Large-scale comparative genomics of *Salmonella enterica* to refine the organization of the global *Salmonella* population structure

Chao Chun Liu¹, William W.L. Hsiao¹,²*

¹ Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada
² Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada

*Correspondence:*
William Hsiao
wwhsiao@sfu.ca

**Keywords:** Bacterial population structure, comparative genomics, *Salmonella*, Phylogenetics, Bacterial typing nomenclature, molecular evolution

1. **Abstract**

Since the introduction of the White-Kauffmann-Le Minor (WKL) scheme for *Salmonella* serotyping, the nomenclature remains the most widely used for reporting the disease prevalence of *Salmonella enterica* across the globe. With the advent of whole genome sequencing (WGS), traditional serotyping has been increasingly replaced by in-silico methods that couple the detection of genetic variations in antigenic determinants with sequence-based typing. However, despite the integration of genomic-based typing by in-silico serotyping tools such as SeqSero2 and SISTR, in-silico serotyping in certain contexts remains ambiguous and insufficiently informative due to polyphyletic serovars. Furthermore, in spite of the widespread acknowledgement of polyphyly from genomic studies, the serotyping nomenclature remains unaltered. To prompt refinements to the *Salmonella* typing nomenclature for disease reporting, we herein performed a systematic characterization of putative polyphyletic serovars and the global *Salmonella* population structure by comparing 180,098 *Salmonella* genomes (representing 723 predicted serovars) from GenomeTrakr and PubMLST databases. We identified a range of core genome MLST typing thresholds that result in stable population structure, potentially suitable as the foundation of a genomic-based typing nomenclature for longitudinal surveillance. From the genomic comparisons of hundreds of predicted serovars, we demonstrated that in-silico serotyping classifications do not consistently reflect the population divergence observed at the genomic level. The organization of *Salmonella* subpopulations based on antigenic determinants can be confounded by homologous recombination and niche adaptation, resulting in shared classification of highly divergent genomes and misleading distinction between highly similar genomes. In consideration of the pivotal role of *Salmonella* serotyping, a compendium of putative polyphyletic serovars was compiled and made publicly available to provide additional context for future interpretations of in-silico serotyping results in disease surveillance settings. To refine the typing nomenclatures used in *Salmonella* surveillance reports, we foresee an improved typing scheme to be a hybrid that integrates both genomic and antigenic information such that the resolution from WGS is leveraged to improve the
precision of subpopulation classifications while preserving the common names defined by the WKL scheme. Lastly, we stress the importance of controlled vocabulary integration for typing information in open data settings in order for the global Salmonella population dynamics to be fully trackable.

2. Impact Statement

Salmonella enterica (S. enterica) is a major foodborne pathogen responsible for an annual incidence rate of more than 90 million cases of foodborne illnesses worldwide. To surveil the high order Salmonella lineages, compare disease prevalence across jurisdictions worldwide, and inform risk assessments, in-silico serotyping has been established as the gold standard for typing the bacteria. However, despite previous Salmonella genomic studies reporting discordance between phylogenomic clades and serovars, refinements have yet been made to the serotyping scheme. Here, we analyzed over 180,000 Salmonella genomes representing 723 predicted serovars to subdivide the population into evolutionarily stable clusters in order to propose a stable organization of the Salmonella population structure that can form the basis of a genomic-based typing scheme for the pathogen. We described numerous instances in which genomes between serotypes are more similar than genomes within a serotype to reflect the inconsistencies of subpopulation classifications based on antigenic determinants. Moreover, we found inconsistencies between predicted serovars and reported serovars which highlighted potential errors in existing in-silico serotyping tools and the need to implement controlled vocabularies for reporting Salmonella subtypes in public databases. The findings of our study aim to motivate the future development of a standardized genomic-based typing nomenclature that more accurately captures the natural populations of S. enterica.

3. Data Summary

The assembly accession numbers of the genomes analyzed in this study (n = 204,952) and the associated metadata (e.g. sampling location, collection date, FTP address for retrieval) are documented in Table S1. The GenomeTrkr genomes were retrieved from the National Center for Biological Information GenBank database. The PubMLST genomes were retrieved using the BIGSdb API.

4. Introduction

Salmonella enterica (S. enterica) is a significant public health burden that causes more than 90 million cases of foodborne infections globally every year (1). Existing surveillance networks rely on globally standardized typing nomenclatures to rapidly share information between laboratories, veterinarians and public health authorities for risk assessments and disease tracking. Since its introduction more than 80 years ago, serotyping has been the predominant method for reporting Salmonella disease prevalence and represents a high-order classification to subdivide the Salmonella population structure into characteristic groups that differ in ecological and clinical features (2). To date, the White-Kauffmann-Le Minor (WKL) scheme subdivides S. enterica into over 2,600 serovars based on variations in the O (somatic) and H (flagellar) antigens (3). Although Salmonella serotyping has been established as a central nomenclature that is useful for disease reporting and risk assessments,
its inconsistencies in the representation of natural microbial populations due to the existence of polyphyletic serovars have challenged its role as the gold standard scheme for longitudinal surveillance (4-5).

Advances in DNA sequencing over the years have opened up new avenues for microbial typing. Inference of sequence types based on genetic variations in 6-10 housekeeping genes known as multi-locus sequence typing (MLST), has shown high concordance with Salmonella serotyping, while achieving greater discriminatory power (5). Highly clonal serovars such as Enteritidis are widely associated with a specific sequence type, while genetically diverse serovars such as Newport and Typhimurium are subdivided into multiple sequence types. Moreover, MLST represents a more robust scheme for natural population characterization compared to antigenic variations, as it overcomes the occasional failure of the serotyping scheme to demonstrate genetic similarity amongst isolates of the same serovar (5-8). In an attempt to develop a novel Salmonella typing scheme that better reflects real and stable subpopulations, Achtman et al. described the concept of eBurst Groups (eBGs) that represent single-linkage chains of sequence types differing by a single locus (5). A total of 138 eBGs was defined and demonstrated effective segregation of polyphyletic serovars into distinguished groups (5). Although eBG designations have been validated across >100,000 Salmonella genomes and determined to be an evolutionarily stable typing scheme (8), its discriminatory power remains limited, as the scheme only accounts for genetic variations in seven housekeeping genes.

