

Worldwide comparison of *Salmonella* Enteritidis genomes with and without the *gyrA* codon substitution D87Y

David J. Bradshaw II

USDA-ARS, National Animal Disease Center

Kaitlin A. Tagg

Centers for Disease Control and Prevention

Bradley L. Bearson

USDA-ARS, National Animal Disease Center

Hattie E. Webb

Centers for Disease Control and Prevention

Shawn M.D. Bearson

`shawn.bearson@usda.gov`

USDA-ARS, National Animal Disease Center

Article

Keywords:

Posted Date: August 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10463527/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Worldwide comparison of *Salmonella* Enteritidis genomes with and without the *gyrA* codon substitution D87Y

David J. Bradshaw II^{1,2*}, Kaitlin A. Tagg³, Bradley L. Bearson⁴, Hattie E. Webb³, Shawn M.D. Bearson^{1*}

¹USDA-ARS, National Animal Disease Center, Ames, Iowa, 50010, U.S.A., Food Safety and Enteric Pathogens Research Unit

²Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee, 37831, U.S.A. , Agricultural Research Service Participation Program

³Centers for Disease Control and Prevention, Atlanta, Georgia, USA, Division of Foodborne, Waterborne, and Environmental Diseases

⁴USDA-ARS, National Laboratory for Agriculture and the Environment, Ames, Iowa, 50011, U.S.A., Agroecosystems Management Research Unit

***Corresponding authors:** Shawn M.D. Bearson (shawn.bearson@usda.gov) & David J. Bradshaw II (david.bradshaw@usda.gov)

Abstract

Salmonella enterica serovar Enteritidis (*S. Enteritidis*) with decreased susceptibility to ciprofloxacin (DSC) has emerged in the United States, driven by NCBI SNP Cluster PDS000065758 (denoted SE65758) strains with GyrA(D87Y). In the current study, metadata analysis of 79,476 *S. Enteritidis* genomes worldwide did not identify a similar increase of DSC Enteritidis in non-USA countries. Gene presence/absence differences from pangenomic analysis were not prominent between USA and non-USA SE65758 genomes with GyrA(D87Y), suggesting genetic content differences are not a major contributor to the USA-associated emergence of DSC Enteritidis. Variation between SE65758 genomes with and without GyrA(D87Y) was also not notable, but significant differences were identified between SE65758 and non-SE65758 genomes with GyrA(D87Y), including phage-like regions unique to each. In general, the data suggest the USA specific emergence of SE65758 genomes with GyrA(D87Y) was not solely driven by GyrA(D87Y) or a specific pangenomic element, warranting additional genomic and phenotypic analyses.

Introduction

Salmonella enterica subspecies *enterica* serovar Enteritidis (*S. Enteritidis*) is one of the most common serovars causing bacterial foodborne illness¹. The U.S. Centers for Disease Control and Prevention (CDC) reports that while the incidence of human salmonellosis in the USA decreased on average from 2016-2018 (16.9 average cases of infection per 100,000 persons) to 2023 (16.6 cases of infection per 100,000 persons), *S. Enteritidis* was the only serovar of concern that had an increase in incidence over the same time period (2.6 on average from 2016-2018 to 3.1 in 2023)². Of particular concern is the rise in the fraction of *S. Enteritidis* classified as having decreased susceptibility to ciprofloxacin (DSC)^{3,4}. Ciprofloxacin is a critically important broad-spectrum antibiotic of the quinolone class used to treat various bacterial infections and is a first-choice drug for severe salmonellosis infections⁵. One of the main mechanisms of decreased susceptibility associated with DSC Enteritidis is a nucleotide substitution at codon 87 in gyrase (GyrA), resulting in an aspartate (Asp; D) to tyrosine (Tyr; Y) amino acid substitution (henceforth referred to as GyrA(D87Y))³. The rise of *S. Enteritidis* genomes with GyrA(D87Y) is highly associated with the PDS000065758 NCBI single nucleotide polymorphism (SNP) Cluster (denoted as SE65758) within the USA (Ford et al. 2026, in review).

S. Enteritidis is a globally distributed serovar, driven by centralized poultry breeding and international trade^{6,7}. It is recognized as a worldwide food safety issue as evidenced by recent studies exploring the global or country-specific population structure and genomic content of *S. Enteritidis* strains (e.g. South Korea, United Kingdom, China, Poland, Singapore, Greece, and Brazil)^{6,8-15}. While a rise in DSC Enteritidis has been reported in the USA, reports of a similar rise outside the USA are limited and have not been explored within the context of association

with specific NCBI SNP Clusters, including SE65758^{6,9,15}. Given the global connectivity of *S. Enteritidis* lineages, we investigated the emergence of DSC *Enteritidis* in the USA, to determine if parallel increases in this resistance profile are occurring in worldwide *S. Enteritidis* populations. A worldwide genomic investigation was conducted by comparing (1) the presence and absence of GyrA(D87Y) across (2) USA and non-USA isolation locations in relation to (3) the inclusion or exclusion of a genome in SE65758. Comparisons were made across genome categories based on prophage-like regions (PLR), plasmids, and pangenomic gene content to determine if genomic differences exist that could contribute to increased fitness, host colonization, or environmental persistence of DSC *Enteritidis* that has emerged as a *Salmonella* strain of concern in the USA.

Results and Discussion

Metadata distributions of Genome Categories: USA/non-USA Isolation Source, GyrA(D87Y) Presence/Absence, and SE65758/non-SE65758 NCBI SNP Cluster

The global distribution and genetic lineages of *S. Enteritidis* has been broadly explored in the literature, but investigation of specific genetic content associated with *S. Enteritidis* genomes with GyrA(D87Y), which confers DSC, is minimal^{6,9,14,15}. To address this knowledge gap, publicly-available *S. Enteritidis* genomes from the NCBI Pathogen Detection Database (PDD) were filtered based on known metadata, in-silico serotyping, and assembly quality control, thereby generating a collection of 195 complete genomes and 79,476 draft genomes; four in-house complete genomes were also included (Figure 1, Supplementary Data 1). Supplementary Data 2 shows the breakdown of various datasets (genomes, plasmid replicons, etc) comparing the three categories of interest (collectively referred to hereafter as “Genome Categories”): 1) USA/non-USA Isolation Source, 2) GyrA(D87Y) Presence/Absence, 3) and SE65758/non-

SE65758 NCBI SNP Cluster. Analysis of two datasets (complete genomes and draft genomes) was performed with a focus on the genetic content of SE65758 genomes with GyrA(D87Y), referred to as SE65758-D87Y genomes, in relation to other Genome Categories.

The complete genomes dataset contained 44 USA genomes (40 PDD and 4 in-house) and 155 non-USA genomes from 17 countries and 6 continents (Supplementary Data 1). Six were SE65758 genomes, including: two in-house USA-based genomes with GyrA(D87Y), one in-house USA-based genome without GyrA(D87Y), one PDD Asia-based genome with GyrA(D87Y), and two PDD Asia-based genomes without GyrA(D87Y). Thirty-one non-SE65758 complete genomes with GyrA(D87Y) were distributed across five SNP Clusters and were all Asia-based. Non-SE65758 complete genomes without GyrA(D87Y) (n=162) were distributed across 48 SNP Clusters. Eight USA and 34 non-USA complete genomes contained codon substitutions for GyrA(D87N), while only non-USA genomes had GyrA(S83F) (n=2) and GyrA(D87G) (n=12) substitutions. No complete genome had multiple codon substitutions for GyrA. The lack of non-SE65758 complete genomes with GyrA(D87Y) and dearth of SE65758 genomes from the USA in the PDD highlights the need for submission of such genomes.

The draft genomes dataset was comprised of 50,719 USA genomes and 28,757 non-USA genomes from 76 countries across 5 continents (Supplementary Data 1). The number of USA genomes with the GyrA(D87Y) substitution was 11,357, with 10,772 of these from SE65758 (21.2% of the 50,719 USA assemblies). Other GyrA codon substitutions in the USA dataset included: D87N (n=2,615), S83Y (n=455), D87G (n=318), S83F (n=68), S83F/D87N (n=7), G81D (n=3), S83Y/D87N (n=3), A119E (n=1), D82N (n=1), and G81C (n=1) (Supplementary Data 1). USA genomes were distributed across 561 SNP Clusters besides SE65758. For non-USA genomes, the GyrA(D87Y) substitution (n=3,037) was distributed across 636 SNP clusters

including 402 genomes within SE65758 (1.4% of the 28,757 non-USA genomes). Non-USA genomes had six other GyrA codon substitutions: S83Y (n=2,005), D87N (n=1,736), D87G (n=740), S83F (n=221), A119E (n=1), and D82N (n=1). The number of assemblies with wildtype GyrA (no codon substitutions) was 35,890 (70.8%) and 21,016 (73.1%) from the USA and non-USA genomes, respectively.

For source attribution, the complete genome dataset contained 104 human-associated (3 in-house), 46 chicken-associated, 36 other source/host, eight other/unspecified clinical, two egg-associated, two swine/pork-associated, and one turkey-associated (1 in-house) genomes (Supplementary Data 1). Human-associated genomes also made up the bulk of the draft genome dataset (67,154; 84.5%), followed by chicken-associated (7,997; 10.1%), other source/host (3,331; 4.2%), other/unspecified clinical (578; 0.73%), egg-associated (147; 0.18%), cow/beef-associated (103; 0.13%), swine/pork-associated (98; 0.12%), and turkey-associated (68; 0.09%).

The prevalence of GyrA(D87Y) increased over time in USA genomes compared to non-USA genomes

Distribution of genomes across USA/non-USA Isolation Locations and GyrA(D87Y) Presence/Absence were queried to assess their association with NCBI SNP Clusters. The availability of *S. Enteritidis* genomes in the USA increased from 2013 (n=331) to 2014 (n=3,217) in the NCBI PDD database (likely due to an increase in sequencing efforts) with the first occurrence of GyrA(D87Y) in 2001 (Figure 2a). The prevalence of GyrA(D87Y) substitutions in USA genomes was low (<6%) until 2018 when the occurrence increased to 12.1% and reached 71.2% in 2025 (Figure 2b). GyrA(D87Y) substitutions in non-USA *S. Enteritidis* genomes were observed since 2004, ranging from 3 to 24.1% for the last two decades (Figure 2c and 2d). The overall decline in the number of genomes in 2024 and 2025 is presumably due to delays in

sequence data availability at the time of the metadata download and probably does not reflect an actual decrease in *S. Enteritidis* isolates (Figures 2a and 2c). An increase in the number of genomes with GyrA(D87Y) was noted in chicken-associated genomes, but the total number of chicken-associated genomes did not increase over time (i.e. replacement, not additive) (Supplementary Figure 1). This replacement pattern was not observed in non-USA genomes (Supplementary Figure 2).

An increase in the presence of GyrA(D87Y) corresponded to an increase in SE65758 in the USA genomes, but not in non-USA genomes

In 2017, >50% of USA genomes with GyrA(D87Y) clustered in SE65758 and the percentage has increased yearly to 99.8% in 2025 (Figure 3a). While SE65758 became the dominant SNP cluster containing genomes with GyrA(D87Y) substitutions in the USA, a similar increase in genomes with GyrA(D87Y) within SE65758 was not observed in non-USA genomes (Figure 3c). Furthermore, non-USA genomes containing the GyrA(D87Y) substitution represented multiple SNP clusters. SE65758 genomes without GyrA(D87Y) were generally found at higher percentages in non-USA genomes (Figure 3d) than in USA genomes (Figure 3b). Figures 3e and f summarize the higher percentage of genomes with GyrA(D87Y) from SE65758 in the USA than in non-USA locations, and the data is further broken down by non-USA countries (UK and Canada) and continents in Supplementary Figure 3. The distribution of SE65758 genomes with and without GyrA(D87Y) over time in USA and non-USA countries is further delineated in Supplementary Figure 4. To summarize, differential patterns exist between the distribution of USA and non-USA genomes based on GyrA(D87Y) Presence/Absence and SE65758/non-65758 SNP Clusters. These differential patterns, particularly the emergence of SE65758-D87Y in the USA, warrant a deeper investigation into their genomic content for unique genetic elements that

may enhance adaptive/advantageous phenotypic features, such as host colonization or environmental persistence, contributing to the emergence of SE65758-D87Y in the USA.

