221. https://doi.org/10.1016/j.envres.2016.05.021.

StataCorp, 2019. Stata Statistical Software: Release 16. StataCorp LLC, College Station, TX.

StataCorp, 2021. ElasticNet – ElasticNet for Prediction and Model Selection. Retrieved from. https://www.stata.com/manuals/lassoelasticnet.pdf. (Accessed 1 August 2021).

Stephen, D.M., Barnett, A.G., 2016. Effect of temperature and precipitation on salmonellosis cases in South-East Queensland, Australia: an observational study. BMJ Open 6 (2), e010204.

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 Technol. 69, 42–49.

Tanui, C.K., Karanth, S.K., Njage, P.M.K., Meng, J., Pradhan, A.K., 2022. Machine learning-based predictive modeling to identify genotypic traits associated with Salmonella enterica disease endpoints in isolates from ground chicken. LWT 154, 112701.

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

Weisburd, D., Wilson, D.B., Wooditch, A., Britt, C., 2021. Count-based regression models. In: Advanced statistics in criminology and criminal justice. Springer, Cham. https://doi.org/10.1007/978-3-030-67738-1_6.

Yun, J., Greiner, M., Höller, C., Messelhäusser, U., Rampp, A., Klein, G., 2016. Association between the ambient temperature and the occurrence of human *Salmonella* and *Campylobacter* infections. Sci. Rep. 6, 28442.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320.