

Comparative genomics of *Salmonella enterica* serovars Paratyphi A, Typhi and Typhimurium reveals distinct profiles of their pangenome, mobile genetic elements, antimicrobial resistance and defense systems repertoire

Charles Coluzzi^{a,b}, Bar Piscon^{c,d}, Sandra Dérozier^a, Hélène Chiapello^a, and Ohad Gal-Mor ^{c,d}

^aUniversité Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France; ^bMicrobial Evolutionary Genomics, Institut Pasteur, Université Paris Cité, CNRS, Paris, France; ^cThe Infectious Diseases Research Laboratory, Sheba Medical Center, Tel-Hashomer, Israel; ^dDepartment of Clinical Microbiology and Immunology, Faculty of Medical & Health Sciences, Tel-Aviv University, Tel-Aviv, Israel

ABSTRACT

Salmonella enterica (*S. enterica*) is a highly ubiquitous and diverse animal and human pathogen. Distinct *S. enterica* serovars may present varying host-specificity and cause different diseases. While the human-restricted serovars *S. Typhi* (STY) and *S. Paratyphi A* (SPA) cause in humans a systemic life-threatening enteric fever, the host-generalist serovar, *S. Typhimurium* (STM) causes in immunocompetent individuals a self-limited gastroenteritis. Here, we have performed whole-genome sequencing and hybrid assembly of new SPA and STY typhoidal strains and took a comparative genomics approach to examine their phylogeny, pangenome structure and accessory genome content in comparison to the reference non-typhoidal serovar, STM. Our results identified previously uncharacterized lineages of SPA and refined the presence and distribution of core pseudogenes in typhoidal serovars. Pangenome analysis showed that while these serovars have a relatively similar core-genome size, the accessory genome of STM is more than four times larger than those of typhoidal *Salmonellae* and that STY and SPA display a more closed pangenome than STM. Unexpectedly, we demonstrate that STY and SPA present distinct differences in their pangenome composition, with a noticeable lower number of prophages, conjugative elements and antimicrobial genes per genome in SPA vs. STY. These results suggest that although SPA and STY are closely related at the DNA level, share a similar lifestyle and cause a symptomatic-indistinguishable disease, their genomic evolution and accessory genomes are markedly different. Moreover, these results may provide genomic explanation to phenotypic and epidemiological differences in antimicrobial resistance profiles associated with these serovars globally.

ARTICLE HISTORY

Received 6 August 2024

Revised 5 March 2025

Accepted 6 May 2025

KEYWORDS


Salmonella enterica;
pangenome; mobile genetic
elements; plasmids;
antimicrobial resistance
genes; pseudogenes

Introduction

The bacterial species *salmonella enterica* (*S. enterica*) is a highly ubiquitous animal and human foodborne pathogen. This diverse species is serologically classified into more than 2,600 different serovars (also known as serotypes) according to surface antigens expressed on their lipopolysaccharide (O antigens), flagella (H antigens) and capsule (Vi antigen) [1], following the Kauffmann-White-Le Minor serotyping scheme [2]. Different *S. enterica* serovars can be clinically categorized according to the disease they cause in humans. Typhoidal *Salmonella* serovars including *S. enterica* serovar Typhi (*S. Typhi*), *S. enterica* serovar Paratyphi A (*S. Paratyphi A*) and *S. enterica* serovar Sendai (*S. Sendai*) infect only humans and higher primates and cause a life-threatening bloodstream infection (bacteremia) and systemic disease, known as enteric fever

that can also be referred to as typhoid (following *S. Typhi* infection) or paratyphoid (in case of a *S. Paratyphi A* infection) fever. This manifestation is a non-inflammatory disease, in which typhoidal *Salmonellae* colonize the spleen, liver, and mesenteric lymph nodes (MLN) [3]. In contrast, many of the non-typhoidal serovars (NTS) like *S. enterica* serovar Typhimurium (*S. Typhimurium*) or *S. enterica* serovar Enteritidis (*S. Enteritidis*) can infect a wide range of animal and human hosts [4]. Infection of immunocompetent humans will lead, in most cases to a short-term gastroenteritis, presented as an acute inflammation of the terminal ileum and colon. Nevertheless, a malfunction of the mucosal barrier in immunocompromised individuals can result in a complication of a potentially fatal bacteremia [5,6]. Overall, *Salmonella* is still a leading cause of foodborne infections, with an

CONTACT Ohad Gal-Mor  Ohad.Gal-Mor@sheba.health.gov.il; Hélène Chiapello  helene.chiapello@inrae.fr

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/21505594.2025.2504658>

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

annual global Salmonellosis burden of 78.7 million cases of gastroenteritis [7] and over 27 million cases of enteric fever [8].

Differences in virulence and disease manifestation caused by different *Salmonella* serovars are yet not fully understood. While studies have tried to explain the prominent clinical differences between typhoidal and NTS serovars, by the distinct presence of key virulence factors such as the typhoid toxins CdtB and PltAB [9] or the type III secretion system effector GtgE [10], one of the emerging concept that is common to all known host-specific typhoidal serovars is genomic decay. This genetically reducing process is characterized by multiple events of gene inactivation (pseudogene formation) and/or complete gene deletion from the genome [11–14]. Pseudogene formation resulted in loss of gene function and complete gene deletion seems to be a genetic signature of host-specific pathogens in comparison to their genetically related host-generalist strains [15–18]. Pioneering comparative genomics approaches have demonstrated that while the generalist *Salmonella* serovar *S. Typhimurium* has a relatively low number of pseudogenes (54 in strain SL1344, or 25 in strain LT2), the typhoidal serovar, *S. Typhi* (strain CT18) chromosome was predicted to encode 204 pseudogenes, comprising about 4.5% of its coding genes [15]. Different *S. Paratyphi* A isolates were also shown to harbor 173 inactivated genes in strain ATCC 9150 [16], or 204 pseudogenes in strain AKU12601 [19], consisting of 4 and 4.8% of annotated coding sequences (CDS), respectively. Moreover, *S. Paratyphi* A (strains AKU12601 and ATCC9150) and *S. Typhi* (strains CT18 and Ty2) were found to share 66 common pseudogenes, which were formed as a result of a convergent evolution due to a similar lifestyle in the same host [19].

Here, we performed whole-genome sequencing (WGS) and hybrid assembly of new *S. Paratyphi* A and *S. Typhi* strains and took a comparative genomics approach to examine the phylogeny, pangenome structure and the accessory genome content of *S. Typhi* (STY) and *S. Paratyphi* A (SPA) in comparison to a reference NTS serovar, *S. Typhimurium* (STM). Our results identified previously uncharacterized lineages of SPA and refined the dominant presence and distribution of pseudogenes in typhoidal serovars. Moreover, we demonstrate that although STY and SPA are considered to be closely related at the DNA level [16,19], they present distinct differences in their content of plasmids, prophages, antimicrobial resistance genes and defense systems. These results suggest that although SPA and STY share a similar lifestyle and cause a symptomatic-indistinguishable disease, their genomic evolution and their accessory genome content are distinct, in a way that may explain phenotypic and global epidemiological differences between these serovars.

Materials and methods

Sequencing, assembly and annotation of SPA and STY isolates

SPA45157 is a clinical isolate of *S. Paratyphi* A, which was isolated in 2009 from a hospitalized patient as part of a paratyphoid outbreak among Israeli travelers in Nepal [20]. SPA45157 genomic DNA was extracted with a GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich-Merck) and sequenced using a MiSeq V3 illumina sequencer and a paired-end library (2×300bp). We obtained 3,137,216 raw reads representing 944,302,618 bases of the SPA45157 genome.

SPA45157 was also sequenced using an Oxford nanopore MinION platform (GridION sequencer) with a flo-min106 (R9.4.1 revD) flow-cell and a sqk-lsk109 (1D) sequencing kit. We obtained a total of 722,848 reads representing 5,755,901,734 bases. In total, more than 98% of the full genome had at least 100 × of coverage, allowing high-quality assembly. Both illumina and long raw reads were quality controlled with Trimmomatic version 0.39. Quality controlled illumina reads were assembled for each isolate with SPAdes version 3.15.3 with the default parameters. These assemblies were further polished using pilon version 1.2.4 with minimum number of flank bases of 10 and kmer size of 47. Resulting contigs and long reads were used to perform a hybrid assembly using Unicycler version 4.7, which used SAMtools version 1.9, and bowtie2 version 2.4.4 modules. The final assembly was then repolished with pilon using illumina sequencing reads. Consecutive runs of pilon were performed until the polishing converged (no bases were further corrected after 3 runs). SPA45157 genome (accession number CP156168) was annotated with prokka version 1.14.6 [21]. Functional annotation was then performed using eggNOG-mapper version 2.1.9 [22] with default parameters.

STY120130191 is a clinical isolate of *S. Typhi*, which was isolated at 2012 from the blood of a patient with typhoid fever. STY120130191 genomic DNA was extracted using GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich-Merck) and SMRT DNA libraries were constructed according to the Pacbio standard protocol with BluePippin length selection. Sequences were generated on a Pacbio RSII instrument using P6-C4 chemistry. Using 3 PacBio SMRT® cells, we obtained 160,468 raw reads consisting 1,887 million bases. Reads that were larger than 22 kb were selected for structural assembly using HGAP3 + Quiver. Pacbio reads with length >500 bp were used only for polishing. Plasmid and genome circularization was performed using Circlator. STY120130191 was also sequenced using an

Illumina HiSeq 3000 paired-end strategy. Illumina reads were used to correct the Pacbio assembly using Pilon [23]. STY120130191 genome (Accession number CP156169 for the chromosome, and CP156170 for the plasmid) was annotated with EugenePP version 1.2 [24] with default parameters. Functional annotation was then performed using InterProScan version 5.15–54.0, COG version COG 2014 [25].

Paratyphi A, typhi and Typhimurium genome datasets

A total of 155 Paratyphi A (SPA) public genomes were downloaded from NCBI RefSeq database (23 May 2021) using NCBI dataset tool version 13.6.0. We added to this dataset the SPA45157 new genome and obtained a dataset of 156 SPA genomes. Since some of the analyses (e.g. pseudogene and MGE identification) required high-quality sequences, we used dRep version 3.2.2 [26] to limit redundancy in the SPA genomes and select the best representative genomes based on assembly quality and genomic distance. Briefly, we chose genomes with high completeness, low contamination, few contigs, and/or high coverage depth, prioritizing these metrics in this order. We selected assemblies with a minimum N50 value of 44,653 bases, maximum 244 contigs, 188 scaffolds, and a maximum proportion of gaps 0.54%. *dRep* computed two successive clustering using genomic distance computed with the fastANI algorithm (k-mer-based approach) [27]. The first average nucleotide identity (ANI) threshold to form primary clusters was set to 99.9% ANI and the second threshold that determines identical genomes in primary clusters was set to 99.99% ANI. The second clustering produced 134 clusters that were used to pick up the best quality representative genomes using assembly quality and cluster centrality. This process permits to remove low-quality genomes that are more than 99.99% identical to at least one genome in the dataset. We finally obtained 134 dereplicated “representative” genomes of SPA including our SPA45157 assembly.

We downloaded NCBI RefSeq public Typhi (STY) and Typhimurium (STM) genomes exhibiting an assembly level equal to “Chromosome” or “Complete” using NCBI dataset tool version 13.6.0. We obtained 242 genomes for Typhi and 443 genomes for Typhimurium. We applied the same dereplication process as described above for SPA and obtained 43 dereplicated STY genomes and 164 dereplicated STM genomes. Three STM genomes (GCA_017094545.1_ASM1709454v1, GCA_017094565.1_ASM1709456v1, GCA_017094585.1_ASM1709458v1)

were removed because of outlier value of genes and pseudogenes.

As an outgroup of our dataset, the reference genome of *Salmonella bongori* (strain N268–08; RefSeq NC_021870.1) was added. Consequently, our final dataset includes 339 dereplicated “representative” genomes composed of: 134 of SPA, 43 of STY, 161 of STM and 1 of *S. bongori* as an outgroup (Table S1).

Pangenome construction and phylogenetic analyses

To perform homogeneous structural annotation of all SPA, STY and STM dereplicated genomes, and *S. bongori*, the Prokka version 1.14.6 [21] was used. We then used *Roary* version 3.13.0 [28] with the *-s* and *-e* parameters to compute the pangenome on the three datasets together and separately (SPA, STY, STM). We then computed the phylogenetic trees using *iqtree* version 2.2.0.3 and the following parameters: *-m* MFP *-nt* 8 *-bb* 1000 and the aligned core genes produced by *Roary* as input dataset.

We computed SPA lineage definition using *fastbaps* (R package version 1.0.8) [29] and the *multi_res_baps* function to perform Bayesian hierarchical clustering of population structure successively at multiple resolutions. We choose *fastbaps* level 2 results that were consistent with Zhou et al. [27] previous study and the Maximum Likelihood Phylogenetic tree. All phylogeny figures were produced using the *ggtree* version 3.8.2 R package.

Pseudogene annotation

We developed a new workflow named *Pseudoscreen* to detect pseudogenes in each serovar genome datasets. The workflow relies on a reference dataset of supposed functional genes composed of the representative genes of the pangenome built with *Roary* version 3.13.0 on all complete 342 dereplicated genomes of the SPA, STY, and STM genomes plus the *S. bongori* genome as an outgroup. For each pangenome family, one representative protein sequence was selected to represent the most common sequence size of each pangenome family. For pangenome families with a binomial size distribution, two sequences were selected. This resulted in a dataset of 9194 non-redundant protein reference sequences representing the pangenome families of serovars SPA, STY and STM.

