Title
Evolutionary Genomics of Salmonella enterica Subspecies

Permalink
https://escholarship.org/uc/item/7vh8682z

Journal
mBio, 4(2)

ISSN
2150-7511

Authors
Desai, P. T
Porwollik, S.
Long, F.
et al.

Publication Date
2013-03-05

DOI
10.1128/mBio.00579-12

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed
Evolutionary Genomics of *Salmonella enterica* Subspecies

Prerak T. Desai, Steffen Porwollik, Fred Long, Pui Cheng, Aye Wollam, Sandra W. Clifton, George M. Weinstock, Michael McClelland

Department of Pathology and Laboratory Medicine, University of California, Irvine, Irvine, California, USA; Vaccine Research Institute of San Diego, San Diego, California, USA; The GENOME Institute, Washington University School of Medicine, St. Louis, Missouri, USA

**ABSTRACT** Six subspecies are currently recognized in *Salmonella enterica*. Subspecies I (subspecies *enterica*) is responsible for nearly all infections in humans and warm-blooded animals, while five other subspecies are isolated principally from cold-blooded animals. We sequenced 21 phylogenetically diverse strains, including two representatives from each of the previously unsequenced five subspecies and 11 diverse new strains from *S. enterica* subspecies *enterica*, to put this species into an evolutionary perspective. The phylogeny of the subspecies was partly obscured by abundant recombination events between lineages and a relatively short period of time within which subspeciation took place. Nevertheless, a variety of different tree-building methods gave congruent evolutionary tree topologies for subspeciation. A total of 285 gene families were identified that were recruited into subspecies *enterica*, and most of these are of unknown function. At least 2,807 gene families were identified in one or more of the other subspecies that are not found in subspecies I or *Salmonella bongori*. Among these gene families were 13 new candidate effectors and 7 new candidate fimbrial clusters. A third complete type III secretion system not present in subspecies *enterica* (I) isolates was found in both strains of subspecies *salamae* (II). Some gene families had complex taxonomies, such as the type VI secretion systems, which were recruited from four different lineages in five of six subspecies. Analysis of nonsynonymous-to-synonymous substitution rates indicated that the more-recently acquired regions in *S. enterica* are undergoing faster fixation rates than the rest of the genome. Recently acquired AT-rich regions, which often encode virulence functions, are under ongoing selection to maintain their high AT content.

**IMPORTANCE** We have sequenced 21 new genomes which encompass the phylogenetic diversity of *Salmonella*, including strains of the previously unsequenced subspecies arizonae, diarizonae, houtenae, salamae, and indica as well as new diverse strains of subspecies *enterica*. We have deduced possible evolutionary paths traversed by this very important zoonotic pathogen and identified novel putative virulence factors that are not found in subspecies I. Gene families gained at the time of the evolution of subspecies *enterica* are of particular interest because they include mechanisms by which this subspecies adapted to warm-blooded hosts.

*Salmonella* spp. cause about 1.3 billion cases of nontyphoidal salmonellosis worldwide each year (1). The economic burden due to salmonellosis in the United States alone is estimated to be ~$2.3 billion annually (2). *Salmonella* is also a major pathogen of domestic animals causing huge economic losses and providing a source of infection for humans (3). Serotyping-based identification of somatic (O) and flagellar (H) antigens was among the first methods used for taxonomic classification of *Salmonella*, and each serovar was initially considered a different species. However, cell surface antigens are sometimes horizontally transferred, a phenomenon that can cause classification of genetically unrelated strains within the same serovar (4, 5). *Salmonella* taxonomy took a major stride when Falkow and colleagues (6) used DNA hybridization to demonstrate that all tested serovars were related at the species level and identified five distinct subgenera within the species. *Salmonella* is now considered to consist of two species, *Salmonella bongori* and *Salmonella enterica*, and *S. enterica* is further classified into six subspecies, arizonae (IIIa), diarizonae (IIIb), houtenae (IV), salamae (II), indica (VI), and *enterica* (I) (7). *S. enterica* subsp. *enterica* (I) strains represent the vast majority of *Salmonella* strains isolated from humans and warm-blooded animals, while all the other subspecies and *S. bongori* are more typically (though not exclusively) isolated from cold-blooded animals (8). Approximately 50 of the nearly 2,600 known *Salmonella* serovars account for ~99% of all clinical isolates of *Salmonella* from humans and domestic mammals (9), and all of these 50 serovars are in subspecies I.

Genome sequencing efforts in *S. enterica* have so far focused on the most prevalent serovars of subspecies *enterica* (1). We have sequenced the genomes of eleven additional members of subspecies *enterica* (I), selected based on their diversity, and those of two different serovars from each of the other five known subspecies. We compared these sequences to the whole-genome sequences of *S. bongori* (10) and to seven previously sequenced subspecies en-
terica (I) strains. We construct phylogenetic hypotheses for these 29 genomes while taking into account the high rate of recombina-
tion among Salmonella strains (11–15). Because acquisition and
loss of genes is a major force driving the evolution of virulence in
Salmonella (16), we modeled the gain and loss of gene families at
each ancestral node. We reconstructed hypothetical gene contents of
the most recent common ancestor (MRCA) of each subspecies.
Gene families gained at the subspecies enterica (I) node may pro-
vide clues to the strategies and virulence factors that contributed
to the formation of a lineage which has evolved to infect prin-
cipally warm-blooded hosts. We also identified gene families under-
going accelerated evolution based on pairwise synonymous-to-
nonsynonymous single nucleotide polymorphism (SNP) ratios.
This group of gene families is particularly interesting because
some of this selection may be driven by newly acquired life strat-
egies or by interactions of the bacterium with the host.

