Downloaded from orbit.dtu.dk on: feb 15, 2022

DTU Library

Machine learning-based predictive modeling to identify genotypic traits associated with Salmonella enterica disease endpoints in isolates from ground chicken

Tanui, Collins K.; Karanth, Shraddha; Njage, Patrick M.K.; Meng, Jianghong; Pradhan, Abani K.

Published in:
LWT

Link to article, DOI:
10.1016/j.lwt.2021.112701

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Tanui, C. K., Karanth, S., 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]. https://doi.org/10.1016/j.lwt.2021.112701

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LWT - Food Science and Technology 154 (2022) 112701

Contents lists available at ScienceDirect

LWT

journal homepage: www.elsevier.com/locate/lwt

Machine learning-based predictive modeling to identify genotypic traits associated with *Salmonella enterica* disease endpoints in isolates from ground chicken

Collins K. Tanui [a,b], Shraddha Karanth [a], Patrick M.K. Njage [c], Jianghong Meng [a,b,d], Abani K. Pradhan [a,b,*]

[a] Department of Nutrition and Food Science, University of Maryland, College Park, MD, 20742, USA
[b] Center for Food Safety and Security Systems, University of Maryland, College Park, MD, 20742, USA
[c] Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark
[d] Joint Institute for Food Safety and Applied Nutrition, University of Maryland, College Park, MD, 20742, USA

ARTICLE INFO

Keywords:
Predictive modeling
Machine learning
Whole genome sequencing
Salmonella

ABSTRACT

As the cost of genome sequencing of foodborne pathogens decreases, it has become possible to sequence a large number of isolates and evaluate those using machine learning algorithms. This study aimed to utilize machine learning algorithms to predict the disease endpoints in untagged *Salmonella* genome sequences isolated from ground chicken. Our models recognized genetic patterns in the test dataset based on our training dataset obtained from an extensive literature review, using a semi-supervised approach. Using known genotypes as input features, the semi-supervised random forest model showed the highest overall accuracy of 0.94 (95% confidence interval: 0.85–0.99), and a Kappa value of 0.82, and predicted 87% of the disease endpoints. The model predicted genes associated with specific disease endpoints that were associated with virulence, which could be used as features in predictive modeling endeavors in the future. Our machine learning approach would be useful in different areas of food safety, including identifying pathogen sources, predicting antibiotic resistance, and risk assessment of foodborne pathogens.

1. Introduction

Salmonella enterica subsp. *enterica* is a ubiquitous, gram-negative, facultative anaerobic bacterium with demonstrated human health implications. According to the United States Centers for Disease Control and Prevention (U.S. CDC), *Salmonella enterica* is responsible for an estimated more than one million foodborne illnesses in the U.S. every year (CDC, 2021). Foodborne salmonellosis, which is linked to the consumption of contaminated foods ranging from meat animals and poultry to produce and nuts (Zhao et al., 2001; Horby et al., 2003; Naugle et al., 2006; Braden, 2006; Danyluk et al., 2007; Scallan et al., 2011; Angelo et al., 2015; Huang et al., 2016), is one of the most common causes of salmonellosis worldwide (CDC, 2021). Due to genetic evolution and the need for adaptation to a diverse range of hosts from warm-blooded mammals to vegetables and fruits, *Salmonella enterica* has high intra-species diversity (Amavisit et al., 2003; Monack, 2012). *Salmonella enterica* subsp. *enterica* alone has over 2500 named serovars, many with markedly different host specificities and virulence capacities, and many of its subspecies and serovars making the jump across hosts to be able to infect humans (Uzzau et al., 2000). Researchers have observed significant inter-and intra-serovar heterogeneity in *Salmonella*-related disease endpoints, such as gastroenteritis, systemic infection, bacteremia, and enteric fever (Majowicz et al., 2010; Mohammed & Cormican, 2016; Cao et al., 2020; Calero-Cáceres et al., 2020). Therefore, the risk associated with the presence of *Salmonella* in foods can be arguably defined as being higher. This understates the importance of understanding and incorporating the variation in virulence among different serovars of *Salmonella* when assessing the risk of salmonellosis.

In the last few years, whole genome sequencing (WGS) has seen significant popularity in the food safety domain. Sequences from pathogenic microorganisms isolated as part of routine surveillance and outbreak investigations have been made publicly available. However, their use in functions other than epidemiological surveillance has been restricted by a distinct lack of associated metadata (Köser et al., 2012;

* 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.lwt.2021.112701
Received 2 August 2021; Received in revised form 22 October 2021; Accepted 23 October 2021
Available online 25 October 2021
0023-6438/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Rantsiou et al., 2018), and heterogeneity in the data collection and processing strategies by the primary investigators. While WGS provides a wealth of information surrounding pathogen virulence, biologists are still trying to identify the linkage between genotypic and phenotypic traits using classical statistics. Therefore, accurate characterization of strains' phenotypes is fundamental for quantitative microbial risk assessment (QMRA), and published research articles have shown possible links between the phenotypic behavior of a number of foodborne pathogens and their genotypic traits (den Besten et al., 2018; Njage, Henri, et al., 2019; Njage, Leekitcharoenphon, & Hald, 2019; Chen et al., 2020). In *Salmonella*, particularly, concrete associations have been identified between genotypic features, such as the *Salmonella* Pathogenicity Islands (SPI)-1 and SPI-2, and disease phenotypes (Zou et al., 2011; Suez et al., 2013; Chen et al., 2020).

Machine learning (ML) has gained popularity in the era of big data in the scientific community (Libbrecht & Noble, 2015; Ching et al., 2018). Whereas mechanistic modeling mainly relies on simplified mathematical formulations to solve issues regarding complex datasets (Baker et al., 2018), ML algorithms extract meaningful features and make predictions based on "learned" patterns in such datasets (Libbrecht & Noble, 2015; Alkema et al., 2016). ML can either be performed in a supervised fashion by classifying, predicting, and interpreting data or via an unsupervised means by unraveling and detecting unique patterns within a given dataset (Tebani et al., 2016). ML is typically applied in situations where large datasets are available and can be related to known outputs of interest (Libbrecht & Noble, 2015). ML algorithms such as random forest (RF), logistic regression, support vector machine (SVM), gradient boosting (GBM), and AdaBoost, have, in fact, seen increasing usage in the biological and microbiological domain (Libbrecht & Noble, 2015; Jordan & Mitchell, 2015; Safae et al., 2018; Wheeler et al., 2018). ML offers an opportunity to overcome challenges associated with linking genotypic information to phenotypic traits that are too complex to model mathematically (Tebani et al., 2016; Baker et al., 2018; Njage, Henri, et al., 2019; Njage, Leekitcharoenphon, et al., 2019). This is particularly the case for phenotyping of foodborne pathogens, where it is difficult to efficiently predict possible phenotypic outcomes from pathogen genetic information using classical mathematical models only, due to a number of challenges, including how to handle heterogeneous data, inherent imbalance in classes due to heterogeneity in sampling and testing motivations, and reliance on arbitrary p-values (Libbrecht & Noble, 2015). Therefore, using these predictive tools with readily available genomes deposited in public databases offers novel opportunities for developing ways to predict specific outcomes.