Since then, increasing adoption of whole genome sequencing (WGS) for microbial characterization accompanied by the development of efficient bioinformatic tools renders strain differentiation based on a greater number of genetic markers feasible. Leveraging the plethora of genetic information made available by WGS, core genome MLST (cgMLST) extends upon traditional MLST by the inclusion of 300-3,000 loci present in 95-99% of Salmonella isolates in its scheme (9). And given the conservative nature of the core genome composition, it represents a stable genomic segment that can be used to infer genealogy (9). Due to its discriminatory power superior to that of classical fingerprinting methods such as pulsed-field gel electrophoresis, phage typing or multi-locus variable-number tandem repeat analysis, cgMLST has been shown to be more effective in the differentiation of closely related strains rendering it a compelling complement for outbreak investigations (10, 11). In addition, due to the aforementioned correlation between MLST and antigenic variations, cgMLST is leveraged by in-silico serotyping workflows such as SISTR (12) and presents valuable information when putative serovars cannot be discriminated based on antigenic determinants alone.

However, the application of cgMLST for the routine tracking of high order Salmonella lineages has not been widely explored, as the serotyping scheme has long been used for diagnostics and longitudinal surveillance. To propose an alternative genomic-based approach to assign high order Salmonella groupings, we characterized stable Salmonella subpopulations from a diverse dataset of 180,098 Salmonella genomes retrieved from public databases, namely GenomeTrakr and PubMLST. We applied a recently proposed statistic called neighbourhood Adjusted Wallace Coefficient (nAWC) that quantifies single-linkage cluster concordance between two neighbouring distance thresholds (T vs. T + 1) in which signatures of cluster stability would arise from the maintenance of high cluster concordance.
across multiple successive thresholds (13). The identification of long-range cgMLST typing thresholds that consistently yield high cluster concordance would inform an ideal cutoff to define a stable subdivision of the *Salmonella* genomes into characteristic groups. Such subdivision would consequently allow the development of a genomic-based typing scheme suitable for longitudinal surveillance.

It remains challenging to adopt a genomic-based typing scheme for longitudinal surveillance given how integral serotyping is for the current state of global surveillance, and decades of research on the epidemiological significance associated with different serovars rendering serotype assignments irreplaceable. Although existing in-silico serotyping tools leverage the correlation between genome-wide variations and serovars for in-silico prediction, this observed correlation has yet been systematically evaluated across most serotypes on a large scale. Hence, to contribute to the effort, we compared genomic clusters to in-silico serovar assignments in order to assess whether isolates of different serovars converge in the same genomic cluster and the extent of sequence dissimilarity amongst isolates of the same serovar. We further demonstrated the utility of existing genome-wide association methods to characterize genetic features associated with microbial lineage divergences to formulate the basis of lineage-specific subtyping schemes. The usage of typing nomenclatures that consistently represent natural bacterial populations will permit a more precise categorization of genomes and avoid the misinterpretation of epidemiological patterns from surveillance data. However, the benefits of genomic-based typing schemes cannot be realized without the community convergence on an universal genomic-based typing scheme and collaborative efforts to characterize the functional differences, ecological adaptations and health risks associated with distinct subpopulations.

5. Methods

5.1. Study dataset

The metadata of the *Salmonella* genomes archived in the GenomeTrakr database (n = 186,312) was retrieved from the National Center for Biotechnology Information (NCBI) Pathogen Detection portal (https://www.ncbi.nlm.nih.gov/pathogens) on Dec 10th, 2019. The metadata of the *Salmonella* genomes archived in the PubMLST database (n = 18,640) were retrieved from the PubMLST database (https://pubmlst.org/) on Nov 9th, 2019. The GenomeTrakr genomes were downloaded from the NCBI GenBank database on December 13th, 2019. The PubMLST genomes were retrieved using the BIGSdb API on December 13th, 2019. The associated metadata of the combined list of genomes (n = 204,952) analyzed in this study is available in Table S1. Assembly statistics such as sequence length and N50 were evaluated using QUAST v5.0.2 (14). Genome completeness and contamination were evaluated using CheckM v1.1.2 (15). The minimum data quality criteria are as follows: genome length > 4 Mbp, genome completeness > 80%, genome contamination < 5%, N50 > 50 Kbp, proportion of N bases < 1%. Assemblies that failed to meet the minimum data quality were removed prior to any downstream analysis. Following cgMLST allele calling, genomes with <5% cgMLST profile completeness were also removed from further analyses. Of the remaining 180,098 *Salmonella* genomes that passed the quality filters, the dataset is composed of 723 different serovars predicted by SISTR v1.1.1 with confidence (12). The criteria for confident SISTR prediction is determined by the presence of O/H antigenic
determinants and >297/330 core genes in the query genome (12). The samples were collected from 137 countries with a sampling date range from 1900 to 2019. As the majority of contributors to the GenomeTrakr network are situated in the United States coupled with routine data synchronization with its United Kingdom's counterpart, Enterobase, 75% of the samples were collected from either the United States or the United Kingdom (16). The samples were sequenced from 4,634 different types of isolation sources distributed across human, environment, animal, and food origins with ~60% of the samples reported as human clinical samples.

5.2. cgMLST and Phylogenetic analysis

The cgMLST profile of each genome was generated using chewBACCA v2.0.16 based on a 3,246 loci scheme (17). The original cgMLST scheme was downloaded from https://zenodo.org/record/1323684 and consisted of 3,255 loci; however, 18 loci appeared to be paralogous pairs and hence one loci of each paralogous pair was arbitrarily removed from the original scheme. Hamming distances were calculated to compare the allelic profile differences and generate a pairwise distance matrix. A neighbour joining tree was constructed from the pairwise distance matrix using Rapidnj v2.3.2 and visualized in GrapeTree v2.1.9 (18, 19). Core genome SNP phylogenomics analysis of London/Winterthur and Chester/Madras genomes was conducted using PhaME v1.0.3 which uses Mummer v3.23 to generate core genome SNP alignments from draft genomes, and FastTree v2.1.9 to construct a maximum likelihood tree (20-22).

5.3. Cluster stability quantification

The adjusted Wallace coefficient was initially proposed to assess cluster concordance and has since been applied to compare the results of different subtyping methods. nAWC represents an extension to the adjusted Wallace coefficient by assessing cluster concordance between neighbouring thresholds that differ by one distance unit. Cluster membership resistance to incremental changes of distance thresholds or alternatively the maintenance of high degree of cluster concordance would consequently reflect cluster stability and render the median value of the cluster stable region suitable for defining WGS-based subtyping nomenclatures for pathogen surveillance.

TreeClust was used to cluster the leaves of the constructed neighbour joining tree to generate single-linkage clusters at thresholds ranging from 1 to 3,246 allelic differences. The resultant cluster assignments were compared between neighbouring thresholds to evaluate nAWC and Shannon entropy using the R scripts available at https://github.com/theInnuendoProject/nAWC.

