Title: Historical analysis of *Salmonella*: trends in outbreaks, genomics, and geographic factors

Running title: Historical analysis of *Salmonella* outbreaks

Tatum S. Katz[a#], Tatum D. Mortimer[b], Kimberly Casares[b], Brittany Henry[b], Amy T. Siceloff[b], Nikki W. Shariat[b], Dayna M. Harhay[a], Tommy L. Wheeler[a]

a. U.S. Department of Agriculture, Agricultural Research Service, Roman L. Hruska U.S. Meat Animal Research Center, Clay Center, NE, United States
b. Department of Population Health, Poultry Diagnostic and Research Center, College of Veterinary Medicine, University of Georgia, Athens, GA, United States

Corresponding author email: tatum.katz@usda.gov

Abstract word count: 136

Text word count excluding references, table footnotes, and figure legends: XX

Abstract

We live in an era where advances in microbial detection, DNA sequencing technology, and the availability of big data have allowed inroads to the field of predictive microbiology. The tools presently in use, as well as those on the horizon, provide an opportunity to examine emerging infectious diseases in ways previously not possible. We sought to develop a holistic understanding of "who, where, and why" certain *Salmonella* strains emerge to cause outbreaks and provide for the beef industry a short-list of warning signs and monitoring recommendations to predict the next outbreak before it happens. Key findings from this analysis are that outbreak serovars emerge from the existing diversity of *Salmonella* in beef; regionality has important implications for serovar management; and currently available genotyping reported by common databases do not necessarily differentiate outbreak strains from FSIS isolates.

Importance

The control of *Salmonella* outbreaks remains a recalcitrant problem in the United States. Complexity from environmental, ecological, evolutionary, and genomic factors makes it a challenge to predict (and therefore control or prevent) these costly events. Previous studies have shown that these factors can contribute to the emergence of outbreaks but have not combined them into a single holistic analysis. Here we introduce new tools for the modeling of *Salmonella* serovars such as Species Distribution Modeling (SDM) to exploit all available information on *Salmonella*. Our study identifies key findings actionable for both academia and industry; specifically, we have identified new in-roads and information gaps in addition to preliminary warning signs that a new outbreak may occur. In total, our findings deepen the current understanding of the complex and multifaceted process through which *Salmonella* causes beef-related outbreaks.

Introduction

Salmonella is a leading bacterial foodborne pathogen in the United States and worldwide. It is estimated that approximately 1.3 million people in the United States suffer from illnesses caused by non-typhoidal *Salmonella* annually (1). Consumption of contaminated beef products is implicated in 6.9% of human salmonellosis cases (2).

Salmonella is represented by over 2,600 different serovars and these can differ significantly in the pathogenicity and host specificity (3–5). In relation to beef, serovars Cerro and Newport are both commonly found in beef products (6), yet have different public health outcomes (i.e. pathogenicity) in that serovar Newport causes many more cases of foodborne illness than serovar Cerro (7). Where both serovars are prevalent, this difference can be explained by differential presence of virulence factors (8). Serovar Dublin is considered host-adapted for cattle, causing significant illness and mortality in calves. Though infrequent, when human outbreaks occur, they typically result in larger numbers of hospitalizations than other salmonellosis outbreaks. As such, a more accurate understanding and prediction of outbreaks must consider *Salmonella* at the serovar level.

Serovars that are prevalent in a certain product and that are also linked to significant human illness represent serovars of concern (SoCs) for that product. SoCs have been defined using epidemiological data (5, 9, 10), and genomic data (8, 10). Despite continued investment in reduction of *Salmonella* on beef products, including regulatory sampling by FSIS, beef-attributed *Salmonella* outbreaks have occurred at an unchanged rate over the past two decades (11).

It is likely that future *Salmonella* regulation will focus on these serovars of concern rather than *Salmonella* prevalence alone. In broilers, USDA-FSIS recently suggested regulation based on three serovars, Enteritidis, Typhimurium, and monophasic Typhimurium. Identification of analogous serovars in beef will be useful in informing industry of the most relevant serovars to target and in the development of appropriate diagnostic tools to detect these serovars. However, there are dynamic shifts in serovars over time, best represented by the emergence of serovar Infantis in poultry in recent years. Of the 42 beef outbreaks that have occurred from 2009 to 2023, 16 different serovars have been implicated (11). This suggests that in beef production as well, there is a need to re-evaluate which serovars are considered serovars of concern over time. Examining how frequently serovars change or shift within beef production has not been previously investigated, and may provide valuable insight into the regional and seasonal dynamics of SoC attributed to beef.

Over the last ten years, predictive modeling methods have gained attention for their ability to identify trends and predictors of infectious disease (12). These early warning systems can incorporate environmental data such as climate and meteorological indicators (13), disease surveillance data, and human illness burden information to identify outbreaks before they are labeled as such (12). Environmental predictors such as meteorological factors (14, 15), epidemiological predictors (16–19), and genomic predictors (15) have been used to estimate salmonellosis outbreaks, incidence, and risk. As diverse as their data sources are their tools: these systems frequently exploit the growing wealth of machine learning and artificial intelligence approaches to generate predictions about how, where, and why outbreaks occur. Techniques include random forest (14), neural networks (17), time series (18, 19), and ensemble approaches (16).

Based on these previous findings and publicly available data, we aimed to combine cross-disciplinary, advanced predictive modeling methods with *Salmonella* surveillance data in the US to identify early warning signs of emergent *Salmonella* beef outbreaks. We sought to address the following questions: (1) Do *Salmonella* serotypes of concern change over time, and if so, over what time scale? (2) Are there genetic factors that drive the emergence of *Salmonella* outbreaks? (3) Can *Salmonella* surveillance data be used to predict the occurrence of outbreaks? (4) Are there geospatial factors that drive the occurrence of the top *Salmonella* serotypes? We used a consortium of approaches including time series, geospatial, and phylogenomic analyses to address these questions.

Results

SoC lists over time.

To understand the rate at which *Salmonella* SoCs change over time, SoC lists based on both the AGNES and outlier SoC definitions of Katz et al. 2024 were generated for one-year staggered rolling windows with durations 2 to 7 years. These lists can be found in Supplementary Table A. We developed a heuristic (ΔSoC) to capture the change in SoC lists between windows while accounting for the length of the window. The higher the value of ΔSoC, the more change in SoCs you get given the duration. The ΔSoC heuristic ranged from 3.96 to 5.15 for the outlier definition, and from 0.94 to 1 for the AGNES definition (Figure 1). Based on the poor performance of the AGNES definition, we chose to move forward only with the outlier definition of SoCs. The duration with the highest heuristic value was 4 years (5.15), followed closely by 5 years (5.12, Figure 1).

Predicting the next outbreak serotype.

We calculated Krippendorf's Alpha to compare our SoC predictions against the following year's outbreaks. We estimated the Krippendorf's Alpha for a four-year moving window to be 0.14 (0.017 - 0.30, 95% confidence interval, Supplementary Table 2). The best value was obtained with a moving window duration of just 2 years, giving an Alpha of 0.22 (0.047 – 0.40 95% CI). Due to the overall low Krippendorf's Alpha values (values must exceed 0.67 to be considered a reliable predictor (20)), we proceeded to search for other possible predictors to enhance our SoC list predictions. All further analyses failed to outperform the two-year moving window with the outlier definition (Krippendorf's Alpha values ranging from negative values indicating high bias in predictions, to 0.16 for hypothesis (2) below).

Temporal indicators of emergent outbreaks.

We developed a Seasonal Auto Regressive Integrated Moving Average (SARIMA) time series model to determine if outbreaks occur on a predictable timetable. These models decompose time series data into their components: an autoregressive component describing how reliant future observations are on past observations; a moving average component describing the regression error; and a seasonal component describing periodicity in the time series (21). The best-fit SARIMA model for human beef-related outbreaks was an ARIMA(0,0,0), also known as a white noise process. This indicates that outbreaks happen randomly and unpredictably. In contrast, the raw beef time series describing the number of positive *Salmonella* FSIS samples produced a best-fit model of

SARIMA(0,1,7)(0,1,12). This model indicates no autoregression, but two seasonal components on a 2 and 7 month timescale. These models and their one-year forecasts can be viewed in Supplementary Figure 1. For the ARIMA model predicting outbreaks using FSIS raw beef sample *Salmonella* positives as a predictor, a white noise ARIMA(0,0,0) was again identified as the best fit model. Interestingly, however, this model fit the outbreak data better than the model without using the raw beef samples (ΔAICc of 93.69, ΔAICc of 2 ~ p-value < 0.05).

For each year's raw beef sampling serotype communities 2017 through 2024, we calculated a series of temporal diversity indices as possible predictors of beef-related outbreaks. These indices were plotted through time and aligned against beef-related outbreaks reported in the CDC NORS dataset (Figure 2). We subsequently identified three points where two outbreaks occurred simultaneously ("worst case scenarios") and used these timepoints to evaluate the indices for predictive ability. We identified that large absolute values of the slope of serotype mean rank shift (changes in abundances) related tightly to these three timepoints, i.e., high rates of "shuffling" of serotype abundances could be a possible indicator of outbreaks.

Genome assembly, typing, and antimicrobial resistance profiling

To provide a consistent dataset of genomic features commonly reported for surveillance isolates, we developed an automated Snakemake workflow for genome assembly, annotation, and typing of sequencing data from FSIS isolates, including *in silico* serotyping and identification of antimicrobial resistance genes with AMRFinderPlus. Following assembly and serotyping, AMR gene detection using AMRFinderPlus identified a diverse repertoire of resistance alleles across *Salmonella* isolates from beef derived sources. The pipeline characterized AMR determinants at both the gene and class level, including β-lactamases, aminoglycoside modifying enzymes, tetracycline resistance genes, and efflux associated loci. Each assembly was screened against the curated AMRFinderPlus reference database to detect both acquired resistance genes and chromosomal mutations linked to antimicrobial resistance phenotypes.

The prevalence and distribution of AMR alleles varied considerably across *Salmonella* serotypes (Figure 3). Serotypes Dublin, Newport, and Typhimurium exhibited the highest overall diversity and frequency of AMR genes, reflecting their association with multidrug resistance in bovine and human cases. Common β-lactamase alleles such as $bla_{\text{TEM-1}}$ and $bla_{\text{CMY-2}}$ were enriched in these serotypes, whereas tetracycline resistance genes (*tet(A)*, *tet(B)*, and *tet(C)*) and sulfonamide resistance genes (*sul1* and *sul2*) were distributed more broadly among multiple serovars. Overall, AMRFinderPlus-based profiling revealed that

Salmonella enterica isolates from beef sources harbor a core set of resistance genes primarily targeting β-lactams, aminoglycosides, and tetracyclines.

Phylogenomic comparison of beef isolates to outbreak isolates for serovars Brandenburg and Saintpaul

To assess if standard genomic comparisons, such as the virulence and antimicrobial resistance (AMR) genotypes provided by NCBI's Pathogen Detection platform, could provide greater insight into the mechanisms underlying *Salmonella* outbreaks, we conducted a phylogenetic analysis with isolates from two previous beef-related outbreaks. Namely, serovar Brandenburg, which caused 139 illnesses in 2022, and serovar Saintpaul, responsible for 18 illnesses in a 2023 multistate outbreak. All isolates collected between 2017 and 2024 by the FSIS as part of the *Salmonella* verification program in beef that were confirmed as serovars Brandenburg or Saintpaul were included in this analysis. A single nucleotide variant (SNV) phylogeny was generated with the FSIS and human clinical outbreak isolates for both serovars, respectively (Figure 7, 8; Supplementary Figure 2, 3). The phylogenetic tree was annotated with the virulence and AMR genotypes, along with the production region each isolate was sampled from. All outbreak isolates clustered together, along with the majority of FSIS isolates, so subsequent phylogenies were generated of the outbreak clades exclusively to better explore the closeness of isolates. Interestingly, for serovar Brandenburg, one outbreak isolate possessed a different AMR genotype from the others. However, there were no further apparent patterns connecting outbreak isolates with the present metadata.

Spatial patterns of top serovars

The overall serotyping results from the FSIS regulatory beef dataset are relatively consistent between the years, with serovars Anatum, Dublin, Montevideo, and Muenchen remaining prevalent throughout (Figure 5A). The serovars shown include those present in the top 10 of at least one region between 2014 – 2024. From the final list (n = 17), all serovars were found annually, with the exception of serovars Brandenburg and Give in 2014, 2016, 2017 and 2014, respectively. Serovar Montevideo was the most abundant overall (n = 609), followed by serovars Anatum (n = 270), Muenchen (n = 232), and Dublin (n = 214). To consider the impact of each beef production region (Figure 5B), the FSIS surveillance results were divided respectively. Region 8 has both the greatest number of states (n = 16) and, subsequently, number of processing establishments sampled (n = 849). Conversely, Region 1 had the lowest number of establishments (n = 119) with three states included. The

serotyping results had greater variance year-to-year when split across the regions, and some regional trends were identified (Figure 5C). For example, 73% (37/51) of Uganda isolates were collected in regions 5, 6, and 8, which suggests a possible northeast signal. Additionally, 59% (35/59) of Give isolates originated from regions 7 and 8, representing the east coast. Of the I 4,[5],12:i:- isolates, 44% (24/54) were found in Region 8.

