# Pan-genome analyses of the species *Salmonella enterica*, and identification of genomic markers predictive for species, subspecies, and serovar.

### Chad R Laing1, Matthew D Whiteside1, and Victor PJ Gannon1

#### 1 National Microbiology Laboratory @ Lethbridge, Public Health Agency of Canada, Lethbridge, Alberta, T1J 3Z4

/f

# Abstract

Food safety is a global concern, with upwards of 2.2 million deaths due to enteric diarrheal disease every year. Molecular diagnostic networks such as PulseNet have greatly increased the ability to rapidly identify and respond to food-related outbreaks. Until recently, identification of these contaminants relied on traditional microbiology laboratory techniques such as biochemical utilization, molecular characterization through the identification of key virulence markers, or banding patterns from rare-cutting restriction enzymes.

Current whole-genome sequencing platforms allow for routine sequencing for surveillance, and during outbreak situations; however, a critical challenge that remains in food safety is linking the genotype of an organism to its corresponding phenotype, as differences in human illness have been observed within lineages of the same bacterial species, and even within clades of these lineages

We have previously developed the software program Panseq, which identifies the pan-genome for a group of sequences, and the SuperPhy platform, which utilizes this pan-genome information to identify virulence and antimicrobial resistance genes within genome sequences, as well as biomarkers (both the presence / absence of genes, and single-nucleotide polymorphisms within shared genes) that are predictive of groups of bacterial strains.

In this study we use the previously described software to identify and examine the pan-genome of *Salmonella enterica*, an enteric pathogen responsible for the loss of more disability adjusted life years than any other enteric pathogen. We identify regions of the pan-genome that are statistically predictive of the species *Salmonella enterica*, its six subspecies, and serotypes within subspecies enterica. The approach defined in this study is broadly applicable to any species or collection of genomes.

# 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 fingerprinting the etiological agents of disease, and keeping nationwide databases of previous fingerprints associated with human disease. Since its inception, PulseNet has relied on Pulsed-Field Gel Electrophoresis (PFGE) for fingerprinting of bacterial pathogens. 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 lack 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 the known virulence factors, are rarely associated with human disease (Waryah *et al.* 2016, Lupolova *et al.* (2016)). Thus, other factors within the genome 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).

Whole-genome sequencing (WGS) has become the de facto standard for the complete characterization for bacterial pathogens for both ongoing surveillance and outbreak investigations (Deng X, den Bakker HC 2016, Franz *et al.* (2016)). It allows clear definition between outbreak-related strains and those from unrelated sources, the ability to identify routes of transmission, and the ability to perform source attribution of bacterial contaminants (Bakker *et al.* 2014). It is currently being utilized worldwide in outbreak and routine surveillance, including the characterization of all *Listeria monocytogenes* isolated in the United States[genometrackr ref], all *Salmonella* isolated by Public Health England as part of routine surveillance (Ashton *et al.* 2016), and a large-scale survey of *Staphylococcus aureus* in continental Europe, which demonstrated the applicability of WGS for the identification of the emergence and spread of clinically relevant *Staphylococcus aureus* (Aanensen *et al.* 2016).

It has been shown that *in silico* prediction of antimicrobial resistance (Zhao *et al.* 2016, Tyson *et al.* (2015), McDermott *et al.* (2016)), serotype (Levine *et al.* 2016, Yoshida *et al.* (2016)), and other traditional sub-typing schemes such as multi-locus sequence typing (Sheppard *et al.* 2012) can be accurately reproduced from bacterial genome sequences. However, given the more complex task of identifying bacterial isolates that are most likely to cause disease in humans, methods that can correctly identify markers that predict such strains from the genome sequence alone are needed. 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 species that contains human-adapted strains responsible for typhoid fever, as well as a large number of non-typhoidal strains responsible for an estimated 93.8 million annual cases of enteric illness worldwide (Gal-Mor *et al.* 2014, Majowicz *et al.* (2010)). We identify *S. enterica* species- and subspecies-specific markers, as well as markers predictive of serotype for subspecies enterica. While this study focused on *S. senterica*, the tools and approach are broadly applicable to any species or collection of genomes.

# Results

## *S. enterica* pan-genome

Initially we determined the size and distribution of the *S. enterica* pan-genome as genome fragments of 1000bp in size. 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 that present in fewer than 100 genomes. 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 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 405 putative *S. enterica* species-specific regions.