We then used our protein reference dataset to detect pseudogenes in all query genomes of the SPA, STY and STM datasets. Briefly, each genome was aligned using *tblastn* version 2.12.0 [30] against every sequence of the reference pangenome dataset.

Every alignment with an identity superior than 95% was post-filtered. We used filters described below to detect altered proteins according to this reference sequence. We considered six categories of events to predict pseudogenes: (i) Frameshift: nucleotide deletion not divisible by 3 that kept more than 80% of the protein compared to the reference; (ii) Premature stop codon: more than 70% of the protein sequence is aligned but includes a premature stop codon; (iii) in-frame medium or large deletions: in-frame deletion of 5 amino acids or more that kept less than 30% of the reference protein; (iv) in-frame small deletion (altered genes or allelic variation): in-frame small deletion of 5 amino acids or less; (v) Truncated protein: more than 30% of the protein was deleted at the beginning or at the end of the protein relative to the reference; (vi) large deletions scattered along the protein: more than 60 amino acids covering one or multiple sites that kept at least 30% of the protein compared to the reference.

Pan-pseudogenome and core-pseudogenome

For each serovar, we looked at each reference gene category (core, soft-core, shell, and cloud genes) for all pseudogenes annotated in the different serovars. When a pangenome category was annotated as pseudogene in more than 99% of the strains of a serovar, it was considered as a core pseudogene for the serovar. Categories pseudogenized in 95% to 99% of the genomes of a certain serovar were considered as soft-core pseudogenes, categories pseudogenized in 15% to 95% of the genomes of a serovar were considered as cloud pseudogenes, and categories pseudogenized in less than 15% of the genomes of a serovar was considered as shell pseudogenes. To compute the core and pan-pseudogenome common to STY and SPA, the same thresholds were applied. For example, core pseudogenes common to SPA and STY correspond to pangenome categories that are pseudogenized in more than 99% of the 177 genomes of the STY/SPA dataset (43 STY genomes + 134 SPA genomes).

Gene ontology (GO) pathway enrichment analysis of pseudogenes

The PANTHER software [31] version 19.0 in the Gene Ontology portal (<https://www.geneontology.org>) was used to perform a functional enrichment analysis of SPA and STY pseudogene families. Briefly, we analyzed a subset of STY and SPA pseudogenes that were represented as annotated functional genes in the reference

Salmonella Typhimurium genome at the PANTHER database and chose the “complete GO biological process” as annotation data set. Functional enrichment analysis was done using the Fischer’s exact test and a standard correction for multiple testing based on False Discovery Rate (FDR). All GO categories with corrected FDR p-values ≤ 0.05 were considered to be significantly enriched in SPA/STY pseudogenes.

Mobile genetic elements content analysis

The presence of prophage was predicted using VirSorter2 [32] version 2.2.3 (default parameters with the “— min-length 1500” option). The quality and the integrity of the viral regions were then assessed using CheckV version 1.0.1 default parameters [33]. After assessment of the quality and the integrity, high quality and complete prophage regions were clustered using VContact2, default parameters and version 0.9.19 [34]. Chi-Squared test was performed to assess whether the distribution of viral clusters differed significantly across various geographic regions. To quantify the strength of these associations, we calculated Cramer’s V score. All analyses were executed using Python, with a significance threshold set at $p < 0.05$. Heatmaps were plotted using seaborn heatmap package.

Protein sequences corresponding to known prophage-associated virulence factors (GipA, GogB, SodC, SopE, SopE2, SseI, SseK3, SspH1, and SspH2) were retrieved from the NCBI database. These sequences were then compared against prophage protein sequences predicted by VirSorter2 and validated with CheckV using BLASTp. Matches exhibiting at least 95% sequence identity and 95% query coverage were considered as virulence factors encoded by prophages.

The type of plasmids was classified as Conjugative (pConj), mobilizable (pMob) or decayed conjugative (pdConj) plasmids using the CONJScan module of MacSyFinder2 [<https://peercommunityjournal.org/articles/10.24072/pcjournal.250/>] version 2.1.1 with the “Plasmid” parameters.

For the detection of Conjugative operon and lone relaxases in the chromosome, we used the CONJScan module of MacSyFinder2 with the “Chromosome” parameters.

Classification of plasmids

The taxonomic classifier of plasmids, COPLA version 1.0, was used to assign plasmids to taxonomic units with default parameters [35].

Defense systems annotation

Defense systems were annotated using DefenseFinder with default parameters (last update 10 October 2023) [36].

Conjugation assay

The donor (*E. coli* K1037/pN3, *E. coli* FS1290/pRP4, *E. coli* DH10B/pCVM29188, *S. Typhimurium* SL1344/pESI, *E. coli* C600/pRK2) and recipient (*S. Typhimurium* LT2, *S. Typhimurium* SL1344 and *S. Paratyphi* A 9150) strains were grown for overnight in selective LB medium. One ml from each culture was concentrated, washed, and resuspended in 100 µL of fresh LB. Donor and the recipient strains were mixed at equal volumes in a test tube, and 20 µL of the conjugation mix or 10 µL of the donor suspension only (as a negative control) were spotted onto an LB agar plate or an LB agar plate and incubated at 37°C. After 6 h, the conjugation mix was scraped from the plate and resuspended in 1 ml saline, and serial dilutions were plated onto LB supplemented with two antibiotics for CFU counting. The plates were incubated at 37°C overnight, and the conjugation frequency was calculated as the number of transconjugant CFUs/donor CFUs under each condition.

Rarefaction curve and heaps' law fit

Rarefaction curves for the pangenome of each serovars were computed using the package vegan for R, version 2.5.6 (<https://CRAN.R-project.org/package=vegan>). The Heaps' law was fitted to each rarefaction curves.

$$F = \kappa N^\beta$$

Where F is the total number of distinct pangenome families and K and β are parameters determined empirically. The associated K and β values were extracted using the “optimize.curve_fit” method from scipy version 11.4 for python [37].

Results

Sequencing, assembly and annotation of new typhoidal strains

To elucidate the genome structure of *S. Paratyphi* A and *S. Typhi*, we have resequenced and hybrid assembled the genomes of *S. Paratyphi* A 45157 (SPA45157) and *S. Typhi* 120130191 (STY120130191) both are clinical low-passage isolates of enteric fever patients. SPA 45157 was isolated in 2009 at the Sheba Medical Center from the

blood of a hospitalized patient, as part of a large paratyphoid outbreak affected a large group of Israeli travelers who have visited Nepal during the fall of 2009 [20]. This strain was found to be similar on the genomic level to multiple SPA strains from lineage A that were predominantly isolated in south east Asia (see below).

STY 120130191 was isolated at 2012 at the Sheba Medical Center from the blood of a patient diagnosed with a typhoid fever, after returning from a backpacking trip to India and Thailand. This strain showed very high sequence similarity to a 2008 STY isolate (ERL082356) that was isolated in India. The genome of SPA45157 was hybrid assembled using both Myseq and Oxford Nanopore sequencing reads and yielded a complete chromosome of 4,554,125 bp (accession number CP156168). The obtained chromosome length was very similar to the genome size of other *S. Paratyphi* A genomes, which were previously reported [38]. The STY120130191 genome (accession number CP156169) was generated using a combined PacBio and Illumina HiSeq 3000 sequencing reads and yielded a complete genome of 4,783,423 bp and an IncFIB replicon of 106,706 bp (accession number CP156170).

Functional and structural gene annotation using Prokka [21] and eggNOG-mapper [22] of SPA45157 and STY120130191 genomes produced 4,307 and 4,745 coding genes, 84 and 82 tRNA, and 22 and 25 ribosomal RNA genes, respectively. Table S2 summarizes the assembly and annotation statistics of these two genomes.

The phylogeny of typhoidal serovars

To shed light over the phylogeny of these strains, their assembled genomes were analyzed together with additional high-quality (Reference Sequences) dereplicated and representative 134 (out of 155 analyzed) SPA, 43 (out of 242 analyzed) STY and 161 (out of 443 analyzed) STM genomes, retrieved from the NCBI RefSeq database (Table S1). As an out-group, we included the genome of *S. bongori* (strain N268-08; RefSeq NC_021870.1). To describe the phylogenetic relations between these strains, a maximum likelihood phylogenetic tree based on 3,368 aligned core genes (alignment length: 73863 nucleotide sites, distinct site patterns: 1,922) shared by all serovars was built. The resulting phylogenetic tree (Figure 1) indicated that the SPA, STY and STM serovars are all grouped into well-resolved separated clades. This result is consistent with previous phylogenomic studies [13,40], indicating that *Salmonella* core-genes generally exhibit

a phylogenetic grouping matching the serologic classification according to the Kauffmann-White-Le Minor serotyping scheme.

To gain further insights into the evolutionary genomics of these serovars, we compared nucleotide diversity using the P_i (π) indice in aligned core genes between all isolates of each serovar. Interestingly, we found that nucleotide diversity varies largely between serovars. The highest nucleotide diversity was indicated in the STM serovar ($P_i = 0.00154$), which demonstrated significantly higher diversity than SPA ($P_i = 8.24 \times 10^{-4}$) and almost 20-fold higher diversity than STY ($P_i = 7.71 \times 10^{-5}$). These results indicated that while STM exhibits the highest genetic diversity, and SPA presents somewhat moderate level of nucleotide diversity, STY is basically a clonal serovar that exhibits very low genetic diversity among isolates.

The population structure of SPA

While the population structure of STY was recently addressed by several studies [41–44], SPA is still an understudied typhoidal serovar. Therefore, we sought to analyze in more details the population structure of SPA and applied the Fastbaps method based on Bayesian Hierarchical clustering [45] to define its Phylogeny. Using this approach, we were able to identify ten genetically distinct lineages, named A to J (Figure 2). Five out of the ten lineages (A, B, F, H, and J) were monophyletic, and three of them are well correlated with the geographic origin of their isolates. Specifically, lineage A contains mainly isolates from Asia, and lineages H and J contain strains that were isolated in South America. Lineages B and F consist of a mixture of isolates from different countries and continents, suggesting global distribution of these strains.



Figure 1. Phylogenetic tree structure of *S. Paratyphi A* ($n = 134$), *Typhi* ($n = 43$) and *Typhimurium* ($n = 161$) representative isolates. *S. bongori* (isolate N268–08) was used as an outgroup. The tree was obtained from aligned core genes ($n = 3368$) of dereplicated Refseq public genomes and maximum likelihood inference calculated with iqtrees [39]. Branches are colored according to serovar (SPA in red; STY in blue; and STM in orange). Internal nodes with support of $>90\%$ are shown in black, while nodes with support of $>50\%$ are shown in gray. Tip colors indicate the geographical origin of isolates, according to NCBI *geo_loc_name* metadata. Year of isolation was added to strain names when it was available from the NCBI associated metadata.

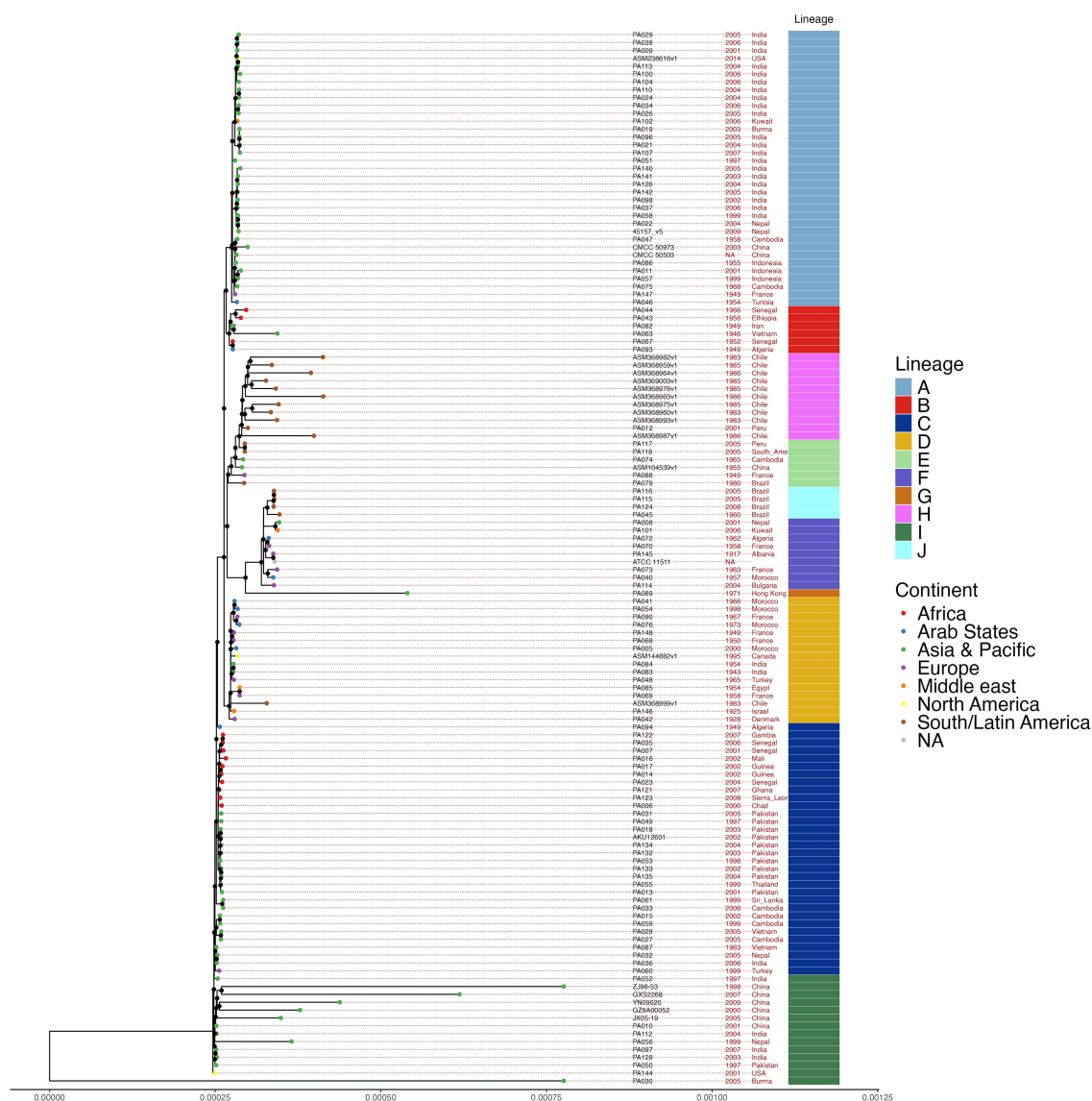


Figure 2. *S. Paratyphi A* phylogeny and classification into lineages. A phylogenetic tree of SPA was obtained from aligned core genes ($n = 3877$) of 134 representative SPA public genomes together with the resequenced SPA45157 genome. A maximum likelihood inference was calculated with *iqtree*. Lineages [A to J] were defined using bayesian hierarchical clustering implemented in the *fastbaps* software [45]. Country and isolation dates of isolates were retrieved from NCBI sample metadata and are indicated in red. Tip colors indicate the geographical origin of the isolates.