Geospatial predictors of top serovars

To examine the possibility of geographic and environmental predictors of *Salmonella* serovar, we utilized Species Distribution Models (SDMs). These models use occurrence data of a species (here, a *Salmonella* serovar) and gridded, geospatial predictors to calculate the probability of the given species' occurrence across the landscape (22). Areas with similar climate to the occurrence data are typically scored with high probability of occurrence, and dissimilar areas are scored with low probability of occurrence. Using the developed SDMs, we generated predictions for the distributions of serovars Agona, Anatum, Cerro, Dublin, Give, I 4,[5],12:i:-, Infantis, Kentucky, Mbandaka, Meleagridis, Montevideo, Muenchen, Muenster, Newport, and Typhimurium for the contiguous United States (Figure 4). Serovar distribution predictions were analyzed for overlap (e.g., how similar are the predicted regions). Overlap values were generally very high, with overlaps ranging from 84% to 98%, indicating that the 15 serovars are largely co-occurrent. However, visual inspection of the predictions reveals important spatial differences not identified by niche overlap calculations (Figure 4).

Overall, the most important predictors (contributing at least 10% of the prediction) for the presence of any of the 15 serovars were, in descending order of frequency, the number of cattle operations, precipitation of the warmest quarter, and annual mean temperature. Serovar distribution models were then separated into "SoC" and "non-SoC" categories to look for differences in important predictors for these two groups. Precipitation of the warmest quarter and the number of cattle operations were frequently important predictors regardless of SoC classification. Only SoC models included precipitation of the wettest quarter as an important predictor, while only non-SoC models included mean temperature of the wettest quarter, annual temperature range, precipitation seasonality, and the number of poultry operations as important predictors (Figure 9).

Sequencing of historical Salmonella *genomes and comparison to FSIS isolates.*

To understand genomic changes of *Salmonella* serovars over time, we used short- and long-read sequencing to generate complete assemblies for 24 isolates (Supplementary

Table C). All assemblies included a single, circular contig to represent the chromosome, and CheckM (23) estimated genome completeness between 98.73-100%. Using PopPunk (24), we found that 22/24 complete genomes clustered with FSIS isolates. SARB27 and SARB56 did not cluster with isolates collected by FSIS. SARB27 (serovar Infantis) also did not cluster with any publicly available complete genomes and represents a novel genotype among complete genomes. SARB27 was collected in the 1970s in Senegal.

Our historical isolates from serovar Newport were assigned to three distinct clusters, and they nested within contemporary diversity from FSIS isolates (Supplementary Figure 4). Similarly, isolates from serovar Saintpaul were assigned to two distinct clusters, and our historical isolates nested within FSIS diversity (Supplementary Figure 5). Serovars Infantis and Anatum were represented by a single cluster, and historical isolates were not distinct from diversity represented by FSIS isolates (Supplementary Figures 6,7). However, SARB13, a serovar Dublin isolate collected in France in 1982, was genetically distinct from serovar Dublin isolates collected by FSIS and the sequenced historical isolates, SARB12 and SM-73-1, that were collected in the United States (Supplementary Figure 8).

Discussion and Conclusions

We examined a diversity of factors to understand the "who, where, and why" of beef-related salmonellosis outbreaks. We explored geographic, epidemiological, ecological, genetic, and evolutionary patterns to identify potential warning signs of future outbreaks. Predicting outbreaks before they happen is a significant challenge due to complex interactions between many factors (13, 14, 25–30). By examining these factors in a holistic way, we can begin to draw conclusions about outbreak predictors that can be exploited to manage outbreaks before they occur.

As *Salmonella* management perspective shifts from total elimination to strategic targeting of the most dangerous serovars, multiple methods of identifying these serovars of concern (SoCs) have emerged (5, 9, 10). We evaluated SoC determination methods for their ability to identify future outbreak serovars, rather than inform on what has already occurred. SoC lists are useful for retrospective risk analysis but have poor predictive power to determine future outbreak serovars when calculated in a moving-window fashion. However, no other hypotheses tested here had higher predictive power than the Katz et al. 2024 (5) outlier rule over a two-year moving window, therefore, we propose this rule as the current best estimate of what serovars may cause beef-related outbreaks in the coming years. Outbreaks are a complex process that cannot be fully described by FSIS sample data and

CDC outbreak data. There is a hidden process that occurs between these two data sources that we are unable to observe, which is likely driving the bulk of the variance in which serovars emerge to cause outbreaks. Identification and then monitoring of this hidden process is required to accurately predict future outbreaks.

Our time series analysis further highlights the existence and importance of this hidden process. While this analysis showed increased *Salmonella* detection on raw beef samples by FSIS appears to happen on a 2- and 7-month cycle and is an important predictor of human beef-related salmonellosis outbreaks, we also observed that outbreaks are best modeled by a white noise process. This indicates that factors which occur between FSIS raw beef sampling and consumption of the product are both critical in determining outbreak timing and are highly variable, resulting in the apparent randomness of outbreaks. Further information on this hidden process is likely to dramatically improve timing prediction of future outbreaks.

Spatial analyses proved to be more fruitful than time series analyses. Our results demonstrate that *Salmonella* serovars in beef appear to be associated with particular geographic regions. These regional trends may be explained in part by variables other than cattle operation distribution, such as precipitation, average temperature during the wettest seasons, and average yearly temperature. While we found there is overlap in the distribution of most serovars, distinct differences in regionality can inform the meat industry on which serovars should be targeted specifically for their locality. Understanding regional trends empowers the industry to develop serovar-specific management strategies to focus resources on the most likely serovars of concern in their production systems.

We found that beef-related *Salmonella* serovar diversity appears consistent; emergence or extirpation of serovars are extremely rare events which could be explained by limited sampling. Serovar communities usually have consistent membership. The biggest changes over time are in the relative abundances of the member serovars. Yet, abundance of a given serovar does not appear to be a strong predictor that it will cause an outbreak in the future. Outbreak strains seem to emerge from pre-existing diversity in beef; specifically, we found no evidence of a new strain emerging around the time of outbreaks. However, evidence suggests that a high rate of “shuffling” of serovar abundances, or mean rank shift, within raw beef sampling data, may predict periods of multiple simultaneous outbreaks.

Phylogenetic and genomic analyses demonstrated that historical outbreak isolates collected in North America nest within diversity of FSIS isolates, indicating that diversity is maintained over time within geographic regions with no obvious clonal replacements. However, we did find that historical isolates collected outside North America (SARB27 and SARB13) are genetically distinct from historical and FSIS isolates of the same serovar,

indicating novel genetic components may be introduced via international transmission. The variable prevalence of AMR-related alleles across serovars highlights the importance of integrating AMR gene data with epidemiological context to monitor the emergence and persistence of resistant lineages within beef production.

In summary, our results demonstrate that SoCs may not so much emerge as they rotate. SoC lists should be updated every two years. While a two-year update will not capture all possible SoCs and may miss critical serovars, this is the best method we found for prediction of future outbreak serovars. Prediction of the timing of new outbreaks remains challenging without data on the hidden process that occurs between FSIS sampling and CDC outbreak detection, yet high incidence of *Salmonella*-positive raw beef samples is an important predictor of outbreaks. Additionally, a high rate of change of mean rank shift year-to-year of serovar communities may help predict time periods with multiple significant outbreaks. We found evidence that regionality is important in determining which serovars are most likely to be detected in raw beef sampling, allowing producers to focus their resources on serovars specific to their locality. Finally, we found that genotyping reported by common databases like Pathogen Detection do not necessarily differentiate outbreak strains from FSIS isolates and that more detailed genomic analysis is needed to understand the genomic mechanisms that lead to outbreak strain emergence.

Methods

Data.

To collect information on beef-related outbreaks, we downloaded data on beef-related Salmonella outbreaks from 2009 to 2024 from the Centers for Disease Control's National Outbreak Reporting System (11) on January 15th, 2025. These data describe outbreaks, defined as two or more cases of similar illnesses from a common exposure (11). Initially, there were 66,713 unique outbreaks in the dataset. We cleaned the data to include only outbreaks of single serovars of *Salmonella enterica* whose exposure route was identified as a beef food item, resulting in 42 outbreaks for analysis.

To understand *Salmonella* populations on raw beef products, we downloaded data from the Raw Beef Sampling program from USDA's Food Safety Inspection Service (FSIS) on February 12th, 2025 (6). These data describe regulatory sampling of pathogens from various raw beef products across establishments in the U.S. From 197,913 total observations, we selected only those samples which were analyzed for *Salmonella*, providing 184,370 samples for analysis.

SoC lists over time.

We generated SoC lists for moving windows of 2 to 7 years using the CDC NORS data and the SoC definitions of Katz et al. 2024. First, we subsetted the NORS data repeatedly for overlapping moving windows of lengths 2 through 7 years. For example, for the two year moving windows, we identified outbreaks occurring in the windows of 2009 to 2010, 2010 to 2011, 2011 to 2012, and so on. For the seven year moving windows, we identified outbreaks in the windows of 2009 to 2015, 2010 to 2016, 2011 to 2017, and so on. For each of these subsets of the NORS outbreak data, we then used the AGNES and outlier SoC definitions from (5) to identify the SoCs for each moving window period.

To evaluate differences between moving window durations, we calculated a heuristic (ΔSoC) as the average Manhattan distances between the SoC lists for each duration, added it to the average number of SoCs identified in each window, and divided by the window duration. When the value of this heuristic is equal to 1, that represents a relative change of one SoC member per year. The larger this value becomes, the more change in SoCs you have between windows, relative to the duration of the window. The intent of this heuristic is to identify the shortest moving window duration which provides the most change in SoCs. We calculated this value for each duration of 2 to 7 years. The duration with the highest value for this heuristic was selected as the best duration.

Predicting the next outbreak serovar

Following identification of the best duration of moving windows for the reclassification of SoC lists, we sought to determine if our duration-based lists could also be used for prediction of future outbreaks. For each window in the best duration, we compared the list of SoCs to outbreaks in the following year using CDC NORS data. For example, if the window ended in 2009, we would compare the SoC list to outbreaks from the year 2010. We compared the SoC list against the outbreak list for agreement using Krippendorff's Alpha, a multilabel and multiclass rater agreement score not unlike Cohen's Kappa (31). Values of 1 indicate perfect agreement, while values of 0 indicate complete lack of agreement (31).

To augment the predictability of the moving window SoC lists described above, we examined the Raw Beef FSIS sampling data to see if it could enhance our predictive power. For this analysis, we further limited the FSIS data to domestic sampling only (project codes MT60, MT60_C, MT64, MT64_C, MT65, MT65_C, and MT43, (32)), resulting in 3251 positive *Salmonella* samples for analysis. We eliminated follow-up domestic samples, as these are no longer unbiased detections, but rather intensive sampling done at a specific facility

following an initial positive sample. We tested the following hypotheses about the process by which new outbreaks emerge: (1) high variance in the number of detections for a given serovar can predict the next outbreak (autocorrelation analysis) (2) the serovar which has the maximum relative abundance can predict the next outbreak; (3) the dominant O-group will be the O-group of the next outbreak; (4) a serovar significantly exceeding its past relative abundance will be the next outbreak (using binomial tests); (5) within O-groups, serovars which exceed their past relative abundance will be the next outbreak. We calculated lists of SoCs to compare against real outbreaks (as above) in the same year, as well as using the best duration identified above to compare against the following year, and finally for a duration of 2 years to predict the following year.

Temporal indicators of emergent outbreaks.

We developed time series models using both the CDC and FSIS datasets to further interrogate how outbreaks and serovar communities change over time. Using the CDC NORS data, we developed a SARIMA model on a 12-month time series. Because the NORS dataset describes detections only, we filled in zeroes for any month:year combination which had no outbreaks reported. We additionally developed a SARIMA model for Salmonella positive raw beef samples using the FSIS dataset to evaluate connections, and a SARIMA model of outbreaks which included the FSIS data as a predictor.

To understand how FSIS Salmonella serovar communities change over time, we calculated a variety of temporal diversity indices to compare against our outbreak time series using FSIS data from the Pathogen Detection Isolates Browser (https://www.ncbi.nlm.nih.gov/pathogens/isolates/, downloaded February 5th, 2025). For each year's serovar communities, we calculated serovar complexity; serovar turnover, appearance, and disappearance; serovar mean rank shift (changes in abundance of the serovars); and the community Euclidean distance to the previous year (similarity of serovar communities, year to year). We aligned these values through time to outbreaks in the CDC NORS dataset to inspect for trends and possible predictors of future outbreaks.