RESULTS AND DISCUSSION

We sequenced 21 new Salmonella genomes, 10 of which were se-
quenced to completion while the remaining genomes were se-
quenced to obtain improved high-quality drafts. The strains and
sequencing statistics are summarized in Table S1 in the supple-
mental material along with information regarding other previ-
ously published strains and species to which these new genomes
were compared.

Among the 21 new genomes, two strains were selected from
each of the five previously unsequenced subspecies. In addition,
11 genomes were selected from 305 strains within subspecies en-
terica (I) that lacked many genes found in S. enterica subsp. en-
terica serovars Typhimurium LT2 and Typhi CT18 as well as
strains representing distant genomovars within a single serovar
based on comparative genomic hybridization (17–19) (data acces-
sible at https://dl.dropbox.com/u/99836585/MMCC_all_CGH
_100407.xlsx).

The fact that some of the genomes have not been sequenced to
completion means that some sequencing errors, misassemblies,
annotation errors due to collapsing of duplicate genes, and dupli-
cated annotations of genes that span contig boundaries still exist
in a few locations in a few genomes in this data set. The analyses we
perform below are designed to mitigate but not eliminate these
limitations. The numbers presented are all estimates constrained
by these caveats, and some analyses, such as strict tests of orthol-
ygy and studies of gene duplication, are not possible on our draft
genomes. Nevertheless, the obtained high-quality drafts permit a
fascinating insight into evolutionary processes during Salmonella
subspeciation.

Phylogenetic analysis. We used three different approaches to
predict phylogenetic relationships between orthologous “core”
regions shared by all subspecies of Salmonella. We first used
Mauve (20) to align the 29 Salmonella genomes included in our
study and used Escherichia coli K-12 as an outgroup. We identified
737,062 SNPs in the “core” regions of ~2.6 Mb that were present
and aligned with high confidence in all taxa. We used these data to
construct a bootstrapped maximum likelihood (ML) tree using
RAXML (version 7.2.6) (21). Figure 1A shows the cladogram con-
structed using this approach. The relationship between the sub-
species was supported in all 1,000 bootstrap replicates using ran-
dom samples of 50% of the SNP data.

To estimate the divergence times of each subspecies, codon
alignments were constructed for 2,025 genes present across all 30
genomes in single copies, as annotated by automated RAST (22)
or annotated in the publically available genomes. A total of
348,642 synonymous SNPs, which did not change the amino acid
sequence, were identified. Figure 1B shows a condensed and lin-
erized version of the tree built using these SNPs (with 1,000 boot-
straps on 50% of the data), calibrated based on a previously esti-
ated 140-million-year divergence time between Salmonella and
E. coli (23). The topology of this tree (Fig. 1B) was in agreement
with the tree of all “core” SNPs (Fig. 1A).

Using synonymous SNPs, it was estimated that subspecies en-
terica (I) diverged from its last common ancestor ~27 mil-
ion years ago and that the most recent common ancestor
(MRCA) of all analyzed subspecies enterica strains arose ~12 mil-
ion years ago. This estimate supports the possibility that the sub-
species evolved long after their respective preferred hosts.

As a distinct alternative strategy to determine phylogeny, we
built individual maximum likelihood (ML) trees for each of the
2,025 core genes. DNA sequences for each potentially orthologous
gene family were aligned using Muscle (version 3.8.1) (24), and
ML trees were constructed using RAxML (version 7.2.6) with
1,000 bootstraps on 50% of the data (21). A consensus tree based
on these 2,025 trees was generated in Phylip (version 3.69) (25)
using the extended majority consensus rule (Fig. 1C). The num-
bers on the internal nodes indicate the fractions of gene trees
which support the partition of the taxa at that node. The low
numbers of compatible trees at each node indicate a high level of
incongruence between individual gene trees and the consensus
tree. Nevertheless, the consensus tree for subspeciation displayed
the same topology as the SNP trees in Fig. 1A and B. There was
further congruence at each terminal subspecies node than at
ancestral nodes. This may either reflect that the subspecies are
acting like incipient species, with a barrier to intraspecies recom-
bination, or indicate that there was insufficient time in which re-
combination could have taken place after the subspecies arose, or
both.

Two mechanisms that can introduce phylogenetic discordance
between orthologous genes in gene trees and species trees are in-
tergenomic recombination (gene conversion) and incomplete line-
age sorting (ILS) (26). Given the relatively short branch lengths
within which the subspecies arose (Fig. 1B), incomplete lineage
sorting as a source of gene tree discordance is possible. However,
gene conversion events can reintroduce previously lost allelic
changes into recombining populations. It is not possible to distin-
guish between these two mechanisms using data from contempo-
rary Salmonella strains.

