The genomic and epidemiological virulence patterns of *Salmonella enterica* serovars

Gavin J. Fenske[1], Jane G. Pouzou[1], Régis Pouillot[1], Daniel D. Taylor[1], Solenne Costard[1], and Francisco J. Zagmutt[1]*

[1]EpiX Analytics, Fort Collins, Colorado, United States of America

*Corresponding Author

Email: fzagmutt@epixanalytics.com

Abstract

The serovars of *Salmonella enterica* display dramatic differences in pathogenesis and host preferences. Grouping *Salmonella* isolates and serovars by their public health risk can provide better S*almonella* control targets along the food chain. We collated a curated set of 12,337 *S. enterica* isolate genomes from human, beef, and bovine sources in the US. After annotating a virulence gene catalog for each isolate, we used unsupervised random forest methods to estimate the proximity (similarity) between isolates based upon the genomic presentation of putative virulence traits We then grouped isolates (virulence clusters) using hierarchical clustering (Ward's method), used non-parametric bootstrapping to assess cluster stability, and externally validated the virulence clusters against epidemiological virulence measures from FoodNet, the National Outbreak Reporting System (NORS), and US federal sampling of beef products. We identified five stable virulence clusters of *S. enterica* serovars. Cluster 1 serovars yielded an annual incidence rate of domestically acquired sporadic cases roughly one and a half times higher than the other four clusters combined. Compared to other clusters, cluster 1 also had a higher proportion of infections leading to hospitalization and was implicated in more foodborne and beef-associated outbreaks, despite being isolated at a similar frequency from beef products as other clusters. We also identified subpopulations within 11 serovars. Remarkably, we found *S.* Infantis and *S.* Typhimurium subpopulations that significantly differed in genome length and clinical case presentation. Further, we found that the presence of the pESI plasmid accounted for the genome length differences between the *S.* Infantis subpopulations. Our results demonstrate that *S. enterica* strains with the highest incidence of human infections share a common virulence repertoire. This work could be used in

combination with foodborne surveillance information to best target serovars of public health concern.

Introduction

Members of *Salmonella enterica* subspecies *enterica* are some of the most ubiquitous agents implicated in foodborne human illnesses. Despite being constituents of the same subspecies, members of *S. enterica* are not only commonly isolated from livestock but also amphibians [1] and wild birds [2]. The wide host range for *S. enterica* makes control of the pathogen exceedingly difficult due to the large number of potential reservoirs. Historically, strains of *S. enterica* have been grouped into units termed serovars based upon serological antigen presentation. While an initial list presented 44 *S. enterica* serovars in 1934 [3], today's descriptions include over 2,500 serovars of *S. enterica* [4]. Nonetheless, in the US only 20 serovars accounted for 69.2% of human *S. enterica* isolates collected in 2016 by the US Centers for Disease Control and Prevention's (CDC) Laboratory-based Enteric Disease Surveillance (LEDS) program [5]. Furthermore, nearly 10% of *S. enterica*'s serovars may be polyphyletic or paraphyletic [6].

To establish infections in disparate hosts, *S. enterica* manipulates common immune functions of higher vertebrates. Indeed, the classic gastroenteritis associated with *S. enterica* infections is the result of the pathogen affecting the host's innate immune system to generate inflammation, subsequently producing unique metabolic niches for *S. enterica* while killing its competitors for reduced substrates in the hindgut [7-9]. Such remarkable expropriation of the hosts immune functions is achieved by virulence genes, many of which are contained within chromosomal elements termed Salmonella Pathogenicity Islands (SPI) [10]. Genes contained within SPI aid in host cell invasion, and subsequent survival and dissemination within and

between Eukaryotic host cells [11,12]. However, serovars display differences in pathogenesis and host-preferences. For example, the human-restricted serovar *S. enterica* ser. Typhi (*S.* Typhi), the etiological agent of typhoid fever, does not typically cause submucosal inflammation and resultant diarrhea in infected patients as with classical salmonellosis, but instead elicits a systemic enteric fever characterized by initial immune evasion [13,14]. *S.* Dublin, a bovine adapted serovar, commonly generates systemic infections in humans and is isolated from blood samples in 61% of human clinical infections as compared to an average of 5% for other *S. enterica* serovars in the US [15]. The general pathogenesis of *S. enterica* is not fully elucidated, and the virulence potential for individual serovars is poorly understood. Furthermore, most studies have focused upon *S.* Typhimurium as a model organism for all *S. enterica* virulence, [8,16-21] which could obfuscate differences between serovars.

Despite the tremendous virulence diversity within *S. enterica*, microbial criteria from the US Food Safety and Inspection Services (FSIS) on important sources of *S. enterica* such as beef and poultry meats target all serovars equally, based on prevalence. Further, traditional surveillance methods can take considerable time to identify emerging serovars of public health concern, thereby delaying food safety intervention implementation [22]. Understanding virulence differences between serovars and identifying emerging virulent serovars in a timely manner can be important for more focused risk management strategies targeting serovars with an inordinate impact on public health while reducing food waste due to recalls.

Previous studies have used genomics to identify serovar groups of public health concern. Karanth et al. analyzed a limited number of genomes and serovars originating from humans, poultry, and swine to characterize virulent serovars [23]. This analysis had the benefit of using the entire genome of Salmonella to group isolates by disease presentation; however, the

computational resources required prevent its application to a large number of isolates. In another study, researchers used single nucleotide polymorphism (SNP) clusters and *S.* Saintpaul as a model to identify virulent isolates [24]. Although using high-resolution genomic methods identified SNP clusters associated with a high proportion of human clinical isolates, *S.* Saintpaul may not be the best serovar model due to its polyphyletic nature [25]. The objective of this study was to develop a computationally efficient genomic approach to group *Salmonella* serovars by their risk to human health, using virulence biomarkers in isolates from humans, beef, and bovine animals. We chose beef as a model foodstuff since US federal monitoring of *Salmonella* in beef is well-established, nationally representative, and beef remains an important vehicle for *S. enterica*. Beef production in the US is more decentralized than poultry and pork production [26] and we e hypothesize that this decentralization may present unique genomic populations arising from geographic separation as previously observed in *S. enterica* serovars [27,28]. Furthermore, *S. enterica* in beef products is understudied compared to other vehicles such as eggs, poultry, and pork meat.

Materials and methods

We used genetic virulence factors as markers to group serovars by their risk to human health. After compiling a curated set of *S. enterica* genomes (n=12,337) from human, bovine, and beef sources, we applied an unsupervised random forest and hierarchical clustering approach to group isolates based upon genomic virulence trait presentation and validated the groups against epidemiological measures including clinical presentation from sequenced isolates collected by the FoodNet active surveillance network (29).

Contig assembly selection and quality criteria

We compiled *S. enterica* assemblies from bovine-associated isolates from three primary sources: 1. BioProject PRJNA242847 (FSIS HACCP samples, accessed 7/13/2021), 2. BioProject PRJNA292666 (FSIS NARMS isolates, accessed 7/13/2021), and 3. BioProject PRJNA292661 (FDA NARMS isolates, accessed 8/25/2021). We collected isolates from sources specified as bovine-associated or beef origin from the metadata for the above BioProjects.

We retrieved *S. enterica* isolates from human clinical cases from BioProject PRJNA230403 (CDC PulseNet, accessed 9/13/2021) and identified sporadic, domestically acquired *S.* enterica isolates from the FoodNet active surveillance network [29]. We did not include outbreak cases from FoodNet since they are not attributed to a particular source in that dataset. Instead, we used the National Outbreak Reporting System (NORS) dataset, a passive system for reporting enteric disease outbreaks in the US, to identify beef-attributed outbreak isolates. We initially defined beef attribution based on the Interagency Food Safety Analytics Collaboration (IFSAC) classification. As beef-associated salmonellosis outbreaks for which clinical isolates were sequenced are limited, we widened the definition of potentially beef-associated illness to include outbreaks which listed beef as an identified contaminated ingredient. We based this inclusion on whether the list of commodities and ingredients per outbreak included beef dishes, even if other possible ingredients could not be ruled out to definitively assign an IFSAC classification. The following text strings were used to identify beef-associated outbreaks: "beef", "burger", "steak", "carne", "kitfo", "ox tongue", "short-rib", "prime rib", "barbacoa". If the IFSAC classification attributed an outbreak to other foods, we did not designate it as a beef-associated outbreak.

We removed isolates from the data set if: 1) No pre-computed assembly was available on NCBI, 2) SKESA v. 2.2 was not used to construct the assembly, 3) The number of contigs representing the assembly was greater than 300, and 4) The contig n50 was less than 25,000 bp. After initial parsing for isolation sources and assembly quality, we included serovars with 50 or more isolates in the analysis. In total, the final analysis set includes 12,337 assemblies and represents 37 serovars.

In silico serovar prediction

We used *Salmonella in silico* Typing Resource (SISTR) [30] with default options to assign putative serovars to each assembly. 1,077 assemblies failed the quality control step within the SISTR software with the same error message "FAIL: Wzx/Wzy genes missing…", but all 330 / 330 genes for the core genome multilocus sequence typing (cgMLST) scheme used within the software were present within these assemblies. We retained assemblies failing QC with the aforementioned error message and because they contained all 330 cgMLST loci for the analysis. S1 table provides full details of the SISTR serotyping results. We excluded from the analysis any assemblies which failed the quality control step and did not have all 330 cgMLST genes.

