frontiers in Microbiology

ORIGINAL RESEARCH
published: 31 July 2017
doi: 10.3389/fmicb.2017.01345

Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar

Chad R. Laing*, Matthew D. Whiteside and Victor P. J. Gannon

National Microbiology Laboratory, Public Health Agency of Canada, Lethbridge, AB, Canada

Food safety is a global concern, with upward of 2.2 million deaths due to enteric disease every year. Current whole-genome sequencing platforms allow routine sequencing of enteric pathogens for surveillance, and during outbreaks; however, a remaining challenge is the identification of genomic markers that are predictive of strain groups that pose the most significant health threats to humans, or that can persist in specific environments. We have previously developed the software program Panseq, which identifies the pan-genome among a group of sequences, and the SuperPhy platform, which utilizes this pan-genome information to identify biomarkers that are predictive of groups of bacterial strains. In this study, we examined the pan-genome of 4893 genomes of *Salmonella enterica*, an enteric pathogen responsible for the loss of more disability adjusted life years than any other enteric pathogen. We identified a pan-genome of 25.3 Mbp, a strict core of 1.5 Mbp present in all genomes, and a conserved core of 3.2 Mbp found in at least 96% of these genomes. We also identified 404 genomic regions of 1000 bp that were specific to the species *S. enterica*. These species-specific regions were found to encode mostly hypothetical proteins, effectors, and other proteins related to virulence. For each of the six *S. enterica* subspecies, markers unique to each were identified. No serovar had pan-genome regions that were present in all of its genomes and absent in all other serovars; however, each serovar did have genomic regions that were universally present among all constituent members, and statistically predictive of the serovar. The phylogeny based on SNPs within the conserved core genome was found to be highly concordant to that produced by a phylogeny using the presence/absence of 1000 bp regions of the entire pan-genome. Future studies could use these predictive regions as components of a vaccine to prevent salmonellosis, as well as in simple and rapid diagnostic tests for both *in silico* and wet-lab applications, with uses ranging from food safety to public health. Lastly, the tools and methods described in this study could be applied as a pan-genomics framework to other population genomic studies seeking to identify markers for other bacterial species and their sub-groups.

Keywords: genomics, pan-genome, *Salmonella*, predictive markers, food safety

OPEN ACCESS

Edited by:
Sandra Torriani,
University of Verona, Italy

Reviewed by:
Jinshui Zheng,
Huazhong Agricultural University, China
Dapeng Wang,
Shanghai Jiao Tong University, China

****Correspondence:***
Chad R. Laing
chadr.laing@canada.ca

Specialty section:
This article was submitted to Food Microbiology, a section of the journal Frontiers in Microbiology

Received: *20 February 2017*
Accepted: *03 July 2017*
Published: *31 July 2017*

Citation:
Laing CR, Whiteside MD and Gannon VPJ (2017) Pan-genome Analyses of the Species Salmonella enterica, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. Front. Microbiol. 8:1345. doi: 10.3389/fmicb.2017.01345

INTRODUCTION

The global burden of bacterial enteric disease, much of it foodborne, results in an estimated 2.2 million deaths per year, and an annual loss of 112,000 disability adjusted life years in the United States alone (Bergholz et al., 2014; Scallan et al., 2015). Nationwide molecular diagnostic networks, such as PulseNet in North America, were designed to enable the rapid identification of outbreaks by genetic fingerprinting the etiological agents of disease, and keeping nationwide databases of genetic fingerprints of specific pathogens associated with human disease. Since its inception, PulseNet has relied on pulsed-field gel electrophoresis (PFGE) for fingerprinting of bacterial pathogens to identify the specific sources of outbreaks and prevent further infections. Using this approach, it has been estimated that PulseNet prevents 277,000 illnesses from bacterial pathogens annually in the United States, reducing the costs associated with medical care and loss of productivity due to worker illness (Scharff et al., 2016).

Despite the usefulness of PulseNet, the PFGE technique itself is often unable to distinguish between related and unrelated strains, due to its reliance on rare-cutting restriction enzyme sites within the genome (Allard et al., 2012). Additionally, the interpretation of the banding patterns among labs requires extensive training and standardization to enable meaningful comparisons. Lastly, the banding patterns provide no information on the actual content of the genomes they represent, so important information regarding human virulence, such as the presence or absence of known toxins, is not available.

Lastly, while the presence of known virulence factors has been correlated with severe human disease in a number of bacterial species, it has also been shown that some lineages or clades within these same species, while possessing specific virulence factors, are rarely associated with human disease (Lupolova et al., 2016; Waryah et al., 2016). Thus, multiple virulence factors, and regulatory genes that influence the expression of key virulence factors, or otherwise modulate the virulence of these strains, need to be taken into consideration when attempting to predict the strains of a bacterial species that are potential human health threats (Opijnen et al., 2012).

Recently, whole-genome sequencing (WGS) has displaced PFGE as the *de facto* standard for the complete characterization of bacterial pathogens, in both ongoing surveillance and outbreak investigations (Deng et al., 2016; Franz et al., 2016). WGS allows clear definition between outbreak-related strains and those from unrelated sources, and it has the ability to identify routes of transmission, and attribute bacterial contaminants to specific sources (den Bakker et al., 2014). It is currently being utilized in reference laboratories worldwide. Examples of its application include the sequencing of all *Listeria monocytogenes* isolated in the United States, all *Salmonella* isolated by the Food and Drug Administration in the USA, and by Public Health England as part of routine surveillance (Ashton et al., 2016), and a large-scale survey of *Staphylococcus aureus* in continental Europe. In the latter study, the applicability of WGS for the identification of the emergence and spread of clinically relevant *Staphylococcus aureus* was demonstrated (Aanensen et al., 2016).

It has also recently been shown that antimicrobial resistance (Tyson et al., 2015; McDermott et al., 2016; Zhao et al., 2016), serovar (Levine et al., 2016; Yoshida et al., 2016b), and the results of other traditional sub-typing schemes such as multi-locus sequence typing (Sheppard et al., 2012) can be accurately predicted *in silico* through the analysis of bacterial genome sequences. However, identifying bacterial isolates that are most likely to cause disease in humans, based on the genome sequence alone, is a more complex task. In addition, markers that can identify bacteria likely to exhibit particular phenotypes, such as the ability to survive in a particular niche, or the ability to tolerate harsh environments such as those found in food processing plants are also required.

We have previously developed the software platform Panseq, for the analyses of thousands of genomes in a pan-genome context, where both the presence/absence of the accessory genome and SNPs within the shared core-genome are computed (Laing et al., 2010). Additionally, we recently released a platform for the predictive genomics of *Escherichia coli*, called SuperPhy, in which markers statistically biased within groups of bacteria, based on any metadata category, can be identified (Whiteside et al., 2016).

In this study we use our previously created software to examine the pan-genome of *Salmonella enterica*, a pathogen that causes an estimated 93.8 million cases of enteric illness worldwide each year (Majowicz et al., 2010; Gal-Mor et al., 2014). The species *S. enterica* is divided into six subspecies: *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae*, and *indica*. Over 99% of human disease caused by *S. enterica* is done so by subspecies *enterica*, with the World Health Organization estimating that *S. enterica* infections from contaminated food alone constitute a loss of 6.43 million disability adjusted life years worldwide, more than any other enteric pathogen (Kirk et al., 2015). Within this bacterial subspecies, are human-adapted strains responsible for typhoid fever, as well as a large number of animal-derived non-typhoidal strains responsible for foodborne illness. In this study, we have identified species- and subspecies-specific markers, as well as markers predictive of serovar for subspecies enterica. While this study focused on *S. enterica*, the tools and approach are broadly applicable to any species or collection of genomes.

MATERIALS AND METHODS

All commands and parameters used to analyze the data and generate the Figures are available as Supplementary File 1. The scripts used for analyses are available at https://github.com/superphy/gamechanger. The following is a summary of the methods used.

Data Collection

All *S. enterica* genomes were downloaded from GenBank in nucleotide fasta format. A full listing of the initial 4939 genomes, including GenBank identifier, subspecies, serovar, the number of species-specific core regions present, the number of contigs, and whether the genome passed the quality filtering steps are listed in Supplementary File 2.

Serovar Identification

Most of the *S. enterica* genomes in GenBank had serovar provided as part of their metadata; however, 321 were missing this designation. The SISTR web-server, as well as the SISTR commandline app were used to predict the serovar for these strains (Yoshida et al., 2016b).

Pan-genome Analyses

Panseq (commit:1d0ab9d37e8e358d266e1d0aa80e9b27f28a1def) was used to identify the pan-genome of the 4939 strains in this study (Laing et al., 2010). Genomes were initially fragmented into 1000 bp segments, and subsequently clustered using cd-hit v.4.6 to remove potential duplicates/paralogs from the analyses using a 90% sequence identity threshold (Fu et al., 2012). Initially Panseq was used to determine the distribution of the pan-genome among the genomes at a 90% sequence identity threshold, from which a "conserved core" was identified. Within the conserved core, Panseq was then used to identify single-nucleotide polymorphisms.

Identification of *S. enterica* Species-Specific Regions

To identify regions that were likely to represent the species as a whole, we initially examined the 211 closed *S. enterica* genomes in GenBank (Supplementary File 2), and identified 3832 regions of 1000 bp that were found in 90% (190) of the 211 closed genomes using Panseq, at a 90% sequence identity threshold. These regions were then screened against the online GenBank nr database using megablast as a first-pass filter with default parameters, searching across bacteria (taxid:2), and excluding all *Salmonella* (taxid:590) hits that had greater than 80% identity across 80% of the query length from the results. The remaining 1482 genomic regions were subsequently screened against the online GenBank nr database of all bacteria (taxid:2), using the blastn algorithm, to identify matches that were missed using the less-specific megablast algorithm, with word size 11, an e-value cutoff of 0.001, and excluding all *Salmonella* (taxid:590). These results were filtered in the same manner, leaving 405 potentially species-specific regions. Lastly, these regions were compared against *Salmonella bongori* genomes in GenBank; one *S. bongori* hit was identified, which left 404 genomic regions present in *S. enterica* but no other bacterial genomic sequences within the GenBank nr database.

