Chapter 1

Insights from Comparative Genomics of the Genus *Salmonella*

Trudy M. Wassenaar, Se-Ran Jun, Visanu Wanchai, Preecha Patumcharoenpol, Intawat Nookaew, Katrina Schlum, Michael R. Leuze and David W. Ussery

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/67131

Abstract

Comparative genomics have become a standard approach to gain insights into the interrelationships of microorganisms. Here, we have applied variable bioinformatic techniques to compare over 200 *Salmonella* genomes. First, we present a tree of all sequenced different members of the *Enterobacteriaceae* family, based on comparison of average amino acid identities. This technique was also applied to zoom in on the genomes of the genus *Salmonella*. The pan and core genomes of this genus were established and compared to experimental data available on the literature that identified essential genes. Difficulties and shortcomings of both approaches are discussed. Metabolic pathways unique for *Salmonella* were identified. Finally, we present an analysis of genes coding for small RNAs, an important part of the genetic repertoire of bacteria that is often ignored. The findings reported here are discussed and compared with available literature.

Keywords: comparative genomics, *Salmonella*, core genome, small RNA, AAI tree

1. Introduction

The genus *Salmonella* belongs to the *Enterobacteriaceae*, a large family within the gamma-proteobacteria to which *E. coli* also belongs. Since its first characterization in 1884 from diseased pigs by scientists working in the group of Daniel Salmon (after whom the genus is named), *Salmonella* species have been known to cause disease, notably typhoid fever and food poisoning. Pathogenic *Salmonella* types can be found in a wide range of animal hosts and often infect humans via contaminated food; they are responsible for more than a million infections in the

United States every year. Infections vary from (long-term) asymptomatic carriage and self-limiting salmonellosis to life-threatening conditions and fatal typhoidal fever [1].

Historically, many species of this genus were recognized, at first based on the clinical symptoms typical for their infections and it was soon recognized that these correlated with their serotype. However, based on sequence analysis, in 1973, it was proposed that all these *Salmonella* serotypes belonged to the same species [2]. This resulted, in 2005, to the designation of *Salmonella enterica* as the type species for the genus, as described by the International Committee on Systematics of Procaryotes [3]. Only one other species is currently formally recognized within the genus: *Salmonella bongori*, which lives in cold-blooded reptiles. *S. enterica* is further divided into six subspecies, of which *S. enterica* subsp. *enterica* is clinically most relevant. The names originally used to describe clinically distinct 'species' live on as serovars or serotypes. All *Salmonella* bacteria are none spore-forming, chemotrophic, facultative anaerobes, which survive in their host intracellularly [1].

The number of *Salmonella* genome sequences available in GenBank is constantly increasing. At the time of writing their number reached five thousand, the vast majority of which were obtained from *S. enterica*. As of September 15, 2016, there were 4934 genomes of this species in GenBank, with three additional genomes from *S. bongori*. Only a small fraction of these genomes are submitted as complete sequences without gaps and fulfilling all criteria set by GenBank for a genome to be listed as 'complete' (201 genomes at the time of writing, corresponding to 4% of the total). In this chapter, we employ whole-genome methods to compare complete *Salmonella* genomes in order to produce insights into the genomic diversity of this genus.

2. *Salmonella* comparative genome analyses

2.1. Genome-based trees

The first approach was aimed to show the overall relatedness of all species belonging to the *Enterobacteriaceae* family, based on their (completely sequenced) genomes. For this, we collected up to ten genome sequences per species, as far as these were available, which led to 255 genome sequences to be compared. The comparison was based on average amino acid identity (AAI) comparison, a method that uses all annotated protein genes in a given genome, producing more robust trees than methods based on direct alignments or concatenated protein sequence alignments [4]. The resulting tree is presented with collapsed branches for redundant species (**Figure 1**). The *Salmonella* genus, shown in red, is positioned on a cluster together with *Citrobacter*, with *Escherichia*/*Shigella* as the closest neighbors. These genera are supposed to have been separated for tens of millions of years [5]. The close relationship between *Citrobacter* and *Salmonella* has been observed before, and it was proposed that recombination between these and to a lesser extent with *Escherichia*, has been frequent in the past, during a process of fragmented speciation [5].

Next, we extracted all 201 complete genomes from the *Salmonella* genus (in May 2016), combined with 164 'nearly completed' genomes. The latter were extracted from GenBank as good quality draft sequences only, retrieved from GenBank when selecting for genomes

http://dx.doi.org/10.5772/67131

of 'chromosome' quality; all contained one contiguous sequence, without gaps. These 365 *Salmonella* genomes represent only a tiny fraction of what is available. Apart from the nearly 5000 *Salmonella* genomes available in GenBank, there are currently more than 62,000 *Salmonella enterica* genomes stored in the Sequence Read Archive. However, in principle, the complete genome sequences should be of high quality and reliable in terms of annotation; therefore, we restricted the analysis to complete genomes.