In the current study, we have utilized a ML approach to investigate the relatedness of isolates from ground chicken and disease phenotypic outcomes using *Salmonella* whole genome sequences. The overall approach and outcome of the study would be useful in food safety and predictive modeling, which further opens avenues to potentially integrate genomic data into risk assessment frameworks and source attribution studies.

2. Materials and methods

2.1. Data collection

2.1.1. Labeled dataset

Salmonella isolates previously identified from human cases of bacteremia, gastroenteritis, and systemic infection were identified from literature (Calero-Cáceres et al., 2020; Cao et al., 2020; Mohammed & Cormican, 2016; Octavia et al., 2019). The curated *Salmonella* isolates were associated with a human endpoint of bacteremia (n = 12), gastroenteritis (n = 9), and systemic infection (n = 7). These pre-labeled isolates with known clinical endpoints (Supplementary Table S1) are important in identifying genotypic patterns associated with each clinical endpoint, in order to extrapolate patterns onto sequences with untagged data. Isolates were included based on the availability of a specific clinical endpoint, as well as published genome sequences. The general lack of availability of the associated disease phenotypes and the stringent search criteria employed were responsible for the low number of tagged isolates included in our study.

2.1.2. Unlabeled dataset

In the U.S. Food and Drug Administration's (U.S. FDA) GenomeTrakr network, a total of 134,733 *Salmonella* isolates were sequenced (as of January 2020), 205 of which were from ground chicken. Samples taken from ground chicken were employed in this study because ground chicken is likely to have higher microbiological loads than whole carcasses and parts (Chen et al., 2014). The sequences were opportunistically selected from among those reported by the GenomeTrakr project with previously reported cases of *Salmonella* infection and were comprised of different serovars (CDC, 2016). Therefore, the unlabeled dataset was comprised of this unlabeled *Salmonella* WGS data (N = 205; Supplementary Table S2) obtained from the National Center for Biotechnology Information's (NCBI) Pathogen Detection database. Additionally, available and important metadata was also collected in order to perform a preliminary phylogenetic analysis. PhyloPhlAn pipeline was used to generate a phylogenetic tree on labeled and unlabeled *Salmonella* strains as described in a previous study (Segata et al., 2013).

2.2. Bioinformatics analysis

The raw reads from the test and training datasets were *de novo* assembled using the Pathosystems Resource Integration Center (PATRIC) webserver employing different strategies (Davis et al., 2020). Short reads were assembled using BayesHammer (Nikolenko et al., 2013). Subsequently, the genomes were assembled using the in-built Velvet (Zerbino & Birney, 2008), IDBA (Peng et al., 2010), and SPAdes (Bankevich et al., 2012) algorithms on PATRIC, and each assembled genome was assigned an assembly score by ARAST (Davis et al., 2020). All algorithms were run on default parameters, and assembly quality was determined by analyzing the QUAST score. The SPAdes assembly algorithm in PATRIC consistently received a higher ARAST ranking compared to the others and was therefore used to assemble the sequences. *De novo* assembled sequences with the highest scores were then subjected to downstream analysis. Genome annotation was performed using the in-built Rapid Annotation and Subsystems Technology toolkit (RASTtk) (Aziz et al., 2008) in the PATRIC web server. Target genes, specifically those coding for *Salmonella* virulence, were identified from an extensive literature survey. Briefly, the presence/absence of these target genes was determined from all labeled and unlabeled annotated genomes, together with their corresponding identity percentages. These were then extracted and tabulated into a matrix. Genes used as the input/predictor dataset in this study are shown in Supplementary Table S3.

2.3. Machine learning-based prediction of disease phenotypes

2.3.1. Overview of the approach to model building/selection and disease prediction using the selected model

Ensemble machine learning algorithms were employed to attempt class prediction based on target gene presence in this study. Model building and the prediction were performed by randomly splitting (70/30) the data into training and test subsets based on the labeled dataset. Models were fitted to the training dataset using the RF, logit boost (LB), GBM, and SVM with radial and linear kernel (SVMR and SVML, respectively) ML algorithms. The best-fitting models were determined by 10-fold cross-validation and selected based on accuracy metrics. The fitted models were then evaluated on the test data set and the best-performing model was used in predicting unknown endpoints of isolates (disease phenotypes). Simply put, several base classifier models

were trained on labeled data to predict three disease endpoints/phenotypes – gastroenteritis, bacteremia, or systemic infection. The trained classifiers were then employed to assign probabilistic class labels to unlabeled genetic data. We propose this modeling approach as a solution to utilizing the host of available unlabeled WGS data in improving predictive models to characterize the hazard through disease endpoint prediction in *Salmonella* from food sources. Here, we have utilized a number of base classifiers, including RF, GBM, and SVM to identify the best-fit classifier. The classifiers tested in our study were selected due to their prior usage in biological sequence classification and disease endpoint prediction (Kotsiantis, 2007; Libbrecht & Noble, 2015; Lupolova et al., 2016; Wheeler et al., 2018). Model building and prediction were performed using the *caret* package in R (v. 4.0.3, R Core Team, 2019; Vienna, Austria). A simplified workflow illustrating the approach employed in this study is presented in Fig. 1 and further explained below this section.

2.3.2. Data exploration and splitting

An exploration of the data revealed a class imbalance between clinical phenotypic outcomes. According to Velez et al. (2007), if imbalanced classes are to be used, then the model will most probably learn from predictors associated with the larger classes (classes with a greater number of samples) (Velez et al., 2007). To overcome this, a *post hoc* sampling technique previously proposed by Kuhn and Johnson (2013) was adapted, using the *caret* package in R to up-sample classes with a low number of samples (Kuhn & Johnson, 2013). Improved accuracy was obtained by up-sampling, which is a method to re-balance unbalanced datasets, or datasets where a particular class is over-represented (or where one or more classes are under-represented), which can introduce bias into the models. In this method, samples are taken with replacement from the classes such that classes with minority samples are equal to those of the majority class. This was performed until each class (bacteremia, gastroenteritis, and systemic infection) had nearly the same number of samples.

2.3.3. Algorithm selection and evaluation

Model training was performed on *Salmonella* isolates with known clinical endpoint (training). Initial model training was performed using a number of ML classification algorithms. Models were subjected to 10-fold cross-validation for hyperparameter tuning and selection of the best-performing model as previously described (Njage, Henri, et al., 2019,Njage, Leekitcharoenphon, et al., 2019). Average accuracy for the ten-fold cross-validations was obtained by comparing both the up-sampled and original datasets (Kuhn & Johnson, 2013; Njage, Henri, et al., 2019; Njage, Leekitcharoenphon, et al., 2019). A confusion matrix is a method to evaluate the model, by determining the balanced accuracy of each model. A confusion matrix is generally used to evaluate the performance of a classification model on a given test dataset for which the true (positive or negative) values are known. The accuracy scores, sensitivity, and specificity of each model in predicting the disease phenotype were determined. The accuracy score calculated from the confusion matrix describes the association between the predicted and actual classes.