Noteworthy, lineages A to G are largely consistent with a previously published analysis using 149 of SPA genomes [46], with one of these lineages (lineage G) containing a single isolate (PA089) only, which was isolated in 1971 in Hong-Kong. Nonetheless, our current analysis expanded this previous study and identified three additional new lineages including, lineage H (contains 11 isolates), lineage I (contains 14 isolates), and lineage J (contains 4 isolates) that were not identified in previous analyses. Supplementary Figure S1 presents a comparison of the SPA lineages between our current analysis and the previous study reported by Zhou et al. [46].

The pangenome of STM, SPA and STY

To understand better the diversity, dynamics and evolution of typhoidal *Salmonellae*, we have constructed the pangenome of STM, SPA and STY, as represented by the 339 annotated complete genomes. The resulted pangenome of the three serovars consists of 15,096 genes, of which 3,368 are core genes (present in more than 99% of the genomes), 312 were defined as soft core genes (present in 95–99% of genomes), 1,215 shell genes (present in 15–95% of the genomes) and 10,201 cloud genes (present in less than 15% of genomes). The cloud genes category represents 67.5% of the pangenome and includes genes that are either specific to only

one serovar or to a subgroup of genomes from one serovar. This large array of cloud genes indicated the variable distribution of a significant pool of accessory genes among the STM, STY and SPA genomes that shapes its intra-serovar genetic diversity. Table S3 summarizes the core, soft core, and the accessory (shell and cloud) genes for STM, SPA, and STY and lists common core and accessory genes shared by STY and SPA.

In order to evaluate differences in gene repertoire between typhoidal and STM genomes, we compared their gene pool size. To do so, we built a pangenome and a core-genome specific to each serovar (Figure 3). Interestingly, while the three analyzed serovars had a relatively similar core-genome size of 3342, 3878 and 4089 genes for STM, SPA and STY, respectively, the total accessory gene pool (cloud+shell genes) of the STM dataset (8290 genes) was more than four times larger than the ones of the typhoidal datasets (1991 and 1912 genes for SPA and STY, respectively). To assess whether this difference could translate into a difference in the openness of the pangenomes,

we computed the pangenome rarefaction curves of each serovar. Pangenome size can either increase a lot with the addition of novel genomes and then it is referred as an open pangenome, or increase very little and then it is considered as closed pangenome. These changes can be quantified by fitting rarefaction curves with the Heaps' law and extracting the values of the γ parameter that vary between 0 (closed) and 1 (open) [47]. While STY ($\gamma_{\text{STY}} = 0.0976689$) and SPA ($\gamma_{\text{SPA}} = 0.0981325$) displayed highly closed pangenome, the γ parameter of STM ($\gamma_{\text{STM}} = 0.235909$) suggested significantly more opened pangenome (Figure 3). These results suggest that since the time of divergence, SPA and STY have had a smaller rate of acquisition of new genes into their ancestral genome in comparison to STM. We concluded from this analysis that STY and SPA exhibit a conservative genomic structure with a relatively low diversity and a scarcity of new gene acquisition. Yet, despite this low acquisition rate, both serovars also harbor genes that are absent from STM and are specific to STY, SPA or both.

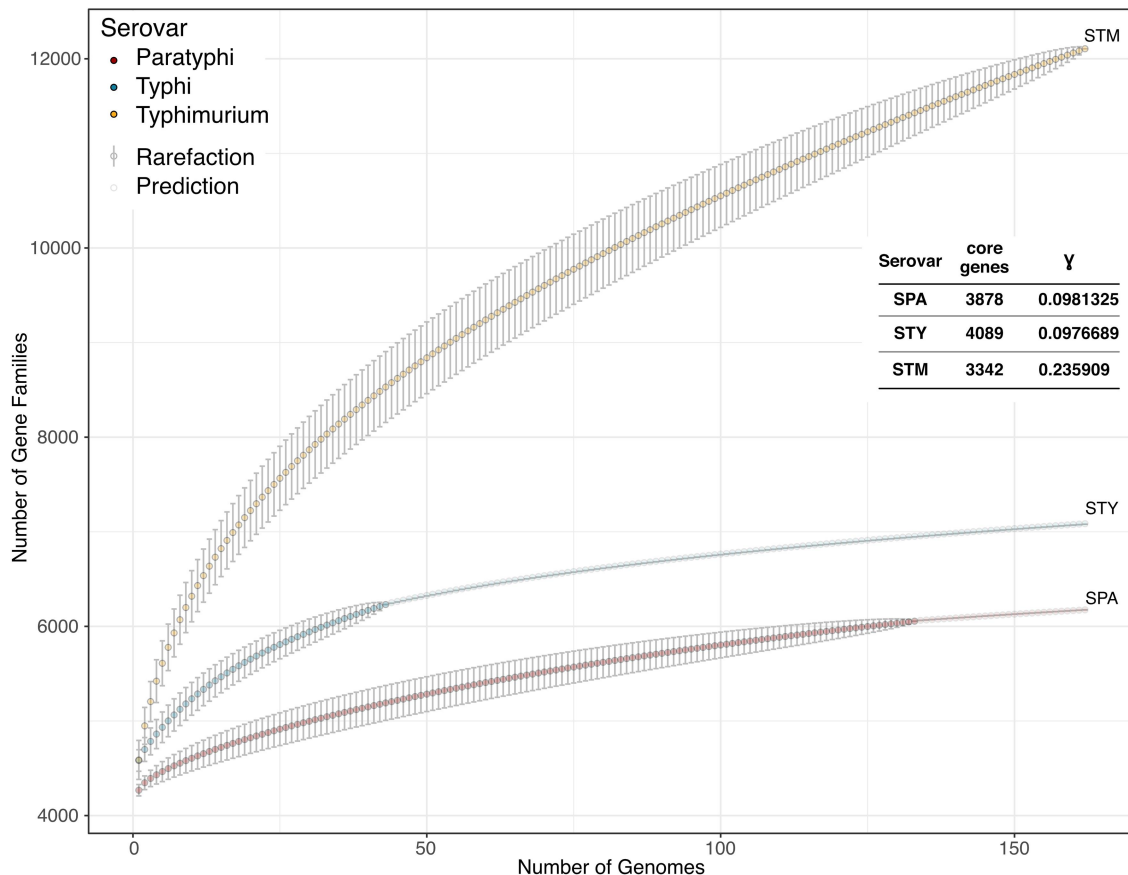


Figure 3. Rarefaction curves of the pangenome of *S. Typhimurium*, *S. Paratyphi A* and *S. Typhi*. Gamma values deduced from the fitted heaps' law curve and the number of core genes for each serovar are shown in the insert table. The rarefaction curves are shown in yellow, red and blue for the genome of STM, SPA, and STY respectively. Values calculated with the accumulation method are shown in gray error bars. For STY and SPA, values predicted using an Arrhenius model to reach STM genome number (164) are shown in light gray circles without error bars.

Different and common pseudogenes repertoire in SPA and STY

Typhoidal serovars are human-specific pathogens and thought to lack an environmental reservoir compared to non-typhoidal serovars. Such lifestyle can lead to reductive genome evolution, as many genes become useless in a narrow host range [19]. Indeed, previous studies have shown a significant degree of genome decay in STY and SPA genomes [16,38,48,49], however this observation was based on a relatively low number of genomes. To gain a broader perspective about genome degradation in SPA, we compared the content of pseudogenes in SPA ($n = 134$), STY ($n = 43$) and STM ($n = 164$) strains. Pseudogenes in each strain were determined using an in-house pipeline (see Materials and Methods section) that was adjusted to detect pseudogenes in *Salmonella*. Pangenome families of pseudogenes were defined as clusters of proteins that showed 80% or more homology and presented signs of pseudogenization, including nonsense mutation, in-frame deletion of more than five amino acids, or truncation of more than 30% of the protein. Overall, we found 2216 pangenome families that were pseudogenized in at least one genome in STM, 743 pangenome families in STY and 1315 pangenome families in SPA. Based on the presence/absence of each pangenome family as pseudogene, we then built a core-pseudogenome and a pan-pseudogenome for each serovar (Figure 4). As it was previously defined for traditional pangenome analyses [50], we considered as core pseudogenome, pseudogenes that were identified in more than 99% of the strains; as soft-core pseudogenome, pseudogenes found in 95–99% of the strains; shell pseudogenome, pseudogenes found in 15–95% of strains; and as cloud pseudogenome, pseudogenes that occurred in less than 15% of the strains. Since STM was used as the reference, it contained no core pseudogene families, i.e. a gene that is missing or inactivated in more than 99% of the STM genomes. However, it revealed 2108 cloud-pseudogene families that were potentially inactivated in one or few genomes. In contrast to STM, STY and SPA were found to carry plenty of core-pseudogene families, with 174/743 (23.4%) of the STY pseudogenes and 137/1315 (10.4%) of the SPA pseudogenes were identified as core pseudogenes (Figure 4; Table S4). Moreover, we built the core pseudogenome of the typhoidal strains, and found 34 core pseudogenes and 13 soft-core pseudogenes common to both STY and SPA (Table S4). The fact that these pangenome families are consistently pseudogenized in both STY and SPA genomes suggest that their function is dispensable or possibly deleterious for their lifestyle as typhoidal pathogens.

Multiple pseudogenes of SPA are known to be involved in adhesion and fimbriae biogenesis (*stfF*, *csgF*, *safD*, *bcfF*, *sinH*, *zirT*), flagella biogenesis and motility (*ycgR*, *hin*, *fliB*) and iron and magnesium homeostasis (*fhuA*, *fhuB*, *fhuE*, *mgtA*).

Interestingly, similar functions were also found to be dominant among the STY pseudogenes including adhesion and fimbriae biogenesis (*csgD*, *htrE*, *sthE*, *steA*, *sinH*), flagella biogenesis and motility (*tsr*) and iron and magnesium homeostasis (*fhuA*, *fhuB*, *fhuE*, *hoxN*) (Table S4).

Gene ontology (GO) pathway enrichment analysis using the PANTHER tool [31] allowed classification of 768 out of 1315 SPA pseudogenes into 45 GO pathways. The most significantly enriched pathways after FDR correction (had the lowest FDR corrected p-value) were small molecule metabolic process (GO:0044281), metabolic process (GO:0008152), and cellular process (GO:0009987) (Table S5). Similar analysis has classified 376 out of 743 STY pseudogenes into 11 GO pathways, while the most significantly enriched pathways after FDR correction were cellular process (GO:0009987), amino acid metabolic process (GO:0006520) and transmembrane transport (GO:0055085) (Table S6).

Finally, we also looked at small in-frame deletions (less than five amino acids) detected by our analysis. Core small in-frame deletions were found in 47 gene families in SPA, and in 37 gene families in STY. Among small in-frame deletions detected in SPA, we found changes in additional motility and chemotaxis genes (*cheM*, *fliK*, and *tar*), invasion and virulence (*sipD*, *rhuM*, *sinI*) and fimbriae biogenesis (*stdA*, *safB*). Related functions were also found in genes with small in-frame deletions in STY including genetic variation in the motility gene (*fliK*); invasion and virulence (*sipD*, *srfC*, *sinI*, and *rhuM*) and fimbriae biogenesis (*fimA*, *safB*, and *stdA*). Since all of these variations are small in-frame deletions, we currently do not know if these gene products are actually functional or not, and therefore, we included them in a separate list as potential pseudogenes (Table S7). Yet, the fact that these genes accumulate more genetic variations in typhoidal strains suggests that they may be under a different evolutionary pressure in SPA and STY in comparison to STM.