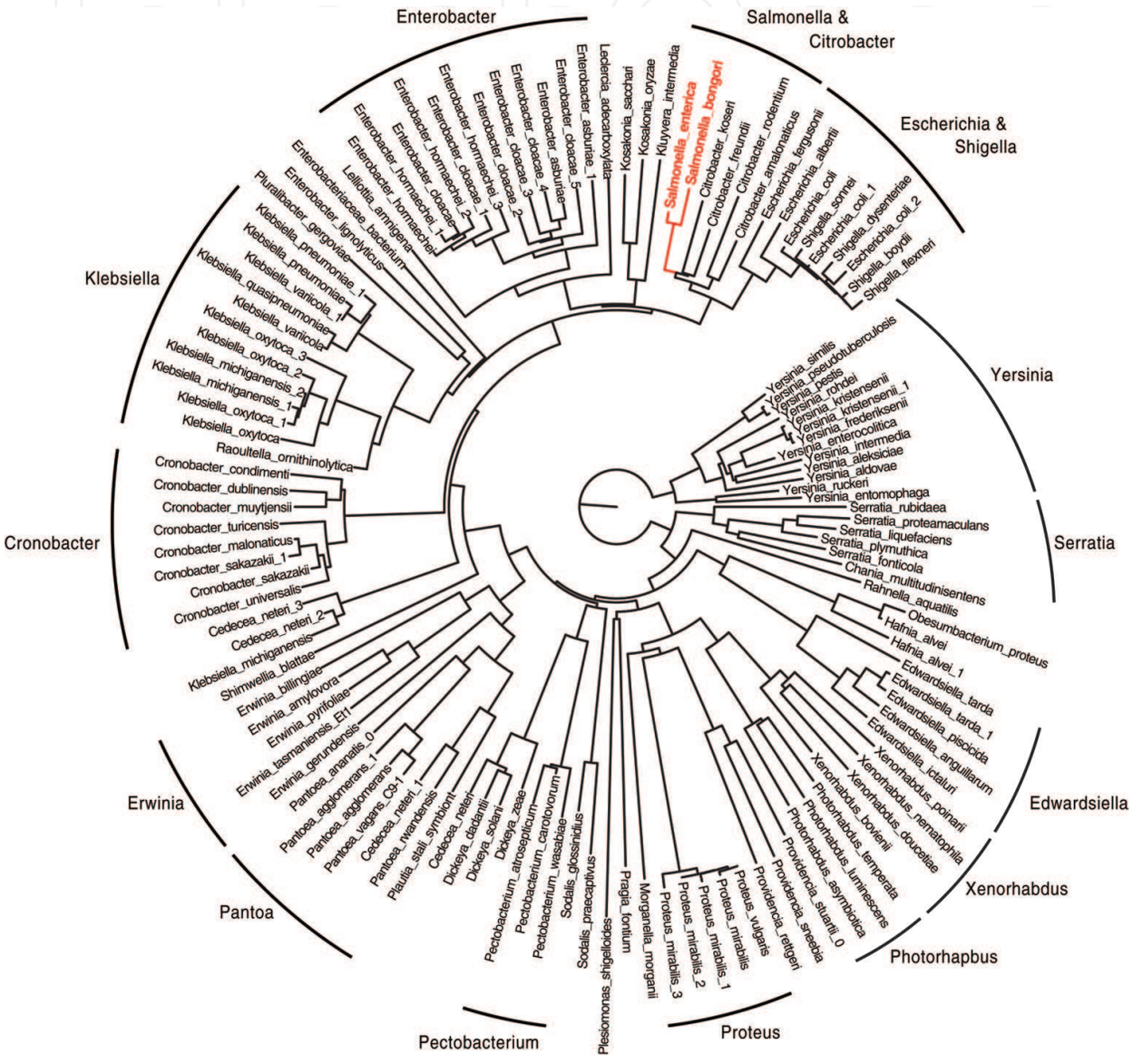

Figure 1. Tree based on average amino acid identity (AAI) of 255 genomes from members of the *Enterobacteraceae*. Branches were collapsed at the species level. The branch with the two *Salmonella* species is colored and some distinct genus clusters are labeled.

An AAI tree was constructed to establish the interrelationship of the 365 complete genomes, representing 33 different serovars including 36 Typhimurium and 6 Typhi genomes. The branches of the AAI tree were collapsed at serovar level. This produced a tree with 62 branches, as shown in **Figure 2**. As can be observed, by and large the tree clustered the genomes according

to serovars, though the separation is not absolute and some serovars end up in mixed clusters. This was to be expected, as the analysis is based on the complete annotated proteome (capturing all protein-coded sequences), while the phenotypic characteristics that determine a serovar are determined by a limited number of genes only, that produce the surface antigens captured by serotyping. Of the 36 *S. enterica* sv. Typhimurium genomes (represented on 13 branches, blue in the figure), 32 cluster together on 10 branches (together with four branches of non-specified serovars), while four are placed on three branches outside the Typhimurium cluster. A distinct cluster is also observed containing the serovars Enteritidis, Pullorum, Gallinarum and Dublin (colored green in the figure) which together are known as 'group D *Salmonella*' [6]. The first three of these are adapted to the chicken host, but serovar Dublin is mostly colonizing cattle, and other serovars frequently found in chickens are placed outside the group D cluster. It has been suggested that the serovars Paratyphi and Choleraesuis, both with a narrow host range (for humans and pigs, respectively) are phylogenetically related, a conclusion that was based on SNP analysis [6]. Indeed, we observe that one Paratyphi genome clusters with a Choleraesuis, but two other Paratyphi and another Choleraesuis genome are more distinct (colored red in **Figure 2**).

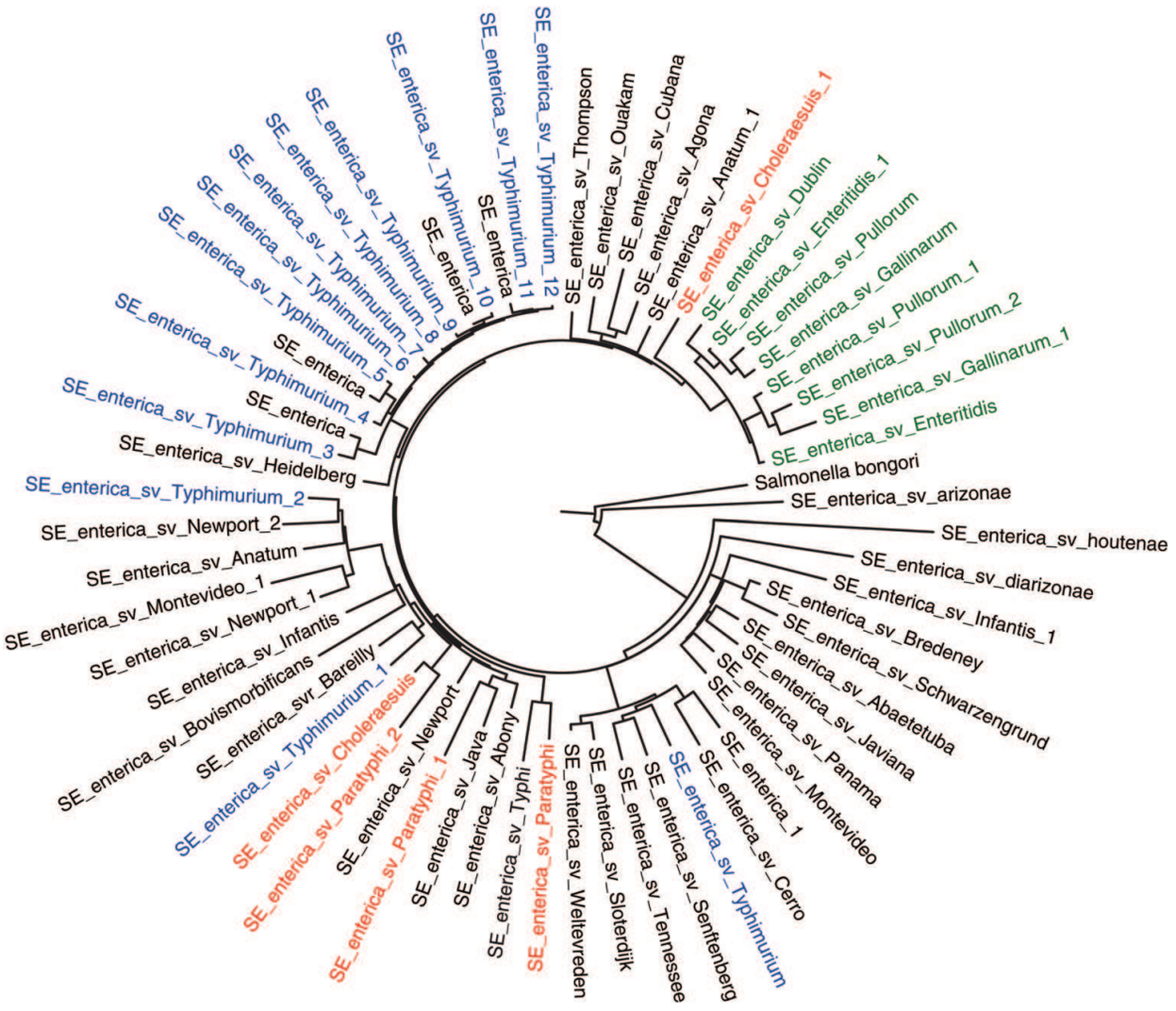

Figure 2. AAI tree of 365 *Salmonella* genomes representing 33 serovars of *S. enterica* (abbreviated as 'SE') subsp *enterica*. Indentical branches were collapsed per serotype. For explanation of the colors, see text.

