Convergent Molecular Evolution of Genomic Cores in Salmonella enterica and Escherichia coli

Sujay Chattopadhyay, Sandip Paul, Dagmara I. Kisiela, Elena V. Linardopoulou and Evgeni V. Sokurenko

J. Bacteriol. 2012, 194(18):5002. DOI: 10.1128/JB.00552-12.
Published Ahead of Print 13 July 2012.

JB
Journals.ASM.org

Convergent Molecular Evolution of Genomic Cores in *Salmonella enterica* and *Escherichia coli*

Sujay Chattopadhyay, Sandip Paul, Dagmara I. Kisiela, Elena V. Linardopoulou, and Evgeni V. Sokurenko

Department of Microbiology, University of Washington, Seattle, Washington, USA

One of the strongest signals of adaptive molecular evolution of proteins is the occurrence of convergent hot spot mutations: repeated changes in the same amino acid positions. We performed a comparative genome-wide analysis of mutation-driven evolution of core (omnipresent) genes in 17 strains of *Salmonella enterica* subspecies I and 22 strains of *Escherichia coli*. More than 20% of core genes in both *Salmonella* and *E. coli* accumulated hot spot mutations, with a predominance of identical changes having recent evolutionary origin. There is a significant overlap in the functional categories of the adaptively evolving genes in both species, although mostly via separate molecular mechanisms. As a strong evidence of the link between adaptive mutations and virulence in *Salmonella*, two human-restricted serovars, Typhi and Paratyphi A, shared the highest number of genes with serovar-specific hot spot mutations. Many of the core genes affected by Typhi/Paratyphi A-specific mutations have known virulence functions. For each species, a list of nonrecombinant core genes (and the hot spot mutations therein) under positive selection is provided.

Bacterial genome evolution is shaped via several mechanisms, such as the acquisition of novel genes by horizontal transfer, homologous recombination, gene deletion, gene amplification, and point mutation. A complex network of these evolutionary mechanisms drives adaptive divergence of microorganisms, especially of bacterial pathogens, in the course of niche differentiation. For instance, uropathogenic *Escherichia coli* behave as commensals in the human intestine but as pathogens in the urinary bladder. Zoonotic pathogens also follow similar evolutionary dynamics, commonly behaving as commensals in animal hosts but as pathogens in humans. The primary focus to understand evolution of bacterial virulence had been on the genes acquired by horizontal gene transfer (in particular, on the acquisition of pathogenicity islands) rather than on the genes shared by pathogenic and nonpathogenic strains. Only recently, pathogenicity-adaptive (pathoadaptive) mutations that result in specific functional changes of genes were also identified as an important mechanism of virulence evolution. However, very limited genome-level analysis has thus far been performed to investigate the role of point mutations in common (core) genes for their potential contribution in the differential adaptation of pathogenic bacterial species.

In the present study, we focused on the adaptive microevolution of *Salmonella enterica* subspecies *enterica* (i.e., subspecies I), one of the most important human pathogens. Unlike *E. coli*, *Salmonella* does not generally circulate as a commensal among humans, and most serovars are transmitted by ingestion of contaminated food or by contact with infected animals. The majority of *Salmonella* infections are self-limited gastroenteritis caused by broad host-range serovars such as Enteriditis or Typhimurium. These serovars are frequently associated with disease in a wide variety of animals. In contrast, a small number of serovars infect only specific animal-hosts and do not persist or cause disease in other vertebrate species. Such narrow-host-range (host-adapted or host-restricted) *Salmonella* typically produces severe invasive infections. For example, human-restricted *Salmonella* (e.g., Typhi and Paratyphi A and C) causes typhoid fever, and avian-restricted *S.* Gallinarum cause fowl typhoid; both diseases are characterized by septicemia over intestinal symptoms.

Our earlier studies reported that virulence habitats impose positive selection on bacteria for point mutations in their core genes to increase pathogenicity (7, 40, 51). This was demonstrated by specifically designed microevolutionary analysis approach, termed zonal phylogeny (ZP), that identifies the occurrence of repeated phylogenetically unlinked mutations at same amino acid positions. The presence of these so-called structural hot spot mutations is considered a very strong evidence of positive selection in particular habitats, indicating repeated selection in different strains that compete for survival in the same or similar environments (20, 34). Using this approach for *E. coli* genomes, we detected the presence of positive selection in a large fraction of core genes that, in part, was pathotype specific (8).

Here, we performed a comparative genome-wide analysis of strains from over a dozen of pathogenic *Salmonella* serovars to evaluate the extent to which mutations in the core genes could contribute to the pathoadaptive evolution. For comparison, we analyzed in parallel genomes from 22 strains of *E. coli*, the bacterial species known to evolve extensively via pathoadaptive mutations.

MATERIALS AND METHODS

Selection of strains. For each species, out of all fully assembled genomes available at the time, the selected strains included those that are clonally distinct, i.e., having different alleles of housekeeping genes used for multilocus sequence typing (MLST) analysis (http://mlst.ucc.ie) (Fig. 1). The only exceptions were two pairs of clonally identical strains, one for *Salmonella* (serovar Paratyphi A strains ATCC 9150 and AKU_12601) and one for *E. coli* (extraintestinal pathogenic *E. coli* [ExPEC] strains S88 and UTI89).

Received 3 April 2012 Accepted 6 July 2012

Published ahead of print 13 July 2012

Address correspondence to Evgeni V. Sokurenko, evs@u.washington.edu, or Sujay Chattopadhyay, sujayc@u.washington.edu.

Supplemental material for this article may be found at http://jb.asm.org/.