5.4. Genome-wide association and functional enrichment

PCA of the cgMLST profiles from the two Enteritidis lineages was conducted in R using the adegenet v2.1.3 package (23). The principal component loadings were analyzed to quantify the allelic contribution to the observed lineage divergence. Evaluation of genome wide association to lineage divergence was conducted using Scoary v1.6.16 that tests the statistical association between the presence or absence of genotypic features to phenotypic
characteristics (24). The cgMLST profiles were transformed into an allele presence/absence matrix such that each unique allele of a loci is considered an independent genotypic feature. The Scoary output was filtered by multiple test corrected p-values to determine the set of alleles significantly associated with lineage identity. Similarly with pan-genome wide association, Roary v3.12.0 was used to generate a pan-genome gene presence/absence matrix and subsequently used as input to Scoary to identify statistical association between gene presence and lineage identity (25). The dN/dS ratio was calculated using the CODEML program part of the PAML v4.9 package (26). Protein functional annotation enrichment was performed by uploading a list of significant genes to the DAVID web service (27). COG category enrichment analysis was performed using the BOG v2.0 package in R (28).

6. Results

6.1. Genomics-based characterization of S. enterica subpopulations

Quantification of cluster stability by nAWC across the range of possible cgMLST distance threshold (T) values revealed two notable features: regions of compositionally identical clusters and cluster destabilizing thresholds (Fig. S1). Here, we define a given cluster stable region as any range of thresholds that maintains an nAWC > 0.99 for a minimum of 5 successive Ts. By this definition, we identified a total of 43 cluster stable regions of which the earliest signature of stability occurred at a median value of $T = 45$ allelic differences and the longest region of stability occurred at a median value of $T = 1,538$ allelic differences (Fig. S1A). The cluster membership dynamics observed with nAWC closely mirrored the changes in Shannon entropy, a commonly used statistic to quantify population structure richness and evenness. The consolidations of large clusters observed at $T > 2,732$ allelic differences demonstrated by frequent oscillations in nAWC coincided with the sharp decline of Shannon entropy (Fig. S1B). The identification of these cluster stable T's provided the means to use genomics data to redefine the high order groupings of closely related Salmonella isolates.

Visualization of the phylogenetic tree constructed from the entire dataset revealed distinctive segregation by serovar assignments, as many serovars such as Enteritidis, Thompson, Heidelberg, Oranienburg, and Braenderup appeared to form monophyletic clades (Fig. S2). Consistent with previous reports, the frequently reported polyphyletic serovars such as Newport and Kentucky segregated into multiple lineages that differed by more than 2,000 allelic differences. Antigenically similar serovars such as variants of Paratyphi B and Typhimurium were observed to share highly similar core genomes, as the respective isolates clustered together in the tree. Another heterogeneous cluster observed in the tree is comprised of Lubbock and Mbandaka (Fig. S2) in which the observation is concordant with a previous study that suggested Lubbock have descended from a Mbandaka ancestor via a recombination event (29).

To subdivide the dataset into the most stable genomic clusters, we searched for the longest range of cgMLST threshold that maintained nAWC > 0.99, which occurred between $T = 1,192$ and $T = 1,885$. The robustness of our cluster definition derives from the fact that the cluster assignments, richness and evenness remain stable across a long range of distance cutoffs. Evaluation of the clusters defined at the median $T$ of the longest region of stability ($T = 1,538$) revealed a total of 1,342 clusters and a high degree of concordance was observed.
between the predicted clusters and in-silico serovar assignments in which 507/723 (70.1%) serovars mapped to one specific genomics-based cluster (Fig. 1). Examples of serovars that demonstrated 1-to-1 mapping to genomic clusters included Heidelberg, Braenderup, Agona, Mbandaka and Paratyphi A. Interestingly, 15 of the top 20 most prevalent serovars that cause human salmonellosis reported by the United States Centers of Disease Control and Prevention (30) did not exhibit assignments to one specific genomic cluster at the most stable T, which reflected the multi-lineage nature of several clinically significant serovars, including Enteritidis, Newport, Typhimurium, i 1,4,[5],12:i:-, Javiana, Infantis, Muenchen, Montevideo, Thompson, Saintpaul, Oranienburg, Mississippi, Bareilly, Paratyphi B var. Java, and Anatum. In contrast to previous reports which often described the serovar Enteritidis to be sequence type specific and highly clonal (31), we found the Enteritidis isolates to segregate into 8 different clusters at the most stable T; however, significant bias was indeed observed in the relative abundance across the Enteritidis clusters in which 33,980 (99.4%) Enteritidis isolates in the dataset were classified under a single cluster, Cluster 841. It is unlikely that the minor clusters were mispredicted due to data quality, as their average genome completeness and contamination were within 1% difference from that of Cluster 841. Instead, these minor clusters could be explained by geographical origins given that the remaining 210 Enteritidis isolates were primarily sampled from Australia (40.5%), United States (27.6%), and United Kingdom (14.3%). To distinguish between uniform and skewed cluster distribution across these multi-lineage serovars, we compared the Shannon entropy of each clinically significant serovar at the most stable T. Of the clinically significant serovars, Bareilly, Javiana, Mississippi, Montevideo, Muenchen, Newport, Oranienburg, Paratyphi B var. Java, Saintpaul, and Thompson more closely approximated uniform cluster distribution indicating that the multi-lineage segregation of these serovars was not due to rare or undersampled distant strains (Fig. S3). Moreover, it was observed that numerous serovars were assigned to the same single-linkage clusters at the most stable T, as only 890/1,342 (66.3%) genomic clusters were found to consist of a single in-silico serovar. Examples included inter-serogroup clustering of serovars Bredeney (B), Give (E1), Kimuenza (B) and Schwarzengrund (B) and serogroup B clustering of serovars Brandenburg, Reading, Madras, and Sandiego.

6.2. Characterization of genomically similar serovars

To comprehensively characterize a list of serovars that share highly similar core genomes, we analyzed the genomics clusters predicted at the earliest signature of stability (T = 45) defined by the first occurrence of 5 successive T with nAWC > 0.99. In this region, a total of 23,748 clusters were predicted, of which 23,584 (99.31%) clusters contained isolates of a single in-silico serovar. Subsequent investigation of the other 164 (0.69%) clusters that were compositionally mixed led to the identification of a total of 65 genomically similar serovar pairs. Of note, low confidence in-silico serovar predictions were ignored when determining compositionally mixed clusters to avoid the inference of artificial relationships due to erroneous in-silico serotyping. Monophasic/diphase variants of Typhimurium and Paratyphi B/Paratyphi B variant Java were amongst the identified serovar pairs that shared highly similar core genomes. To evaluate the correlation between antigenic and genomic variations, the antigenic formulas of the identified genomically similar serovar pairs were
compared. If the antigenic loci and genomic variations are correlated, one should expect the identified serovar pairs to share similar antigenic combinations given that the pairs are clustered based on their core genomes. Given this rationale, we identified 19 serovar pairs that differed in serogroups and 4 serovar pairs that differed by two or more antigenic types (Fig. 2). However, there were numerous genetically similar serovar pairs previously reported to share identical pulse-field gel electrophoresis patterns, but were not identified in this study. It was discovered that the insensitivity is due to consistent discrepancies between SISTR serovar predictions and the reported serovars in the metadata (Table S2). Notably, SISTR is unable to distinguish between the O-6 antigen serovar pairs reported by Mikoleit et al. (32) and consequently led to its inability to detect these genetically similar O-6 antigen serovar pairs.