Spatial patterns of top serovars

To gain a better understanding of any regional serovar trends, we divided the FSIS raw beef sampling results into eight production regions, as set by The Beef Industry Food Safety Council (BIFSCo). Only samples collected as part of standard regulatory operations were included (project codes "MT43", "MT60", "MT60_C", "MT64", "MT65", and "MT65_C"), spanning from 2014 to 2024. The regions were defined as follows: 1, Washington, Oregon,

Idaho; 2, California, Nevada; 3, Arizona, New Mexico, Texas; 4, Wyoming, Montana, Utah, Colorado; 5, South Dakota, Nebraska, North Dakota, Minnesota, Wisconsin; 6, Missouri, Iowa, Kansas; 7, Arkansas, Louisiana, Alabama, Georgia, Mississippi, Florida, South Carolina, North Carolina, Oklahoma, Tennessee; 8, Michigan, Illinois, Indiana, Ohio, Kentucky, West Virginia, Maryland, Virginia, Pennsylvania, New York, New Jersey, Vermont, Maine, New Hampshire, Connecticut, Massachusetts, Rhode Island, Washington D.C., Delaware. Samples collected from outside of the continental United States were not included in the overall analysis. Only serovars that were ranked in the top 10 annually for the serotyping results are shown (n = 17).

Geospatial predictors of top serovars

We used the FSIS raw beef sampling data as our occurrence points for the Species Distribution Models (SDMs). These samples are collected at processing establishments rather than the point of origin of the animal. Therefore, we limited our occurrence samples to those associated with project codes “MT60”, “MT64”, “MT60_C”, and “MT64_C”, which are trim (60) and ground beef components (64) and are usually sourced locally (KR personal communication). We eliminated the “MT43” and “MT65” project code samples that correspond to ground beef and bench trim samples, which usually come from locations very far from the processing establishment (KR personal communication). This resulted in 1239 positive *Salmonella* isolates for analysis. We used 21 predictors for analysis, 19 of which were the WorldClim Bioclimatic variables that describe monthly temperature and rainfall values in biologically meaningful ways (33). The remaining predictors, the number of cattle operations in the county and the number of poultry operations in the county, were collected from the National Agricultural Statistics Service’s 2022 AgCencus (https://www.nass.usda.gov/Publications/AgCensus/2022/, downloaded August 26th 2025). These predictors and the occurrence points were cropped to the contiguous United States. We limited our analysis to just serovars with at least 21 occurrence points to ensure model overfitting does not occur. These 15 serovars were Agona, Anatum, Cerro, Dublin, Give, I 4,[5],12:i:-, Infantis, Kentucky, Mbandaka, Meleagridis, Montevideo, Muenchen, Muenster, Newport, and Typhimurium. For each of the 15 serovars, we used the algorithm MaxEnt (34) to generate an SDM from the provided predictors.

Genome assembly, typing and antimicrobial resistance profiling pipeline using Snakemake workflow management system.

To enable genomic and phylogeographic analyses required for identifying factors contributing to disease emergence, we developed a modular Snakemake pipeline (v7.32.3) (35) that automates the assembly, quality assessment, typing, and downstream comparative genomics of *Salmonella enterica* isolates from beef derived sources. The workflow was designed for deployment on the Georgia Advanced Computing Resource Center (GACRC) using SLURM job scheduling. Each step of the workflow is containerized through Conda environments to ensure reproducibility and portability across computing environments. All software dependencies are managed and downloaded using Miniforge (https://github.com/conda-forge/miniforge).

Raw sequencing data accessions and associated metadata were obtained from the NCBI Pathogen Detection isolates browser, filtered for *Salmonella* enterica isolates recovered from beef production environments, carcass trim, ground beef, and slaughter facilities collected from 2016 onward. The Snakemake workflow orchestrates each stage of the analysis through a directed acyclic graph (DAG), ensuring that input and output dependencies are explicitly tracked, and re-runs are defined without recomputing the entire workflow. The pipeline is configured via a YAML file specifying HPC resources (4 cores, 4000 MB per job, and a 10-hour wall time per rule) and a default partition of batch on SLURM. Each rule automatically generates benchmarking files, logs and runtime statistics for transparency and reproducibility.

Sequencing reads were downloaded using SRA toolkit, and raw reads were preprocessed using fastp (v0.23.4) (36) for adapter trimming and quality filtering prior to assembly. Then trimmed reads were assembled using SPAdes (v3.15.5) (37) with default k-mer sizes (k=21, 33, 55, 77, 99, 127) optimized for bacterial genomes. To ensure reliability of assemblies, the workflow applied coverage-based filtering and evaluated completeness metrics with QUAST (v5.2.0) (https://github.com/ablab/quast), capturing contig count, N50, total assembly length, and GC content.

Assemblies were subjected to multiple typing and functional annotation tools to capture serovar identity, sequence type, and antimicrobial resistance profiles. Serotype prediction was performed using SeqSero2 (v1.3.1) (38), which integrates k-mer-based and gene-targeted approaches to identify*Salmonella* serovars from assembled genomes. Multilocus sequence typing (MLST) was conducted using the MLST software (v2.23.0) (https://github.com/tseemann/mlst) following the Achtman scheme, assigning allelic profiles across seven housekeeping genes to provide insights into lineage and population structure. Antimicrobial resistance (AMR) genes were identified using AMRFinderPlus (v3.12.8) (39), which detects acquired resistance genes, point mutations, and stress response elements through curated reference databases.

Whole genome sequencing, assembly, and typing of historical Salmonella *isolates.*

Twenty-four *Salmonella enterica* isolates were streaked for isolation on TSA from frozen stocks. One colony was inoculated into 10 mL TSB and incubated at 37°C. High molecular weight DNA was extracted with Qiagen Genomic DNA buffer kits and 100g columns. DNA was reconstituted in 250 µL sterile Low TE (10 mM TRIS, 0.1 mM EDTA pH 8.0).

Long-read sequencing was performed using Oxford Nanopore Technologies (ONT). DNA was prepared for sequencing using the Native Barcoding Kit 24 V14 (SQK-NBD114.24). The library was sequenced with a FLO-MIN114 flow cell in a MinION Mk1B device with MinKNOW v 25.05.14.

Paired-end, Illumina sequencing was performed at SeqCoast Genomics. Libraries were prepared using the Illumina DNA Prep tagmentation kit with Illumina Unique Dual Indexes. Libraries were sequenced on the Illumina NextSeq2000, which produced 2X150bp paired-end reads. Demultiplexing, trimming, and run analytics were performed using DRAGEN v4.2.7.

ONT reads were basecalled and demultiplexed using dorado v 0.8.3 (https://github.com/nanoporetech/dorado) with the dna_r10.4.1_e8.2_400bps_sup@v5.0.0 model. We used Filtlong v 0.2.1 (https://github.com/rrwick/Filtlong) to remove reads less than 1000 nucleotides in length and 10% of reads with the lowest quality. Filtered ONT reads were assembled using Autocycler v 0.5.0 (40), which subsampled reads and generated a consensus assembly from assemblies produced by Canu v 2.3 (41), Flye v 0.3-r179 (42), Miniasm v 0.3-r179 (43), NextDenovo v 2.5.2 (44), and Raven v 1.8.3 (45). Assemblies were polished with Illumina reads using pypolca v 0.4.0 (46). Contigs less than 500 nucleotides in length were removed from final assemblies. Assembly completeness was assessed with CheckM v 1.2.4 (23) using the lineage workflow.

Assemblies were annotated with bakta v 1.10.1 (47). We used AMRFinderPlus v 4.0.22 (39) with the curated *Salmonella* database to identify antimicrobial resistance and virulence associated genes and alleles. Plasmids were identified and typed using MOB-suite v 3.1.9 (48). In silico serotyping was performed from assemblies using SeqSero2 v 1.3.1 (38) with the k-mer workflow.

Single nucleotide variant (SNV) phylogeny of serovar Brandenburg and Saintpaul beef regulatory and human clinical outbreak isolates.

The sampling results for the FSIS *Salmonella* surveillance program of raw beef products for fiscal years 2014 – 2024 were downloaded on June 4, 2025 (https://www.fsis.usda.gov/science-data/data-sets-visualizations/laboratory-sampling-data). All *Salmonella*-positive isolates with WGS data available that were serotyped as either serovar Brandenburg (n = 47) or serovar Saintpaul (n = 12) were chosen for this analysis. For genomic comparison to beef-related outbreak isolates of these two serovars, five isolates of the outbreak strain were included as well. The paired-end reads for all selected isolates were downloaded from the NCBI Sequence Read Archive, then *de novo* assembled using SPAdes v4.1.0 (37) and annotated with bakta v1.11.0 (47) using the Prodigal training file for *Salmonella enterica* (https://github.com/B-UMMI/chewBBACA/tree/master/CHEWBBACA/prodigal_training_files). Additionally, one closed genome for serovars Brandenburg (GCA_004328765.1) and Saintpaul (GCA_001952995.1) were used as reference for the following phylogenetic analysis and downloaded from NCBI.

To generate a SNV phylogeny for both serovars, all alignments were done using snippy v4.6.0 (https://github.com/tseemann/snippy). Assembled genomes were aligned to the corresponding reference genome for the respective serovar then a core genome alignment was generated and Gubbins v3.3.5 (49) was used to remove recombination regions. The phylogenetic tree produced by Gubbins was visualized and annotated using iTOL v7 (50). Additional genomic characterizations included screening for the presence of plasmids or AMR genes using PlasmidFinder v2.1.6 (51) (minimum coverage: 75%, minimum identity: 95%) and AMRFinder v4 (39) (database 2025-07-16.1; minimum coverage: 50%, minimum identity: 90%), respectively. The virulence genotypes were provided by the NCBI Pathogen Detection platform. Subsequent phylogenies were generated using the above methods and a subset of the dataset for both serovars, namely all isolates within the original clade containing the outbreak strain.

Phylogenetic analysis of historical Salmonella *isolates with isolates collected by FSIS.*

All complete *Salmonella enterica* genomes available in NCBI as of September 10, 2025 were downloaded using the NCBI Datasets command line tool v 16.29.0 (52). PopPUNK v 2.7.2 (24) was used to generate a database of complete genomes, including genomes sequenced in this work; HDBSCAN was used for model fitting. Assembled genomes from isolates collected by FSIS were queried against this database to assign genomic clusters. For each cluster containing at least one newly sequenced historical genome and FSIS genomes, we performed phylogenetic analysis. Reads trimmed with fastp v 0.24.0 (36).

were mapped to a complete genome from each cluster using snippy v 4.6.0 (https://github.com/tseemann/snippy). A core alignment for each cluster was generated with snippy-core. Gubbins v 3.3.5 (49) was used to identify recombination events and estimate a recombination-corrected core genome phylogeny.

Acknowledgements

We wish to thank Joanna VanDenBoom for administrative support. The use of product and company names is necessary to accurately report the methods and results; however, the United States Department of Agriculture (USDA) neither guarantees nor warrants the standard of the products, and the use of names by the USDA implies no approval of the product to the exclusion of others that may also be suitable. The USDA is an equal opportunity provider and employer.

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by funds from the Meat Institute MEAT INSTITUTE GRANT NUMBERS ETC GO HERE and the U.S. Department of Agriculture, Agricultural Research Service CRIS project 3040-42000-020-00D.

Figures, Tables, and Captions

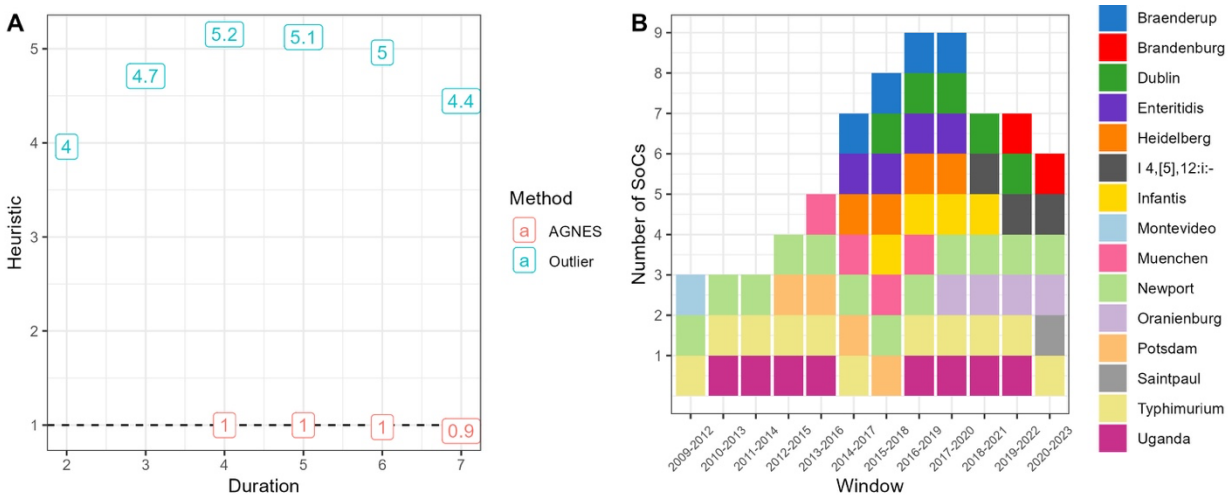


Figure 1.