The putative function of these regions was determined by screening them across the GenBank nr database using blastx with "max hits:10," "taxid limit:1236 (gammaproteobacteria)," and an "e-value threshold: 0.001." The best matching hit above a 90% sequence identity threshold was used for the putative functional assignment.

Identification of Subspecies- and Serovar-Specific Regions

The Fisher's Exact test, using the Bonferroni correction for multiple testing was applied as in the SuperPhy platform (Whiteside et al., 2016), implemented here as the standalone program feht[1]. The input for the program was Supplementary File 2, which contained metadata for all the strains, as well as the binary_table.txt output file from the Panseq analyses, which denotes the presence/absence of each 1000 bp pan-genome region among all the strains.

S. enterica Phylogenetic Analyses

The phylogeny based on SNPs within the core genome was generated using RAxML v8.2.9, with the snp.phylip output file from Panseq (Stamatakis, 2014). The phylogeny based on the presence/absence of the pan-genome was also generated using RAxML v8.2.9, with the binary.phylip output file from Panseq.

Generation of Figures and Tables

The R-statistical language v3.3.2 was used to generate the summary Figures and Tables (R Core Team, 2016). The R-scripts and all others used for the analyses can be found at https://github.com/superphy/gamechanger/tree/master/src. The ggtree package for R was used in the generation of the phylogenetic tree images (Yu et al., 2016).

RESULTS

S. enterica Pan-genome

We initially determined the size and distribution of the *S. enterica* pan-genome as genome fragments of 1000 bp in size, across the 4939 genome sequences of this study, which are summarized by subspecies in **Table 1**, and within subspecies *enterica* by serovar in **Table 2**. As can be seen in **Figure 1**, the pan-genome comprised of 4939 *S. enterica* genomes was found to be 25.3 Mbp in size, with 70% of the pan-genome present in fewer than 100 strains. Conversely, the core genome was found to be 1.5 Mbp in size, with all but 200 genomes (96%) containing 3.2 Mbp of shared genomic core. Only 17% of the pan-genome was found in greater than 100 genomes, but fewer than 4739 genomes.

S. enterica Species-Specific Regions

To identify regions of *S. enterica* that were likely to be shared among most genomes of the species, we examined all 211 closed genomes of *S. enterica* in GenBank, looking for genomic regions that were present in at least 190 (90%) of these genomes. We

TABLE 1 | The frequency of the subspecies observed within the study set of 4936 *Salmonella enterica* genomes, prior to any quality filtering.

Subspecies	No.
enterica	4913
arizonae	7
diarizonae	7
houtenae	4
salamae	4
indica	1

[1] https://github.com/chadlaing/feht

TABLE 2 | The serovars with more than 20 representatives in the current study set of 4936 *Salmonella enterica* genomes, and their frequency, prior to any quality filtering.

Serovar	No.
Typhi	1977
Typhimurium	758
Enteritidis	413
Heidelberg	201
Paratyphi	158
Kentucky	155
Agona	136
Weltevreden	120
Bareilly	106
Newport	82
Tennessee	77
Montevideo	69
Saintpaul	48
Infantis	39
Senftenberg	35
Bovismorbificans	34
Hadar	33
Muenchen	30
Anatum	27
Schwarzengrund	27
Dublin	24
Cerro	21

The list of all serovars and their frequency within the current study is available as Supplementary File 2.

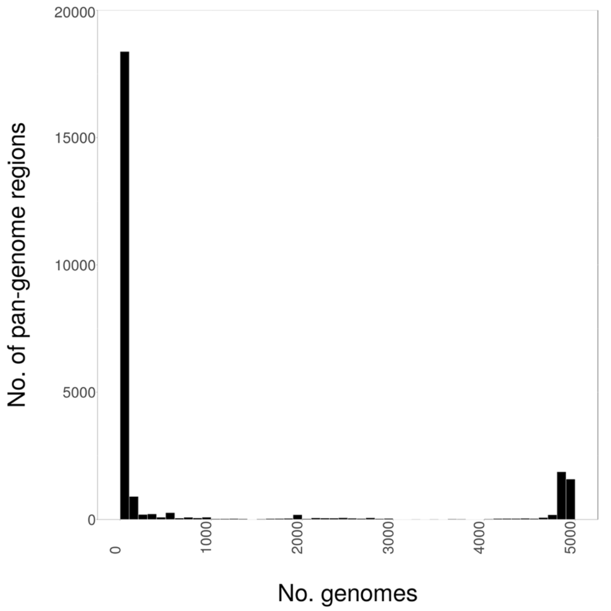

FIGURE 1 | The distribution of the *Salmonella enterica* pan-genome, as 1000 bp fragments, among 4939 whole-genome sequences (WGSs).

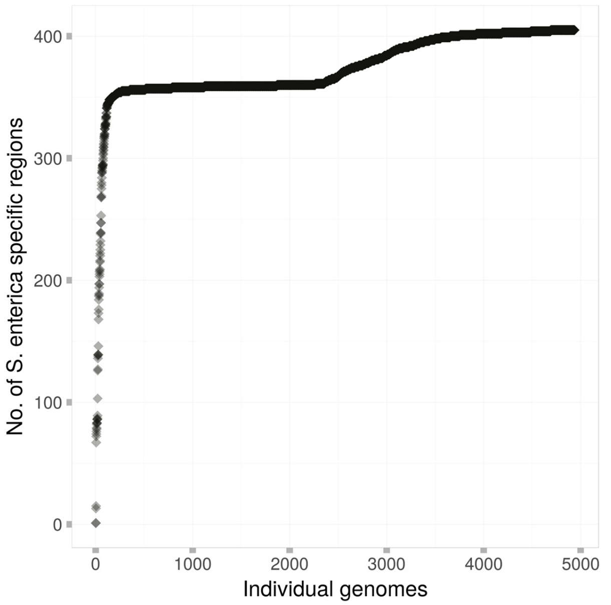

FIGURE 2 | The carriage of the 404 *S. enterica* species-specific regions among each of the 4939 genomes of this study. Each dot represents a single *S. enterica* genome, which are arranged in order from those that contain the fewest species-specific regions to those that contain the most.

identified 3832 regions of 1000 bp that were present in at least 90% of the closed genomes. These regions were subsequently screened against the GenBank nr database, and any present in non-*Salmonella* genomes were removed, leaving 404 putative *S. enterica* species-specific regions (Supplementary File 3).

Figure 2 shows the carriage of these 404 regions among the 4939 genomes of this study. All but 105 genomes contained at least 330 of these putative *S. enterica* specific regions. A stark difference in carriage of these species-specific markers was observed, with 4742 genomes containing at least 350 species-specific markers, while only 2674 genomes contained 360 or more species-specific markers.

Quality Filtering for Subsequent Analyses

To ensure the quality of the genomes in use for subsequent analyses, we plotted carriage of the 404 species-specific regions versus the number of contigs that each sequenced genome was comprised of (**Figure 3**). As can be seen, the two genomes marked in yellow contained only one, and the same, species-specific region each, despite being comprised of relatively few contigs. Subsequent searches against the GenBank nr database identified these two genomes as *Citrobacter* spp. contamination, mislabeled as *S. enterica* (GCA_001570325 and GCA_001570345). The "*Salmonella enterica* species-specific region" found in both of the contaminant *Citrobacter* genomes, did not match any other *Citrobacter* spp. in GenBank above the thresholds used for determining presence/absence in this study. However, due to the presence of this region in what have been identified as *Citrobacter* genomes, the region was removed from subsequent analyses.

The majority of genomes (4913) were from subspecies *enterica*, with genomes from the five other *S. enterica* subspecies present in drastically fewer numbers (**Table 1**). All closed

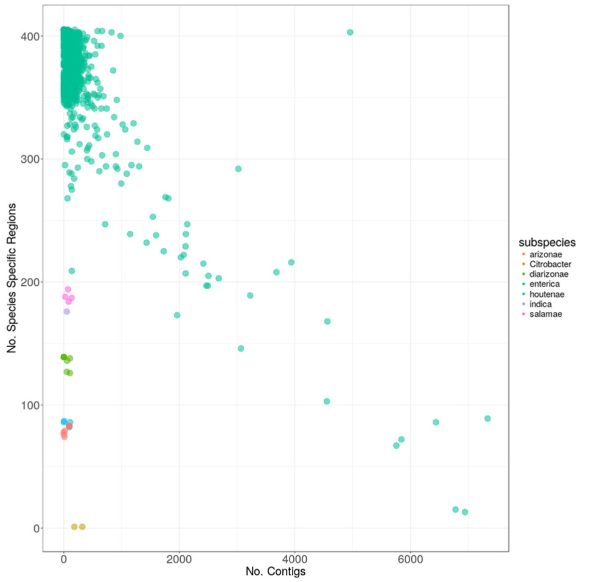

FIGURE 3 | The carriage of the 404 *S. enterica* species-specific regions, versus the number of contigs for each of the 4936 genomes. Colors indicate the subspecies within *S. enterica* as follows: red: arizonae, lime: diarizonae, teal: enterica, blue: houtenae, lavender: indica, magenta: salamae and yellow: sample with *Citrobacter* contamination.

genomes from subspecies *enterica* contained greater than 250 species-specific regions, which was more than the genomes from any other subspecies, with the exception of *enterica* genomes that were of poor quality and comprised of many 1000s of contigs (**Figure 3**). Genomes from subspecies *houtenae* and *arizonae* contained fewer than 100 species-specific regions, while genomes from *diarizonae*, *indica*, and *salamae* contained between 100 and 200 species-specific regions. All regions were screened against *S. bongori* to ensure specificity to *S. enterica*; one region was found to also be present in genomes from *S. bongori* and was removed from further analyses.