Distribution of plasmid replicons were unrelated to Genome Categories

Plasmids are mobile genetic elements that can contribute to increased antibiotic resistance, virulence, colonization, and persistence in *Salmonella* and have been implicated as contributing factors in foodborne outbreaks¹⁶. Plasmids within the genome datasets were assessed in two ways: (1) plasmid replicon distributions within the full dataset across Genome Categories and (2) plasmid-associated genes with differential representation across Genome Categories in the representative draft genomes dataset. First, PlasmidFinder identified 74 types of plasmid replicons across 79,476 draft genomes with 1,054 different combinations of plasmid replicons present within the genomes (Supplementary Data 2). IncXI was present in at least 25% of all non-SE65758 categories but less than 5% of all SE65758 categories, corresponding with the findings from a previous study focusing on DSC Enteritidis¹⁷. IncFIB(S) and IncFIS(S) were present in at least 88% of all genomes across all eight Genome Categories. Broadly, the distribution of plasmid replicons was generally unrelated to Genome Categories and do not implicate plasmid replicon-association with the increase in SE65758-D87Y genomes. In the second plasmid assessment method, specific plasmid-associated genes that were statistically different between Genome Categories are described in the pangenomic analysis below.

Distribution of five phage-like regions associated with Genome Categories

Bacteriophages that integrate into host bacterial genomes can provide novel functions that increase bacterial fitness, protect the bacteria from lytic phages, and mediate horizontal gene transfer, thereby shaping the genetic evolution of bacteria¹⁸. Identification of prophages in bacterial genomes is an evolving field, and various studies have described the capacity of

different tools to recognize prophage regions as inconsistent^{19,20}. This study focused on identifying differences in established genetic elements of prophage-associated regions between Genome Categories rather than the prospective effects of the mobile genetic elements. Because studies have shown that phage identification tools are more accurate on complete genomes, prophage-like-regions were first identified in complete genomes (n=682 PLRs), clustered into viral operational taxonomic units (vOTUs) (n=67), and used to query the representative draft genomes dataset²¹.

BLASTn determined that 52 out of 67 vOTUs identified in the complete genomes dataset were present in the representative draft genomes dataset (n=9,354) (i.e. passed Minimum Information about an Uncultivated Virus Genome (MIUViG) thresholds of >95% sequence identity and 85% coverage), indicating that not all PLRs identified in the complete genomes were present in the representative draft genome dataset. According to Panaroo gene designations, the representative draft genomes had one core ($\geq 99\%$ of genomes), no soft core ($< 99\%$ and $\geq 95\%$), thirteen shell ($< 95\%$ and $\geq 15\%$), and 38 cloud vOTUs ($< 15\%$) (Supplementary Data 3). Five high-quality shell vOTUs had distributions strongly related to Genome Categories. For instance, CP087495.1_pp_3 (42,124 bp) was only found in 5 out of 1,812 (0.28%) SE65758 genomes but in 387/407 (95.1%) non-SE65758 genomes with GyrA(D87Y). This vOTU was found within the CP136141.1_pp_4 vOTU (78,310 bp), per reciprocal vOTU alignments, which had a similar but less prominent pattern; it was present in three SE65758 genomes, and 267 (65.6%) non-SE65758 genomes with GyrA(D87Y). The two vOTUs do not have sequence similarity to the ESSI-2 phage which was previously identified as not associated with DSC Enteritidis¹⁷. Instead, ESSI-2 is most aligned (100% sequence identity; 94% sequence coverage)

with the CP007247.1_pp_4 PLR which was found in non-USA (n=53) or USA (n=689) non-SE65758 representative draft genomes without GyrA(D87Y), substantiating previous results.

Three vOTUs (CP160171.1_pp_1, CP173222.1_pp_3, CP007249.2_pp_3) had an opposite pattern: each were identified in 85.9% or more of SE65758 genomes with and without GyrA(D87Y) (1,556/1,812) but only two each in the non-SE65758 genomes with GyrA(D87Y) (2/407, 0.49%). CP173222.1_pp_3 (33,465 bp) was found within both CP160171.1_pp_1 (71,069 bp) and CP007249.2_pp_3 (53,702 bp). CP007249.2_pp_3 was also found within CP160171.1_pp_1. To summarize, while the MIUViG thresholds identified five vOTUs of interest, further refinement suggested the presence of two distinct phage regions of interest, one more highly represented in SE65758 genomes with or without GyrA(D87Y) and the other in non-SE65758 genomes with GyrA(D87Y). These results prompted further pangenome investigation of specific vOTU-associated genes with statistical differences between Genome Categories.

Gene distributions from complete and representative draft genomes were more related to SNP Cluster than to Isolation Location or GyrA(D87Y) status

Pangenome analysis was used to gain an overall understanding of the distribution of genes across genomes as well as compare gene distribution patterns across categories of interest. Panaroo pangenome analysis identified 14,038 genes across representative draft genomes and all complete genomes (n=9,553 total) with 4,071 core ($\geq 99\%$), 190 soft core ($< 99\%$ and $\geq 95\%$), 212 shell ($< 95\%$ and $\geq 15\%$), and 9,565 cloud genes ($< 15\%$) (Supplementary Data 4).

Permutational analysis of variance (PERMANOVA) showed that genome content was significantly different across all Genome Category groups (n=8) according to the presence/absence of genes ($R^2 = 0.07418$, Pseudo F-Score = 109.25, p-value = 0.0001). Pairwise

PERMANOVA analysis results across all possible Genome Category comparisons (n=28) are presented in Supplementary Table 1. Pairwise PERMANOVA results on Genome Category comparisons of interest (n=7) revealed that SE65758 genomes with GyrA(D87Y) were not significantly different from SE65758 genomes without GyrA(D87Y) within both USA and non-USA isolation locations (Table 1). USA SE65758-D87Y genomes were also not significantly different from the non-USA SE65758-D87Y genomes. Non-USA non-SE65758 genomes with or without GyrA(D87Y) were significantly different from Non-USA SE65758 genomes with GyrA(D87Y). ; Similarly, USA non-SE65758 genomes with GyrA(D87Y) were significantly different from USA SE65758 genomes with GyrA(D87Y), though the p-value for the USA SE65758 genomes with GyrA(D87Y) versus the USA non-SE65758 without GyrA(D87Y) comparison could not be calculated. Principal coordinates analysis (PCoA) showed the same within Isolation Location (USA or non-USA) results: SE65758 genomes clustered together regardless of the GyrA variant and clustered away from non-SE65758 genomes with GyrA(D87Y) (Figure 4). Non-SE65758 genomes without GyrA(D87Y) had a wide distribution reflecting a high dispersion of genes within the group and thus were not statistically assessed.

To identify genes that were statistically higher in one Genome Category versus another, PySEER was used to analyze five different comparisons. The non-SE65758 genomes without GyrA(D87Y) were not considered in any PySEER comparison, due to their wide genomic variations as established by PERMANOVA and PCoA analysis. A summary of the number of variants (genes) tested and found to be significantly different via filter or likelihood ratio test (LRT) p-values (unadjusted or adjusted) across all five comparisons is found in Table 2 and the results associated with all genes tested by PySEER is found in Supplementary Data 5.

In Table 2, comparison between USA versus non-USA SE65758-D87Y genomes identified two genes encoding hypothetical proteins (group_3228 and group_1792) statistically higher (LRT p-value) in non-USA SE65758-D87Y genomes (77.1% and 14.6%, respectively) compared to USA SE65758-D87Y genomes (51.2% and 1.4%, respectively) (Comparison 1) (Supplementary Data 6). Additionally, *intR* (prophage integrase), a vOTU-associated gene, was statistically higher (filter p-value) in USA SE65758-D87Y genomes (99.2%) than in non-USA SE65758-D87Y genomes (91.7%). Comparison 2 between USA SE65758 genomes with versus without GyrA(D87Y) identified two genes encoding hypothetical proteins (group_3118 and group_3843) both associated with vOTU CP136141.1_pp_4, that were significantly higher (LRT p-value) in USA SE65758 genomes without GyrA(D87Y) (97.1% and 98.5%, respectively) versus those with GyrA(D87Y) (86.9% and 86.7%, respectively). Also, a plasmid-associated gene (*dsbA*; a disulfide interchange protein) was statistically higher (filter p-value) in genomes with GyrA(D87Y) (97.6%) than genomes without GyrA(D87Y) (89.7%). In Comparison 3, no genes were statistically different between non-USA SE65758 genomes with versus without GyrA((D87Y).

Genetic differences between SE65758 and non-SE65758 genomes with GyrA(D87Y) were driven by vOTU-associated genes

The number of statistically different genes between SE65758 and non-SE65758 genomes with GyrA(D87Y) within USA and non-USA countries were 173 (Comparison 4) and 157 (Comparison 5), respectively via LRT-p-values (Supplementary Data 6). One-hundred and thirty-three of these genes were found in both USA and non-USA comparisons. All 24 genes unique to the non-USA comparison between SE65758 and non-SE65758 genomes with GyrA(D87Y) (Comparison 5) were plasmid-associated genes exclusively found in the non-SE65758 genomes,

but none occurred in more than 9% of those genomes; 17 of these genes were also vOTU-associated. Similarly, 21 of the 40 genes exclusive to the USA comparison (Comparison 4) were plasmid-associated genes that occurred in less than 1% of the SE65758 genomes but did not occur in more than 18.75% of the non-SE65758 genomes. One USA-exclusive comparison plasmid-associated gene (group_6304 – hypothetical protein) was present at higher percentages, but in both types of genomes with GyrA(D87Y) (Comparison 4): 99.4% of SE65758 genomes and 91.25% of non-SE65758 genomes. An unnamed replication initiator protein (group_5959) was exclusively found in USA SE65758 genomes (17.5%) but was neither PLR- nor plasmid-associated. While the 29 plasmid-associated genes that were significantly different in both USA and non-USA comparisons (Comparisons 4 and 5) were identified in less than 3% of SE65758 genomes, none were present in greater than 50% of the non-SE65758 genomes. In general, the PySEER results at the gene level supported the aforementioned plasmid replicon results, and while plasmid-associated genes were statistically significant, the magnitude of variation between the genome categories was underwhelming. Thus, it is unlikely that a plasmid provided a fitness advantage that would explain the emergence of *S. Enteritidis* strains within SE65758-D87Y.

The remaining genes (n=17 out of 40) exclusive to the USA SE65758 versus non-SE65758 comparison (Comparison 4) were vOTU-associated genes found only in the non-SE65758 genomes, but none occurred in more than 24% of those genomes. The vOTU-associated gene *pefI* (encoding the transcriptional regulator PefI) was found in 95% of USA and 99.2% of non-USA non-SE65758 genomes, while only present in 28.8% of USA and 12.5% of non-USA SE65758 genomes (Comparisons 4 and 5). PefI was not associated with a previously described Genome Category-associated vOTU.

Ninety-six vOTU-associated genes identified in both of the SE65758 versus non-SE65758 genomes with GyrA(D87Y) comparisons (Comparisons 4 (USA) and 5 (non-USA)) were associated with the previously described vOTUs. Fifty-five genes were observed in a higher percentage of non-SE65758 genomes (USA: 92.5% to 97.5%; non-USA: 98.3% to 99.2%); while none of the genes were identified within the USA SE65758 genomes with GyrA(D87Y), they were present in 2.1% of the non-USA SE65758 genomes with GyrA(D87Y). Fifty-three of the 55 genes were found in both non-SE65758-associated vOTUs (CP087495.1_pp_3 and CP136141.1_pp_4). Gene products were mostly associated with prophage maintenance including unnamed lysis and structural tail or head proteins, but also with functions including a nuclease (*yajD*), transcriptional regulators (*perC*; *ymfL*), a crossover junction endodeoxyribonuclease (*rusA*), a lipoprotein (*rzoQ*), a molecular chaperon (*dnaJ*), an antitermination protein (*quuQ*) and a recombination protein (*rdgC*). The remaining two genes, a gene encoding a hypothetical protein (group_4497) and the type III secretion system effector arginine glycosyltransferase (*sseK3* also referred to as *nleB* or *sseK*), were exclusively found within the CP136141.1_pp_4 non-SE65758-associated vOTU.

Conversely, 40 of the vOTU-associated genes were significantly higher in SE65758-D87Y genomes and found in the three aforementioned SE65758-associated vOTUs (CP160171.1_pp_1, CP173222.1_pp_3, CP007249.2_pp_3); one additional gene encoding a RES domain-containing protein (group_2654) was only present in the CP160171.1_pp_1 and CP173222.1_pp_3 vOTUs. While all genes were present in greater than 99% of SE65758-D87Y genomes, their presence varied in the non-SE65758 genomes with GyrA(D87Y) (USA: 2.5% to 27.5%; non-USA: 0.3% to 10.9%). Most of these genes were similarly related to phage maintenance (*gpD*), phage structure (*gpX* and *gpV*), phage lysis (*lysB*), phage

recombination/integration (*xerD*), lysozyme activity (*rrrD*), phage portals (*yjcR*), but also included a gene encoding a levan regulator (*rlsC*).