Further phylogenetic analysis of two genomically similar serovar pairs (London/Winterthur and Chester/Madras) that demonstrated unexpected clustering did not reveal evidence of a recent common descent for isolates of the same serotype. Phylogenomic analysis of all publicly available genomes for the serovar pair London (n = 412) and Winterthur (n = 2) revealed significant genetic divergence amongst serovar London and that the Winterthur isolates have likely descended from an ancestor common to a specific London lineage; however, the two Winterthur isolates were segregated and did not share a common branch in the phylogeny (Fig. S3). Similarly, the analysis of the serovar pair Chester (n = 511) and Madras (n = 6) revealed the divergence of Madras into two distant lineages of which only one appeared compositionally homogeneous (Fig. S4). Our observations of dissimilar antigens synthesized by genomically similar isolates and the inconsistent phylogenetic groupings amongst antigenically identical isolates represent examples in which predicted serovars do not represent observed phylogenetic relationships.

6.3. Characterization of putative non-monophyletic serovars

It is widely described in the literature that serovars have distinctive ecological and clinical features. Moreover, it has been demonstrated that serovars are genetically consistent based on MLST with a few exceptions. However, these observations have not been evaluated at a larger scale and thus may not be representative of the true genetic diversity of each serovar. Hence, to evaluate the correlation between in-silico serotyping and the lineage divergences inferred by WGS, we systematically compared the genomic similarity amongst isolates of the same in-silico serovar assignments to identify putative polyphyletic serovars and confirm whether isolates of the same serovar are in fact more similar than isolates of other serovars. Of the 738 serovars predicted by SISTR in the study dataset, 201 serovars were represented by a single sequence in the dataset and hence could not be evaluated.

We searched for putative non-monophyletic serovars by identifying occurrences in which inter-serovar comparisons yielded greater sequence identity than intra-serovar comparisons at any $T$ between the earliest signature of stability ($T = 45$) and the maximum threshold ($T = 3,246$). By this approach, we found 265/537 (49.3%) putative non-monophyletic serovars including Enteritidis, Typhimurium, Newport, Kentucky, and Javiana. Subsequent manual literature curation indicated that 18% of the evaluated serovars had been previously reported to be polyphyletic or paraphyletic (5, 33-41). Furthermore, the evaluation of the maximum distance required to link genomes of the same in-silico serovar
assignments in a single cluster revealed genomes of 209/537 (38.9%) serovars required a distance of >2,000 allelic differences to be linked in a single cluster (Fig. 3). The higher than previously reported prevalence of putative polyphyletic serovars and the multitude of multi-lineage serovars demonstrated from our collective analyses express the complex nature of the evolutionary histories of *Salmonella* which cannot be comprehensively depicted by antigenic determinants. Moreover, there is also evidence to suggest that our estimation of the proportion of putative non-monophyletic serovars likely represents an underestimation of the true proportion, as serovar sample size was found to be a statistically significant predictor of non-monophyly (Fig. S5).

6.4. Analysis of cluster destabilizing thresholds

Significant changes in nAWC by one unit incremental change in T signify the consolidation of diverged lineages that differed in a set of loci. Hence, the analysis of converging clusters at cluster destabilizing thresholds can inform evolutionary-driven adaptations and divergence between diverged *Salmonella* subpopulations. Here, we analyzed a single cluster destabilizing threshold at $T = 157$ allelic differences and determined that the decrease in nAWC was due to the convergence of two Enteritidis clusters named Cluster 3139 ($n = 10,546$) and Cluster 3140 ($n = 21,398$). This was further confirmed by recomputing nAWC at $T = 157$ with the exclusion of the Enteritidis isolates from the predicted clusters which recovered an nAWC above 0.99. Comparing the geographic distribution of the two Enteritidis clusters revealed a clear geographical segregation of the two lineages of which Cluster 3139 isolates predominated in Asia (log ratio: 3.13), South America (log ratio: 3.87) and Oceania (log ratio: 4.16), while Cluster 3140 isolates predominated in North America (log ratio: 2.08) (Fig. 4A). As an orthogonal validation, principal component analysis (PCA) of the two clusters’ cgMLST profiles similarly revealed the genetic divergence of the two subpopulations that can be explained by the first principal component alone (Fig. 4B). The inclusion of the second principal component revealed the partition of Cluster 3140 into additional subpopulation structures that are suggestive of greater diversification in Cluster 3140 isolates compared to Cluster 3139 which is concordant with previous observations by Zheng et al. (42).

Subsequent analysis of the principal component loadings identified 284 different alleles involving 168 loci that showed significant contributions (loading threshold > 0.025) to the observed genetic segregation of the two Enteritidis subpopulations. An independent analysis of core genome allelic association to cluster identity using Scoary suggested a higher number of alleles in which 388 alleles involving 183 loci demonstrated significant statistical association (adjusted p-value < 0.05) to Enteritidis subpopulation classification. High concordance was observed between the two methods, as all 168 loci predicted by PCA were also identified in the core genome allelic association analysis. Notably amongst the significant loci, the cpsG gene matched the minimal SNP-based subtyping scheme previously described by Zheng et al. (42) who proposed the differentiation of the two major Enteritidis lineages based on two genes (cpsG and citT) in the *Salmonella* genome (42). Mann-Whitney U-test and DAVID were used to assess functional and pathway enrichment of significant genes which revealed enrichment in COG category E (amino acid transport and metabolism) and membrane protein genes, respectively (Fig. 4C and 4D). Targeted assessment of
Comparison of SISTR in-silico serotyping to reported serovars

The increased adoption of WGS for pathogen characterization and the unsustainable costs associated with traditional serology have prompted the replacement of traditional serotyping with in-silico serotyping (43). However, both in-silico serotyping tools such as SISTR (12) or SeqSero2 (44) and the serovar reports in public sequence databases are not flawless due to variable sequencing data quality and human errors, leading to the dissemination of false information and misinterpretations. Misclassification by in-silico serotyping could be a major source of error in this study leading to the overestimation of genetic variability of serovars and hence there is a critical need for a systematic evaluation of in-silico serotyping. According to a previous study that evaluated the performance of in-silico serotyping and the reliability of public database metadata across 42,400 genomes (45), it was reported that SISTR predictions showed 91.91% concordance rate with reported serovars. Moreover, the results discordance was largely attributed to misreports in records and seldom in-silico misclassifications (45).