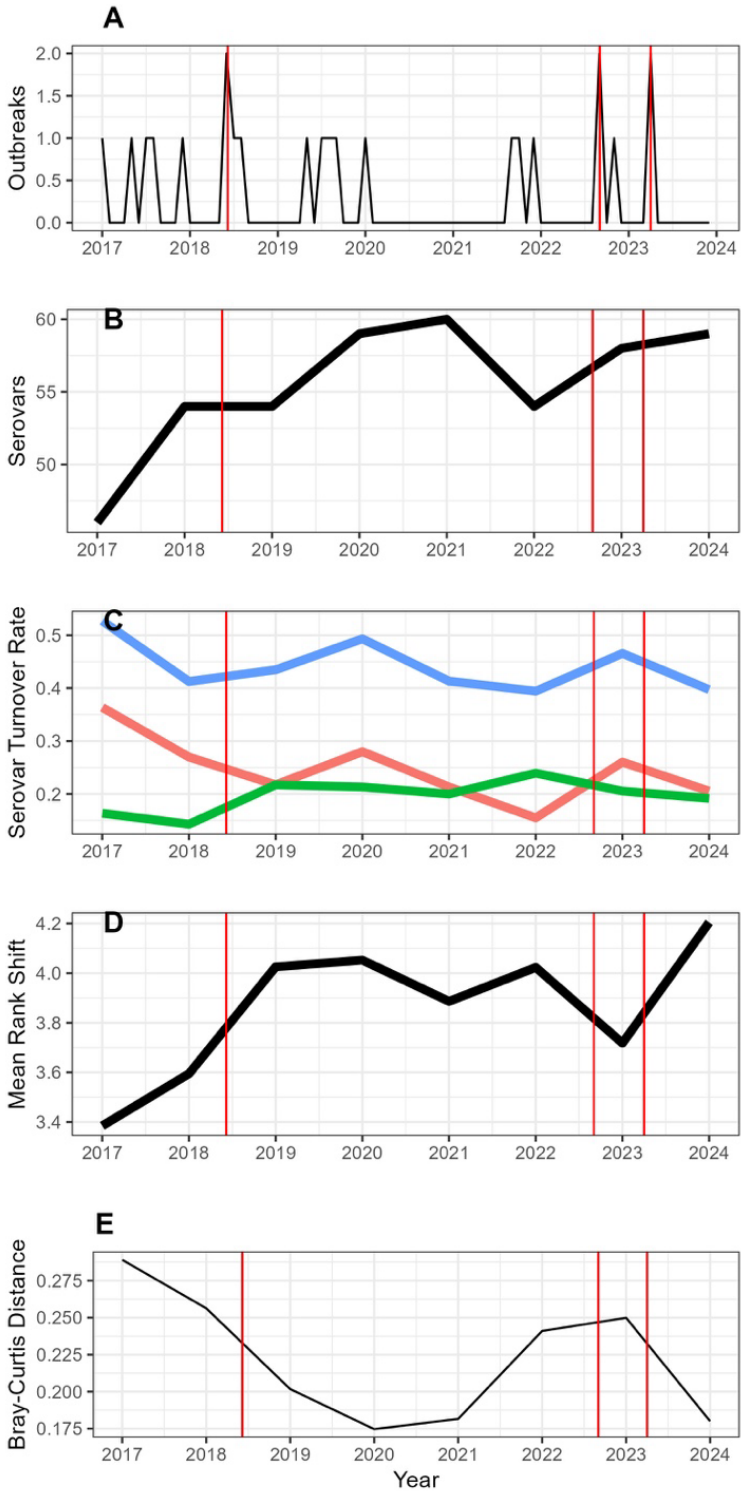

A
Outbreaks
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
2020
2021
2022
2023
2024
B
Serovars
60
55
50
C
Serovar Turnover Rate
0.5
0.4
0.3
0.2
D
Mean Rank Shift
4.2
4.0
3.8
3.6
3.4
E
Bray-Curtis Distance
0.275
0.250
0.225
0.200
0.175
Year

Figure 2.

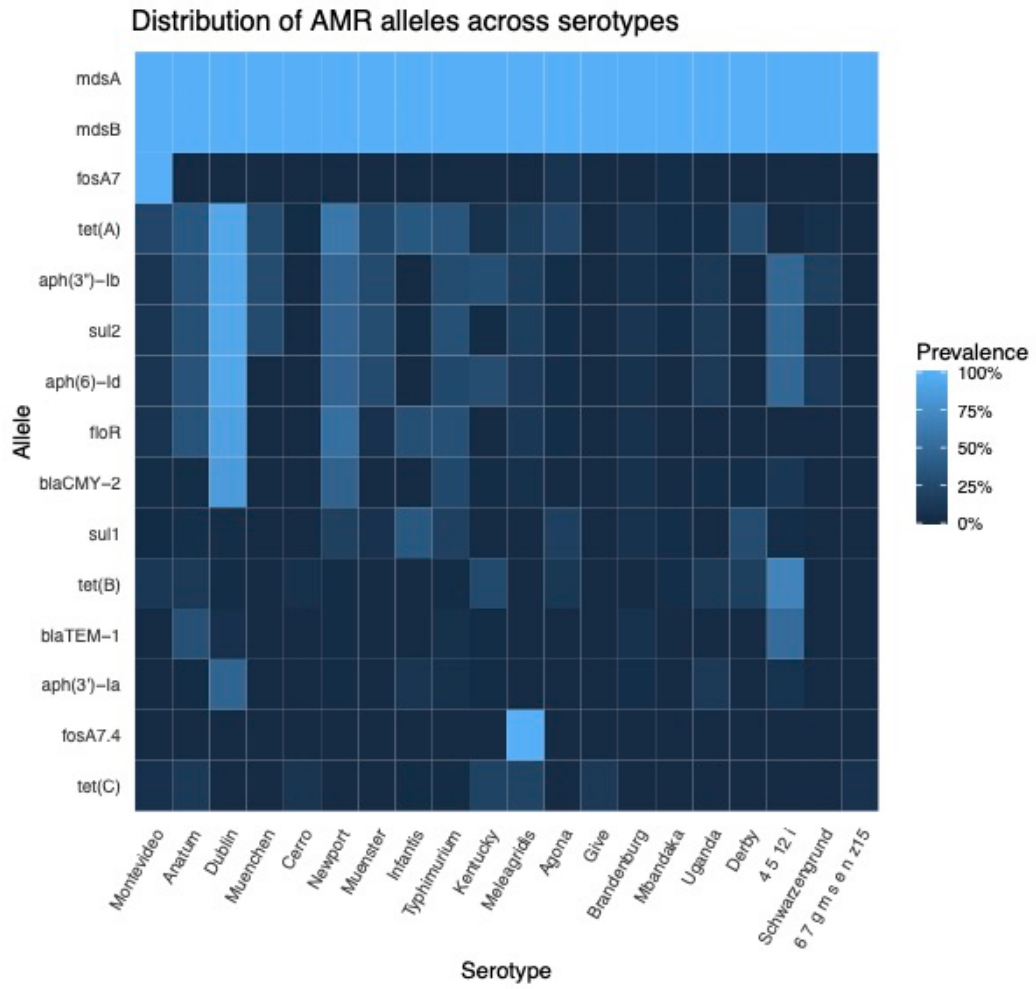

Distribution of AMR alleles across serotypes
mdsA
mdsB
fosA7
tet(A)
aph(3")−Ib
sul2
aph(6)−Id
floR
blaCMY−2
sul1
tet(B)
blaTEM−1
aph(3')−Ia
fosA7.4
tet(C)
Allele
Montevideo
Anatum
Dublin
Muenchen
Cerro
Newport
Muenster
Infantis
Typhimurium
Kentucky
Meleagridis
Agona
Give
Brandenburg
Mbandaka
Uganda
Derby
4 5 12 i
Schwarzengrund
6 7 g m s e n z15
Serotype
Prevalence
100%
75%
50%
25%
0%

Figure 3.

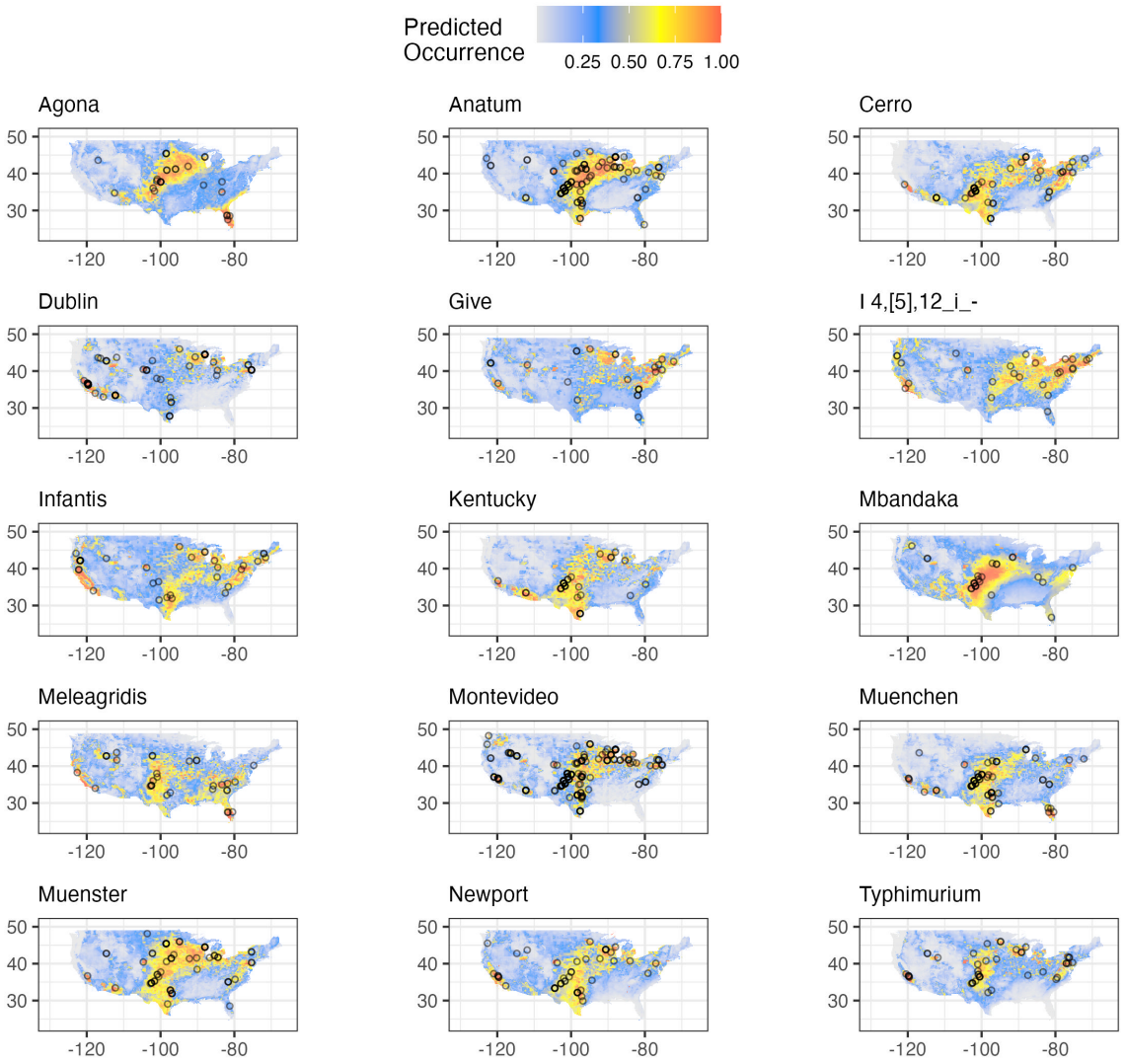

Predicted Occurrence
0.25 0.50 0.75 1.00
Agona
Anatum
Cerro
Dublin
Give
I 4,[5],12_i_-
Infantis
Kentucky
Mbandaka
Meleagridis
Montevideo
Muenchen
Muenster
Newport
Typhimurium
50
40
30
-120
-100
-80

Figure 4.