doi:10.1128/JB.00552-12

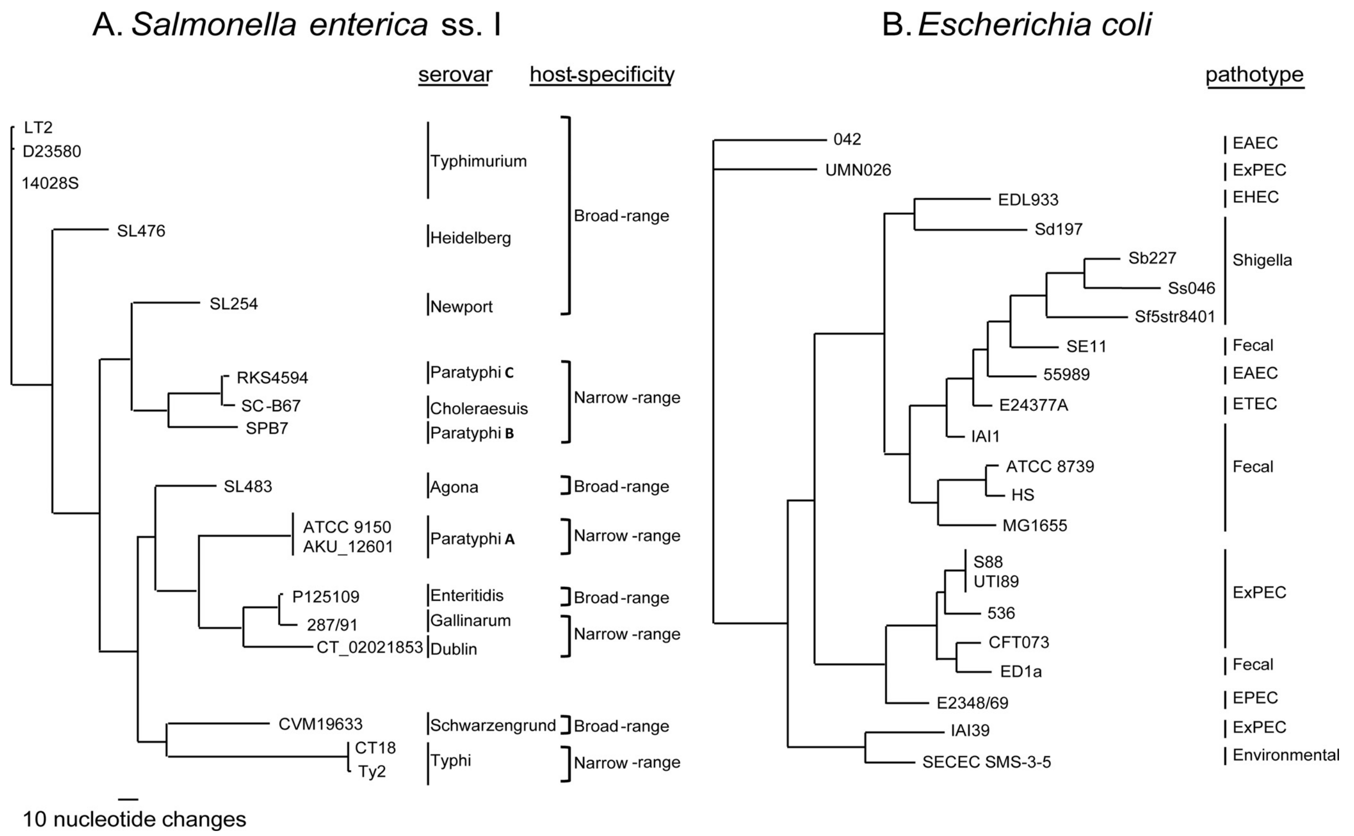


FIG 1 Phylograms of concatenated sequences of seven housekeeping genes of *S. enterica* subsp. I (A) and *E. coli*/*Shigella* (B). The housekeeping genes used in MLST were selected for analysis. The *Salmonella* genes were *aroC* (chorismate synthase), *dnaN* (DNA polymerase III, β-subunit), *hemD* (uroporphyrinogen III synthase), *hisD* (histidinal dehydrogenase), *purE* (phosphoribosylaminoimidazole carboxylase), *sucA* (2-oxoglutarate dehydrogenase decarboxylase), and *thrA* (aspartokinase I). The *E. coli* genes were *adk* (adenylate kinase), *fumC* (fumarate hydratase class II), *gyrB* (DNA gyrase subunit B), *icd* (isocitrate dehydrogenase), *mdh* (malate dehydrogenase), *purA* (adenylosuccinate synthetase), and *recA* (recombinase A). Serovars and host specificity ranges were shown for the *Salmonella* strains, while pathotypes were mentioned for the *E. coli* strains. EAEC, ExPEC, EHEC, ETEC, and EPEC stand for enteroaggregative, extraintestinal pathogenic, enterohemorrhagic, enterotoxigenic, and enteropathogenic *E. coli*, respectively.

Phylogenetic analysis and extraction of core genes. To assess the evolutionary relationship of strains in each species, the corresponding set of seven housekeeping gene sequences were concatenated for each strain and used for reconstructing a maximum-likelihood (ML) phylogenetic tree, using the general time reversible (GTR) substitution model with estimated base frequencies site-specific by codon position distribution implemented in PAUP* (45) to depict an average neutral diversity among the strains. The pairwise p-distance matrix was calculated using MEGA (46).

During the extraction of core genes using annotations of one reference strain in each species, we excluded all annotated pseudogenes in reference genome. The BLAST hits with either internal stop codon or non-ACGT characters were also excluded. To construct each gene data set, we performed stand-alone BLAST against the rest of the genomes and extracted homologs with nucleotide sequence identity as well as sequence length coverage values of ≥95% relative to the reference sequence. For each of the extracted core genes, BLAST was repeated against all genomes (including the reference) using the same threshold values for sequence identity and coverage, and any gene with multiple copies in any genome was excluded as probable nonorthologs from further analysis.

Molecular evolutionary analysis. Rates of nonsynonymous (*dN*) and synonymous (*dS*) mutations were computed by using mutation-fraction method (33). To assess any significant difference between *dN* and *dS*, we approached nonparametric bootstrap procedure (14) by simulating 1,000 data sets by sampling with replacement DNA sites of the multiple sequence alignment, thereby disrupting the codon structure of real data set. This led us to calculate the distribution of *dN*/*dS* under the null hypothesis of neutrality (*dN*/*dS* = 1). The rank of the observed *dN*/*dS* in the distribution of simulated data set values suggested any presence of positive (i.e., a higher observed value for >950 cases), negative (i.e., a higher observed value for <50 cases), or neutral selection at a 95% significance level.

For the analysis of core gene data sets, zonal phylogeny software (ZPS) (6) was used to perform CLUSTAL W-based sequence alignment, followed by PAUP*-based, ML tree topology reconstruction for the genome-wide gene data sets in batch mode, using the GTR substitution model with estimated base frequencies site specific by codon position distribution. ZPS reconstructs an unrooted protein phylogram from the corresponding DNA phylogram, distinguishing two categories of protein variants: those encoded by multiple alleles with synonymous diversity (i.e., evolutionarily long-term variants) and those encoded by one allele only (i.e., evolutionarily recent variants). Separate scripts were written to calculate the frequencies of parallel and coincidental hot spot mutations of different nature (long-term or recent hot spots), the frequency and sharing of hot spot mutations within or between pathotypes or serovars.

We detected probable recombination events using software package PhiPack (1) that included three recombination-detection statistics: pairwise homoplasy index (Phi), maximum χ^2 (MaxChi), and neighbor similarity score (NSS). A gene was considered to be recombinant if *P* values for all of the three statistics were <0.1 (4).

Simulations. EvolveAGene3 (18) was used to perform random simulations of mutations on nonrecombinant core genes with hot spots. For each gene, the reference genome sequence was considered as the root sequence to simulate the data set and to generate a random tree topology where each branch had equal probability to lead to either a terminal node or an internal node. For simulations, average branch lengths and average selection on amino acid replacements (i.e., *dN*/*dS*) were estimated from corresponding real data set phylogeny. No indels were allowed in the simulated data sets. Selection over sequence, as well as over branches along the tree, was set to be constant with the default modifier value of 1.

Analysis of functional category enrichment. Functional annotation clustering for different sets of candidate genes under positive selection was performed using DAVID (10). For the analysis, a "medium" classification stringency was used. Annotation clusters with an enrichment score of >0.5 and a *P* value of <0.05 were selected as enriched functional categories.

Detection of interspecies orthologs. Orthologous protein-coding genes between *Salmonella* and *E. coli* were detected by using the Roundup Orthology Database (http://roundup.hms.harvard.edu/ [9]), in which the orthologs were computed using reciprocal smallest distance algorithm (50). A "retrieve" option of this database was used with stringent BLAST E-value of 1e-20 and a global pairwise sequence divergence of 0.2 to recover a set of orthologs between *Escherichia coli* CFT073 and *Salmonella enterica* serovar Typhi Ty2 genomes. From this set of orthologs, genes accumulating recent hot spot mutations in both species were identified.

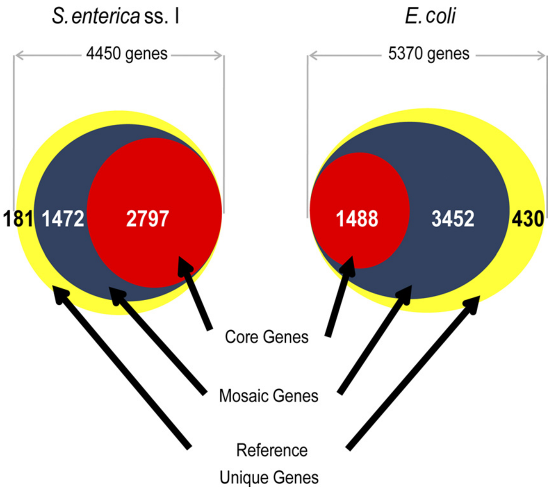


FIG 2 Venn diagrams showing unique (to reference genome), mosaic, and core genes in *S. enterica* subsp. I and *E. coli*.