Salmonella genomes undergo frequent intergenomic recombi-
nations (11–15) and thereby disrupt the clonal inheritance pattern
within the recombined regions. Clonal Frame (27) takes recombi-
nation into account during inference of phylogenetic relation-
ships. Clonal Frame also attempts to reconstruct the mutation and
recombination events that give rise to a particular lineage on the
branches of the phylogenetic tree. The program is computa-
tionally intensive; therefore, it was necessary to use only a subset of all
available data. Core genome regions of greater than 10 kb that
were shared by all 30 taxa were extracted from Mauve-based
whole-genome alignments. This resulted in extraction of 8 blocks
whose total alignment size was 104,029 bp, representing ~4% of
the core genome, which was used as an input for Clonal Frame.
Two replicate runs of 200,000 Markov chain Monte Carlo
(MCMC) iterations were performed. The tree topology for the
subspecies obtained from Clonal Frame (Fig. 1D) predicted a branching order of the subspecies similar to that of the other tree-building methods and further suggested that, within subspecies *enterica* (I), the previously defined clade B is a separate ancient lineage (13, 28, 29). Five of the subspecies *enterica* (I) genomes we sequenced proved to be new members of clade B. No additional consensus was achieved for the branching order within the strains we used in subspecies *enterica* (I).

**Recombination.** The importance of recombination in the evolution of *S. enterica* is well established (11–15). Clonal Frame analysis suggested that recombination played an equally important role as mutation in the evolution of the subspecies (recombination-to-mutation ratio \[r/m\] = 0.94). We used the Recombination Detection Program (RDP) (30) to determine the most likely regions of recombination in each genome. This tool suggested hundreds of such events per genome (data not shown). As an illustration, we annotated the potential gene conversion events in just one strain, the archetype strain for *S. enterica* subsp. *enterica* (I) serovar Typhimurium—strain LT2 (31). We annotated only the recombination events in which the potential source of the sequence in LT2 was not within subspecies *enterica* (I). This information is presented in Table S2 in the supplemental material. In brief, approximately 120 conversion events encompassing around 347 kb (336 genes) of *S. Typhimurium* LT2 sequence were predicted to have a source outside subspecies *enterica* (I).

**Gain and loss of genes during the evolution of *Salmonella*.**

Gain and loss of genes is one of the major mechanisms that drives diversification within bacteria (32). Homologous gene families among the 29 *Salmonella* genomes were identified using FastOrtho (33), a reimplementation of OrthoMCL (34), and is available at [http://enews.patricbrc.org/fastortho/](http://enews.patricbrc.org/fastortho/). For the 29 *Salmonella* genomes covering the known phylogenetic diversity of *Salmonella*, we identified 11,443 FastOrtho gene families (including orthologs and close paralogs) in the pangenome, of which 3,221 gene families were conserved across all analyzed strains. Figure 2 shows the rate of identification of new gene families observed...
with each new sequence, based on a rarefaction curve that was calculated using EstimateS (35), where the number of gene families equals $4,922.9 \times (\text{number of genomes})^{0.2492}$ and $R^2 = 0.9985$. This approximation, which is highly reliant on the level of diversity already sampled, estimates that about 100 ± 63 (mean ± SD) new families would be observed in the next additional Salmonella genome sequence.

Some of the genomes we analyzed are incomplete, leading to some annotation inconsistencies, especially at contig boundaries and in multicity genes. To reduce the rate of false negatives (i.e., calling a gene absent because it was not annotated), we augmented the phyletic pattern with a tblatx analysis as detailed in Materials and Methods. The augmented phyletic pattern for all identified gene families (see Table S2 in the supplemental material) was created based on the presence and absence of a homologous gene across all strains. The DNA sequence of an exemplar for each family is included in Table S2, and Table S3 contains the locus tags of all annotated genes that belong to each gene family. These data were used to model gain and loss of gene families at all nodes of the phylogenetic hypothesis (Fig. 1A) by computing posterior probabilities for gain and loss of gene families at inner nodes using Count (36). Gitools (37) was then used to estimate statistical over-representation of gene families gained and lost at each ancestral node within specific gene sets. Figure 3 shows the gain and loss of specific gene sets that were statistically enriched (false-discovery rate [FDR], ≤0.1) at major ancestral nodes. Figure 4 summarizes the presence/absence profile of important gene sets across the 30 strains analyzed. A complete list of all predicted gene family gains and losses can be found in Table S4. Using Count (36), 3,379 gene families within the observed pangenome were predicted to have been gained more than once (homoplasies) across the evolutionary tree. Hence, at least ~30% of the pangenome appears to be shared within the genus by multiple horizontal transfers. Gene sets enriched for harboring homoplastic gene families are indicated in Fig. 4. This list is dominated by virulence factors (SPI-7, -8, -10, -12, -13, -14, and -19 and the ye-rsinia bacter cluster) and includes eleven fimbrial clusters and six uncharacterized islands. Metabolic traits enriched for homoplastic gene families include allantoin utilization and inositol catabolism.

Due to some fragmented genes being annotated twice at contig boundaries and the collapse of some regions of known duplication in the incomplete genomes, we refrained from modeling gene duplications onto the phylogenetic tree. However, because fimbrial gene clusters are of general interest due to their many paralogs and their suspected role in pathoadaptation, we manually curated the phyletic pattern of genes in fimbrial gene families, taking into account genomic context whenever possible (see Table S5 in the supplemental material). For all other genes, the individual gene trees provided in Table S2 can be used to resolve the paralogs on a gene-by-gene basis.