Within subspecies *enterica*, a negative linear relationship was observed among the number of species-specific regions contained within a genome, and the number of contigs the genome was comprised of, with the worst-case genome (GCA_000495155) being comprised of 6945 contigs, but containing only 13 species-specific regions. Other genomes such as *S. enterica* Bovismorbificans strain GCA_001114865 contained both few contigs (140) as well as fewer species-specific regions (209) than other *enterica* genomes. Additional searches discovered sequencing gaps within the genome totaling over 464 Kbp. A final outlier genome harbored nearly 5000 contigs, but also contained 403 of the species-specific regions. It was determined by searching the GenBank database, that this sequence (GCA_000765055) was actually a combination of multiple genomes in a single file.

Given the above information, all genomes from the five subspecies other than *enterica* were included in subsequent analyses, while the thresholds for inclusion of *enterica* genomes were set at a maximum of 1000 contigs, and a minimum of 250 species-specific regions. Following this quality filtering, 43 genomes were removed, leaving 4870 *S. enterica* subspecies *enterica* genomes for the following analyses.

Phylogeny of *S. enterica* Using the Conserved Core Genome

Based on the distribution of the pan-genome presented in **Figure 1**, the "conserved core" of *S. enterica* was set at being present in more that 4500 genomes, to fully capture the conserved genomic regions within the species. A phylogeny based on the SNPs among these shared regions was created, and is shown along with the distribution of the *S. enterica* species-specific regions in **Figure 4**. As can be seen, the majority of the genomes are subspecies *enterica*, and the other five subspecies are relatively more distant in the order of *indica*, *salamae*, *houtenae*, *diarizonae*, and *arizonae*. However, the order of subspecies in declining number of species-specific regions is: *enterica*, *diarizonae*, *salamae*, *indica*, *houtenae*, and *arizonae*, which is shown in **Figure 3**.

The serovar distribution within subspecies *enterica* was shown to be largely concordant with phylogeny, as demonstrated in **Figure 5**, where the 10 most abundant serovars in the current study are highlighted. However, not all serovars clustered as monophyletic groups, as can be seen with serovar Bareilly; nor were all clades found to be comprised of single serovars, demonstrated by the clade containing genomes of serovars Bareilly and Agona.

The large clades within the phylogenetic tree also demonstrate clade-specific patterns of presence/absence for the 404 species-specific markers. Among the most abundant serovars, Typhimurium, Heidelberg, Newport, and Enteritidis were found to contain the most species-specific markers, and grouped together near the center of the tree. Likewise, serovars Agona, Welevreden, and Kentucky contained fewer species-specific regions, and group together near the bottom of the tree, closer to the non-*enterica* sub-species genomes.

Table 3 considers all serovars with at least 10 members in the dataset, and the average number of species-specific markers per serovar. As can be seen, the serovars with the largest average number of species-specific regions were: Enteritidis (401.7), Anatum (401.5), Muenchen (400.5), Hadar (400.3), and Typhimurium (400.1); conversely, the serovars with the fewest average number of species-specific regions were: Derby (360.7), Montevideo (360.1), Typhi (358.1), Bovismorbificans (355.3), and Cerro (342.0).

Phylogeny of *S. enterica* Using the Pan-genome

A phylogeny based on the presence/absence of the pan-genome among the 4893 *S. enterica* genomes was created, and is shown along with the distribution of the *S. enterica* species-specific regions in **Figure 6**. As can be seen this phylogeny based on the presence/absence of the entire 25.3 Mbp pan-genome is highly concordant with the phylogeny based on the SNPs found in the conserved core of the same strains (**Figure 5**). In both trees the serovars cluster together and in the same relation to each

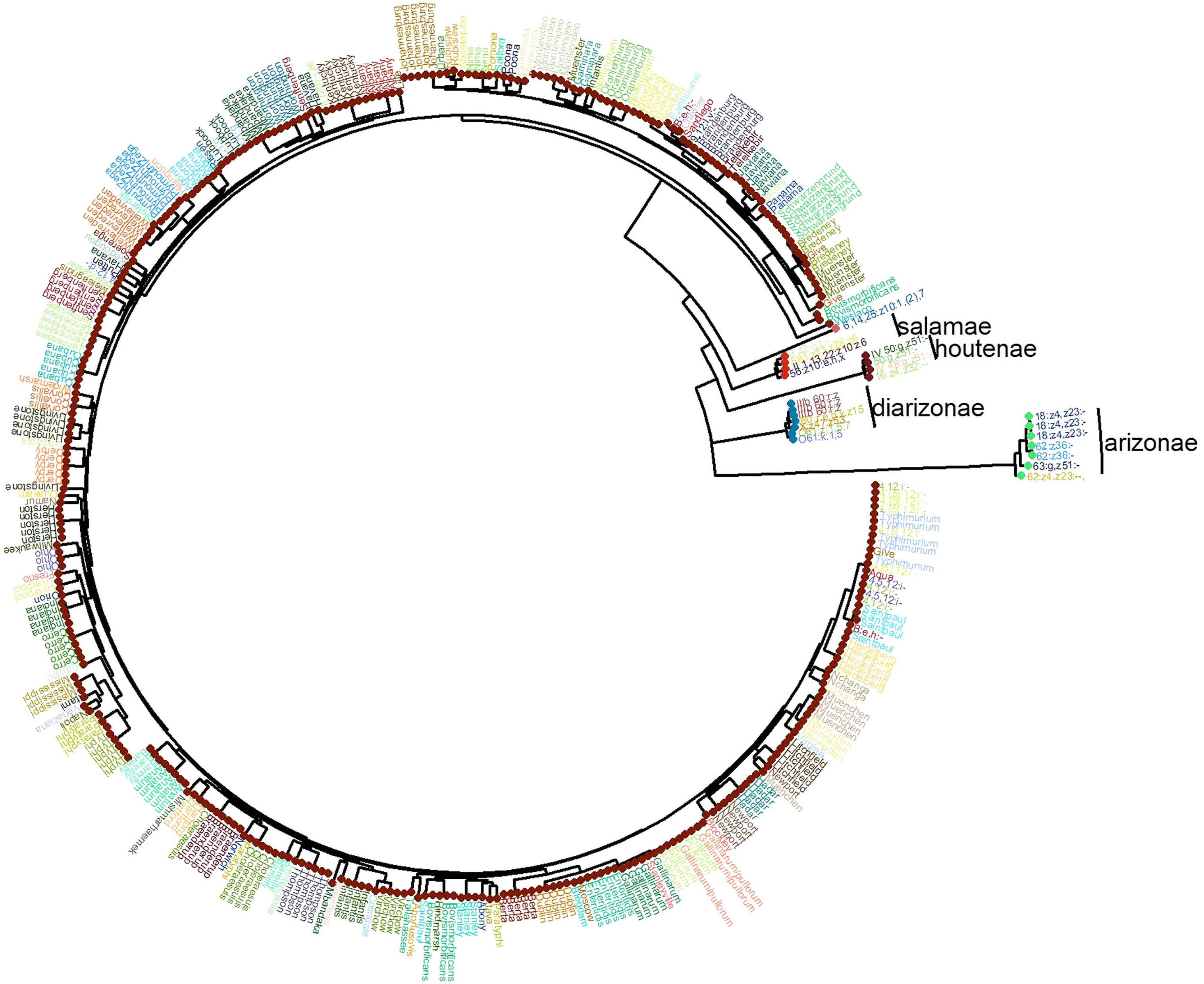

FIGURE 4 | The phylogeny of the 4893 *S. enterica* genomes post quality-filtering, and limiting the number of genomes from each serovar to five. The name of each serovar is presented as text, and the six subspecies are shown as colored circles as follows: teal: arizonae, blue: diarizonae, dark orange: enterica, peach: houtenae, dark green: indica, light orange: salamae.

