High frequency of hotspot mutations in core genes of *Escherichia coli* due to short-term positive selection

Sujay Chattopadhyay[a], Scott J. Weissman[b], Vladimir N. Minin[c], Thomas A. Russo[d,e,f], Daniel E. Dykhuizen[g], and Evgeni V. Sokurenko[a,1]

[a]Department of Microbiology, [b]Division of Infectious Diseases, Department of Pediatrics, and [c]Department of Statistics, University of Washington, Seattle, WA 98195; [d]Veterans Administration Western New York Healthcare System, [e]Department of Medicine, and [f]Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, State University of New York, Buffalo, NY 14260; and [g]Department of Ecology and Evolution, University at Stony Brook, State University of New York, Stony Brook, NY 11794

Communicated by John R. Roth, University of California, Davis, CA, June 5, 2009 (received for review January 7, 2009)

Core genes comprising the ubiquitous backbone of bacterial genomes are not subject to frequent horizontal transfer and generally are not thought to contribute to the adaptive evolution of bacterial pathogens. We determined, however, that at least one-third and possibly more than one-half of the core genes in *Escherichia coli* genomes are targeted by repeated replacement substitutions in the same amino acid positions—hotspot mutations. Occurrence of hotspot mutations is driven by positive selection, as their rate is significantly higher than expected by random chance alone, and neither intragenic recombination nor increased mutability can explain the observed patterns. Also, commensal *E. coli* strains have a significantly lower frequency of mutated genes and mutations per genome than pathogenic strains. *E. coli* strains causing extra-intestinal infections accumulate hotspot mutations at the highest rate, whereas the highest total number of mutated genes has been found among *Shigella* isolates, suggesting the pathoadaptive nature of such mutations. The vast majority of hotspot mutations are of recent evolutionary origin, implying short-term positive selection, where adaptive mutations emerge repeatedly but are not sustained in natural circulation for long. Such pattern of dynamics is consistent with source-sink model of virulence evolution.

commensal | *E. coli* core genome | extra-intestinal | pathoadaptive evolution | *Shigella*

Adaptive evolution leading to phenotypic diversification within a species is driven by 2 major genetic mechanisms, mutation and horizontal gene transfer. Within eukaryotic and some viral species, the mutational mechanism predominates, as variability in the presence/absence of genes is relatively rare among representatives of the same species (1, 2). Thus, large numbers of eukaryotic and viral genetic studies are devoted to finding associations between the presence of mutations in specific genes and phenotypes such as resistance or predisposition to certain diseases among humans or the ability to cause epidemic or chronic diseases among viruses (3, 4). In contrast, with bacterial genomes, the importance of mutations to within-species evolution has been overshadowed by the presumably overwhelming importance of genes that are transferred horizontally: plasmids, prophages, or chromosomal insertions (islands) (5, 6). Thus, as multiple representative genomes from individual bacterial species have become publicly available, a major focus of their comparative analysis has been the association of presence/absence of specific genes with particular bacterial phenotypes or within-species ecotypes (7, 8). Here, however, we show that based on the occurrence of replacement hotspot mutations, adaptive mutations accumulate in a substantial fraction of core genes of *Escherichia coli* genomes that, unlike horizontally transferred genes, are present in all strains.

E. coli is a frequent commensal inhabitant of the large intestine of humans and other mammals, but nonetheless causes a variety of intestinal and extra-intestinal infections, some of which may be life-threatening. For example, the publicly available, complete *E. coli* genomes encompass commensal strains along with clinical isolates from cases of dysentery (shigellosis), hemolytic-uremic syndrome, and urinary tract infection, among others. The cross-genomic studies of *E. coli* undertaken thus far demonstrate rather strikingly that the majority of genes in a given strain are not present in all *E. coli* genomes, being horizontally transferred into select lineages (9–11). Although certain associations between presence/absence of genes and isolation origin of strains (ecotype) have been established, even strains isolated from the same type of infection exhibit high levels of genome plasticity. Some studies have examined occurrence of mutations, in particular single nucleotide polymorphisms (SNPs), under positive selection in the genes shared by different genomes (12, 13). Although such loci have been identified and some correlations with strain origins (pathogenicity) have been proposed, the number of candidate loci has been very low. Such paucity may be due to the fact that common population genetic analyses—such as Tajima D (14), Fu & Li D^* (15), or the ratio of nonsynonymous to synonymous mutation rates (dN/dS) in genes (16), individual codons (17), or along phylogenetic branches (18)—detect relatively long-term evolutionary forces resulting in high fractions of adaptive mutations. This limitation is of critical importance, as it results in overlooking genes with minimal numbers of positively selected SNPs, especially when these have occurred in diverse allelic backgrounds. Even a single SNP, when leading to an amino acid replacement, can have a significant adaptive effect on the function of a coded protein and provide means for rapid adaptation of bacterial populations to niche-specific conditions (19). Such adaptive point mutations commonly occur in bacteria that cause clinical infections in human or other hosts, and are termed pathoadaptive mutations. Mutator strains are especially prone to rapid acquisition of adaptive mutations (20).