**Gene families gained by the most recent common ancestor of all Salmonella subspecies.** Count (36) predicted that the most recent common ancestor (MRCA) of all Salmonella subspecies gained an estimated 657 gene families (including orthologs and close paralogs) while diverging from a common ancestor with *Escherichia*. A total of 454 of these gene families are present in all 29 analyzed Salmonella genomes. A total of 347 out of the 657 genes were estimated to be gained in 90 islands of two or more genes. Important acquisitions from a virulence perspective included SPI-1 (encoding a type III secretion system [T3SS] and effectors required for invasion of eukaryotic cells [38]), SPI-4 (a type I secretion system for toxin delivery [38]), five genes in SPI-5 (T3SS effectors [38]), the tetrathionate respiration cluster (39), and the Bcf fimbrial cluster required for colonization of Peyer’s patches (40). Other genes gained included 9 glycosyl hydrolases, 41 genes with cyclic AMP receptor protein (CRP) binding sites (genes involved in carbohydrate catabolism under catabolite repression), and 9 genes (three operons) coding for anaerobic respiration.
FIG 3 Gene sets gained and lost at major ancestral nodes. Only statistically overrepresented gene sets (FDR < 10%) are shown. Subspecies enterica (I) is to the left of the vertical red line. Each cell is colored based on the ratio of genes gained/lost to the total number of genes in the gene set. See Table S7 in the supplemental material for identification of genes belonging to each gene set.

ductases (other than the tetrahionate reductase), which enable use of alternate electron acceptors. Hence, after diverging from E. coli, the most recent common ancestor of all Salmonella subspecies appears to have gained genes that enable utilization of different carbohydrates under anaerobic conditions. A total of 224 hypothetical proteins with no currently attributed function were also gained at this node, of which 145 were conserved in all 29 Salmonella strains analyzed. The latter class is a particularly intriguing subgroup for further study.

Genes gained and lost by the hypothetical S. bongori ancestor. Using an assumption of maximum parsimony, S. bongori retained all ancestral genes gained by the MRCA of Salmonella but lost 63 genes ancestral to the MRCA of Salmonella and Escherichia. These losses included islands required for ethanolamine utilization (STM2454-STM2470), fucose catabolism (STM2974-STM2979), and nitrate and nitrite ammonification (STM4277-STM4280). Hence, S. bongori either failed to gain or lost specific carbon utilization and anaerobic respiratory pathways that may contribute to survival in an inflamed gut (41). This species also gained five fimbrial clusters. S. bongori gained a phylogenetically distinct type VI secretion system (54736.3.peg.1295-54736.3.peg.1308) which seems to remain unique to S. bongori (10). In addition to the T3SS effectors gained by the MRCA of Salmonella, S. bongori gained seven additional effectors, all of which were apparently gained in multiple other lineages (homoplasies). These data are illustrated in Table S6 in the supplemental material, which lists the absence/presence status of all potential effector gene families identified in Salmonella.

Gene families gained by the most recent common ancestor of the species S. enterica. Under the assumption of maximum parsimony, when S. enterica separated from the MRCA of Salmonella, it gained 213 gene families (including orthologs and close paralogs), of which about 120 gene families were gained in 15 islands of two or more genes, for a total of approximately 108 insertion events. Gene families gained in islands of three or more included SPI-2 (STM1391-STM1422), which triggers colitis in the warm-blooded Salmonella host (42) and enables survival and replication in macrophages (42, 43), a vitamin B12 synthesis and propanediol utilization cluster (STM2019-STM2058) (44), a salmochelin-mediated iron acquisition cluster (STM2773-STM2777) that results in salmochelin-mediated lipocalin resistance in an inflamed environment (45), a putative sialic acid metabolism cluster (STM1127-STM1131), and three more islands (Fig. 3) encoding putative metabolic clusters. Although the putative sialic acid metabolism cluster has not been associated with virulence, its gain might also provide a metabolic advantage in the usually sialic acid-rich environment of the gut (46). A total of 71 hypothetical proteins were also gained at the node, of which 38 were conserved across all 28 S. enterica strains. The latter genes may define new capabilities that are preserved in all S. enterica strains.

Gene families gained and lost by the MRCA of S. enterica subsp. enterica (I). Among all the subspecies, strains in subspecies enterica (I) cause the vast majority of infections in warm-blooded animals. Hence, genes gained at this ancestral node might be important for the evolution of virulence in mammalian and bird hosts. A total of 285 gene families (including orthologs and close paralogs) were modeled as being gained by the MRCA of subspecies enterica (I), of which 167 were unique to that subspecies while 118 genes may have been acquired by multiple lineages through horizontal transfer (homoplasies) (see Table S4 in the supplemen-
FIG 4 Phyletic distribution of major groups of genes. Gene classes that include novel genes are shown in red. The number of genes in a gene set and the proportion of homoplastic genes within each gene set are shown on the right. The FDR (<10%) for homoplaspy enrichment is in green. Serovars of subspecies enterica (I) are shown to the left of the vertical red line. Each cell is shaded based on the ratio of genes present, compared to the total number of genes in the gene set. See Table S7 in the supplemental material for identification of genes belonging to each gene set.
tal material). A total of 185 gene families were gained in 47 islands of two or more genes; hence, subspecies enterica (I) gained 285 gene families in about 147 events.

The majority of genes gained by the common ancestor of subspecies enterica (I) had little or no homology to known functions. Among identifiable genes, notable gains without any homoplasies included SPI-6 (a type VI secretion system and Saf fimbriae), a 2-aminoethylphosphonate metabolism cluster (STM0426-STM0432), two islands (STM3779-STM3785 and STM4065-STM4080) needed for unknown carbohydrate utilization, and the autoinducer 2 (AI-2) transport and processing (STM4072-STM4079) cluster. No subspecies enterica (I)-specific effectors were discovered in this analysis, indicating that the acquisition of effectors may not have been part of the driving force in the foundation of this subspecies.