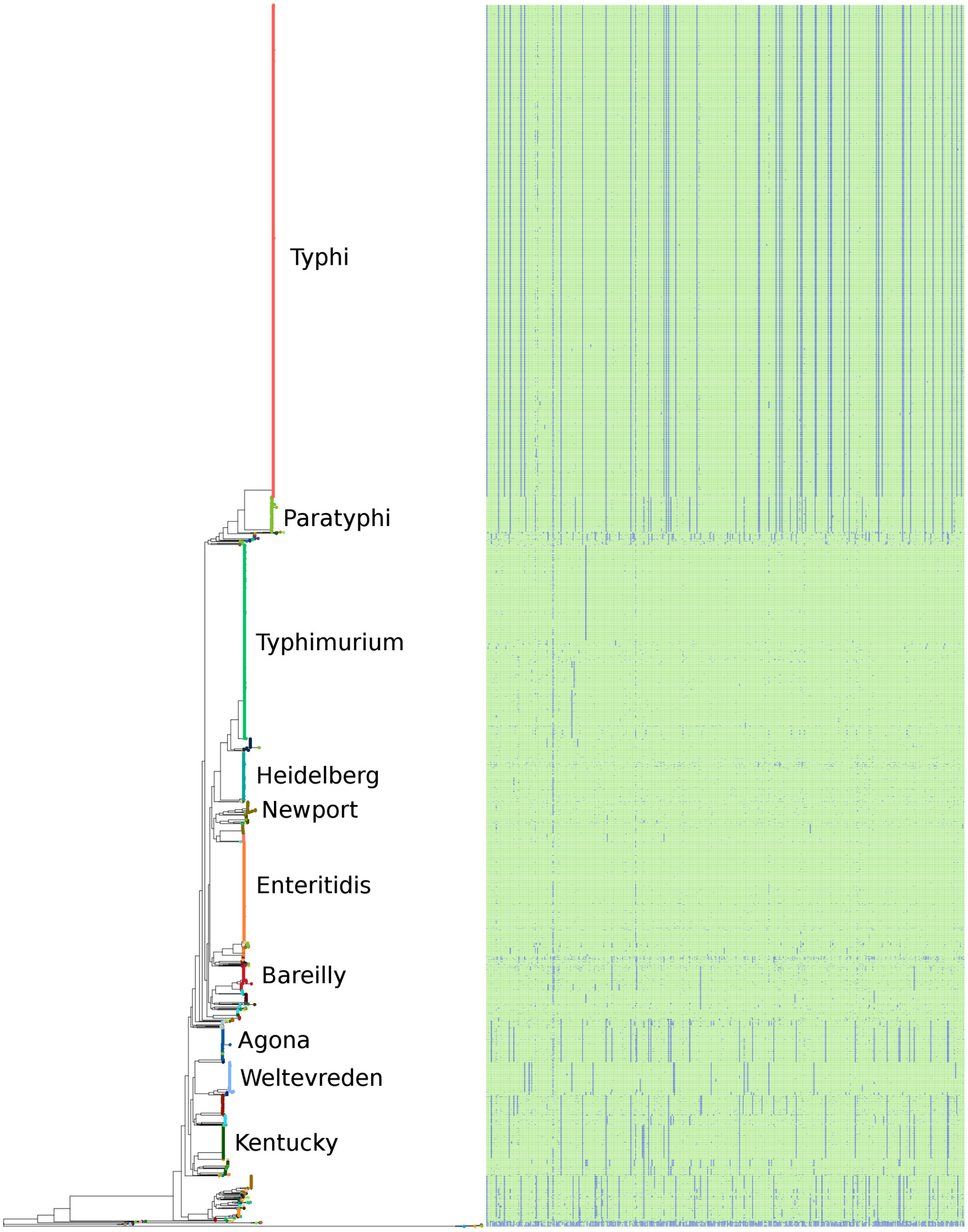

FIGURE 5 | The phylogeny of the 4893 *S. enterica* genomes post quality-filtering based on SNPs found within the conserved core genome. The 10 most abundant serovars of subspecies enterica in the current study (Agona, Bareilly, Enteritidis, Heidelberg, Kentucky, Newport, Paratyphi, Typhi, Typhimurium, Weltevreden) are labeled on the tree. The matrix to the right of the phylogeny represents the 404 species-specific regions, with blue being the absence of a region, and green being the presence of a region, for each of the genomes of the study.

other, for example serovars Typhi and Paratyphi strains form a discrete monophyletic clade. However, the branch lengths in the pan-genome tree are larger than those in the conserved SNP tree, due to the larger variation among the presence/absence of the pan-genome than to sequence variation among shared core regions.

Identification of a Minimum Set of Species-Specific Genomic Markers

Within the 404 species-specific markers, none were specific for any of the subspecies. That is, a marker was always present in genomes from at least two subspecies.

TABLE 3 | The average number of species-specific genomic regions found among serovars of subspecies *enterica*, that contained at least 10 representative genomes, within the 4870 quality filtered subspecies *enterica* genomes of this study.

Serovar	Average no. species-specific regions
Enteritidis	401.7
Anatum	401.5
Muenchen	400.5
Hadar	400.3
Typhimurium	400.1
Newport	399.8
Thompson	399.7
Saintpaul	399.6
Heidelberg	397.4
Dublin	395.2
Infantis	394.9
Braenderup	392.8
Weltevreden	390.0
Bareilly	388.5
Kentucky	380.3
Plymouth/Zega	377.9
Senftenberg	376.5
Mbandaka	374.5
Lubbock	374.1
Reading	370.4
Agona	369.5
Tennessee	368.3
Schwarzengrund	362.3
Paratyphi	361.5
Derby	360.7
Montevideo	360.1
Typhi	358.1
Bovismorbificans	355.3
Cerro	342.0

We next determined that the presence of a minimum set of two genomic regions was required to unambiguously identify genomes of *S. enterica*, within the 4893 genomes of the current study. A combination of two genomic regions were all that was required, and two such markers that were also present in the most *S. enterica* genomes were found at the following locations within the Typhimurium reference genome LT2: (1336001.. 1337000) and (2467001.. 2468000) (Supplementary File 3). All members of *S. enterica* examined contained at least one of these markers, but many other combinations within the 404 species-specific markers are also possible.

Putative Functional Identification of the *S. enterica* Species-Specific Regions

The putative function of the 404 quality-filtered *S. enterica* species-specific regions were determined from the GenBank nr database. The annotation of each of the 404 regions is available as Supplementary File 1. **Table 4** summarizes the frequency of functional annotation categories, after annotating each region with the single best match. As can be seen, hypothetical proteins accounted for the majority (64) of the 404 annotations, with secreted effector and membrane proteins being the next most frequent category among the species-specific regions. Other membrane, transport, and secretion proteins were observed. The species-specific regions also included proteins involved in core metabolic functions, protein and DNA synthesis, and response to stress.

Identification of Subspecies-Specific Markers from the Pan-genome

Having identified species-specific markers, we employed the same techniques, utilizing the presence/absence of all pan-genome markers, just as was carried out in identifying the 404 species-specific ones, to identify subspecies-specific markers. The number of markers that were completely unique to a subspecies is given in **Table 5**. Subspecies *arizonae* contained the most unique markers, at 207, and *enterica* contained the least, at 9.

Identification of Universal Serovar Markers within *Subspecies enterica* from the Pan-genome

Subspecies *enterica* genomes were the vast majority of those available, so we attempted to identify serovar-specific markers for the top 10 serovars, in the same manner that we identified subspecies-specific markers. We found that there were no genomic markers that uniquely defined any of the serovars based on their presence or absence; however, there were a number of genomic regions that were universally present or absent among serovars, as well as statistically over- or under- represented with respect to all other serovar genomes from this study; they are shown in **Table 6**.

To further assess the validity of these markers, a dataset comprised of 3948 genomes from EnteroBase[2], was selected to have an identical number of strains belonging to each of nine serovars in our GenBank dataset. The EnteroBase dataset was used to test the predictive markers we identified from the GenBank dataset in the first part of the study. The results of this comparison are shown in **Figure 7**. As can be seen, the markers were well-conserved among the EnteroBase dataset, with eight of the nine serovars having a subset of the predictive markers present among all of the test genomes; serovar Typhimurium had a marker subset that was present in all but one of the test genomes.

DISCUSSION

S. enterica Pan-genome

Previous examinations of the *S. enterica* pan-genome were based on relatively small datasets of 45 and 73 genomes (Jacobsen

[2] https://enterobase.warwick.ac.uk/species/index/senterica

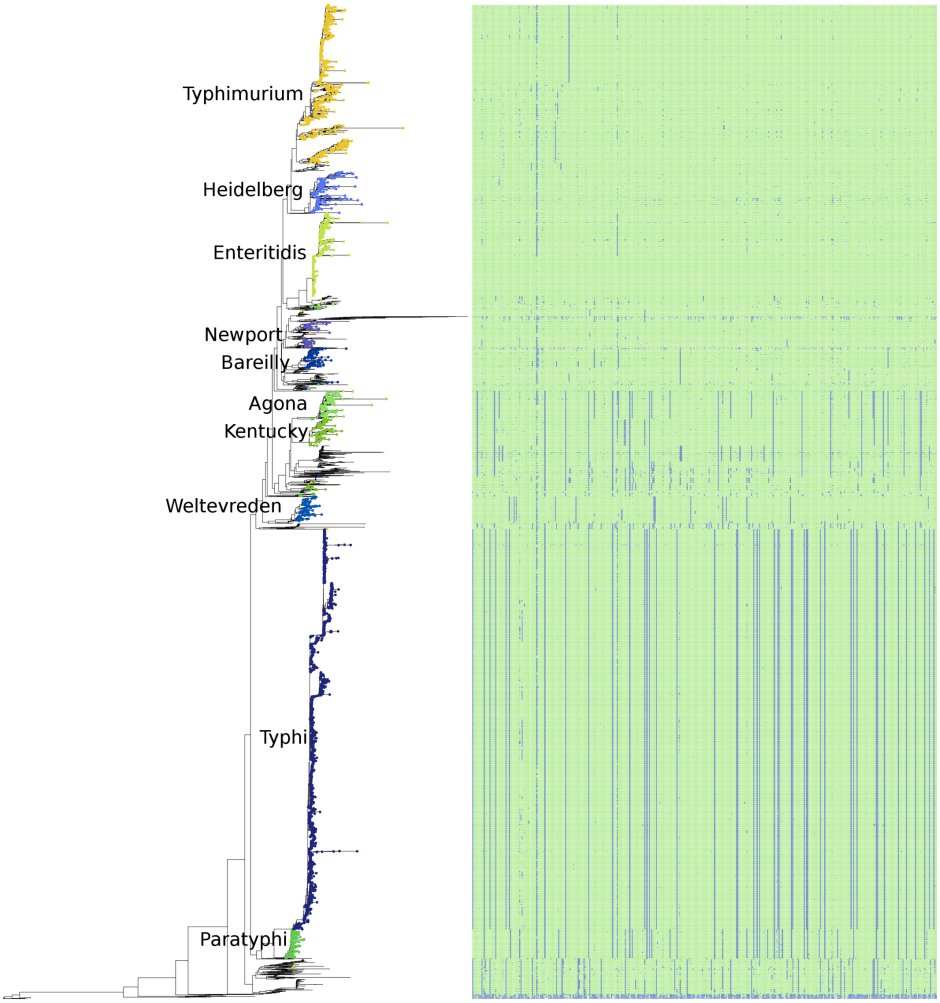

FIGURE 6 | The phylogeny of the 4893 *S. enterica* genomes post quality-filtering based on the presence/absence of the entire pan-genome as 1000 bp fragments. The 10 most abundant serovars of subspecies enterica in the current study (Agona, Bareilly, Enteritidis, Heidelberg, Kentucky, Newport, Paratyphi, Typhi, Typhimurium, Weltevreden) are labeled on the tree. The matrix to the right of the phylogeny represents the 404 species-specific regions, with blue being the absence of a region, and green being the presence of a region, for each of the genomes of the study.

et al., 2011; Leekitcharoenphon et al., 2012). While others have analyzed 1000s of *S. enterica* genomes, the analyses were not conducted to examine the population structure. For example, in demonstrating the software program Roary, 1000 *S.* Typhi genomes were used to test the program (Page et al., 2015). Likewise, the GenomeTrackR project utilized 32 *S. enterica* genomes to identify a *S. enterica* core, which was subsequently used as the basis for genetic distance estimates for nearly 20,000 genomes (Pettengill et al., 2016).