RESULTS AND DISCUSSION

Decreased nucleotide diversity but an increased number of core genes in *Salmonella* relative to *E. coli*. *Salmonella* and *E. coli* strains analyzed here were phylogenetically diverse based on concatenated sequence of seven housekeeping genes: for *Salmonella*, *aroC*, *dnaN*, *hemD*, *hisD*, *purE*, *sucA*, and *thrA* loci (total, 8,985 bp), and for *E. coli*, *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA* (total, 9,093 bp). These genes are present in all strains of the corresponding species and internal fragments of them are used for their MLST analysis (http://mlst.ucc.ie). The *Salmonella* set had 16 clonally distinct strains each represented by a unique MLST sequence type, with no specific clustering of serovars based on the host specificity range (Fig. 1A). The *E. coli* was represented by 21 clonally distinct strains and, although some *E. coli* strains of same origin (e.g., ExPEC, *Shigella* or commensal) clustered in a somewhat distinct fashion, the clades were not fully pathotype specific (Fig. 1B). The pairwise nucleotide diversity of the *Salmonella* genes was 1.0% ± 0.1%, whereas the *E. coli* genes diversity was 1.5% ± 0.1%.

We next extracted protein-coding genes shared by all strains within each species. As a reference for the BLAST analysis, we selected for each species a genome with well-annotated assembly and with well-preserved backbone gene-set (i.e., without too many genes deleted). Since we performed a phylogenetic analysis of each gene, it is not important to select either a commensal or a laboratory strain as a reference. Genomes of *S.* Typhimurium strain LT2 (a total of 4,451 open reading frames [ORFs]) and ExPEC strain CFT073 (5,378 ORFs), respectively, were used. Threshold values of ≥95% for both nucleotide sequence identity and length coverage were used to extract the orthologous gene copies. A total of 2,797 genes were found in all strains of *Salmonella* and were defined as the species core genes (Fig. 2). Among the rest, 1,472 genes were defined as mosaic genes since they were shared by only some strains in the species, and 181 genes were uniquely found only in the reference strain LT2 (Fig. 2). In *E. coli*, 1,488 genes were identified as species core, 3,452 genes were of mosaic nature, and 430 were unique to the reference strain CFT073 (Fig. 2).

Among the core genes, 15 *Salmonella* and 5 *E. coli* genes were found in multiple copies in various positions in the chromosome in at least one strain (see Table S1 in the supplemental material) and were excluded from the analysis as potentially nonorthologous. The rest were found in one copy and in syntenic position (sharing at least one neighboring locus) in every genome of the corresponding species and were considered to be of both core and orthologous nature for each species.

The almost 2-fold greater number of core genes in *Salmonella* was in contrast to average size of genomes analyzed, which was somewhat lower for the *Salmonella* strains (4.78 Mb on average, ranging from 4.58 to 4.89 Mb) than *E. coli* (4.94 Mb on average, ranging from 4.37 to 5.53 Mb). One possible explanation for the larger genomic core in *Salmonella* strains is their closer evolutionary relatedness. Indeed, subspecies I of *Salmonella* is considered to have evolved from the rest of the species by the acquisition of horizontally transferred genomic elements, such as *Salmonella* pathogenicity islands SPI-3 (17 kb), SPI-4 (25 kb), and SPI-5 (9 kb) (15). Thus, parts of the genomes that are overall of a mosaic nature in *Salmonella* species are likely to be shared by all of the subspecies I strains and are defined here as core genes. This explanation is consistent with the fact that frequency of mosaic genes (defined here as shared by some strains only) was significantly lower in the *Salmonella* strains (33%) than *E. coli* strains (64%). Only 20 of the core genes in *Salmonella* and 7 genes in *E. coli* were identical in all strains of the corresponding species (see Table S2 in the supplemental material). In other core genes, the average nucleotide diversity values were estimated to be 1.0% ± 0.01% for *Salmonella* and 1.6% ± 0.02% for *E. coli* (Table 1, $P < 0.0001$), i.e., very similar to the diversity of corresponding seven MLST genes above.

The lower diversity of the *Salmonella* genes is also likely to reflect the fact that all of the analyzed strains belonged to just one major phylogenetic clade within the species: *S. enterica* subsp. *enterica* or I. The subspecies I strains are considered to be adapted to warm-blooded animals and cause the majority of *Salmonella*-related diseases in humans, whereas strains from the other subspecies II, IIIa, IIIb, IV, and VI are adapted to cold-blooded animals and cause few infections in humans. In contrast, the *E. coli* strains analyzed represented all major phylogenetic clades of the species that are divided into the so-called *E. coli* reference (ECOR) phy-

TABLE 1 Comparative analysis of diversity and hot spot mutations in core genes between *S. enterica* subspecies I and *E. coli*

Parameter	*S. enterica*	*E. coli*
No. of core orthologs		
Total	2,782	1,483
No. identical	20	7
No. with ≥4 alleles	2,646	1,404
Nucleotide diversity (mean no. ± SD)		
MLST genes, π	0.01 ± 0.001	0.015 ± 0.0001
Core genes, π	0.01 ± 0.0001	0.016 ± 0.0001
Core genes, *dN*	0.003 ± 0.0001	0.004 ± 0.0001
Core genes, *dS*	0.034 ± 0.0003	0.055 ± 0.0008
Total no. of genes with hot spot mutations of different nature		
Parallel only	577	323
Coincidental only	97	85
Mixed	175	188
No. of nonrecombinant genes with hot spot mutations		
Parallel only	447	166
Coincidental only	87	61
Mixed	128	90
No. of simulated data set genes with hot spot mutations		
Parallel only	3	8
Coincidental only	30	40
Mixed	5	10
Evolutionary time origin of genes with hot spot mutations (%)		
Long-term only	14.6	10.7
Recent only	73.2	69.9
Mixed	12.2	19.4

logenetic groups A, B1, B2, and D and a diverse number of pathotypes.

To test whether there could be indeed a direct correlation between the reduced diversity and increased number of core genes (as observed in *Salmonella*), we examined the diversity of core genes in a subgroup of six *E. coli* strains that form phylogenetic group B2 (Fig. 1B). The average pairwise nucleotide diversity of the B2 strains was 0.6% ± 0.1%, based on the MLST genes, i.e., lower than the MLST diversity of either *E. coli* or *Salmonella* strains. At the same time, in the group B2 strains, a total of 3,447 core genes were identified (see Fig. S1 in the supplemental material); this was significantly higher than the number of core genes found in *E. coli* or even in *Salmonella* species. Thus, resemblance in the pattern of core gene diversity between *Salmonella* and B2 *E. coli* strains supports the hypothesis that a relatively large core genome in *Salmonella* subspecies could be due to the relatively close relatedness of the strains.

The variability of core genes in both *Salmonella* and *E. coli* is primarily due to synonymous mutations rather than to nonsynonymous changes. The nucleotide diversity of coding genes is both nonsynonymous (amino acid replacement) and synonymous (structurally silent) in nature. The number of nonsynonymous or synonymous changes per nonsynonymous or synonymous site, respectively, defines the nonsynonymous mutation rate (*dN*) or synonymous mutation rate (*dS*) in a gene. For *Salmonella* core genes, the average *dN* (0.3% ± 0.01%) and *dS* (3.4% ± 0.03%) were lower than the corresponding values for *E. coli* ($dN =$ 0.4% ± 0.01% and $dS =$ 5.5% ±0.08%) (Table 1). However, because *dS* was >10-fold higher than *dN* for both species and the differences in *dS* between the species (calculated as $dS_{E.\ coli} - dS_{S.\ enterica}/dS_{S.\ enterica}$) were more pronounced than in *dN* (i.e., $dN_{E.\ coli} - dN_{S.\ enterica}/dN_{S.\ enterica}$), the synonymous mutations contributed significantly more ($P < 0.01$) than the amino acid replacement mutations into the differences in the overall nucleotide diversity between *Salmonella* and *E. coli*.