Seven genes present in both USA and non-USA comparisons were not associated with plasmids or vOTUs. Six genes encoding two hypothetical proteins (group_1163 and group_4170), three restriction endonucleases (group_1811, group_3054, and group_4131), and one DNA methyltransferase (group_2623) were significantly higher in SE65758-D87Y genomes (USA: 91.3% to 100%; non-USA: 100%) than non-SE65758 genomes with GyrA(D87Y) (USA: 51.3% to 56.3%; non-USA: 63.4% to 68.2%). A host cell division inhibitor Icd-like protein (group_4393) was significantly higher in non-SE65758 genomes with GyrA(D87Y) (USA: 81.3%; non-USA: 87.4%) than SE65758-D87Y genomes (USA: 2.3%; non-USA: 6.25%).

In summary, the emergence of *S. Enteritidis* genomes with GyrA(D87Y) substitutions in the USA was associated with a concurrent temporal rise in an NCBI SNP Cluster of interest (PDS000065758 denoted as SE65758). However, metadata investigations did not reveal a similar escalation of SE65758-D87Y strains in non-USA countries over the same time period.

Pangenomic analysis to evaluate genetic differences between Genome Categories of interest (USA/non-USA Isolation Source, GyrA(D87Y) Presence/Absence, and SE65758/non-SE65758 NCBI SNP Cluster) revealed little to no statistical differences between USA and non-USA SE65758-D87Y genomes (Comparison 1), suggesting that genetic content differences may not be a major driver behind the prominence of SE65758-D87Y strains in the USA. Furthermore, emergence of DSC *Enteritidis* cannot solely be accounted for by the presence of GyrA(D87Y); otherwise, the rise of DSC *Enteritidis* would not be so strongly associated with a specific NCBI SNP Cluster (SE65758). Likewise, DSC *Enteritidis* emergence is not simply explained by the SE65758 SNP cluster; otherwise, SE65758 genomes with or without GyrA(D87Y) would be

increasing. Genetic differences associated with SE65758-D87Y were queried by controlling for each aspect of its Genomic Category: inclusion in SE65758 (Comparisons 2 and 3) and presence of GyrA(D87Y) (Comparisons 4 and 5). Differences in genetic content (overall and at the gene level) were minimal between SE65758 genomes with and without GyrA(D87Y) regardless of USA/non-USA Isolation Location (Comparisons 2 and 3). Genetic differences between SE65758 genomes and non-SE65758 genomes with GyrA(D87Y) (USA and non-USA) were evident (Comparison 4 and 5) and frequently associated with prophage-like regions containing prophage maintenance and structural genes. Future investigation of these phage-like regions via phenotypic analyses is warranted, and further investigation of other genome variations such as single nucleotide polymorphisms (SNPs), insertions, and deletions may identify additional genetic differences between Genome Categories. Identification of genetic variations in *Salmonella* isolates of concern informs future research exploration for potential fitness advantages that may contribute to their emergence in poultry production and human health.

Materials and Methods

Data Management and Statistical Analysis

RStudio (v2024.12.1+; R v4.5.1) was used for data processing, graphic generation, and statistical analysis^{22,23}. The RStudio environment was loaded with the following main libraries: tidyverse (v. 2.0.0), ggpubr (v0.6.3), ggtext (v0.1.2), ggnewscale (v0.5.2), lubridate (v. 1.9.3), scales (v.1.4.0), parallel (v. 4.4.1), and doParallel (v. 1.0.17), foreach (v. 1.5.2), vegan (v2.7-1), pairwiseAdonis (v0.4.1), phyloseq (v1.52.0), igraph (v2.1.4)²⁴⁻³⁶. Command line scripts, R scripts, and R Studio session information including additional supportive libraries used and their versions are available on GitHub (https://github.com/USDA-FSEPRU/worldwide_dsc_enteritidis.git). Figure 1 depicts the overall workflow for this study.

Dataset Selection and Filtering

The metadata for all *Salmonella enterica* genomes from the NCBI Pathogen Detection Database (PDD) that were identified as *S. Enteritidis* within the Computed types, Serovar, and Scientific Name columns were downloaded on 11/14/25³⁷. The downloaded dataset had 319 complete and 145,434 draft assemblies. When filtered for known collection year, known isolation location, defined NCBI SNP Cluster, known isolation source, and defined NCBI assembly and SRR accession numbers (draft assemblies only), the dataset was reduced to 204 complete and 80,017 draft assemblies. Assemblies were downloaded from NCBI (ncbi-genome-download v0.3.3), underwent *in silico* serovar confirmation (2/3 of SeqSero2 using the assembly kmer-based method (v1.2.1), SISTR (v.1.1.1), and Mash (v.2.3)) and were quality checked for completeness ($\geq 95\%$) and contamination ($\leq 5\%$) with CheckM (v. 1.2.2) resulting in 195 complete and 79,476 draft PDD assemblies used for further analysis³⁸⁻⁴². Four in-house complete assemblies were also added to the dataset resulting in a total of 199 complete genomes in the dataset⁴³. All assemblies were annotated with Bakta (v.1.11.0)⁴⁴. Plasmid replicons were identified with PlasmidFinder v2.1.6⁴⁵. Contig lengths of all complete genomes (n=199) were determined with seqkit (v2.6.1)⁴⁶.

Metadata Analysis

An additional metadata column (Isolation Source (Summarized)) that classified isolation source information into general categories was generated based upon the keyword searches querying the data contained within the Isolation source, Host, Isolation type, Source type, and IFSAC category columns from the NCBI PDD metadata in a defined order (Supplementary Data 7). Additionally, the isolation location was summarized into continents or the countries with the top three number of genomes (USA, United Kingdom, and Canada) (Isolation

Location(Summarize)) (Supplementary Data 7). The presence of a GyrA amino acid substitution in an assembly was based on querying the “AMR.genotypes” information for each assembly for GyrA substitutions present in PointFinder (A67P, D72G, V73I, G81C, G81S, G81H, G81D, D82G, D82N, S83Y, S83F, S83A, D87N, D87G, D87Y, D87K, L98V, A119S, A119E, A119V, A131G, E139A)⁴⁷. Most statistical analysis was based on comparing genomes across three main metadata categories that were generated based on the presence or absence of the GyrA D87Y substitution (GyrA(D87Y) Presence/Absence), whether the genome was within the PDS000065758 NCBI SNP cluster or not (denoted as SE65758/non-SE65758 NCBI SNP Cluster), and whether the isolation location was within the USA or not (USA/non-USA Isolation Location). Collectively these three categories will be known as “Genome Category”.

Phage-like Region Analysis of Complete Genomes

PLRs in complete genomes were identified with PhiSpy (v5.0.6)^{19,48}. Their completeness, contamination, and identities as prophages were determined with CheckV (v1.0.3)⁴⁹. PLRs were clustered into representative prophage-like regions aka viral operational taxonomic units (vOTUs) with MMSeqs2 (v18.8cc5c) according to the Minimum Information about an Uncultivated Virus Genome (MIUViG) thresholds for phage clustering ($\geq 95\%$ nucleotide identity over 85% coverage) (MMSeqs2 parameters: --min-seq-id 0.95 -c 0.85 --cov-mode 0)⁵⁰⁻⁵². These thresholds were based on a comparison of sequences that were available at the time of their analysis in NCBI RefSeq10 and IMG/VR. The coverage mode was reciprocal meaning that both the query PLR and the subject PLR had to have $> 85\%$ coverage to one another instead of just one to the other (i.e. just the query PLR or the subject vOTU). PhiSpy identified 682 PLRs across 199 complete genomes that were clustered into 67 vOTUs (Supplementary Data 8). CheckV identified two of the vOTUs as complete, 35 as high-quality, 8 as medium-quality, 17 as

low-quality, and two with undetermined quality. vOTUs were aligned to one another with BLASTn (v2.14.1) to determine if any vOTU was a subsequence of another, though only subject coverage was assessed instead of reciprocal coverage⁵³. Nineteen vOTUs had BLAST alignments to other vOTUs that passed MIUViG thresholds, suggesting they may have been shorter versions of other vOTUs (Supplementary Data 9).

Draft Genome Accessory Genome Cluster

Representative genomes were selected from the main dataset to ease computational management and visualize isolate genomic interrelatedness; as a part of this process, pangenome analysis was performed to identify accessory genome genes that could be used in clustering. All draft genomes were run in batches in Panaroo (v.1.6.0) and merged to create a pangenome (panaroo-merge)⁵⁴. Panaroo pangenome analysis identified 19,002 genes across the 79,476 draft genomes with 4,043 core genes (present in $\geq 99\%$ of draft genomes), 94 soft core genes ($< 99\%$ and $\geq 95\%$), 224 shell genes ($< 95\%$ and $\geq 15\%$), and 14,641 cloud genes ($< 15\%$) (Supplementary Data 10). The number of core genes found in all genomes was 3,776, while the number of cloud genes found in one genome (singletons) was 2,243. The presence and absence matrix of non-singleton cloud and shell accessory genes (n=12,622) was used to create a co-occurrence matrix which was used to generate a weighed graph (igraph v 2.1.4)^{35,36}. In this graph, genomes represented the nodes and the number of shared accessory genes between genomes represented the weighted edges. The weighted graph was used as an input into Leiden clustering to determine accessory genome clusters⁵⁵. Analysis had to be adjusted to account for tool limitations which includes some R libraries being unable to work with datasets with more than 2^{31} values (2,147,483,647). To accomplish this, principles from the Jaccard Index Network Analysis (JINA) tool were applied to choose the best edge threshold to use for filtering⁵⁶. Edge filtering allows for

removal of potentially spurious or weakly connected assemblies. This dataset had 79,476 genomes; thus, the max number of possible connections (edges) was 3,158,217,288. The minimum edge filtering value that reduced the number of edges with non-zero values to less than the 2^{31} limit was 97. Edge filtering values from 97 to the third quartile of edges (167) were trialed to generate clusters and evaluate them based on two JINA-based clustering filters: the number of genomes not in large clusters (clusters with >4 genomes) should be less than 10% of the total dataset (7,947.6) and the number of large clusters should be less than the natural log of the total number of genomes in the dataset ($\ln(79,476) = 11.28$). Forty-one weighted edge thresholds (97-138 overlapping genes) passed the number of large clusters filter; only 35 (97-131 overlapping genes) of these passed the number of non-large cluster genomes filter (Supplementary Data 11).

In order to balance clustering the most genomes possible into large clusters and cluster quality, the 35 filter-passing edge value thresholds were ranked (1 being the highest and thus best value) based on the Leiden cluster quality value and the ratio of genomes in large clusters to those in small clusters (<5 genomes) and singletons (i.e. genomes without any edges to other genomes). The cluster quality had a generally positive trend over increasing edge filter thresholds while the large versus non-large cluster genomes ratio had a negative trend. Per JINA principles, both values should be maximized when choosing an edge filtering threshold, thus the edge threshold with the smallest summed rank, and thus the best balance of the ratio and quality, was selected for further analysis. The edge threshold with the lowest rank was a filtering threshold of 107. The summed rank of this edge threshold was 34 with the ratio (30.79) ranked 11th out of 35 qualifying edge filters, while the cluster quality (0.23) ranked 23rd. Filtering the edges with this threshold led to 2,495 singletons (3.14% of all genomes), 5 genomes (0.0063%)

assigned to two small clusters, and 76,976 (96.85%) genomes distributed across five large clusters. The distribution of genomes in these clusters across Genome Categories is visualized in Supplementary Figure 5 and detailed in Supplementary Data 2.

Representative Genome Selection

Genomes were grouped by relevant metadata categories: 1) summarized isolation source, 2) summarized isolation location, 3) NCBI SNP cluster, 4) GyrA substitution or wildtype designation, and 5) accessory genome Leiden cluster (from Panaroo analysis)⁵⁵. This led to 76,976 genomes from large accessory genome clusters being distributed across 2,374 groups; the classification of genomes as representative or not and identification of what group they were in is found in Supplementary Data 1. Ten percent of the genomes within each group (rounded up to the nearest whole number, or 1 if there were less than 10 genomes) were randomly selected to generate a representative genomes dataset of 9,354 draft genomes that was used for pangenome and phage-like region analysis of draft genomes. The distribution of representative draft genomes across Genome Categories is noted in Supplementary Data 2.

Phage-like Region Analysis of Representative Draft Genomes

All representative draft assemblies (subject) were queried for the presence of the complete genome vOTUs (n=67) (query) with BLASTn (v. 2.14.1)⁵³. BLASTn parameters were run with default settings besides max target sequences set to 5, identity set to 95%, and HSP (High Scoring Pair) query coverage set to 85% to match MIUViG thresholds⁵⁰. The use of HSP coverage ensured that the calculated coverages were on a per alignment region basis instead of the total of all alignments to an assembly (query coverage).

Pangenome Analysis of Representative Draft and Complete Genomes

All PDD assemblies from representative draft genomes (n=9,354) and complete genomes (n=195) as well as the four in-house complete genomes were used in pangenome analysis (Panaroo). A gene presence/absence-based Jaccard distance index was used in permutational analysis of variance (PERMANOVA) statistical analysis and visualizing the distribution of genomes with PCoA analysis with phyloseq R package³⁴.