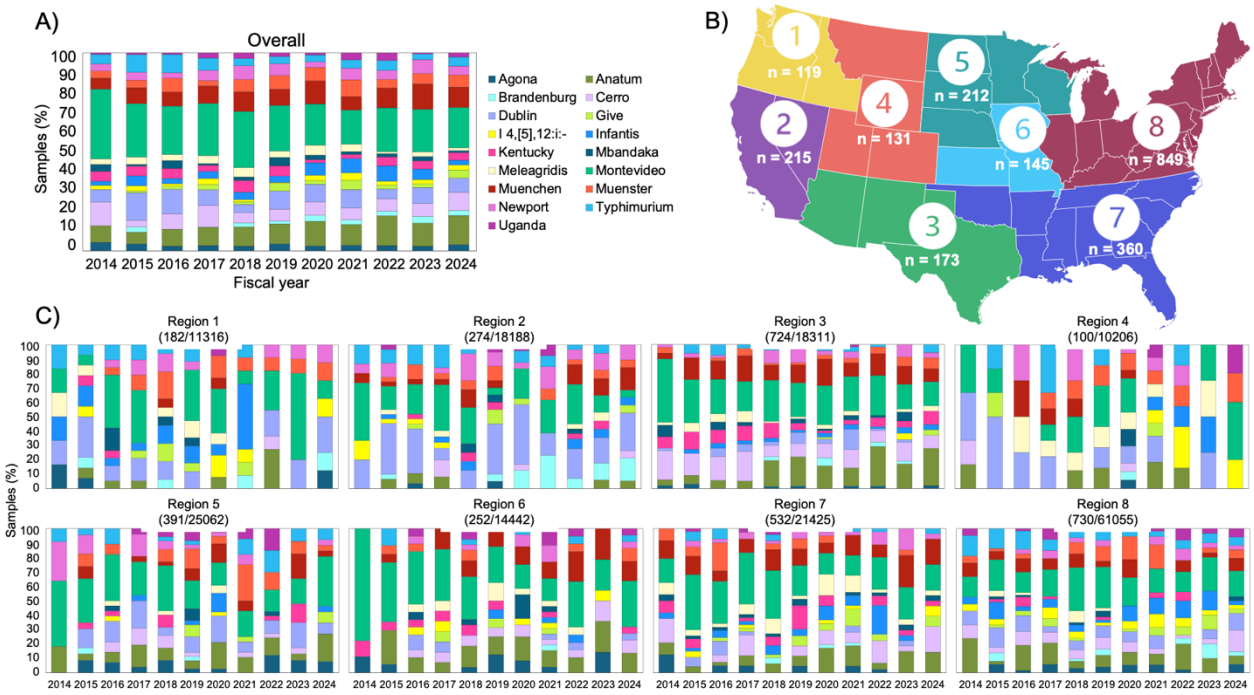


Figure 5.

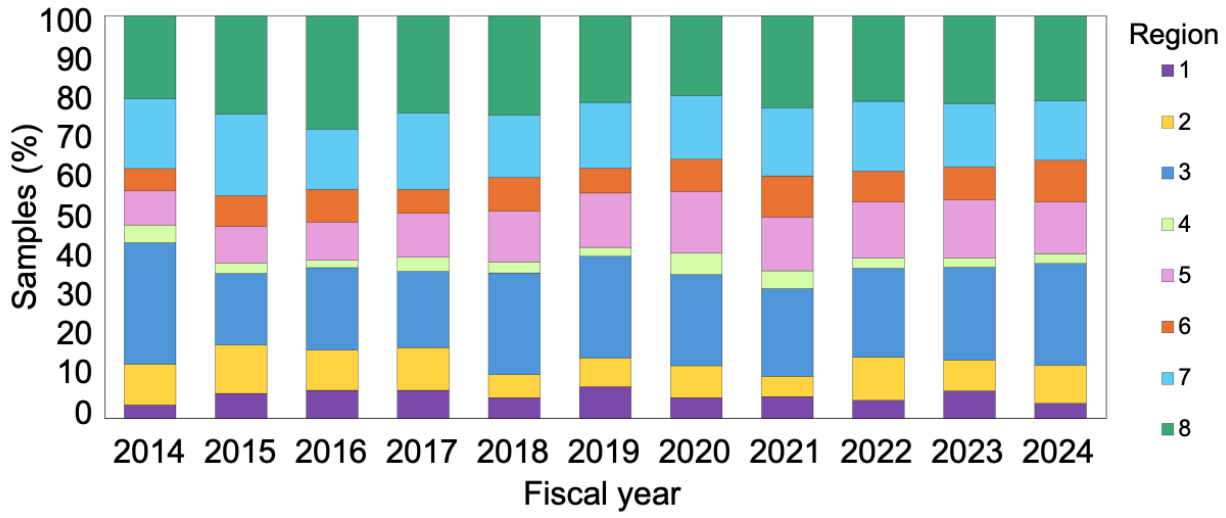


Figure 6.

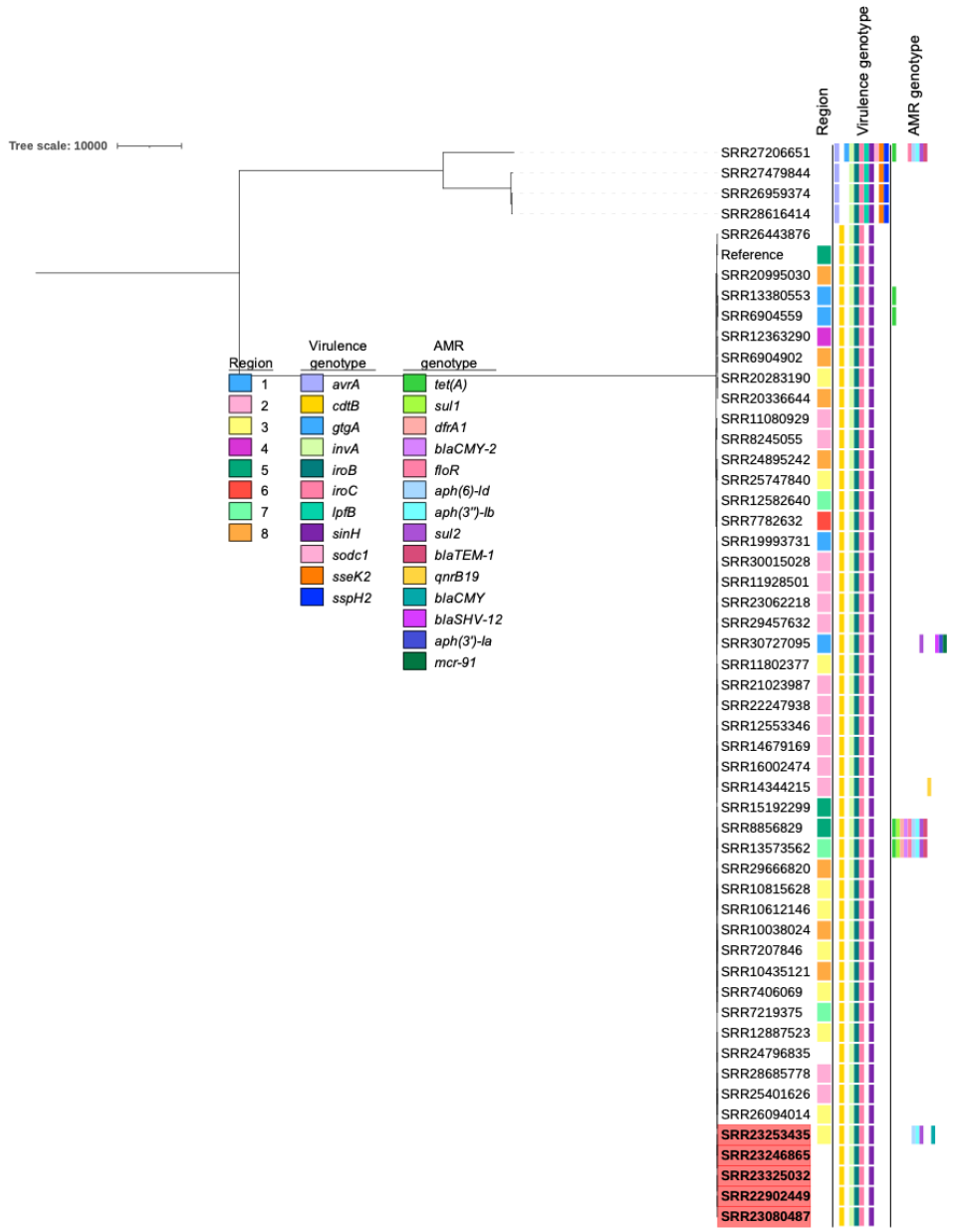

Tree scale: 10000
Region
Virulence genotype
AMR genotype
SRR27206651
SRR27479844
SRR26959374
SRR28616414
SRR26443876
Reference
SRR20995030
SRR13380553
SRR6904559
SRR12363290
SRR6904902
SRR20283190
SRR20336644
SRR11080929
SRR8245055
SRR24895242
SRR25747840
SRR12582640
SRR7782632
SRR19993731
SRR30015028
SRR11928501
SRR23062218
SRR29457632
SRR30727095
SRR11802377
SRR21023987
SRR22247938
SRR12553346
SRR14679169
SRR16002474
SRR14344215
SRR15192299
SRR8856829
SRR13573562
SRR29666820
SRR10815628
SRR10612146
SRR10038024
SRR7207846
SRR10435121
SRR7406069
SRR7219375
SRR12887523
SRR24796835
SRR28685778
SRR25401626
SRR26094014
SRR23253435
SRR23246865
SRR23325032
SRR22902449
SRR23080487
Region
1
2
3
4
5
6
7
8
Virulence genotype
avrA
cdtB
gtgA
invA
iroB
iroC
lpfB
sinH
sodc1
sseK2
sspH2
AMR genotype
tet(A)
sul1
dfrA1
blaCMY-2
floR
aph(6)-Id
aph(3")-Ib
sul2
blaTEM-1
qnrB19
blaCMY
blaSHV-12
aph(3')-Ia
mcr-91

Figure 7.

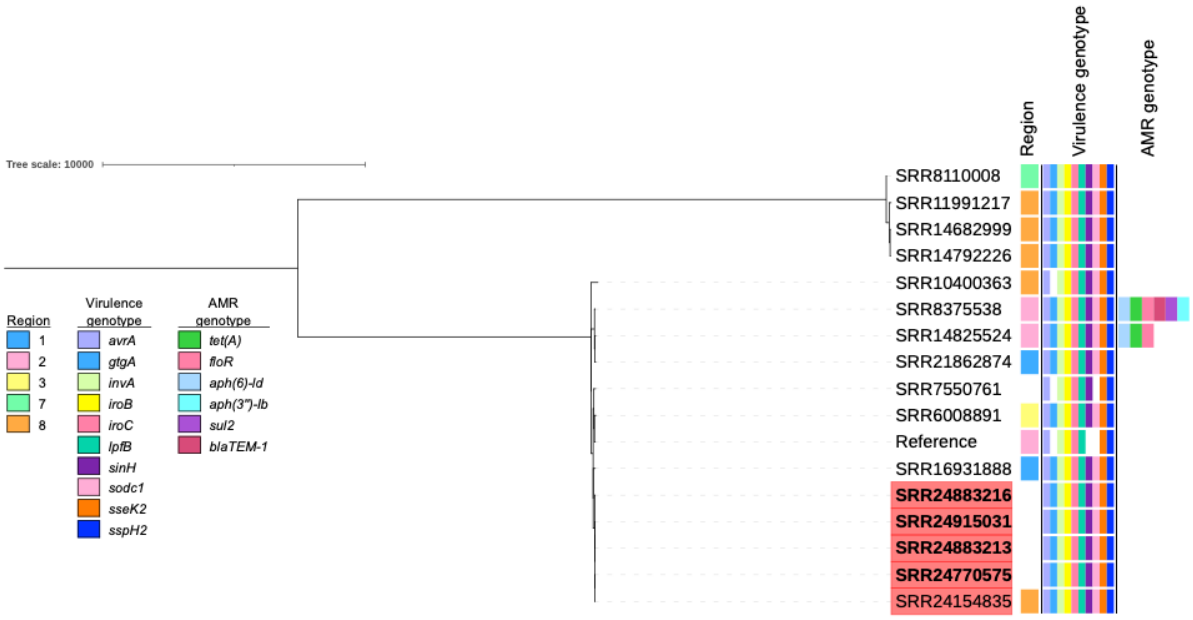

Tree scale: 10000
Region
Virulence genotype
AMR genotype
SRR8110008
SRR11991217
SRR14682999
SRR14792226
SRR10400363
SRR8375538
SRR14825524
SRR21862874
SRR7550761
SRR6008891
Reference
SRR16931888
SRR24883216
SRR24915031
SRR24883213
SRR24770575
SRR24154835
Region
1
2
3
7
8
Virulence genotype
avrA
gtgA
invA
iroB
iroC
lpfB
sinH
sodc1
sseK2
sspH2
AMR genotype
tet(A)
floR
aph(6)-Id
aph(3")-Ib
sul2
blaTEM-1

Figure 8.