Furthermore, the Kappa value was used to evaluate the agreement between class distributions. Usually, Kappa values range between −1 and 1: −1 indicates no agreement whereas 1 suggests a perfect agreement between the predicted and observed classes. Algorithms with Kappa value < 0.40, 0.40–0.70, and >0.75 were categorized as poor, fair to good, and excellent, respectively, as described previously (Fleiss et al., 2003). The prediction accuracy of phenotypic outcomes was evaluated using the sensitivity and specificity value (Altman & Bland, 1994).

2.3.4. Disease phenotype prediction

The best performing model was used to predict possible disease phenotypes in the unlabeled dataset (i.e., labeling). The chosen model would learn specific genetic patterns from the labeled dataset and predict possible phenotypic outcomes in the unlabeled dataset. The trained model was employed to assign probabilistic classes/labels to the unlabeled dataset (*Salmonella* isolates from ground chicken without endpoint data). The trained ML model classifies the untagged isolates based on the posterior probability (π_{ij}) that isolate i belongs to the jth disease endpoint. Posterior classification of each isolate with unlabeled disease endpoints was then performed by calculating their individual posterior

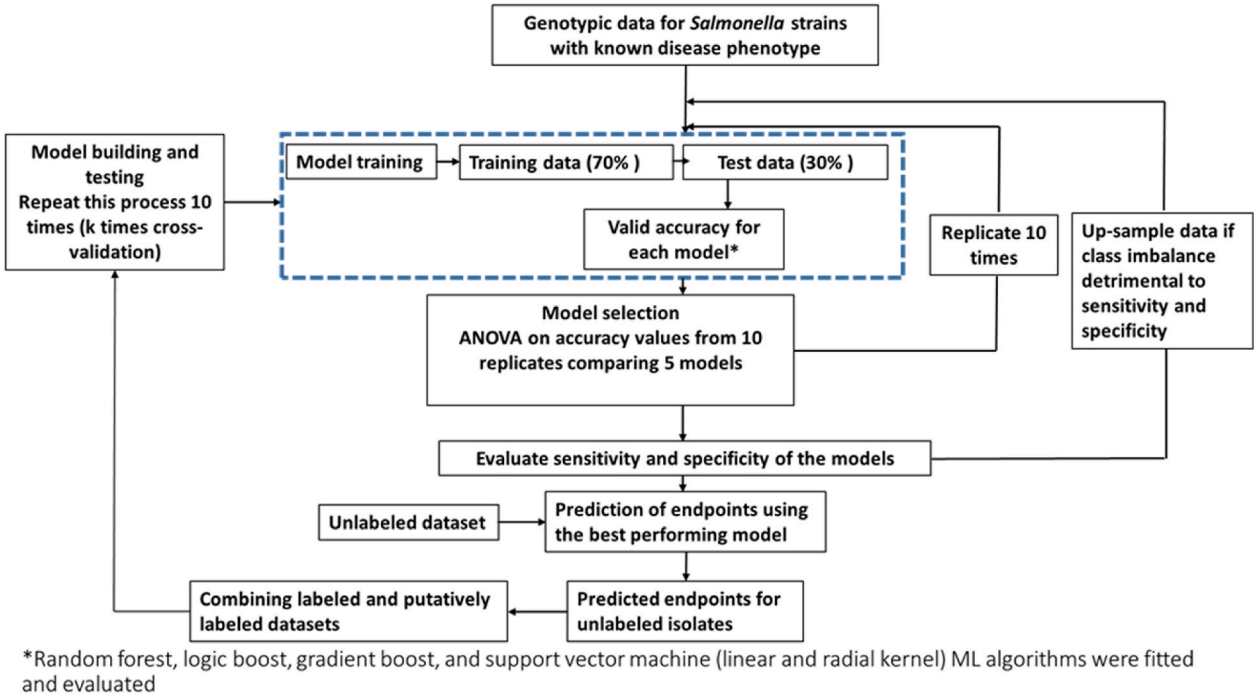

Fig. 1. Overview of prediction strategy. Semi-supervised machine learning-based predictive modeling, *Salmonella* isolates with known phenotypes were used for initial model training. *Salmonella* isolated from ground chicken were used for subsequent outcome prediction and model re-training and testing.

probabilities (π_{ij}) which express how likely the *i*th isolate is to belong to disease endpoint *j*, considering the observed set of genes for that isolate. The classification rule used was: Classify isolate *i* into disease endpoint *j* if and only if $\pi_{ij} = \max_k (\pi_{ij})$ which means to classify into the disease endpoint to which untagged isolate *i* is most likely to belong.

2.4. Variable importance

Genes encoding for virulence determinants responsible for bacteremia, gastroenteritis, and systemic infection, as well as redundant variable features that may most contribute to predictive model accuracy, were identified. The best-performing model was used to select features with variable importance for different outcomes and integrate the association between selected predictors. Therefore, important features were selected using the logit boost and random forest models and compared to those selected using the Boruta algorithm which is a model-independent wrapper algorithm (Kursa, 2014). These approaches assisted in the identification of important genes that are either strongly or weakly related to the disease outcomes.

3. Results

3.1. Identification of target genes for initial matrix development

Assembled and annotated *Salmonella* sequences from ground chicken for the labeled and unlabeled datasets had an average genome size of 4.8 Mb (Supplementary Tables S1 and S2). The predicted genes were searched against the protein family database, Pfam 5.5 (Uni Prot Consortium 2018). The COG database was used to find putative orthologs in other completed genomes (Tatusov et al., 2001). Target genes for all *Salmonella* isolates in the PATRIC database were analyzed and the genes' % identity were obtained and used in downstream analysis. It is worth noting that some genes were either present or absent in isolates. Phylogenetic analysis of our dataset established that the strains were distributed across the major *Salmonella* serovars, namely Dublin, Typhimurium, Infantis, Kentucky, and Enteritidis, associated with disease outcomes (Supplementary Fig. S1).

3.2. Predictive modeling

3.2.1. Model selection

A supervised approach was employed to develop a model that could predict the disease phenotype in unlabeled *Salmonella* isolates based on "training" received from a pre-labeled dataset. In our study, 28 *Salmonella* isolates with known phenotypes (systemic (7), gastroenteritis (9), bacteremia (12)) were employed for model training and testing, and 205 *Salmonella* isolates (not associated with any disease phenotype) from ground chicken were employed for class label prediction and subsequently added to the labeled set for model retesting. The model predictors were *Salmonella* target genes (n = 384) associated with virulence (Supplementary Table S1). The performance of the five models was compared by the average model accuracy obtained from ten iterations. The average accuracy and validation accuracy for all 10 iterations are shown in Table 1. Random forest and logit boost were the best performing models. Though RF and LB had the highest mean accuracies, there was no significant difference ($p > 0.05$) between their means and that of other models (GBM, SVMR, SVML) (Table 1).