PySEER (v1.4.1) was used to conduct a genome-wide association study (GWAS) based on a MASH-based distance matrix (v2.3) generated with full genomes (--distances; --mds metric) to take into account population structure and the presence/absence of genes (--pres) across various pairs of strain types^{57,58}. The GWAS was run on five comparisons (--phenotypes): (1) USA versus non-USA SE65758 genomes with GyrA(D87Y), (2) USA SE65758 genomes with and without GyrA(D87Y), (3) non-USA SE65758 genomes with and without GyrA(D87Y), (4) USA SE65758 genomes with GyrA(D87Y) versus nonSE65758 genomes with GyrA(D87Y), and (5) non-USA SE65758 genomes with GyrA(D87Y) versus nonSE65758 genomes with GyrA(D87Y). By default, PySEER will only test variants, genes in this case, with allele frequencies (af) or presences greater than 1% (--min-af) and less than 99% (--max-af) of the genomes in the comparison. The minimum allele frequency was modified if 1% of genomes in comparison lead to variants present in 4 or less genomes passing allele frequency filters per developer recommendations. The adjusted p-value was calculated by dividing 0.05 by the number of patterns tested (i.e. Bonferroni multi-testing adjustment). Filtered p-value refers to the initial p-value calculated by PySEER using its chi² test-based algorithm while the likelihood ratio test (LRT) p-value was calculated by considering population structure to account for confounding effects due to relatedness between samples. Variants that passed the adjusted-LRT p-value were further evaluated. To integrate multiple data types to gain a better understanding of

potential genomic origins of genes, complete genome annotations were used to label pangenome genes tested by PySEER as plasmid-associated (Supplementary Data 12) and vOTU-associated (Supplementary Data 13). The former was accomplished by linking pangenome genes to specific plasmid genes found in the complete genome dataset, which had 246 plasmids distributed across 141 complete genomes ranging from one to four plasmids per genome (Supplementary Data 1). The latter was accomplished by using the previously mentioned PhiSpy PLR annotations. These labels do not guarantee a pangenome gene was plasmid- or vOTU-associated, just that the pangenome gene was associated with one of these gene types in a complete genome. Gene clusters without an annotated gene name were given “group_XXX” names by PySEER.

Data Availability

The GenBank assemblies (Supplementary Data 1) used in this study can be found in the United States National Center for Biotechnology Information (NCBI) Assembly Database (<https://www.ncbi.nlm.nih.gov/home/genomes/>). The metadata that supports these assemblies can be found at the NCBI Pathogen Detection Database (<https://www.ncbi.nlm.nih.gov/pathogens/>). Other datasets and metadata generated during the current study not present in the Supplementary Data are available from the corresponding authors on reasonable request.

Code Availability

The underlying code for this study is available in GitHub and can be accessed via this link (https://github.com/USDA-FSEPRU/worldwide_dsc_enteritidis.git).

Acknowledgements

The authors are greatly appreciative of Samuel Whelan for his outstanding manuscript editing contributions and Dr. Laura Ford for her general guidance and her epidemiology and genomic expertise.

This research utilized an appointment to the U.S. Department of Agriculture (USDA), Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the USDA. ORISE is managed by ORAU under DOE contract number DE-SC0014664. Funding for open access charge: USDA, ARS.

Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendations or endorsement by the USDA or the Centers for Disease Control and Prevention (CDC). The U.S. Department of

Agriculture is an equal opportunity provider and employer. The findings and conclusions of this article are those of the authors and do not necessarily represent the views of USDA-ARS or CDC.

Author Contributions

D.B. provided bioinformatics expertise, and was involved with project design, data processing, and results analysis and interpretation. K.T. provided bioinformatics expertise and was involved with project design, support for data analysis, and results analysis and interpretation. B.B. provided gene/protein annotation expertise was involved with project design and results analysis and interpretation. H.W. provided bioinformatics expertise and was involved with project design and results analysis and interpretation. S.B. was the project supervisor, provided microbiology expertise, and was involved with project design and results analysis and interpretation. D.B. wrote the main manuscript text. All authors were involved in manuscript writing and revision. All authors have read and approved the manuscript.

Competing Interests

All authors declare no financial or non-financial competing interests.

Funding

This work was supported by the United States Department of Agriculture (USDA), Agricultural Research Service (ARS) appropriated funds from CRIS project 5030-31000-007-000-D.

541 References

- 542 1 Walter, E. J. S. et al. Foodborne Illness Acquired in the United States—Major Pathogens,
543 2019. *Emerg. Infect. Dis.* **31**, 669-677 (2019).
544 <https://doi.org/https://doi.org/10.3201/eid3104.240913>
- 545 2 Shah, H. J. et al. Reported Incidence of Infections Caused by Pathogens Transmitted
546 Commonly Through Food: Impact of Increased Use of Culture-Independent Diagnostic
547 Tests — Foodborne Diseases Active Surveillance Network, 1996–2023. *MMWR Morb*
548 *Mortal Wkly. Rep.* **73**, 584-593 (2024).
549 <https://doi.org/https://doi.org/10.15585/mmwr.mm7326a1>
- 550 3 NARMS. *NARMS Interim Data Updates*, <[https://www.fda.gov/animal-](https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/narms-interim-data-updates)
551 [veterinary/national-antimicrobial-resistance-monitoring-system/narms-interim-data-](https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/narms-interim-data-updates)
552 [updates](https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/narms-interim-data-updates)> (2025).
- 553 4 Wang, Y. et al. A global atlas and drivers of antimicrobial resistance in *Salmonella* during
554 1900-2023. *Nat. Commun.* **16**, 4611 (2025). <https://doi.org/10.1038/s41467-025-59758-3>
- 555 5 Sivanandy, P. et al. A systematic review of recent outbreaks and the efficacy and safety of
556 drugs approved for the treatment of *Salmonella* infections. *IJID Regions* **14**, 100516
557 (2025). <https://doi.org/https://doi.org/10.1016/j.ijregi.2024.100516>
- 558 6 Li, S., He, Y., Mann, D. A. & Deng, X. Global spread of *Salmonella* Enteritidis via
559 centralized sourcing and international trade of poultry breeding stocks. *Nat. Commun.* **12**,
560 5109 (2021). <https://doi.org/10.1038/s41467-021-25319-7>
- 561 7 Ranjan, A. et al. *Salmonella* Infections: Global Trends and Emerging Challenges.
562 *Microorganisms* **14**, 816 (2026). <https://doi.org/10.3390/microorganisms14040816>
- 563 8 Shin, E., La, T. M., Yoo, J., Kim, J. & Hyeon, J. Y. Emergence of Distinct *Salmonella*
564 *enterica* Serovar Enteritidis Lineage since 2020, South Korea. *Emerg. Infect. Dis.* **31**,
565 1386-1393 (2025). <https://doi.org/10.3201/eid3107.250043>
- 566 9 Zheng, X. et al. An MDR *Salmonella* Enteritidis sublineage associated with
567 gastroenteritis outbreaks and invasive disease in China. *J. Infect.* **90**, 106421 (2025).
568 <https://doi.org/https://doi.org/10.1016/j.jinf.2025.106421>
- 569 10 Alikhan, N.-F. et al. Dynamics of *Salmonella enterica* and antimicrobial resistance in the
570 Brazilian poultry industry and global impacts on public health. *PLoS Genet.* **18**,
571 e1010174 (2022). <https://doi.org/10.1371/journal.pgen.1010174>
- 572 11 Polemis, M., Sideroglou, T., Chrysostomou, A. & Mandilara, G. D. First Data on WGS-
573 Based Typing and Antimicrobial Resistance of Human *Salmonella* Enteritidis Isolates in
574 Greece. *Antibiotics* **13**, 708 (2024). <https://doi.org/10.3390/antibiotics13080708>
- 575 12 Aung, K. T. et al. Characterisation of *Salmonella* Enteritidis ST11 and ST1925
576 Associated with Human Intestinal and Extra-Intestinal Infections in Singapore. *Int. J.*
577 *Env. Res. Public Health* **19**, 5671 (2022). <https://doi.org/10.3390/ijerph19095671>
- 578 13 Kamińska, E. et al. Molecular Characteristics of Epidemiologically Successful
579 *Salmonella* Enteritidis in Poland. *Transbound. Emerg. Dis.* **2025**, 5598487 (2025).
580 <https://doi.org/10.1155/tbed/5598487>
- 581 14 Luo, L. et al. Elucidation of global and national genomic epidemiology of *Salmonella*
582 *enterica* serovar Enteritidis through multilevel genome typing. *Microb. Genom.* **7** (2021).
583 <https://doi.org/10.1099/mgen.0.000605>

- 15 Feasey, N. A. et al. Distinct *Salmonella* Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. *Nat. Genet.* **48**, 1211-1217 (2016). <https://doi.org/10.1038/ng.3644>
- 16 Robertson, J., Schonfeld, J., Bessonov, K., Bastedo, P. & Nash, J. H. A global survey of *Salmonella* plasmids and their associations with antimicrobial resistance. *Microb. Genom.* **9**, 001002 (2023). <https://doi.org/10.1099/mgen.0.001002>
- 17 Oladeinde, A. et al. The fitness of *Salmonella* Enteritidis in poultry is strongly shaped by interactions between bacteriophages, plasmids, and the microbial community. *ASM Anim. Microbiol.* **1**, e00008-00026 (2026). <https://doi.org/doi:10.1128/asmam.00008-26>
- 18 Wendling, C. C., Vasse, M. & Wielgoss, S. Phage quest: a beginner's guide to explore viral diversity in the prokaryotic world. *Brief. Bioinform.* **26** (2025). <https://doi.org/10.1093/bib/bbaf449>
- 19 Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126-e126 (2012). <https://doi.org/10.1093/nar/gks406>
- 20 Schackart, K. E., Graham, J. B., Ponsero, A. J. & Hurwitz, B. L. Evaluation of computational phage detection tools for metagenomic datasets. *Front. Microbiol.* **14**, 1078760 (2023). <https://doi.org/10.3389/fmicb.2023.1078760>
- 21 Turner, D., Adriaenssens, E. M., Tolstoy, I. & Kropinski, A. M. Phage Annotation Guide: Guidelines for Assembly and High-Quality Annotation. *Phage (New Rochelle)* **2**, 170-182 (2021). <https://doi.org/10.1089/phage.2021.0013>
- 22 Posit Team. *RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA.* <http://www.posit.co/>. (2024).
- 23 R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, <https://www.r-project.org/> (2024).
- 24 Wickham, H. et al. Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019). <https://doi.org/10.21105/joss.01686>
- 25 Kassambara, A. *ggpubr: 'ggplot2' Based Publication Ready Plots.* <https://CRAN.R-project.org/package=ggpubr>
- 26 Wilke, C. O., & Wiernik, B. M. ggtext: Improved Text Rendering Support for 'ggplot2' v. 0.1.2 (2022). 10.32614/CRAN.package.ggtext
- 27 Campitelli, E. ggnewscale: Multiple Fill and Colour Scales in 'ggplot2' v. 0.5.2 (2025). doi:10.5281/zenodo.254376234
- 28 Grolemund, G. & Wickham, H. Dates and Times Made Easy with lubridate. *J. Stat. Softw.* **40**, 1-25 (2011). <https://doi.org/10.18637/jss.v040.i03>
- 29 Wickham, H., Pedersen, T. L. & Seidel, D. *scales: Scale Functions for Visualization.* <https://scales.r-lib.org>
- 30 Microsoft Corporation & Weston, S. *doParallel: Foreach Parallel Adaptor for the 'parallel' Package.* 10.32614/CRAN.package.doParallel
- 31 Microsoft Corporation & Weston, S. *foreach: Provides Foreach Looping Construct.* <https://github.com/RevolutionAnalytics/foreach>
- 32 Oksanen, J. et al. *vegan: Community Ecology Package.* 10.32614/CRAN.package.vegan
- 33 Martinez Arbizu, P. *pairwiseAdonis: Pairwise Multilevel Comparison using Adonis.*
- 34 McMurdie, P. J. & Holmes, S. phyloseq: An R package for reproducible interactive analysis and graph ics of microbiome census data. *PLoS One* **8**, e61217 (2013). <https://doi.org/10.1371/journal.pone.0061217>