Distribution of prophages in STM, STY, and SPA genomes

Mobile genetic elements drive the horizontal transfer of adaptive traits across prokaryotes. They allow the bacteria to quickly acquire new genes and functions, which may eventually facilitate their divergence into different

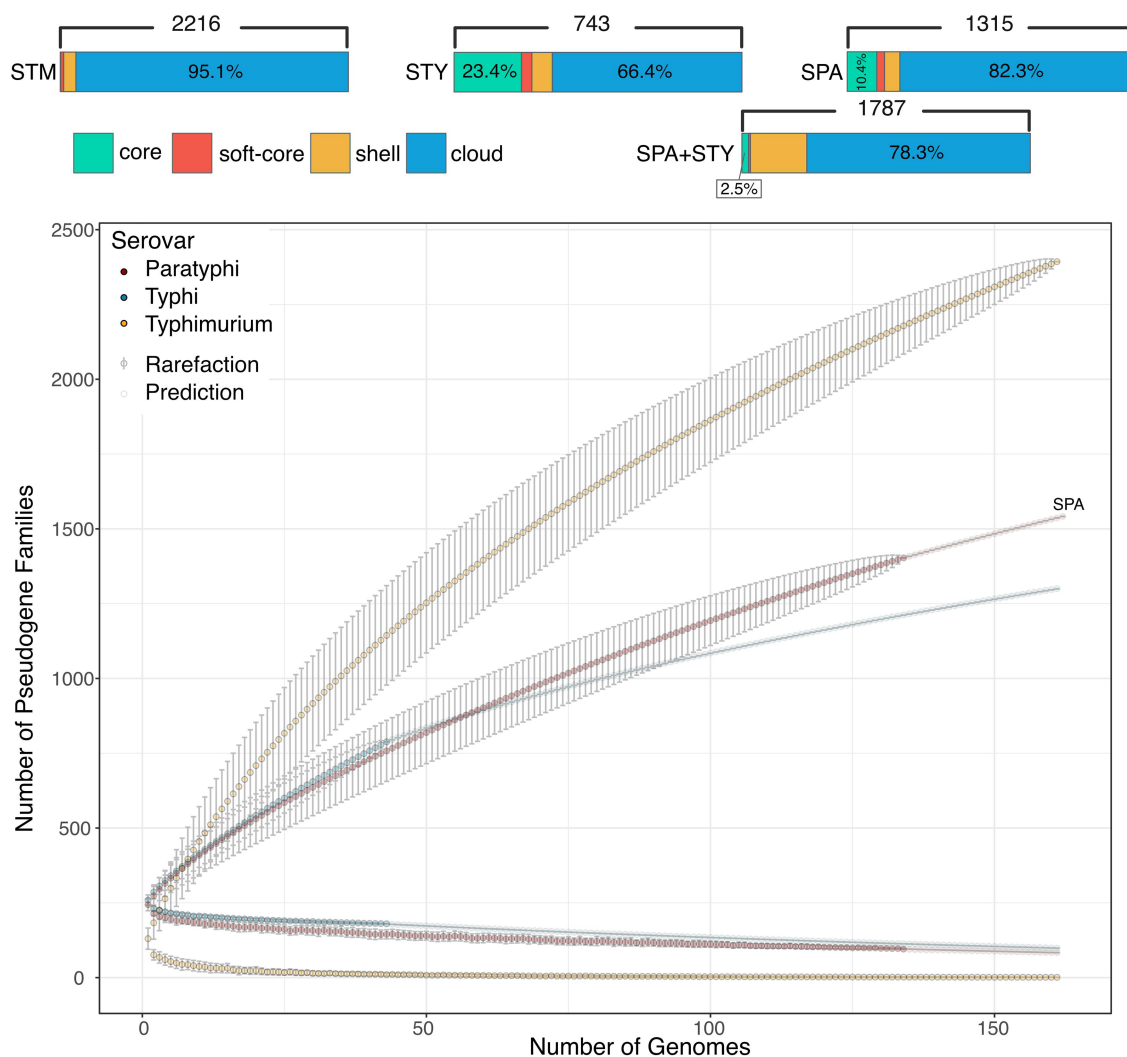


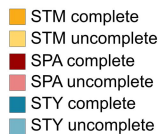
Figure 4. Number of core, cloud, shell and soft-core pseudogenes and rarefaction curves for the core and pan-pseudogenome of *S. Typhimurium*, *S. Paratyphi A* and *S. Typhi*. The top panels show the number of core, cloud, shell and soft-core pseudogenes for the STM, SPA, STY and their numbers in a combined SPA+STY dataset. The rarefaction curves are shown in yellow, red and blue for STM, SPA, and STY, respectively. For STY and SPA, values predicted to reach STM genome number (164) using either an Arrhenius model for pangenome curves or an exponential decay function for the core genome curves are shown by discontinuous points without error bars.

species [51,52]. Therefore, we were inspired to test if the identified difference in pangenome openness between the serovars could be explained by a different content of mobile genetic elements. For that end, we specifically analyzed the presence and distribution of prophages, conjugative and mobilizable elements (extrachromosomal or integrated) in the three serovars.

To be conservative, we accounted only for complete, highly confident prophage predictions yielded by *virsorter2* [32]. This way, we detected 263 complete prophages in the 164 STM genomes, 87 complete prophages in the 43 STY genomes, and 131 complete prophages in the 134 SPA genomes analyzed (Figure S2, Table S8). The maximum number of complete prophage sequences found in one STM strain was

five, which was identified in the clinical (human) stool isolate AUSMDU00027944. Overall, STM and STY presented similar numbers of complete prophage, with an average of 1.6 and 1.8 phages per genome, respectively (not significantly different by Mann – Whitney U-test), and both contain more complete prophages per genome than SPA, which harbored on average only 0.97 complete prophage per genome (Figure 5(a)) (Mann–Whitney U, p-values = 7.600e-06 and 5.845e-06, compared to STM and STY, respectively).

We also examined the diversity of prophages among STM, SPA, and STY strains. The overall prophages repertoire identified in this dataset can be classified into 70 viral clusters (VCs), showing high diversity of



and may suggest that the lack of an environmental reservoir limits horizontal acquisition of accessory genes into their genome. This notion was also reinforced by testing specifically the distribution of known phage-encoded virulence factors in complete and incomplete prophage elements (phage remnants) as predicted by Virsorter and CheckV. These elements can still carry virulence-associated genes that could play a role during host infection. Here, we tested the distribution of the virulence factor genes *gipA* (Gifsy-1 encoded invasion protein), *gogB* (Gifsy-1 encoded anti-inflammatory effector), *sodC* (Gifsy-2 encoded periplasmic [Cu,Zn]-superoxide dismutase), *sopE* (SopEphi

encoded guanine exchange factor), *sopE2* (phage-like element encoded guanine exchange factor), *sseI* (Gifsy-3 encoded E3 ubiquitin ligase), *sseK3* (ST64B encoded invasion effector), *sspH1* (Gifsy-2 encoded actin polymerization inhibitor), and *sspH2* (phage-like element encoded actin polymerization inhibitor). This analysis indicated that while a significant portion of STM genomes carries phage-encoded virulence factors at varying frequencies, their distribution among SPA and STY genomes is very limited. SPA genomes were found to carry mainly *SopE* (few genomes also carry *SopE2*) and STY genomes harbor mainly *SopE* and *SodC* (Figure 5(c)).

Next, we tested possible correlation between the presence of specific VCs and the geographic location of the isolates using Chi-Squared Test and Cramer's V score. This analysis identified statistically significant association between certain VCs and the origin country of some isolates. For example, VC_0_15 was found enriched in USA and China STM strains, VC_1_6 among Chile SPA strains, VC_39_2 in China and Australia STM strains and VC_7_4 in STM China, Australia, Taiwan and USA strains (Figure S3). These results suggest that at least some prophages are more prevalent in specific countries than others are and that prophages distribution can be geography dependent.

Finally, it is noteworthy that, while having a low diversity of phages, typhoidal serovars share five common VCs that were not found in STM. This could be due to vertical inheritance of selected prophages that were present in their ancient genome before STY and SPA diverged from their common ancestor or that STY and SPA can be infected by few, closely related prophages due to their similar lifestyle.

Distribution of plasmids and mobile genetic elements in STM, STY, and SPA

We next classified the identified plasmids into phylogenetically coherent taxonomic groups. Overall, 299 plasmids were assigned into 42 plasmid taxonomic units (PTUs) (Figure 6(a), Table S10), showing a high diversity of plasmids among the three serovars of our dataset. However, this diversity was not distributed equally among all serovars. Out of the identified PTUs repertoire, 76% (32/42) different PTUs were found in STM, 40% (17/42) in SPA and only 17% (7/42) in STY (Figure 6). Nineteen PTUs were specific to STM only, six PTUs were specific to SPA (PTU-E68, PTU-BAC44, PTU-E9, PTU-E39, PTU-N2/3, PTU-Bac19), and only three PTUs were specific to STY (PTU-N1, PTU-E50, PTU-E80). Of note, one PTU (PTU-Y), was found in both STY and SPA, but not in

STM, suggesting that this PTU may be adapted to typhoidal serovars. This high diversity of plasmids found in STM is in agreement with the openness of its pangenome and with the high diversity of accessory genes found in STM, compared to STY and SPA. Furthermore, the low diversity of plasmids in STY might explain why, despite having a similar number of MGE per genome as STM, STY has a more closed pangenome.

Interestingly, two PTUs, PTU-X1 (IncX1) and PTU-HI1A (IncHI1A) were found in all three serovars in low, but similar frequencies (Table S10), suggesting that these plasmids may be beneficial or stable in both typhoidal and NTS serovars. Furthermore, the three most prevalent PTUs in *Salmonella* were all specific to STM, including PTU-FS (IncFII/IncFIB), PTU-I1 (IncI1), and PTU-HI2 (IncHI2). These PTUs contain plasmids such as pST90-2 [53], pESI [54], and pJXP9 [55] respectively, and are known to be associated with multidrug-resistance strains.

Similarly, we looked at the distribution of conjugative and mobilizable systems (collectively referred to as conjugative elements) in the chromosomes and plasmids of the three serovars. Altogether, this analysis revealed 227 conjugative elements in the 164 STM genomes, 51 in the 43 STY genomes, but only 15 in the 134 SPA genomes analyzed (Figure S4). Interestingly, most of the STM (78%) and almost all STY (93%) genomes contained at least one conjugative element. In many cases, STM genomes harbored multiple conjugative elements, with 40.8% of them harbored two or more conjugative elements (Figure S4 and Table S11). In contrast, conjugative elements were rather scarce in SPA genomes, with only 9.7% (14/134 genomes) were found to encode one or more conjugative elements. Consequently, like for prophages, STY and STM had a similar number of conjugative elements per genome, and both contained more conjugative elements per genome than SPA as shown in Figure 6(b) (Mann-Whitney U, $p \leq 1.108e-24$ and $p \leq 1.406e-30$, compared to STM and STY, respectively). Taken together, these results indicate that SPA genomes have either acquired much less mobile genetic elements than STY and STM, or that these mobile genetic elements are not maintained for a long time in SPA genomes.

One option that could have potentially explained the low frequency of conjugative elements in SPA genomes is an impaired ability of this serovar to perform conjugation, possibly due to a very long O-antigen chains that characterize this serovar [56], which may interfere with DNA transfer via conjugation. To examine this hypothesis experimentally, the conjugation frequency of five different conjugative plasmids from various Inc