Figure 2 shows the carriage of these 405 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 405 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 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 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 thousands of contigs (Table 2). Genomes from subspecies houtenae and arizonae contained fewer than 100 species-specific regions, while genomes from diarizone, indica, and salamae contained between 100 and 200 species-specific regions. All regions were screened against *Salmonella 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 totalling over 464 Kbp. A final outlier genome harbored nearly 5000 contigs, but also contained 403 of the species-specific regions. It was determined 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* 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 greater than 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 order of declining 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 ten 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 403 species-specific markers. Among the most abundant serovars Typhimiurium, Heidelberg, Newport, and Enteritidis were found to contain the most species-specific markers, and group 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 4 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 6. As can be seen this phylogeny based on the presence / absence of the entire 25.3 Mbp pan-genome is highly concordant to the phylogeny based on the SNPs found in the conserved core of the same strains (Figure 6). In both trees, the serovars cluster together and in the same relation to each 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 regions.

## Identification of a minimum set of species-specific genomic markers for the identification of subspecies *S. enterica*.

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

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. The two genomic regions that met this requirement and were also present in the most genomes are highlighted in Table X. All members of *S. enterica* examined contained at least one of these markers, but many other combinations within the 404 species-specific markers are possible.

## Putative functional identification of the *S. enterica* species-specific regions

The putative function of the 403 quality-filtered *S. enterica* species-specific regions were determined form the GenBank nr database. The annotation of each of the 403 regions is available as Supplementary File 1. Table 3 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 403 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 the 404 species-specific ones, to identify subpecies-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 identified serovar-specific markers for the top five 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 conserved in their presence or absence among serovars, even though theses same markers were both present and absent in genomes of other serotypes. The number of markers universal for presence or absence among the enterica serovars is shown in Table 6.

# Discussion

## *S. enterica* pan-genome

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).

Previous examinations of the *S. enterica* pan-genome were based on relatively small dataset of 45 and 73 genomes (Jacobsen *et al.* 2011,Leekitcharoenphon *et al.* (2012)). While others have analyzed thousands of *S. enterica* genomes, the analyses have not been 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, which given an average gene size of 1000bp is ~1500 genes, and a much larger pan-genome at ~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 *Escherichia 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 labelled, and poor quality assemblies, can greatly affect the size, analyses, and composition of the pan-genome. Software tools to evaluate assembly quality 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.* (2016)), but the benefit of a pan-genome analysis is that it is schema free.

## *S. enterica* species-specific regions

The host intestinal environment consists of a multitude of bacterial species competing for scare nutritional sources such as carbohydrates, direct antagonistic competition with other bacterial cells, and competition for access to the host 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 intestinal disruption from 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 capitalize on the fluctuating environment (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 [Severi2008]. In this study we identified n-acetylneuraminic acid mutarotases as species-specific genomic regions, along with sialic acid transporters. It is likely 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 levels of bacterial colonization of the host gut (Ng *et al.* 2013).

The ability to utilize sialic acids has 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 diverse 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 be 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 into 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, Bijlsma2005). 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-kB signalling 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 help cue the bacteria 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 be competitive 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 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 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 give a good overview of what make it such an effective pathogen and intestinal inhabitant.

## Specific regions for subspecies and serovar

The phylogenetic relationship of the six *S. enterica* subspecies has been previously described, which the current study recapitulates (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 diarizone 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 arizone 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 serotype 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 serotype 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 serotypes.

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 upwards 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 (Klemm *et al.* 2016, Sabbagh *et al.* (2010)). 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 not undergoing selection for genomic change are much harder to individually 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 homologues in the core genome to infer evolutionary relationships (Treangen *et al.* 2014), as paralogues and horizontally transfered 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 across the domain 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. And while it may at first seem like the concordance between a phylogeny based on core genome SNPs and the presence / absence of pan-genome regions would be high, recent studies have shown just that. 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 portion of all genomes 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.

# Tables

Table 1. The frequency of the subspecies observed within the study set of 4937 *Salmonella enterica* genomes, prior to any quality filtering.