2.2. Essential genes based on published gene inactivation studies

What makes a *Salmonella* a *Salmonella*? There are of course particular biochemical characteristics that can be used for identification, but can we recognize a set of genes that are always conserved, required and necessary for a *Salmonella* to be called that? And how many of those genes would be essential for growth and survival of the bacteria? These questions are addressed in this and the next session. Here, we start with genes proposed to be essential for survival under laboratory conditions, based on experimental data.

Traditionally, targeted mutagenesis has been used to determine if a gene from a given *Salmonella* strain were essential for infection, an approach that restricted the analyses to low numbers of genes only. An alternative approach was published in 2004 (based on previously developed techniques) to identify larger numbers of essential genes, by insertion of conditional lethal mutations into random gene fragments in a *S. typhimurium* strain [7]. The conditional switch used here was growth temperature, while tetracycline-dependent expression was used by others [8], although they only reported findings for four essential genes. A few years later, transposon (Tn) mutagenesis combined with high-throughput sequencing became available and this was applied to *S. enterica* strains [9–12]. Typically, in this approach mutants are screened for growth in LB broth. With a sufficiently high density of transposon insertions, genes that have not received insertions can be considered essential, as their inactivation had resulted in mutants unable to multiply under the conditions applied. Yet another approach was followed by Thiele and coworkers, who used metabolic reconstruction (MR) to extract a list of essential genes in *S*. Typhimurium that could be possible drug targets [13].

The experimental approaches reported in the literature are not without difficulties, as realized by their authors. For instance, polarity of transposon insertions in operons containing multiple genes can result in genes being scored as essential only because they are positioned downstream of an inactivated essential gene; attempts have been made to correct for this. Gene orthologs can further complicate findings, whereby one copy of an essential gene can be inactivated as long as a second copy remains intact. When an obtained mutant library is cultured for several generations, some mutants that originally survived will be removed from the population because their deletions are disadvantageous though not directly lethal. Such genes are typically scored as being under strong selection, an analysis that has been performed for *S*. Typhimurium strain ATCC 14028 and *S*. Thyphi strain Ty2 [11].

That experimental wet-laboratory data can be controversial is demonstrated by the fact that 26 of the 28 genes in *S*. Typhimurium strain ATCC 14028 that Knuth and coworkers reported as essential [7] could nevertheless be inactivated by site-directed mutagenesis [14].

Some research groups selected for conditions more closely resembling natural conditions of infection, for instance growth at 42°C instead of 37°C, to resemble the body temperature of mice that *S*. Typhimurium would typically encounter, or in the presence of bile acid ([10], work conducted with strain ATCC 14028). Exposure to low pH has also been tested [8]. Moreover, even 'essential' genes can often endure a transposon insertion without complete loss of function. If only those genes would be scored as essential that were truly resistant to Tn insertions from high-throughput mutagenesis, the essential gene pool would be very small indeed: only 96 genes from *S*. Typhi strain Ty2 and 57 genes from *S*. Typhimurium strain SL3261 remained

free of Tn insertions under conditions that were considered to have reached Tn saturation [12]. Thus, a small number of insertions can be permitted, even in genes considered essential for life in laboratory medium. Since the chance to receive a Tn insertion depends on gene length, a highly variable parameter, the number of observed insertions needs to be corrected for gene length [9]. This produces an insertion index, where the number of observed insertions is divided by gene length. In addition, a likelihood can be calculated from the ratio of observed versus expected number of Tn insertions, to predict the chance of a gene being essential [9, 12]. For this approach, a cutoff value is required, to bin genes as either essential or not. The problem with this is that the used parameter (likelihood P value, Tn-insertion index or both) is a continuously increasing value. This makes the choice of the cutoff inevitably arbitrarily: There is no biological reason why genes bordering this cutoff would or would not be essential.

To illustrate the difficulty, we plotted the P value reported by Barquist and colleagues [12], who provided the most elaborate list of Tn mutants available to date (**Figure 3**). Panel A of the figure shows how the P value of all genes of *S*. Typhimurium steadily increases. Similar results are obtained for *S*. Typhi (not shown), and even for those genes that have very low P values, there is a continuous increase, as shown in Panel B. Note that in this figure, the log10 value was plotted for clarity, and the cutoff value corresponding to a P value of <0.05 is indicated by the red line. Clearly, this value is artificial, since there is no noticeable increment around this value.

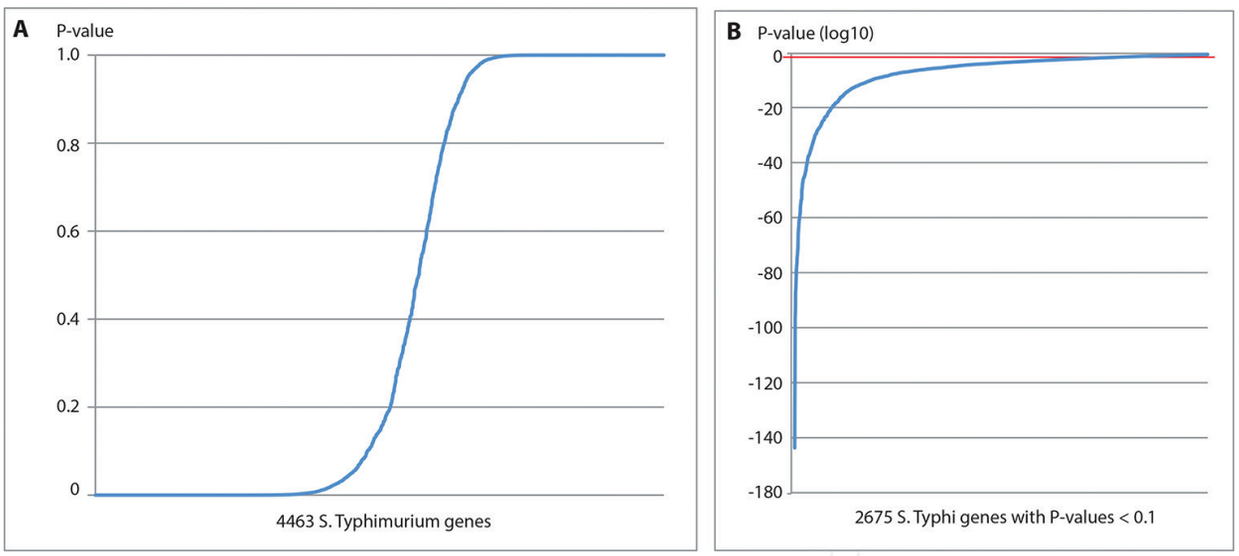

Figure 3. The continuous increase of P values of Tn insertions. In Panel A, P values of all 4463 genes of *S*. Typhimurium are plotted. In Panel B, a selection of 2675 *S*. Typhimurium genes is shown with P values >0 but <0.1, plotted for the exponent (log10) of the P values for clarity. The red line indicates the cutoff of P < 0.05, corresponding with a log10 value of −1.3 that was used by the authors. Data after Ref. [12].