Here, we took the opportunity to evaluate 102,857 public Salmonella genomes archived with serovar information to assess the consistency between SISTR in-silico serotyping and reported serotypes found in public databases. Based on strict string matching, an overall 90% concordance between predicted serovars and reported serovars was observed. We noticed a higher rate of discrepancy in lower quality genome assemblies (< 297/330 hard core genes) in which discrepancies were detected in 32% of lower quality genomes compared to 9.6% in high quality genomes. Further analysis of the discrepancies in high quality genomes revealed sources of inconsistencies unrelated to data quality. 67% of the observed serovar discrepancies in high quality genomes occurred due to the lack of data harmonization for serovar information in public sequence databases. The inconsistent use of different phrases to describe the same serovar and spelling errors were frequently observed. A prominent example is the serovar monophasic variant of Typhimurium which was reported in 8 different ways in the metadata including “I 1,4,[5],12:i:-”, “typhimurium monophasic”, “typhimurium - monophasic”, “4,[5],12:i:-”, “1,4,[5],12:i:-”, “I 4,5,12:i:-”, “4,5,12:i-”, and “4,5,12:i-”. However, the remaining 33% of the discrepancies are likely unrelated to data quality or semantics, and instead related to potential errors in in-silico serovar prediction or reports in public data. Examples include the aforementioned O-6 antigen serovar pairs and 1,238 other unique mismatches listed in Table S4. Following the correction for discrepancies due to data quality and semantics, the overall concordance rate between serotyping reports and predictions increased to 96%, suggesting that misclassification by in-silico serotyping has a minor effect on our analyses.

7. Discussion
From the core genome allele comparisons of a diverse *Salmonella* dataset of 180,098 genomes, we demonstrated that the global population structure of the pathogen can be arranged into 1,342 stable subpopulations. We confirmed the validity of our population structure analysis based on in-silico serovar classifications in which a crude correlation between serotyping and genomic variation was observed. Moreover, the observation that many serovars subdivided into multiple lineages was consistent with our expectation, considering that the WKL scheme was intended to provide a low resolution typing for *Salmonella* (5). However, the inconsistencies of *Salmonella* serotyping became apparent when the extent of genetic variability in individual serovars was evaluated. Collectively, we highlighted two major types of serotyping inconsistencies: 1) segregation of closely related strains into unrelated antigenic groups and 2) inter-serovar genetic distance exceeding that of intra-serovar comparisons. Importantly, our observations suggested that the inconsistent serovars do not represent groupings of genetically related isolates and thus their corresponding isolates likely do not share the same epidemiological or clinical implications. Our findings are concordant with a previous report that described the failure of the serovar Senftenberg to represent a classification of genetically related isolates (46). Specifically, different subsets of the serovar Senftenberg isolates were reported to share either the same sequence type (ST14) with the serovar Westhampton, or the same sequence type (ST185) with another serovar, Dessau (46). For future work, we suggest a systematic analysis of serotyping inconsistencies in a phylogenetic context to confirm the non-monophyletic nature of the serovars that we have identified solely based on genetic distances.

Indeed although a correlation exists between antigenic and genetic variations, caution must be made surrounding specific serovars to avoid misinterpretations of genetic, biological and epidemiological features based on serotyping. It can be postulated that the occasional discrepancy between antigenic and genome-wide variation is attributable to the homologous recombination of genes involved in antigen synthesis which confounds the inference of genealogy. Recently, it has been reported that the O-antigen biosynthesis genes are situated in a genomic island with distinctive GC content rendering it a hot spot for homologous recombination and horizontal gene transfer (47). The discovery of over 200 serovars that had not been reported as putative non-monophyletic classifications is further indicative of the negligence of genomic investigations for many serovars thus far. Moreover, previous studies often selected a minor subset of genomes to represent a given serovar leading to insufficient coverage of genetic diversity and consequently the assumption of monophyly (48, 49). Hence, the collective analysis of a highly genetically diverse dataset on the scale of hundreds of thousands of genomes remains valuable to deduce a more holistic view of bacterial population structures.

To contribute to the knowledge base of *Salmonella* serovar classification, we compiled a public resource (Table S5) that documents the putative non-monophyletic serovars, as well as the serovars that belong to the same lineage as the putative non-monophyletic serovars. The deficiencies of serotyping highlighted in this study can be addressed by WGS given its capability to survey genome-wide markers and thereby empower genomics researchers to model the divergence of bacterial populations with greater precision and resolution. By subdividing *Salmonella* into 1,342 groups based on the most stable T, a genomic-based typing scheme can be developed. A single sequence can be selected from
each stable cluster to construct a set of reference sequences that would collectively represent the global *Salmonella* population structure. Subsequent alignments of query sequences to the reference sequences would thus inform the subtypes of the query genomes based on the most similar reference.

In fact, a similar solution has already been implemented in Enterobase that integrates the application of cgMLST thresholds to define stable *Salmonella* clusters represented by hierarchical cluster codes (16). The Enterobase cgMLST thresholds have been specifically chosen with the intention to reflect natural *Salmonella* subpopulations such as subspecies, superlineages, and eBGs (16). The integration of hierarchical cluster code assignments in Enterobase represents an initial step towards the establishment of a standardized WGS-based typing scheme for *Salmonella*. However, a global community consensus on a standardized genomic-based typing scheme remains critical to encourage the usage of a consistent language to share *Salmonella* population dynamics across jurisdictions and international borders. Moreover, community consensus of a core or whole genome MLST scheme is also necessary to converge on a fixed set of genome-wide loci for comparison and reinforce a standardized allele number nomenclature. In particular, the Chewie Nomenclature Server (https://chewbbaca.online/) could be a highly promising avenue to synchronize local schema with globally maintained schema and centralize a collection of gene-by-gene subtyping schema for microbial pathogens (50).

The importance of standardization and data harmonization cannot be undermined, as the lack of controlled vocabularies in open data settings hinders data interpretability and summarization. The high degree of metadata inconsistencies in public foodborne pathogen databases is evident from the frequent discrepancies between SISTR predictions and reported serovars solely due to semantics. The unrestricted entry of serovar information in GenomeTrakr and PubMLST hinders an accurate estimation of the number of serovars in the study dataset based on the reported serovars in the metadata alone. Moreover, it renders the estimated 90% concordance rate between SISTR predictions and reported serovars to likely represent an underestimation of the true concordance rate. The integration of controlled vocabularies for subtype information in public databases would be highly valuable for future studies that intend to analyze public *Salmonella* sequences by eliminating the necessity for time-consuming data harmonization and thus reinforcing analysis reproducibility.