3.2.2. Model evaluation

Model evaluation was performed by analyzing the overall accuracy and confusion matrix statistics generated by the RF algorithm. Prior to up-sampling, the model accuracy for RF was 0.94 (95% confidence interval: 0.85–0.99), with a Kappa value of 0.82, both of which indicated a good fit. Improved accuracy was obtained by up-sampling the reduced phenotypic classes with replacement. Results showed that up-sampling improved the model performance significantly (95% confidence interval: 0.98–1.00). A Kappa value of 1.00 showed that this model performed substantially well, as described by Landis and Koch (1977). Sensitivity for bacteremia, gastroenteritis and systemic infection were 1.0, 1.0, and 0.4, respectively. The balanced accuracies for bacteremia, gastroenteritis and systemic infection were 0.86, 1, and 0.7, respectively. The final RF model was used to analyze the combined labeled and unlabeled datasets using a 70–30% train-test split and subsequently predict genetic patterns indicative of specific outcomes. Ten-fold cross-validation was applied to estimate the model performance (Pang et al., 2018; Njage, Henri, et al., 2019, Njage, Leekitcharoenphon, & Hald, 2019). Model accuracy was 0.99 (95% confidence interval: 0.98, 0.99), obtained after 11 iterations. The Kappa value was 0.99, indicating our model performed well according to Landis & Koch, (1977) or excellent as proposed by Fleiss et al. (2003). Sensitivity and specificity values were between 0.98 and 1. These results were consistent with the results of other tested models based on 10-fold cross-validation.

Table 1
Model performance of different machine learning algorithms.

Model	GBM	RF	SVMR	SVML	LB
Average Accuracy	0.988 ± 0.01[a]	0.992 ± 0.09[a]	0.977 ± 0.1[a]	0.986 ± 0.007[a]	0.991 ± 0.01[a]
Valid accuracy	0.993	0.993	0.974	0.987	1

GBM = gradient boosting.
RF = random forest.
SVMR = support vector machine with radial kernel.
SVML = support vector machine with linear kernel.
LB = logit boost.
[a] Denotes that there was no significant difference ($p > 0.05$).

3.3. Disease phenotype prediction

The best performing model (RF algorithm) was used to predict disease phenotype in 87% (178) of our included samples. The main predicted disease phenotype was bacteremia, comprising 87.64% of the included isolates, followed by gastroenteritis (6.74%) and systemic infection (5.62%). Furthermore, a plot of the probabilities of each isolate *i* belonging to disease endpoint j indicated that outcome bacteremia was the most observed compared to gastroenteritis and systemic infection (Fig. 2) as captured by the model.

3.4. Identification of important virulence gene predictors

The top twenty predictor variables sorted by maximum importance across the classes are shown in Table 2. Results indicated that a number of virulence determinants were predicted to be important features for gastroenteritis, bacteremia, and systemic infection (Table 2). Genes coding for virulence factors involved in invasion, adhesion, ecological competition, transcriptional regulators, and antibiotic resistance were found to be important predictors of the disease phenotype. Our results indicated that genes coding for proteins associated with *Salmonella* inter/intra competition, invasion, environmental sensors, and multidrug resistance showed higher probabilities as predictors of both gastroenteritis and systemic infections, while those coding for survival, antimicrobial resistance, and other hypothetical proteins were associated with all disease phenotypes (Table 2).

4. Discussion

Recent years have seen a significant increase in the availability of genomic data for various foodborne pathogens, primarily due to the improvement in sequencing technologies. This has provided researchers with an unprecedented opportunity to understand the genotypic and phenotypic variations in infectivity, virulence, and disease outcomes between different strains of the same bacterial species. However, there is

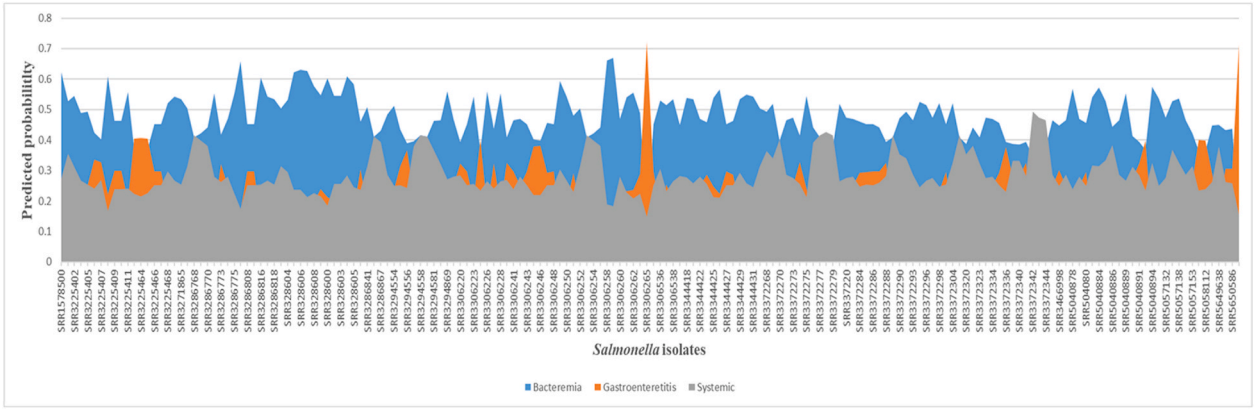

Fig. 2. Distribution of individual disease phenotype among isolates. A higher number of *Salmonella* isolates were highly predictive of the bacteremia (blue) disease phenotype, compared to gastroenteritis (orange) and systemic infection (gray), based on target gene expression. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Top twenty most important variables/genes identified by the random forest model, sorted by maximum importance across the classes.

Gene Name	Protein Name
invJ	Antigen presentation protein SpaN
hilE	Negative regulator that acts upon HilD
sopE	Guanine nucleotide exchange factor SopE
dgcJ (YeaJ)	Diguanylate cyclase
arcR	Potential AcrAB operon repressor
yigG	Hypothetical protein
elaD	Putative cytoplasmic protein
sseJ	Secreted effector protein
mgrB	PhoP/PhoQ regulator MgrB
murA	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase
ramR	Negative regulator the MarAB
bacA	Bacitracin resistance
sigD/sopB	Effector protein (inositol phosphate phosphatase)
ramA	Transcriptional activator RamA
cpxA	Two-component system sensor histidine kinase CpxA
mdtB	Multidrug transporter subunit MdtB
cysB	Transcriptional regulator
mdsA	Putative cation efflux pump
mdtG	Multidrug transporter subunit MdtG
gyrA	DNA gyrase subunit A

Note: The numbers denote importance based on accuracies of the prediction of the disease phenotypes by each feature (target gene).

a lack of metadata and useful biological information associated with the sequenced bacterial strains, which has hindered their widespread use by researchers. Based on patterns recognized from human cases with defined (tagged) endpoints, researchers are trying to link potential phenotypic outcomes to unlabeled sets of genetic data obtained from food sources. Here, we have utilized a machine learning strategy to predict the outcomes in unlabeled data, and subsequently employ such data to identify genomic patterns associated with specific phenotypic outcomes (in this case, disease outcome).