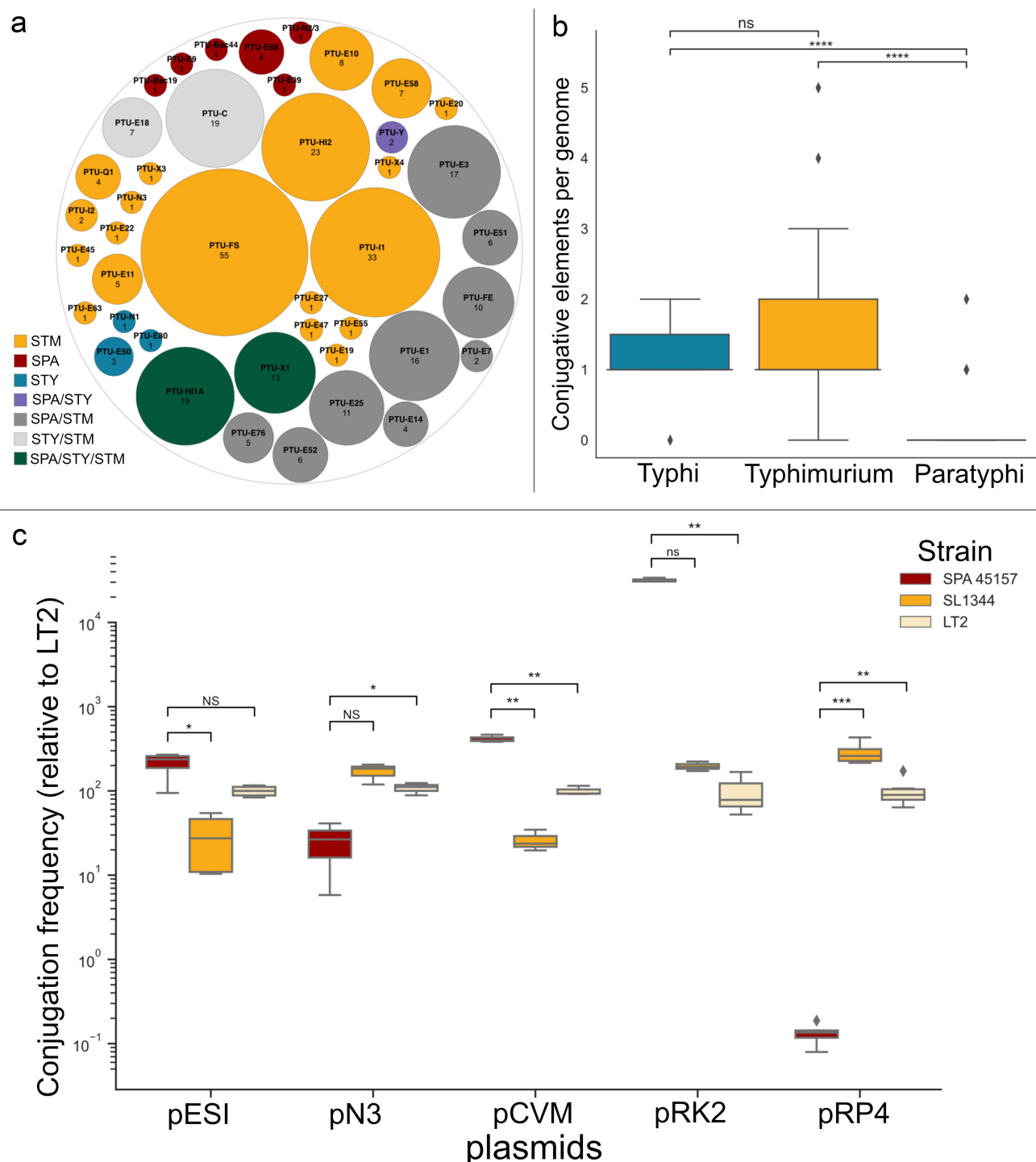


Figure 6. Distribution of conjugative elements and diversity of the plasmids in *S. Typhimurium*, *S. Paratyphi A* and *S. Typhi*. (a) The diversity of plasmid taxonomic units (PTUs) per serovar is shown and color-coded based on its distribution between serovars. The circles represent the different PTUs and their size is proportional to the number of plasmids in the family. The name of the family and the number of the elements in the family is shown in the center of each circle. (b) The number of conjugative elements corresponds to complete conjugative systems present either in the chromosome or in plasmids of each genome per serovar. The horizontal bar represents the median value, while the lower and upper hinges correspond to the first and third quartiles. The whiskers extend from the hinge to 1.5 times the range between the first and third quartile. Data beyond these values are shown as dots. The statistical significance was calculated by the Mann–Whitney U test. ns, $p > 0.05$; ****, $p \leq 0.0001$. (c) Conjugation frequency of different plasmids into STM and SPA. The conjugation frequencies of different plasmids were tested by mating assay into SPA (strain 9150) and to two STM strains (SL1344 and LT2). Conjugative plasmids tested belong to different PTUs: pN3, PTU-N1; pESI, PTU-I1; pRK2, PTU-E; pCVM, PTU-I1; pRP4, PTU-P1. The horizontal bar represents the median value, while the lower and upper hinges correspond to the first and third quartiles. The whiskers extend from the hinge to 1.5 times the range between the first and third quartile. Data beyond these values are shown as dots. The statistical significance was calculated by the Mann–Whitney U test. ns, $p > 0.05$; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$.

groups including pESI [54], pN3 [57], pCVM [58], pRK2 [59] and pRP4 [60] were compared between two strains of STM (LT2 and SL1344) and SPA 9150. As shown in Figure 6(c), with the exception of the plasmid pRP4, that was indeed conjugated into SPA in a much lower frequency than to the examined STM strains, other conjugative plasmids were conjugated under laboratory conditions at similar or at slightly higher frequencies into SPA than to STM. These results suggested that the differences in the distribution of conjugation elements are not the results of an impaired conjugation ability of SPA, but due to a different reason. Alternative possibilities may include the presence of more stringent defense systems against foreign DNA in SPA or that SPA is less exposed to environmental pressures such as antibiotics, in a way that lessen the selective advantage conferred by conjugative plasmids carrying antibiotic resistance genes. These possibilities are further deliberated in the Discussion section.

Differences in the distribution of antimicrobial resistance genes between serovars

To assess if the differences in the distribution of accessory genomes and its openness between serovars affect their antibiotic resistance potential, we compared the distribution of antimicrobial resistance genes in STM, SPA, and STY (Figure 8, Table S12). Out of the 1041 AMR genes we detected in our analysis in all serovars, 60% (625 genes) were carried by plasmids while 40% (416 genes) were chromosomally encoded. Interestingly, antimicrobial resistance due to horizontal acquisition of genes was more prevalent in STM and STY than in SPA. For instance, 72% of the STM genomes and 30% of the STY genomes analyzed were resistant to sulfonamide (due to acquisition of *sul1*, *sul2* and/or *sul3* genes), but only 3.8% of the SPA genomes harbored these genes. Similarly, 63% of the STM and 23% of the STY, but only 3.7% of the SPA genomes were resistant to streptomycin (due to acquisition of *sat2*). Overall, SPA genomes were found to harbor the lowest prevalence of AMR genes and the distribution of all the AMR genes examined (Table S12) never exceeded 4% of the SPA genomes (Figure 7).

Next, we asked if these AMR genes were carried by the plasmids associated with each serovar. Many were actually located on PTU-HI2 (250), PTU-FS (37), and PTU-II (33), which are STM-specific PTUs, and on PTU-C (111) that was found in both STM and STY. Similarly, when we analyzed the distribution of *qnr* and *oqx* genes, conferring resistance to quinolones, we found that 19% of the STM genomes were predicted to be quinolone-resistant due to acquisition of *qnr* or

oqx genes, but only 9% of STY genomes and 0.7% of the SPA genomes were harbored these genes (Figure 7). Nevertheless, when we analyzed the distribution of point mutations in the core genes *gyrA* (encoding DNA gyrase) and *parC* (encoding topoisomerase IV), we found that while 44% of STY and 37% of SPA genomes acquired point mutations potentially leading to quinolone resistance, only 22% of the STM genomes possessed these mutations. These results suggest that while quinolone resistance is abundant in the three serovars, it is more associated with an AMR gene acquisition in STM, but with mainly point mutations in STY and SPA. Taken together, these results suggest that STM acquires antibiotic resistance genes through acquisition of AMR genes through horizontal gene transfer and that this environmental reservoir might be less accessible to SPA and STY. In consequence, SPA and STY accumulate more point mutations to face antibiotic pressure. Moreover, comparison between SPA and STY indicated a significantly higher frequency of AMR genes distributed in STY than in SPA genomes.

Diverse distribution of defense systems in STM, STY and SPA

The difference in the genetic content of plasmid and prophage sequences between typhoidal and non-typhoidal serovars could potentially be the result of differences in the presence and distributions of defense systems. To test this hypothesis, we investigated the diversity of defense systems in the *Salmonella* isolates cohort using Defense Finder (Figure 8, Table S13). This analysis detected not less than 3017 defense systems in the 342 genomes represented in our dataset, including 1699 defense systems in STM, 1078 systems in SPA and 330 systems in STY. Among all defense systems identified, the vast majority of which (39%; 1176) were restriction modification (R-M) systems. STM was found to encode the highest number of defense systems per genome, followed by STY and SPA (Figure 8, Mann-Whitney U, P-values ≤ 0.0005). In terms of their diversity, while all three serovars encoded RM and CRISPR-Cas systems, STM harbors a much higher diversity of different defense systems, including 29 that are completely absent from STY or SPA (i.e. STM-specific). As many defense systems are carried by MGEs [61], this higher diversity is in-line with the higher diversity of MGE (especially unique prophages, Figure 5b) found in the STM genomes. Nevertheless, we found that some defense systems including PD-T4-3, Druantia and RST_3HP are almost exclusively specific to SPA and STY and are absent from STM. On the other hand,

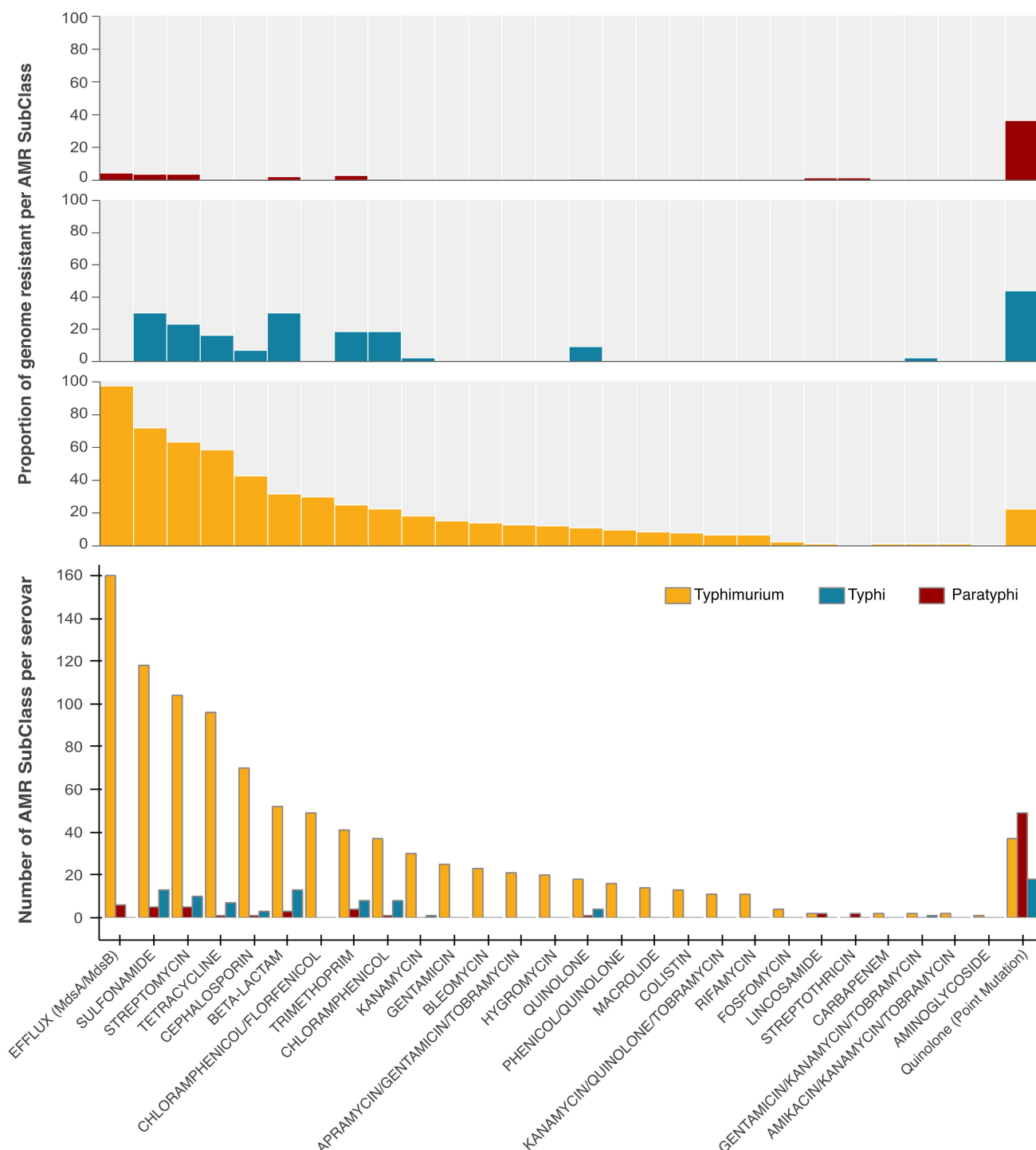


Figure 7. Prevalence and distribution of antibiotic resistance genes in STM, STY, and SPA. The top three charts represent the proportion of genomes from each serovar that encode AMR genes. The bottom graph shows the total number of antibiotic resistance genes found in each serovar for every antibiotic subclass. Point mutations conferring resistance to quinolone are shown on the right end of each graph.

this analysis showed that the defense systems Paris, BREX, Mokosh, Retron and PD-T4-1 are specific to STM and are missing from SPA and STY genomes.