The role of SPI-6 in macrophage survival and cell invasion has been documented previously (44, 47), while the 2-aminoethylphosphonate cluster has been reported to be essential for long-term persistence in mice (48). STM3779-STM3785 is an AT-rich island and encodes a complete phosphotransferase system (PTS) with unknown specificity, along with a transcriptional regulator and a fructose-bisphosphate aldolase. This island was reported to be under histone-like nucleoid structuring protein (H-NS)-mediated repression (49). An analysis of all publicly available expression data for S. Typhimurium using Colombos (50) suggested that the expression of this island was induced upon treatment with H$_2$O$_2$ (51). Hence, this island might be important during intracellular survival of Salmonella.

STM4065-STM4071 codes for a permease, a transcriptional regulator, and metabolic enzymes for utilization of unknown carbohydrates. Colombos analysis revealed that expression of this cluster was induced during swimming motility compared to swarming motility in S. Typhimurium (Gene Expression Omnibus data set GSE1633 at [http://www.ncbi.nlm.nih.gov/geo/](http://www.ncbi.nlm.nih.gov/geo/)). A recent screen of a complete single-gene deletion library in mice revealed that four genes in this island were of functional importance for S. Typhimurium survival in the liver of mice (M. McClelland, unpublished data).

The role of AI-2 in Salmonella signaling and metabolism is not clear (52). The only known regulon under direct AI-2-mediated transcriptional control codes for active uptake and processing of AI-2 (53). Colombos analysis suggested that the AI-2 loci might be under arcA-mediated negative regulation in anaerobic conditions (54). LuxS-mediated production of AI-2 is conserved across all Salmonella spp., but the mechanism to sense AI-2 is present only in subspecies enterica (I) and absent in all other subspecies as well as in S. bongori. However, orthologs for AI-2 transport and processing are present in many other Enterobacteriaceae.

Gene families specific to other subspecies. A total of 2,807 gene families, which were found in one or more of the other five S. enterica subspecies but not present in any member of S. enterica subsp. enterica (I) or S. bongori, were identified. The majority of these gene families had little or no homology to known functions. A comprehensive list can be easily extracted from data presented in Table S2 in the supplemental material. Novel gene clusters encoding putative virulence functions include the following:

- **SPI-20** (55), which encodes a type VI secretion system present only in subsp. arizonae (IIIa)
- **SPI-21** (55), which encodes a type VI secretion system present only in subsp. diarizonae (IIIb)

- The locus of enterocyte effacement (LEE) (56), which—within Salmonella—is specific to subsp. salamae (II). The locus harbors a complete T3SS and is homologous and syntenic to the LEE in E. coli O157:H7. However, homologs of only 5 of the 24 known effectors translocated through this machinery in E. coli O157:H7 (57, 58), namely, tir, nleD, cseAB, cseD2, and cseT, were present in subsp. salamae.

- Seven new fimbrial operons (named sib, sic, sas, sdf, sbs, sti, and ssf) (Fig. 4)

- A novel family of cytotoxic distending toxins (523839.5.peg.1805, 523839.5.peg.1806, 523839.5.peg.1807) which was specific to subsp. arizonae (IIIa) and subsp. diarizonae (IIIb). This cluster codes for all three subunits of a functional multimeric toxin. The cluster is homologous to the alt cluster found in E. coli (59).

A total of 13 new putative effector families (see Table S6 in the supplemental material). Eight of these were specific to subsp. salamae (II), five of which were part of the LEE. One of these eight effectors (523838.4.peg.2994) may encode a chimeric protein with parts similar to EspF from E. coli O103:H25 and SopA from subsp. enterica (I).

**Gene families with higher evolutionary rates.** Different parts of the genomes diverge at different rates under different evolutionary pressures. To identify proteins evolving at a higher rate than the rest of the pan genome, pairwise ratios of synonymous substitutions per synonymous site to nonsynonymous substitutions per nonsynonymous site (dS/dN) were calculated for all protein coding genes within each gene family for all pairs of taxa. The pairwise ratios within each gene family were averaged to assign a mean evolutionary rate to each family. Using this approach, we observed that recent protein coding gene acquisitions evolve faster than ancestral genes after normalizing for the rate of neutral divergence. Figure 5 shows the distribution of mean log$_{10}$ values (dS/dN) for all nonhomoplastic genes that were recruited into the genomes at different times. There was a statistically significant difference ($P < 0.01$) in the evolutionary rate of genes gained at different evolutionary time points. Younger proteins evolved faster than ancient proteins. We also calculated the averages of log$_{10}$ values (dS/dN) within only subspecies enterica (I), in which this same trend was also observed (data not shown). Genes acquired through horizontal gene transfer (HGT) have previously been observed to evolve faster than core genes in E. coli (60).

An enrichment analysis of the pan genome ranked by mean log$_{10}$ values was performed to estimate the overabundance of fast-evolving proteins within gene sets grouped by a variety of criteria. This analysis (Table 1) suggested a higher proportion of fast-evolving proteins among genes with abnormal GC content, genes encoding plasmid-related functions (type IV secretion, toxin-antitoxin systems), and genes important for the interaction of Salmonella with the host. Among the known SPIs, SPI-2, SPI-6, SPI-7, SPI-8, SPI-10, and SPI-12 were also enriched for fast-evolving proteins.