We have recently developed a molecular evolution tool, zonal phylogeny (ZP) analysis, specifically designed to identify footprints of positive selection in genes with minimal changes per unique sequence variant (i.e., allele), by differentiating them from variations that are either selectively neutral or detrimental (21–23). Among several approaches used by ZP is the identification of replacement hotspot mutations—repeated, phylogenetically unlinked mutations in the same amino acid position. Occurrence of hotspot mutations is strong evidence of positive selection, indicating that replacement of a specific amino acid provides an adaptive advantage under particular conditions and, thus, is repeatedly selected in different allelic backgrounds of clones competing for survival in the same environments (24, 25).

Here, we analyzed the occurrence of replacement hotspot mutations in core genes that are omnipresent in the current set of *E. coli* genomes, to evaluate the extent to which hotspot SNPs target

Author contributions: S.C., V.N.M., D.E.D., and E.V.S. designed research; S.C., S.J.W., V.N.M., and T.A.R. performed research; S.C., D.E.D., and E.V.S. analyzed data; and S.C., S.J.W., V.N.M., T.A.R., D.E.D., and E.V.S. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: evs@u.washington.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0906217106/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.0906217106

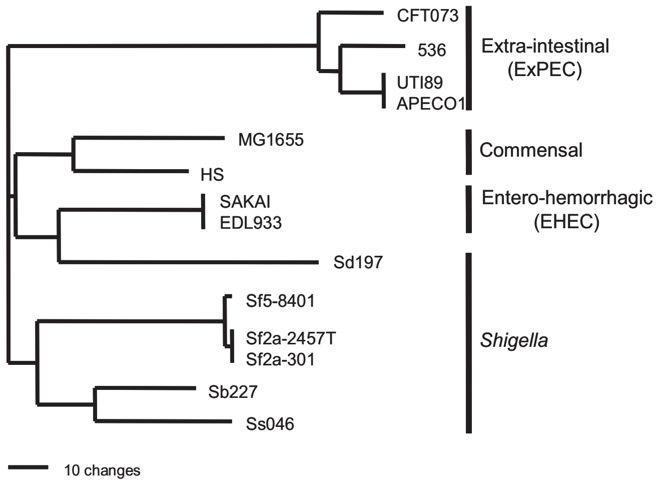

Fig. 1. Phylogram of concatenated sequences of 7 housekeeping genes (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, *recA*) for 14 *E. coli/Shigella* genomes analyzed.

genes during within-species microevolution of *E. coli* and to what extent they might be associated with the ability of strains to cause different types of clinical infection.

Results

Clonal Analysis of Sequenced Strains. Occurrence of hotspot SNPs was first analyzed across 14 assembled *E. coli* genomes available in public databases: 4 extra-intestinal pathogens (ExPEC), 6 *Shigella*, 2 enterohemorrhagic (EHEC) pathogens, and 2 commensal isolates. Phylogenetic relationship of the strains was established based on complete coding regions of 7 housekeeping genes frequently used for multilocus sequence typing (MLST) analysis of *E. coli*: *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA* (ref. 26; http://mlst.ucc.ie).

The ExPEC isolates formed a distinct cluster with average pairwise nucleotide diversity $\pi = 0.004 \pm 0.001$ (Fig. 1), reflecting their shared origin in the B2 phylogenetic group of *E. coli* (27). Avian pathogenic strain APECO1 and uropathogen UTI89 were identical over all 7 genes and thus were clonally related. As expected, the 2 O157:H7 EHEC isolates were also clonally related. Also as expected, strains of *Shigella flexneri* were clonally related, with 2 serotype 2a strains, 2457T and 301, having identical sequences, and serotype 5 strain 8401 differing from the former by one nonsynonymous change in *adk* and two synonymous changes in *icd*. The rest of the strains were all clonally distinct and relatively diverse ($\pi = 0.007–0.012$). In the subsequent analysis, the genomes of all strains have been included, unless otherwise specified.