Most of these defense systems are known anti-phage factors [62] and many were found to be encoded on

MGEs, specially prophages. Therefore, it is possible that this difference in defense system content is directly linked with the presence of serovar-specific prophages. To test this hypothesis, we studied the genetic location of the identified defense systems. Interestingly, we

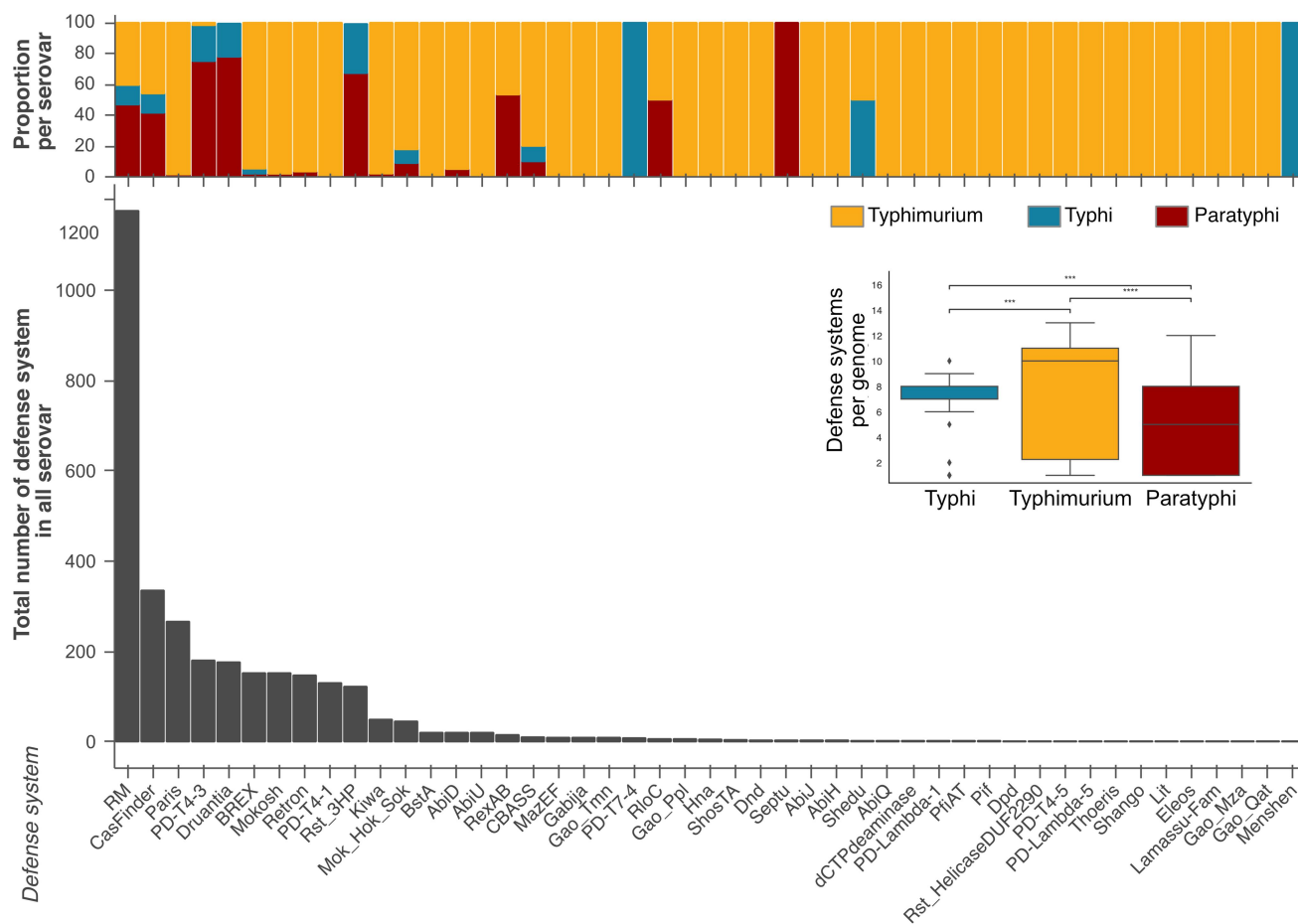


Figure 8. Prevalence and diversity of the defense systems in serovars Typhimurium, Paratyphi A, and Typhi. The histogram shows the distribution of the different defense systems found among STM, SPA and STY. The top part represents the distribution of each defense system among the three serovars. The box plot insert shows the number of defense systems per genome for each serovar. The horizontal bar represents the median values, while the lower and upper hinges correspond to the first and third quartiles. The whiskers extend from the hinge to 1.5 times the range between the first and third quartile. Data beyond these values are shown as dots. The statistically significant differences between plots are shown at the top (***, $p \leq 1.00e-03$; ****, $p \leq 1.00e-04$, by Mann-Whitney U test).

found that the defense systems specific to STM, Paris and PD-T4-1, Kiwa and AbiU were all encoded within prophages specific to STM (VC35-0, 37-0, 2-11 and 7-4), while the defense systems Rst_3HP and Septu, which are present only in typhoidal serovars were encoded by prophages specific to STY and SPA (VC1-0 and VC5-0). Therefore, these results suggest that STM, SPA and STY share RM and CRISPR-Cas systems, but also harbor specific defense systems arsenal, which may shape the difference in prophage content identified between serovars.

Discussion

Salmonella enterica serovar Paratyphi A is an understudied causative agent of enteric fever, with an increasing prevalence in some endemic regions in parts of Asia [63–66]. Revealing its global phylogeny, genome

structure, AMR gene composition and pseudogene repertoire are important for understanding the evolution of this pathogen and identifying effective measures to reduce enteric fever burden.

Previous population structure studies that were focused on SPA have reported that this serovar has a considerable regional differences with the emergence of seven distinct lineages, while each one of these lineages have originated in a specific geographical location [46,67]. Our global phylogeny analysis has broadened these analyses and updated these reports, by clustering the SPA genomes into ten defined lineages, while capturing higher population diversity of serovar Paratyphi A than previously reported [19,38,48,67].

It is broadly accepted that a genomic hallmark of the typhoidal *Salmonella* serovars STY and SPA is the accumulation of high number pseudogenes in their genome in comparison to the broad-host serovar STM

[15,19]. This difference was thought to result from the adaptation to the human host and a genetic drift due to a population bottleneck affect that have occurred during their evolution [15,19,49,68]. Analyzing multiple genomes of SPA allowed us to evaluate more accurately the composition of core and flexible pseudogenes in this serovar. Overall, we found that among 3878 SPA core genes, 137 (3.53%) are *bone fide* core pseudogenes and 47 (1.21%) are potentially pseudogenes due to small in-frame deletions of less than five amino acids. Similar analysis performed on STY indicated that among 4089 core STY genes, 174 (4.25%) are *bone fide* core pseudogenes and 37 (0.90%) are potentially pseudogenes due to small in-frame deletions of less than five amino acids. Although further experimental characterization is still required for functional confirmation of these small in-frame deletions, one of them, *fief* (that is also called *yiiP*) was recently confirmed experimentally to be inactive in STY [69], providing further experimental support to the possibility that at least some of these changes lead to gene inactivation.

Many of the inactivated genes are mapped to fimbrial, motility and chemotaxis genes, lysogenic phages and genes encoded within *Salmonella* pathogenicity islands (SPIs). This estimation is similar to previous analyses, which suggested that 4.8% of annotated CDSs in SPA [19] and 4.5% of annotated CDS in STY [16,70] are pseudogenes. While previous pseudogene predictions in different STY genomes indicated that the number of pseudogenes in STY genomes vary remarkably from 150 to 233 [44], our current analysis is based on higher number of genomes and therefore allowed better assignment of inactivated genes into core, cloud, shell and soft-core pseudogene groups. Together, these reports suggest that gene inactivation may be an important driving force of genome variation among typhoidal *Salmonella*.

Pseudogenes formation was also shown to shape the evolutionary trajectory of the STM sequence type 313 (ST313), a leading causative agent of invasive nontyphoidal *Salmonella* infections in Africa that was shown to harbor multiple gene inactivation events in important virulence genes [71]. Noteworthy, some of its characterized pseudogenes are also inactivated in STY (e.g. *macB*, encoding a putative ABC transport protein) and SPA (e.g. *pipD*, encoding a T3SS secreted effector and *csgD*, encoding a biofilm regulator). These common examples demonstrate possible convergence evolution in different *Salmonella* lineages towards invasive or systemic lifestyle.

Interestingly, although a significant level of pseudogene formation was detected in SPA, we could not

identify any evidence for genome size reduction indicating that pseudogenization does not lead to genome downsizing as was previously suggested for other human pathogens including *Mycobacterium leprae* [72] and *Rickettsia prowazekii* [73].

Constructing a core-gene multiple alignment of STM, SPA, and STY serovars indicated that the genetic diversity is lower in typhoidal serovars compared to STM. These results are concurring with the openness of the pangenome of the three serovars, suggesting that pan-genome of STY and SPA are more closed than the one of STM. Collectively, these results are in agreement with previous reports, indicating that SPA does not tend to acquire new genes and that this phenomenon may contribute to its conservative genomic structure and clonal population structure [38,67]. Nevertheless, since our current study includes more diverse genomes from more geographical regions, it provides additional support to this notion and reinforces the observation that SPA possesses a closed pangenome.

Comparison of the nucleotide diversity (π) of the core genome of STM, STY and SPA indicated that STY presents the lowest level of nucleotide diversity among these three serovars and that this serovar demonstrates a clonal population structure. This conclusion is well aligned with the notion that STY is a genetically monomorphic pathogen with a slow mutation rate and infrequent recombination events [43,74]. These results also suggest that STY is under relatively little selective pressure from its human host [44] and that typhoid in humans is a relatively new disease [75].

Bacterial core genome consists of conserved genes with constant presence, which are involved in cellular functions. In contrast, the accessory genome comprises the genes that vary between specific strains, providing bacteria the ability to adapt to particular environments or lifestyles including the acquisition of pathogenic traits [76–78]. Here, we showed that while the three analyzed serovars had a relatively similar core-genome size, the accessory genome of STM (8290 genes) was more than four times larger than that of the typhoidal datasets (1991 and 1912 genes for SPA and STY, respectively).

MGEs play a key role in driving genetic diversity and shaping the evolutionary routes of bacteria, while enabling them to adapt to various environmental challenges and life styles [61]. Here we compared the composition of conjugative elements, plasmids and prophages between the three studied *Salmonella* serovars. Our analyses showed significantly higher occurrence of MGEs in STM and STY than in SPA and higher diversity of MGE in STM compared to their diversity in SPA and STY. These results suggest

that SPA genomes have either acquired much less mobile genetic elements than STY and STM, or that these MGEs are not maintained for a long time in SPA genome. The lower frequency of SPA MGEs can therefore be explained by the presence of more stringent defense systems against foreign DNA. It is possible that SPA have developed strong defense mechanisms against foreign DNA making it more difficult for MGEs to establish and persist in the SPA population under natural conditions. Alternative explanation can be that while MGEs like plasmids and ICEs typically impose a metabolic burden on their bacterial hosts, SPA might have limited exposure to antibiotics compared to STM due to the lack of environmental reservoirs. Reduced exposure to antibiotics and other environmental stressors could lessen the selective advantage conferred by MGEs carrying antibiotic resistance genes and that the fitness cost of maintaining these elements might outweigh their potential benefits leading to their loss over time. Differences in the frequency of MGEs and AMR genes between SPA and STY may be explained by the presence of “hotspots” for integration of AMR regions in the STY chromosome as suggested by recent reports [79,80].

The high diversity of plasmids found in STM is in agreement with the openness of its pangenome and with the high gene pool diversity accessible for STM, compared to STY and SPA. Furthermore, the low diversity of plasmids and prophages in STY might explain why, despite having a similar number of MGEs per genome as STM, STY has a more closed pangenome, which is less diverse than STM. The low prevalence of plasmids found in SPA is in agreement with other recent studies, reporting that most strains from serovar Paratyphi A lack any plasmid contigs [67,81].

In addition to the lower frequency of plasmids and prophages in SPA, our analysis indicated significantly lower frequency of AMR genes in SPA genomes in comparison to STM. Taken together, these findings suggest that STM frequently acquires MGEs and AMR genes through horizontal gene transfer events and that this pool of accessory genes might be less accessible to or less stable in SPA. In consequence, SPA accumulates more point mutations in core genes than horizontal acquisition of AMR genes to face antibiotic pressure. On the phenotypic level, unlike STM or STY, in which antibiotic resistance is often associated with the self-transmissible IncHI1 plasmids [82–84], most SPA isolates are susceptible to antimicrobials, including the traditional first-line antibiotics (ampicillin, chloramphenicol and sulfonamides). Nevertheless, decreased

fluoroquinolone susceptibility, which is often conferred by point mutations within the Quinolone Resistance Determining Region (QRDR) of the DNA gyrase (*gyrA*) or the topoisomerase IV (*parC*) genes, has been reported in recent years for SPA [48,66,67].

The low frequency and diversity of MGEs in SPA might also be the result of a certain repertoire of defense systems associated with SPA genomes that reduces the acquisition or the maintenance of MGEs carrying AMR genes into the SPA genome. Interestingly, the Septu defense system [85] was found only in SPA, but not in other STM or STY genomes and additional three systems including PD-T4-3 [86], Druantia [87], and Rst_3HP [88] were found mostly in SPA genomes, with only a limited presence in STY. Therefore, differences in the composition of defense systems may suggest that each serovar appears to harbor a distinct collection of defense factors, that are shaped by its interaction with the environment or host, as was previously observed for distinct *E. coli* phylogroups [89]. Therefore, it is possible that a specific combination of defense systems in SPA contributes to the limited presence of MGEs and AMR genes in the SPA genomes. Possible demonstration for such effect was given by the different conjugation frequency of plasmid pRP4 that was shown to be conjugated at much lower frequencies into a SPA strain than into STM strains. Nevertheless, further mechanistic examination using a wider array of MGEs, including naturally circulating plasmids and phages is required to test this hypothesis.

Despite interesting and novel findings highlighted in this report, one of its limitations is the relatively low number genomes that were included in the final analyses in comparison to recent studies that have analyzed the population structure of SPA [48] and STY [43,79,90]. In this context, it is important to note that the final list of genomes included in this study resulted from a rationalized dereplication step that was applied to reduce redundancy and include the best quality representative genomes from a larger cohort of genomes. Additional confine that affected our sample size is the necessity to include only complete (or close to complete) high-quality sequences for a reliable analysis of pseudogenes, AMR sequences and MGEs, that may be affected by sequencing or assembly errors.