Virulence gene annotation

We collated a custom database of putative virulence factors from *Salmonella*, *Escherichia*, *Shigella*, and *Yersinia* from the virulence factor database (VFDB) (accessed 9/13/2021) [31] and putative virulence factors from *Salmonella*, *Escherichia*, and *Shigella* from PATRIC (accessed 9/13/2021) [32]. Next, we combined amino acid sequences of the open reading frames (ORF) with a reference proteome of *Salmonella* Typhimurium LT2 (https://www.uniprot.org/proteomes/UP000001014) and made the database non-redundant by clustering the open reading frames at 0.90 global identity using cd-hit [33]. We passed the

resultant database to Prokka [34] using the "--proteins" option to specify the database as the primary annotation database in the software pipeline. We then parsed gene annotations from the VFDB and PATRIC non-redundant database from the resultant Prokka annotation tables. Additionally, to ensure consistent ORF predictions between assemblies, we trained a model using Prodigal [35] on the chromosome of the reference Salmonella Typhimurium LT2 assembly ASM694v2 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000006945.2/) and passed the training file to Prokka using the command "--prodigaltf".

pESI plasmid identification

The pESI plasmid is an emerging concern in some *S. enterica* serovars, namely *S.* Infantis [36]. To confirm the presence of this plasmid, we complied an additional nucleotide database of 13 marker genes previously used to identify pESI plasmids in *S. enterica* contig assemblies from two sources [36,37]. We then conducted a nucleotide BLAST search against the database using the software Abricate [38] and defined positive hits as a percent identify of ≥ 95% and percent coverage of ≥ 95%. We presumed that an isolate contained the pESI plasmid if the contig assembly was positive for the pESI specific *repA* gene and contained at least five additional marker genes.

Random forest model construction

We used virulence gene annotations from the resultant Prokka outputs and constructed a count matrix of virulence genes. We excluded putative virulence loci with a prevalence of greater than 0.95 or which were not present in at least 10 (0.0008) assemblies. To generate row similarity (isolate relatedness), we fit an unsupervised random forest to the count matrix of virulence loci (assemblies x virulence factors) using the randomForest [39] package in R[40].

The random forest model contained 50,000 trees and used 60 features (columns, virulence loci) at every split.

Grouping isolates and assessing cluster stability

We converted the row-wise proximity matrix from the random forest model to a distance matrix (1 – similarity) and subjected it to agglomerative clustering using Ward's method [41]. We conducted two analyses: 1) k = 5, used to group serovars by virulence and 2) k = 37, using the same number of clusters as serotypes in the data to identify possible subpopulations within serovars with different virulence gene catalogues. We defined cluster stability for both analyses as a Jaccard similarity of≥ 0.75 [42] for 10,000 non-parametric bootstrap samples. For the main serovar clustering analysis, we choose k=5 as it was the highest value for which all clusters were stable. The k=37 version of the analysis tested whether distribution of virulence factor combinations is more similar within each serotype than between serotypes and, therefore, whether serotype is a reasonable representation of these virulence differences. We defined serovar subpopulations as serovars with at least two of these populations annotated into different clusters (k = 37 clustering) and with at least two of the populations representing greater than or equal to 0.20 of the total serovar population. Clustering and bootstrapping were performed using the clusterboot function from the fpc [43] package in R.

Epidemiological indicators

We estimated epidemiological indicators for both virulence clusters and serovars using sporadic and domestically acquired cases from 2016 – 2019. We excluded outbreak-associated cases to decrease bias due to outbreak size and removed travel-related cases to exclude foodstuffs from regions that may have different *S. enterica* population structures than the US and are not good indicators of US consumer exposure to *S. enterica* in food. We calculated the

196 proportion of positive samples for *Salmonella*, individual serovars, and each cluster in beef
197 products from FSIS regulatory testing programs: MT43 (raw ground beef), MT60
198 (manufacturing trimmings) and MT64 (Components other than Trim) for 2016 – 2019. We
199 modeled all proportions using *Beta*(s+0.5, n-s+0.5) (eqn. 1), with a Jeffery's prior (*Beta*(0.5, 0.5)
200 as a Bayesian conjugate to the Binomial distribution[44]. Table 1 lists parameters *s* and *n* used to
201 model these proportions. We used 1M Monte Carlo simulations to estimate and compare
202 posterior distributions using numerical integration, with a 99% confidence level for statistical
203 significance.
204

205 **Table 1. Description of parameters used for calculation of epidemiological indicator**
206 **estimations using Equation 1**

Model variable / Epidemiological indicator	Parameter description	
	s (numerator)	*n* (denominator)
Salmonella proportion positive in beef (p_+)	(s_+): number of FSIS samples positive for Salmonella	(n_+): number of FSIS samples taken
Hospitalization proportion	$H_{k:}$ number of hospitalizations from cluster k (k=1…5)	I_k : number of total infections in FoodNet from cluster k (k=1…5)
Extraintestinal Infection proportion	EI_k: number of extraintestinal infection from cluster k (k=1…5)	
Mortality Proportion	M_k: number of deaths from cluster k (k=1…5)	
Proportion of gene presence within respective cluster group ($p_{g,k}$)	$n_{k,g}$: number of isolates from cluster (k=Cluster 1 vs. clusters 2-5) with gene g (g=1…182)	n_k number of isolates from cluster k (k=Cluster 1 vs. clusters 2-5)

207

208 ## Incidence of domestically acquired sporadic cases

209 We modeled the incidence of domestically acquired sporadic cases per 100k people per
210 year (λ_{ij}) for serovar *j* in FoodNet state *i* using the Bayesian conjugate for a Poisson rate with
211 Jeffrey's prior *Gamma*(0.5, 0.00001)(44), hence $Gamma(\alpha_i + 0.5,\ \beta_i t + 0.00001)$(eqn. 2),
212 where α_{ij} is the serovar case totals per state and β_i is the FoodNet catchment area population for
213 *t*=4 years. Sporadic cases are defined as illnesses which were not linked to a known outbreak.

The catchment area of FoodNet is not evenly distributed between the 10 participating states, so the population-weighted mean serovar incidence (λ_j) for the study period was $\sum_{i=1}^{10} \lambda_{ij} p_i$, where λ_{ij} is the FoodNet state-specific mean serovar incidence and p_i is the state catchment proportion of total FoodNet population. Virulence cluster population-weighted average incidence, λ_k, is the sum of the clusters' constituent serovars incidence rates for the study period ($\lambda_k = \sum_{j=1}^{j_k} \lambda_j$).

Serovar proportion positive in beef

We determined the proportion of *Salmonella* positives (p_+) following eqn. 1, with s_+ number of samples positive for *Salmonella* and n_+ total number of samples from FSIS testing. We estimated the proportion of serovar *j* isolated from beef products (X_j) with a Dirichlet distribution, ($Dir(a_j)$ eqn. 3), where a_j is the number of FSIS isolates from serovar *j*. We excluded serovars without any positive isolates in FSIS testing and retained all serovars from the testing program (including those not included in the analysis set).

Serovar proportion positive (p_{j_+}) was taken as the product of total *Salmonella* proportion positive (p_+) and the serovar proportion of the total *Salmonella* population from eqn. 3 (X_j). Finally, we derived the cluster proportion positive (p_{k_+}) as the summation of the cluster's constituent serovars.

Hospitalization, Extraintestinal Infection, and Mortality Proportions

We determined the proportion of infections with a certain outcome (i.e., hospitalization, extraintestinal infections, and mortality) for each cluster *k* (k=1…5) and for cluster 1 vs the combinations of others. We defined extraintestinal infections as having "URINE", "BLOOD", "ORTHO", "ABSCESS", "OTHER STERILE SITE" and "CSF" isolation sources. We modeled all the proportions using eqn. 1, with parameters described in Table 1.

Differential gene carriage

To identify genes differential between cluster 1 and combined clusters 2-5, we trained a supervised random forest model (ntree = 5,000, features to try at split = 13) to classify isolates into two groups: cluster 1 and clusters 2-5. We extracted variable importance from the random forest model and defied gene importance using the mean decrease of Gini impurity. As with other proportions, we used eqn. 1 to model the proportion of gene presence $(p_{g,k})$ within respective cluster group (1 vs. 2-5) for each of the virulence genes used in the random forest. The relative frequency (RF) of a given gene was the resultant ratio of proportions of gene presence ($RF = p_{g_1} / p_{g_{(2-5)}}$).

Code and data availability

Aggregated data and code written for this analysis are available in our online repository: link will be provided upon publication (private links shared with editor and reviewer). FoodNet data is used with permission from the Centers for Disease Control and Prevention and although raw data may not be shared, code written from aggregated inputs is provided in our online repository.

Results

Genome assemblies analyzed

The Pathogen Detection Network hosted by NCBI contains over 400,000 sequenced *Salmonella* isolates from various sources and contributors. From these, we extracted 53,849 isolates from specific sampling programs, further reduced to a final analysis set to establish an analysis set of 12,337 *S. enterica* assemblies of human clinical cases in the US and bovine and

beef associated isolates, representing 37 major serovars (Fig 1). Approximately 55% (6,751) assemblies are from US human clinical infections with the remaining 45% (5,586) representing isolates from bovine animals and beef products. The metadata for the genomes analyzed is provided in S2 Table.

Fig 1. Analysis set of genomes. Description of the *S. enterica* genome assemblies considered, and exclusion and inclusion criteria applied to generate the analysis set.

Clustering serovars using isolate virulence gene catalogues

To establish clusters of serovars, we identified virulence genes from each assembly and compiled them into a count matrix, trained an unsupervised random forest model to approximate similarity between isolate virulence gene catalogues (Fig 2), and subjected the resultant isolate similarity matrix to agglomerative clustering to identify clusters with subsequent non-parametric bootstrapping to validate cluster stability.