Previous estimates placed the core-genome size of *S. enterica* at ∼2800 gene families, and the pan-genome at ∼10,000 gene families (Jacobsen et al., 2011). The current study identified a strict core of 1.5 Mbp, and a conserved core of 3.2 Mbp shared among 96% of the genomes, which given an average gene size of 1000 bp is ∼1500 and ∼3200 genes respectively, with a much larger pan-genome of ∼25,300 genes. Previous analyses found *S. enterica* to have a closed pan-genome (Jacobsen et al., 2011), and thus the rate of discovery for new genomic regions would decrease for each new genome of the species sequenced (Tettelin et al., 2005).

In line with *S. enterica* having a closed pan-genome, when we compared it to *E. coli*, a related bacterial species with an open pan-genome (Tettelin et al., 2005), we found that the *E. coli* pan-genome was larger (37.4 Mbp), despite the fact that the *E. coli* study used less than half the number of strains in the current *Salmonella enterica* study. Additionally, more of the pan-genome of *S. enterica* was distributed among more genomes than in *E. coli* (Whiteside et al., 2016). Specifically, in *S. enterica* 70% of the pan-genome was found to belong to 100 or fewer of the genomes examined, while in *E. coli* 80% of the pan-genome was found in 100 or fewer genomes.

It should be noted that erroneously labeled, and poor quality assemblies, can greatly affect the size, analyses, and composition of the pan-genome. Software tools to evaluate assembly quality

TABLE 4 | The putative function of the *S. enterica* species-specific regions for functions that were identified more than once, utilizing the best hit for each region.

Putative protein function	Frequency
Hypothetical	64
Secreted effector	10
Membrane	7
Secretion system apparatus	5
Uncharacterized	5
Fimbrial	5
Pathogenicity island 2 effector	4
Fimbrial assembly	4
Outer membrane usher	4
mfs transporter	3
Oxidoreductase	3
Histidine kinase	3
Putative inner membrane	3
Putative cytoplasmic	3
lysr family transcriptional regulator	3
Transcriptional regulator	2
Permease	2
Outer membrane	2
Type III secretion	2
Phosphoglycerate transport	2
arac family transcriptional regulator	2
Conserved hypothetical	2
Methyl-accepting chemotaxis	2
Hybrid sensor histidine kinase/response regulator	2
Glycosyl transferase, partial	2
Phenylacetaldehyde dehydrogenase	2
Pathogenicity island 1 effector	2
n-Acetylneuraminic acid mutarotase, partial	2
Type III secretion system	2
Transcriptional regulator, partial	2
Cytoplasmic	2
Fimbrial chaperone	2
Putative sialic acid transporter	2

The complete list of all putative functions is available as Supplementary File 3.

TABLE 5 | The number of subspecies-specific pan-genome markers that were universally present or absent among members of the subspecies, and not absent or present among genomes from any other subspecies.

Subspecies	No. markers
arizonae	207
diarizonae	93
enterica	9
houtenae	134
indica	192
salamae	135

TABLE 6 | The number of pan-genome regions that were universally present and absent, as well as statistically over- or under-represented in comparison to all other genomes, within the 10 most abundant serovars within the 4870 subspecies *enterica* genomes of this study.

Serovar	No. universally present	No. universally absent
Typhi	288	2720
Typhimurium	41	698
Enteritidis	18	440
Heidelberg	121	840
Paratyphi	65	202
Kentucky	177	331
Agona	161	638
Weltevreden	426	608
Bareilly	87	436
Newport	226	360

have been created to help researchers identify bad data. These include QUAST (Gurevich et al., 2013), which summarizes the assembly statistics including average contig size and number of contigs; as well as CGAL (Rahman and Pachter, 2013), which uses a likelihood approach to infer assembly quality rather than summary statistics. As demonstrated in the current study, having a known set of species-specific genome regions can facilitate rapid quality assessment and filtering of genome assemblies. Others have proposed whole-genome MLST for this purpose as well (Babenko et al., 2016; Yoshida et al., 2016b), but the benefit of a pan-genome analysis is that it is schema free, requiring no agreed upon reference set or central repository of alleles.

S. enterica Species-Specific Regions

Previous studies have identified gene targets that are useful in the identification of *Salmonella*. These include the *fimA* gene (Cohen et al., 1996), *hilA* (Guo et al., 2000), *invA* (Malorny et al., 2003), *ttr* (Malorny et al., 2004), and *ssaN* (Chen et al., 2010). Other markers, and combinations thereof have been developed for use in RT-PCR (Postollec et al., 2011), and other detection platforms such as loop-mediated isothermal amplification (Kokkinos et al., 2014). Additionally, the identification of serovar based on allelic variation in somatic and flagellar genes has previously been conducted, with at least four laboratory methods currently available [the *Salmonella* genoserotyping assay (Yoshida et al., 2014), and the commerical assays: *Salmonella* Serogenotyping Assay, Check&Trace *Salmonella*, and xMAP *Salmonella* serotyping assay], capable of identifying over 100 of the most common *Salmonella enterica* serovars in some cases (Yoshida et al., 2016a). The recently released software, the *Salmonella in silico* typing resource (SISTR), is capable of providing *Salmonella* serovar prediction from WGSs for 90% (2,190) of all serovars (Yoshida et al., 2016b).

Despite the utility of the previously mentioned methods, previous marker-discover studies have used at most 100s of *Salmonella* strains, while the current study examines nearly 5000. Further, the current study analyzes the entire pan-genome for predictive markers, and identified over 400 that were specific to the species, as well as others being predictive for both subspecies and serovar.

The host intestinal environment consists of a multitude of bacterial species competing for scarce nutritional sources such as carbohydrates, direct antagonistic competition with other bacterial cells, and competition for access to the host

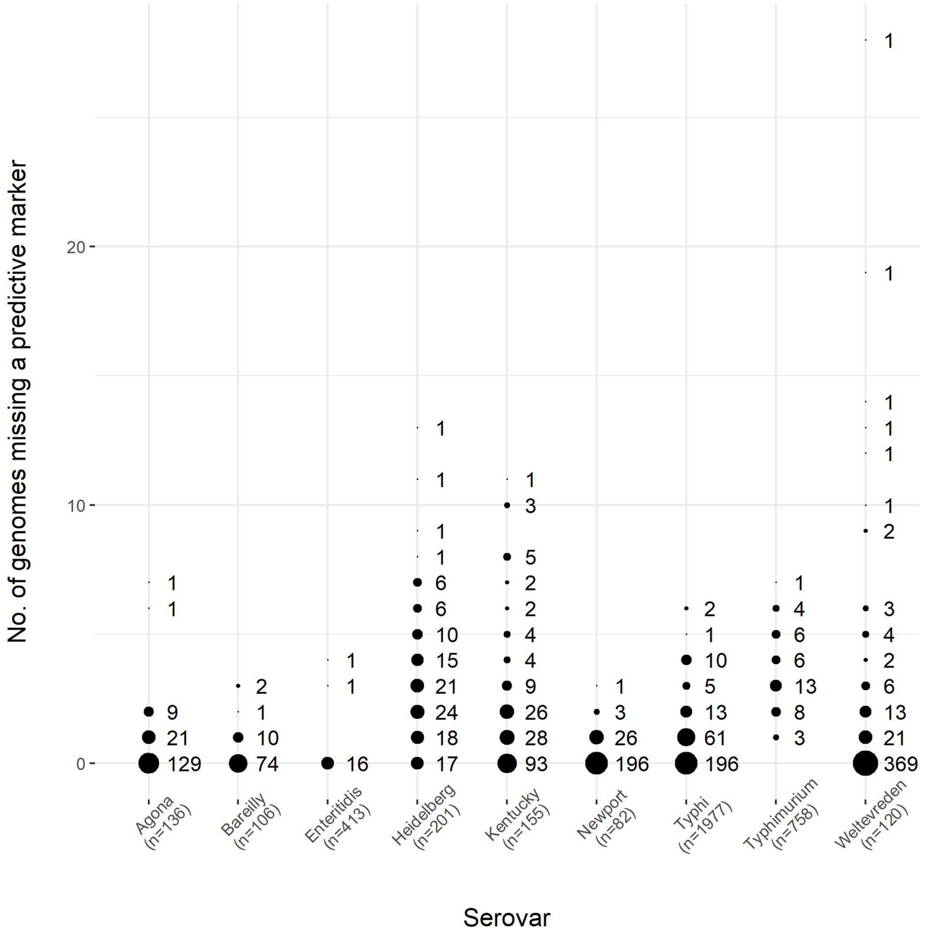

FIGURE 7 | The number of predictive markers from the GenBank dataset found within the EnteroBase dataset for nine serovars of *S. enterica*, which encompassed a test set of 3948 genomes. The number of genomes for each serovar was the same between the GenBank and EnteroBase datasets, as shown in **Table 6**. The size of the circles is proportional to the number of predictive markers from the GenBank dataset found in the EnteroBase dataset. The number of genomes for each serovar is given in the horizontal axis label. Using serovar Agona as an example, there were 136 genomes in both the GenBank and EnteroBase datasets, and 129 of the 161 predictive markers from the GenBank dataset were found in all of the genomes from the EnteroBase dataset, whereas 21 of the GenBank predictive markers were found in all but one (135) of the EnteroBase genomes examined.

intestine, where stable attachment and colonization of the local environment are possible (Sana et al., 2016). The normal intestinal microflora offer protection to the host against enteric pathogens such as *S. enterica*, but disruption of the intestinal environment by virulence factors and effector proteins secreted by the pathogen itself, or external factors including antibiotics, have been shown to alter the composition of the microbiota, and allow pathogens such as *S. enterica* to proliferate (Ng et al., 2013).