Conclusions

We present an updated population structure of SPA and show previously uncharacterized lineages of this serovar. Our genomic comparisons between STY and SPA across multiple genomes indicated that 5.15%, and

4.74% of STY and SPA core genes, respectively are pseudogenes, demonstrating a significant genomic decay in these serovars. Moreover, we demonstrated that although STY and SPA are considered to be closely related at the DNA level [16,19], they present distinct differences in the content of plasmids, prophages, antimicrobial resistance genes and defense systems. These results suggest that although SPA and STY share a similar lifestyle and cause symptomatic-indistinguishable disease, their genomic evolution and possibly their access to accessory gene pools is different.

Acknowledgements

We are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing and storage resources". MIGALE is part of the Institut Français de Bioinformatique (ANR-11-INBS-0013).

Author contributions statement

Conceptualization: HC, CC and OGM; data curation: BP, CC and SD; investigation: BP, CC and SD; writing-the draft: CC, OGM and HC; writing-review and editing: HC and OGM. All authors have read and approved the final manuscript and agree to be personally accountable for author's own contributions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Infect-Era program, SalHostTrop project ("Understanding the Human-Restricted Host Tropism of Typhoidal Salmonella", 2016–2020). The work at the Gal-Mor laboratory was supported by Infect-Era/Chief Scientist Ministry of Health under grant number 3–12435; the German-Israeli Foundation for Scientific Research and Development (GIF) under grant number I-41-416.6-2018; and the Israel Science Foundation (ISF) under grant numbers 2616/18 and 1228/23.

Data availability statement

The genome of SPA45157 is available from NCBI under accession number CP156168. The chromosome of STY120130191 is available from NCBI under accession number CP156169 and its associated IncFIB replicon of 106,706 bp under accession number CP156170. The data that support the findings of this study as well as its associated supplementary materials are openly available in <https://doi.org/10.6084/m9.figshare.26496382.v2>.

Ethical approval statement

Collecting and sequencing clinical *Salmonella* isolates (SPA45157 and STY120130191) were conducted under approval number 7072–09-SMC of the Sheba Medical Center ethics committee. Sequencing data was generated and analyzed anonymously and are not associated with any patient identifying details. All other *Salmonella enterica* genome sequences used in this study were retrieved from the National Center for Biotechnology Information (NCBI) GenBank database. These genomes are publicly available and do not require additional ethical approval for secondary bioinformatics analysis. No human subjects or personally identifiable information were involved in this research.

ORCID

Ohad Gal-Mor  <http://orcid.org/0000-0002-2540-6790>

References

- [1] Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, et al. Supplement 2008–2010 (no. 48) to the white-Kauffmann-Le minor scheme. *Res Microbiol.* 2014;165(7):526–530. doi: [10.1016/j.resmic.2014.07.004](https://doi.org/10.1016/j.resmic.2014.07.004)
- [2] Grimont P, Weill F-X. Antigenic Formulae of the *Salmonella* serovars. 9th ed. Paris: WHO Collaborating Centre for Reference and Research on *Salmonella*; Institute Pasteur. 2007. p 1–166.
- [3] Wang BX, Butler DS, Hamblin M, et al. One species, different diseases: the unique molecular mechanisms that underlie the pathogenesis of typhoidal *Salmonella* infections. *Curr Opin Microbiol.* 2023;72:102262. doi: [10.1016/j.mib.2022.102262](https://doi.org/10.1016/j.mib.2022.102262)
- [4] Uzzau S, Brown DJ, Wallis T, et al. Host adapted serotypes of *Salmonella enterica*. *Epidemiol Infect.* 2000;125(2):229–255. doi: [10.1017/S0950268899004379](https://doi.org/10.1017/S0950268899004379)
- [5] Gal-Mor O. Persistent infection and long-term carriage of Typhoidal and nontyphoidal *Salmonellae*. *Clin Microbiol Rev.* 2019;32(1). doi: [10.1128/CMR.00088-18](https://doi.org/10.1128/CMR.00088-18)
- [6] Gal-Mor O, Boyle EC, Grassl GA. Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front Microbiol.* 2014;5:391. doi: [10.3389/fmicb.2014.00391](https://doi.org/10.3389/fmicb.2014.00391)
- [7] Havelaar AH, Kirk MD, Torgerson PR, et al. World health organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *PLOS Med.* 2015;12(12):e1001923. doi: [10.1371/journal.pmed.1001923](https://doi.org/10.1371/journal.pmed.1001923)
- [8] Crump JA, Luby SP, Mintz ED. The global burden of typhoid fever. *Bull World Health Organ.* 2004;82(5):346–353.
- [9] Deng L, Song J, Gao X, et al. Host adaptation of a bacterial toxin from the human pathogen *Salmonella* Typhi. *Cell.* 2014;159(6):1290–1299. doi: [10.1016/j.cell.2014.10.057](https://doi.org/10.1016/j.cell.2014.10.057)
- [10] Spano S, Galan JE. A Rab32-dependent pathway contributes to *Salmonella* typhi host restriction. *Science.* 2012;338(6109):960–963. doi: [10.1126/science.1229224](https://doi.org/10.1126/science.1229224)

- [11] Bakker HC, Switt AI, Cummings CA, et al. A whole-genome single nucleotide polymorphism-based approach to trace and identify outbreaks linked to a common *Salmonella enterica* subsp. *enterica* serovar Montevideo pulsed-field gel electrophoresis type. *Appl Environ Microbiol.* 2011;77(24):8648–8655. doi: [10.1128/AEM.06538-11](https://doi.org/10.1128/AEM.06538-11)
- [12] Hiyoshi H, Wangdi T, Lock G, et al. Mechanisms to evade the phagocyte respiratory burst arose by convergent evolution in Typhoidal *Salmonella* Serovars. *Cell Rep.* 2018;22(7):1787–1797. doi: [10.1016/j.celrep.2018.01.016](https://doi.org/10.1016/j.celrep.2018.01.016)
- [13] Langridge GC, Fookes M, Connor TR, et al. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proc Natl Acad Sci U S A.* 2015;112(3):863–868. doi: [10.1073/pnas.1416707112](https://doi.org/10.1073/pnas.1416707112)
- [14] Stevens MP, Kingsley RA. *Salmonella* pathogenesis and host-adaptation in farmed animals. *Curr Opin Microbiol.* 2021;63:52–58. doi: [10.1016/j.mib.2021.05.013](https://doi.org/10.1016/j.mib.2021.05.013)
- [15] Parkhill J, Dougan G, James KD, et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature.* 2001;413(6858):848–852. doi: [10.1038/35101607](https://doi.org/10.1038/35101607)
- [16] McClelland M, Sanderson KE, Clifton SW, et al. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of *Salmonella enterica* that cause typhoid. *Nat Genet.* 2004;36(12):1268–1274. doi: [10.1038/ng1470](https://doi.org/10.1038/ng1470)
- [17] Andersson JO, Andersson SG. Genome degradation is an ongoing process in *Rickettsia*. *Mol Biol Evol.* 1999;16(9):1178–1191. doi: [10.1093/oxfordjournals.molbev.a026208](https://doi.org/10.1093/oxfordjournals.molbev.a026208)
- [18] Thomson NR, Howard S, Wren BW, et al. The complete genome sequence and comparative genome analysis of the high pathogenicity *Yersinia enterocolitica* strain 8081. *PLOS Genet.* 2006;2(12):e206. doi: [10.1371/journal.pgen.0020206](https://doi.org/10.1371/journal.pgen.0020206)
- [19] Holt KE, Thomson NR, Wain J, et al. Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars paratyphi A and Typhi. *BMC Genomics.* 2009;10(1):36. doi: [10.1186/1471-2164-10-36](https://doi.org/10.1186/1471-2164-10-36)
- [20] Gal-Mor O, Suez J, Elhadad D, et al. Molecular and cellular characterization of a *Salmonella enterica* serovar paratyphi a outbreak strain and the human immune response to infection. *Clin Vaccine Immunol.* 2012;19(2):146–156. doi: [10.1128/CVI.05468-11](https://doi.org/10.1128/CVI.05468-11)
- [21] Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* 2014;30(14):2068–2069. doi: [10.1093/bioinformatics/btu153](https://doi.org/10.1093/bioinformatics/btu153)
- [22] Cantalapiedra CP, Hernandez-Plaza A, Letunic I, et al. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol.* 2021;38(12):5825–5829. doi: [10.1093/molbev/msab293](https://doi.org/10.1093/molbev/msab293)
- [23] Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLOS ONE.* 2014;9(11):e112963. doi: [10.1371/journal.pone.0112963](https://doi.org/10.1371/journal.pone.0112963)
- [24] Sallet E, Gouzy J, Schiex T. EuGene-PP: a next-generation automated annotation pipeline for prokaryotic genomes. *Bioinformatics.* 2014;30(18):2659–2661. doi: [10.1093/bioinformatics/btu366](https://doi.org/10.1093/bioinformatics/btu366)
- [25] Galperin MY, Makarova KS, Wolf YI, et al. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* 2015;43(D1):D261–269. doi: [10.1093/nar/gku1223](https://doi.org/10.1093/nar/gku1223)
- [26] Olm MR, Brown CT, Brooks B, et al. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *Isme J.* 2017;11(12):2864–2868. doi: [10.1038/ismej.2017.126](https://doi.org/10.1038/ismej.2017.126)
- [27] Jain C, Rodriguez RL, Phillippy AM, et al. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun.* 2018;9(1):5114. doi: [10.1038/s41467-018-07641-9](https://doi.org/10.1038/s41467-018-07641-9)
- [28] Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics.* 2015;31(22):3691–3693. doi: [10.1093/bioinformatics/btv421](https://doi.org/10.1093/bioinformatics/btv421)
- [29] Cheng L, Connor TR, Siren J, et al. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol.* 2013;30(5):1224–1228. doi: [10.1093/molbev/mst028](https://doi.org/10.1093/molbev/mst028)
- [30] Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10(1):421. doi: [10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421)
- [31] Thomas PD, Ebert D, Muruganujan A, et al. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* 2022;31(1):8–22. doi: [10.1002/pro.4218](https://doi.org/10.1002/pro.4218)
- [32] Guo J, Bolduc B, Zayed AA, et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome.* 2021;9(1):37. doi: [10.1186/s40168-020-00990-y](https://doi.org/10.1186/s40168-020-00990-y)
- [33] Nayfach S, Camargo AP, Schulz F, et al. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol.* 2021;39(5):578–585. doi: [10.1038/s41587-020-00774-7](https://doi.org/10.1038/s41587-020-00774-7)
- [34] Bin Jang H, Bolduc B, Zablocki O, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol.* 2019;37(6):632–639. doi: [10.1038/s41587-019-0100-8](https://doi.org/10.1038/s41587-019-0100-8)
- [35] Redondo-Salvo S, Bartomeus-Penalver R, Vielva L, et al. COPLA, a taxonomic classifier of plasmids. *BMC Bioinformatics.* 2021;22(1):390. doi: [10.1186/s12859-021-04299-x](https://doi.org/10.1186/s12859-021-04299-x)
- [36] Tesson F, Herve A, Mordret E, et al. Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun.* 2022;13(1):2561. doi: [10.1038/s41467-022-30269-9](https://doi.org/10.1038/s41467-022-30269-9)
- [37] Virtanen P, Gommers R, Oliphant TE, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261–272. doi: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2)
- [38] Liang W, Zhao Y, Chen C, et al. Pan-genomic analysis provides insights into the genomic variation and evolution of *Salmonella paratyphi a*. *PLOS ONE.* 2012;7(9):e45346. doi: [10.1371/journal.pone.0045346](https://doi.org/10.1371/journal.pone.0045346)
- [39] Yu G. Using ggtree to visualize data on tree-like structures. *Curr Protoc Bioinformatics.* 2020;69(1):e96. doi: [10.1002/cpbi.96](https://doi.org/10.1002/cpbi.96)
- [40] Pornsukarom S, van Vliet AHM, Thakur S. Whole genome sequencing analysis of multiple *Salmonella* serovars provides insights into phylogenetic relatedness, antimicrobial resistance, and virulence markers