Nucleotide Diversity of Core Genes. Due to its large size and well-annotated assembly, the genome of the uropathogenic strain CFT073 was designated as the reference genome. After exclusion of pseudogenes, a total of 5,377 ORFs of CFT073 was used for a BLAST search of orthologous genes among the remaining 13 strains. Highly homologous (≥95% nucleotide sequence identity) copies of 1,516 genes were found in all 14 strains and were defined as core genes (Fig. 2*A*). Of the core genes, all but 3 were present in a single copy in each genome and determined to be of orthologous nature. The 3 core genes with multiple copies in at least 1 strain [*tufB* (elongation factor Tu), c4351 (conserved hypothetical protein), and *slpA* (peptidyl-prolyl *cis-trans* isomerase)] were excluded from further analysis.

In the analyzed data set of 1,513 core genes, each gene was represented by an average of 8.6 alleles. Only 9 genes were identical in all 14 strains (see Table S1), whereas no gene had a different allele in each of the 14 strains. Out of 1,504 core genes with at least 2 alleles, average $\pi = 0.015$, with rates of nonsynonymous (*dN*) and synonymous (*dS*) substitutions being 0.004 and 0.052, respectively ($dN/dS = 0.08$). The z-test (at $P < 0.05$) comparing the distribution of *dN* and *dS* (28) showed only 34 (2%) of core genes with *dN* significantly higher than *dS* ($dN/dS \gg 1$), with bootstrap analysis confirming $dN/dS \gg 1$ in only 20 genes (*SI Materials and Methods* and Table S2).

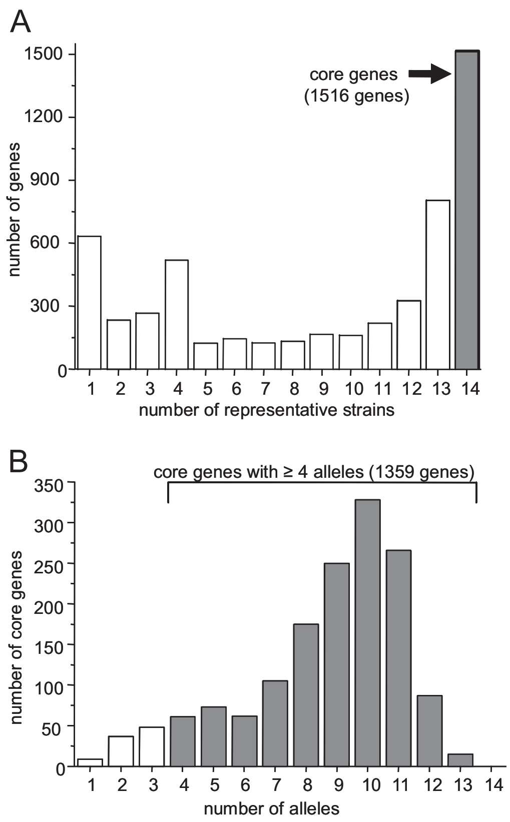

Fig. 2. (*A*) Frequency distribution of genes (of reference genome CFT073) across 14 genomes. The genes represented by only 1 genome are unique to CFT073, whereas the ones represented by all 14 genomes (shown as gray bar) are termed "core" genes. (*B*) Allelic diversity of core genes. The genes with at least 4 alleles (shown as gray bars) could be used for the maximum-likelihood based phylogenetic reconstruction and the following ZP-analysis.

Thus, according to the conservative *dN*/*dS* statistic, candidate genes under positive selection amount to only 1.3% of core genes, even less than the fraction expected by chance at the $P < 0.05$ significance level.

Occurrence of Hotspot Mutations Under Positive Selection. Because the occurrence of replacement hotspot mutations, repeated independent changes in the same amino acid position, could be strong evidence of positive selection, we assessed their occurrence in the core genes. Identical changes in the same amino acid position are defined as parallel hotspots. Different amino acid replacements in the same position are defined as coincidental hotspots.

As parallel hotspots in different alleles are identical in nature but not linked by (i.e., not present in) a common ancestor, they result in homoplasy on the phylogenetic trees that, in turn, can be assessed by the analysis of consistency index (CI) of the corresponding gene (29). To construct an unrooted phylogenetic tree, a gene is required to have at least 4 alleles, as were present for 1,359 core genes (Fig. 2*B*). The CI analysis of the gene trees suggested that 25% of these are affected by homoplasy, with CI < 1 (Fig. S1). Because such substantial frequency of homoplasy suggests frequent occurrence of hotspot mutations, we investigated this possibility further using ZP software (ZPS) that is able to identify the number and nature of

Table 1. Occurrence of hotspot mutations in core genes relative to simulated datasets