Generating WGS data for important microorganisms has prompted widespread interest in ML since genomic data provides rich information for ML algorithms to extract essential patterns and build predictive models (L'Heureux et al., 2017; Chen et al., 2020). In this study, the best performing model (RF) predicted the disease phenotype accurately in 87% of our included samples. Phenotype bacteremia was predicted more often compared to gastroenteritis and systemic infection in part because of the class imbalance in the dataset used in this study. While studies indicate that non-typhoidal *Salmonella* (NTS) serovars could have broad host ranges or elicit specific disease symptoms in different hosts (McClelland et al., 2001; Jajere, 2019), gastroenteritis accounts for a majority of the NTS-related illnesses (Majowicz et al., 2010).

Genomic comparisons among bacterial strains can reveal their intrinsic similarities and differences, enabling a better understanding of the observed phenotypic traits (Edwards et al., 2002; Amavisit et al., 2003; Mohammed & Cormican, 2016). The availability of massive amounts of molecular data due to the advent and increased usage of WGS allows unraveling certain genetic patterns and manifested phenotypic traits in microorganisms. However, our incomplete knowledge of the complex molecular mechanisms employed by pathogenic microorganisms to cause infections poses a major challenge in translating such genotypic data to associated phenotypic traits (Brul et al., 2012; Tebani et al., 2016; Haddad et al., 2018). Advanced machine learning offers us an unprecedented opportunity to interpret these large and complex molecular datasets (Libbrecht & Noble, 2015; Haddad et al., 2018; van Heyningen, 2019). Normally, if well trained, ML models can 'learn' or 'recognize' important genotypic patterns in a dataset associated with a given phenotypic trait (Libbrecht & Noble, 2015; Farrell et al., 2018; Wheeler et al., 2018). This study suggests that machine learning in support of genomic-based microbial risk assessment can predict *Salmonella* phenotypic outcomes from a given WGS data.

The ability to generate genomic data and to build predictive models based on such large-scale data has been complex (Jordan & Mitchell, 2015; L'Heureux et al., 2017; Zhou et al., 2017). In the current study, we utilized machine learning to extrapolate patterns onto genomic data that had no associated disease outcomes and predicted the possible outcomes. It is important to note that predicting disease outcomes based on complex molecular data has been difficult due to the lack of consistent and usable metadata. Therefore, our findings demonstrate that it may be possible to use ML for predictive modeling by utilizing untapped genomic data.

In this study, the average accuracies of different ML algorithms were compared, and the best-performing algorithm was used to train, test, and analyze our untagged dataset. The use of various ML algorithms in predicting biological endpoints or disease phenotypic outcomes has been fast gaining momentum (Libbrecht & Noble, 2015; Zhou et al., 2017; Njage, Henri, et al., 2019; Njage, Leekitcharoenphon, et al., 2019). Particularly, these models have been employed in making sense of genetic data from a wide range of sources (Kuhn, 2008; Machado et al., 2015). Here, we chose to test these models to classify *Salmonella* into one of three disease phenotypes or endpoints based on its genetic composition. Our results showed that RF had the highest average accuracy, followed by LB, although the difference was not statistically significant ($p > 0.05$). The sensitivity, specificity, and balanced accuracy of disease phenotypic outcome prediction by the final RF model ranged between 0.4 and 1, which was comparable to performances of ML techniques in the prediction of microbial infection outcomes based on WGS as reported previously (Njage, Henri, et al., 2019, Njage,

Leekitcharoenphon, et al., 2019). The prediction of phenotypic outcomes is often a major objective of studies with molecular data, as seen in a previous study predicting cancer subtypes using genomic markers (Wu et al., 2003). RF is a popular tree-based ensemble ML tool because it is highly data-adaptive, applicable to "large p (predictor variables), small n (observations)" problems, able to account for correlation as well as interactions among features (Breiman, 2001; Lin & Jeon, 2006). Therefore, it is a suitable tool used to predict phenotypic outcomes using genomic data. In our study, RF was reasonably accurate in classifications based on virulence (target) genes influencing disease phenotype in *Salmonella* serovars.

Genomic comparisons between *de novo* assembled sequences of the 205 *Salmonella* isolates isolated from ground chicken revealed some genetic similarities and differences. As expected, more than 90% of the *Salmonella* genes are part of the core genome and other genes are acquired through horizontal gene transfer, which might be for better adaptability and diversity among the serovars (McClelland et al., 2001; Parkhill et al., 2001; Amavisit et al., 2003; Fu et al., 2015; Gupta et al., 2019; Chen et al., 2020). Even though *Salmonella* serovars are closely related genetically, there are slight gene variations between them (Amavisit et al., 2003). Significant genetic changes are usually observed in the "variable" genome, with minor single nucleotide polymorphisms also causing differences in overall *Salmonella* pathogenicity and survival (Chen et al., 2020). Many *Salmonella* serovars cause specific disease symptoms in different hosts or have different host ranges (McClelland et al., 2001; Jajere, 2019). *S.* Typhimurium and *S.* Enteritidis, for instance, have a wide host range and infect mice, humans, and chicken, causing either bacteremia, systemic infection, or gastroenteritis (Majowicz et al., 2010; Mohammed & Cormican, 2016; Cao et al., 2020; Calero-Cáceres et al., 2020). *S.* Typhi on the other hand is host-specific and only infects humans, causing typhoid (Parkhill et al., 2001). *Salmonella enterica* is an extremely diverse species, comprising more than 2500 named serovars, designated for their distinctive antigen presentation and pathogenicity profiles. Some *Salmonella* species are known to be virulent pathogens, while not colonizing/infecting humans (Timme et al., 2013). A phylogenetic analysis was performed to determine the diversity of our isolates. Our results indicated that both model building and prediction data sets were distributed across the major *Salmonella* serovars previously associated with human disease outcomes. The machine learning model predicted unique virulence features that may explain the difference in phenotypic traits observed among the *Salmonella* isolates. These predictors enable a better understanding of the pathogens' molecular complexity and phenotypic outcomes (den Besten et al., 2018; Fritsch et al., 2019). In addition to accurate prediction of disease phenotypes, ML enables feature selection by identifying sub-sets of virulence genes whose expression patterns may significantly correlate with different disease phenotypes.

Mutation of the *invJ* gene (antigen presentation gene, associated with gastroenteritis) rendered *S.* Typhimurium mutants defective for entry into cultured epithelial cells; however, these mutants were not affected in their ability to adhere to epithelial cells (Collazo et al., 1995). Similarly, the HilE protein, which was associated with gastroenteritis and systemic infection, interacts with the HilD protein to negatively regulate the expression of the *hilA* gene, and also plays a role in the invasive phenotype in *S.* Typhimurium (Baxter et al., 2003; Lou et al., 2019). SopE is an effector protein that contributes to intracellular replication, promotes bacterial entry, and induces inflammation, and diarrhea (Wood et al., 1996; Humphreys et al., 2012). Interestingly, we found that *sseJ* and *yeaJ* genes predicted systemic disease and gastroenteritis endpoints, which corresponded well with the literature. For instance, the putative *sseJ* and *yeaJ* genes are required by *S.* Typhimurium to initiate infection and, unexpectedly, to persist systemically within the host (Lawley et al., 2006). Additionally, the *sopB* gene is required for membrane fission and damage to epithelial barrier function during the invasion (Raffatellu et al., 2005; Zhang et al., 2002). In our study, this gene predicted both systemic infection and gastroenteritis. Finally, *murA*, *mdtB*, *mdsA*, *mdtG*, and *bacA* are important genes previously identified as conferring antimicrobial resistance to *Salmonella* (Jebastin & Narayanan, 2019; Nishino et al., 2007).