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

Table 2. The serovars with more than 20 representatives in the current study set of 4937 *Salmonella enterica* genomes, and their frequency, prior to any quality filtering. The list of all serovars and their frequency within the current study is available as Supplementary File 2.

|  |  |
| --- | --- |
| 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 |

Table 3. 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. The complete list of all putative functions is available as Supplemental File 1.

|  |  |
| --- | --- |
| Putative protein function | Frequency |
| hypothetical | 64 |
| secreted effector | 10 |
| membrane | 7 |
| secretion system apparatus | 5 |
| uncharacterised | 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 |

Table 4. 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 subpecies 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 |
| I | 399.0 |
| 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 |
| C1:g | 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 |

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 ten most abundant serotypes 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 |

# Figures

1. Figure 1 The distribution of the *Salmonella enterica* pan-genome, as 1000bp fragments, among 4939 whole-genome sequences.
2. Figure 2 The carriage of the 405 *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.
3. Figure 3 The carriage of the 405 *S. enterica* species-specific regions, and the number of contigs that each of the 4939 genomes of this study were distributed amongst. Colours indicate the subspecies within *S. enterica*, or contamination from ther species as follows: red: arizonae, yellow: *Citrobacter* contamination, lime: diarizonae, teal: enterica, blue: houtenae, lavender: indica, magenta: salamae.
4. Figure 4 The phylogeny of the 4893 *S. enterica* genomes post quality-filtering. The six subspecies are highlighted as follows: red: arizonae, yellow: diarizonae, lime: enterica, teal: houtenae, blue: indica, magenta: salamae. The matrix to the right of the phylogeny represents the 404 species-specific regions, with red being the absence of the region, and black being the presence of the region, for each of the genomes of the study.
5. Figure 5 The phylogeny of the 4893 *S. enterica* genomes post quality-filtering based on SNPs found within the conserved core genome. The ten most abundant serotypes of subspecies enterica in the current study are highlighted as follows: dark green: Agona, dark orange: Bareilly, purple: Enteritidis, magenta: Heidelberg, light green: Kentucky, yellow: Newport, brown: Paratyphi, grey: Typhi, pink: Typhimurium, light blue: Weltevreden. The matrix to the right of the phylogeny represents the 404 species-specific regions, with red being the absence of the region, and black being the presence of the region, for each of the genomes of the study.
6. Figure 6 The phylogeny of the 4893 *S. enterica* genomes post quality-filtering based on the presence / absence of the entire pan-genome as 1000bp fragments. The ten most abundant serotypes of subspecies enterica in the current study are highlighted as follows: dark green: Agona, dark orange: Bareilly, purple: Enteritidis, magenta: Heidelberg, light green: Kentucky, yellow: Newport, brown: Paratyphi, grey: Typhi, pink: Typhimurium, light blue: Weltevreden. The matrix to the right of the phylogeny represents the 404 species-specific regions, with red being the absence of the region, and black being the presence of the region, for each of the genomes of the study.

Aanensen DM, Feil EJ, and Holden MTG *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.

Allard MW, Luo Y, and Strain E *et al.* 2012. High resolution clustering of Salmonella enterica serovar Montevideo strains using a next- generation sequencing approach. *BMC Genomics* **13**: 32.

Ashton PM, Nair S, and Peters TM *et al.* 2016. Identification of <i>Salmonella</i> for public health surveillance using whole genome sequencing. *PeerJ* **4**: e1752.

Babenko D, Azizov I, and Toleman M. 2016. wgMLST as a standardized tool for assessing the quality of genome assembly data. *International Journal of Infectious Diseases* **45**: 329.

Bakker HC den, Allard MW, and Bopp D *et al.* 2014. Rapid whole-genome sequencing for surveillance of salmonella enterica serovar Enteritidis. *Emerging Infectious Diseases* **20**: 1306–14.

Bergholz TM, Moreno Switt AI, and Wiedmann M. 2014. Omics approaches in food safety: Fulfilling the promise? *Trends in Microbiology* **22**: 275–81.

Brunet YR, Khodr A, and Logger L *et al.* 2015. H-NS silencing of the salmonella pathogenicity island 6-encoded type VI secretion system limits salmonella enterica serovar typhimurium interbacterial killing. *Infection and Immunity* **83**: 2738–50.

Chai SJ, White PL, and Lathrop SL *et al.* 2012. Salmonella enterica serotype enteritidis: Increasing incidence of domestically acquired infections. *Clinical Infectious Diseases* **54**.