Nutritional competition exists for free metabolic compounds, such as carbohydrates that are readily available, as well as others that are sequestered in forms such as the intestinal mucus, which is composed of sialic sugar acids (McDonald et al., 2016). In the gut, these sugar acids exists as a conjugate in the alpha form, which to be useful for bacteria such as *Salmonella*, need to be converted to the beta form by a mutarotase enzyme (Severi et al., 2008). In this study, we identified n-acetylneuraminic acid mutarotase genes as species-specific genomic regions, along with sialic acid transporter genes. It is possible the presence of these systems allow *S. enterica* to more efficiently compete with the host microbiota by efficiently utilizing scarce metabolic sources.

It was also previously found that sialic acid on the surface of host colon cells increased colonization by *S.* Typhi, and disialylation of these cells reduced the adherence of the *Salmonella* strains by 41% (Sakarya et al., 2010). This was also demonstrated in *S.* Typhimurium, where following antibiotic treatment, the presence of free sialic acid increased, and the ability to utilize it was correlated with higher levels of bacterial colonization of the host gut (Ng et al., 2013).

Enzymes that utilize sialic acids have previously been shown to be present in 452 bacterial species, including other pathogens such as *Vibrio cholerae*, but the genomic regions found in the current study were sufficiently unique at the nucleotide level to be determinative for *S. enterica* (McDonald et al., 2016).

In addition to species-specific regions used to gain a metabolic advantage, a number of secretion system and effector proteins were identified as diagnostic of *S. enterica*. These included components of the Type VI secretion system (T6SS), which is a contact-dependent, syringe-like secretion system that allows *S. enterica* to directly kill other competing bacteria that it comes into physical contact with (Brunet et al., 2015), and is encoded on the *Salmonella* Pathogenicity Island 6 (Sana et al., 2016). It has been demonstrated that silencing the T6SS via H-NS repression (histone-like nucleoid structuring), reduces inter-bacterial killing of *S. enterica* (Brunet et al., 2015). It was also previously shown that commensal bacteria are killed by *S. enterica* in a T6SS-dependent manner, that the T6SS was required for *Salmonella* to establish infection in the host gut, and that increased concentrations of bile salts resulted in a concomitant increase in T6SS anti-bacterial activity (Sana et al., 2016). The T6SS itself has been shown to have been independently acquired from four separate lineages within five of the six *S. enterica* subspecies (Desai et al., 2013).

Like the T6SS, the type III secretion system (T3SS) found within *S. enterica* is a syringe like apparatus that injects effector proteins into host cells (Kubori et al., 2000). There are two T3SS found within *S. enterica*: the first is encoded on the *Salmonella* Pathogenicity Island 1 (SPI1) and is required for invasion of host cells; the second is encoded on *Salmonella* Pathogenicity Island 2 (SPI2), and is required for survival and proliferation within the host macrophage cells (Hensel et al., 1998; Bijlsma and Groisman, 2005). The innate host immune system utilizes the inflammatory response to help reduce the proliferation of bacterial pathogens (Sun et al., 2016). *S. enterica* has developed a means of regulating host inflammation via the SPI1 T3SS, whereby secreted effector proteins target the NF-κB signaling pathway, reduce inflammation and host tissue damage, and allow increased *S. enterica* propagation within the host. *S. enterica* also relies on free long-chain fatty acids within the host to regulate T3SS expression, and provide a cue to the bacteria to up-regulate genes necessary for host intestinal colonization (Golubeva et al., 2016).

The current study identified many secretion system and effector proteins as being species-specific, as well as proteins for attachment to the host, such as fimbriae. These proteins allow *S. enterica* to compete within the intestinal environment, and take up residence within the host, where it can proliferate.

Effector proteins and other virulence factors aid in the colonization of the host, and are frequently horizontally acquired and are present on mobile elements such as integrated bacteriophages (Moreno Switt et al., 2013). Previous work identified clusters of phages that carried virulence factors such as adhesins and antimicrobial resistance determinants within *S. enterica* (Moreno Switt et al., 2013).

Additionally, many of the genes associated with bacteriophage in *S. enterica* have been found to be of the putative and hypothetical class (Penadés et al., 2015). The current study identified a large accessory gene pool that contained many hypothetical and putative genes, which were also the most abundant category of species-specific genomic regions. The proteins of putative and unknown function may aid in colonizing warm-blooded animals, or specific animal or environmental niches. Previous studies identified genotype/phenotype correlations of *S.* Typhimurium that had particular gene complements associated with specific food sources (Hayden et al., 2016). The same study also postulated that specific phage repertoires may give phylogenetically distant strains a similar accessory gene content, and therefore similar niche specificity. Previously, 285 gene families were identified as being recruited into *S. enterica*, where most of these genes had unknown function, but were postulated to be important for its survival and infection of its host (Desai et al., 2013). It is therefore not surprising to find that the most abundant species-specific category of genomic regions are those of unknown or putative function; they likely represent genes enhancing the ability of *S. enterica* to propagate within warm-blooded animals, but they have not yet been fully characterized. The other genomic regions diagnostic of *S. enterica* include means for disseminating these fitness genes within the population, competing for resources in the host, and attaching and proliferating. The *S. enterica* species-specific regions likely give a good overview of the factors responsible for making it such an effective pathogen and intestinal inhabitant.

Specific Regions for Subspecies and Serovar

The current study recapitulates the phylogenetic relationship of the six *S. enterica* subspecies that has been previously described by others (Desai et al., 2013). However, the number of species-specific regions found within each subspecies does not follow the same pattern. For example, *diarizonae* is more distantly related to *enterica* than subspecies *indica*, but contains more species-specific regions, and the branch lengths on the tree are shorter. This indicates that although the *diarizonae* strains diverged longer ago than the houtenae strains, they have accumulated less genomic change. Both subspecies *diarizonae* and *houtenae* strains are associated with reptile-acquired salmonellosis (Schroter et al., 2004; Horvath et al., 2016), but the differences in genomic change may reflect the specific reptile niches that each inhabit.

Genomic regions specific to each subspecies were identified, the presence of which were unambiguously indicative of each subspecies. The most abundant subspecies in the current analyses, *enterica*, had the fewest specific markers present (9), while the most distantly related subspecies *arizonae*, had the most specific markers (207). These results indicate that just as core genome size decreases with the number of genomes examined, so too do the number of markers “core” to each subspecies. As more genomes in subspecies *arizonae* and closely related subspecies are examined, we would expect fewer genomic regions to remain specific for the subspecies. This has important implications for designing a set of markers indicative for subspecies, indicating that a group of redundant markers should be used, and that a sampling of the diversity within a subspecies is first required to identify genomic regions that are truly core.

This was also observed within serovar for subspecies *enterica* strains. The original study examining the pan-genome of *S. enterica* used a set of 45 genomes and was able to identify unique gene families for each serovar examined, with Enteritidis having the fewest (29), and Typhi having the most (349) (Jacobsen et al., 2011). The results of the current study showed no unique genomic regions for any of the serovars with a sample set of 4893 quality filtered genomes. Although genomic regions universally present for each serovar were observed, and followed the same pattern with Enteritidis having the fewest (18), and Typhi having the most (288), these regions were also observed among genomes of other serovars, even though they were statistically over-represented for the serovar in question. The presence of these predictive markers in nearly all of the genomes within the EnteroBase test dataset indicates that the markers are robust, indicative of serovar, and could be combined to determine the likelihood of a genome being of a particular serovar.

When examining the average number of the 404 species-specific regions found among the *enterica* serovars, it was interesting to observe that Enteritidis, which had the fewest number of universal genomic regions, had the highest average number of species-specific regions; likewise Typhi, which had the most universally shared genomic regions, had one of the lowest averages of species-specific regions present. These results indicate that Enteritidis is the serovar that is closest to being the "core" example of a *S. enterica* genome, while Typhi is the serovar that is the most divergent. *S.* Enteritidis is the most common cause of enteric *Salmonella* infection, causing upward of one quarter of all infections, and is prevalent in chickens as well as their eggs (Chai et al., 2012). Conversely, *S.* Typhi is a human adapted serovar, responsible for Typhoid fever, and observed to have undergone genome degradation, rearrangement, and acquisition through horizontal gene-transfer, as it has evolved within its human host (Sabbagh et al., 2010; Klemm et al., 2016). It thus appears that genomic change enabling adaptation to a host creates a genomic pool that distinguishes a group from others of the same species. At the same time, genetically similar serovars that maintain a broad host range do not undergo as much selection for genomic change are much harder to distinguish as separate groups, but much easier to identify as members of the subspecies.

Core and Pan-genome Comparison

Most phylogentic studies focus on variation within homologs in the core genome to infer evolutionary relationships (Treangen et al., 2014), as paralogs and horizontally transferred elements confound the evolutionary signal found in genes obtained through vertical descent over time (Gabaldón and Koonin, 2013). While this approach is undoubtedly useful for long-term evolutionary analyses, when attempting to identify phenotypic linkages between phylogenetic clades, the accessory genome needs to be taken into account, as non-ubiquitous genomic regions allow different groups within the species to occupy and thrive in specific niches (Polz et al., 2013). Additionally, it has recently been shown that regulatory switching to non-homologous regulatory regions acquired via horizontal gene transfer happens in many bacteria (Oren et al., 2014). It was further shown that regulatory regions can move without the genes they regulate moving, and that at least 16% of the differences in expression observed within an *E. coli* population were explained by this regulatory switching.