**Evolution of differences in GC content across the genome.** The ancestral orthologous regions shared by Salmonella and E. coli have evolved to contain a different GC content, with Salmonella genomes being relatively more GC rich (see Fig. S1 in the supple-
mental material). In contrast, laterally transferred islands acquired by *Salmonella* since divergence from *E. coli*, which often contain virulence functions, are usually very AT rich compared to the rest of the genome (49). Genetic acquisitions within the *Salmonella* pangenome after the speciation of *Salmonella* from *E. coli* are enriched for AT-rich regions (Fig. S2). An H-NS-mediated regulatory mechanism controls the transcription of AT-rich regions (49). Hence, newly acquired AT regions might need to remain AT rich or become even more AT rich if they were to rely on this regulatory mechanism. Table 2 shows the maximum composite likelihood (MCL) estimate of the pattern of nucleotide substitution of three sets of genes; the concatenated alignment of 2,025 genes shared by all the *Salmonella* and *E. coli* strains (the core), the AT-rich SPI-1, and the similarly AT-rich SPI-4. The core has evolved from a lower GC to maintain a higher GC in *Salmonella*, as evidenced by a higher AT-to-GC substitution rate than GC-to-AT substitution rate (46.6% versus 39.1%). In contrast, the two AT-rich islands (SPI-1, SPI-4) have a substitution ratio that maintains them at their current level of AT richness (37.8% versus 47.2% for SPI-1 and 33.3% versus 49.2% for SPI-4), consistent with selective pressure to maintain regulation by the silencer of AT-rich regions, H-NS.

We investigated SPI-1 T3SS homologs in a wide variety of species. *Enterobacteriaceae*, such as *Escherichia*, *Shigella*, *Citrobacter*, and *Erwinia*, are closely related to *Salmonella*, and the T3SS homologs are similarly AT rich. However, in certain other high-GC organisms, such as *Pseudomonas*, *Burkholderia*, *Xanthomonas*, *Chromobacterium*, and *Bordetella*, xenologs for the T3SS have GC contents similar to that of their host genome (data not shown). Hence, even though H-NS homologs are widely distributed (61) and present in all the above-mentioned organisms, different groups of bacteria seem to tolerate foreign genes with distinct GC content via different mechanisms.

**Conclusion.** We predicted the evolutionary descent of each *Salmonella enterica* subspecies using three distinct approaches. All three methods were congruent in predicting the subspeciation topology. Highly incongruent individual gene trees suggested significant ancestral recombination during the evolution of the subspecies. We also identified lineage-specific gene clusters and genes undergoing accelerated evolution.

**TABLE 1** Statistically enriched gene sets containing a disproportionately high number of genes evolving faster than the rest of the pangenome

<table>
<thead>
<tr>
<th>Gene set</th>
<th>FDR q value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GC between 30 and 40</td>
<td>0</td>
</tr>
<tr>
<td>GC less than 30</td>
<td>0</td>
</tr>
<tr>
<td>GC above 65</td>
<td>0</td>
</tr>
<tr>
<td>TSSS effectors</td>
<td>0</td>
</tr>
<tr>
<td>Fimbrial operons</td>
<td>0</td>
</tr>
<tr>
<td>Type IV secretion and conjugative transfer</td>
<td>0</td>
</tr>
<tr>
<td>Type VI secretion systems</td>
<td>0.01</td>
</tr>
<tr>
<td>SPI-8</td>
<td>0.02</td>
</tr>
<tr>
<td>Spv locus</td>
<td>0.02</td>
</tr>
<tr>
<td>Toxin-antitoxin systems (other than RelBE and MazEF)</td>
<td>0.02</td>
</tr>
<tr>
<td>SPI-6</td>
<td>0.04</td>
</tr>
<tr>
<td>SPI-12</td>
<td>0.05</td>
</tr>
<tr>
<td>SPI-2</td>
<td>0.05</td>
</tr>
<tr>
<td>Srf-Fim</td>
<td>0.05</td>
</tr>
<tr>
<td>Tcf-Fim</td>
<td>0.06</td>
</tr>
<tr>
<td>Sfa-Fim</td>
<td>0.06</td>
</tr>
<tr>
<td><em>Salmonella</em> island STM0340-STM0348</td>
<td>0.1</td>
</tr>
<tr>
<td>Siderophore yersiniabactin biosynthesis</td>
<td>0.1</td>
</tr>
<tr>
<td>Sta-Fim</td>
<td>0.12</td>
</tr>
<tr>
<td>SPI-14</td>
<td>0.23</td>
</tr>
<tr>
<td><em>Salmonella</em> island STM0715-STM0728</td>
<td>0.24</td>
</tr>
</tbody>
</table>

**TABLE 2** AT-to-GC and GC-to-AT nucleotide substitutions in 29 *Salmonella* genomes, compared to the predicted state of the common ancestor