- 35 Antonov, M. et al. Igraph enables fast and robust network analysis across programming languages. *arXiv preprint arXiv:2311.10260* (2023).
<https://doi.org/10.48550/arXiv.2311.10260>
- 36 Csárdi, G. et al. *igraph: Network Analysis and Visualization in R*.
10.32614/CRAN.package.igraph
- 37 The NCBI Pathogen Detection Project, <<https://www.ncbi.nlm.nih.gov/pathogens/>> (2016).
- 38 Blin, K. ncbi-genome-download (Zenodo, 2023). <https://doi.org/10.5281/zenodo.8192486>
- 39 Zhang, S. et al. SeqSero2: Rapid and Improved *Salmonella* Serotype Determination Using Whole-Genome Sequencing Data. *Appl. Environ. Microbiol.* **85**, e01746-01719 (2019). <https://doi.org/10.1128/AEM.01746-19>
- 40 Yoshida, C. E. et al. The *Salmonella* in silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS One* **11**, e0147101 (2016). <https://doi.org/10.1371/journal.pone.0147101>
- 41 Li, W. et al. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res.* **49**, D1020-d1028 (2021).
<https://doi.org/10.1093/nar/gkaa1105>
- 42 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015). <https://doi.org/10.1101/gr.186072.114>
- 43 Bradshaw, D. J. et al. Three complete *Salmonella enterica* serovar Enteritidis genomes from the United States with and without a GyrA D87Y substitution. *Microbiol. Resour. Announc.* **0**, e00130-00126 (2026). <https://doi.org/doi:10.1128/mra.00130-26>
- 44 Schwengers, O. et al. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb. Genom.* **7** (2021).
<https://doi.org/10.1099/mgen.0.000685>
- 45 Carattoli, A. et al. PlasmidFinder and pMLST: in silico detection and typing of plasmids. *Antimicrob. Agents Chemother.* **58**, 3895-3903 (2014). <https://doi.org/10.1128/aac.02412-14>
- 46 Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *iMeta* **3**, e191 (2024). <https://doi.org/10.1002/imt2.191>
- 47 Zankari, E. et al. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J. Antimicrob. Chemother.* **72**, 2764-2768 (2017). <https://doi.org/10.1093/jac/dkx217>
- 48 Akhter S., Aziz R. K., & Edwards R. A. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res.* **40**, e126 (2012). <https://doi.org/10.1093/nar/gks406>
- 49 Nayfach, S. et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578-585 (2021).
<https://doi.org/10.1038/s41587-020-00774-7>
- 50 Roux, S. et al. Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* **37**, 29-37 (2019). <https://doi.org/10.1038/nbt.4306>
- 51 Stopnisek, N., Hedžet, S., Accetto, T. & Rupnik, M. Insights into diversity, host-range, and temporal stability of Bacteroides and Phocaeicola prophages. *BMC Microbiol.* **25**, 92 (2025). <https://doi.org/10.1186/s12866-025-03827-6>

- 52 Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026-1028 (2017).
<https://doi.org/10.1038/nbt.3988>
- 53 Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinf.* **10**, 1-9 (2009).
<https://doi.org/10.1186/1471-2105-10-421>
- 54 Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020). <https://doi.org/10.1186/s13059-020-02090-4>
- 55 Trachsel, J. M., Bearson, B. L., Brunelle, B. W. & Bearson, S. M. D. Relationship and distribution of *Salmonella enterica* serovar I 4,[5],12:i:- strain sequences in the NCBI Pathogen Detection database. *BMC Genom.* **23**, 268 (2022).
<https://doi.org/10.1186/s12864-022-08458-z>
- 56 Peñil-Celis, A. et al. Mobile genetic elements define the non-random structure of the *Salmonella enterica* serovar Typhi pangenome. *mSystems* **9**, e00365-00324 (2024).
<https://doi.org/doi:10.1128/msystems.00365-24>
- 57 Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016). <https://doi.org/10.1186/s13059-016-0997-x>
- 58 Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310-4312 (2018). <https://doi.org/10.1093/bioinformatics/bty539>

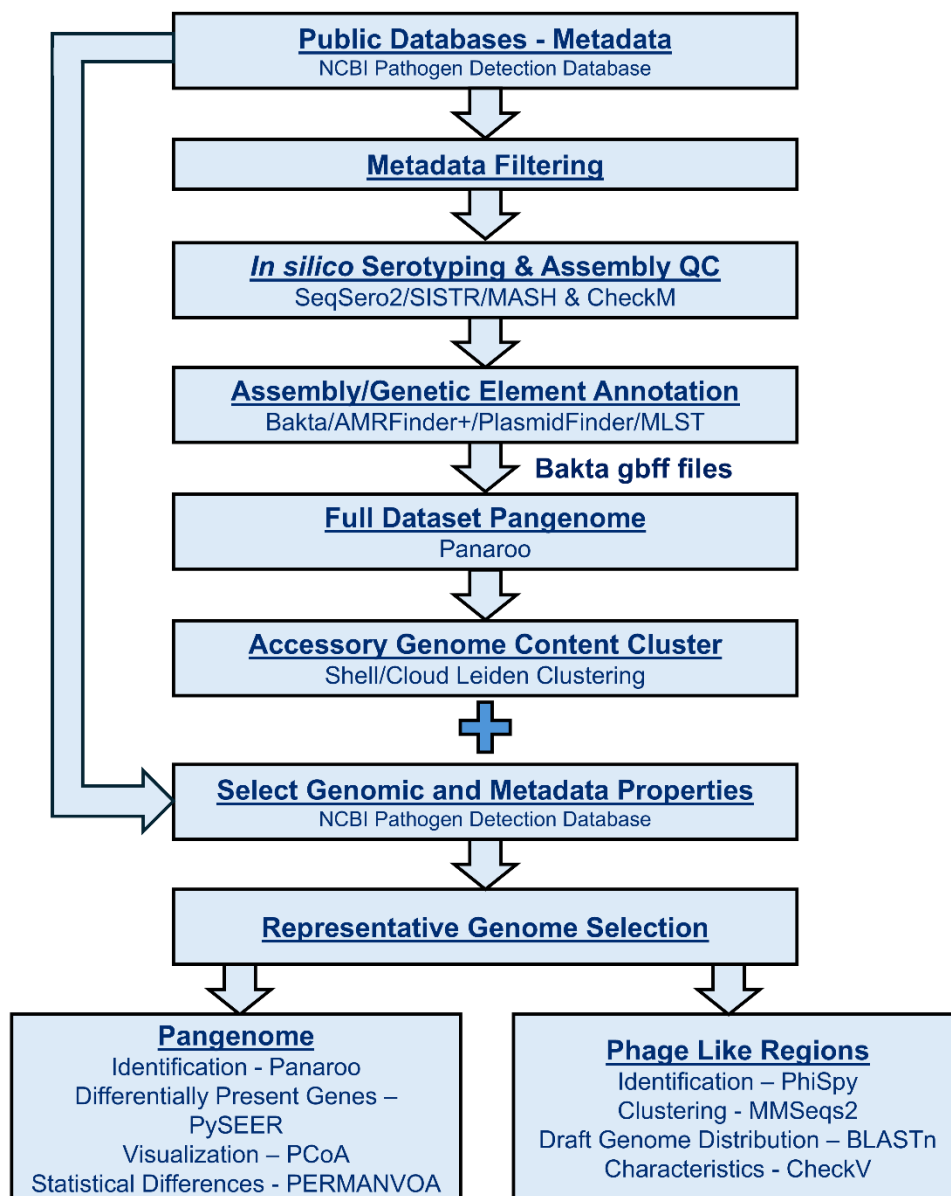


Figure 1 – Data Analysis Workflow

Flow diagram showing the general steps and tools used for data analysis. NCBI stands for National Center for Biotechnology Information, QC for quality check, PERMANVOA for permutational analysis of variance, PCoA for principal coordinates analysis, and SNPs for single nucleotide polymorphisms. Tool versions can be found in the Materials and Methods section.

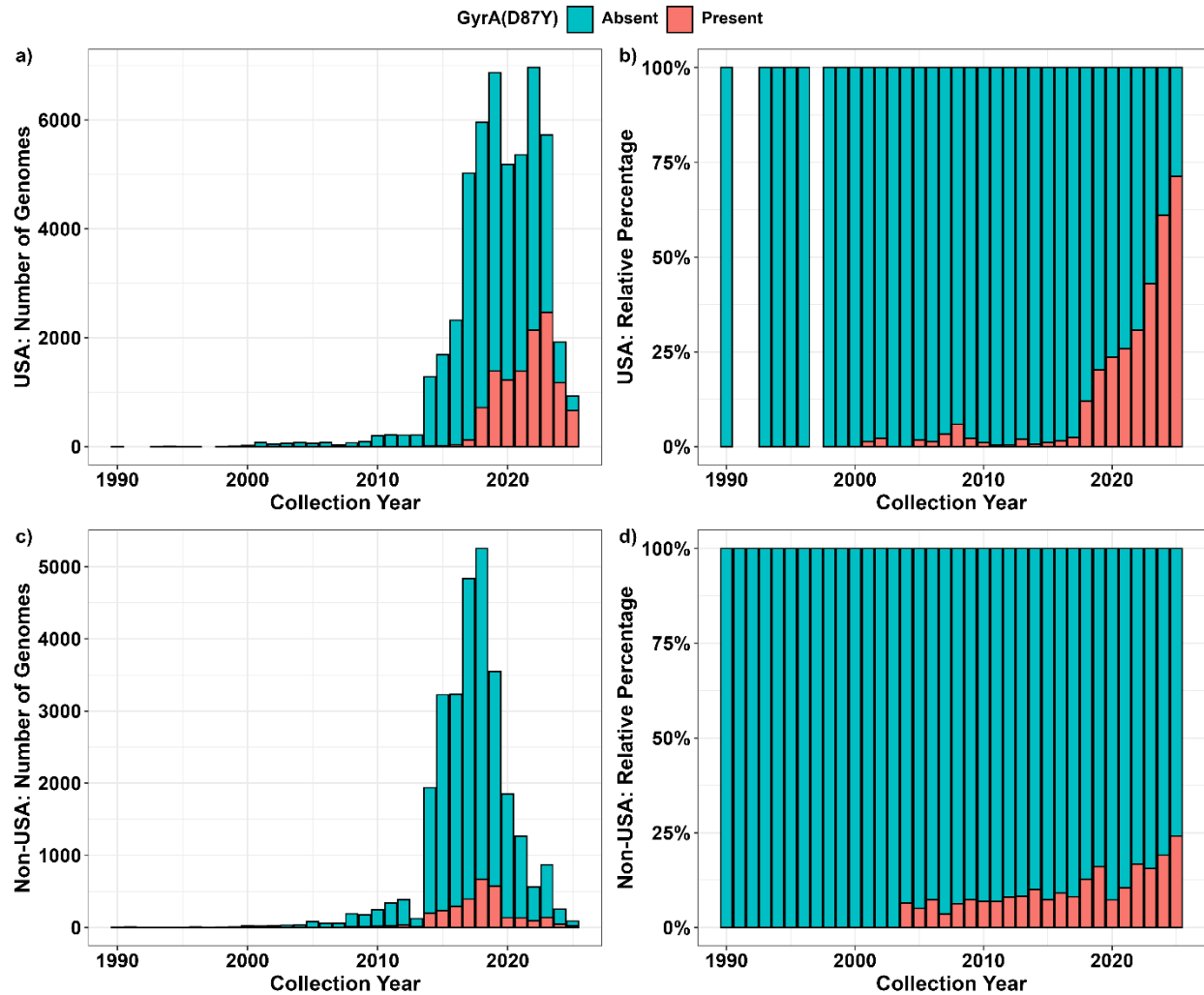


Figure 2 – GyrA(D87Y) Presence/Absence Over Time in USA versus Non-USA Genomes

The number (a, c) and relative percentage (b, d) of *Salmonella enterica* subspecies *enterica* serovar Enteritidis genomes (n=79,476 total) isolated from USA (a, b) and non-USA (c, d) locations with (red) or without (blue) the GyrA(D87Y) substitution over time.