Salmonella pathogenicity islands (SPIs) encode many genes that are responsible for pathogenicity (Jajere, 2019). Our results indicated that target genes were distributed across SPI-1-5. SPI-1 genes are primarily required for bacterial invasion into epithelial cells of the intestine, while SPI-2, 3, and 4 are primarily needed for bacterial growth and survival within the host, manifesting in the systemic phase of the disease. The SPI-5 virulence genes were recently shown to mediate the inflammation and chloride secretion characterizing the enteric phase of the disease (Marcus et al., 2000; Sırıken, 2013). Successful invasion and colonization of *Salmonella* in humans can result in a number of clinical outcomes ranging from moderate (gastroenteritis) to severe (systemic infection) (Edwards et al., 2002; Jajere, 2019; Kadhim, 2020; Mohammed & Cormican, 2016). These different clinical outcomes have been previously associated with the expression of different sets of genes (Jajere, 2019; Mohammed & Cormican, 2016). Our results indicate a possible association between clinical outcomes and target genes, as the prediction algorithm indicated that different sets of target genes were responsible for either a specific disease or more than one disease outcome in humans, which is in line with the results presented in prior studies. Therefore, these findings and models provide a step towards linking genotypic traits of any sequenced *Salmonella* isolates extracted from food source to disease phenotypes. Additionally, the identified variable importance genes would be helpful as potential features to provide more insights into predictive modeling and risk assessment studies.

In this study, we made an initial attempt in utilizing a semi-supervised learning approach to predicting clinical outcomes of *Salmonella* infection using whole genome sequencing data. It is evident from prior studies (Wheeler et al., 2018; Njage, Henri, et al., 2019, Njage, Leekitcharoenphon, et al., 2019) that ML can be used to predict disease outcomes given a large amount of standardized genomic data. Currently, predictive model development is hindered by the lack of genomic data with standardized data. Whereas in this study, we have utilized a semi-supervised ML method to generate reasonable associations between gene presence/absence and clinical outcome using unlabeled data, our results are dependent on the available data. Due to our stringent search criteria for the labeled dataset for initial model development, our initial model was built on a small sample set and as such the results should be interpreted with caution. Despite this, we consider our approach is reasonable, and the associations would be strengthened with the availability of standardized phenotype data or metadata in the future.

In conclusion, we utilized a machine learning model for predicting outcomes from untagged genomic data extracted from ground chicken. Food safety issues have a considerable impact on public health. Machine learning techniques can be used to identify patterns and trends impacting food safety. Genome sequencing is increasingly being used in the field of food safety, specifically in microbial tracking and outbreak investigation. Such data, combined with the availability of standardized metadata, could be analyzed by ML models to predict antibiotic resistance patterns, pathogen source attribution, foodborne outbreak investigation, and risk assessment in the future.

CRediT authorship contribution statement

Collins K. Tanui: Conceptualization, Methodology, Data curation, Formal analysis, Investigation, Writing – original draft, preparation, Visualization. **Shraddha Karanth:** Conceptualization, Methodology, Writing – review & editing. **Patrick M.K. Njage:** Writing – review & editing. **Jianghong Meng:** 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.

Acknowledgments

This study was supported through a grant from the Maryland Agricultural Experiment Station (MAES) and in part 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.lwt.2021.112701.

References

Alkema, W., Boekhorst, J., Wels, M., & Van Hijum, S. A. F. T. (2016). Microbial bioinformatics for food safety and production. *Briefings in Bioinformatics, 17*, 283–292. https://doi.org/10.1093/bib/bbv034

Altman, D. G., & Bland, j. M. (1994). Statistics notes: Diagnostic tests 2: Predictive values. *BMJ, 309*, 102. https://doi.org/10.1136/bmj.309.6947.102

Amavisit, P., Lightfoot, D., Browning, G. F., & Markham, P. F. (2003). Variation between pathogenic serovars within *Salmonella* pathogenicity islands. *Journal of Bacteriology, 185*, 3624–3635. https://doi.org/10.1128/JB.185.12.3624-3635.2003

Angelo, K. M., Chu, A., Anand, M., et al. (2015). Outbreak of *Salmonella* Newport infections linked to cucumbers–United States, 2014. *MMWR Morb Mortal Wkly Rep, 64*, 144–147.

Aziz, R. K., Bartels, D., Best, A. A., et al. (2008). The RAST server: Rapid annotations using Subsystems Technology. *BMC Genomics, 9*, 75. https://doi.org/10.1186/1471-2164-9-75

Baker, R. E., Peña, J.-M., Jayamohan, J., & Jérusalem, A. (2018). Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biology Letters, 14*, 20170660. https://doi.org/10.1098/rsbl.2017.0660

Bankevich, A., Nurk, S., Antipov, D., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology, 19*, 455–477. https://doi.org/10.1089/cmb.2012.0021

Baxter, M. A., Fahlen, T. F., Wilson, R. L., & Jones, B. D. (2003). HilE interacts with *hilD* and negatively regulates *hilA* transcription and expression of the *Salmonella enterica* serovar Typhimurium invasive phenotype. *Infection and Immunity, 71*, 1295–1305. https://doi.org/10.1128/IAI.71.3.1295-1305.2003

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., , … Stevens, R., et al. (2020). The PATRIC bioinformatics Resource center: Expanding data and analysis capabilities. *Nucleic Acids Research, 48*, D606–D612. https://doi.org/10.1093/nar/gkz943

den Besten, H. M. W., Amézquita, A., Bover-Cid, S., et al. (2018). Next generation of microbiological risk assessment: Potential of omics data for exposure assessment. *International Journal of Food Microbiology, 287*, 18–27. https://doi.org/10.1016/J.IJFOODMICRO.2017.10.006

Braden, C. R. (2006). *Salmonella enterica* serotype Enteritidis and eggs: A national epidemic in the United States. *Clinical Infectious Diseases, 43*, 512–517. https://doi.org/10.1086/505973

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Brul, S., Bassett, J., Cook, P., et al. (2012). "Omics" technologies in quantitative microbial risk assessment. *Trends in Food Science & Technology, 27*, 12–24. https://doi.org/10.1016/j.tifs.2012.04.004

Calero-Cáceres, W., Villacís, J., Ishida, M., et al. (2020). Whole-genome sequencing of *Salmonella enterica* serovar Infantis and Kentucky isolates obtained from layer poultry farms in Ecuador. *Microbiol Resour Announc, 9*. https://doi.org/10.1128/mra.00091-20

Cao, G., Balkey, M., Jin, Q., et al. (2020). Draft genome sequences of 30 *Salmonella enterica* serovar Enteritidis isolates associated with multiple outbreaks in Brazil. *Microbiol Resour Announc, 9*. https://doi.org/10.1128/mra.01580-19

CDC. (2016). *Outbreaks involving Salmonella | CDC*. CDC. https://www.cdc.gov/Salmonella/outbreaks.html. (Accessed 14 September 2020) Accessed.