A slightly different picture emerges when the Tn-insertion index is plotted, as shown in **Figure 4**. Although the increase in this index is also continuous, the shape of the obtained curve is slightly sigmoidal at the beginning, suggesting a trend toward saturation of the index value around 0.03, before it increases again. This trend is stronger for *S*. Typhi (Panel 4A) than for

S. Typhimurium (Panel 4B). Based on these findings, a cutoff value of 0.25 and 0.03 for the Tn index, respectively, might be appropriate for these species. We therefore recorded genes with a Tn index <0.25 for *S*. Typhi (n = 545 genes) and with a Tn index <0.30 for *S*. Typhimurium (n = 445), based on the data from Barquist and coworkers [12]. The Tn index of these genes is shown in Panels C and D of **Figure 4**. We further recorded the genes that Barquist and colleagues had originally selected (301 genes from *S*. Typhi and 299 for *S*. Typhimurium) which contained a reanalysis of the data from Langridge [9], as well as all genes previously identified as 'essential' by Knuth [7], Khatiwari [10], Canals [11] and Thiele [13], regardless of whether such genes were successfully inactivated by others. This produced an 'all inclusive' list of 847 genes putatively essential for growth and survival, or under strong selection, in LB medium. Relatively few genes were consistently recorded as essential by all or most authors; most genes were found in two independent approaches or were single findings (results not shown).

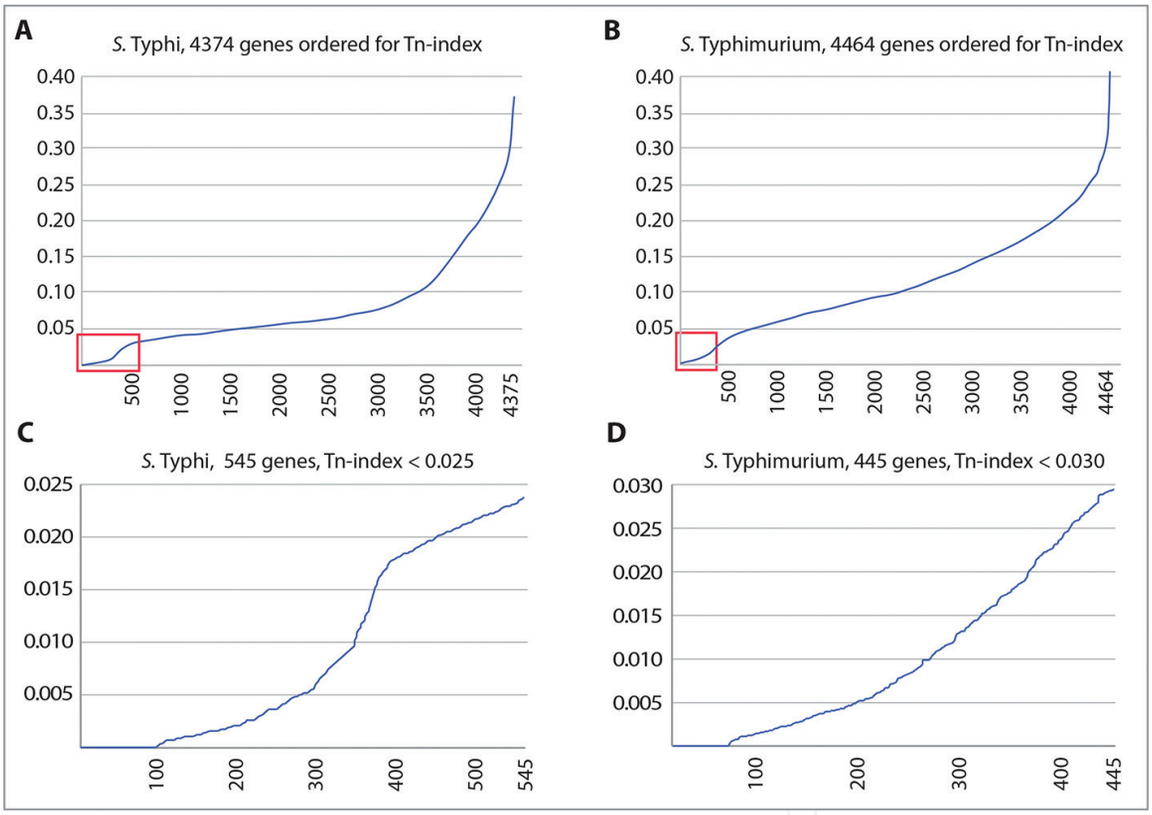

Figure 4. Analysis of transposon insertion frequency for genes of *S*. Typhi (left) and *S*. Typhimurium (right), based on data published by [12]. In Panels A and B, all genes are sorted for Tn index. The bottom Panels C and D show an enlargement of the part in the red square of A and B, respectively. For more explanation, see text.

A word of caution is needed here. It turned out to be rather cumbersome to identify the genes mentioned in the original published data (mostly using the supplementary tables provided with the publications) and to compare the findings with those of others, because genes were mostly described by gene names, which are by no means suitable as unique identifiers. For instance, the large operon for LPS-biosynthesis is called *waa* in *S*. Typhi but *rfb* in

S. Typhimurium; the essential gene *mrdA* of *E. coli* is called that in *S*. typhimurium, but it is *pbpA* in *S*. Typhimurium. The gene that is called *ribE* in both *Salmonella* genomes is essential, but it is called *ribC* in *E. coli*, while *ribE* in the latter species is called *ribH* in *Salmonella* (also essential). This makes it very risky to assume two genes are the same if they have the same name, or different if they do not. In most reports, a short protein functional description is provided, which can assist in correct identification, but many genes have very general functional characteristics, or are of unknown function. In such cases, the only way to identify which gene was meant is to use the gene location, but even that information does not always prove to be sufficient, for instance, when authors have re-annotated a genome but did not make this annotation public.

In conclusion, it is tedious and sometimes impossible to connect the findings from one study to those of another. Genes scored as 'essential' by one group can be inactivated without consequences on viability by another group. Moreover, most so-called essential genes endure a low number of transposon insertions without the loss of viability.

2.3. Conserved genes found in the core genome of *Salmonella enterica*

The second approach to identify essential genes in *Salmonella* is based on bioinformatical analysis of published genome sequences. If a gene is essential for growth, one can expect it to be strictly conserved between genomes, so a comparison on gene conservation can identify possible candidates. This is also not a completely unambiguous approach and depends on a number of choices that have to be made. For instance, one must define homologs between genomes in order to assess if genes are conserved, but this requires a defined percentage of homology that must be allowed and required for genes to be combined into a gene family. In addition, how should one deal with very short open reading frames, in other words, what is the minimum length of genes included, without adding too many artificial short open reading frames? And should one use original gene annotations, which is a transparent procedure that is easily reproducible, or is it better to re-annotate genomes using a standardized procedure to reduce variation? The latter approach produces more robust data as it no longer depends on variable gene calling, but it is less transparent when the used re-annotations are not made public. When core genomes are being defined from a set of highly different organisms, it may be required to allow for genes that are missing in a low number of analyzed genomes. However, when dealing with a single species, one could apply a strict requirement of presence in all genomes to produce a realistic core, especially if only fully sequenced genomes, re-annotated with a standardized algorithm, are included.

For this chapter, we decided to use publically available annotations, to aim for maximum transparency, and we further illustrate the effect of different core genome definitions. The core genome was established based on the annotations of the 362 completely sequenced *Salmonella enterica* genomes that were used to construct **Figure 2**, complemented with the three *S. bongori* genomes. Protein-coding genes were binned into gene families by the use of the program USEARCH [15] such that members of each family have at least 50% sequence identity and at least 50% alignment length of the best hit against the centroid of the family. Using a strict definition of required presence in all analyzed genomes, a so-called 100% core genome could

be identified that consisted of 1061 gene families. Although this seems an impressive number, it is lower than expected, probably because of variations in the used gene annotations. Based on our experience with core-genome determination from many bacterial genera, we were expecting the core genome of *S. enterica* to be larger, as the species contains relatively closely related organisms. Thus, we relaxed the requirement to allow gene presence in 344 or 95% of the investigated genomes. This produced a core genome of 3499 gene families, a size that is comparable with the preliminary core established for thousands of sequenced *Salmonella* genomes (S-R Jun and DW Ussery, unpublished data). We also constructed the core genome for *S. bongori*, but with only three genomes available, this core is relatively large, as a core genome usually decreases with an increasing number of included genomes. For the core genome of the complete *Salmonella* genus, these two datasets were combined. The results are summarized in **Table 1**.