<table>
<thead>
<tr>
<th>Rate of change to:</th>
<th>AT</th>
<th>GC</th>
<th>% equilibrium GC</th>
<th>% observed GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>46.64</td>
<td>54.36</td>
<td>54.52</td>
<td></td>
</tr>
<tr>
<td>GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPI-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>37.77</td>
<td>44.47</td>
<td>44.65</td>
<td></td>
</tr>
<tr>
<td>GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPI-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>33.33</td>
<td>40.4</td>
<td>40.57</td>
<td></td>
</tr>
<tr>
<td>GC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Three sets are shown: (i) the concatenated core codon alignments, (ii) the AT-rich SPI-1, and (iii) the AT-rich SPI-4. Maximum composite likelihood estimates were compared for 2,025 single-copy genes, the AT-rich SPI-1 (300998 to 3046643 in LT2), and SPI-4 (4477857 to 450818 in LT2). The observed GC is the average GC content of all sequences compared. % equilibrium GC equals the rate of AT-to-GC substitutions divided by the sum of the rate of AT-to-GC substitutions and the rate of GC-to-AT substitutions. Gaps and missing data were excluded from the analysis.
clusters, iron acquisition systems, cytolethal distending toxins, and anaerobic respiratory reductases.

Genes recruited at the subspecies enterica (I) node may ultimately give clues to the adaptation of this lineage into warm-blooded animals. However, of 167 nonhomologous genes acquired at the subspecies enterica (I) node, 61 are hypothetical proteins and 93 have poorly defined putative functions. Functional characterization of these genes may reveal new mechanistic details of how subspecies enterica (I) adapted to warm-blooded hosts.

MATERIALS AND METHODS

Growth conditions and DNA isolation. Bacterial strains were grown under standard conditions, at 37°C in Luria broth, with aeration. Genomic DNA for sequencing was prepared from stationary cultures using the GenElute genomic DNA isolation kit for bacteria (Sigma) by closely following the manufacturer’s recommendations.

Genome sequencing and assembly. Genomes were sequenced and assembled at The Genome Institute (Washington University). Sequencing technology used and coverage obtained for each genome are shown in Supplementary Table S1. The sequencing reads for S. enterica serovar Paratyphi B BAA 1250, S. enterica subsp. diarizonae BAA639, and S. enterica serovar Enteritidis BAA1734 were assembled using the PCAP assembly program (62), while the rest of the genomes were assembled using the Newbler assembly program (454 Life Sciences, CT). All assemblies were run through a contamination screening process, which involves generation of %GC plots based on the assembly consensus, blasting the assembly consensus against the nucleotide database (NCBI) and ribosomal DNA database (63), identification and removal of contaminating sequences, and regeneration of new assembly output files after contamination removal. The finishing phase, which includes closure of gaps, resolution of ambiguous bases, and correction of misassembled regions, was completed for the assemblies of Salmonella enterica subsp. indica BAA1576, Salmonella enterica subsp. houtenae BAA1580, S. Enteritidis BAA1587, S. Paratyphi B BAA 1250, S. enterica subsp. diarizonae BAA639, and S. enterica subsp. arizonae BAA731. The remaining genome sequences were improved to high-quality drafts by automated and manual review and editing of the sequence data and orientation of contigs.

Genome annotation and pangenome analysis. For each genome, contigs from the de novo assemblies were ordered and concatenated into a pseudocotig, using the S. Typhimurium LT2 genome as the reference. Mummer (64) was used to align and order the contigs to the reference, and the alignment was parsed using custom Perl scripts. The ordered contigs were concatenated by inserting 50 N between contig breaks. These sequences were submitted to RAST (22) for automated annotation. Sieve (65) was used for de novo prediction of putative T3SS effectors. Putative effectors predicted using Sieve were also analyzed using BPBAac (66), T3MM (58), and Effective T3 (67). Genes either annotated as effectors by RAST based on homology to known effector families or predicted to be effectors by at least 2 of the 4 methods tested were considered to be putative T3SS effectors.

Orthologous and paralogous gene families were initially constructed based on RAST annotations. An “all-against-all” tblastx alignment was constructed using BLAT (68). The blat matrix was used as an input to construct orthologous/paralogous gene families using FastOrtho (http://enews.patricbrc.org/fastortho/), which is a faster reimplementation of OrthoMCL (34). A representative nucleotide sequence from each gene family was collated with strain-specific genes not present in any gene families. This set of representative sequences was again subjected to a tblastx analysis against the whole-genome nucleotide sequences to identify candidate genes that may have been missed by RAST. A gene and/or close paralog was considered “present” if it had at least 90% nucleotide sequence identity covering at least 20% of the sequence length. The phylectic table generated from the tblastx analysis was consolidated with the phylectic table generated from the OrthoMCL analysis to compute a comprehensive pangenome matrix. Rarefaction curves for pangenome size estimates were calculated based on the phylectic pattern, using EstimateS (version 8.2.0) (35).

A representative nucleotide sequence from each gene family was used for annotation augmentation using a variety of different methods. SignalP (69) was used to annotate sequences for presence of signal peptides, while TMHMM (70) was used to predict presence of transmembrane helices. Representative sequences from individual gene families were also annotated using KAAS (71) to overlay KEGG pathways onto the pangenome. Gene sets for metabolic pathways were constructed based on the KEGG and SEED annotations (22) from RAST. Other gene sets were constructed based on Virulence Factor Database (VFDB) annotations (72) and manual curation. Membership of each gene family within individual gene sets is listed in Table S7 in the supplemental material.

Gain/loss analysis. The pangenome matrix was used as an input for Count (36) to calculate posterior probabilities for gain and loss of each gene family across all nodes of the phylogenetic hypothesis. The posterior probability matrix was converted into a binary matrix using a threshold of 0.75. This binary matrix was used to estimate overabundance of gain or loss of gene families within specific gene sets across major ancestral nodes using Gotools (37). Within Gotools, the Fisher exact test was used to estimate the significance of enrichment of specific gene sets, and false-discovery rates (FDR) were used to adjust the P values for multiple comparisons.