Candidates	All hotspots	Parallel hotspots only	Coincidental hotspots only	Mixed, parallel, and coincidental hotspots
No. of genes with hotspots	290	167	75	48
No. of recombinants	99	54	16*	29*
No. of nonrecombinants	191	113*	59*	19*
Frequency of hotspots in nonrecombinants				
Real datasets	0.243	0.163	0.080	N/A
Simulated datasets	0.033 ± 0.002	0.006 ± 0.001	0.027 ± 0.002	N/A

Number of genes followed by asterisk (*) indicates genes selected as candidates under positive selection. The frequency of hotspot mutations was calculated as a ratio of the number of (all or parallel or coincidental) hotspot mutations to the total number of replacement mutations. For each of the 191 nonrecombinant genes, 10 simulations were performed. N/A, not applicable.

replacement hotspot mutations using protein trees reconstructed from the corresponding DNA trees (23).

According to the ZPS analysis, a total of 290 genes contained replacement hotspot mutations, with an average of 4.5 ± 0.2 hotspot mutations and 2.0 ± 0.1 affected amino acid positions per gene, with 644 being parallel and 277 coincidental. Only parallel mutations were found in 167 genes, whereas only coincidental ones were found in 75 genes (Table 1). Both parallel and coincidental mutations occurred in 48 genes; such overlap is significantly greater ($P < 0.001$ according to 2×2 χ^2 test) than would be expected from chance alone, indicating that genes having one type of hotspot mutations are prone to have the other type as well.

Parallel hotspot polymorphisms may occur as point mutations, or such changes may arise by recombination, thereby leading to overestimation of the frequency of parallel hotspots as independent mutational events. Of 290 genes with hotspots, recombination was detected in 99 genes (Table 1). In the remaining nonrecombinant data set, genes with parallel mutations were still 2-fold more numerous than coincidental ones. To assess what patterns can be expected if mutations would be occurring randomly, we subjected each of the 191 nonrecombinant genes with hotspots to 10 rounds of simulated mutation under the null hypothesis of no selection (*SI Materials and Methods*). We found a significantly smaller frequency of hotspot mutations in the simulated data sets for either parallel or coincidental ones than in the real data sets ($P < 0.0001$; Table 1) and, in contrast to the predominance of parallel mutations in real data sets, coincidental mutations were 5 times more common than parallel ones in the simulated data sets.

The higher than expected rate of overall or specific types of hotspots might reflect increased mutation rate in specific nucleotide positions, like SNP hypermutability of CpG dinucleotides with methylated cytosines in eukaryotic genomes (30). No such point mutation hypermutability so far has been documented among prokaryotic microbes, and the relative pattern of substitution rates in all dinucleotides was found to be similar in genes with or without hotspot mutations (Fig. S2). The only hypermutable sites known in bacteria are short tandem repeats (STR) that are prone to small deletions/insertions, not SNPs (31–33).

Taken together, these results show that hotspot mutations in the core genes are acquired at a significantly higher rate than expected under neutrality, indicating that the observed hotspot mutations are the result of positive selection. For further study, 236 genes were designated as selection-candidate genes (marked by asterisks in Table 1 and Table S3). This group includes all genes with coincidental mutations, as these mutations are always true hotspot events, and genes with parallel-only mutations that show no evidence of recombination. As a conservative approach, recombinant genes with parallel-only mutations were not included as selection-candidate loci.

Recent Versus Long-Term Protein Variants. Protein variants that are sustained in nature without change over evolutionarily long periods of time are likely to be encoded by multiple phylogenetically linked alleles that differ by synonymous mutations (reflecting a molecular clock). In contrast, protein variants that have emerged only recently are likely to be encoded by single alleles, i.e., exhibiting no silent diversification among different strains. ZPS allows identification of long-term versus recent nonsynonymous variation in specific genes by collapsing silent changes along the DNA tree into corresponding protein variant nodes (23).

On average, a core gene encoded 1.7 long-term and 3.7 recent protein variants (Table 2). The long-term variants were encoded by significantly higher numbers of alleles and carried by significantly higher numbers of strains than recent variants. For 236 selection-candidate genes, the average number of recent variants was significantly higher (6.0 variants per gene, see Table 2). In 27 (11.5%) selection-candidate genes, all hotspot mutations occurred only in long-term protein variants (long-term hotspots); in 196 (83.4%) genes, all hotspot mutations occurred along the branches that led to the emergence of recent protein variants (recent hotspots); in the remaining 13 genes, hotspot mutations were found in both long-term and recent protein variants. Thus, the vast majority of hotspot mutations in candidate genes under selection have emerged relatively recently.