Table 1 further lists that 11 genes from the 95% core were not annotated in the reference genome of the species typestrain *S. enterica* subsp *enterica* Typhimurium LT2. Originally, this number was much higher: There appeared to be 141 of the 3499 core genes missing in the annotated *S*. Typhimurium LT2 genome. However, when the DNA sequences of these genes were checked against the reference genome, 130 were actually present but not annotated. Thus, only 11 core genes remained that appear to be truly missing in the reference genome. This number did not change for core gene families based on *S. enterica* or the complete *Salmonella* genome (**Table 1**).

Dataset	Core genome size in 100% of dataset	Core genome size in 95% of dataset	Number of core genes missing in reference genome
362 *S. enterica* genomes	1061 gene families	3499 gene families	11 core genes out of 3499 are missing in *S*. Typhimurium LT2
3 *S. bongori* genomes	3368 gene families	3368 gene families	n.a.
365 *Salmonella* genomes	1009 gene families	3470 gene families	11 core genes out of 3470 are missing in *S*. Typhimurium LT2

Table 1. Core genome analysis based on 365 *Salmonella* genome sequences.

It was further checked if core gene families in the reference genome contained multiple entries, in other words, whether those core gene families contained orthologs or paralogs. This was the case for 120 gene families. When the function of these gene copies is interchangeable, these orthologs can be considered as 'back-up' copies, possibly maintained in the genome to protect against loss of essential function; alternatively, the genome can contain orthologs to allow for a higher production of the gene product. The multiple copies of the ribosomal RNA genes would be a nice example of the latter, though they are not captured in our core genome analysis, which was restricted to protein-coding genes only. To give another example, multiple copies of ferric enterobactin (enterochelin) transporters were found. Such orthologs of essential genes can complicate the outcome of in vitro mutagenesis analyses, as discussed above. However, not all orthologous genes are duplicated because they are essential, so it is not a predictive characteristic.

The genomes used for **Table 1** were not only used to select conserved core genomes, but also to define the pan genome, containing all gene families of the *Salmonella* genus. This is visually represented in **Figure 5**. The pan genome increases in size until approximately 180 genomes have been added, at which stage it reaches a plateau and is hardly affected by addition of further *S. enterica* genomes. It increases again when *S. enterica* Infantis and especially when *S. bongori* genomes are added, as these introduce novel gene families to the pan genome. Panel B of **Figure 5** illustrates the validity of defining a 95% core, instead of applying the strict requirement of presence in 100% of all genomes. The 100% core genome steadily decreases with the cumulative addition of the genomes analyzed here (the order of the genomes is the same as for Panel A) and decreases sharply to approximately 1000 gene families after addition of the *S. bongori* genomes. Instead, in the 95%, core genome is quite robust and remains more or less constant at around 3470 gene families (**Figure 5**).

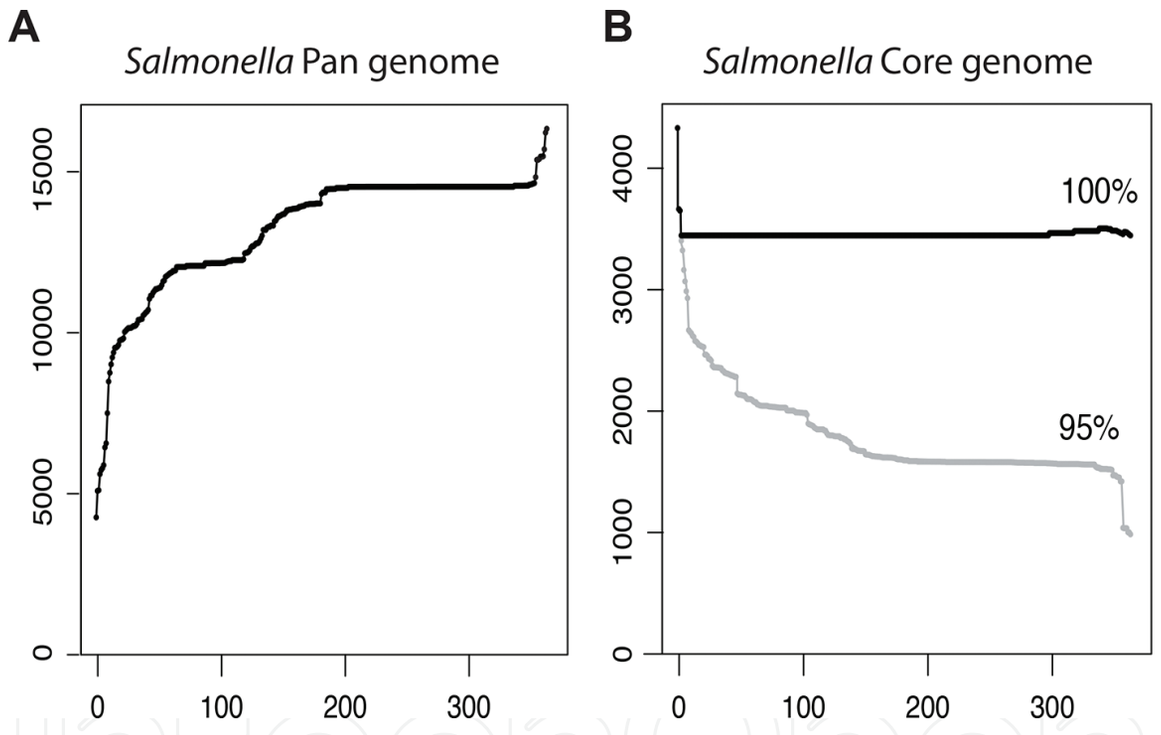

Figure 5. Pan-core plots based on 365 *Salmonella* genomes. Panel A shows the pan genome of *Salmonella*, with *S. bongori* added last. Panel B shows the core genome of the 365 *Salmonella* genomes with 95 and 100% conservation.

As was discussed in the previous section, the literature findings on essential genes are often controversial, for reasons discussed, while core genome determination is also not without caveats. Importantly, one can assume that all genes required for growth in LB medium must be conserved in all genomes and thus be part of the core, though the reverse may not be true: Not all core genes will be essential for growth and survival under these laboratory conditions. Therefore, we checked which of the essential genes reported in the literature were actually present in the core genome. For this, we used the 95% core genome, though core genes missing in the original annotation of the reference genome of *S*. Typhimurium LT2 were added manually. A total of 683 core genes could with reasonable confidence be identified that at least by one approach was found as putatively essential (results not shown). Conversely, of the

870 genes that were identified as essential by any of the methods discussed in the previous section, 694 were identified as part of the 95% core. The least reliable prediction of 'essential' genes turned out to be a low P value of Tn insertion, as this contained the highest fraction of genes that were not part of the core.