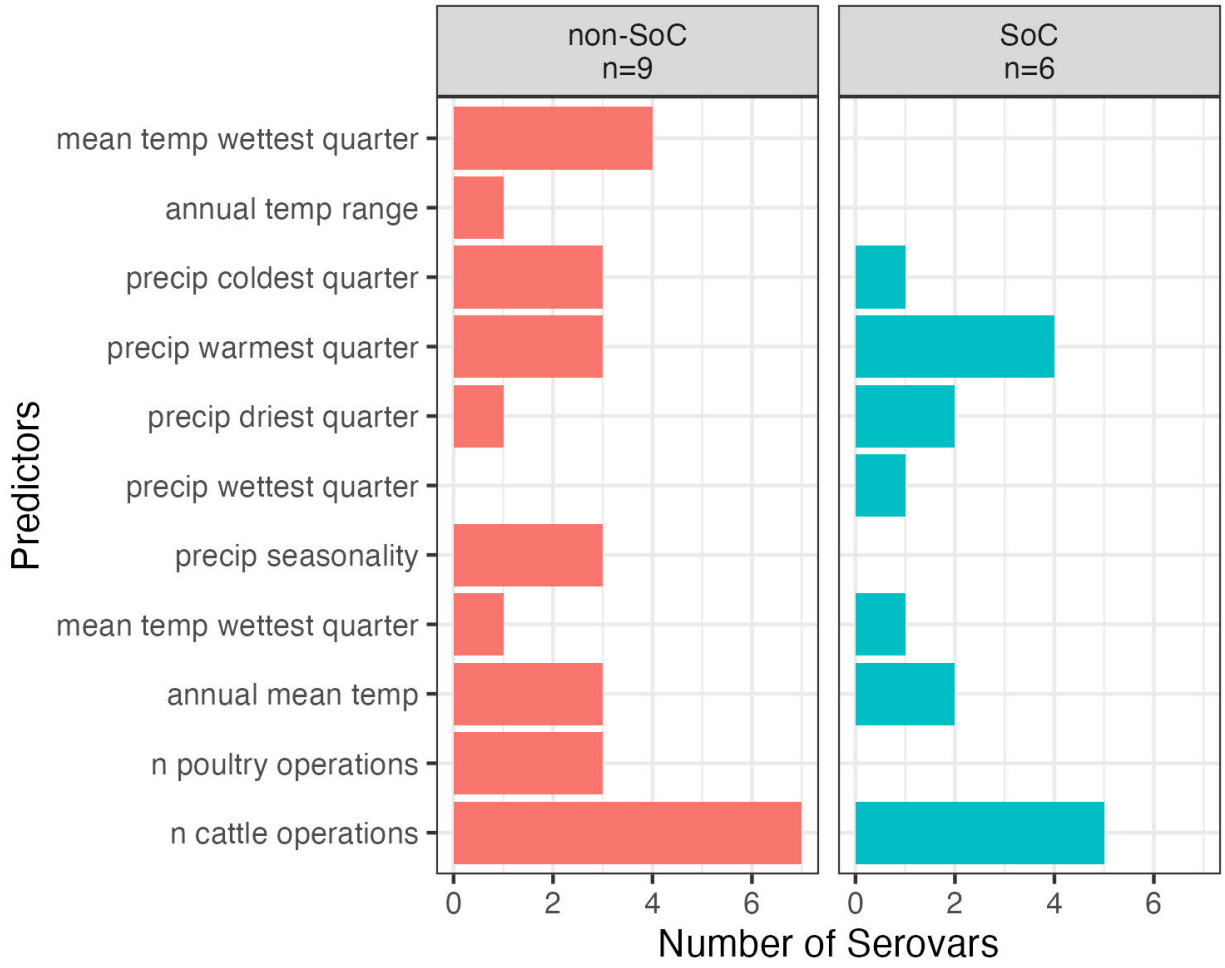
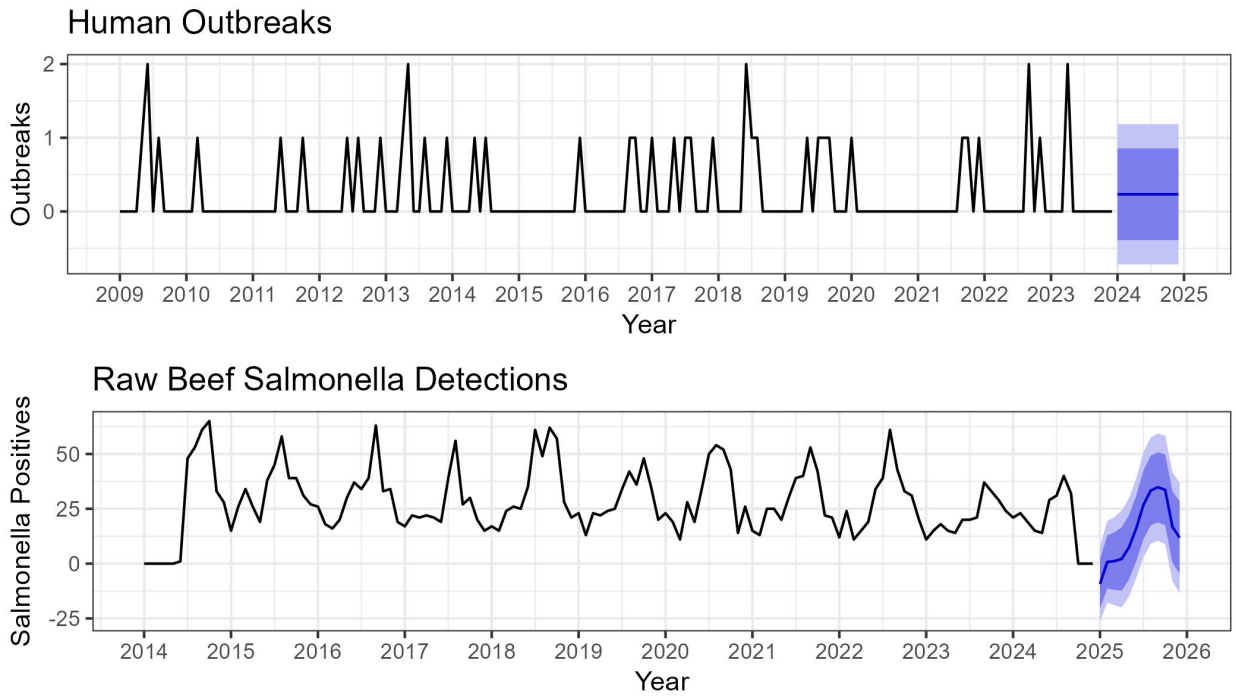
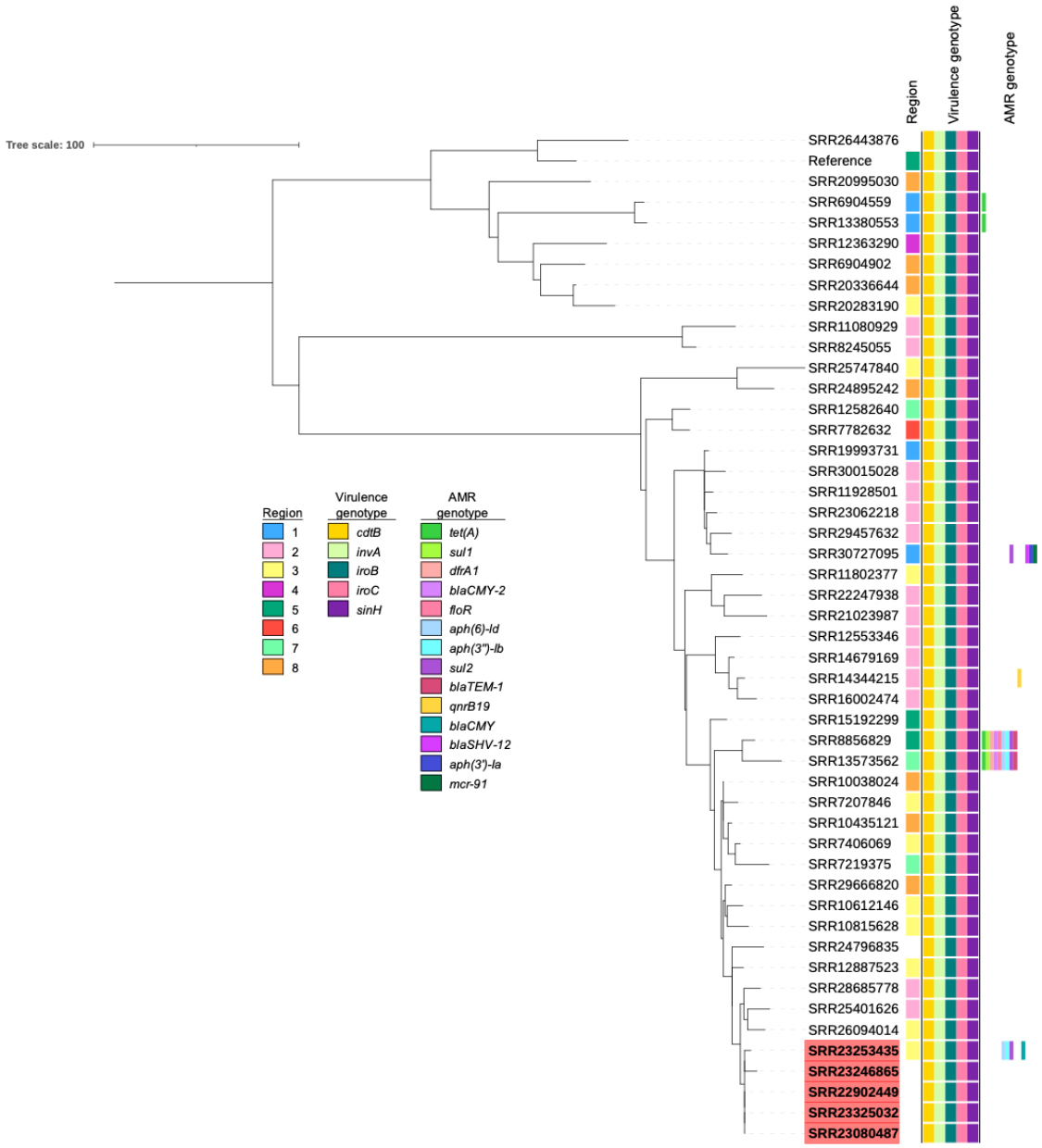

non-SoC
n=9
SoC
n=6
Predictors
mean temp wettest quarter
annual temp range
precip coldest quarter
precip warmest quarter
precip driest quarter
precip wettest quarter
precip seasonality
mean temp wettest quarter
annual mean temp
n poultry operations
n cattle operations
0
2
4
6
0
2
4
6
Number of Serovars

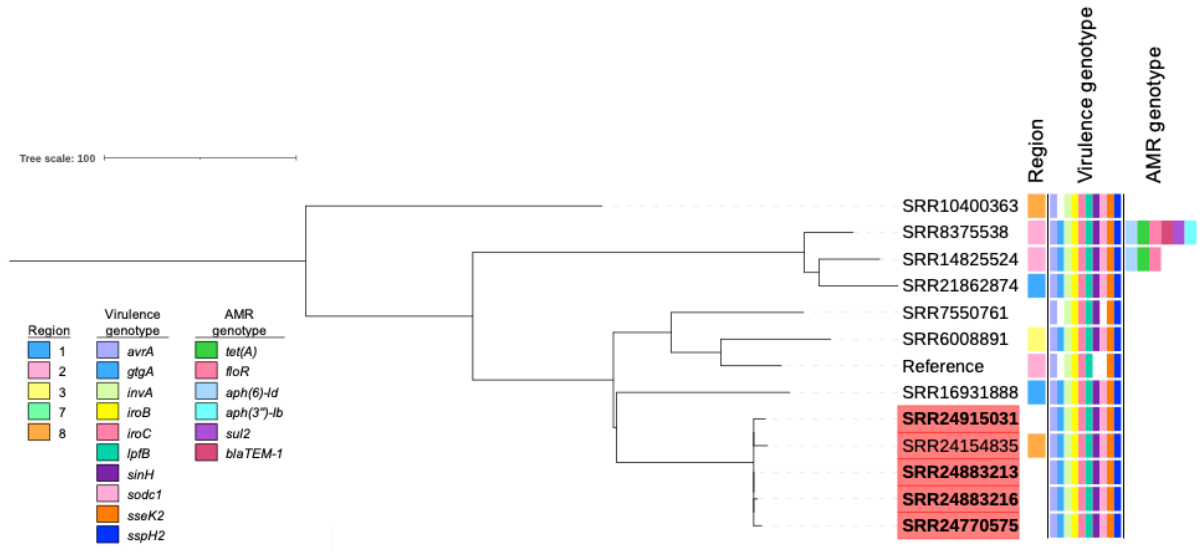
Figure 9.