Deng X, den Bakker HC HR. 2016. Genomic Epidemiology: Whole-Genome-Sequencing-Powered Surveillance and Outbreak Investigation of Foodborne Bacterial Pathogens. *Annu Rev Food Sci Technol*: 1–22.

Desai PT, Porwollik S, and Long F *et al.* 2013. Evolutionary Genomics of Salmonella enterica Subspecies. *mBio* **4**: 1–13.

Franz E, Gras LM, and Dallman T. 2016. Significance of whole genome sequencing for surveillance, source attribution and microbial risk assessment of foodborne pathogens. *Current Opinion in Food Science* **8**: 74–9.

Gabaldón T and Koonin EV. 2013. Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics* **14**: 360–6.

Gal-Mor O, Boyle EC, and Grassl GA. 2014. Same species, different diseases: How and why typhoidal and non-typhoidal Salmonella enterica serovars differ. *Frontiers in Microbiology* **5**: 1–10.

Golubeva YA, Ellermeier JR, Chubiz JEC, and Slauch JM. 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**: 1–9.

Gurevich A, Saveliev V, Vyahhi N, and Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **29**: 1072–5.

Hayden HS, Matamouros S, and Hager KR *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**: 1–11.

Hensel M, Shea JE, and Waterman SR *et al.* 1998. Genes encoding putative effector proteins of the type III secretion system of <i>Salmonella</i> pathogenicity island 2 are required for bacterial virulence and proliferation in macrophages. *Molecular Microbiology* **30**: 163–74.

Horvath L, Kraft M, and Fostiropoulos K *et al.* 2016. <i>Salmonella enterica</i> Subspecies <i>diarizonae</i> Maxillary Sinusitis in a Snake Handler: First Report. *Open Forum Infectious Diseases* **3**: ofw066.

Jacobsen A, Hendriksen RS, and Aaresturp FM *et al.* 2011. The Salmonella enterica Pan-genome. *Microbial Ecology* **62**: 487–504.

Kirk MD, Pires SM, and Black RE *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 Medicine* **12**: 1–21.

Klemm EJ, Gkrania-Klotsas E, and Hadfield J *et al.* 2016. Emergence of host-adapted Salmonella Enteritidis through rapid evolution in an immunocompromised host. *Nature Microbiology* **1**: 1–6.

Kubori T, Sukhan A, Aizawa SI, and Galán JE. 2000. Molecular characterization and assembly of the needle complex of the Salmonella typhimurium type III protein secretion system. *Proceedings of the National Academy of Sciences of the United States of America* **97**: 10225–30.

Laing C, Buchanan C, and Taboada EN *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.

Leekitcharoenphon P, Lukjancenko O, and Friis C *et al.* 2012. Genomic variation in Salmonella enterica core genes for epidemiological typing. *BMC Genomics* **13**: 88.

Levine MM, Stinear T, and Holt KE *et al.* 2016. In silico serotyping of E. coli from short read data identifies limited novel O-loci but extlinensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microbial Genomics*: 1–14.

Lupolova N, Dallman TJ, and Matthews L *et al.* 2016. Support vector machine applied to predict the zoonotic potential of <i>E. coli</i> O157 cattle isolates. *Proceedings of the National Academy of Sciences* **113**: 201606567.

Majowicz SE, Musto J, and Scallan E *et al.* 2010. The Global Burden of Nontyphoidal Salmonella Gastroenteritis. *Clinical Infectious Diseases* **50**: 882–9.

McDermott PF, Tyson GH, and Kabera C *et al.* 2016. The use of whole genome sequencing for detecting antimicrobial resistance in nontyphoidal Salmonella. *Antimicrobial Agents and Chemotherapy* **60**: AAC.01030–16.

McDonald ND, Lubin JB, Chowdhury N, and Boyd EF. 2016. Host-derived sialic acids are an important nutrient source required for optimal bacterial fitness In Vivo. *mBio* **7**: 1–10.

McNally A, Oren Y, and Kelly D *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 Genetics* **12**: e1006280.

Moreno Switt AI, Orsi RH, and Bakker HC den *et al.* 2013. Genomic characterization provides new insight into Salmonella phage diversity. *BMC genomics* **14**: 481.