2.4. How close is *S*. Typhimurium to *E. coli*?

This chapter started with a comparison of all *Enterobacteriaceae*, to illustrate the close relationship between *Salmonella, Citrobacter* and *Escherichia*. But how close are *Salmonella* and *Escherichia*, in terms of conserved proteins? To address this question, the core genes of *S. enterica* Typhimurium LT2 (the type strain of the species) were compared to the core genes recently defined for *E. coli* (using the same definitions and parameters) [16], which we applied to the species typestrain *E. coli* DSM 30083. As reported in **Table 1**, the 95% core genome of all *Salmonella* comprises 3470 gene families, of which 11 are missing in Typhimurium LT2. This strain thus contains 3459 core gene families, while the *E. coli* typestrain contains 3100 core gene families. When these were compared, it was found that 2615 of these are shared, which corresponds to 75.6% of the *S*. Typhimurium LT2 core gene families, 84.4% of *E. coli* DSM 30083 and 66.3% of the total gene families assessed for these two species. This is illustrated in Panel A of **Figure 6**. The definition for gene families applied here is the same as for **Table 1** and **Figure 5**, but as explained above, this requires a defined cutoff for sequence similarity. The biological function of proteins is mostly defined by their functional domains, which is sometimes only a fraction of the total protein sequence. Thus, we narrowed this analysis down, to define the common core genome based on functional domains only, using Pfam domains. Since a Pfam domain is not described for all core genes, there were fewer domains captured in this comparison (2416 for *S*. typhimurium LT2 and 2263 for *E. coli* DSM 30083). Panel B of **Figure 6** shows that there are 2142 shared protein domains, corresponding to 88.7% of the *S*. Typhimurium LT2 core proteins, 94.7% of the *E. coli* DSM 30083 core proteins, and 84.4% of the total number of functional domains captured here. Interestingly, the fractions of shared core genes and shared functional domains are larger for the *E. coli* typestrain than for the *Salmonella enterica* typestrain. We believe this is caused by the larger diversity of the *E. coli* species, compared to *S. enterica*. As a consequence, the core genome of *E. coli* is smaller, even at 95%, which means a larger fraction of these is shared with *S. enterica*.

We further investigated the functions of the *Salmonella* core gene families in *S. Typhimurium* LT2 and found that most of them related to cellular metabolism. The core genome of *S. Typhimurium* LT2 was mapped to the genome-scale metabolic model SMT_v1.0 [13], which resulted in a total of 1271 genes and 2545 metabolic reactions. As shown in Panel C of **Figure 6**, 1012 genes from the *S. Typhimurium* LT2 core genome have a metabolic function (~80% of total genes in the model) and these account for 2358 metabolic reactions (93% of total reactions in the model). When comparing this with the *E. coli* core genome, *S. Typhimurium* LT2 has 156 unique metabolic genes, responsible for 452 metabolic reactions. The unique metabolic reactions that were identified here are mostly involved in transport systems across the inner membrane as well as the outer membrane (porins), specific transport of inorganic ions, and the recycling of lipopolysaccharide biosynthesis components. Such analyses can share light on the biochemical and metabolic properties that *Salmonella* is specialized in, related to its intracellular lifestyle.

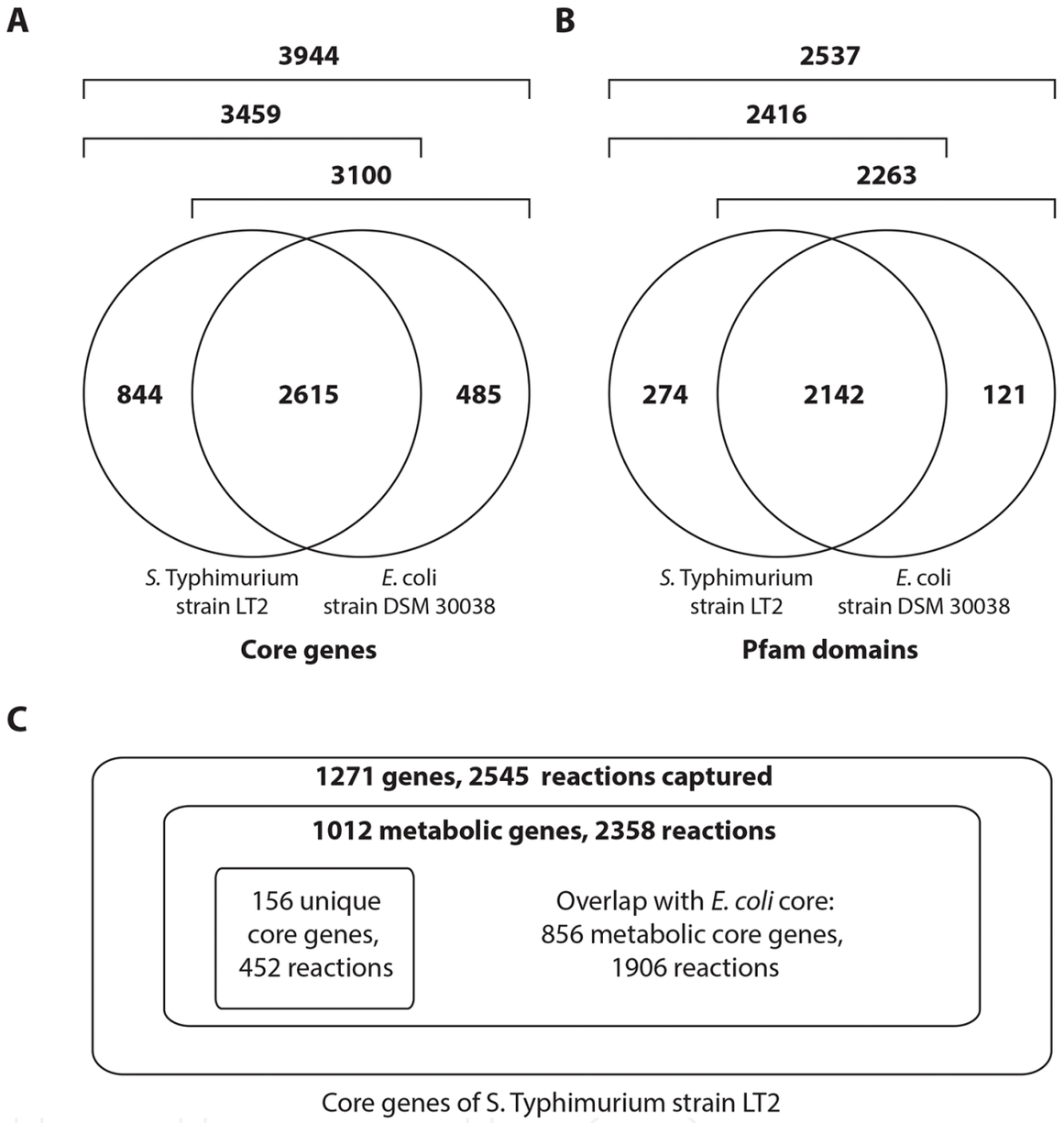

Figure 6. Comparison of *Salmonella* and *E. coli* core genes, using the type strains for both species. Panel A shows the size and overlap of the core gene families. Panel B shows the comparison using PfamA domains. Panel C summarizes how many metabolic pathways are shared in the *Salmonella* and *E. coli* cores.