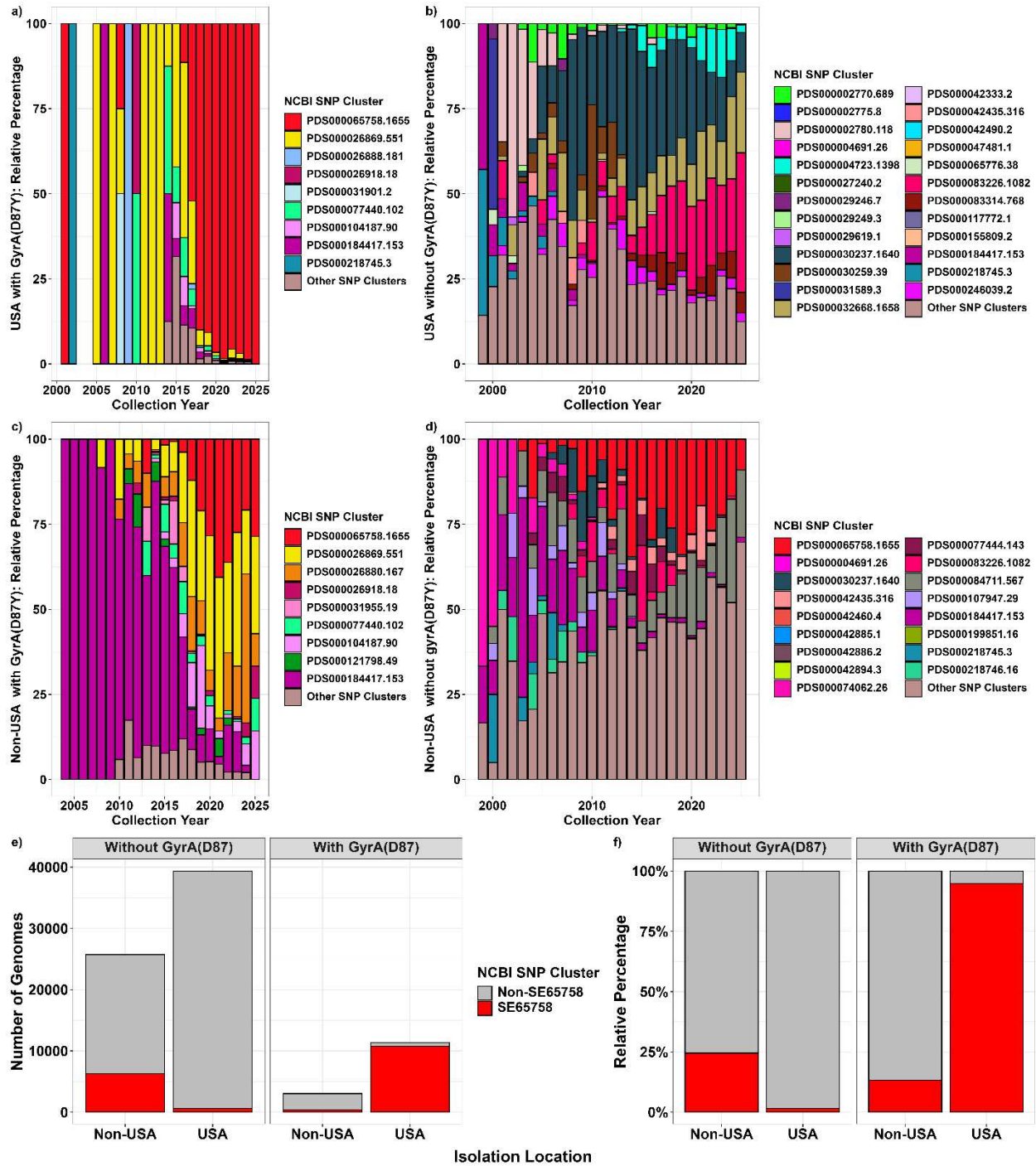


Figure 3 – Summary of Genomes With or Without GyrA(D87Y) by NCBI SNP Cluster for USA and Non-USA Isolation Locations

The relative percentage of NCBI SNP clusters of *Salmonella enterica* subspecies *enterica* serovar Enteritidis genomes (n=79,476 total) over time in USA (a, b) and non-USA (c, d) genomes with (a, c) and without (b, d) the GyrA(D87Y) substitution. In Figures a-d, color represents the SNP cluster with the SNP cluster of interest (PDS000065758 aka SE65758) in red

715 at the top of the legend, and all SNP clusters with an average relative percentage of less than 1%
716 across all collection years combined into Other SNP clusters at the bottom of the legend (rosy
717 brown). Figures e and f summarizes the number (e) and relative percentage (f) of *Salmonella*
718 *enterica* subspecies *enterica* serovar Enteritidis genomes by USA/non-USA Isolation Location
719 by GyrA(D87Y) Presence/Absence by SE65758/non-SE65758 NCBI SNP Cluster (red =
720 SE65758, gray = all other SNP clusters aka non-SE65758).

721

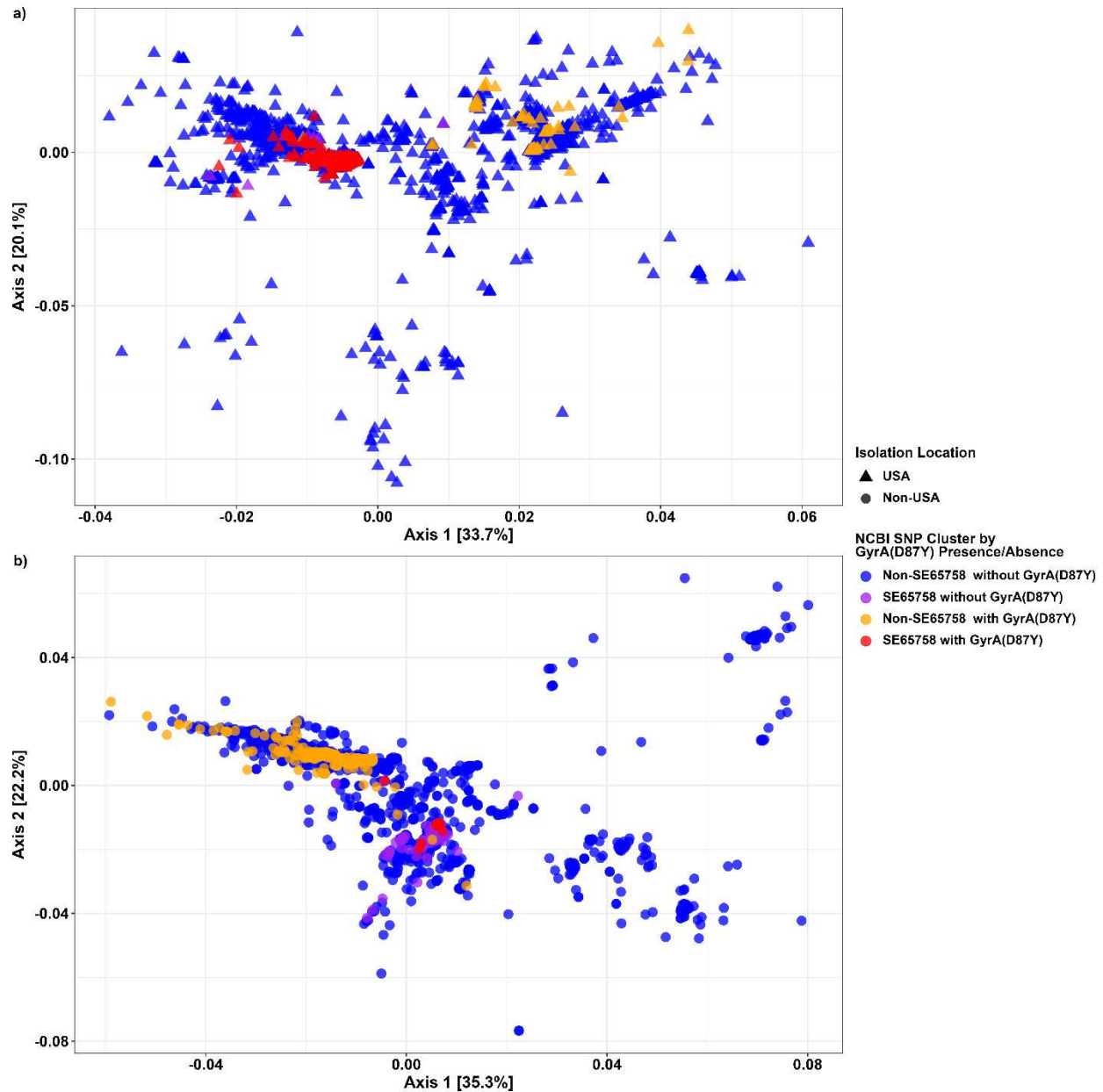


Figure 4 –Jaccard-Based Principal Coordinates Analysis of the Representative Genomes based on the Presence/Absence of Genes in the Pangenome

Jaccard distance matrices based on the presence/absence of pangenome genes were used as ordination inputs for principal coordinates analysis. Analysis was conducted on representative draft (n=9,354) and complete (n=199) *Salmonella enterica* subspecies *enterica* serovar Enteritidis genomes from the USA (a) (triangles) and Non-USA countries (b) (circles). Color represents the presence/absence of the GyrA(D87Y) substitution (with and without) and the NCBI SNP Cluster (SE65758 aka PDS00065758 NCBI SNP cluster and non-SE65758) of each genome.

732 Table 1 –PERMANOVA Comparisons of the Presence/Absence of Genes within Representative

733 Genomes

Genome Category1 ^a	Genome Category 2 ^a	Sums of Squares	Pseudo-F Statistic	R ²	p-value	Adjusted p-value (All) ^c	Adjusted p-value (Focused) ^d
USA SE65758 with GyrA(D87Y)	USA non-SE65758 without GyrA(D87Y)	-1.83E-02 ^b	-33.97 ^b	-0.0063 ^b	ND ^e	ND ^e	ND ^e
USA SE65758 with GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	8.36E-05	4.65	0.0041	0.0118	0.3304	0.0826
USA SE65758 with GyrA(D87Y)	USA non-SE65758 with GyrA(D87Y)	5.59E-02	1279.93	0.53	0.0001	0.0028	0.0007
USA SE65758 with GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	8.08E-05	4.92	0.0044	0.0115	0.322	0.0805
Non-USA SE65758 with GyrA(D87Y)	Non-USA non-SE65758 without GyrA(D87Y)	8.60E-03	6.18	0.0021	0.0006	0.0168	0.0042
Non-USA SE65758 with GyrA(D87Y)	Non-USA SE65758 without GyrA(D87Y)	2.07E-05	0.64	0.00093	0.5761	1	1
Non-USA SE65758 with GyrA(D87Y)	Non-USA non-SE65758 with GyrA(D87Y)	3.70E-02	102.83	0.2029	0.0001	0.0028	0.0007

734 Permutational analysis of variance (PERMANOVA) was run on a Jaccard-based distance matrix

735 generated from a gene presence/absence matrix created from pangenomic analysis (Panaroo

736 v.1.6.0) of a dataset containing 9,553 genomes.

737 ^aAnalysis was run on all combinations (n=28) of the eight Genome Category groups based on

738 three metadata categories: USA/non-USA Isolation Location, GyrA(D87Y) Presence/Absence,

739 and NCBI SNP Cluster (SE65758 aka PDS000065758 NCBI SNP cluster or non-SE65758); only

740 the seven combinations of interest are in Table 1, see Supplementary Table 1 for full results.

741 ^bNegative Sum of Squares, Pseudo-F Statistic, and R² values indicate analysis could not be

742 completed likely due to high dispersion of genes within a group.

743 ^cAdjusted p-value (All) is the Bonferroni adjusted p-value based on all combinations (n=28).

744 ^dAdjusted p-value (Focused) is the Bonferroni adjusted p-value based on the 7 combinations of

745 interest.

746 ^eNot Determined; this likely occurred due to high dispersion of genes within a group, thus

747 PERMANOVA analysis could not be completed.

748 Table 2 – Statistical Analysis of Genetic Differences between Genome Categories of Interest

Comparison	Comparison Number	Patterns Tested ^a	Adjusted P-Value	Number of Significantly Different Genes			
				Default Filtered P-value ^b	Adjusted Filtered P-value ^c	Default LRT P-value ^d	Adjusted LRT P-value ^e
USA SE65758 genomes with GyrA(D87Y) (n=1062) vs Non-USA SE65758 genomes with GyrA(D87Y) (n=48)	1	100	0.0005	18	3	21	2
USA SE65758 genomes with GyrA(D87Y) (n=1062) vs USA SE65758 genomes without GyrA(D87Y) (n=68)	2	105	0.00047619	11	1	21	2
Non-USA SE65758 genomes with GyrA(D87Y) (n=48) vs Non-USA SE65758 genomes without GyrA(D87Y) (n=640)	3	161	0.000310559	4	0	5	0
USA SE65758 genomes with GyrA(D87Y) (n=1062) vs USA Non-SE65758 genomes with GyrA(D87Y) (n=80)	4	281	0.000177936	182	173	184	173
Non-USA SE65758 genomes with GyrA(D87Y) (n=48) vs Non-USA Non-SE65758 genomes with GyrA(D87Y) (n=358)	5	712	7.02247E-05	288	128	482	157

749 PySEER (v1.4.1) was used to conduct a genome-wide association study (GWAS) based on the presence/absence of genes across
750 Genome Categories of interest; this table is a summary of these results, see Supplementary Data 6 for full results.
751 Genome Categories are based on three metadata categories: USA/non-USA Isolation Location, GyrA(D87Y) Presence/Absence, and
752 NCBI SNP Cluster (SE65758 aka PDS000065758 NCBI SNP cluster or non-SE65758).
753 By default, PySEER will remove any variant with allele frequencies (af) or presences less than 1% (--min-af) and more than 99% (--
754 max-af) of the genomes in the comparison being tested.
755 The inputs for the pangenome genes-based GWAS were: a MASH-based distance matrix (v2.3) generated with all complete and
756 representative draft genomes (n=9,553) (--distances; --mds metric) and a gene presence/absence matrix (--pres).
757 ^aPatterns Tested refers to number of samples tested for differences after filtering for minimum and maximum allele frequency.