As defined above, across all shared genes, the average rate of nonsynonymous mutations is much lower than the rate of synonymous changes, indicating a prevalence of negative selection against amino acid changes. When the *dN* and *dS* values of each core gene were compared by using the z-test (at $P < 0.05$), *dN*/*dS* was significantly less than 1 in the vast majority of individual genes, reflecting the overall predominance of synonymous changes. We found only 11 genes in *Salmonella* (0.4% of the total core genes) that showed *dN* values significantly higher than the *dS* values ($dN/dS \gg 1$), i.e., well within the range of the gene fraction randomly expected at a $P < 0.05$ significance level. Although the number of genes with $dN/dS \gg 1$ was 7-fold higher in *E. coli* (77 genes, $P < 0.0001$), the overall fraction (5.2%) was found to barely exceed the randomly expected range. Moreover, when a more rigorous nonparametric bootstrap analysis was applied to both sets of the genes with $dN/dS \gg 1$, the numbers dropped significantly, with only 1 *Salmonella* and 24 *E. coli* genes showing a *dN* significantly higher than the *dS* (see Table S3 in the supplemental material).

Thus, according to the *dN*/*dS* test, very few of the core genes in *Salmonella* and *E. coli* have accumulated replacement mutations under positive selection. One needs to note, however, that the *dN*/*dS* test is considered to be highly conservative and mostly detects an action of positive selection over a relatively long period of time (e.g., between species) but not on a microevolutionary (within-species) scale.

High frequency of amino acid hot spot mutations in the core genomes of *Salmonella* and *E. coli*. We next examined the action of positive selection on amino acid changes in core genes by the detection of convergent mutations, i.e., repeated, phylogenetically unlinked nucleotide changes that result in mutation of the same amino acid positions. These so-called hot spot mutations can be of two types: (i) "parallel mutations" that lead to a same amino acid change in the same position and (ii) "coincidental mutations" that lead to a different amino acid change in the same position. The convergent evolution (evolutionarily independent acquisition of the same biological trait) is a very strong indicator of adaptive nature of the associated phenotypic changes, either at the organismic or at the molecular level (20, 34).

Detection of hot spot mutations is based on sequence phylogeny and not via a mere mismatch with the reference genome sequences. Since this approach requires the reconstruction of phylogenetic trees, it can be done only in genes that have four or more alleles. A total of 2,646 of the *Salmonella* genes and 1,404 of the *E. coli* genes met such a criterion. A total of 849 *Salmonella* and 596 *E. coli* core genes accumulated hot spot changes. In both species, the rate of parallel hot spot changes was significantly higher than the rate of coincidental mutations (Table 1).

The parallel hot spot mutations could not arise due to a horizontal gene transfer of entire gene, followed by further allelic diversification, because they would be phylogenetically linked and, thus, not show up as a hot spot change in our analysis. However, gene changes that appear as parallel hot spot changes could be due to homologous intragenic recombination, i.e., exchange of small internal regions in orthologous genes. So, to remove alleles with such potentially false (nonmutational) hot spot changes, recombination detection analysis was performed for each gene with the hot spot-like mutation. Recombination was detected in 187 (22%) *Salmonella* genes and 279 (47%) *E. coli* genes with apparent hot spots, indicating significantly ($P < 0.0001$) higher rate of homologous recombination in the latter. The increased recombination rate in *E. coli* has been reported in previous studies (12, 17, 47, 49, 52, 54). Genes affected by recombination (even if detected just in one of the alleles) were removed from the analysis, leaving 662 *Salmonella* genes (24% of the total core genes) and 317 *E. coli* (21%) where all hot spot changes were likely to be truly mutational in nature (see Tables S4 and S5 in the supplemental material).

Despite the removal of alleles with potentially false parallel mutations, the frequency of genes with parallel mutations remained significantly higher than those with coincidental mutations (Table 1). In contrast, if hot spot mutations accumulate randomly, the probability of a same mutation in the same position (i.e., parallel hot spot mutations) would be significantly lower than that of a different mutation in the same position (i.e., coincidental hot spot mutations). Therefore, the predominance of parallel hot spot changes cannot be expected from chance alone. To confirm this, we performed 10 rounds of random simulations of mutational changes in each of the nonrecombinant genes. As expected, in simulated data sets the coincidental mutation frequency was much higher than the parallel mutation frequency (Table 1). However, even the simulated coincidental mutation frequency was well below ($P < 0.0001$) the one in real data sets of both *Salmonella* and *E. coli* (Table 1). Taken together, the simulation analysis strongly suggests that the hot spot mutations in both *Salmonella* and *E. coli* were acquired not randomly but under positive selection, i.e., the changes were adaptive in nature.

Also supportive of the function of positive selection is the fact that genes acquiring hot spot mutations tend to do so multiple times; there were, on average, 5.0 ± 0.2 hot spot mutations in 2.1 ± 0.1 positions per gene in *Salmonella* and 6.1 ± 0.4 hot spot mutations in 2.5 ± 0.1 positions per gene in *E. coli*.

Finally, one of the distinctive characteristics of functionally adaptive hot spot mutations in bacterial genes is their evolutionarily recent nature, reflecting a recent emergence of the pathogenic lineages and/or their relative evolutionary instability. The latter could be due to the fitness trade-off functional mutations in so-called "source-sink" dynamics of the pathogen microevolution (41). The "source" is defined as an evolutionarily stable, reservoir habitat (usually of environmental or of commensal nature), while pathogen-targeted compartments are evolutionary "sinks" representing unstable, transient habitats. The sink habitats, therefore, are compartments where bacteria are introduced occasionally and then experience selection pressures to improve their ability as pathogen. To define whether or not the hot spot mutations were recent from evolutionary perspectives, we determined whether or not, respectively, their acquisition was followed by accumulation of synonymous changes in the corresponding alleles. This analysis was done by using the zonal phylogeny software tool (6, 40). In both *Salmonella* and *E. coli*, the vast majority of hot spot mutations (in 85.4% and 89.3% of genes with hot spots, respectively) were of evolutionarily recent origin (Table 1), i.e., the pattern typical for pathoadaptive mutations in bacterial pathogens.

Taken together, these results indicate that the hot spot mutations in both *Salmonella* and *E. coli* were acquired at a significantly higher rate than expected under neutrality and with patterns strongly suggestive of the action of positive selection.

***Salmonella* and *E. coli* share functional trajectories of adaptive evolution by hot spot mutations.** Overall, ca. 44% *Salmonella* and 48% *E. coli* genes with recent hot spot mutations were found to have a defined function, while the rest were annotated only with putative functions. We performed functional annotation-based clustering analysis using DAVID (10) to identify what functional categories of the adaptively evolving proteins are significantly enriched (or overrepresented), i.e., are targeted more frequently than expected randomly.

In *Salmonella*, the hot spot-affected genes belonged to a total of 78 functional categories, 18 of which were significantly enriched (Fig. 3, left). One major enriched category was secretion system apparatus proteins encoded by genes in two pathogenicity islands: *invG* and *prgI* in SPI-1 and *ssaC*, *ssaJ*, *ssaQ*, and *ssaT* in SPI-2 (19, 28). Some other notable genes demonstrated to be necessary for the *Salmonella* virulence included *pagC* (encoding two-component protein for survival in the macrophage [31]), *purG* (encoding a purine biosynthetic enzyme [30]), *eutG* (encoding an enzyme for carbon source degradation [43]), *nfo* (endonuclease IV counteracting DNA damage by host nitric oxide [35, 44]), and *rfaK* (encoding lipopolysaccharide biosynthesis enzyme for colonization on chicken alimentary tract [48]), etc. Several other genes were earlier shown to be associated with virulence in other bacterial pathogens, such as *ubiF* in *E. coli* (involved in ubiquinone and other terpenoid-quinone biosynthesis [27]), *kdpA* in *Mycobacterium tuberculosis* (two-component transcriptional regulator gene [36]), *pyrC* in *Bacillus anthracis* (encoding pyrimidine metabolism protein critical for growth [38]), and *hisB* in *Alternaria citri* (involved in amino acid biosynthesis [23]), etc.