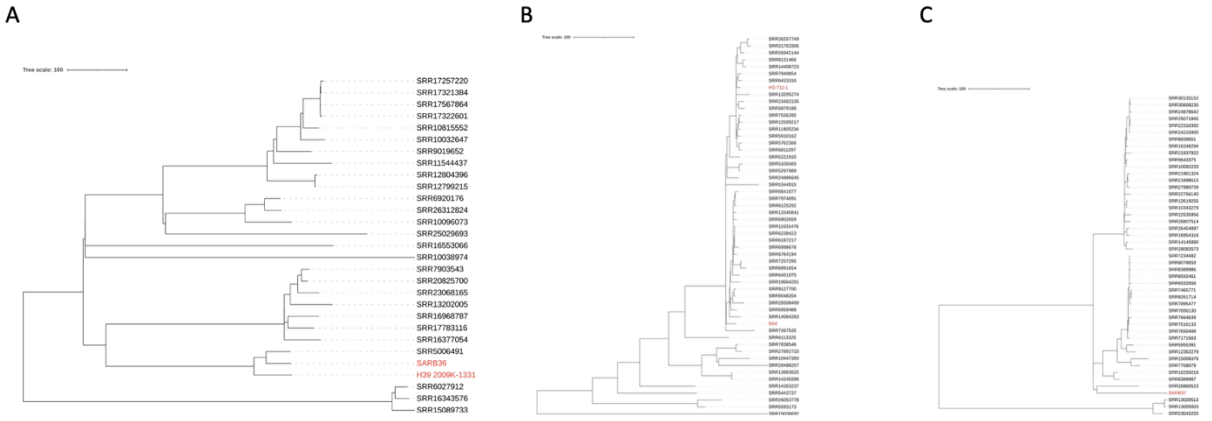
Supplementary Figure 1.



Supplementary Figure 2.

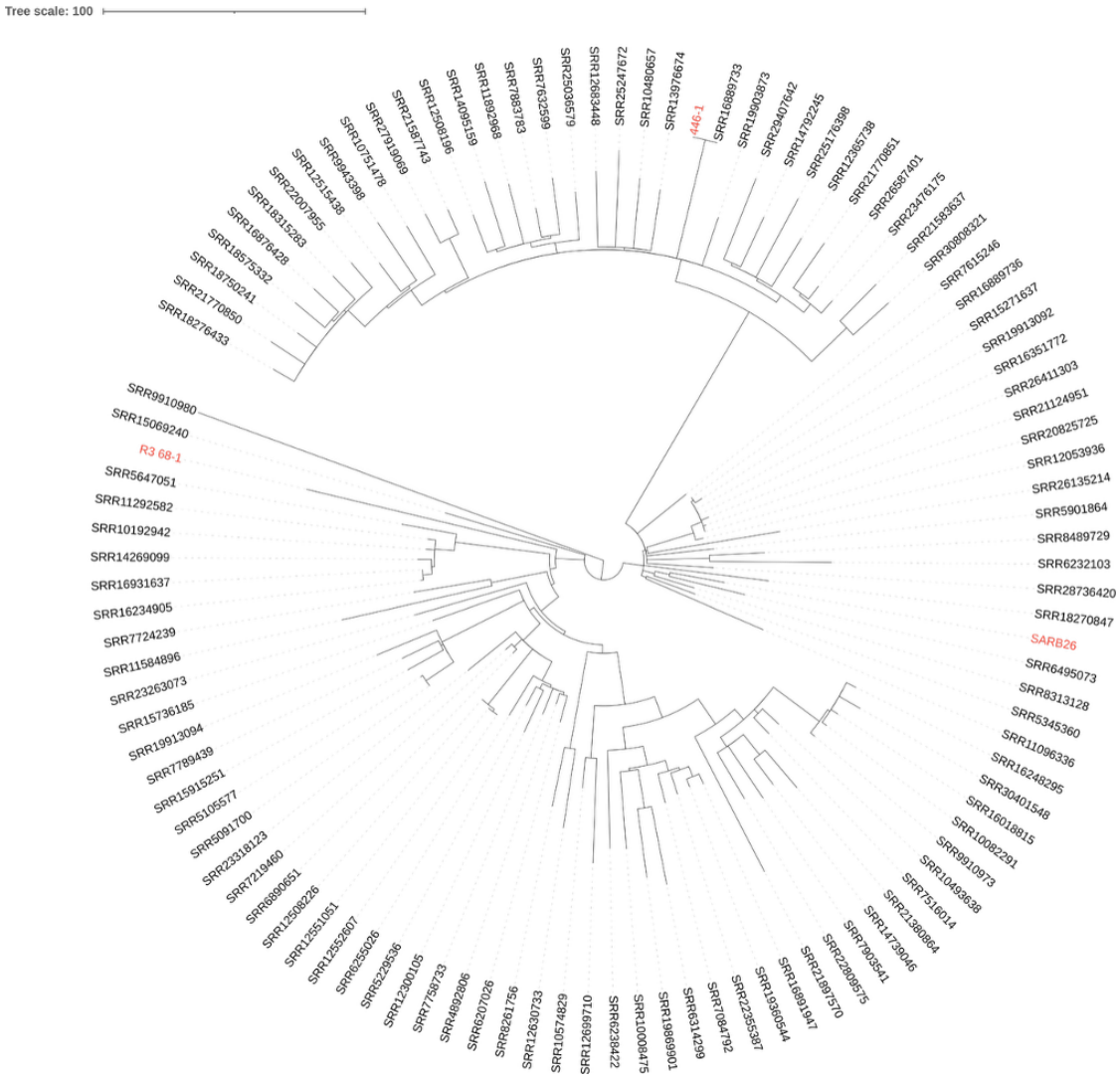


Supplementary Figure 3.

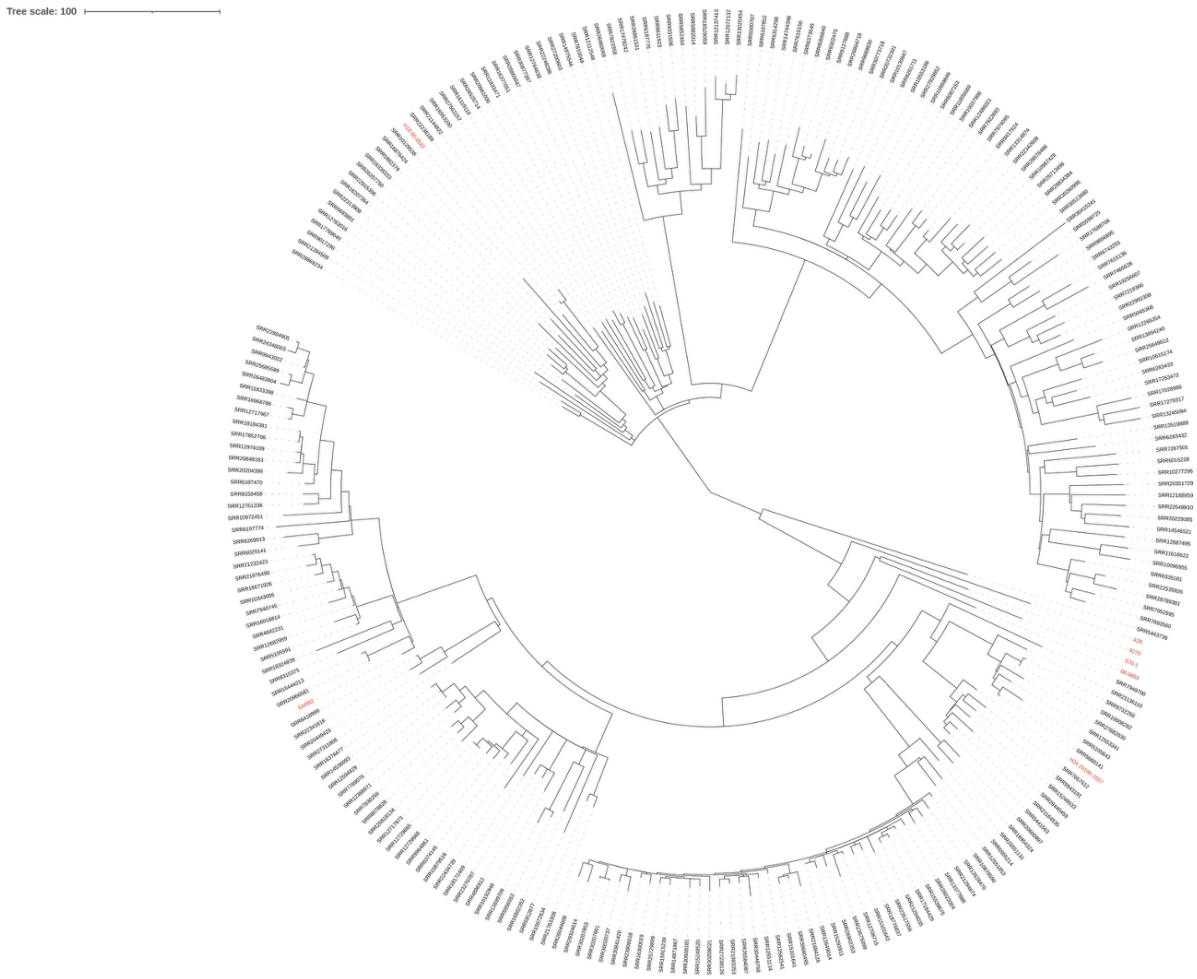


Supplementary Figure 4.

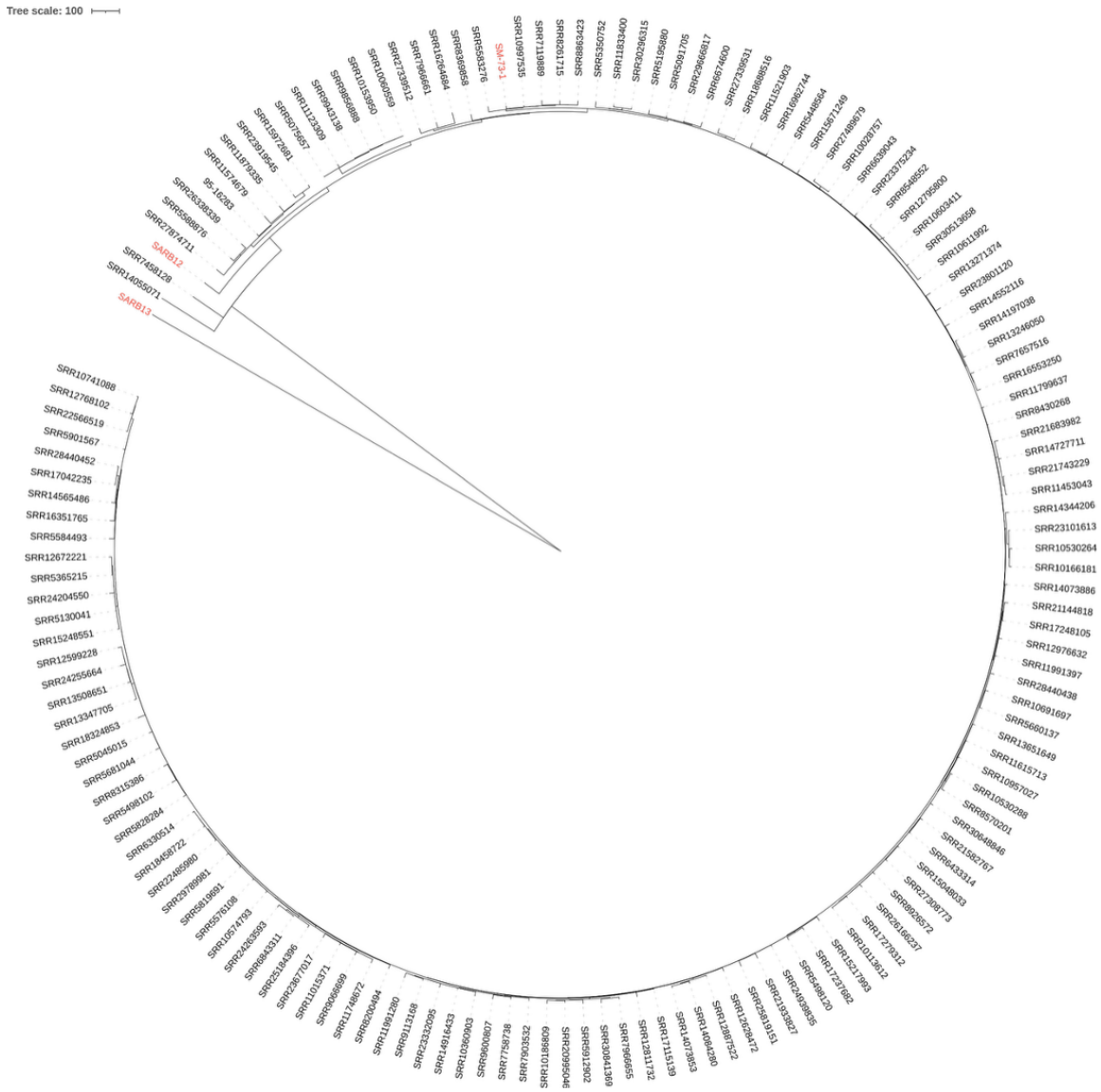
Supplementary Figure 5.



Supplementary Figure 6.



Supplementary Figure 7.



Supplementary Figure 8.