We identified five stable clusters of *S. enterica* isolates (Fig 3A), with the majority of serovar isolates resident within the same clusters (mean within serovar cluster proportion = 0.96). (Fig 3B). However, *S.* Reading (cluster 1: n = 28 (0.47), cluster 3: n = 32 (0.53)), *S.* Saintpaul (cluster 3: n = 134 (0.66), cluster 1: n = 68 (0.34)), and *S.* I 1,4,[5],12:b:- (cluster 1: n = 53 (0.66), cluster 3: n = 30 (0.34)) had at least 33% of total serovar isolates in two different virulence clusters. The five virulence clusters are of uneven size (Fig 3C) with cluster 1 containing almost 10 times more assemblies than cluster 2. We attempted to decrease the size of cluster 1 by introducing a sixth cluster. However, the sixth cluster was unstable (bootstrap Jaccard similarity = 0.515) and cluster 4 was split, not cluster 1, indicating that the variance (Ward's method used to cluster) within cluster 1 is less than that of cluster 4 despite its much larger size (S1 Fig). Interestingly, cluster 2 is comprised of only *S.* Javiana and the cluster

homogeny of *S.* Javiana was preserved with the addition of the sixth cluster (S1 Fig). Assembly cluster designations are provided in Table S3.

Fig 2. Conceptual model of virulence cluster development. First, we downloaded contig assemblies and quality controlled for fragmentation followed by the identification of virulence genes. We then fit an unsupervised random forest model to the isolate level virulence gene catalogues to approximate relatedness. We converted the resultant similarity matrix to a distance matrix (1 – similarity) and clustered using Ward's method. We identified five stable clusters and validated using non-parametric bootstrapping.

Fig 3. Description of the five virulence clusters. (A) Dendrogram depicting the hierarchical relationship between 12,337 *S. enterica* genome assemblies based upon virulence gene carriage with the five virulence clusters superimposed on top. (B) Heatmap of serovar proportion within each of the five respective virulence clusters. Rows are clustered using Ward's method. (C) Characteristics of the five virulence clusters: cluster stability - Jaccard similarity of 10,000 non-parametric bootstraps, Number of Genomes - depicting the number of S. enterica genomes constituent in each cluster, and number of serovars (within cluster serovar proportion > 0.5) in each cluster.

General epidemiological characteristics of virulence clusters

To investigate if the genomic virulence clusters correspond to clinical case presentation, we computed basic epidemiological characteristics per cluster for 2016-2019 as proxies for virulence phenotypes: proportion positive in beef products, number of outbreaks, incidence of domestically acquired sporadic cases per 100k people per year, hospitalization proportion given infection, extraintestinal infection proportion given infection, and mortality proportion given infection. We computed the results by virulence cluster (S4 Table) and by serovar (S5 Table).

Not every *S. enterica* captured during surveillance programs in the US is subjected to sequencing, therefore we attributed cases from a given serovar to the cluster to which the highest proportion of serovar isolate was assigned (e.g., 98.5% of *S.* Typhimurium isolates were resident in cluster 1, therefore all cases of *S.* Typhimurium in the datasets were allocated to cluster 1). Cluster 1 serovars have the highest incidence rate of domestically-acquired sporadic cases (5.9 cases per 100k population per year, 99% CrI: 5.77 – 6.06) (Fig 4A), approximately 1.5x higher than that of clusters 2-5 combined during 2016 – 2019 (incidence rate ratio: 1.5, 99% CrI: 1.44 – 1.55). Moreover, infections from serovars in cluster 1 had a higher proportion of hospitalizations than serovars in cluster 2 (relative frequency (RF): 1.10, 99% CrI: 1.002 – 1.200), cluster 4 (RF: 1.15, 99% CrI: 1.029 – 1.296), and cluster 5 (RF: 1.17, 99% CrI: 1.058 – 1.288) (Fig 4BThe cluster 1 proportion positive in beef products was less than half of clusters 3-5 (proportion positive ratio: 0.44, 99% CrI: 0.366 – 0.528) (Fig 4C). However, cluster 1 serovars were implicated in the highest proportion of total foodborne outbreaks and beef associated outbreaks in the US from 2016 – 2019 (Fig 4D),generating approximately 2.5x more beef associated outbreaks (20 vs. 8) than clusters 3-5 combined (There were no cluster 2 isolates found in beef sampling or in beef associated outbreaks). Additionally, cluster 1 serovars were involved in approximately 1.47x more foodborne outbreaks than clusters 2-5 combined from 2016 - 2019 (285 vs. 194).

Fig 4. Epidemiological indicators of the five virulence clusters for the study period 2016-2019. (A) Incidence of domestically acquired sporadic cases per 100k population per year by virulence cluster. (B) Proportion of clinical infections resulting in hospitalization by virulence cluster. (C) Proportion positive estimates in FSIS testing of US beef products. No isolates from

cluster 2 (comprised solely of *S.* Javiana) were retrieved from 2016 – 2019. (D) Proportion of total US foodborne and beef-associated outbreaks attributed to serovars in the analysis set.

Differential Carriage of Virulence Loci Between Cluster 1 and Clusters 2-5

Cluster 1 serovars yielded the largest number of foodborne outbreaks and the highest incidence rate of domestically acquired sporadic cases. Additionally, a clear bifurcation exists between cluster 1 and clusters 2-5 (Fig 3A). Therefore, we sought to identify virulence genes differentially present between clusters 1 and clusters 2-5. We estimated the proportion of gene presence for the top 20 most differential genes, as determined by Gini impurity, in cluster 1 and clusters 2-5 combined and derived their RF (Table 2). The top two genes (*f17d-D*, and *f17d-C*) which best differentiated cluster 1 from the others (highest mean decrease of Gini impurity) were found in much lower proportion in cluster 1 (*f17d-D*, RF: 0.07, 99% CrI: 0.057 – 0.075)(*f17d-C*, RF: 0.043, 99% CrI: 0.036 – 0.051) (Table 2). The full list of RF for each virulence gene tested, mean decrease of Gini impurity, and gene metadata are provided in S6 Table.

Table 2. Importance and relative frequency of select virulence loci between cluster 1 and clusters 2-5.

Gene	Mean Decrease Gini Impurity	Relative Frequency (Cluster 1 / Clusters 2-5)	Lower Bound (0.005)	Upper Bound (0.995)
f17d-D	504.64	0.0657	0.05743	0.07459
f17d-C	455.07	0.0432	0.03618	0.05085
rfbI	405.35	4.3706	4.12059	4.64112
rfbG	394.86	4.3704	4.12070	4.64015
rfbH	336.41	4.3579	4.10740	4.62863
csaB	279.77	0.0013	0.00032	0.00323

cfaA	266.06	0.0011	0.00021	0.00285
lpfE	211.35	3.1146	2.97288	3.26624
faeD	209.61	0.0004	0.00001	0.00162
csbC	207.19	0.0033	0.00145	0.00608
lpfD	205.67	3.1131	2.97157	3.26366
faeE	204.28	0.0004	0.00001	0.00166
sseI/srfH	202.37	12.5251	11.06216	14.21786
lpfA	179.38	3.1130	2.97120	3.26442
lpfC	173.75	3.0919	2.95155	3.24155
lpfB	161.87	3.0812	2.94158	3.22994
rfbC	140.29	2.7218	2.60974	2.84109
faeC	112.30	0.0004	0.00001	0.00186
rfbD	109.33	2.3555	2.26954	2.44697
rfbP	104.21	2.3741	2.28674	2.46708

Table 2 legend. The top 20 most differential virulence genes between cluster 1 and clusters 2-5 as identified by a supervised random forest model. Genes are arranged in descending order of mean decrease of Gini impurity. RF of gene carriage is the proportion of gene presence in for cluster 1 divided by that in clusters 2-5 combined.

Within serovar virulence subpopulations

Horizontal gene transfer molds virulence gene carriage, especially within SPI [45,46]. We hypothesized that horizontal gene transfer may lead to virulence subpopulations that could be identified using random forest methods otherwise missed in more traditional alignment-based phylogeny methods. To test this hypothesis, we increased the number of clusters to correspond to the number of serovars (k = 37). If no virulence subpopulations are present (within serovar variance is less than between serovar variance), each of the 37 clusters should contain a majority of one serovar (see methods). However, we found 11 serovars with virulence subpopulations (Table 3). The full list of subpopulation designations is provided in S7 Table. To test if virulence subpopulations may correspond to phenotypic differences in case presentation, we computed the proportion of clinical infections resulting in extraintestinal infections for each serovar

subpopulation for sequenced strains with case presentation in the FoodNet surveillance system. Two serovars yielded significant differences in invasiveness between serovar subpopulations. *S.* Infantis split into two subpopulations (subpopulation 18: n = 145, subpopulation 20: n = 243) as shown in Fig 5A. The genome assembly size for subpopulation 18 isolates was significantly longer (4.98 Mb vs. 4.68 Mb, p-value < 2.2E-16, Mann-Whitney U test) (Fig 5B) than isolates from subpopulation 20. Of the 388 *S.* Infantis genome assemblies in the analysis set, 242 had associated clinical presentation data from FoodNet split evenly between the two subpopulations (n = 121, n = 121). Isolates from subpopulation 18 were more than twice as likely to result in extraintestinal clinical infections than isolates from subpopulation 20 (RF: 2.06, 99% CrI: 1.122 – 3.778) (Fig 5C). There was an association between subpopulation 18 isolates and older patients (median age 56.1 years) when compared to subpopulation 20 isolates (median age 36.4 years) (p-value: 5.00E-6, Mann-Whitney U test) (Fig 5D).

We hypothesized that the approximately 300kb difference between the assembly lengths of the *S.* Infantis subpopulations may be due to the presence of the pESI plasmid previously identified in *S.* Infantis(36). After checking all isolates for the presence of this plasmid,144 out of 145 *S.* Infantis isolates annotated to subpopulation 18 and 0 out of 243 isolates from subpopulation 20 were putatively positive for pESI plasmids. Only one isolate, a *S.* Muenchen, was putatively positive for the pESI plasmid outside of the *S.* Infantis 18 subpopulation.