To demonstrate the additional utility of nAWC to reveal cluster consolidation dynamics and inform evolutionary histories, we analyzed the genomic differences between two geographically segregated Enteritidis lineages that converged at a cluster destabilizing threshold ($T = 157$). The pronounced allele frequency bias across different geographical regions demonstrated from genome-wide association analyses suggested that the observed lineage divergence could be a consequence of niche adaptation or founder effects. A previous study on the evolutionary history of *Salmonella* Enteritidis had estimated the most recent common ancestor of present-day Enteritidis lineages to have existed in the 1600s when cross-continental overseas trades were increasingly embraced (51). Hence, an ancestral European Enteritidis subpopulation may possibly have migrated to America during the era of colonial trade and evolved in reproductive isolation leading to the fixation of the lineage-specific alleles observed in Cluster 3139. However, the enrichment of allelic bias in metabolic and membrane transport functions could reflect the local adaptation of the
Enteritidis subpopulations to the variabilities in regional nutrient availability, a phenomenon that has been previously reported in a metabolic analysis of Enteritidis lineages in high- and low-income settings (52). We recommend a future analysis of in-vitro metabolic activities of the two Enteritidis populations and correlating differential metabolic activities to the divergence of lineage-specific alleles to test our hypotheses.

In the genomics era, the development of efficient bioinformatics tools and high-throughput sequencing have revolutionized our understanding about microbial evolution and biology. Here, we continued to echo the value of understanding by proposing novel ways to define the global Salmonella population structure and discovering novel genetic patterns that differentiate microbial subpopulations. Our study is a further demonstration of how global genomic data sharing benefits scientific research. The open accessibility of large-scale genomic datasets with minimal metadata fields will foster new directions to investigate molecular evolution and microbial diversity. With increasing integration of genomics in infectious disease surveillance worldwide, we encourage joint efforts across scientific research and public health to standardize a genomic-based typing nomenclature for global disease reporting such that Salmonella classifications consistently reflect both genetic relatedness and epidemiological significance.

8. Author Statements

8.1. Authors and contributors

C.L.: Conceptualization, formal analysis, methodology, validation, software, visualization, funding acquisition, and writing – original draft

W.H.: Project administration, conceptualization, supervision, funding acquisition, and writing – review and editing

8.2. Conflicts of interest

The authors declare that there are no conflicts of interest.

8.3. Funding information

This work has been supported by grants to W.H. from the Canadian Institutes of Health Research (Project Grant Number: PJT-159456), the Michael Smith Foundation for Health Research Scholar Program, and Genome BC and Genome Canada (Grant Number: 286GET). Computational resources were provided by Compute Canada via Canada Research Chair and Canadian Foundation for Innovation grants to Dr. Jennifer Gardy. C.L. was supported by the Canadian Graduate Scholarship – Master’s award and the Natural Sciences and Engineering Research Council of Canada - Collaborative Research and Training Experience program.

8.4. Acknowledgements

The authors thank Zohaib Anwar for his constructive feedback on the original manuscript and the SISTR developers: John Nash, Ed Taboada, and James Robertson for their insightful discussions on the subject matter. The authors express gratitude to all the contributors to GenomeTrakr and PubMLST whose data enabled this large scale analysis.
References

1. Marus JR, Magee MJ, Manikonda K, Nichols MC. Outbreaks of Salmonella enterica infections linked to animal contact: Demographic and outbreak characteristics and comparison to foodborne outbreaks-United States, 2009-2014. Zoonoses Public Health. 2019 Jun;66(4):370–6.

2. Eng S-K, Pusparajah P, Ab Mutalib N-S, Ser H-L, Chan K-G, Lee L-H. Salmonella: A review on pathogenesis, epidemiology and antibiotic resistance. Frontiers in Life Science. 2015 Jul 3;8(3):284–93.

3. Dalyan Cilo B, Karakeçili F, Güleşen R, Levent B, Ozakin C, Gedikoğlu S. Determination of Salmonella serotypes by conventional and molecular methods. Mikrobiyol Bul. 2013 Oct;47(4):693–701.

4. Wattiau P, Boland C, Bertrand S. Methodologies for Salmonella enterica subsp. enterica subtyping: gold standards and alternatives. Appl Environ Microbiol. 2011 Nov;77(22):7877–85.

5. Achtman M, Wain J, Weill F-X, Nair S, Zhou Z, Sangal V, et al. Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. PLoS Pathog. 2012 Jun 21;8(6):e1002776.

6. Sévellec Y, Vignaud M-L, Granier SA, Lailler R, Feurer C, Le Hello S, et al. Polyphyletic Nature of Salmonella enterica Serotype Derby and Lineage-Specific Host-Association Revealed by Genome-Wide Analysis. Front Microbiol. 2018 May 17;9:891.

7. Zheng J, Luo Y, Reed E, Bell R, Brown EW, Hoffmann M. Whole-Genome Comparative Analysis of Salmonella enterica Serovar Newport Strains Reveals Lineage-Specific Divergence. Genome Biol Evol. 2017 Apr 1;9(4):1047–50.

8. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet. 2018 Apr;14(4):e1007261.

9. Pearce ME, Alikhan N-F, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. Int J Food Microbiol. 2018 Jun 2;274:1–11.

10. Radomski N, Cadel-Six S, Cherchame E, Felten A, Barbet P, Palma F, et al. A Simple and Robust Statistical Method to Define Genetic Relatedness of Samples Related to Outbreaks at the Genomic Scale - Application to Retrospective Salmonella Foodborne Outbreak Investigations. Front Microbiol. 2019 Oct 24;10:2413.

11. Mohammed M, Thapa S. Evaluation of WGS-subtyping methods for epidemiological surveillance of foodborne salmonellosis. One Health Outlook. 2020 Jul 6;2:13.

12. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, 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. 2016 Jan 22;11(1):e0147101.

13. Barker DOR, Carriço JA, Kruczkiewicz P, Palma F, Rossi M, Taboada EN. Rapid Identification of Stable Clusters in Bacterial Populations Using the Adjusted Wallace Coefficient [Internet]. bioRxiv. 2018 [cited 2021 Sep 29]. p. 299347. Available from: https://www.biorxiv.org/content/10.1101/299347v1
Available from: http://dx.doi.org/10.1128/MRA.01450-18

Nodes of Complete Bugarel Park issue): W169–75.