2.5. Conserved RNAs across 201 *Salmonella* genomes

So far, all analyses were based on the annotated proteomes of the *Salmonella* genomes, but genes that code for RNA as the final product should not be ignored. A genome annotation would not be complete without its ribosomal RNA genes, coding for 5S, 16S and 23S RNA, as well as the tRNA genes. *Salmonella enterica* contains 7 *rrn* operons, which is more than can be found in many bacterial species but certainly is not a maximum, as some soil bacteria can contain up to 15 copies of the rRNA genes. The number of *rrn* copies of bacterial species

has been related to their capacity to change their metabolism to use available resources [17]. Although it is often assumed that these gene duplications are all identical, in fact some degree of sequence variation can be observed, even within a genome. For *Salmonella*, it was reported that the gene encoding 16S rRNA (which is typically used for taxonomic description) is conserved for 97% only [18]. The gene coding for 23S rRNA is also not strictly conserved in *Salmonella*, as it contains both point mutations and indels [19].

The number of tRNA genes present in the *Salmonella* reference genome is 85, representing 47 different tRNA molecules that together cover the 40 required anticodons [20]. These numbers can vary between genomes and serovars. But these are not the only bacterial genes that are never translated into protein. In addition to essential RNA genes such as the gene coding for tmRNA (transfer-messenger RNA, required for correct protein translation), it is now recognized that bacterial genomes contain a large number of small RNA genes (sRNA) that are not always annotated. These are often involved in post-transcriptional regulation of gene expression [21]. As a final analysis, we decided to assess the conservation of these, incorrectly neglected, RNA genes.

The bioinformatic analysis performed was based on a publication where transcription start sites were identified from 31 *Salmonella* genomes [22]. We analyzed those 113 RNA genes in the 201 completely sequenced genomes. For this analysis, we excluded the nearly completed sequences that had been included in the analyses resulting in **Figures 2** and **5**, because genome assembly is biased toward protein-coding regions, so that regions on which sRNA genes may reside are likely to be missed, unless a genome is truly completed. For comparison, eight other *Enterobacteriaceae* were included. The results are presented in a matrix heat map (**Figure 7**). Based on their sRNA content, most of the genomes neatly clustered according to their serotype, with only few exceptions. Interestingly, the genomes of strains FORC-015 and FORC-020, which are annotated as Typhimurium, are placed outside the Typhimurium cluster in **Figure 7**, and these were also placed outside the main Typhimurium cluster in the AAI tree of **Figure 2**. Thus, it can be questioned if the serotype of these two strains was correctly identified. That most of the *Salmonella* genomes are nicely clustered according to their serotype in **Figure 7** is surprising, as the nonprotein coding sRNA genes analyzed here do not have a specific role in expression of surface antigens. The correlation identified here is in line with a publication that sRNA genes can be used as targets for serotype-specific PCR detection of Typhi and Paratyphi [23]. It was recently described that some sRNA genes of *S.* Typhimurium are under regulation of Sigma 28, and there is extensive cross talk between genes of the *Salmonella* pathogenicity pathways SPI1 and SPI2 and particular sRNA genes [24]. In this context, it is surprising that the sRNA genes are so strongly conserved throughout the *Salmonella* genomes (illustrated by the dominant red in **Figure 7**), whereas the presence of SPIs widely varies across serotypes [24]. This suggests that sRNA genes are strongly conserved and may well belong to the collection of essential genes, though this has not yet been experimentally demonstrated. The analysis further showed that the sRNA genes are specific for the *Salmonella* genus, and bear relatively little resemblance with the other *Enterobacteriacea* members included at the bottom of the figure.

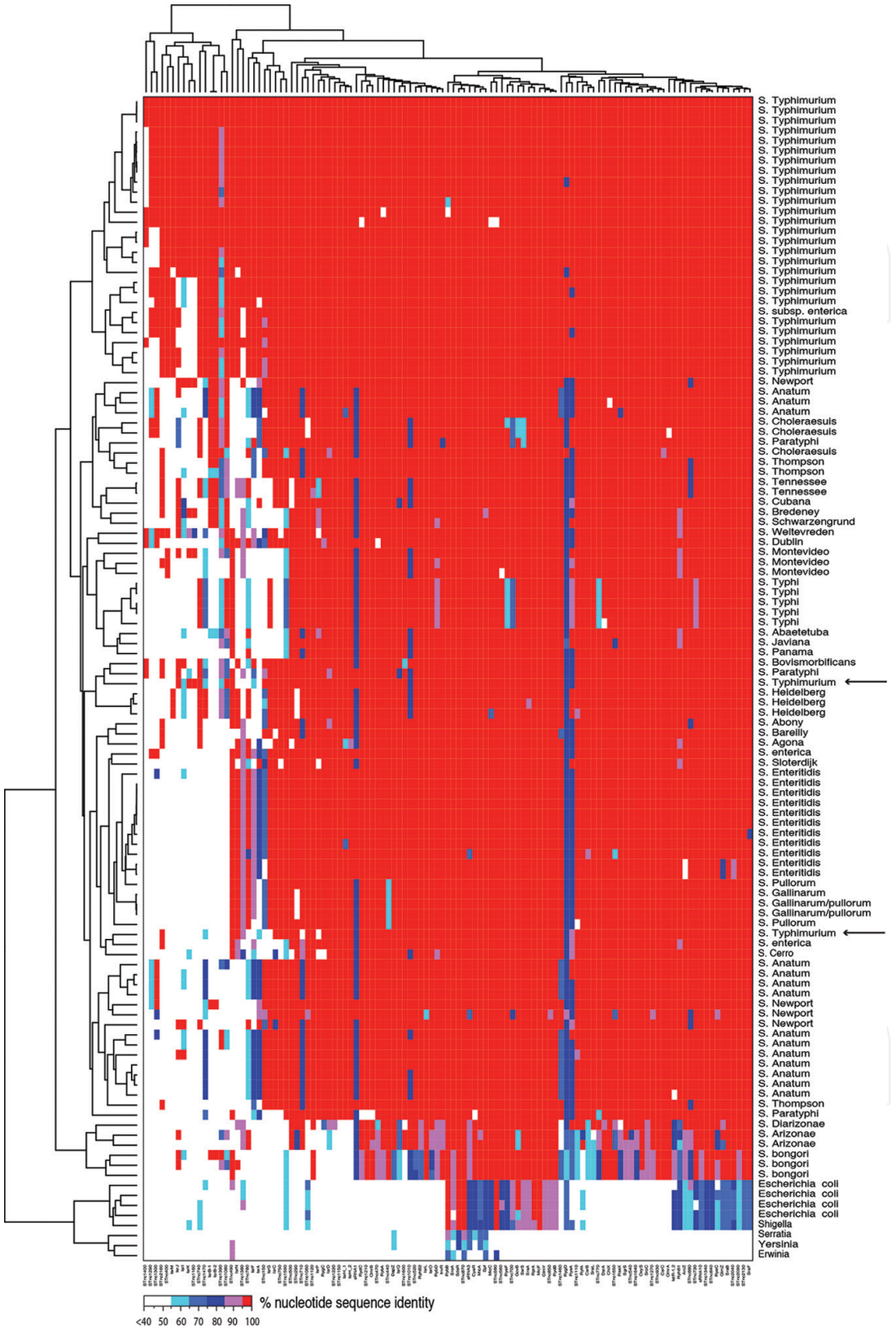

Figure 7. Conserved sRNAs across 201 *Salmonella* genomes. The tree to the left mostly clusters serotypes together, based on their sRNA genes. Two wrongly placed *S*. Typhimurium genomes are pointed out by the arrows to the right. The tree at the top identifies clusters of related sRNA genes. The eight genomes at the bottom are from other *Enterobacteriaceae*.