Two subpopulations represented approximately 85% of the total *S.* Typhimurium population in the analysis set, which we analyzed further (Fig 6A). Similar to the *S.* Infantis subpopulations, the two subpopulations yielded significantly different genome assembly lengths (subpopulation 2: 4.90 Mb, subpopulation 18: 4.85 Mb, p-value < 2.2E-16, Mann-Whitney U test) (Fig 6B). However, the assembly difference of approximately 5kb between the *S.*

Typhimurium subpopulations is far less dramatic than the approximately 300kb difference observed between *S.* Infantis subpopulations. 668 of the 937 *S.* Typhimurium isolates in subpopulations 2 (n = 359) and 16 (n = 309) have clinical case presentation data. Subpopulation 2 isolates presented as double the extraintestinal infections than subpopulation 16 isolates (RF 2.11, 99% CrI: 1.109 – 4.016) (Fig 6C). In contrast with the *S.* Infantis subpopulations, the age of patients was not significantly different between the two subpopulations (p-value 0.97, Mann-Whitney U test) (Fig 6D).

Fig 5. Description of two *S.* Infantis virulence subpopulations. (A) Dendrogram highlighting the locations of the two *S.* Infantis virulence subpopulations within the greater population of 12,337 *S. enterica* isolates. (B) Histograms of the assembly lengths for the respective subpopulations. (C) Proportion of extraintestinal infections among illnesses caused by the two subpopulations (FoodNet data). (D) Boxplots of the distribution of patient age in infections caused by the two subpopulations (FoodNet data).

Fig 6. Description of two *S.* Typhimurium virulence subpopulations. (A) Dendrogram highlighting the locations of the two *S.* Typhimurium virulence subpopulations within the greater population of 12,337 *S. enterica* isolates. (B) Histograms of the assembly lengths for the respective subpopulations. (C) Proportion of extraintestinal infections among illnesses caused by the two subpopulations (FoodNet data). (D) Boxplots of the distribution of patient age in infections caused by the two subpopulations (FoodNet data).

Table 3. Within serovar virulence subpopulations.

	Subpopulation A	Subpopulation B

Serovar	Total Serovar Count	Subpopulation ID	Genome Count	Proportion of Total Serovar Population	Subpopulation ID	Genome Count	Proportion of Total Serovar Population
I 1,4,[5],12:b:-	83	30	30	0.36	29	52	0.63
I 1,4,[5],12:i:-	530	9	294	0.55	2	163	0.31
Infantis	388	20	243	0.63	18	145	0.37
Kentucky	169	32	89	0.53	21	80	0.47
Montevideo	1341	14	503	0.38	11	838	0.62
Muenchen	392	12	133	0.34	6	195	0.50
Newport	1751	6	427	0.24	1	926	0.53
Oranienburg	137	14	32	0.23	11	103	0.75
Reading	60	37	28	0.47	29	28	0.47
Saintpaul	202	29	53	0.26	27	134	0.66
Typhimurium	1106	16	477	0.43	2	460	0.42

Table 3 legend: Serovars found to contain at least two subpopulations, each representing greater than 0.20 of the total serovar population. Subpopulations were identified by increasing the number of clusters to match the number of serovars (k = 37). Provided is the subpopulation ID's, the number of genomes resident within each subcluster, and the proportion of the total population the subcluster represents. Note, that the subpopulations may not represent the total combined population of the serovar in the analysis set.

Discussion

The pathogenesis of *S. enterica* is not completely understood. Furthermore, how different serovars generate distinct disease pathologies is not well-defined either. To better understand how serovars group together based on virulence gene carriage, we used an unsupervised random forest, which allowed for rapid identification of serovars of public health concern. Compared to

methods used in previous studies [23,24], this scalable genomic approach allowed us to generate a measure of relatedness for a large number of *S. enterica* isolates in a computationally efficient manner and group them using established hierarchical clustering methods [41]. While we considered other clustering methods such as logistic principal component analysis and k-means clustering, we chose the unsupervised random forest approach because it is more robust to outliers, non-parametric, and aggregates results from many models rather than basing inference on a single, "best" model. Our method cannot be read as a traditional phylogeny of evolutionary process but rather as a snapshot of the current virulence potential of more than 12,000 isolates retrieved from humans, bovine animals, and beef products. We did not employ a traditional phylogeny as such classical alignment methods (core genome alignment, read mapping to a reference assembly, etc.) are computationally intensive given the number of samples we analyzed. Additionally, for this analysis, we were less interested in the evolutionary development of virulence, but rather in the current state of potential virulence that consumers are exposed to through beef products.

We contend that this method is pertinent to virulence loci found with SPI as the regions are subject to horizontal gene transfer [45,46]. Common methods to differentiate serovars typically rely on the alignment of core genes or single nucleotide polymorphisms (SNP) identified against reference assemblies [6,24,47,48]. These methods must rely on *post hoc* analysis to determine if two evolutionary similar strains have acquired virulence genes which may correspond to differences in case presentation as witnessed in *S.* Infantis and *S.* Typhimurium. Demarcating isolates by the presence/absence of virulence genes identified a cluster of serovars (cluster 1) that accounts for a large proportion of sporadic cases and beef-associated and total foodborne outbreaks compared to the other clusters combined. The higher

441 occurrence of beef-associated outbreaks occurs despite a much lower frequency of isolation from
442 regulatory beef samples relative to serovars from the other clusters combined. Our method, in
443 combination with quantitative risk assessment techniques could be used to account for the
444 relative exposure to serovars (e.g., via different food consumption) and the resultant probability
445 of disease.

446 FoodNet isolates are the basis of the incidence calculation, and the dataset does not
447 provide source attribution. Therefore, it is probable that serovars from the sporadic human
448 clinical data set are from multiple exposure sources (poultry, beef, vegetables, etc.). However,
449 even with the expected wide source range of *S. enterica* isolates, most serovars resided in one of
450 the five major virulence clusters (including beef and bovine isolates) suggesting that basal
451 virulence gene carriage is conserved within serovars across sources.

452 Cluster 1 serovars generate more human infections than serovars in the other clusters.
453 Our supervised random forest model identified virulence genes involved in attachment to host
454 cells or outer membrane structure as the most differential genes between cluster 1 and 2-5 (as
455 measured by mean decrease Gini impurity). The F17 fimbriae genes *f17-C* and *f17-D*, coli
456 surface antigen operon (*csa*) and the *fae* fimbria operon were absent from cluster 1. All three
457 operons originated from *Escherichia* virulence databases. Use of only putative *Salmonella*
458 virulence factors from PATRIC [32] and the Virulence Factor Database [31] would not have
459 annotated the open reading frames highlighting the need for expanding the putative virulence
460 factors of *S. enterica* outside of the genera to members of Enterobacteriaceae. However, we
461 found two operons, *lpf* (long polar fimbriae) and *rfb*, in significantly higher proportions in cluster
462 1 than clusters 2-5. The higher proportion of *lpf* genes is notable as the operon has been
463 associated with *S. enterica* binding the Peyer's patches, namely the M-cells found within the

lymphoid organs. [49,50]. This may potentiate the higher infectivity of cluster 1 serovars as recent work with the Type Three Secretion systems (T3SS) of *S. enterica* (involved with the introduction of effector proteins to the cytoplasm of host cells) suggest that the structure does not penetrate the cytoplasmic membrane like a syringe, but requires tension and adopts a "tent-pole" like structure [51]. If tension is required for the function of the T3SS, enhanced binding to M-cells mediated by the *lpf* operon may be one reason cluster 1 has a higher incidence rate of domestically acquired sporadic cases. The *lpf* operon has not only been implicated in *S. enterica* infections; strains of *Escherichia coli* O157:H7 with mutations in the *lpf* operon show decreased attachment and colonization in both in vitro [52] and in vivo [53] models again show the interplay of virulence mechanisms between Enterobacteriaceae. The roles of *rfb* genes are not as well investigated as the *lpf* operon but appear to be involved in the biosynthesis of the O-antigen and lipopolysaccharide structuring. A recent report suggests that the full complement of *rfb* genes leads to higher virulence in experimentally infected chickens [54].

Examining virulence gene catalogues not only identified large, serovar level clusters but also, by altering the cluster number (k value), virulence subpopulations within serovars. With the current method, it cannot be ascertained whether the virulence subpopulations represent polyphyletic clades within serovars as it cannot be interpretated as a phylogeny. However, by applying a top-down approach, the presence of increased virulence capacity can be readily identified. The two subpopulations of *S.* Infantis present over a two-fold difference in probability of extraintestinal infections. *S.* Infantis has been rapidly increasing in incidence in Israel and previous studies suggest that the addition of a virulence megaplasmid pESI could be responsible [55]. The mean difference between the two subpopulations was approximately 300kb, similar in length to the pESI plasmid (280 kb), and querying *S.* Infantis isolates against a database of

marker genes revealed that isolates in the subpopulation with longer assemblies are putatively positive for pESI presence. In addition, the pESI plasmid carries genes necessary for the synthesis of yersiniabactin [55], a siderophore dependent iron uptake system commonly observed in *Yersinia pestis*. The eight genes comprising the *ybt* operon are resident in every strain of the higher invasive cluster of *S.* Infantis, and only two out of 243 isolates from the lower invasive cluster contain the operon. Iron is an essential nutrient for *S. enterica* replication during systemic infections [56]. A previous study suggests that co-infections of Malaria and *S. enterica* leads to more systemic infections as excess iron is released upon the lysis of red blood cells, liberating the metal for use by *S. enterica* [57]. Increased iron availability, due to the addition of yersiniabactin, may be one factor for the almost double rate of extraintestinal infections of *S.* Infantis cluster 18 compared to *S.* Infantis infections without this plasmid.