In *E. coli*, of 48 functional categories represented by the hot spot genes, 13 were found to be significantly enriched or overrepresented (Fig. 3B). Most of the enriched categories (and corresponding genes) were same as reported by us previously (8). As in *Salmonella*, they mostly included metabolic genes. Interestingly, nine enriched clusters were common to both *Salmonella* and *E. coli* species (Fig. 3, cross-linked), showing overlap that was significantly higher than would be randomly expected ($P < 0.001$). This suggests that both species evolve via similar physiologic pathways. This is not surprising considering that both bacteria share close common ancestry and occupy similar habitats. However, while the enriched functional categories were very similar, the genes affected were mostly different. The common enriched clusters included 113 genes in *Salmonella* and 74 genes in *E. coli*. Although most of these genes in one species had orthologous copies in the other one (68 of the *Salmonella* genes and 56 of the *E. coli* genes), only 10 of the common orthologous genes were affected by hot spot mutations in both species (see Table S6 in the supplemental material). Still, in four of them—*murF* (UDP-*N*-acetylmuramoyl-tripeptide--D-alanyl-D-alanine ligase), *trmA* [tRNA (uracil-5-)-methyltransferase], *creB* (transcriptional regulatory protein), and *holB* (DNA polymerase III, δ′ subunit)—mutations affected the same amino acid positions. Thus, at least in some

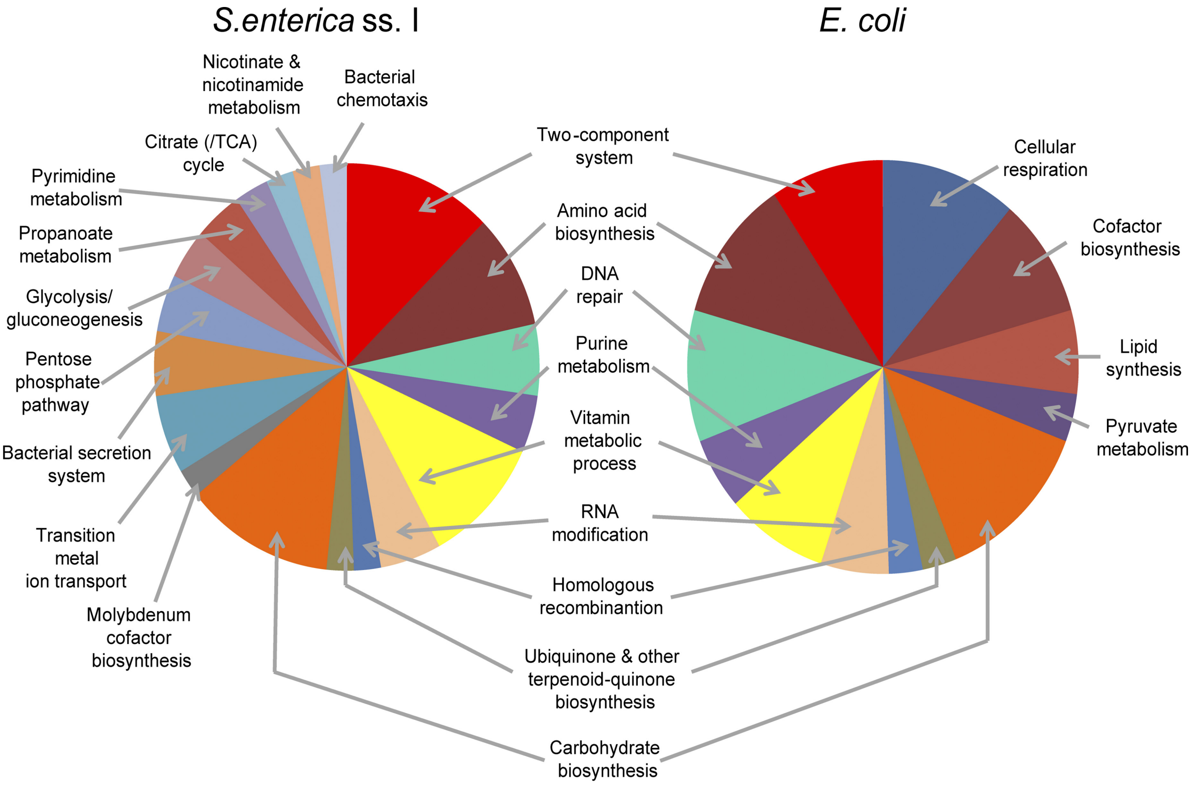


FIG 3 Pie charts showing significantly ($P < 0.05$) overrepresented functional categories of candidate proteins with recent hot spot (adaptive) mutations in *S. enterica* subsp. I and *E. coli*. Common overrepresented functional categories are mentioned in the center.

cases, there is a convergent evolution in the two species not only on the level of common functional categories but also at the molecular level.

A recent work suggests that synonymous diversity fluctuates by >20-fold among genes across different regions in *E. coli* genome (29). Such diversity is contributed to by heterogeneity in the neutral mutation rate, leading to distinct genetic regions of mutational "hot spots" and mutational "cold spots." We compared the enriched functional clusters encoded by genes in hot spots or cold spots and those encoded by our set of nonrecombinant *E. coli* genes with structural hot spot mutations. In either the hot spot or cold spot gene regions, the overlap was in one gene category each (amino acid biosynthesis and cellular respiration, respectively). This indicates that the accumulation of structural hot spot mutations does not correlate with the synonymous diversity of the genes. Thus, they can primarily be attributed to selective factors and not to any underlying behavior of DNA sequences such as heterogeneity due to a neutral mutation rate. However, we cannot exclude completely that the occurrence of at least some hot spot mutations is due to a relaxed purifying selection or an increased mutation rate at specific nucleotide positions.

Pathotype-specific convergent evolution of *Salmonella* Typhi and Paratyphi A. We next compared the genes with recent hot spot mutations in different pathotypes of each species. Within two main pathotypes of the *Salmonella* serovars—broad host range and narrow host range (Fig. 1A)—the average numbers of the recent hot spot genes were not significantly different (Fig. 4). Similarly, we found no significant difference between the two major pathotypes of *E. coli*—ExPEC and *Shigella*—that were represented by multiple genomes in the databases. However, both *E. coli* pathotypes had more hot spot genes than did *E. coli* fecal strains (Fig. 4). This is in full accordance with our previous study on a smaller number of *E. coli* genomes, showing that pathogenic lineages accumulate larger number of adaptive mutations than nonclinical strains (8). Unlike *E. coli*, however, all *Salmonella* strains were of clinical origin, and no comparison with nonclinical strains could be made.