3. Conclusions

Based on genomic average amino acid identity (AAI), *Salmonella* genomes appear as a distinct clade within the enterics, closely related to the *Citrobacter* genus. The serovars of *S. enterica* subsp. *enterica* generally cluster together when analyzed for AAI. There is a stable core set of about 3400 gene families, found in nearly all *Salmonella enterica* genomes, and these genes are on average 99% or more identical to each other across all the *Salmonella* genomes. Further, many of these genes seem to be involved in metabolic processes, and the core genes account for about 80% of the total genes of the *Salmonella* genome-scale metabolic model. Finally, we examined small RNA conservation and found the same clustering of outlier genomes (e.g., particular *S*. Typhimurium strains) that were observed in the AAI analysis.

Acknowledgements

This work has been funded in part by The Arkansas Research Alliance and UAMS.

Author details

Trudy M. Wassenaar[1]*, Se-Ran Jun[2], Visanu Wanchai[2], Preecha Patumcharoenpol[2], Intawat Nookaew[2], Katrina Schlum[3], Michael R. Leuze[3] and David W. Ussery[2]

*Address all correspondence to: trudy@mmgc.eu

1 Molecular Microbiology and Genomics Consultants, Zotzenheim, Germany

2 Department of BioMedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA

3 Computing Science and Mathematics Division, Oak Ridge National Labs, Oak Ridge, Tennessee, USA

References

[1] Dekker JP, Frank KM. *Salmonella, Shigella* and *Yersinia*. Clin Lab Med. 2015;**35**:225–246. doi:10.1016/j.cll.2015.02.002

[2] Crosa JH, Brenner DJ, Ewing WH, Falkow S. Molecular relationships among the *Salmonelleae*. J Bacteriol. 1973;**115**:307–315.

[3] International Committee on Systematics of Prokaryotes. The type species of the genus *Salmonella Lignieres* 1900 is *Salmonella enterica* (ex Kauffmann and Edwards 1952) Le Minor and Popoff 1987, with the type strain LT2T and conservation of the epithet *enterica* in *Salmonella enterica* over all earlier epithets that may be applied to this species. Opinion 80. Int J Syst Evol Microbiol. 2005;**55**:519–520. doi:10.1099/ijs.0.63579-0

[4] Konstantinidis KT, Tiedje JM. Towards a genome-based taxonomy for prokaryotes. J Bacteriol. 2005;**187**:6258–6264. doi:10.1128/JB.187.18.6258-6264.2005

[5] Retchless AC, Lawrence JG. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proc Natl Acad Sci USA. 2010;**107**:11453–11458. doi:10.1073/pnas.1001291107

[6] Foley SL, Johnson TJ, Ricke SC, Nayak R, Danzeisen J. *Salmonella* pathogenicity and host adaptation in chicken-associated serovars. Microbiol Mol Biol Rev. 2013;**77**:582–607. doi:10.1128/MMBR.00015-13

[7] Knuth K, Niesalla H, Hueck CJ, Fuchs TM. Large-scale identification of essential *Salmonella* genes by trapping lethal insertions. Mol Microbiol. 2004;**51**:1729–1744.

[8] Hidalgo AA, Trombert AN, Castro-Alonso JC, Santiviago CA, Tesser BR, Youderian P, Mora GC. Insertions of mini-Tn10 transposon T-POP in *Salmonella enterica* sv. typhi. Genetics. 2004;**167**:1069–1077. doi:10.1534/genetics.104.026682

[9] Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK. Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants. Genome Res. 2009; **19**:2308–2316. doi:10.1101/gr.097097.109

[10] Khatiwara A, Jiang T, Sung SS, Dawoud T, Kim JN, Bhattacharya D, Kim HB, Ricke SC, Kwon YM. Genome scanning for conditionally essential genes in *Salmonella enterica* Serotype Typhimurium. Appl Environ Microbiol. 2012;**78**:3098–3107. doi:10.1128/AEM.06865-11

[11] Canals R, Xia XQ, Fronick C, Clifton SW, Ahmer BM, Andrews-Polymenis HL, Porwollik S, McClelland M. High-throughput comparison of gene fitness among related bacteria. BMC Genomics. 2012;**13**:212. doi:10.1186/1471-2164-13-212

[12] Barquist L, Langridge GC, Turner DJ, Phan MD, Turner AK, Bateman A, Parkhill J, Wain J, Gardner PP. A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium. Nucleic Acids Res; 2013;**41**:4549–4564. doi:10.1093/nar/gkt148

[13] Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming RM, Hsiung CA, De Keersmaecker SC, Liao YC, Marchal K, Mo ML, Özdemir E, Raghunathan A, Reed JL, Shin SI, Sigurbjörnsdóttir S, Steinmann J, Sudarsan S, Swainston N, Thijs IM, Zengler K, Palsson BO, Adkins JN, Bumann D. A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella Typhimurium* LT2. BMC Syst Biol. 2011;**5**:8. doi:10.1186/1752-0509-5-8

[14] Santiviago CA, Reynolds MM, Porwollik S, Choi SH, Long F, Andrews-Polymenis HL, McClelland M. Analysis of pools of targeted *Salmonella* deletion mutants identifies novel genes affecting fitness during competitive infection in mice. PLoS Pathog. 2009;**5**(7):e1000477. doi:10.1371/journal.ppat.1000477

[15] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;**26**:2460–2461. doi:10.1093/bioinformatics/btq461

[16] Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;**15**:141–161. doi:10.1007/s10142-015-0433-4

[17] Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol. 2000;**66**:1328–1333.

[18] Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. PLoS One. 2013;**8**(2):e57923. doi:10.1371/journal.pone.0057923

[19] Pei A, Nossa CW, Chokshi P, Blaser MJ, Yang L, Rosmarin DM, Pei Z. Diversity of 23S rRNA genes within individual prokaryotic genomes. PLoS One 2009;**4**(5):e5437. doi:10.1371/journal.pone.0005437

[20] Withers M, Wernisch L, dos Reis M. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. RNA. 2006;**12**:933–942. doi:10.1261/rna.2272306

[21] Papenfort K, Vogel J. Regulatory RNA in bacterial pathogens. Cell Host Microbe. 2010;**8**:116–127. doi:10.1016/j.chom.2010.06.008

[22] Kröger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, Colgan A, Leekitcharoenphon P, Langridge GC, Lohan AJ, Loftus B, Lucchini S, Ussery DW, Dorman CJ, Thomson NR, Vogel J, Hinton JC. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proc Natl Acad Sci USA. 2012;**109**:1277–1286. doi:10.1073/pnas.1201061109

[23] Nithya R, Ahmed SA, Hoe CH, Gopinath SC, Citartan M, Chinni SV, Lee LP, Rozhdestvensky TS, Tang TH. Non-protein coding RNA genes as the novel diagnostic markers for the discrimination of *Salmonella* species using PCR. PLoS One. 2015;**10**(3):e0118668. doi:10.1371/journal.pone.0118668

[24] Blondel CJ, Jiménez JC, Contreras I, Santiviago CA. Comparative genomic analysis uncovers 3 novel loci encoding type six secretion systems differentially distributed in *Salmonella* serotypes. BMC Genomics. 2009;**10**:354. doi:10.1186/1471-2164-10-354

IntechOpen