The methods employed here cannot identify virulence changes due to sequence variations within virulence loci. Variants of the *macA* and *macB* genes in African strains of *S.* Typhimurium sequence-type 313 may have higher invasiveness in human patients and increased survival against challenge with antimicrobial peptides [58]. Others have identified virulence gene alleles that may correspond to pathogenicity differences [59]. The method employed identifies virulence genes against a non-redundant database using BLASTP, so alleles with variation less than 10% sequence identity will be collapsed into the same gene annotation. Furthermore, we did not consider pseudogene formation of virulence genes. Previous work suggests that pseudogenes in *S. enterica* genomes do not follow neutral evolution (random genetic drift, as in many Eukaryotes) but are readily lost from the chromosome [60]. However, pseudogene formation of the *sseI/srfH* secreted effector protein (higher carriage in cluster 1, Table 2) leads to hyperdissementation of ST313 *S.* Typhimurium in experimentally infected mice [61]. The role of

pseudogene formation and the pathogenesis needs more study, and the addition of pseudogene information could further improve virulence classifications. Additionally, we chose to focus our analysis on human, bovine, and beef isolates from the US. It is probable given the diversity of *S. enterica* that all virulence patterns and serovar subpopulations are not represented in this work.

S. enterica is a diverse pathogen. Yet, most risk assessments and food safety regulations informed by these assessments only separate Typhoidal Salmonellosis and non-Typhoidal Salmonellosis, treating serovars as a homogenous unit [62-64]. However, our results suggests that strains with the highest incidence of domestically acquired sporadic cases and outbreaks of human infections share a common virulence repertoire. Control and surveillance programs should devote resources to identifying and eliminating the major reservoirs of clinically relevant serovars. Furthermore, serovar virulence cannot be considered homogenous in all cases as observed with *S.* Infantis and *S.* Typhimurium. Although attributing virulence to specific genes was beyond the scope of this study, our analysis could inform further research to identify *Salmonella* genes associated with severe illness.

Acknowledgements

G.F, J.P., D.T., S.C., and F.J. are employed by EpiX Analytics. R.P. is a Senior Scientific Advisor for EpiX Analytics. This work utilized the Summit supercomputer, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University. The Summit supercomputer is a joint effort of the University of Colorado Boulder and Colorado State University. FoodNet Data: The findings and conclusions in this report are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

1. Srikantiah P, Lay JC, Hand S, Crump JA, Campbell J, Van Duyne MS, et al. Salmonella enterica serotype Javiana infections associated with amphibian contact, Mississippi, 2001. Epidemiol Infect. 2004 Apr;132(2):273–81.

2. Lawson B, de Pinna E, Horton RA, Macgregor SK, John SK, Chantrey J, et al. Epidemiological Evidence That Garden Birds Are a Source of Human Salmonellosis in England and Wales. PLOS ONE. 2014 Feb 26;9(2):e88968.

3. Salmonella Subcommittee of the Nomenclature Committee of the International Society for Microbiology. The Genus Salmonella Lignières, 1900. J Hyg (Lond). 1934 Oct;34(3):333–50.

4. Grimont P, Weill FX. Antigenic Formulae of the Salmonella serovars, (9th ed.) Paris: WHO Collaborating Centre for Reference and Research on Salmonella. Inst Pasteur. 2007 Jan 1;1–166.

5. Centers for Disease Control and Prevention (CDC). National Salmonella Surveillance Annual Report, 2016. Atlanta, Georgia: US Department of Health and Human Services: CDC; 2018.

6. Worley J, Meng J, Allard MW, Brown EW, Timme RE. Salmonella enterica Phylogeny Based on Whole-Genome Sequencing Reveals Two New Clades and Novel Patterns of Horizontally Acquired Genetic Elements. mBio. 2018 Nov 27;9(6):e02303-18.

7. Rivera-Chávez F, Bäumler AJ. The Pyromaniac Inside You: Salmonella Metabolism in the Host Gut. Annu Rev Microbiol. 2015 Oct 15;69(1):31–48.

8. Thiennimitr P, Winter SE, Winter MG, Xavier MN, Tolstikov V, Huseby DL, et al. Intestinal inflammation allows Salmonella to use ethanolamine to compete with the microbiota. Proc Natl Acad Sci U S A. 2011/10/03 ed. 2011 Oct 18;108(42):17480–5.

9. Drumo R, Pesciaroli M, Ruggeri J, Tarantino M, Chirullo B, Pistoia C, et al. Salmonella enterica Serovar Typhimurium Exploits Inflammation to Modify Swine Intestinal Microbiota. Front Cell Infect Microbiol [Internet]. 2016;5. Available from: https://www.frontiersin.org/article/10.3389/fcimb.2015.00106

10. Marcus SL, Brumell JH, Pfeifer CG, Finlay BB. Salmonella pathogenicity islands: big virulence in small packages. Microbes Infect. 2000 Feb;2(2):145–56.

11. Lorkowski M, Felipe-López A, Danzer CA, Hansmeier N, Hensel M. Salmonella enterica invasion of polarized epithelial cells is a highly cooperative effort. Infect Immun. 2014/04/07 ed. 2014 Jun;82(6):2657–67.

12. Steele-Mortimer O, Brumell JH, Knodler LA, Méresse S, Lopez A, Finlay BB. The invasion-associated type III secretion system of Salmonella enterica serovar Typhimurium is

necessary for intracellular proliferation and vacuole biogenesis in epithelial cells. Cell Microbiol. 2002 Jan 1;4(1):43–54.

13. Kurtz JR, Goggins JA, McLachlan JB. Salmonella infection: Interplay between the bacteria and host immune system. Immunol Lett. 2017/07/15 ed. 2017 Oct;190:42–50.

14. Gal-Mor O, Boyle EC, Grassl GA. Same species, different diseases: how and why typhoidal and non-typhoidal Salmonella enterica serovars differ. Front Microbiol [Internet]. 2014;5. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2014.00391

15. Harvey RR, Friedman CR, Crim SM, Judd M, Barrett KA, Tolar B, et al. Epidemiology of Salmonella enterica Serotype Dublin Infections among Humans, United States, 1968-2013. Emerg Infect Dis. 2017 Sep;23(9):1493–501.

16. Ramachandran G, Panda A, Higginson EE, Ateh E, Lipsky MM, Sen S, et al. Virulence of invasive Salmonella Typhimurium ST313 in animal models of infection. PLoS Negl Trop Dis. 2017 Aug 4;11(8):e0005697.

17. Jiang L, Wang P, Song X, Zhang H, Ma S, Wang J, et al. Salmonella Typhimurium reprograms macrophage metabolism via T3SS effector SopE2 to promote intracellular replication and virulence. Nat Commun. 2021 Feb 9;12(1):879.

18. Cheng RA, Eade CR, Wiedmann M. Embracing Diversity: Differences in Virulence Mechanisms, Disease Severity, and Host Adaptations Contribute to the Success of Nontyphoidal Salmonella as a Foodborne Pathogen. Front Microbiol [Internet]. 2019;10. Available from: https://www.frontiersin.org/article/10.3389/fmicb.2019.01368

19. Faber F, Thiennimitr P, Spiga L, Byndloss MX, Litvak Y, Lawhon S, et al. Respiration of Microbiota-Derived 1,2-propanediol Drives Salmonella Expansion during Colitis. PLOS Pathog. 2017 Jan 5;13(1):e1006129.

20. Hannemann S, Galán JE. Salmonella enterica serovar-specific transcriptional reprogramming of infected cells. PLOS Pathog. 2017 Jul 24;13(7):e1006532.

21. Pulford CV, Perez-Sepulveda BM, Canals R, Bevington JA, Bengtsson RJ, Wenner N, et al. Stepwise evolution of Salmonella Typhimurium ST313 causing bloodstream infection in Africa. Nat Microbiol. 2021 Mar 1;6(3):327–38.

22. Ebel ED, Williams MS, Schlosser WD. Estimating the Type II error of detecting changes in foodborne illnesses via public health surveillance. Microb Risk Anal. 2017 Dec 1;7:1–7.

23. Karanth S, Tanui CK, Meng J, Pradhan AK. Exploring the predictive capability of advanced machine learning in identifying severe disease phenotype in Salmonella enterica. Food Res Int. 2022 Jan 1;151:110817.

24. Chen R, Cheng RA, Wiedmann M, Orsi RH. Development of a Genomics-Based Approach To Identify Putative Hypervirulent Nontyphoidal Salmonella Isolates: Salmonella enterica Serovar Saintpaul as a Model. mSphere. 2022 Feb 23;7(1):e0073021.

603 25. National Advisory Committee on Microbiological Criteria for Foods (NACMCF). Response
604 to Questions Posed by the Food Safety and Inspection Service: Enhancing Salmonella
605 Control in Poultry Products. 2022. Available from:
606 https://www.fsis.usda.gov/sites/default/files/media_file/documents/NACMCF_Salmonella-
607 Poultry_Response_for_Committee_Review.pdf

608 26. Ward C. Vertical Integration Comparison: Beef, Pork, and Poultry. Western Agricultural
609 Economics Association; 1997. Available from:
610 https://EconPapers.repec.org/RePEc:ags:waeare:35759

611 27. Palma F, Manfreda G, Silva M, Parisi A, Barker DOR, Taboada EN, et al. Genome-wide
612 identification of geographical segregated genetic markers in Salmonella enterica serovar
613 Typhimurium variant 4,[5],12:i:-. Sci Rep. 2018 Oct 15;8(1):15251.