Ng KM, Ferreyra J a, and Higginbottom SK *et al.* 2013. Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature* **502**: 96–9.

Opijnen TV, Camilli A, Opijnen TV, and Camilli A. 2012. A fine scale phenotype − genotype virulence map of a bacterial pathogen A fine scale phenotype – genotype virulence map of a bacterial pathogen.: 2541–51.

Oren Y, Smith MB, and Johns NI *et al.* 2014. Transfer of noncoding DNA drives regulatory rewiring in bacteria. *Proceedings of the National Academy of Sciences* **111**: 16112–7.

Page AJ, Cummins CA, and Hunt M *et al.* 2015. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**: 3691–3.

Penadés JR, Chen J, and Quiles-Puchalt N *et al.* 2015. Bacteriophage-mediated spread of bacterial virulence genes. *Current Opinion in Microbiology* **23**: 171–8.

Pettengill JB, Pightling AW, and Baugher JD *et al.* 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.

Polz MF, Alm EJ, and Hanage WP. 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics* **29**: 170–5.

Rahman A and Pachter L. 2013. CGAL: computing genome assembly likelihoods. *Genome Biol* **14**: R8.

Sabbagh SC, Forest CG, and Lepage C *et al.* 2010. So similar, yet so different: Uncovering distinctive features in the genomes of Salmonella enterica serovars Typhimurium and Typhi. *FEMS Microbiology Letters* **305**: 1–13.

Sakarya S, Göktürk C, Öztürk T, and Ertugrul MB. 2010. Sialic acid is required for nonspecific adherence of Salmonella enterica ssp. enterica serovar Typhi on Caco-2 cells. *FEMS Immunology and Medical Microbiology* **58**: 330–5.

Sana TG, Flaugnatti N, and Lugo KA *et al.* 2016. Salmonella Typhimurium utilizes a T6SS-mediated antibacterial weapon to establish in the host gut. *Proceedings of the National Academy of Sciences of the United States of America* **113**: E5044–51.

Scallan E, Hoekstra RM, and Mahon BE *et al.* 2015. An assessment of the human health impact of seven leading foodborne pathogens in the United States using disability adjusted life years. *Epidemiology and Infection* **143**: 2795–804.

Scharff RL, Besser J, and Sharp DJ *et al.* 2016. An Economic Evaluation of PulseNet: A Network for Foodborne Disease Surveillance. *American Journal of Preventive Medicine* **50**: S66–73.

Schroter M, Roggentin P, and Hofmann J *et al.* 2004. Pet Snakes as a Reservoir for Salmonella enterica subsp. diarizonae (Serogroup IIIb): a Prospective Study. *Applied and Environmental Microbiology* **70**: 613–5.

Sheppard SK, Jolley KA, and Maiden MCJ. 2012. A gene-by-gene approach to bacterial population genomics: Whole genome MLST of Campylobacter. *Genes* **3**: 261–77.

Sun H, Kamanova J, Lara-Tejero M, and Galán JE. 2016. A Family of Salmonella Type III Secretion Effector Proteins Selectively Targets the NF-B Signaling Pathway to Preserve Host Homeostasis. *PLoS Pathogens* **12**: 1–19.

Tettelin H, Masignani V, and Cieslewicz MJ *et al.* 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 13950–5.

Treangen TJ, Ondov BD, Koren S, and Phillippy AM. 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology* **15**: 524.

Tyson GH, McDermott PF, and Li C *et al.* 2015. WGS accurately predicts antimicrobial resistance in Escherichia coli. *Journal of Antimicrobial Chemotherapy* **70**: 2763–9.

Waryah CB, Gogoi-Tiwari J, and Wells K *et al.* 2016. Diversity of Virulence Factors Associated with West Australian Methicillin-Sensitive <i>Staphylococcus aureus</i> Isolates of Human Origin. *BioMed Research International* **2016**: 1–10.

Whiteside MD, Laing CR, and Manji A *et al.* 2016. SuperPhy: predictive genomics for the bacterial pathogen Escherichia coli. *BMC microbiology* **16**: 65.

Yoshida CE, Kruczkiewicz P, and Laing CR *et al.* 2016. 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.

Zhao S, Tyson GH, and Chen Y *et al.* 2016. Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in <i>Campylobacter</i> spp. *Applied and Environmental Microbiology* **82**: 459–66.