The significant difference in the frequency of hot spot mutations among clinical and nonclinical isolates of *E. coli* indicates

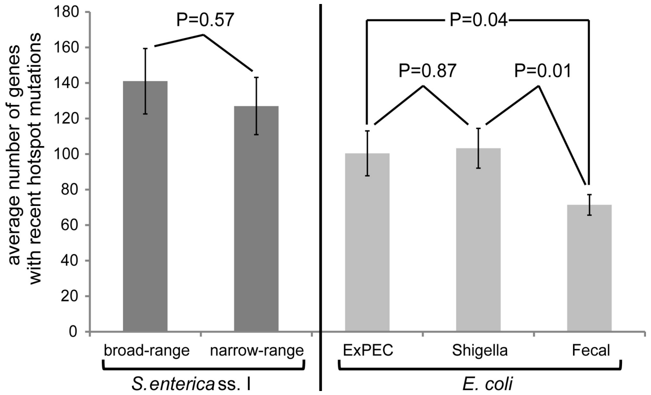
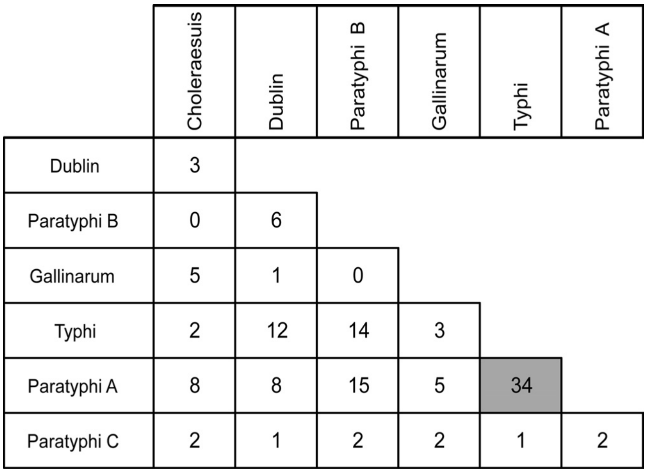


FIG 4 Average number of genes with recent hot spot mutations shared among isolates of broad-range and narrow-range serovars of *S. enterica* subsp. I and among ExPEC, *Shigella*, and fecal isolates of *E. coli*. The error bars denote standard errors, and *P* values were calculated using the z-test.



	Choleraesuis	Dublin	Paratyphi B	Gallinarum	Typhi	Paratyphi A
Dublin	3					
Paratyphi B	0	6				
Gallinarum	5	1	0			
Typhi	2	12	14	3		
Paratyphi A	8	8	15	5	34	
Paratyphi C	2	1	2	2	1	2

FIG 5 Matrix of number of genes with recent hot spot mutations shared exclusively between each *Salmonella* serovar pair in the narrow-range category. The number of genes that shared recent hot spot mutations exclusively between Typhi and Paratyphi A was found to be significantly higher ($P < 0.001$) than the rest and is highlighted in gray.

that the organisms isolated from infected hosts could have been under selection for pathogenicity. Recurrent mutations in the same-pathotype isolates suggest an adaptive advantage of such changes, possibly leading to improved survival as a pathogen.

We then performed a comparative analysis of the distribution of hot spot genes within the narrow-host-range pathotype composed of serovars that cause the most severe, usually invasive infections in humans and/or animals. Pairwise, the narrow-host-range serovars shared on an average 6 genes with hot spot mutations, but the range was very wide (from 0 to 34) (Fig. 5). Interestingly, Typhi and Paratyphi A, two human-restricted serovars, shared the highest number of mutated genes (34 genes) compared to the rest of the host-restricted serovars ($P < 0.01$). This frequency of shared hot spot genes between Typhi and Paratyphi A was 6-fold higher than the random expectation ($P < 0.001$).

The Typhi/Paratyphi A-specific hot spot genes (Table 2) are spread across the genomes (see Fig. S2 in the supplemental material), indicating that the adaptive mutations target a broad spectrum of genes and are not limited to specific gene clusters. These two serovars are equipped to infect and, sometimes, coinfect (22) the same tissues of the human host. Therefore, the infection of common compartments is expected to exert similar selective pressures.

A previous study (11) showed that Typhi and Paratyphi A are phylogenetically distant members of *S. enterica*, similar to what we found with respect to MLST phylogeny (Fig. 1A). It was shown previously (11) that the divergence of the majority of Typhi and Paratyphi A genes was relatively high, i.e., 1.2% (which is above the average pairwise diversity of *Salmonella* core genes observed in our study), indicating independent evolutionary origins of the serovars. However, it was also demonstrated that in the rest, low-diversity genes (comprising of about a quarter of the genome) were commonly a signal of homologous recombination between the serovars, indicating that pathogenic similarity of Typhi and Paratyphi A could be in part due to their direct genetic exchange (11).

Interestingly, the 34 genes with Typhi/Paratyphi A-specific hot spot mutations that we found here were not of recombinant origin, and their average diversity was 1.19% ± 0.06%, i.e., equivalent to the high-divergence regions between Typhi and Paratyphi A determined previously. Thus, although recombination could contribute to the convergent evolution of Typhi and Paratyphi serovars, our study indicates an important role for point mutations in the process of adaptive convergence of these two humans-restricted serovars that cause the most severe types of *Salmonella* infection.

Among the genes with Typhi/Paratyphi A-specific hot spot mutations, 10 genes belonged to two over-represented functional clusters: propanoate metabolism and signal proteins. It was proposed previously that the persistence of pathogenic strains in the intestine may be facilitated by the catabolism of abundant fatty acids such as propanoate (32). In particular, phosphotransacetylase (*pta*) that represents the cluster plays an important role in *Salmonella* virulence (24). Repression of this enzyme results in the accumulation of acetyl coenzyme A (acetyl-CoA), allowing growth retardation along with repression of signaling through the PhoP/PhoQ two-component global virulence regulator in Gram-negative pathogens (5, 16, 26).

In the list of signal proteins, the proteins with known virulence-related functions included ferrioxamine receptor (*foxA*), transglycosylase of penicillin-binding protein (*pbpC*), and a response regulator (*yfbS*). For example, the ferrioxamine binding and uptake, in association with the *tonB* protein, may play a role in the intestinal colonization of pathogenic bacteria (25).

Although other Typhi/Paratyphi A-specific hot spot genes did not form enriched functional categories, some of them—*entD*, *pipB*, *rnb*, *recC*, *aroE*, and *dsdA* in particular—are well known to be important for survival and virulence of *Salmonella*. The enterochelin synthetase component D encoded by *entD* is known to be a part of the iron acquisition system used by pathogenic bacteria for virulence in the iron-limiting stage in human host (21, 53). The *Salmonella* pathogenicity island (SPI-5)-encoded protein PipB is a well-known virulence factor in *Salmonella*, for which the mutant phenotype was shown to have attenuated virulence in mice (28). It is known that exoribonucleases are involved in virulence in *Salmonella* (13), and Rnb (RNase II) is one of such hydrolytic enzyme. *S. enterica* mutants lacking the RecBC function are avirulent in mice and unable to grow inside macrophages since they are highly sensitive to oxidative compounds synthesized by macrophages (2). Therefore, recombination is required for the repair of oxidative damage within the macrophage to survive, and RecC is one of the important subunits of the homologous recombination machinery RecBCD (3). Gene *aroE* is involved in the aromatic amino acid biosynthesis pathway in *Salmonella*, and mutations in this pathway have been the basis of live attenuated *S. Typhi* vaccines (39, 42). Finally, there are reports that mutations in *dsdA*, which encodes a positive regulator of D-serine deaminase, enhance the ability of uropathogenic strain to infect the bladders and kidneys of mice (37).