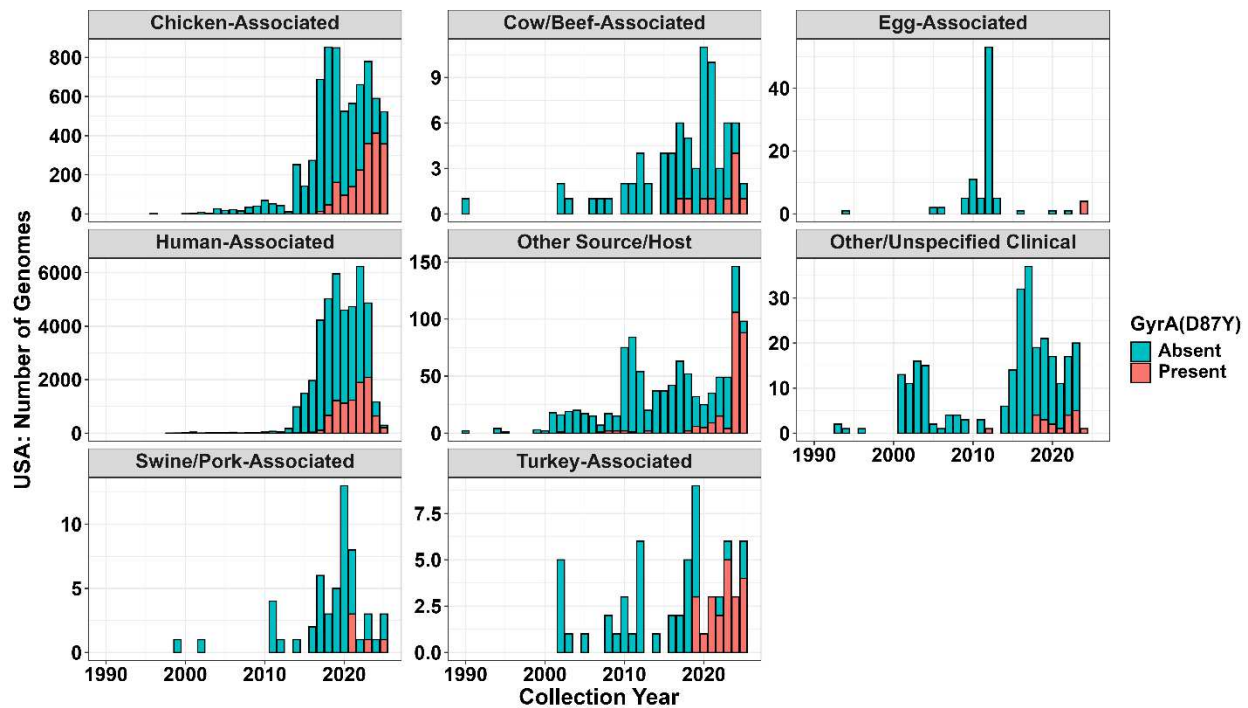
758 ^bFiltered p-value refers to the initial p-value calculated by PySEER using its chi2 test-based algorithm.

759 ^cThe adjusted p-value was calculated by dividing 0.05 by the number of patterns tested (i.e. Bonferroni multi-testing adjustment)

760 ^dThe likelihood ratio test (LRT) p-value was calculated by considering population structure to account for confounding effects due to
761 relatedness between samples.

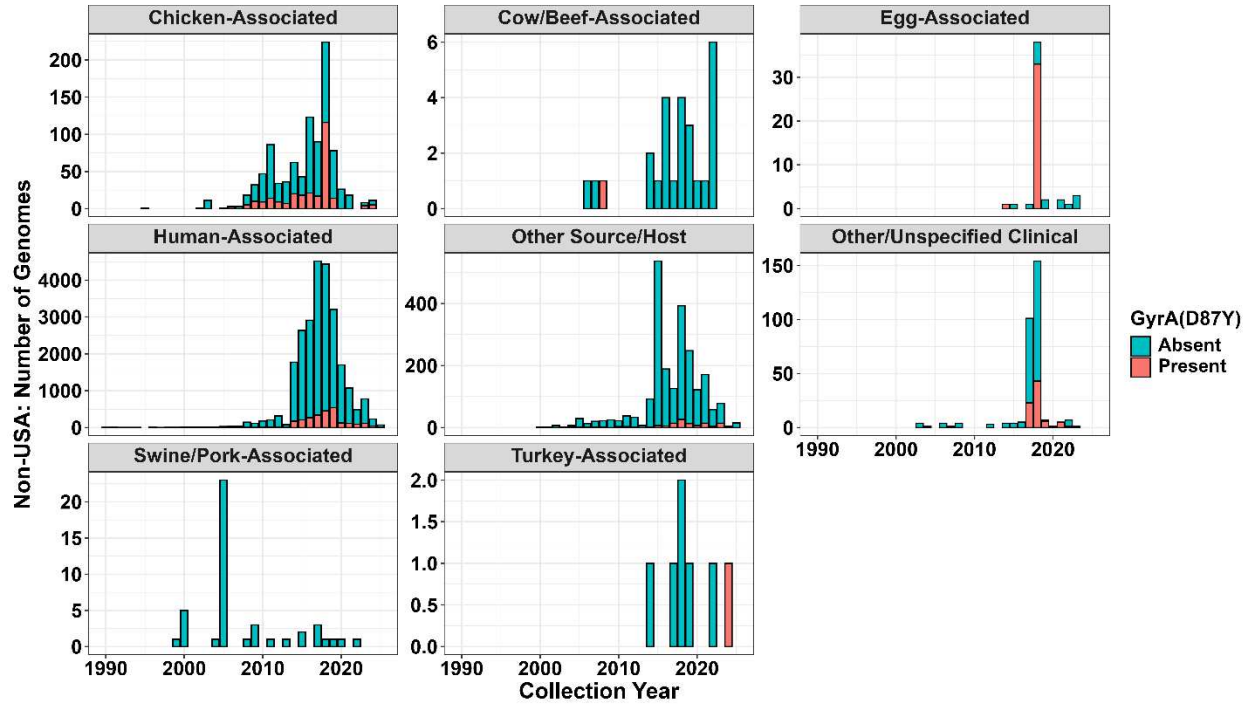
762 .

Supplementary Information



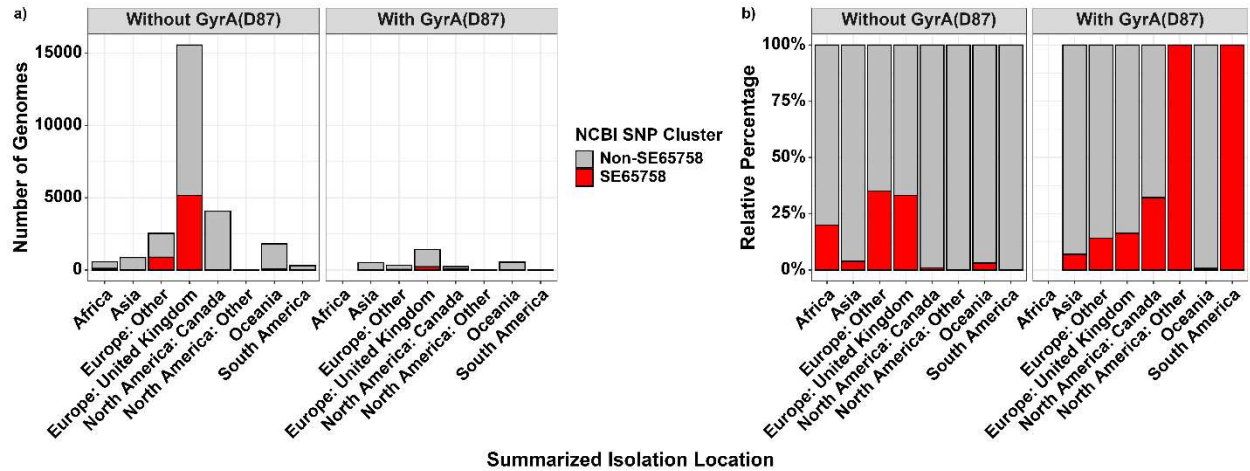
Supplementary Figure 1 - USA Genomes by GyrA(D87Y) Presence/Absence by Isolation Source (Summarized) Over Time

Stacked bar graphs showing the number of USA-based genomes with and without the D87Y gyrase (GyrA) substitution over time by Isolation Source (Summarized). Isolation Source (Summarized) was generated based upon the keyword searches querying the data contained within the Isolation source, Host, Isolation type, Source type, and IFSAC category columns from the NCBI Pathogen Detection Database metadata in a defined order. See Supplementary Data 1 for more information. The presence of the GyrA(D87Y) substitution in a genome was based on querying the “AMR.genotypes” NCBI Pathogen Detection Database metadata.



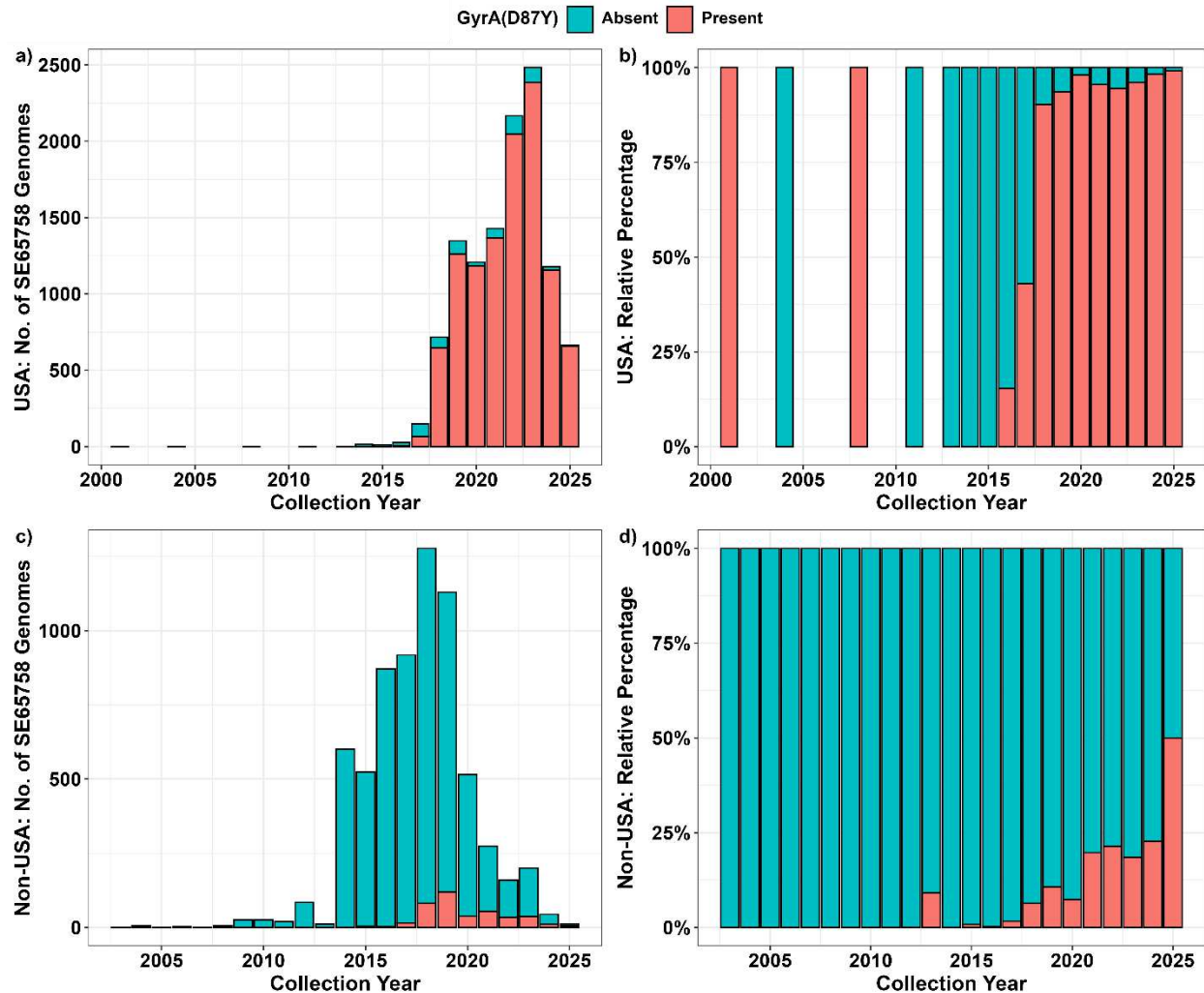
Supplementary Figure 2 – Non-USA Genomes by GyrA(D87Y) Presence/Absence by Isolation Source (Summarized) Over Time

Stacked bar graphs showing the number of Non-USA-based genomes with and without the D87Y gyrase (GyrA) substitution over time by Isolation Source (Summarized). Isolation Source (Summarized) was generated based upon the keyword searches querying the data contained within the Isolation source, Host, Isolation type, Source type, and IFSAC category columns from the NCBI Pathogen Detection Database metadata in a defined order. See Supplementary Data 1 for more information. The presence of the GyrA(D87Y) substitution in a genome was based on querying the “AMR.genotypes” NCBI Pathogen Detection Database metadata.