Resources:

- Huang 2007 Aug;24(8):1586–91.
- Yang 15;31(22):3691–3.
- Brynildsrud Bioinformatics. 2008 Jun 1;24(11):1403–5.
- Jombart with profiles instead of a distance matrix. Mol Biol Evol. 2009 Jul;26(7):1641–50.
- Price 30;5(2):R12.
- Versatile Kurtz microbial tree of life. Sci Rep. 2020 Feb 3;10(1):1723.
- Shakya pathogens. Genome Res. 2018 Sep;28(9):1395–404.
- GrapeTree: Zhou Bioinformatics. Springer Berlin Heidelberg; 2008. p. 113–22.
- Silva Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, et al. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microb Genom [Internet]. 2018 Mar;4(3). Available from: http://dx.doi.org/10.1099/mgen.0.000166
- Simonsen M, Mailund T, Pedersen CNS. Rapid Neighbour-Joining. In: Algorithms in Bioinformatics. Springer Berlin Heidelberg; 2008. p. 113–22.
- Zhou Z, Alikhan N-F, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Res. 2018 Sep;28(9):1395–404.
- Shakya M, Ahmed SA, Davenport KW, Flynn MC, Lo C-C, Chain PSG. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Sci Rep. 2020 Feb 3;10(1):1723.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004 Jan 30;5(2):R12.
- Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009 Jul;26(7):1641–50.
- Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008 Jun 1;24(11):1403–5.
- Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016 Nov 25;17(1):238.
- Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015 Nov 15;31(22):3691–3.
- Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug;24(8):1586–91.
- Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W169–75.
- Park J, Taslim C, Lin S. BOG: R-package for Bacterium and virus analysis of Orthologous Groups. Comput Struct Biotechnol J. 2015 May 21;13:366–9.
- Bugarel M, Cook PW, den Bakker HC, Harhay D, Nightingale KK, Loneragan GH. Complete Genome Sequences of Four Salmonella enterica Strains (Including Those of Serotypes Montevideo, Mbandaka, and Lubbock) Isolated from Peripheral Lymph Nodes of Healthy Cattle. Microbiol Resour Announc [Internet]. 2019 Jan;8(2). Available from: http://dx.doi.org/10.1128/MRA.01450-18
30. Centers for Disease Control and Prevention (CDC) National Salmonella Surveillance Annual Report, 2016. Atlanta, Georgia: National Salmonella Surveillance Annual Report, 2016: US Department of Health and Human Services CDC 2018; 2018.
31. Peters TM, Berghold C, Brown D, Coia J, Dionisi AM, Echeita A, et al. Relationship of pulsed-field profiles with key phage types of Salmonella enterica serotype Enteritidis in Europe: results of an international multi-centre study. Epidemiol Infect. 2007 Nov;135(8):1274–81.
32. Mikoleit M, Van Duyne MS, Halpin J, McGlinchey B, Fields PL. Variable expression of O:61 in Salmonella group C2. J Clin Microbiol. 2012 Dec;50(12):4098–9.
33. Achtman M, Zhou Z, Alikhan N-F, Tyne W, Parkhill J, Cormican M, et al. Genomic diversity of Salmonella enterica -The UoWUCC 10K genomes project. Wellcome Open Res. 2020;5:223.
34. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, et al. Phylogenetic diversity of the enteric pathogen Salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. Genome Biol Evol. 2013;5(11):2109–23.
35. 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 [Internet]. 2018 Nov 27;9(6). Available from: http://dx.doi.org/10.1128/mBio.02303-18
36. Yin Z, Liu J, Du B, Ruan H-H, Huo Y-X, Du Y, et al. Whole-Genome-Based Survey for Polyphyletic Serovars of Salmonella enterica subsp. enterica Provides New Insights into Public Health Surveillance. Int J Mol Sci [Internet]. 2020 Jul 23;21(15). Available from: http://dx.doi.org/10.3390/ijms21155226
37. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ. 2016 Apr 5;4:e1752
38. Xu F, Ge C, Luo H, Li S, Wiedmann M, Deng X, et al. Evaluation of real-time nanopore sequencing for Salmonella serotype prediction. Food Microbiol. 2020 Aug;89:103452.
39. Hudson LK, Constantine-Renna L, Thomas L, Moore C, Qian X, Garman K, et al. Genomic characterization and phylogenetic analysis of Salmonella enterica serovar Javiana. PeerJ. 2020 Nov 20;8:e10256.
40. Old DC, Rankin SC, Crichton PB. Assessment of strain relatedness among Salmonella serotypes Salinatis, Duisburg, and Sandiego by biotyping, ribotyping, IS200 fingerprinting, and pulsed-field gel electrophoresis. J Clin Microbiol. 1999 Jun;37(6):1687–92.
41. Sévellec Y, Vignaud M-L, Granier SA, Lailler R, Feurer C, Le Hello S, et al. Polyphyletic Nature of Salmonella enterica Serotype Derby and Lineage-Specific Host-Association Revealed by Genome-Wide Analysis. Front Microbiol. 2018 May 17;9:891.
42. Zheng J, Pettengill J, Strain E, Allard MW, Ahmed R, Zhao S, et al. Genetic diversity and evolution of Salmonella enterica serovar Enteritidis strains with different phage types. J Clin Microbiol. 2014 May;52(5):1490–500.
43. Banerji S, Simon S, Tille A, Fruth A, Flieger A. Genome-based Salmonella serotyping as the new gold standard. Sci Rep. 2020 Mar 9;10(1):4333.

44. Zhang S, den Bakker HC, Li S, Chen J, Dinsmore BA, Lane C, et al. SeqSero2: Rapid and Improved Salmonella Serotype Determination Using Whole-Genome Sequencing Data. Appl Environ Microbiol [Internet]. 2019 Dec 1;85(23). Available from: http://dx.doi.org/10.1128/AEM.01746-19

45. Robertson J, Yoshida C, Kruczkiewicz P, Nadon C, Nichani A, Taboada EN, et al. Comprehensive assessment of the quality of Salmonella whole genome sequence data available in public sequence databases using the Salmonella in silico Typing Resource (SISTR). Microb Genom [Internet]. 2018 Feb;4(2). Available from: http://dx.doi.org/10.1099/mgen.0.000151

46. Mooijman KA. The fifteenth CRL-Salmonella workshop : 27 June 2010, Saint Malo, France. 2011 Jan 10 [cited 2021 Sep 29]; Available from: http://www.rivm.nl/bibliotheek/rapporten/330604019.html

47. Seif Y, Monk JM, Machado H, Kavvas E, Palsson BO. Systems Biology and Pangenome of Salmonella O-Antigens. MBio [Internet]. 2019 Aug 27;10(4). Available from: http://dx.doi.org/10.1128/mBio.01247-19

48. Timme RE, Pettengill JB, Allard MW, Strain E, Barrangou R, Wehnes C, et al. Phylogenetic diversity of the enteric pathogen Salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. Genome Biol Evol. 2013;5(11):2109–23.