614 28. Fenske GJ, Thachil A, McDonough PL, Glaser A, Scaria J. Geography Shapes the
615 Population Genomics of Salmonella enterica Dublin. Genome Biol Evol. 2019 Aug
616 1;11(8):2220–31.

617 29. CDC. Foodborne Diseases Active Surveillance Network. Atlanta, Georgia: US Department
618 of Health and Human Services: CDC;

619 30. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, et al. The
620 Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly
621 Typing and Subtyping Draft Salmonella Genome Assemblies. PLOS ONE. 2016 Jan
622 22;11(1):e0147101.

623 31. Liu B, Zheng D, Jin Q, Chen L, Yang J. VFDB 2019: a comparative pathogenomic platform
624 with an interactive web interface. Nucleic Acids Res. 2019 Jan 8;47(D1):D687–92.

625 32. Mao C, Abraham D, Wattam AR, Wilson MJC, Shukla M, Yoo HS, et al. Curation,
626 integration and visualization of bacterial virulence factors in PATRIC. Bioinformatics. 2015
627 Jan 15;31(2):252–8.

628 33. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
629 sequencing data. Bioinformatics. 2012/10/11 ed. 2012 Dec 1;28(23):3150–2.

630 34. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul
631 15;30(14):2068–9.

632 35. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic
633 gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar
634 8;11(1):119.

635 36. McMillan EA, Wasilenko JL, Tagg KA, Chen JC, Simmons M, Gupta SK, et al. Carriage
636 and Gene Content Variability of the pESI-Like Plasmid Associated with Salmonella Infantis
637 Recently Established in United States Poultry Production. Genes. 2020 Dec 18;11(12):1516.

638 37. Franco A, Leekitcharoenphon P, Feltrin F, Alba P, Cordaro G, Iurescia M, et al. Emergence
639 of a Clonal Lineage of Multidrug-Resistant ESBL-Producing Salmonella Infantis
640 Transmitted from Broilers and Broiler Meat to Humans in Italy between 2011 and 2014.
641 PLoS ONE. 2015 Dec 30;10(12):e0144802.

642 38. Seemann T. ABRicate. 2022. [cited 2 December 2022]. Database: github [Internet].
643 Available from: https://github.com/tseemann/abricate

644 39. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–
645 22.

646 40. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna,
647 Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-
648 project.org/

649 41. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc.
650 1963;58(301):236–44.

651 42. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal. 2007 Sep
652 15;52(1):258–71.

653 43. Hennig C. fpc: Flexible Procedures for Clustering. 2020. Available from: https://CRAN.R-
654 project.org/package=fpc

655 44. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis.
656 3rd ed. Boca Raton: CRC Press; 2013.

657 45. Lerminiaux NA, MacKenzie KD, Cameron ADS. Salmonella Pathogenicity Island 1 (SPI-1):
658 The Evolution and Stabilization of a Core Genomic Type Three Secretion System.
659 Microorganisms. 2020 Apr 16;8(4):576.

660 46. Brown EW, Bell R, Zhang G, Timme R, Zheng J, Hammack TS, et al. Salmonella Genomics
661 in Public Health and Food Safety. EcoSal Plus. 2021 Dec 15;9(2):eESP00082020.

662 47. Pornsukarom S, van Vliet AHM, Thakur S. Whole genome sequencing analysis of multiple
663 Salmonella serovars provides insights into phylogenetic relatedness, antimicrobial resistance,
664 and virulence markers across humans, food animals and agriculture environmental sources.
665 BMC Genomics. 2018 Nov 6;19(1):801.

666 48. Petrovska L, Mather AE, AbuOun M, Branchu P, Harris SR, Connor T, et al. Microevolution
667 of Monophasic Salmonella Typhimurium during Epidemic, United Kingdom, 2005-2010.
668 Emerg Infect Dis. 2016 Apr;22(4):617–24.

669 49. Bäumler AJ, Tsolis RM, Heffron F. The lpf fimbrial operon mediates adhesion of Salmonella
670 typhimurium to murine Peyer's patches. Proc Natl Acad Sci U S A. 1996 Jan 9;93(1):279–
671 83.

672 50. Gonzales AM, Wilde S, Roland KL. New Insights into the Roles of Long Polar Fimbriae and
673 Stg Fimbriae in Salmonella Interactions with Enterocytes and M Cells. Infect Immun. 2017
674 Aug 18;85(9):e00172-17.

675 51. Park D, Lara-Tejero M, Waxham MN, Li W, Hu B, Galán JE, et al. Visualization of the type
676 III secretion mediated Salmonella-host cell interface using cryo-electron tomography. eLife.
677 2018 Oct 3;7.

678 52. Torres AG, Kanack KJ, Tutt CB, Popov V, Kaper JB. Characterization of the second long
679 polar (LP) fimbriae of Escherichia coli O157:H7 and distribution of LP fimbriae in other
680 pathogenic E. coli strains. FEMS Microbiol Lett. 2004 Sep 1;238(2):333–44.

681 53. Jordan DM, Cornick N, Torres AG, Dean-Nystrom EA, Kaper JB, Moon HW. Long polar
682 fimbriae contribute to colonization by Escherichia coli O157:H7 in vivo. Infect Immun. 2004
683 Oct;72(10):6168–71.

684 54. Gao R, Huang H, Hamel J, Levesque RC, Goodridge LD, Ogunremi D. Application of a
685 High-Throughput Targeted Sequence AmpliSeq Procedure to Assess the Presence and
686 Variants of Virulence Genes in Salmonella. Microorganisms. 2022 Feb 5;10(2).

687 55. Aviv G, Tsyba K, Steck N, Salmon-Divon M, Cornelius A, Rahav G, et al. A unique
688 megaplasmid contributes to stress tolerance and pathogenicity of an emergent Salmonella
689 enterica serovar Infantis strain. Environ Microbiol. 2014 Apr 1;16(4):977–94.

690 56. Nairz M, Ferring-Appel D, Casarrubea D, Sonnweber T, Viatte L, Schroll A, et al. Iron
691 Regulatory Proteins Mediate Host Resistance to Salmonella Infection. Cell Host Microbe.
692 2015 Aug 12;18(2):254–61.

693 57. van Santen S, de Mast Q, Swinkels DW, van der Ven AJAM. The iron link between malaria
694 and invasive non-typhoid Salmonella infections. Trends Parasitol. 2013 May;29(5):220–7.

695 58. Honeycutt JD, Wenner N, Li Y, Brewer SM, Massis LM, Brubaker SW, et al. Genetic
696 variation in the MacAB-TolC efflux pump influences pathogenesis of invasive Salmonella
697 isolates from Africa. PLOS Pathog. 2020 Aug 24;16(8):e1008763.

698 59. Rakov AV, Mastriani E, Liu SL, Schifferli DM. Association of Salmonella virulence factor
699 alleles with intestinal and invasive serovars. BMC Genomics. 2019 May 28;20(1):429.

700 60. Kuo CH, Ochman H. The Extinction Dynamics of Bacterial Pseudogenes. PLOS Genet.
701 2010 Aug 5;6(8):e1001050.

702 61. Carden SE, Walker GT, Honeycutt J, Lugo K, Pham T, Jacobson A, et al. Pseudogenization
703 of the Secreted Effector Gene sseI Confers Rapid Systemic Dissemination of S.
704 Typhimurium ST313 within Migratory Dendritic Cells. Cell Host Microbe. 2017 Feb
705 8;21(2):182–94.

706 62. Food Safety Inspection Service (FSIS). Public Health Effects of Performance Standards for
707 Ground Beef and Beef Manufacturing Trimmings. Washington D.C.: US Department of
708 Agriculture: FSIS; 2019.

709 63. Food Safety Inspection Service (FSIS). Public Health Effects of Raw Chicken Parts and
710 Comminuted Chicken and Turkey Performance Standards. Washington D.C.: US
711 Department of Agriculture: FSIS; 2015.

712 64. Lambertini E, Ruzante JM, Kowalcyk, BB. The Public Health Impact of Implementing a
713 Concentration□Based Microbiological Criterion for Controlling Salmonella in Ground
714 Turkey. Risk Analysis. 2021 Aug; 41(8):1376-95.

715

716 Supporting information

717 **S1 Fig**. **Addition of a sixth virulence cluster**. (A) Dendrogram depicting the hierarchical
718 relationship between 12,337 S. enterica genome assemblies based upon virulence gene carriage
719 with six virulence clusters superimposed on top. (B) Heatmap of serovar proportion within each
720 of the six respective virulence clusters. Rows are clustered using Ward's method. (C)
721 Characteristics of the six virulence clusters: cluster stability - Jaccard similarity of 10,000 non-
722 parametric bootstraps, Number of Genomes - depicting the number of *S. enterica* genomes
723 constituent in each cluster, and number of serovars (within cluster serovar proportion > 0.5) in
724 each cluster.

725

726 **S1 Table**. **Output of SISTR serovar prediction**. Full results of the in silico serovar prediction
727 for the analysis set genomes from the SISTR software.

728

729 **S2 Table**. **Metadata for the analysis set of genomes.** Detailed metadata for the contig
730 assemblies used in the analysis including BioSample, BioProject, and SRA accession numbers.