Conclusions. The primary goal of the present study was to show that, in both *Salmonella* subspecies I and *E. coli*, a substantial number of core genes are likely to acquire point mutations under positive selection. Since these mutations would then be functionally adaptive, they might play potentially significant role in the evolution of virulence of these bacterial species, i.e., to be pathoadaptive in nature. Due to limitations

TABLE 2 Genes with recent hot spot mutations shared between *Salmonella* serovars Typhi and Paratyphi A[a]

							Salmonella serovar mutation	
Gene	Product	Length (bp)	Functional category[b]	No. of syn	No. of nonsyn	HS pos. (aa)	Typhi	Paratyphi A
foxA	Ferrioxamine receptor	2,109	Signal protein*/metal ion transport	15	5	163	A→S	A→S
phnX	2-Aminoethylphosphonate transport	813	Metal ion binding	3	5	77	S→A	S→A
						114	A→T	A→T
ylaB	Putative diguanylate cyclase/phosphodiesterase	1,551	Membrane	17	6	193	P→H	P→S
entD	Enterochelin synthetase, component D	705	Metal ion transport	5	9	68	T→A	T→A
sucD	Succinyl-CoA synthetase, alpha subunit	870	Propanoate metabolism*	22	3	10	K→N	K→R
ycaM	Putative APC family, amino acid transporter	1,422	Membrane transporter	15	1	188	T→A	T→A
pipB	Pathogenicity island-encoded protein SPI5	876	Membrane	5	8	55	R→W	R→Q
scsD	Suppression of copper sensitivity protein	507	Signal protein*	6	2	100	D→G	D→G
STM1123	Putative periplasmic protein	396	Periplasmic	5	0	54	L→I	L→I
yceB	Putative outer membrane lipoprotein	561	Signal protein*	6	0	56	V→A	V→A
ptsG	Glucose-specific IIBC component of the sugar-specific PTS family	1,434	Signal protein*	6	7	175	I→V	I→V
nagZ	Putative glycosyl hydrolase	1,026	Peptidoglycan biosynthetic process	9	0	199	T→A	T→A
rnb	RNase II	1,935	RNA binding	15	1	244	T→A	T→A
						312	A→V	A→V
STM2196	Putative D-serine dehydratase	1,368	Gluconeogenesis	11	2	323	I→M	I→M
yfbS	Putative response regulator	1,827	Signal protein*/metal ion transport	24	5	46	P→S	P→S
pta	Phosphotransacetylase	2,145	Propanoate metabolism*	20	0	371	E→D	E→D
yfdC	Putative transport	942	Membrane transporter	8	2	256	L→F	L→F
yfdZ	Putative aminotransferase	1,239	Transferase	13	2	4	F→S	F→S
STM2529	Putative anaerobic dimethyl sulfoxide reductase	630	Metal ion binding	8	6	150	P→S	P→L
pbpC	Transglycosylase of penicillin-binding protein	2,316	Signal protein*	16	7	268	Q→R	Q→R
						269	N→S	N→S
ygbI	Putative regulatory protein, deoR family	765	Transcriptional regulator/ DNA binding	4	1	135	D→G	D→G
recC	Exonuclease V, subunit	3,372	Exodeoxy-RNase	37	5	225	R→K	R→K
tdcE	Pyruvate formate-lyase 4	2,295	Propanoate metabolism*	23	5	97	D→E	D→N
aroE	Dehydroshikimate reductase	819	Oxidoreductase	6	4	3	T→A	T→I
fic	Putative cell filamentation protein	603	Transferase	11	3	171	K→N	K→N
STM3549	Putative inner membrane protein	972	Membrane	12	5	318	R→K	R→K
yhjH	Putative diguanylate cyclase/phosphodiesterase	768	Nucleotide binding	8	6	319	I→M	I→M
kdgK	Ketodeoxygluconokinase	930	Transferase	7	3	199	T→I	T→I
selB	Selenocysteinyl-tRNA-specific translation	1,851	Nucleotide binding	27	7	205	Q→K	Q→K
dsdA	D-Serine deaminase	1,323	Amino acid Metabolic process	20	9	362	A→D	A→D
						399	A→V	A→T
yjbR	Putative cytoplasmic protein	357	Cytoplasmic	7	1	69	P→S	P→S
ytfN	Putative periplasmic protein	3,780	Signal protein*	38	14	187	D→E	D→E
STM4549	Putative cytoplasmic protein	459	Cytoplasmic	5	1	12	D→A	D→E
yjjV	Putative hydrolase	774	Endodeoxy RNase	12	7	143	H→R	H→R
						170	R→Q	R→Q

[a] "HS pos. (aa)" indicates the amino acid position that accumulated hot spot (HS) mutations, whereas "No. of Syn" and "No. of nonsyn" denote the numbers of synonymous and nonsynonymous differences, respectively, in the pairwise alignment of Typhi and Paratyphi A sequences.

[b] Asterisks (*) indicate statistically over-represented ($P < 0.05$) protein functional groups.

of space, we did not discuss in detail exactly how the putative pathoadaptive mutations could affect the function of specific genes, based on their locations in the proteins for which structure-function data are available. We also did not analyze genes that are not of the core nature based on our relatively strict definition but are still shared by at least some of the isolates. Such mosaic genes are also likely to accumulate point mutations under positive selection. Considering the known or expected role of these (usually horizontally transferred) genes in virulence evolution, the mutations therein could significantly affect pathogenicity of both *Salmonella* and *E. coli*.

Due to a large amount of information that could be obtained from the mutational variability of bacterial genes across the entire genome of a species, it is not practical to perform in-depth analysis by a single research group. It certainly should be a combined effort by many different laboratories in the field, specializing in studying specific bacterial traits or specific aspects of bacterial evolution and population genetics.

Therefore, we provided here detailed lists of nonrecombinant core genes with hot spot mutations in both *Salmonella* subspecies I (see Table S4 in the supplemental material) and *E. coli* (see Table S5 in the supplemental material). Such a gene list, we believe, is an important resource for identifying genes with naturally occurring adaptive amino acid variations for functional assays to decipher any physiologic and virulence significance. The identification of genes undergoing positive selection in specific strains (or strain groups) within and between species is crucial for the understanding of common mechanisms of pathogenesis and adaptive evolution in microbial pathogens.