49. Sharma P, Dahiya S, Balaji V, Kanga A, Panda P, Das R, et al. Typhoidal Salmonellae: Use of Multi-Locus Sequence Typing to Determine Population Structure. PLoS One. 2016 Sep 12;11(9):e0162530.

50. Mamede R, Vila-Cerqueira P, Silva M, Carriço JA, Ramirez M. Chewie Nomenclature Server (chewie-NS): a deployable nomenclature server for easy sharing of core and whole genome MLST schemas. Nucleic Acids Res. 2021 Jan 8;49(D1):D660–6.

51. Deng X, Desai PT, den Bakker HC, Mikoleit M, Tolar B, Trees E, et al. Genomic epidemiology of Salmonella enterica serotype Enteritidis based on population structure of prevalent lineages. Emerg Infect Dis. 2014 Sep;20(9):1481–9.

52. Feasey NA, Hadfield J, Keddy KH, Dallman TJ, Jacobs J, Deng X, et al. Distinct Salmonella Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings. Nat Genet. 2016 Oct;48(10):1211–7.

9. **Figure Legends**

**Figure 1.** Unrooted Neighbour-joining tree constructed from 3,246-loci cgMLST profiles of 180,098 Salmonella enterica genomes with color annotations reflecting the combination of serovar and genomic cluster classifications. The tree is visualized using GrapeTree (43).

**Figure 2.** Antigenic formula comparisons of 65 genomically similar serovar pairs.

**Figure 3.** Distribution of monophyletic and non-monophyletic serovars by the maximum cgMLST allelic distance required to merge all genomes of the same serovar in a single-linkage cluster.
**Figure 4.** Genomic analysis of geographically segregated Enteritidis subpopulations. A) Geographical distribution of Enteritidis isolates classified in Clusters 3139 and 3140 with the proportion calculated by (isolate count/total Enteritidis isolates count in the dataset). B) PCA analysis of the cgMLST profiles of isolates in the two subpopulations. C and D) Functional annotation enrichment in the top COG categories and UniProt keywords ranked by adjusted p-values. E) dN/dS ratio of the membrane protein genes found to exhibit lineage-specific variation between the two Enteritidis subpopulations.

**Figure S1.** nAWC (A) and Shannon entropy (B) plots calculated across the range of all possible cgMLST distance thresholds.

**Figure S2.** Unrooted neighbour-joining tree constructed from 3,246-loci cgMLST profiles of 180,098 *Salmonella enterica* genomes with color annotations reflecting serovar classifications only. The tree is visualized using GrapeTree (43).

**Figure S3.** Unrooted maximum likelihood tree of serovars London and Winterthur genomes visualized in GrapeTree. The branch lengths in the visualization are not to scale of the actual lengths.

**Figure S4.** Unrooted maximum likelihood tree of serovars Chester and Madras genomes visualized in GrapeTree. The branch lengths in the visualization are not to scale of the actual lengths.

**Figure S5.** Sample size density distribution of predicted monophyletic and non-monophyletic serovars. The summary table shows various statistics including the statistical significance and regression coefficient of serovar sample size as a predictor of monophyly based on logistic regression.
| H1 Antigen | H2 Antigen | Serogroup |
|------------|------------|-----------|
| lubbock    | bareilly   | mbandaka  |
| hartford   | i c1:y:-   | i c1:y:-   |
| io:67:m,t,e,n,z15 | madras | edinburg  |
| brazzaville | i h:y:-   | hindmarsh |
| carrau     | bovismorbificans | lattenkamp |
| i o:51:z4,z24:- | heidelberg | langensalza |
| i c1:l,z13:- | i 1,4,[5],12:r:- | i 1,4,[5],12:r:- |
| bonn       | tamberma   | richmond  |
| i c1:z10:- | richmond   | kenya     |
| i c1:z4,z23:- | mbandaka | i c1:l,v:- |
| i c1:z41:e,n,z15 | orion    | i 1:4:1:- |
| elisabethville | c1:c:-    | typhimurium |
| i e1:y:-   | choleraesuis | t:-,l,v: |
| choleraeaeusis | richmond | winterthur |
| i 1,4,[5],12:- | meleagridis | meleagridis |
| i c1:-,l,v:- | london | vitkin |
| bareilly   | i e1:1:-   | potsdam   |
| london     | i m:1:v:-  | ohio      |
| calabar    | i c1:1:v:- | i f:1:4:- |
| i e1:1:v:- | i c1:b:-   | newport   |
| i m:1:v:-  | i c1:e,h:- | i o:57:z10:1,7 |
| i c2-c3:e,h:- | guinea | welltvedren |
| i c2-c3:d:- | i c1:y:- | mikawasima |
| i o:59:y:1,5 | i 1,4,[5],12:d:- | muenchen |
| i 1,4,[5],12:d:- | brunei | kingabwa |
| i c1:1:-   | agouevey | stanley   |
| give       | i 1,4,[5],12:- | i r:1:- |
| macclesfield | c2-c3:y:- | i c2-c3:y:- |
| chester    | cubana     | moscow    |
| i 1,4,[5],12:- | madras | eko |
| i anatum   | infantis  | i d1:l,z28:- |
| i 1,4,[5],12:- | javiana | i d1:l,z13:- |
| i c1:r:-   | napoli    | holcomb |
| paratyphi b | i c2-c3:l:v:- | i c1:r:- |
| ibadan     | virchow    | paratyphi b var. java |
| enteritidis | mississippi | moscow |
| blockley   | haardt    | paratyphi b var. java |
| koesssen   | panama    | moscow    |
| bergen     | i o:51:i:e,n,z15 | mississippi |
| i c1:k:-   | thompson  | moscow    |
| altona     | kentucky  | i c2-c3:1:v:- |
| ib: e,h:- | saintpaul | paratyphi b var. java |
| brandenburg | i b:1:v:- | kentucky |
| i c2-c3:1:- | reading | reading |
| i b: e,h:- | i e1:e,fi:- | macclesfield |
| enteritidis | i e1:e,fi:- | macclesfield |
| term            | estimate | std.error | statistic | p.value |
|-----------------|----------|-----------|-----------|---------|
| Intercept       | 0.26853  | 0.09557   | 2.80993   | 0.00496 |
| Sample Size     | -0.00214 | 0.00053   | -4.07289  | 0.00005 |

**Density**

- Monophyletic
- Non-monophyletic

**Log transformed sample size**