CDC. (2021). *Outbreaks involving Salmonella | CDC*. CDC. https://www.cdc.gov/Salmonella/outbreaks.html. (Accessed 25 February 2021) Accessed.

Chen, X., Bauermeister, L. J., Hill, G. N., et al. (2014). Efficacy of various antimicrobials on reduction of *Salmonella* and *Campylobacter* and quality attributes of ground chicken obtained from poultry parts treated in a postchill decontamination tank. *Journal of Food Protection, 77*, 1882–1888. https://doi.org/10.4315/0362-028X.JFP-14-114

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. *J Agric Food Res, 2*, 100045. https://doi.org/10.1016/j.jafr.2020.100045

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface, 15*. https://doi.org/10.1098/rsif.2017.0387

Collazo, C. M., Zierler, M. K., & Gatan, J. E. (1995). Functional analysis of the *Salmonella* Typhimurium invasion genes *invl* and *invJ* and identification of a target of the protein secretion apparatus encoded in the inv locus. *Molecular Microbiology, 15*, 25–38. https://doi.org/10.1111/j.1365-2958.1995.tb02218.x

Danyluk, M. D., Jones, T. M., Abd, S. J., et al. (2007). Prevalence and amounts of *Salmonella* found on raw California almonds. *Journal of Food Protection, 70*, 820–827. https://doi.org/10.4315/0362-028X-70.4.820

Edwards, R. A., Olsen, G. J., & Maloy, S. R. (2002). Comparative genomics of closely related *Salmonellae*. *Trends in Microbiology, 10*, 94–99.

Farrell, F., Soyer, O., & Quince, C. (2018). Machine learning based prediction of functional capabilities in metagenomically assembled microbial genomes. *bioRxiv*. https://doi.org/10.1101/307157, 307157.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for Rates and Proportions*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Fritsch, L., Felten, A., Palma, F., et al. (2019). 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. https://doi.org/10.1016/J.IJFOODMICRO.2018.11.028

Fu, S., Octavia, S., Tanaka, M. M., et al. (2015). Defining the core genome of *Salmonella enterica* serovar Typhimurium for genomic surveillance and epidemiological typing. *Journal of Clinical Microbiology, 53*, 2530–2538. https://doi.org/10.1128/JCM.03407-14

Gupta, S. K., Sharma, P., McMillan, E. A., et al. (2019). Genomic comparison of diverse *Salmonella* serovars isolated from swine. *PLoS One, 14*, Article e0224518. https://doi.org/10.1371/journal.pone.0224518

Haddad, N., Johnson, N., Kathariou, S., et al. (2018). Next generation microbiological risk assessment—potential of omics data for hazard characterisation. *International Journal of Food Microbiology, 287*, 28–39. https://doi.org/10.1016/J.IJFOODMICRO.2018.04.015

van Heyningen, V. (2019). Genome sequencing—the dawn of a game-changing era. *Heredity, 123*, 58–66.

Horby, P. W., O'Brien, S. J., Adak, G. K., et al. (2003). A national outbreak of multi-resistant *Salmonella enterica* serovar Typhimurium definitive phage type (DT) 104 associated with consumption of lettuce. *Epidemiology and Infection, 130*, 169–178. https://doi.org/10.1017/S0950268802008063

Huang, J., Zong, Q., Zhao, F., et al. (2016). Quantitative surveys of *Salmonella* and *Campylobacter* on retail raw chicken in Yangzhou, China. *Food Control, 59*, 68–73. https://doi.org/10.1016/J.FOODCONT.2015.05.009

Humphreys, D., Davidson, A., Hume, P. J., & Koronakis, V. (2012). *Salmonella* virulence effector SopE and host GEF ARNO cooperate to recruit and activate WAVE to trigger bacterial invasion. *Cell Host & Microbe, 11*, 129–139. https://doi.org/10.1016/j.chom.2012.01.006

Jajere, S. M. (2019). A review of *Salmonella enterica* with particular focus on the pathogenicity and virulence factors, host specificity and adaptation and antimicrobial resistance including multidrug resistance. *Veterinary World, 12*, 504–521.

Jebastin, T., & Narayanan, S. (2019). In silico epitope identification of unique multidrug resistance proteins from *Salmonella* Typhi for vaccine development. *Computational Biology and Chemistry, 78*, 74–80. https://doi.org/10.1016/j.compbiolchem.2018.11.020

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*, 255–260.

Kadhim, H. M. (2020). Review of pathogenicity and virulence determinants in *Salmonella*. *EurAsian J Biosci, 14*, 377–381.

Köser, C. U., Ellington, M. J., Cartwright, E. J. P., et al. (2012). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens, 8*. https://doi.org/10.1371/journal.ppat.1002824

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Inform, 31*, 249–268.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Kursa, M. B. (2014). Robustness of Random Forest-based gene selection methods. *BMC Bioinformatics, 15*. https://doi.org/10.1186/1471-2105-15-8

Lawley, T. D., Chan, K., Thompson, L. J., et al. (2006). Genome-wide screen for *Salmonella* genes required for long-term systemic infection of the mouse. *PLoS Pathogens, 2*, 87–100. https://doi.org/10.1371/journal.ppat.0020011

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine learning with big data: Challenges and approaches. *IEEE Access, 5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics, 16*, 321–332.

Lin, Y., & Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association, 101*, 578–590. https://doi.org/10.1198/016214505000001230

Lou, L., Zhang, P., Piao, R., & Wang, Y. (2019). *Salmonella* pathogenicity island 1 (SPI-1) and its complex regulatory network. *Front. Cell. Infect. Microbiol., 9*.

Lupolova, N., Dallman, T. J., Matthews, L., et al. (2016). Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proceedings of the*

National Academy of Sciences of the U S A, 113, 11312–11317. https://doi.org/10.1073/pnas.1606567113

Machado, G., Mendoza, M. R., & Corbellini, L. G. (2015). What variables are important in predicting bovine viral diarrhea virus? A random forest approach. *Veterinary Research, 46*, 85. https://doi.org/10.1186/s13567-015-0219-7

Majowicz, S. E., Musto, J., Scallan, E., et al. (2010). The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clinical Infectious Diseases, 50*, 882–889. https://doi.org/10.1086/650733

Marcus S, L., Brumell J, H., Pfeifer C, G., & Finlay, B. B. (2000). *Salmonella* pathogenicity islands: Big virulence in small packages. *Microbes and Infection, 2*, 145–156. https://doi.org/10.1016/S1286-4579(00)00273-2

McClelland, M., Sanderson, K. E., Spieth, J., et al. (2001). Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature, 413*, 852–856. https://doi.org/10.1038/35101614

Mohammed, M., & Cormican, M. (2016). Whole genome sequencing provides insights into the genetic determinants of invasiveness in *Salmonella* Dublin. *Epidemiology and Infection, 144*, 2430–2439. https://doi.org/10.1017/S0950268816000492

Monack, D. M. (2012). *Salmonella* persistence and transmission strategies. *Current Opinion in Microbiology, 15*, 100–107.