Any gene family gained more than once across the phylogenetic tree was marked as a homoplastic gene family. This binary information was used as an input with Gotools (37) to assess enrichment of homoplastic genes within specific gene sets.

Evolutionary rate of gene families. Individual gene sequences (DNA and amino acid sequences) for all genes in gene families predicted by FastOrtho were extracted using custom Perl scripts. Codon alignments within gene families were constructed using PAL2NAL (73). Amino acid alignments needed as input for PAL2NAL were constructed using Muscle (24). The dS/dN ratios for all possible pairwise comparisons within a gene family were calculated based on the codon alignments using SNAP (74). Mean dS/dN ratios were assigned for individual gene families by averaging all pairwise ratios within each family. The pangenome was ranked based on log10 dS/dN ratios, and enrichment analysis for gene sets containing relatively fast-evolving gene families was conducted using gene set enrichment analysis (GSEA) (75). Within GSEA, false-discovery rates were used to adjust P values for multiple comparisons.

Phylogenetic analysis. Whole-genome alignments for all 30 strains were constructed using Mauve (version 2.3.1) (20). The SNP matrix was parsed to remove positions with gaps and unknown characters using custom Perl scripts. This matrix was used to infer an ML tree using RAxML (version 7.2.6) (21). The DNA sequences within each family were aligned using Muscle (24), and the alignments were used as an input to construct individual gene trees using RAxML (version 7.2.6) (21). As a distinct alternative, a majority rule consensus tree of all gene family trees that were conserved across all genomes in exactly one copy (2,025 gene families) was constructed using Phylip (version 3.69) (25). A concatenated codon alignment was also constructed for the above-mentioned 2,025 genes. All synonymous positions from the concatenated codon alignment were extracted using DnaSP (version 5/10/01) (76). An ML tree was constructed from the synonymous SNP matrix using RAxML (version 7.2.6) (21). Divergence times for the subspecies were estimated by calibrating the tree based on a divergence time of 140 million years for E. coli and Salmonella, as estimated by Ochman et al. (23) using MEGA (version 5.1) (77).

Recombination analysis. Orthologous regions that were conserved across all genomes and at least 10 kb in length were extracted from the Mauve-based multiple-genome alignments. This resulted in extraction of eight fragments totaling 104 kb. This alignment was used as input for Clonal Frame (version 1.2) (27). Two independent runs, each consisting of 200,000 MCMC iterations, in which the first 100,000 iterations were discarded as burn-in, were performed. The runs were compared for con-
vergence using the Gelman and Rubin statistic (78), which was found to be satisfactory.

As a second alternative, pairwise alignments of S. enterica serovar Typhi strain CT18 and all strains of non-subspecies enterica genomes, including S. bongori, were constructed using Mauve (version 2.3.1) (20) with S. Typhimurium LT2 as a reference. The pairwise alignments were parsed to show orthologous bases for all positions in the S. Typhimurium LT2 genome using custom Perl scripts. Regions in any strain where an orthologous region in S. Typhimurium LT2 was absent were discarded. This pseudo-multigenome alignment was used as input for the Recombination Detection Program (RDP) (version 4.16) (30) to detect recombination tracts in S. Typhimurium LT2. Nine different algorithms, RDP, Geneconv (79), Bootscan (80), Maxchi (81), chimaera (82), SiScan (83), PhyloPro (84), LARD (85), and 3Seq (86), were used to determine recombination breakpoints across the alignment. Recombination events detected within the S. Typhimurium LT2 genome by at least 2 methods were annotated as possible recombination events.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00579-12/-/DCSupplemental.

Table S1, PDF file, 0 MB.
Table S2, PDF file, 0 MB.
Table S1, XLSX file, 0.1 MB.
Table S2, XLSX file, 12 MB.
Table S3, XLSX file, 1.9 MB.
Table S4, XLSX file, 16.3 MB.
Table S5, XLSX file, 0.4 MB.
Table S6, XLSX file, 0.1 MB.
Table S7, XLSX file, 0.2 MB.

ACKNOWLEDGMENTS

This work was supported in part by NIH grants AI039557, AI052237, AI073971, AI075093, AI077645, and AI083646, USDA grants 2009-03579 and 2011-67017-30127, and the Binational Agricultural Research and Development Fund.

We thank Ken Sanderson for supplying strains used in this study and Patti Fields and Monica Ponder for advice on strain selection.

REFERENCES


Desai et al.


Evolutionary Genomics of *Salmonella enterica* Subspecies

Prerak T. Desai,a Steffen Porwollik,a,b Fred Long,a Pui Cheng,a Aye Wollam,c Veena Bhonagiri-Palsikar,c Kymberlie Hallsworth-Pepin,c Sandra W. Clifton,c George M. Weinstock,c Michael McClelland,a,b

Department of Pathology and Laboratory Medicine, University of California, Irvine, Irvine, California, USAa; Vaccine Research Institute of San Diego, San Diego, California, USAb; The Genome Institute, Washington University School of Medicine, St. Louis, Missouri, USAc

Volume 4, no. 2, doi:10.1128/mBio.00579-12, 2013. Two authors were not listed. The byline should appear as shown above.