Figure 1 - (A) results of calculating the heuristic for the AGNES and Outlier methods over windows of durations 2 to 7 years. AGNES values could not be calculated for years 2 and 3 due to the SoC definition itself. (B) visual representation of SoCs identified over time on a 4-year rolling window using the Outlier method.

Figure 2 - Temporal diversity indices for early detection of outbreaks. Red lines indicate, in order, Infantis and Newport outbreaks in June 2018, Typhimurium and Newport outbreaks in September 2022, and Typhimurium and Saintpaul outbreaks in April 2023. (A) number of

beef-related *Salmonella* outbreaks from CDC NORS dataset. (B) number of serovars detected during raw beef sampling by FSIS. (C) Serovar turnover rates, where the green line is the disappearance rate, the red line is the appearance rate, and the blue line is the total turnover rate. (D) Mean rank shift of serovars, where higher numbers indicate greater shuffling of relative abundance ranks. (E) Bray-Curtis distance between each group of serovars compared to the previous month's.

Figure 3 – Distribution of antimicrobial resistance associated alleles across *Salmonella* serovars. Prevalence of genes detected by AMRFinderPlus within each serovar is indicated on a scale from 0% (dark blue) to 100% (light blue).

Figure 4 - Predictions of serovar distributions based on geospatial predictors. Individual circles represent a single sample of the given serovar.

Figure 5 – Top serovars reported in beef products sampled by FSIS from 2014 to 2024 and regional differences. A) Overall results for the top 17 serovars found in beef products in the continental United States. B) Map of production regions with total number of establishments sampled shown. C) Regional serotyping results across the eight production regions. Number of *Salmonella*-positive and total samples collected listed in parenthesis.

Figure 6 – Number of *Salmonella*-positive regulatory beef samples collected in each production region from 2014 to 2024. Results are shown as percentages of the total number of positive samples per year.

Figure 7 – High genetic relatedness of human clinical isolates of serovar Brandenburg and beef-related isolates. Isolates identified as the outbreak strain are highlighted in red, with human clinical isolates collected by the CDC in bold. The remaining isolates were collected by FSIS as part of the beef regulatory surveillance program.

Figure 8 – High genetic relatedness of human clinical isolates of serovar Saintpaul and beef-related isolates. Isolates identified as the outbreak strain are highlighted in red, with human clinical isolates collected by the CDC in bold. The remaining isolates were collected by FSIS as part of the beef regulatory surveillance program

Figure 9 - Most important predictors for the distributions of SoC and non-SoC serovars. Only predictors which contributed 10% or more to the distribution predictions were included.

Supplementary Figure 1 – Time series and forecasting of beef-attributed outbreaks in humans and *Salmonella* detections on raw beef by FSIS.

Supplementary Figure 2 – Subsampled phylogenetic tree of clade containing serovar Brandenburg outbreak isolates. Isolates identified as the outbreak strain are highlighted in red, with human clinical isolates collected by the CDC in bold. The remaining isolates were collected by FSIS as part of the beef regulatory surveillance program.

Supplementary Figure 3 – Subsampled phylogenetic tree of clade containing serovar Saintpaul outbreak isolates. Isolates identified as the outbreak strain are highlighted in red, with human clinical isolates collected by the CDC in bold. The remaining isolates were collected by FSIS as part of the beef regulatory surveillance program.

Supplementary Figure 4 – Phylogenies of FSIS and historical isolates from serotype Newport clusters. Tips associated with historical isolates are indicated in red.

Supplementary Figure 5 – Phylogenies of FSIS and historical isolates from serotype Saintpaul clusters. Tips associated with historical isolates are indicated in red.

Supplementary Figure 6 – Phylogeny of FSIS and historical isolates from serotype Infantis. Tips associated with historical isolates are indicated in red.

Supplementary Figure 7 – Phylogeny of FSIS and historical isolates from serotype Anatum. Tips associated with historical isolates are indicated in red.

Supplementary Figure 8 – Phylogeny of FSIS and historical isolates from serotype Dublin. Tips associated with historical isolates are indicated in red.

Supplementary Table 1. SoC lists for rolling windows of durations 2 to 7 years for the years 2009 through 2023.

Supplementary Table 2. Krippendorf's Alpha values for rolling windows of durations 2 to 7 years.

Supplementary Table 3. Metadata and genome assembly metrics for sequenced historical *Salmonella* strains.

References

1. Scallan Walter EJ, Cui Z, Tierney R, Griffin PM, Hoekstra RM, Payne DC, Rose EB, Devine C, Namwase AS, Mirza SA, Kambhampati AK, Straily A, Bruce BB. 2025. Foodborne Illness Acquired in the United States—Major Pathogens, 2019. Emerg Infect Dis 31:669–677.
2. Interagency Food Safety Analytics Collaboration. 2024. Foodborne Illness Source Attribution Estimates for Salmonella, Escherichia Coli O157, and Listeria Monocytogenes – United States, 2022.
3. Cheng RA, Eade CR, Wiedmann M. 2019. Embracing diversity: Differences in virulence mechanisms, disease severity, and host adaptations contribute to the success of nontyphoidal salmonella as a foodborne pathogen. Front Microbiol 10.
4. Uzzau S, Brown DJ, Wallis T, Rubino S, Leori G, Bernard S, Casadesús J, Platt DJ, Olsen JE. 2000. Host adapted serotypes of Salmonella enterica. Epidemiol Infect https://doi.org/10.1017/S0950268899004379.
5. Katz TS, Harhay DM, Schmidt JW, Wheeler TL. 2024. Identifying a list of Salmonella serotypes of concern to target for reducing risk of salmonellosis. Front Microbiol 15.
6. Food Safety and Inspection Service. 2025. Raw Beef Sampling. https://www.fsis.usda.gov/news-events/publications/raw-beef-sampling.
7. Centers for Disease Control and Prevention. 2022. BEAM Dashboard. https://www.cdc.gov/ncezid/dfwed/BEAM-dashboard.html.
8. Harhay DM, Brader KD, Katz TS, Harhay GP, Bono JL, Bosilevac JM, Wheeler TL. 2024. A novel approach for detecting Salmonella enterica strains frequently attributed to human illness—development and validation of the highly pathogenic Salmonella (HPS) multiplex PCR assay. Front Microbiol 15.
9. Marshall KE, Cui Z, Gleason BL, Hartley C, Wise ME, Bruce BB, Griffin PM. 2024. An Approach to Describe Salmonella Serotypes of Concern for Outbreaks: Using Burden and Trajectory of Outbreak-related Illnesses Associated with Meat and Poultry. J Food Prot 87.
10. Fenske GJ, Pouzou JG, Pouillot R, Taylor DD, Costard S, Zagmutt FJ. 2023. The genomic and epidemiological virulence patterns of Salmonella enterica serovars in the United States. PLoS One 18:1–21.
11. Centers for Disease Control and Prevention. 2023. National Outbreak Reporting System. U.S. Department of Health and Human Services, CDC, Atlanta, Georgia.
12. Oeschger TM, McCloskey DS, Buchmann RM, Choubal AM, Boza JM, Mehta S, Erickson D. 2021. Early Warning Diagnostics for Emerging Infectious Diseases in Developing into Late-Stage Pandemics. Acc Chem Res 54:3656–3666.
13. Morin CW, Semenza JC, Trtanj JM, Glass GE, Boyer C, Ebi KL. 2018. Unexplored Opportunities: Use of Climate- and Weather-Driven Early Warning Systems to

Reduce the Burden of Infectious Diseases. Curr Environ Health Rep. Springer https://doi.org/10.1007/s40572-018-0221-0.
14. Hwang D, Rothrock MJ, Pang H, Guo M, Mishra A. 2020. Predicting Salmonella prevalence associated with meteorological factors in pastured poultry farms in southeastern United States. Science of the Total Environment 713.
15. Karanth S, Patel J, Shirmohammadi A, Pradhan AK. 2023. Machine learning to predict foodborne salmonellosis outbreaks based on genome characteristics and meteorological trends. Curr Res Food Sci 6.
16. White AE, Jackson C, Kisselburgh H, Ledbetter C, Walter ES. 2022. Using Outbreak Data for Hypothesis Generation: A Vehicle Prediction Tool for Disease Outbreaks Caused by Salmonella and Shiga Toxin-Producing Escherichia coli. Foodborne Pathog Dis 19:281–289.
17. Akil L, Ahmad A. 2015. Salmonella infections modelling in Mississippi using neural network and geographical information system (GIS). BMJ Open e009255.
18. Zacher B, Czogiel I. 2022. Supervised learning using routine surveillance data improves outbreak detection of Salmonella and Campylobacter infections in Germany. PLoS One 17.
19. Rojas F, Ibacache-Quiroga C. 2020. A forecast model for prevention of foodborne outbreaks of non-typhoidal salmonellosis. PeerJ 8.
20. Marzi G, Balzano M, Marchiori D. 2024. K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. MethodsX 12.
21. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. 2015. Time Series Analysis: Forecasting and Controlhttps://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021, 5th ed. Wiley Series in Probability and Statistics.
22. Phillips SJ, Anderson RP, Schapire RE. 2006. Maximum entropy modeling of species geographic distributions. Ecol Modell 190:231–259.
23. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.
24. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316.
25. Castano-Duque L, Avila A, Mack BM, Winzeler HE, Blackstock JM, Lebar MD, Moore GG, Owens PR, Mehl HL, Su J, Lindsay J, Rajasekaran K. 2025. Prediction of aflatoxin contamination outbreaks in Texas corn using mechanistic and machine learning models. Front Microbiol 16.

26. Garcia-Vozmediano A, Maurella C, Ceballos LA, Crescio E, Meo R, Martelli W, Pitti M, Lombardi D, Meloni D, Pasqualini C, Ru G. 2024. Machine learning approach as an early warning system to prevent foodborne Salmonella outbreaks in northwestern Italy. Vet Res 55:72.
27. Bisola Oluwafadekemi Adegoke, Tolulope Odugbose, Christiana Adeyemi. 2024. Data analytics for predicting disease outbreaks: A review of models and tools. International Journal of Life Science Research Updates 2:001–009.
28. Ziaur Rahman S, Senthil R, Ramalingam V, Gopal R. 2023. Predicting Infectious Disease Outbreaks with Machine Learning and Epidemiological Data. Journal of Advanced Zoology 44:110–121.
29. Scarpino S V., Petri G. 2019. On the predictability of infectious disease outbreaks. Nat Commun 10.
30. Foluke Ekundayo. 2024. Using machine learning to predict disease outbreaks and enhance public health surveillance. World Journal of Advanced Research and Reviews 24:794–811.
31. Cohen J. 1960. A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES. Educ Psychol Meas.
32. 2024. FSIS Directive 10010.1: Sampling verification activities for Shiga Toxin-producing Escherichia coli in raw beef products. United States Department of Agriculture Food Safety and Inspection Service, Washington, D.C.
33. Fick SE, Hijmans RJ. 2017. WorldClim Version2. International Journal of Climatology. http://worldclim.org/version2. Retrieved 17 October 2017.
34. Phillips SJ, Dudik M, Schapire RE. 2025. Maxent software for modeling species niches and distributions. 3.4.1. https://biodiversityinformatics.amnh.org/open_source/maxent/.
35. Wick RR, Howden BP, Stinear TP. 2025. Autocycler: long-read consensus assembly for bacterial genomes. Bioinformatics 41.
36. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res 27:722–736.
37. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37:540–546.
38. Li H. 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32:2103–2110.
39. Hu J, Wang Z, Sun Z, Hu B, Ayoola AO, Liang F, Li J, Sandoval JR, Cooper DN, Ye K, Ruan J, Xiao C-L, Wang D, Wu D-D, Wang S. 2024. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. Genome Biol 25:107.

40. Vaser R, Šikić M. 2021. Time- and memory-efficient genome assembly with Raven. Nat Comput Sci 1:332–336.
41. Bouras G, Judd LM, Edwards RA, Vreugde S, Stinear TP, Wick RR. 2024. How low can you go? Short-read polishing of Oxford Nanopore bacterial genome assemblies. Microb Genom 10.
42. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. 2021. Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb Genom 7.
43. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 11:12728.
44. Robertson J, Nash JHE. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. Microb Genom 4.
45. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, Lauer AC, Fields PI, Deng X. 2019. SeqSero2: Rapid and improved salmonella serotype determination using whole-genome sequencing data. Appl Environ Microbiol 85.
46. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. Curr Protoc Bioinformatics 70.
47. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15–e15.
48. Letunic I, Bork P. 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res 52:W78–W82.
49. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. *In Silico* Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrob Agents Chemother 58:3895–3903.
50. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, Tsuchiya MTN, Schuler GD, Zhang X, Torcivia J, Ketter A, Breen L, Cothran J, Bajwa H, Tinne J, Meric PA, Hlavina W, Schneider VA. 2024. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data 11.
51. Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: An ultra-fast all-in-one FASTQ preprocessor, p. i884–i890. *In* Bioinformatics. Oxford University Press.
52. Köster J, Rahmann S. 2012. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 28:2520–2522.