REFERENCES

1. **Bruen TC, Philippe H, Bryant D.** 2006. A simple and robust statistical test for detecting the presence of recombination. Genetics **172**:2665–2681.
2. **Buchmeier NA, Lipps CJ, So MY, Heffron F.** 1993. Recombination-deficient mutants of *Salmonella typhimurium* are avirulent and sensitive to the oxidative burst of macrophages. Mol. Microbiol. **7**:933–936.
3. **Cano DA, Pucciarelli MG, Garcia-del Portillo F, Casadesus J.** 2002. Role of the RecBCD recombination pathway in *Salmonella* virulence. J. Bacteriol. **184**:592–595.
4. **Chan CX, Beiko RG, Ragan MA.** 2007. A two-phase strategy for detecting recombination in nucleotide sequences. South Afr. Comp. J. **38**:20–27.
5. **Chang DE, Shin S, Rhee JS, Pan JG.** 1999. Acetate metabolism in a *pta* mutant of *Escherichia coli* W3110: importance of maintaining acetyl coenzyme A flux for growth and survival. J. Bacteriol. **181**:6656–6663.
6. **Chattopadhyay S, Dykhuizen DE, Sokurenko EV.** 2007. ZPS: visualization of recent adaptive evolution of proteins. BMC Bioinform. **8**:187. doi:10.1186/1471-2105-8-187.
7. **Chattopadhyay S, et al.** 2007. Haplotype diversity in "source-sink" dynamics of *Escherichia coli* urovirulence. J. Mol. Evol. **64**:204–214.
8. **Chattopadhyay S, et al.** 2009. High frequency of hot spot mutations in core genes of *Escherichia coli* due to short-term positive selection. Proc. Natl. Acad. Sci. U. S. A. **106**:12412–12417.
9. **Deluca TF, et al.** 2006. Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics **22**:2044–2046.
10. **Dennis G, Jr, et al.** 2003. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. **4**:P3.
11. **Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D.** 2007. A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res. **17**:61–68.
12. **Didelot X, et al.** 2011. Recombination and population structure in *Salmonella enterica*. PLoS Genet. **7**:e1002191. doi:10.1371/journal.pgen.1002191.
13. **Domingues S, et al.** 2009. Biochemical characterization of the RNase II family of exoribonucleases from the human pathogens *Salmonella typhimurium* and *Streptococcus pneumoniae*. Biochemistry **48**:11848–11857.
14. **Efron B.** 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. **7**:1–26.
15. **Francisco Garcia-Del Portillo FG.** 2000. *Salmonella* pathogenicity islands (SPIs), p 26–34. *In* Cary JW, Linz JE, Bhatnagar D (ed), Microbial foodborne diseases: mechanisms of pathogenesis and toxin synthesis. Technomic Publishing Company, Inc, Lancaster, PA.
16. **Groisman EA.** 2001. The pleiotropic two-component regulatory system PhoP-PhoQ. J. Bacteriol. **183**:1835–1842.
17. **Guttman DS, Dykhuizen DE.** 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266**:1380–1383.
18. **Hall BG.** 2008. Simulating DNA coding sequence evolution with EvolveAGene 3. Mol. Biol. Evol. **25**:688–695.
19. **Hansen-Wester I, Hensel M.** 2001. *Salmonella* pathogenicity islands encoding type III secretion systems. Microbes Infect. Inst. Pasteur **3**:549–559.
20. **Hughes AL, Nei M.** 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.
21. **Janke B, Dobrindt U, Hacker J, Blum-Oehler G.** 2001. A subtractive hybridization analysis of genomic differences between the uropathogenic *Escherichia coli* strain 536 and the *E. coli* K-12 strain MG1655. FEMS Microbiol. Lett. **199**:61–66.
22. **Joshi S, Wattal C, Sharma A, Prasad KJ.** 2002. Mixed *Salmonella* infection: a case report. Indian J. Med. Microbiol. **20**:113–114.
23. **Katoh H, Isshiki A, Masunaka A, Yamamoto H, Akimitsu K.** 2006. A virulence-reducing mutation in the postharvest citrus pathogen, *Alternaria citri*. Phytopathology **96**:934–940.
24. **Kim YR, Brinsmade SR, Yang Z, Escalante-Semerena J, Fierer J.** 2006. Mutation of phosphotransacetylase but not isocitrate lyase reduces the virulence of *Salmonella enterica* serovar Typhimurium in mice. Infection and Immunity **74**:2498–2502.
25. **Kingsley RA, et al.** 1999. Ferrioxamine-mediated iron(III) utilization by *Salmonella enterica*. Appl. Environ. Microbiol. **65**:1610–1618.
26. **Lesley JA, Waldburger CD.** 2003. Repression of *Escherichia coli* PhoP-PhoQ signaling by acetate reveals a regulatory role for acetyl coenzyme A. J. Bacteriol. **185**:2563–2570.
27. **Ma C, et al.** 2010. Energy production genes *sucB* and *ubiF* are involved in persister survival and tolerance to multiple antibiotics and stresses in *Escherichia coli*. FEMS Microbiol. Lett. **303**:33–40.
28. **Marcus SL, Brumell JH, Pfeifer CG, Finlay BB.** 2000. *Salmonella* pathogenicity islands: big virulence in small packages. Microbes Infect. Inst. Pasteur **2**:145–156.
29. **Martincorena I, Seshasayee AS, Luscombe NM.** 2012. Evidence of non-random mutation rates suggests an evolutionary risk management strategy. Nature **485**:95–98.
30. **McFarland WC, Stocker BA.** 1987. Effect of different purine auxotrophic mutations on mouse-virulence of a Vi-positive strain of *Salmonella dublin* and of two strains of *Salmonella typhimurium*. Microb. Pathog. **3**:129–141.
31. **Miller SI, Kukral AM, Mekalanos JJ.** 1989. A two-component regulatory system (*phoP phoQ*) controls *Salmonella typhimurium* virulence. Proc. Natl. Acad. Sci. U. S. A. **86**:5054–5058.
32. **Munoz-Elias EJ, McKinney JD.** 2006. Carbon metabolism of intracellular bacteria. Cell. Microbiol. **8**:10–22.
33. **Nei M, Gojobori T.** 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.
34. **Philippe N, Crozat E, Lenski RE, Schneider D.** 2007. Evolution of global regulatory networks during a long-term experiment with *Escherichia coli*. Bioessays **29**:846–860.
35. **Richardson AR, et al.** 2009. The base excision repair system of *Salmonella enterica* serovar Typhimurium counteracts DNA damage by host nitric oxide. PLoS Pathog. **5**:e1000451. doi:10.1371/journal.ppat.1000451.
36. **Rickman L, et al.** 2005. A member of the cAMP receptor protein family of transcription regulators in *Mycobacterium tuberculosis* is required for virulence in mice and controls transcription of the *rpfA* gene coding for a resuscitation promoting factor. Mol. Microbiol. **56**:1274–1286.
37. **Roesch PL, et al.** 2003. Uropathogenic *Escherichia coli* use d-serine deaminase to modulate infection of the murine urinary tract. Mol. Microbiol. **49**:55–67.
38. **Samant S, et al.** 2008. Nucleotide biosynthesis is critical for growth of bacteria in human blood. PLoS Pathog. **4**:e37. doi:10.1371/journal.ppat.0040037.
39. **Sheikh A, et al.** 2011. In vivo expression of *Salmonella enterica* serotype Typhi genes in the blood of patients with typhoid fever in Bangladesh. PLoS Neglected Trop. Dis. **5**:e1419. doi:10.1371/journal.pntd.0001419.
40. **Sokurenko EV, et al.** 2004. Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. Mol. Biol. Evol. **21**:1373–1383.
41. **Sokurenko EV, Gomulkiewicz R, Dykhuizen DE.** 2006. Source-sink dynamics of virulence evolution. Nat. Rev. Microbiol. **4**:548–555.
42. **Stocker BA.** 1988. Auxotrophic *Salmonella typhi* as live vaccine. Vaccine **6**:141–145.
43. **Stojiljkovic I, Baumler AJ, Heffron F.** 1995. Ethanolamine utilization in *Salmonella typhimurium*: nucleotide sequence, protein expression, and mutational analysis of the *cchA cchB eutE eutJ eutG eutH* gene cluster. J. Bacteriol. **177**:1357–1366.
44. **Suvarnapunya AE, Lagasse HA, Stein MA.** 2003. The role of DNA base excision repair in the pathogenesis of *Salmonella enterica* serovar Typhimurium. Mol. Microbiol. **48**:549–559.
45. **Swofford DL.** 2000. PAUP*: phylogenetic analysis using parsimony and other methods (software). Sinauer Associates, Sunderland, MA.
46. **Tamura K, Dudley J, Nei M, Kumar S.** 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. **24**:1596–1599.
47. **Touchon M, et al.** 2009. Organized genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5:e1000344. doi:10.1371/journal.pgen.1000344.
48. **Turner AK, Lovell MA, Hulme SD, Zhang-Barber L, Barrow PA.** 1998. Identification of *Salmonella typhimurium* genes required for colonization

of the chicken alimentary tract and for virulence in newly hatched chicks. Infect. Immun. **66**:2099–2106.

49. **Walk ST, et al.** 2009. Cryptic lineages of the genus *Escherichia*. Appl. Environ. Microbiol. **75**:6534–6544.
50. **Wall DP, Fraser HB, Hirsh AE.** 2003. Detecting putative orthologs. Bioinformatics **19**:1710–1711.
51. **Weissman SJ, et al.** 2007. Differential stability and trade-off effects of pathoadaptive mutations in the *Escherichia coli* FimH adhesin. Infect. Immun. **75**:3548–3555.
52. **Wirth T, et al.** 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. **60**:1136–1151.
53. **Woodrow GC, Young IG, Gibson F.** 1979. Biosynthesis of enterochelin in *Escherichia coli* K-12: separation of the polypeptides coded for by the *entD*, *entE*, *entF*, and *entG* genes. Biochim. Biophys. Acta **582**:145–153.
54. **Zahrt TC, Maloy S.** 1997. Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. Proc. Natl. Acad. Sci. U. S. A. **94**:9786–9791.