- across humans, food animals and agriculture environmental sources. *BMC Genomics*. 2018;19(1):801. doi: [10.1186/s12864-018-5137-4](https://doi.org/10.1186/s12864-018-5137-4)
- [41] Thilliez G, Mashe T, Chaibva BV, et al. Population structure of *Salmonella enterica* Typhi in Harare, Zimbabwe (2012–19) before typhoid conjugate vaccine roll-out: a genomic epidemiology study. *Lancet Microbe*. 2023;4(12):e1005–e1014. doi: [10.1016/S2666-5247\(23\)00214-8](https://doi.org/10.1016/S2666-5247(23)00214-8)
- [42] Dyson ZA, Malau E, Horwood PF, et al. Whole genome sequence analysis of *Salmonella* Typhi in Papua New Guinea reveals an established population of genotype 2.1.7 sensitive to antimicrobials. *PLOS Negl Trop Dis*. 2022;16(3):e0010306. doi: [10.1371/journal.pntd.0010306](https://doi.org/10.1371/journal.pntd.0010306)
- [43] Dyson ZA, Holt KE. Five years of GenoTyphi: updates to the global salmonella typhi genotyping framework. *J Infect Dis*. 2021;224(12 Suppl 2):S775–S780. doi: [10.1093/infdis/jiab414](https://doi.org/10.1093/infdis/jiab414)
- [44] Yap KP, Thong KL. *Salmonella* typhi genomics: envisaging the future of typhoid eradication. *Trop Med Int Health*. 2017;22(8):918–925. doi: [10.1111/tmi.12899](https://doi.org/10.1111/tmi.12899)
- [45] Tonkin-Hill G, Lees JA, Bentley SD, et al. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res*. 2019;47(11):5539–5549. doi: [10.1093/nar/gkz361](https://doi.org/10.1093/nar/gkz361)
- [46] Zhou Z, McCann A, Weill FX, et al. Achtman M: transient Darwinian selection in *Salmonella enterica* serovar paratyphi a during 450 years of global spread of enteric fever. *Proc Natl Acad Sci U S A*. 2014;111(33):12199–12204. doi: [10.1073/pnas.1411012111](https://doi.org/10.1073/pnas.1411012111)
- [47] Tettelin H, Riley D, Cattuto C, et al. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol*. 2008;11(5):472–477. doi: [10.1016/j.mib.2008.09.006](https://doi.org/10.1016/j.mib.2008.09.006)
- [48] Jacob JJ, Pragasam AK, Vasudevan K, et al. Genomic analysis unveils genome degradation events and gene flux in the emergence and persistence of *S. Paratyphi* a lineages. *PLOS Pathog*. 2023;19(4):e1010650. doi: [10.1371/journal.ppat.1010650](https://doi.org/10.1371/journal.ppat.1010650)
- [49] Deng W, Liou SR, Plunkett G, et al. Comparative genomics of *Salmonella enterica* serovar Typhi strains Ty2 and CT18. *J Bacteriol*. 2003;185(7):2330–2337. doi: [10.1128/JB.185.7.2330-2337.2003](https://doi.org/10.1128/JB.185.7.2330-2337.2003)
- [50] Brockhurst MA, Harrison E, Hall JPJ, et al. The ecology and evolution of pangenomes. *Curr Biol*. 2019;29(20):R1094–R1103. doi: [10.1016/j.cub.2019.08.012](https://doi.org/10.1016/j.cub.2019.08.012)
- [51] Hall JPJ, Brockhurst MA, Harrison E. Sampling the mobile gene pool: innovation via horizontal gene transfer in bacteria. *Philos Trans R Soc Lond B Biol Sci*. 2017;372(1735):20160424. doi: [10.1098/rstb.2016.0424](https://doi.org/10.1098/rstb.2016.0424)
- [52] de la Cruz F, Davies J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol*. 2000;8(3):128–133. doi: [10.1016/S0966-842X\(00\)01703-0](https://doi.org/10.1016/S0966-842X(00)01703-0)
- [53] Chen Z, Kuang D, Xu X, et al. Genomic analyses of multidrug-resistant *Salmonella* Indiana, typhimurium, and enteritidis isolates using MinION and MiSeq sequencing technologies. *PLOS ONE*. 2020;15(7):e0235641. doi: [10.1371/journal.pone.0235641](https://doi.org/10.1371/journal.pone.0235641)
- [54] Aviv G, Tsyba K, Steck N, et al. A unique megaplasmid contributes to stress tolerance and pathogenicity of an emergent *Salmonella enterica* serovar infantis strain. *Environ Microbiol*. 2014;16(4):977–994. doi: [10.1111/1462-2920.12351](https://doi.org/10.1111/1462-2920.12351)
- [55] Zhang JF, Fang LX, Chang MX, et al. A trade-off for maintenance of multidrug-resistant IncHI2 plasmids in *Salmonella enterica* serovar Typhimurium through adaptive evolution. *mSystems*. 2022;7(5):e0024822. doi: [10.1128/msystems.00248-22](https://doi.org/10.1128/msystems.00248-22)
- [56] Mylona E, Sanchez-Garrido J, Hoang Thu TN, et al. Very long O-antigen chains of *Salmonella* Paratyphi a inhibit inflammasome activation and pyroptotic cell death. *Cell Microbiol*. 2021;23(5):e13306. doi: [10.1111/cmi.13306](https://doi.org/10.1111/cmi.13306)
- [57] Humphrey B, Thomson NR, Thomas CM, et al. Fitness of *Escherichia coli* strains carrying expressed and partially silent IncN and IncP1 plasmids. *BMC Microbiol*. 2012;12(1):53. doi: [10.1186/1471-2180-12-53](https://doi.org/10.1186/1471-2180-12-53)
- [58] Fricke WF, McDermott PF, Mammel MK, et al. Antimicrobial resistance-conferring plasmids with similarity to virulence plasmids from avian pathogenic *Escherichia coli* strains in *Salmonella enterica* serovar Kentucky isolates from poultry. *Appl Environ Microbiol*. 2009;75(18):5963–5971. doi: [10.1128/AEM.00786-09](https://doi.org/10.1128/AEM.00786-09)
- [59] Pansegrau W, Lanka E, Barth PT, et al. Complete nucleotide sequence of Birmingham IncPa plasmids. *J Mol Biol*. 1994;239(5):623–663. doi: [10.1006/jmbi.1994.1404](https://doi.org/10.1006/jmbi.1994.1404)
- [60] Fangman WL, Novick A. Mutant bacteria showing efficient utilization of thymidine. *J Bacteriol*. 1966;91(6):2390–2391. doi: [10.1128/jb.91.6.2390-2391.1966](https://doi.org/10.1128/jb.91.6.2390-2391.1966)
- [61] Rocha EPC, Bikard D. Microbial defenses against mobile genetic elements and viruses: who defends whom from what? *PLOS Biol*. 2022;20(1):e3001514. doi: [10.1371/journal.pbio.3001514](https://doi.org/10.1371/journal.pbio.3001514)
- [62] Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. *Nat Rev Microbiol*. 2023;21(10):686–700. doi: [10.1038/s41579-023-00934-x](https://doi.org/10.1038/s41579-023-00934-x)
- [63] Arndt MB, Mosites EM, Tian M, et al. Estimating the burden of paratyphoid a in Asia and Africa. *PLOS Negl Trop Dis*. 2014;8(6):e2925. doi: [10.1371/journal.pntd.0002925](https://doi.org/10.1371/journal.pntd.0002925)
- [64] Sahastrabuddhe S, Carbis R, Wierzb TF, et al. Increasing rates of *Salmonella* Paratyphi a and the current status of its vaccine development. *Expert Rev Vaccines*. 2013;12(9):1021–1031. doi: [10.1586/14760584.2013.825450](https://doi.org/10.1586/14760584.2013.825450)
- [65] Connor BA, Schwartz E. Typhoid and paratyphoid fever in travellers. *Lancet Infect Dis*. 2005;5(10):623–628. doi: [10.1016/S1473-3099\(05\)70239-5](https://doi.org/10.1016/S1473-3099(05)70239-5)
- [66] Browne AJ, Kashef Hamadani BH, Kumaran EAP, et al. Drug-resistant enteric fever worldwide, 1990 to 2018: a systematic review and meta-analysis. *BMC Med*. 2020;18(1):1. doi: [10.1186/s12916-019-1443-1](https://doi.org/10.1186/s12916-019-1443-1)
- [67] Rahman SIA, Nguyen TNT, Khanam F, et al. Genetic diversity of *Salmonella* Paratyphi a isolated from enteric fever patients in Bangladesh from 2008 to 2018. *PLOS Negl Trop Dis*. 2021;15(10):e0009748. doi: [10.1371/journal.pntd.0009748](https://doi.org/10.1371/journal.pntd.0009748)
- [68] McClelland M, Sanderson KE, Spieth J, et al. Complete genome sequence of *Salmonella enterica* serovar Typhimurium LT2. *Nature*. 2001;413(6858):852–856. doi: [10.1038/35101614](https://doi.org/10.1038/35101614)

- [69] Wang BX, Leshchiner D, Luo L, et al. High-throughput fitness experiments reveal specific vulnerabilities of human-adapted *Salmonella* during stress and infection. *Nat Genet.* **2024**;56(6):1288–1299. doi: [10.1038/s41588-024-01779-7](https://doi.org/10.1038/s41588-024-01779-7)
- [70] Holt KE, Parkhill J, Mazzoni CJ, et al. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet.* **2008**;40(8):987–993. doi: [10.1038/ng.195](https://doi.org/10.1038/ng.195)
- [71] Pulford CV, Perez-Sepulveda BM, Canals R, et al. Stepwise evolution of *Salmonella* Typhimurium ST313 causing bloodstream infection in Africa. *Nat Microbiol.* **2021**;6(3):327–338. doi: [10.1038/s41564-020-00836-1](https://doi.org/10.1038/s41564-020-00836-1)
- [72] Cole ST, Eiglmeier K, Parkhill J, et al. Massive gene decay in the leprosy bacillus. *Nature.* **2001**;409(6823):1007–1011. doi: [10.1038/35059006](https://doi.org/10.1038/35059006)
- [73] Andersson JO, Andersson SG. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. *Mol Biol Evol.* **2001**;18(5):829–839. doi: [10.1093/oxfordjournals.molbev.a003864](https://doi.org/10.1093/oxfordjournals.molbev.a003864)
- [74] Wong VK, Baker S, Pickard DJ, et al. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental transmission events. *Nat Genet.* **2015**;47(6):632–639. doi: [10.1038/ng.3281](https://doi.org/10.1038/ng.3281)
- [75] Dougan G, Baker S. *Salmonella enterica* serovar typhi and the pathogenesis of typhoid fever. *Annu Rev Microbiol.* **2014**;68(1):317–336. doi: [10.1146/annurev-micro-091313-103739](https://doi.org/10.1146/annurev-micro-091313-103739)
- [76] Hacker J, Carniel E. Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* **2001**;2(5):376–381. doi: [10.1093/embo-reports/kve097](https://doi.org/10.1093/embo-reports/kve097)
- [77] Abby S, Daubin V. Comparative genomics and the evolution of prokaryotes. *Trends Microbiol.* **2007**;15(3):135–141. doi: [10.1016/j.tim.2007.01.007](https://doi.org/10.1016/j.tim.2007.01.007)
- [78] Dobrindt U, Hochhut B, Hentschel U, et al. Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol.* **2004**;2(5):414–424. doi: [10.1038/nrmicro884](https://doi.org/10.1038/nrmicro884)
- [79] Penil-Celis A, Tagg KA, Webb HE, et al. Mobile genetic elements define the non-random structure of the *Salmonella enterica* serovar Typhi pangenome. *mSystems.* **2024**;9(8):e0036524. doi: [10.1128/msystems.00365-24](https://doi.org/10.1128/msystems.00365-24)
- [80] Nair S, Chattaway M, Langridge GC, et al. ESBL-producing strains isolated from imported cases of enteric fever in England and Wales reveal multiple chromosomal integrations of blaCTX-M-15 in XDR *Salmonella* Typhi. *J Antimicrob Chemother.* **2021**;76(6):1459–1466. doi: [10.1093/jac/dkab049](https://doi.org/10.1093/jac/dkab049)
- [81] Kushwaha SK, Anand A, Wu Y, et al. Genomic plasticity is a blueprint of diversity in *Salmonella* lineages. *bioRxiv.* **2023**.
- [82] Britto CD, Dyson ZA, Duchene S, et al. Laboratory and molecular surveillance of paediatric typhoidal *Salmonella* in Nepal: antimicrobial resistance and implications for vaccine policy. *PLOS Negl Trop Dis.* **2018**;12(4):e0006408. doi: [10.1371/journal.pntd.0006408](https://doi.org/10.1371/journal.pntd.0006408)
- [83] Rahman SIA, Dyson ZA, Klemm EJ, et al. Population structure and antimicrobial resistance patterns of *Salmonella* Typhi isolates in urban Dhaka, Bangladesh from 2004 to 2016. *PLOS Negl Trop Dis.* **2020**;14(2):e0008036. doi: [10.1371/journal.pntd.0008036](https://doi.org/10.1371/journal.pntd.0008036)
- [84] Park SE, Pham DT, Boinett C, et al. The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. *Nat Commun.* **2018**;9(1):5094. doi: [10.1038/s41467-018-07370-z](https://doi.org/10.1038/s41467-018-07370-z)
- [85] Doron S, Melamed S, Ofir G, et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science.* **2018**;359(6379):359(6379). doi: [10.1126/science.aar4120](https://doi.org/10.1126/science.aar4120)
- [86] Vassallo CN, Doering CR, Littlehale ML, et al. A functional selection reveals previously undetected anti-phage defence systems in the *E. coli* pangenome. *Nat Microbiol.* **2022**;7(10):1568–1579. doi: [10.1038/s41564-022-01219-4](https://doi.org/10.1038/s41564-022-01219-4)
- [87] Gao L, Altae-Tran H, Bohning F, et al. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science.* **2020**;369(6507):1077–1084. doi: [10.1126/science.aba0372](https://doi.org/10.1126/science.aba0372)
- [88] Rousset F, Depardieu F, Miele S, et al. Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe.* **2022**;30(5):740–753.e5. doi: [10.1016/j.chom.2022.02.018](https://doi.org/10.1016/j.chom.2022.02.018)
- [89] Wu Y, Garushyants SK, van den Hurk A, et al. Bacterial defense systems exhibit synergistic anti-phage activity. *Cell Host Microbe.* **2024**;32(4):557–572.e6. doi: [10.1016/j.chom.2024.01.015](https://doi.org/10.1016/j.chom.2024.01.015)
- [90] Dyson ZA, Ashton PM, Khanam F, et al. Pathogen diversity and antimicrobial resistance transmission of *Salmonella enterica* serovars Typhi and Paratyphi a in Bangladesh, Nepal, and Malawi: a genomic epidemiological study. *Lancet Microbe.* **2024**;5(8):100841. doi: [10.1016/S2666-5247\(24\)00047-8](https://doi.org/10.1016/S2666-5247(24)00047-8)