Occurrence of Recent Hotspot Mutations in Various *E. coli* Ecotypes. The highest overall number of genes affected by recent hotspot mutations was found in the *Shigella* ecotype (Table 3), with the

Table 2. Long-term and recent protein variants in all genes and in genes with hotspots

Dataset		No. per gene	
		Long-term	Recent
1,359 genes analyzed	Protein variants	1.7 ± 0.03	3.7 ± 0.07
	Alleles	5.3 ± 0.06	3.7 ± 0.07
	Representative strains	8.5 ± 0.10	5.3 ± 0.09
236 genes, hotspot-based selection candidates	Protein variants	1.8 ± 0.07	6.0 ± 0.15
	Alleles	3.9 ± 0.14	6.0 ± 0.15
	Representative strains	5.6 ± 0.21	8.3 ± 0.20

Table 3. Ecotype-specific prevalence of recent hotspot mutations and their rate of accumulation

Ecotype	Total no. of genes with recent hotspots	Avg. no. of genes with recent hotspots	Avg. no. of recent hotspot mutations	Rate of recent hotspot accumulation
ExPEC	113	82 ± 1.1	127.5 ± 1.9	4995.2 ± 848.0
Shigella	170	81.5 ± 2.5	108.7 ± 1.9	1354.2 ± 119.7
EHEC	84	84 ± 0.0	116 ± 0.0	N/A
Commensal	98	68.5 ± 1.5	91.0 ± 6.0	1246.0 ± 140.9

For EHEC ecotype, the two strains being clonally related, the rate was incalculable as denoted by N/A (not applicable).

highest number among individual isolates occurring in the *Shigella dysenteriae* strain (Table S4). The highest average number of recent hotspot mutations was accumulated among ExPEC strains (Table 3), with the highest number among individual isolates found in the strain UTI89 (Table S4). The lowest average numbers of the mutated genes or mutations were found in the commensal strains.

To evaluate the rate of accumulation of genes with recent hotspots within specific ecotypes, the number of genes with recent hotspots was calculated by pairwise comparison of strains belonging to the same ecotype and then normalized according to the evolutionary divergence of the strains, based on π as determined from the Fig. 1 phylogeny (Fig. S3). Because π equals (or approaches) zero in all or almost all genes with hotspots, and the mutations were of identical nature in clonally related strains, only UTI89 and *S. flexneri* 8401 strains were used for the within-ecotype analysis of ExPEC and *Shigella*, respectively, whereas the 2 clonally related EHEC strains had to be excluded from the analysis. Based on the diversity normalized values, isolates of the ExPEC ecotype accumulated genes with recent hotspots at the highest rate, exceeding by 4-fold the corresponding rate in commensal strains (Table 3; $P < 0.001$).

At least half of the genes with recent hotspots had a defined or putative function (Table S3) with various functional categories being enriched across different ecotypes (Fig. 3). There was an overlap between different ecotypes in the genes affected by recent hotspots (Fig. S4), but only the overlap between ExPEC and *Shigella* was significantly higher than random expectation ($P = 0.0001$). Overall, 36 genes have acquired recent hotspot mutations in all ecotypes, with only the metal-binding proteins being significantly enriched among them. Categories enriched only in a particular pathotype included response to DNA damage, helix-hairpin-helix DNA binding proteins (class I) and purine metabolism proteins (*Shigella* only), tRNA processing proteins (ExPEC only), and vitamin metabolism-related proteins (commensals only).

Dependence of the Number of Genes with Hotspots on the Number of Genomes Analyzed. Of 10 clonally distinct genomes analyzed above, random sets of 7, 8, and 9 genomes (7 samples each) were reanalyzed for the number of genes with hotspots. The average numbers of genes with either recent or long-term hotspots were increased more or less linearly when the number of genomes in the data set increased from 7 to 10 (Fig. 4, circle points). Inclusion of an additional clonally distinct genome [model urosepsis isolate CP9 from phylogenetic group B2 (*SI Materials and Methods*)] produced a further increase in genes with both hotspot categories without distinct signs of saturation, especially for the genes with recent hotspots (Fig. 4, triangle points).