Naugle, A. L., Barlow, K. E., Eblen, D. R., et al. (2006). Food safety and inspection service testing for *Salmonella* in selected raw meat and poultry products in the United States. *Journal of Food Protection, 69*, 2607–2614. https://doi.org/10.4315/0362-028X-69.11.2607

Nikolenko, S. I., Korobeynikov, A. I., & Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics, 14*, S7. https://doi.org/10.1186/1471-2164-14-S1-S7

Nishino, K., Nikaido, E., & Yamaguchi, A. (2007). Regulation of multidrug efflux systems involved in multidrug and metal resistance of *Salmonella enterica* serovar Typhimurium. *Journal of Bacteriology, 189*, 9066–9075. https://doi.org/10.1128/JB.01045-07

Njage, P. M. K., Henri, C., Leekitcharoenphon, P., et al. (2019a). Machine learning methods as a tool for predicting risk of illness applying next-generation sequencing data. *Risk Analysis, 39*, 1397–1413. https://doi.org/10.1111/risa.13239

Njage, P. M. K., Leekitcharoenphon, P., & Hald, T. (2019b). 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. https://doi.org/10.1016/j.ijfoodmicro.2018.11.016

Octavia, S., Zulaina, S., Seet, S. K., et al. (2019). Whole-genome sequencing of the rare *Salmonella enterica* serovar anfo isolated from food handlers. *Journal of Medical Microbiology, 68*, 429–431. https://doi.org/10.1099/jmm.0.000934

Pang, H., McEgan, R., Micallef, S. A., & Pradhan, A. K. (2018). Evaluation of meteorological factors associated with pre-harvest contamination risk of generic *Escherichia coli* in a mixed produce and dairy farm. *Food Control, 85*, 135–143. https://doi.org/10.1016/j.foodcont.2017.08.003

Parkhill, J., Dougan, G., James, K. D., et al. (2001). Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature, 413*, 848–852. https://doi.org/10.1038/35101607

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2010). *Idba – a Practical iterative de Bruijn Graph de novo assembler* (pp. 426–440). Berlin, Heidelberg: Springer.

Raffatellu, M., Wilson, R. P., Chessa, D., et al. (2005). SipA, SopA, SopB, SopD, and SopE2 contribute to *Salmonella enterica* serotype Typhimurium invasion of epithelial cells. *Infection and Immunity, 73*, 146–154. https://doi.org/10.1128/IAI.73.1.146-154.2005

Rantsiou, K., Kathariou, S., Winkler, A., et al. (2018). Next generation microbiological risk assessment: Opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. *International Journal of Food Microbiology, 287*, 3–9. https://doi.org/10.1016/j.ijfoodmicro.2017.11.007

Safae, L., Habib, B El, & Abderrahim, T. (2018). A review of machine learning algorithms for web page classification. In *Colloquium in information science and Technology* (pp. 220–226). CIST.

Scallan, E., Hoekstra, R. M., Angulo, F. J., et al. (2011). Foodborne illness acquired in the United States-Major pathogens. *Emerging Infectious Diseases, 17*, 7–15. https://doi.org/10.3201/eid1701.P11101

Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature Communications, 4*, 2304. https://doi.org/10.1038/ncomms3304

Sırıken, B. (2013). *Salmonella* pathogenicity islands. *Microbiol Bul, 47*, 181–188.

Suez, J., Porwollik, S., Dagan, A., et al. (2013). Virulence gene profiling and pathogenicity characterization of non-typhoidal *Salmonella* accounted for invasive disease in humans. *PLoS One, 8*, 58449. https://doi.org/10.1371/journal.pone.0058449

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., et al. (2001). The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research, 29*, 22–28. https://doi.org/10.1093/nar/29.1.22

Tebani, A., Afonso, C., Marret, S., & Bekri, S. (2016). Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations. *International Journal of Molecular Sciences, 17*, 1555.

Timme, R. E., Pettengill, J. B., Allard, M. W., et al. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol, 5*, 2109–2123. https://doi.org/10.1093/gbe/evt159

UniProt Consortium, T. (2018). Erratum: UniProt: The universal protein knowledgebase (nucleic acids research (2017) 45 D1 (D158-D169)). *Nucleic Acids Research, 46*, 2699.

Uzzau, S., Brown, D., Wallis, T., Rubino, S., Leori, G., Bernard, S., et al. (2000). Host adapted serotypes of Salmonella enterica. *Epidemiology and Infection, 125*(2), 229–255. https://doi.org/10.1017/S0950268899004379

Velez, D. R., White, B. C., Motsinger, A. A., et al. (2007). A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology, 31*, 306–315. https://doi.org/10.1002/gepi.20211

Wheeler, N. E., Gardner, P. P., & Barquist, L. (2018). Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. *PLoS Genetics, 14*, Article e1007333. https://doi.org/10.1371/journal.pgen.1007333

Wood, M. W., Rosqvist, R., Mullan, P. B., et al. (1996). SopE, a secreted protein of *Salmonella* Dublin, is translocated into the target eukaryotic cell via a sip-dependent mechanism and promotes bacterial entry. *Molecular Microbiology, 22*, 327–338. https://doi.org/10.1046/j.1365-2958.1996.00116.x

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., et al. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics, 19*, 1636–1643. https://doi.org/10.1093/bioinformatics/btg210

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research, 18*, 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, S., Santos, R. L., Tsolis, R. M., et al. (2002). The *Salmonella enterica* serotype Typhimurium effector proteins SipA, SopA, SopB, SopD, and SopE2 act in concert to induce diarrhea in calves. *Infection and Immunity, 70*, 3843–3855. https://doi.org/10.1128/IAI.70.7.3843-3855.2002

Zhao, C., Ge, B., De Villena, J., et al. (2001). Prevalence of *Campylobacter* spp., *Escherichia coli*, and *Salmonella* serovars in retail chicken, Turkey, pork, and beef from the Greater Washington, D.C., area. *Applied and Environmental Microbiology, 67*, 5431–5436. https://doi.org/10.1128/AEM.67.12.5431-5436.2001

Zhou, L., Pan, S., Wang, J., & Vasilakos, A. V. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing, 237*, 350–361. https://doi.org/10.1016/j.neucom.2017.01.026

Zou, W., Al-Khaldi, S. F., Branham, W. S., et al. (2011). Microarray analysis of virulence gene profiles in *Salmonella* serovars from food/food animal environment. *J Infect Dev Ctries, 5*, 94–105. https://doi.org/10.3855/jidc.1396