It is therefore prudent to examine both the accessory genome, and not just genes, but non-coding DNA as well, as both have been shown to influence gene expression, and niche specificity. Recent studies have shown that the concordance between a phylogeny based on core genome SNPs and the presence/absence of pan-genome regions is high. For example, in a study examining *E. coli* lineage ST131, the core and accessory genomes showed high concordance, and the combined analyses of both allowed the analyses of the evolution of the *E. coli* lineage at a resolution not possible if only a restricted portion of the genome had been considered (McNally et al., 2016). The current study shows the same concordant relationship within *S. enterica* between the core and accessory genome, indicating that the accessory genome is not just randomly acquired genomic material, but that selection within specific niches establishes a complement of genes and regulatory elements that enable the survival of the *S. enterica* strains present. It also suggests that to understand why particular clades are more virulent, or possess a particular phenotype, a pan-genomic approach should be used in comparative analyses.

CONCLUSION

We examined a quality filtered set of 4893 genomes, the largest pan-genomic study of the *S. enterica* species to date. We identified a pan-genome of 25.3 Mbp, a strict core of 1.5 Mbp present in all genomes, and a conserved core of 3.2 Mbp found in at least 96% of the genomes in this study. In addition we identified 404 species-specific regions, within which a minimum set of two was required to unambiguously identify a genome as being part of the species *S. enterica*. These species-specific regions were found to have functions related to the propagation in and colonization of the host, including the utilization of sialic acid in intestinal mucus, secretion systems for attachment to the host, and the killing of other host microbiota. Within subspecies *enterica*, the species-specific regions were found most frequently in serovar Enteritidis. Each of the six subspecies was found to have genomic regions specific to it; however, the number of subspecies-specific regions appeared to be correlated with the level of sampling of the diversity within the subspecies. No serovar had pan-genome regions that were present in all of its representative genomes and absent in all other serovar genomes; however, each serovar did have genomic regions that were universally present among all constituent members, and statistically predictive of the serovar. *S.* Typhi, which is host-adapted to humans, was found to have the most universal markers predictive of its serovar. The phylogeny based on SNPs within the conserved core genome was found to be highly concordant to that produced by a phylogeny using the presence/absence of the entire pan-genome, and both agreed with phylogenies previously reported for *S. enterica*. Together, the core and accessory genome offered a more complete picture of the diversity within the genomes than either alone. The genomic regions identified in this study that are predictive of the species *S. enterica*, its six subspecies, and the serovar groups within subspecies *enterica*, could be developed into simple and rapid diagnostic tests, with uses ranging from food safety to public health. Additionally, the tools and methods described in this study could be generally applicable as a pan-genomics framework for future population studies, or those looking for genotype/phenotype linkages.

AUTHOR CONTRIBUTIONS

CL: designed the experiments, analyzed the data, and wrote the manuscript. MW: designed the experiments, and wrote the manuscript. VG: designed the experiments, and wrote the manuscript.

ACKNOWLEDGMENT

Thanks to Peter Kruczkiewicz of the Public Health Agency of Canada for providing the curated metadata for the EnteroBase genomes used in this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01345/full#supplementary-material

REFERENCES

Aanensen, D. M., Feil, E. J., Holden, M. T. G., Dordel, J., Yeats, C. A., Fedosejev, A., et al. (2016). Whole-genome sequencing for routine pathogen surveillance in public health: a population snapshot of invasive *Staphylococcus aureus* in Europe. *mBio* 7:e00444-16. doi: 10.1128/mBio.00444-16

Allard, M. W., Luo, Y., Strain, E., Li, C., Keys, C. E., Son, I., et al. (2012). High resolution clustering of *Salmonella enterica* serovar Montevideo strains using a next-generation sequencing approach. *BMC Genomics* 13:32. doi: 10.1186/1471-2164-13-32

Ashton, P. M., Nair, S., Peters, T. M., Bale, J. A., Powell, D. G., Painset, A., et al. (2016). Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ* 4:e1752. doi: 10.7717/peerj.1752

Babenko, D., Azizov, I., and Toleman, M. (2016). wgMLST as a standardized tool for assessing the quality of genome assembly data. *Int. J. Infect. Dis.* 45:329. doi: 10.1016/j.ijid.2016.02.714

Bergholz, T. M., Moreno Switt, A. I., and Wiedmann, M. (2014). Omics approaches in food safety: Fulfilling the promise? *Trends Microbiol.* 22, 275–281. doi: 10.1016/j.tim.2014.01.006

Bijlsma, J. J., and Groisman, E. A. (2005). The PhoP/PhoQ system controls the intramacrophage type three secretion system of *Salmonella enterica*. *Mol. Microbiol.* 57, 85–96. doi: 10.1111/j.1365-2958.2005.04668.x

Brunet, Y. R., Khodr, A., Logger, L., Aussel, L., Mignot, T., Rimsky, S., et al. (2015). H-NS silencing of the *Salmonella* pathogenicity island 6-encoded type VI secretion system limits *Salmonella enterica* serovar typhimurium interbacterial killing. *Infect. Immun.* 83, 2738–2750. doi: 10.1128/IAI.00198-15

Chai, S. J., White, P. L., Lathrop, S. L., Solghan, S. M., Medus, C., McGlinchey, B. M., et al. (2012). *Salmonella enterica* serotype enteritidis: increasing incidence of domestically acquired infections. *Clin. Infect. Dis.* 54(Suppl. 5), S488–S497. doi: 10.1093/cid/cis231

Chen, J., Zhang, L., Paoli, G. C., Shi, C., Tu, S. I., and Shi, X. (2010). A real-time PCR method for the detection of *Salmonella enterica* from food using a target sequence identified by comparative genomic analysis. *Int. J. Food Microbiol.* 137, 168–174. doi: 10.1016/j.ijfoodmicro.2009.12.004

Cohen, H., Mechanda, S., and Lin, W. (1996). PCR amplification of the fimA gene sequence of *Salmonella* typhimurium, a specific method for detection of *Salmonella* spp. *Appl. Environ. Microbiol.* 62, 4303–4308.

den Bakker, H. C., Allard, M. W., Bopp, D., Brown, E. W., Fontana, J., Iqbal, Z., et al. (2014). Rapid whole-genome sequencing for surveillance of *Salmonella enterica* serovar enteritidis. *Emerg. Infect. Dis.* 20, 1306–1314. doi: 10.3201/eid2008.131399

Deng, X., den Bakker, H. C., and Hendriksen, R. S. (2016). Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu. Rev. Food Sci. Technol.* 7, 353–374. doi: 10.1146/annurev-food-041715-033259

Desai, P. T., Porwollik, S., Long, F., Cheng, P., Wollam, A., Bhonagiri-Palsikar, V., et al. (2013). Evolutionary genomics of *Salmonella enterica* subspecies. *mBio* 4:e00579-12. doi: 10.1128/mBio.00579-12

Franz, E., Gras, L. M., and Dallman, T. (2016). Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Curr. Opin. Food Sci.* 8, 74–79. doi: 10.1016/j.cofs.2016.04.004

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi: 10.1093/bioinformatics/bts565

Gabaldón, T., and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nat. Rev. Genet.* 14, 360–366. doi: 10.1038/nrg3456

Gal-Mor, O., Boyle, E. C., and Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front. Microbiol.* 5:391. doi: 10.3389/fmicb.2014.00391

Golubeva, Y. A., Ellermeier, J. R., Chubiz, J. E. C., and Slauch, J. M. (2016). Intestinal long-chain fatty acids act as a direct signal to modulate expression of the *Salmonella* pathogenicity island 1 type III secretion system. *mBio* 7:e02170-15. doi: 10.1128/mBio.02170-15

Guo, X., Chen, J., Beuchat, L. R., and Robert, E. (2000). PCR detection of *Salmonella enterica* serotype montevideo in and on raw tomatoes using primers derived from *hilA*. *Appl. Environ. Microbiol.* 66, 5248–5252. doi: 10.1128/AEM.66.12.5248-5252.2000.Updated

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

Hayden, H. S., Matamouros, S., Hager, K. R., Brittnacher, M. J., Rohmer, L., Radey, M. C., et al. (2016). Genomic analysis of *Salmonella enterica* serovar Typhimurium characterizes strain diversity for recent U.S. salmonellosis cases and identifies mutations linked to loss of fitness under nitrosative and oxidative stress. *mBio* 7:e00154-16. doi: 10.1128/mBio.00154-16

Hensel, M., Shea, J. E., Waterman, S. R., Mundy, R., Nikolaus, T., Banks, G., et al. (1998). Genes encoding putative effector proteins of the type III secretion system of *Salmonella* pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages. *Mol. Microbiol.* 30, 163–174. doi: 10.1046/j.1365-2958.1998.01047.x

Horvath, L., Kraft, M., Fostiropoulos, K., Falkowski, A., and Tarr, P. E. (2016). *Salmonella enterica* subspecies *diarizonae* maxillary sinusitis in a snake handler: first report. *Open Forum Infect. Dis.* 3:ofw066. doi: 10.1093/ofid/ofw066

Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W., and Friis, C. (2011). The *Salmonella enterica* Pan-genome. *Microb. Ecol.* 62, 487–504. doi: 10.1007/s00248-011-9880-1

Kirk, M. D., Pires, S. M., Black, R. E., Caipo, M., Crump, J. A., Devleesschauwer, B., et al. (2015). World Health Organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. *PLoS Med.* 12:e1001921. doi: 10.1371/journal.pmed.1001921

Klemm, E. J., Gkrania-Klotsas, E., Hadfield, J., Forbester, J. L., Harris, S. R., Hale, C., et al. (2016). Emergence of host-adapted *Salmonella* Enteritidis through rapid evolution in an immunocompromised host. *Nat. Microbiol.* 1:15023. doi: 10.1038/NMICROBIOL.2015.23

Kokkinos, P. A., Ziros, P. G., Bellou, M., and Vantarakis, A. (2014). Loop-Mediated Isothermal Amplification (LAMP) for the detection of *Salmonella* in food. *Food Anal. Methods* 7, 512–526. doi: 10.1007/s12161-013-9748-8