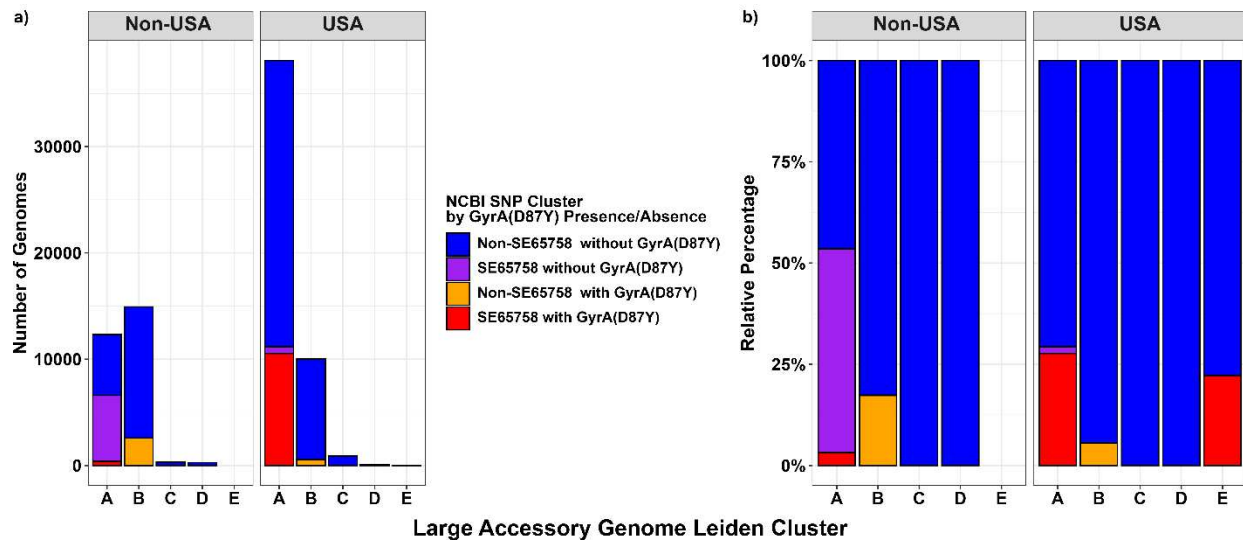
Supplementary Figure 3 - Non-USA Isolation Location by GyrA(D87Y) Presence/Absence by SE65758/non-SE65758 NCBI SNP Cluster

The number (a) and relative percentage (b) of Non-USA *Salmonella enterica* subspecies *enterica* serovar Enteritidis genomes (n=28,757) by summarized non-USA isolation location by GyrA(D87Y) substitution carriage by summarized NCBI SNP cluster (red = PDS000065758 aka SE65758, gray = all other SNP clusters aka non-SE65758).



Supplementary Figure 4 – GyrA(D87Y) Presence/Absence Over Time in USA versus Non-USA SE65758 Genomes

The number (a, c) and relative percentage (b, d) of *Salmonella enterica* subspecies *enterica* serovar Enteritidis genomes within the PDS000065758 NCBI SNP cluster (SE65758) (n=18,099 total) isolated from USA (a, b) and non-USA (c, d) locations with (red) or without (blue) the GyrA(D87Y) substitution over time.



Supplementary Figure 5 - Distribution of USA and Non-USA Genome Categories Across Large Accessory Genome Leiden Clusters

The number (a) and relative percentage (b) of *Salmonella enterica* subspecies *enterica* serovar Enteritidis USA and Non-USA genomes by GyrA(D87Y) Presence/Absence by SE65758/non-SE65758 NCBI SNP cluster (either within the PDS000065758 NCBI SNP cluster (SE65758) or not (non-SE65758)) across large accessory genome clusters. Clusters were generated by conducting Leiden clustering on a graph of draft genomes weighted by the number of overlapping cloud and accessory pangenome genes (present in <95% of all genomes). Only draft genomes with a minimum of 107 overlapping genes were clustered (n=76,976). Only clusters with 5 or more genomes (large) were considered for visualization. Leiden clustering also generated two small clusters with 3 (Cluster F) and 2 genomes (Cluster G).

Supplementary Table 1 – Full Representative Genomes Gene Presence/Absence-based PERMANOVA Results

Genome Category 1 ^a	Genome Category 2 ^a	Sums of Squares	Pseudo-F Statistic	R ²	p-value	Adjusted p-value ^c
USA non-SE65758 without GyrA(D87Y)	Non-USA non-SE65758 without GyrA(D87Y)	3.06E-01	316.25	0.042	0.0001	0.0028
USA non-SE65758 without GyrA(D87Y)	USA non-SE65758 with GyrA(D87Y)	4.35E-02	65.69	0.015	0.0001	0.0028
USA non-SE65758 without GyrA(D87Y)	Non-USA non-SE65758 with GyrA(D87Y)	2.18E-01	337.64	0.067	0.0001	0.0028
Non-USA SE65758 without GyrA(D87Y)	Non-USA non-SE65758 without GyrA(D87Y)	1.23E-01	105.30	0.029	0.0001	0.0028
Non-USA SE65758 without GyrA(D87Y)	USA non-SE65758 with GyrA(D87Y)	5.49E-02	730.17	0.50	0.0001	0.0028
Non-USA SE65758 without GyrA(D87Y)	USA SE65758 with GyrA(D87Y)	8.07E-04	34.79	0.020	0.0001	0.0028
Non-USA SE65758 without GyrA(D87Y)	Non-USA non-SE65758 with GyrA(D87Y)	2.15E-01	1281.86	0.56	0.0001	0.0028
Non-USA non-SE65758 without GyrA(D87Y)	USA non-SE65758 with GyrA(D87Y)	1.95E-02	14.09	0.0046	0.0001	0.0028
Non-USA non-SE65758 without GyrA(D87Y)	USA SE65758 with GyrA(D87Y)	1.63E-01	155.72	0.037	0.0001	0.0028
Non-USA non-SE65758 without GyrA(D87Y)	Non-USA non-SE65758 with GyrA(D87Y)	9.54E-02	73.11	0.022	0.0001	0.0028
Non-USA non-SE65758 without GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	1.56E-02	11.29	0.0037	0.0001	0.0028
USA non-SE65758 with GyrA(D87Y)	USA SE65758 with GyrA(D87Y)	5.59E-02	1279.93	0.53	0.0001	0.0028
USA non-SE65758 with GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	2.15E-02	82.81	0.40	0.0001	0.0028
USA non-SE65758 with GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	2.83E-02	118.74	0.45	0.0001	0.0028
USA SE65758 with GyrA(D87Y)	Non-USA non-SE65758 with GyrA(D87Y)	2.46E-01	2146.80	0.60	0.0001	0.0028
Non-USA non-SE65758 with GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	3.70E-02	102.83	0.20	0.0001	0.0028
Non-USA non-SE65758 with GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	5.31E-02	152.44	0.26	0.0001	0.0028
Non-USA non-SE65758 without GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	8.60E-03	6.18	0.0021	0.0006	0.0168
USA SE65758 with GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	8.08E-05	4.92	0.0044	0.0115	0.322
USA non-SE65758 with GyrA(D87Y)	Non-USA non-SE65758 with GyrA(D87Y)	1.30E-03	3.21	0.0073	0.0118	0.3304
USA SE65758 with GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	8.36E-05	4.65	0.0041	0.0118	0.3304
USA non-SE65758 without GyrA(D87Y)	Non-USA SE65758 without GyrA(D87Y)	-5.16E-03 ^b	-8.82 ^b	-0.0018 ^b	ND ^d	ND ^d
USA non-SE65758 without GyrA(D87Y)	USA SE65758 with GyrA(D87Y)	-1.83E-02 ^b	-33.97 ^b	-0.0063 ^b	ND ^d	ND ^d
USA non-SE65758 without GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	-1.50E-03 ^b	-2.27 ^b	-0.00052 ^b	ND ^d	ND ^d
USA non-SE65758 without GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	-6.87E-04 ^b	-1.05 ^b	-0.00024 ^b	ND ^d	ND ^d

Non-USA SE65758 without GyrA(D87Y)	Non-USA SE65758 with GyrA(D87Y)	2.07E-05	0.64	0.00093	0.5761	1
Non-USA SE65758 without GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	9.84E-05	2.84	0.0040	0.0523	1
Non-USA SE65758 with GyrA(D87Y)	USA SE65758 without GyrA(D87Y)	2.21E-05	0.78	0.0068	0.6143	1

Permutational analysis of variance (PERMANOVA) analysis was run on a Jaccard-based distance matrix generated from a gene presence/absence matrix created from pangenomic analysis (Panaroo v.1.6.0) of a dataset of 9,553 genomes.

SE65758 refers to the PDS000065758 NCBI SNP Cluster associated with decreased susceptibility to ciprofloxacin *Salmonella enterica* subspecies *enterica* serovar Enteritidis and GyrA(D87Y) to the gyrase aspartate (Asp; D) to tyrosine (Tyr; Y) amino acid substitution.

^aAnalysis was run on all combinations (n=28) of USA/non-USA Isolation Location by GyrA(D87Y) Presence/Absence by SE65758/non-SE65758 NCBI SNP Cluster (n=8 groups) aka Genome Category

^bNegative Sum of Squares, Pseudo-F Statistic, and R² values indicate analysis could not be completed likely due to high dispersion of genes within a group

^cAdjusted p-value (All) is the Bonferroni adjusted p-value based on all combinations (n=28)

^dNot Determined; this likely occurred due to high dispersion of genes within a group, thus PERMANOVA analysis could not be completed.

Supplementary Data 1. Filtered Dataset Metadata

Metadata associated with all genomes used in analysis after filtering for known metadata categories, in-silico serotyping, and assembly quality.

Supplementary Data 2. Various Datasets by USA/non-USA Isolation Location by GyrA(D87Y) Presence/Absence by SE65758/non-SE65758 NCBI SNP Cluster

Counts associated with Genome Categories (USA/non-USA Isolation Location, GyrA(D87Y) Presence/Absence, and SE65758/non-SE65758 NCBI SNP Cluster) across various datasets.

Supplementary Data 3. Presence/Absence of vOTUs across Representative Draft Genomes

Distribution of prophage-like regions (PLRs) across representative draft genomes based on alignment to viral operational taxonomic units (vOTU) identified in the complete genome dataset.

Supplementary Data 4. Representative Genomes Pangenome Gene Presence/Absence Table

Presence/absence matrix of merged paralog pangenome genes across representative draft genomes.

Supplementary Data 5. Full PySEER Representative Genomes Pangenome Gene Results

Statistics associated with a genome-wide association study (GWAS) to determine which pangenome genes were statistically different between Genome Categories.

Supplementary Data 6. Statistically Significant Pangenome Genes Between Genome Categories

Genes considered to be statistically different between the elements of each comparison via either filter or likelihood ratio test (LRT) adjusted p-values.

Supplementary Data 7. Isolation Source and Location (Summarized) Designation Protocol

Table of the keywords used to define subcategories associated with Isolation Source (Summarized), Isolation Location (Summarized), Isolation Location (USA Region), and Isolation Location (USA Division).

Supplementary Data 8. Complete Genomes PhiSpy, MMSEQs vOTU Designation, and CheckV statistics per PLR

Statistical results associated with PhiSpy, MMSEqs, and CheckV analysis of the complete genome dataset to identify and classify prophage-like regions (PLRs).

Supplementary Data 9. Pairwise Complete Genome vOTUs BLAST Results Passing MIUViG Thresholds

BLASTn statistics associated with aligning viral operational taxonomic units (vOTU) against each other to determine if there were any overlapping sequences.

Supplementary Data 10. Full Draft Genome Dataset Pangenome Gene Presence/Absence Matrix

Presence/absence matrix of merged paralog pangenome genes across all draft genomes in the dataset.

Supplementary Data 11. Weighted Edge Filter Trialing Results

Leiden clustering results associated with filtering the draft genome dataset accessory genome matrix for different thresholds of overlapping genes between genomes to determine the best threshold for further analysis.

Supplementary Data 12. Complete Genome Plasmid Genes Linked to Pangenome Genes

Complete genome plasmid genes were linked to pangenome gene designations to interpret what statistically different genes across Genome Categories were plasmid-associated.

Supplementary Data 13. Complete Genome Prophage-like Region Genes Linked to Pangenome Genes

Complete genome prophage-like region (PLR) genes were linked to pangenome gene designations to interpret what statistically different genes across Genome Categories were PLR-associated.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryData.zip](#)