731

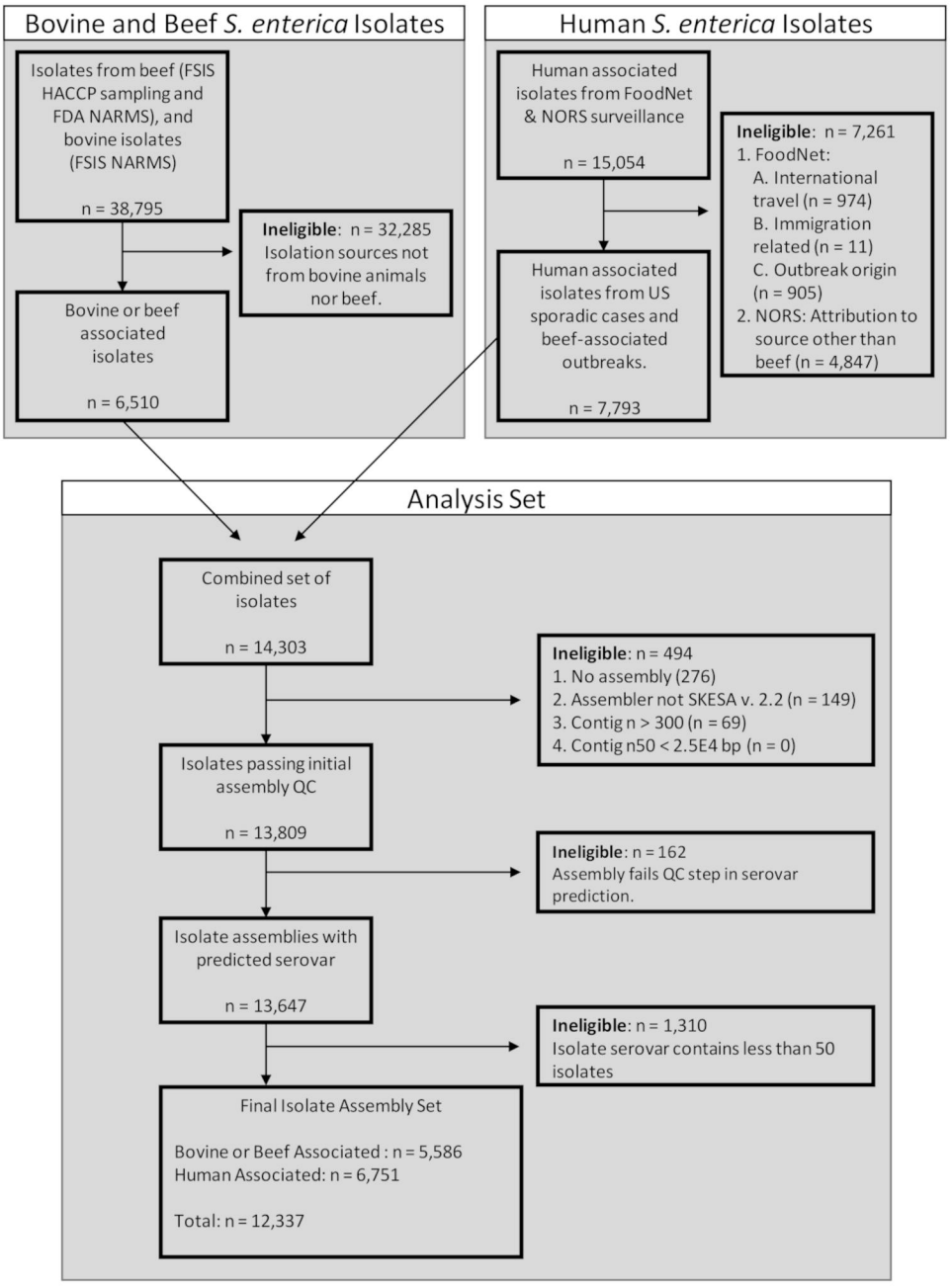
S3 Table. Isolate virulence cluster designations. Virulence cluster designations (k = 5) for the 12,337 contig assemblies in the analysis set including putative serovar designation.

S4 Table. Epidemiological indicators computed for each virulence cluster. Estimates of: incidence of domestically acquired sporadic cases per 100k people per year, hospitalization proportion given infection, proportion positive in FSIS testing of US beef products (MT43, MT60, MT64), extraintestinal proportion given infection, and mortality proportion given infection.

S5 Table. Epidemiological indicators computed for each serovar. Estimates of: incidence of domestically acquired sporadic cases per 100k people per year, hospitalization proportion given infection, proportion positive in FSIS testing of US beef products, extraintestinal proportion given infection, and mortality proportion given infection.

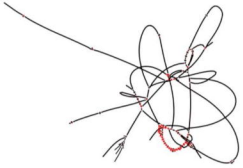
S6 Table. **Differential putative virulence loci between cluster 1 and clusters 2-5.** Relative frequency, mean decrease of Gini impurity, and basic gene metadata for virulence factors tested between cluster 1 and clusters 2-5 combined.

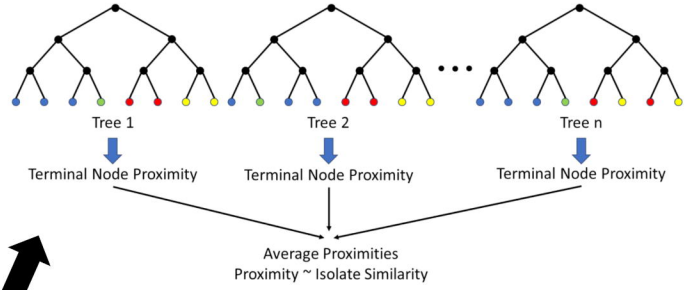
S7 Table. Isolate virulence subpopulation cluster designations. Subpopulation cluster designations (k = 37) for the 12,337 contig assemblies in the analysis set.


Bovine and Beef S. enterica Isolates
Isolates from beef (FSIS HACCP sampling and FDA NARMS), and bovine isolates (FSIS NARMS)
n = 38,795
Ineligible: n = 32,285
Isolation sources not from bovine animals nor beef.
Bovine or beef associated isolates
n = 6,510
Human S. enterica Isolates
Human associated isolates from FoodNet & NORS surveillance
n = 15,054
Ineligible: n = 7,261
1. FoodNet:
A. International travel (n = 974)
B. Immigration related (n = 11)
C. Outbreak origin (n = 905)
2. NORS: Attribution to source other than beef (n = 4,847)
Human associated isolates from US sporadic cases and beef-associated outbreaks.
n = 7,793
Analysis Set
Combined set of isolates
n = 14,303
Ineligible: n = 494
1. No assembly (276)
2. Assembler not SKESA v. 2.2 (n = 149)
3. Contig n > 300 (n = 69)
4. Contig n50 < 2.5E4 bp (n = 0)
Isolates passing initial assembly QC
n = 13,809
Ineligible: n = 162
Assembly fails QC step in serovar prediction.
Isolate assemblies with predicted serovar
n = 13,647
Ineligible: n = 1,310
Isolate serovar contains less than 50 isolates
Final Isolate Assembly Set
Bovine or Beef Associated : n = 5,586
Human Associated: n = 6,751
Total: n = 12,337

Compute Isolate Relatedness
(Unsupervised Random Forest)

Acquire and Quality Control Genome Assemblies

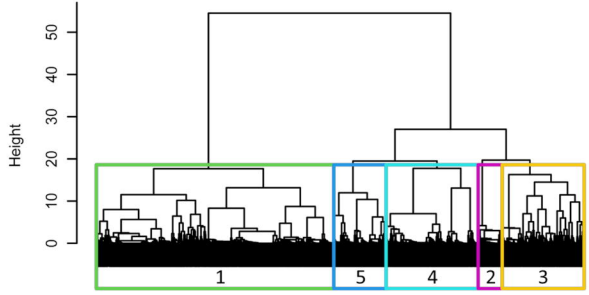




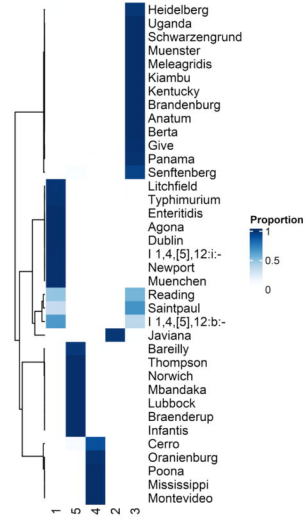
Identify Virulence Genes

Isolate A	11111100010011200101111
Isolate B	11110001110100021111000 1
Isolate C	01110001112000111111110 1
Isolate D	11111100010011200100000
Isolate E	11101000010111200101101
Isolate F	11111100010011200100111
Isolate G	11110001110100021111000 1
Isolate H	11110001110100021111111 2

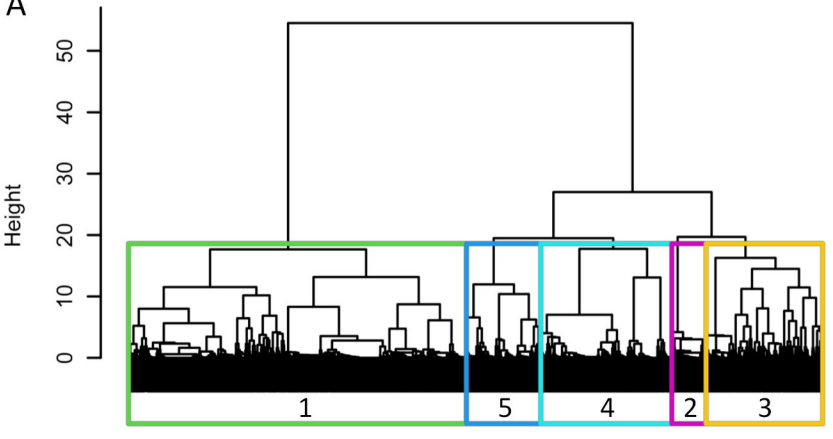
Isolate Cluster Generation
(Hierarchical clustering and non-parametric bootstrapping)