As further genomes became publicly available during the course of this study, we included 4 additional clonally distinct strains (*SI Materials and Methods*) for estimation of numbers of genes with hotspots. The numbers of genes with recent and long-term hotspots increased (Fig. 4, diamond points), but signs of saturation were now detectable in both categories. The sigmoidal mode of increase in the numbers was reinforced by inclusion of 8 more clonally distinct genomes (*SI Materials and Methods*) that became publicly available most recently (Fig. 4, asterisk points). The number of genes with recent hotspots remained predominant, outnumbering the genes with only long-term hotspots by almost 5:1 (361 genes vs. 75 genes) ratio. Interestingly, the number of core, omnipresent genes in *E. coli* has dropped 10%, from 1,516 to 1,363 genes, as the number of genomes increased from the original 10 clonally distinct genomes to 23 at the end of this study. At the same time, the number of core genes that could be analyzed by ZPS (with 4 or more alleles) has dropped only 3.5%, from 1,359 to 1,311 genes. Thus, about one-

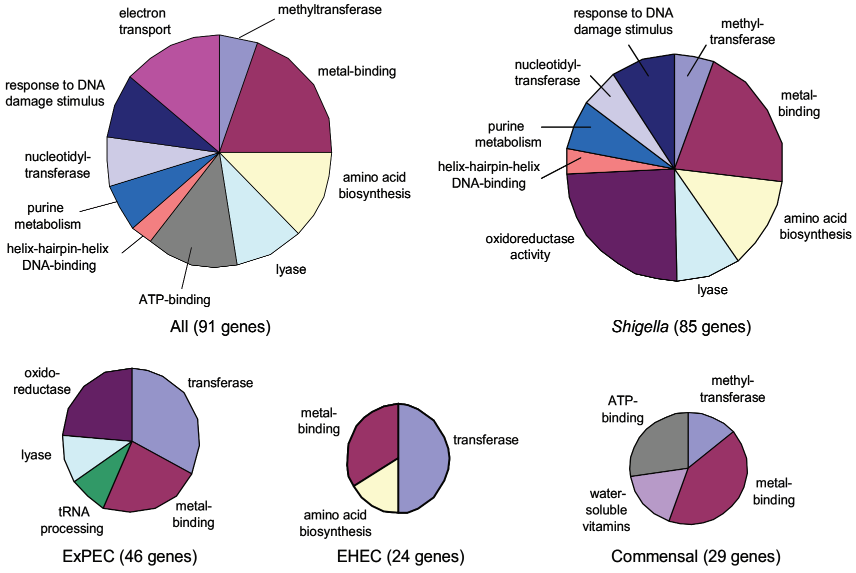

Fig. 3. Pie charts showing significantly ($P < 0.05$) enriched functional categories of candidate proteins with adaptive mutations (based on presence of recent hotspot mutations) in 4 ecotypes using DAVID (stringency, medium). For each pie chart, the number in parentheses denotes the number of selection-candidate genes included in the enriched categories.

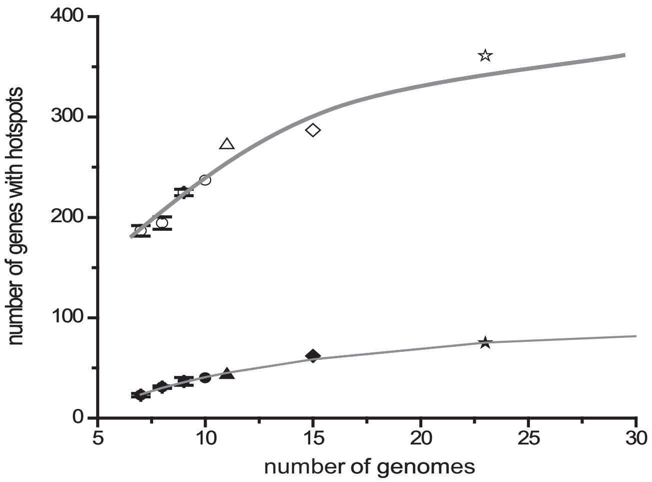

Fig. 4. Dependence of the number of genes with recent (shown by open points) or long-term (shown by filled points) hotspots in *E. coli* on the number of genomes analyzed. The best-fit curve based on data sets of 10 (circles), 11 (triangles), 15 (diamonds), 23 (asterisks) clonally unlinked genomes along with smaller randomly selected data sets of 7, 8, and 9 genomes (circles with error bars) was found to be sigmoidal [$y = (a + b\sqrt{x})/(1 + \sqrt{x})$] with $R^2 = 0.97$ for recent hotspots and $R^2 = 0.99$ for long-term hotspots.

third of core genes were affected by either recent or long-term hotspot mutations in the 23 *E. coli* genomes analyzed.

Assuming continuous addition of *E. coli* genomes for the analysis, the number of core genes is expected to stay at ≈1,300 genes. The number of core genes with recent and long-term hotspots could be projected to saturate at ≈600 and 150 genes, respectively, i.e., affecting about one-half of the core genes. Ninety percent saturation level is expected to be reached at ≈200 genomes.