Kubori, T., Sukhan, A., Aizawa, S. I., and Galán, J. E. (2000). Molecular characterization and assembly of the needle complex of the *Salmonella* typhimurium type III protein secretion system. *Proc. Natl. Acad. Sci. U.S.A.* 97, 10225–10230. doi: 10.1073/pnas.170128997

Laing, C., Buchanan, C., Taboada, E. N., Zhang, Y., Kropinski, A., Villegas, A., et al. (2010). Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* 11:461. doi: 10.1186/1471-2105-11-461

Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M., and Ussery, D. W. (2012). Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* 13:88. doi: 10.1186/1471-2164-13-88

Levine, M. M., Stinear, T., Holt, K. E., Robins-Browne, R. M., Ingle, D. J., Kuzevski, A., et al. (2016). *In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb. Genomics* 2:e000064. doi: 10.1099/mgen.0.000064

Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L., and Gally, D. L. (2016). Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11312–11317. doi: 10.1073/pnas.1606567113

Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O'Brien, S. J., et al. (2010). The global burden of nontyphoidal *Salmonella* gastroenteritis. *Clin. Infect. Dis.* 50, 882–889. doi: 10.1086/650733

Malorny, B., Hoorfar, J., Bunge, C., and Helmuth, R. (2003). Multicenter validation of the analytical accuracy of *Salmonella* PCR: towards an international standard. *Appl. Environ. Microbiol.* 69, 290–296. doi: 10.1128/AEM.69.1.290

Malorny, B., Paccassoni, E., Fach, P., Martin, A., Helmuth, R., and Bunge, C. (2004). Diagnostic real-time PCR for detection of *Salmonella* in food. *Appl. Environ. Microbiol.* 70, 7046–7052. doi: 10.1128/AEM.70.12.7046

McDermott, P. F., Tyson, G. H., Kabera, C., Chen, Y., Li, C., Folster, J. P., et al. (2016). The use of whole genome sequencing for detecting antimicrobial resistance in nontyphoidal *Salmonella*. *Antimicrob. Agents Chemother.* 60, 5515–5520. doi: 10.1128/AAC.01030-16

McDonald, N. D., Lubin, J. B., Chowdhury, N., and Boyd, E. F. (2016). Host-derived sialic acids are an important nutrient source required for optimal bacterial fitness in vivo. *mBio* 7:e02237-15. doi: 10.1128/mBio.02237-15

McNally, A., Oren, Y., Kelly, D., Pascoe, B., Dunn, S., Sreecharan, T., et al. (2016). Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution of bacterial populations. *PLoS Genet.* 12:e1006280. doi: 10.1371/journal.pgen.1006280

Moreno Switt, A. I., Orsi, R. H., den Bakker, H. C., Vongkamjan, K., Altier, C., and Wiedmann, M. (2013). Genomic characterization provides new insight into *Salmonella* phage diversity. *BMC Genomics* 14:481. doi: 10.1186/1471-2164-14-481

Ng, K. M., Ferreyra, J. A., Higginbottom, S. K., Lynch, J. B., Kashyap, P. C., Gopinath, S., et al. (2013). Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* 502, 96–99. doi: 10.1038/nature12503

Opijnen, T. V., Camilli, A., Opijnen, T. V., and Camilli, A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. *Genome Res.* 22, 2541–2551. doi: 10.1101/gr.137430.112

Oren, Y., Smith, M. B., Johns, N. I., Kaplan Zeevi, M., Biran, D., Ron, E. Z., et al. (2014). Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16112–16117. doi: 10.1073/pnas.1413272111

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31, 3691–3693. doi: 10.1093/bioinformatics/btv421

Penadés, J. R., Chen, J., Quiles-Puchalt, N., Carpena, N., and Novick, R. P. (2015). Bacteriophage-mediated spread of bacterial virulence genes. *Curr. Opin. Microbiol.* 23, 171–178. doi: 10.1016/j.mib.2014.11.019

Pettengill, J. B., Pightling, A. W., Baugher, J. D., Rand, H., and Strain, E. (2016). Real-time pathogen detection in the era of whole-genome sequencing and big data: comparison of k-mer and site-based methods for inferring the genetic distances among tens of thousands of *Salmonella* samples. *PLoS ONE* 11:e0166162. doi: 10.1371/journal.pone.0166162

Polz, M. F., Alm, E. J., and Hanage, W. P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* 29, 170–175. doi: 10.1016/j.tig.2012.12.006

Postollec, F., Falentin, H., Pavan, S., Combrisson, J., and Sohier, D. (2011). Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol.* 28, 848–861. doi: 10.1016/j.fm.2011.02.008

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rahman, A., and Pachter, L. (2013). CGAL: computing genome assembly likelihoods. *Genome Biol.* 14:R8. doi: 10.1186/gb-2013-14-1-r8

Sabbagh, S. C., Forest, C. G., Lepage, C., Leclerc, J. M., and Daigle, F. (2010). So similar, yet so different: uncovering distinctive features in the genomes of *Salmonella enterica* serovars Typhimurium and Typhi. *FEMS Microbiol. Lett.* 305, 1–13. doi: 10.1111/j.1574-6968.2010.01904.x

Sakarya, S., Göktürk, C., Öztürk, T., and Ertugrul, M. B. (2010). Sialic acid is required for nonspecific adherence of *Salmonella enterica* ssp. *enterica* serovar Typhi on Caco-2 cells. *FEMS Immunol. Med. Microbiol.* 58, 330–335. doi: 10.1111/j.1574-695X.2010.00650.x

Sana, T. G., Flaugnatti, N., Lugo, K. A., Lam, L. H., Jacobson, A., Baylot, V., et al. (2016). *Salmonella* Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. *Proc. Natl. Acad. Sci. U.S.A.* 113, E5044–E5051. doi: 10.1073/pnas.1608858113

Scallan, E., Hoekstra, R. M., Mahon, B. E., Jones, T. F., and Griffin, P. M. (2015). An assessment of the human health impact of seven leading foodborne pathogens in the United States using disability adjusted life years. *Epidemiol. Infect.* 143, 2795–2804. doi: 10.1017/S0950268814003185

Scharff, R. L., Besser, J., Sharp, D. J., Jones, T. F., Peter, G. S., and Hedberg, C. W. (2016). An economic evaluation of PulseNet: a network for foodborne disease surveillance. *Am. J. Prev. Med.* 50, S66–S73. doi: 10.1016/j.amepre.2015.09.018

Schroter, M., Roggentin, P., Hofmann, J., Speicher, A., Laufs, R., and Mack, D. (2004). Pet snakes as a reservoir for *Salmonella enterica* subsp. *diarizonae* (serogroup IIIb): a prospective study. *Appl. Environ. Microbiol.* 70, 613–615. doi: 10.1128/AEM.70.1.613-615.2004

Severi, E., Müller, A., Potts, J. R., Leech, A., Williamson, D., Wilson, K. S., et al. (2008). Sialic acid mutarotation is catalyzed by the *Escherichia coli* β-propeller protein YjhT. *J. Biol. Chem.* 283, 4841–4849. doi: 10.1074/jbc.M707822200

Sheppard, S. K., Jolley, K. A., and Maiden, M. C. J. (2012). A gene-by-gene approach to bacterial population genomics: whole genome MLST of *Campylobacter*. *Genes* 3, 261–277. doi: 10.3390/genes3020261

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Sun, H., Kamanova, J., Lara-Tejero, M., and Galán, J. E. (2016). A Family of *Salmonella* type III secretion effector proteins selectively targets the NF-κB signaling pathway to preserve host homeostasis. *PLoS Pathog.* 12:e1005484. doi: 10.1371/journal.ppat.1005484

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Treangen, T. J., Ondov, B. D., Koren, S., and Phillippy, A. M. (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* 15:524. doi: 10.1186/s13059-014-0524-x

Tyson, G. H., McDermott, P. F., Li, C., Chen, Y., Tadesse, D. A., Mukherjee, S., et al. (2015). WGS accurately predicts antimicrobial resistance in *Escherichia coli*. *J. Antimicrob. Chemother.* 70, 2763–2769. doi: 10.1093/jac/dkv186

Waryah, C. B., Gogoi-Tiwari, J., Wells, K., Eto, K. Y., Masoumi, E., Costantino, P., et al. (2016). Diversity of virulence factors associated with West Australian methicillin-sensitive *Staphylococcus aureus* isolates of human origin. *BioMed Res. Int.* 2016:8651918. doi: 10.1155/2016/8651918

Whiteside, M. D., Laing, C. R., Manji, A., Kruczkiewicz, P., Taboada, E. N., and Gannon, V. P. J. (2016). SuperPhy: predictive genomics for the bacterial pathogen *Escherichia coli*. *BMC Microbiol.* 16:65. doi: 10.1186/s12866-016-0680-0

Yoshida, C., Gurnik, S., Ahmad, A., Blimkie, T., Murphy, S. A., Kropinski, A. M., et al. (2016a). Evaluation of molecular methods for identification of *Salmonella* serovars. *J. Clin. Microbiol.* 54, 1992–1998. doi: 10.1128/JCM.00262-16

Yoshida, C., Kruczkiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P. J., Nash, J. H. E., et al. (2016b). 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. doi: 10.1371/journal.pone.0147101

Yoshida, C., Lingohr, E. J., Trognitz, F., MacLaren, N., Rosano, A., Murphy, S. A., et al. (2014). Multi-laboratory evaluation of the rapid genoserotyping array (SGSA) for the identification of *Salmonella* serovars. *Diagn. Microbiol. Infect. Dis.* 80, 185–190. doi: 10.1016/j.diagmicrobio.2014.08.006

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T.-Y. (2016). GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* 8, 28–36. doi: 10.1111/2041-210X.12628

Zhao, S., Tyson, G. H., Chen, Y., Li, C., Mukherjee, S., Young, S., et al. (2016). Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in *Campylobacter* spp. *Appl. Environ. Microbiol.* 82, 459–466. doi: 10.1128/AEM.02873-15

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.