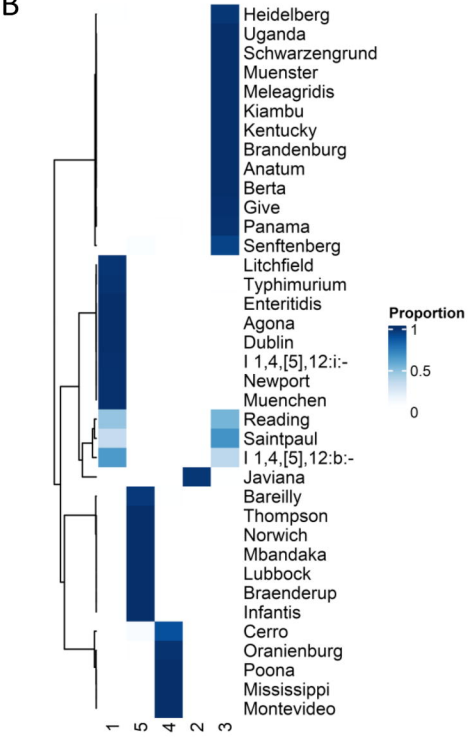
Serovar Virulence Clusters



A

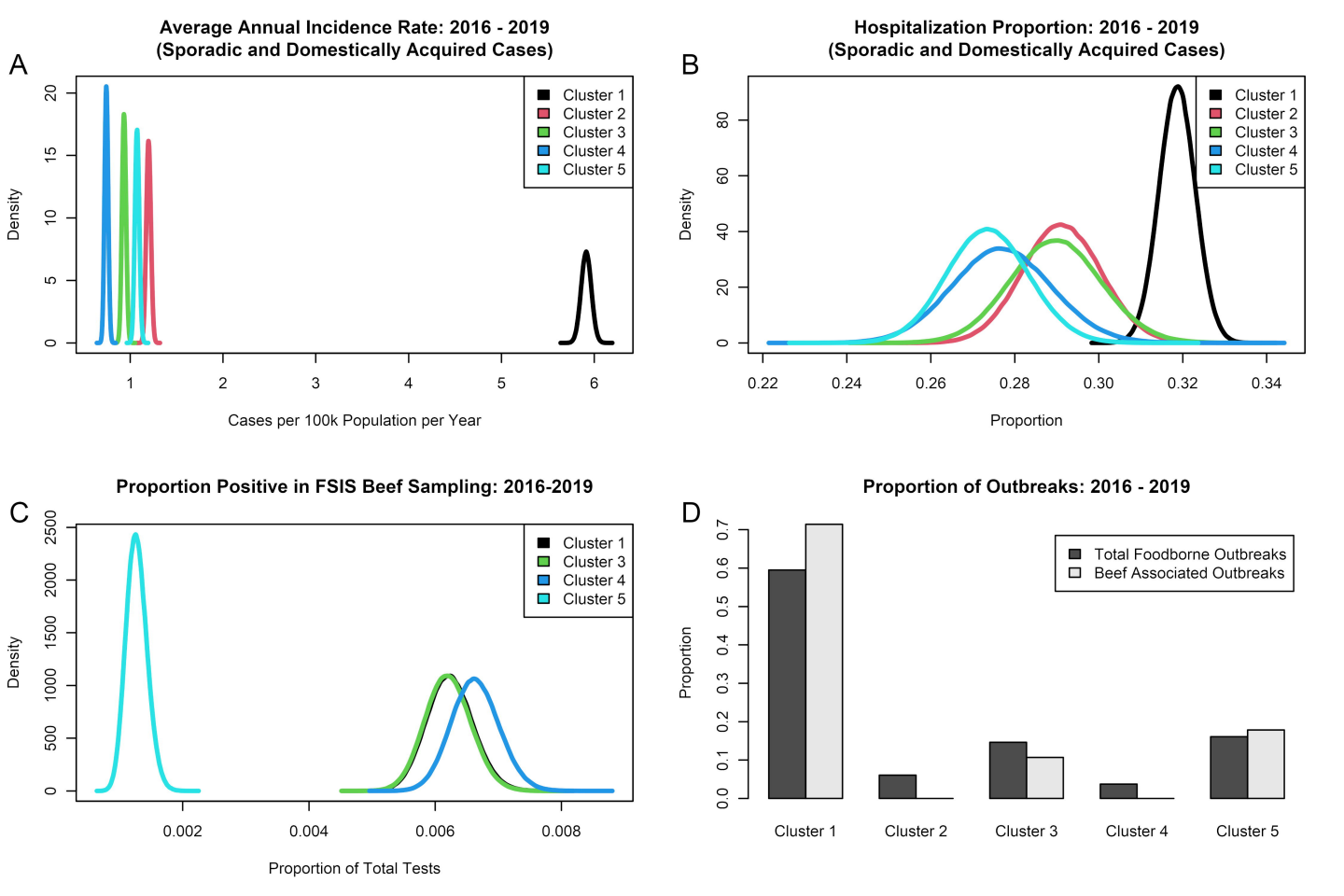


B

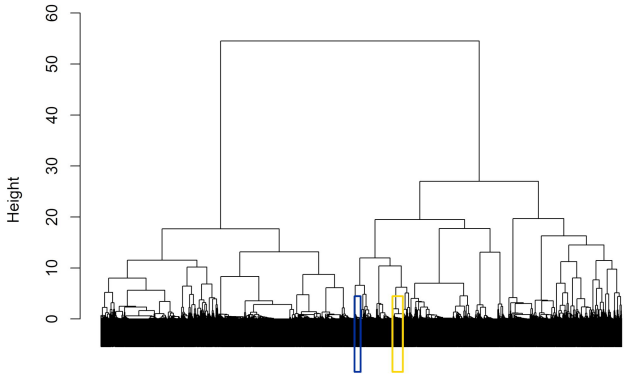


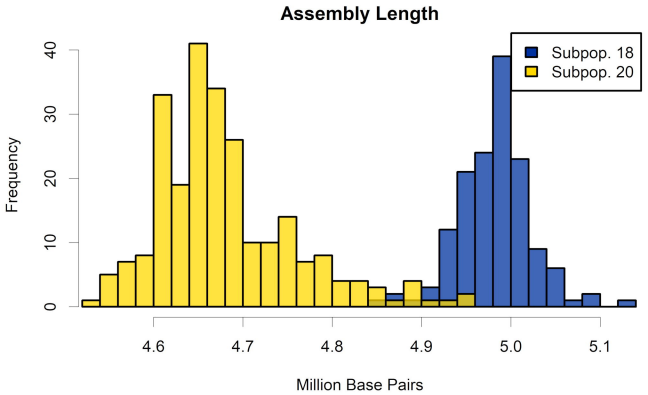
C

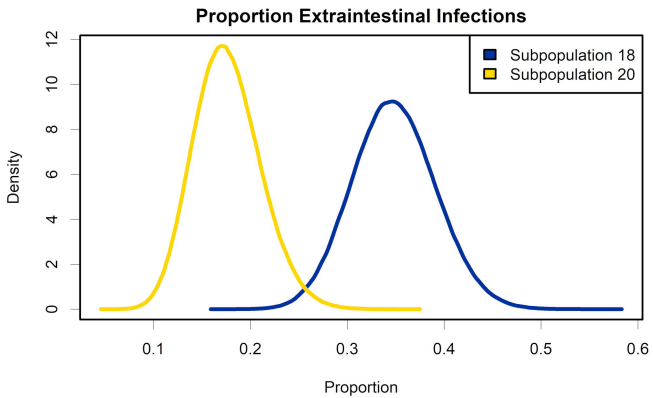
Cluster	Cluster Stability (Jaccard Similarity)	Number of Genomes	Serovar Count (prop > 0.50)
1	0.960	6006	9
2	0.887	608	1
3	0.838	2073	15
4	0.995	2329	5
5	0.987	1321	7

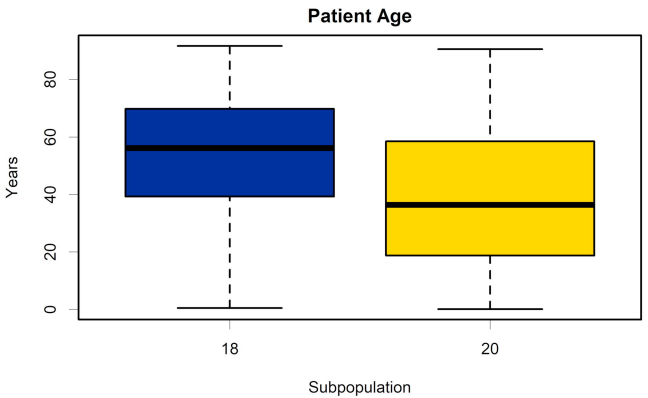

A
Average Annual Incidence Rate: 2016 - 2019
(Sporadic and Domestically Acquired Cases)
Density
Cases per 100k Population per Year
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
B
Hospitalization Proportion: 2016 - 2019
(Sporadic and Domestically Acquired Cases)
Density
Proportion
C
Proportion Positive in FSIS Beef Sampling: 2016-2019
Density
Proportion of Total Tests
D
Proportion of Outbreaks: 2016 - 2019
Proportion
Total Foodborne Outbreaks
Beef Associated Outbreaks

S. Infantis


Height
0
10
20
30
40
50
60


Assembly Length
Subpop. 18
Subpop. 20
Frequency
0
10
20
30
40
4.6
4.7
4.8
4.9
5.0
5.1
Million Base Pairs


Proportion Extraintestinal Infections
Subpopulation 18
Subpopulation 20
Density
0
2
4
6
8
10
12
0.1
0.2
0.3
0.4
0.5
0.6
Proportion


Patient Age
Years
0
20
40
60
80
18
20
Subpopulation

S. Typhimurium