Discussion

Of the protein coding genes in the *E. coli* genome, 28% were present in all 14 strains studied in detail here. Among these core genes, one-fifth possessed replacement hotspot mutations. The predominance of parallel over coincidental mutations in nonrecombinant genes strongly suggests nonneutral evolution of core genes that favors repeated mutations of identical nature in the same amino acid position. Besides, the differential frequency of hotspot mutations among different ecotypes provides additional support for the adaptive significance of hotspots. Neither of these results would be expected from a neutral model or an elevated mutation rate at specific nucleotide sites.

Previous studies that have simulated natural evolution in vitro demonstrated occurrence of replacement hotspot mutations in bacteriophages (34) and antibiotic-resistant *E. coli* (35). These results suggest that a broad range of environmental challenges can select for the same structural changes, leading to the increased frequency of hotspot changes across different lineages. The high frequency of hotspot mutations might also reflect the possibility that the evolving proteins have structural and functional constraints on residues capable of responding to selection. One also can suggest, however, that in proteins with a limited number of evolvable sites, the number of observed hotspot mutations might not be statistically different from the number expected due to random probability alone. However, simulations of mutation on all of the nonrecombinant core genes showed that, to observe statistically significant greater frequency of hotspot mutations in our real data sets over the simulated ones, the minimum length of proteins should be above 100 aa (Fig. S5), i.e., well below the size of a vast majority of the proteins with hotspot mutations (average size, 276 aa) that were identified here.

The predominance of parallel mutations suggests that the repeated replacement SNPs are under selection to modify the function of encoded proteins in a specific, fine-tuned manner, as has been observed with genes encoding adhesive proteins in *E. coli* (21, 36). In contrast, if the selection were operating to eliminate protein function or epitope recognition by antibodies, selection would likely result in coincidental mutations, i.e., multiple types of replacement in specific positions that are critical for an overall functional or structural integrity of the protein.

Despite the predominance of parallel mutations in nonrecombinant genes, recombination appears to be responsible at least for a portion of the parallel hotspot polymorphisms. Indeed, the highest proportion of recombinants was observed in genes with both parallel and coincidental hotspots. Because such genes are very likely to be under selection for hotspot changes, the increased recombination could reflect the efficient action of positive selection to provide for hotspot polymorphisms through recombination as well as through mutation. It should be noted, however, that an excess of parallel mutations can create the illusion of recombination and confuse all available recombination detection algorithms. Therefore, it is possible that some of the genes that we classified as recombinants and excluded from further analysis were, in fact, false positives induced by hotspot mutations.

There is a significantly higher frequency of recent hotspot mutations compared with long-term ones. The recent origin of protein variants may reflect either (*i*) modification in proteins important for the evolutionarily recent expansion of bacterial populations into recently available habitats for the species, and/or (*ii*) repeated adaptation to transient habitats or environmental conditions that do not support bacterial populations over the long term. Under the latter, so-called source-sink dynamics, SNPs that are adaptive for evolutionarily unstable environments emerge continuously, but are selected out of the population due to tradeoff effects in the original, stable habitat, where the long-term protein variants have selective advantage (22, 37). Thus, protein variants with recent hotspot mutations could be functionally significant for relatively short-term adaptation.

It is expected that pathogenic isolates, particularly ExPEC, would be exposed to a greater diversity of niches outside of what is considered to be the primary reservoir habitat of *E. coli*, the large intestine of healthy humans (6, 38). Consequently, pathogens may be expected to show more ecotype-specific adaptive evolution. The observed high frequency of recent hotspot mutations among pathogenic isolates (especially in ExPEC) confirms this expectation. Even if horizontally transferred genes (e.g., chromosomal islands) are major contributors to the niche expansion of pathogenic strains, competition in the habitats should lead to adaptive mutations in backbone genes as well (37, 39). The need for mutational adaptation of existing genes is reflected in the common occurrence of mutator phenotypes among bacterial pathogens (40, 41). However, none of strains examined here has been reported to be a mutator.

The overlap observed among different ecotypes in the genes affected by recent hotspot mutations could be due to physical overlap in the habitats where the selective pressures for hotspot acquisition are applied. Alternatively, hotspot changes in certain genes may have same adaptive effects, but in entirely distinct environments. For example, the significant overlap between ExPEC and *Shigella* strains may reflect similarities in certain aspects of their pathogenesis, e.g., intracellular invasion (albeit in different cell types and via different mechanisms), occasional systemic spread, or the need to evade strong innate immune responses. Determining what genes or gene categories are under positive selection in specific pathogens can provide valuable information on the mechanism of pathogenesis. However, due to the low number of strains analyzed here in detail, the association conclusions are likely to be premature.

Just as human population genomics have been used to associate genetic variation with disease susceptibility (3), microbial population genomics can be used to associate genetic variation with the ability of pathogens to cause disease in humans, even before the functions of genes under selection are fully understood. Finally, selection footprints will provide the research community with loci for functional studies powered by comparison of naturally occurring adaptive variants, thus facilitating the characterization of structure and function according to physiologic and virulence perspectives.

Materials and Methods

Genomes Analyzed. In this paper, we analyzed in detail 14 genomes: 2 commensals, MG1655 (accession no. U00096) and HS (NC_009800) from phylogenetic groups A and AxB1, respectively; 2 EHEC isolates, Sakai (NC_002695) and EDL933 (NC_002655), both from phylogenetic group ABD; 4 ExPEC isolates, CFT073 (NC_004431), 536 (NC_008253), UTI89 (NC_007946), and APECO1 (NC_008563), all from phylogenetic group B2; and the *Shigella* genomes, *S. boydii* Sb227 (NC_007613), *S. sonnei* Ss046 (NC_007384), *S. dysenteriae* (NC_007606), *S. flexneri* 5–8401 (NC_008258), *S. flexneri* 2a-2457T (NC_004741), and *S. flexneri* 2a-301 (NC_004337). The first 2 genomes are from phylogenetic group B1, with the rest from group ABD.

Phylogenetic Analysis and Extraction of Core Genes. Seven housekeeping gene sequences (*adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, *recA*) 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* (42) to depict an average neutral diversity among the strains. The pairwise *p*-distance matrix was computed using MEGA (43).

Using CFT073 as the reference genome (9), we extracted all annotated protein-coding gene sequence (CDS) regions, excluding pseudogenes. To construct the gene data sets, stand-alone BLAST was performed against the genomes of the remaining 13 strains for homologs (nucleotide sequence identity ≥95%) of each of the CFT073 CDS regions. For each of the 1,516 core genes from CFT073, BLAST was performed against all 14 genomes (including the reference CFT073 genome), and multiple copies in any genome were reported if the subject sequence was at least 97.5% in length and 95% in identity to the query sequence.

Software package PhiPack (44) was used for detecting probable recombination events. This package included 3 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 3 statistics were <0.1 (45).

Molecular Evolutionary Analysis. Rates of synonymous (*dS*) and nonsynonymous (*dN*) mutations for each core gene were calculated by using mutation-fraction method of Nei and Gojobori (16). The details of nonparametric bootstrapping to compute the distribution of *dN*/*dS* are provided in *SI Materials and Methods*.

For each core gene, ZPS (23) was used to perform ClustalW-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. ZP reconstructs an unrooted protein phylogram from the corresponding DNA phylogram, distinguishing 2 categories of protein variants—those encoded by multiple alleles with silent diversity (i.e., evolutionarily long-term variants) and those encoded by 1 allele only (i.e., evolutionarily recent variants). Separate scripts were written to compute the consistency index (29), the frequencies of parallel and coincidental hotspot mutations of different nature (long-term or recent hotspots), the frequency of hotspot mutations for any strain or ecotype, and the sharing of hotspot mutations between any strain pairs and ecotype pairs.

Simulations. Random simulations of mutation were performed on 191 nonrecombinant core genes with hotspots using EvolveAGene 3 (46). The details of the procedure are provided in *SI Materials and Methods*.

Analysis of Functional Category Enrichment. Functional annotation clustering for different sets of candidate genes under positive selection was performed using DAVID (47). For the analysis, ''medium'' classification stringency was used. Annotation clusters with enrichment score >0.5 and *P* value <0.05 were selected as enriched functional categories.

ACKNOWLEDGMENTS. This work was supported by the National Institutes of Health; a Veterans Administration Merit Review grant from the Office of Research and Development, Medical Research Service, Department of Veterans Affairs; and a U.S. Army Medical Research Acquisition Activity under contract W81XWH-05-1-0627.

1. Andersson JO (2005) Lateral gene transfer in eukaryotes. *Cell Mol Life Sci* 262:1182–1197.
2. Ghedin E, et al. (2005) Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* 437:1162–1166.
3. Shastry BS (2007) SNPs in disease gene mapping, medicinal drug development, and evolution. *J Hum Genet* 52:871–880.
4. Gabriel G, et al. (2005) The viral polymerase mediates adaptation of an avian influenza virus to a mammalian host. *Proc Natl Acad Sci USA* 102:18590–18595.
5. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* 2:414–424.
6. Sokurenko EV, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: Gene loss and variation in bacterial pathogens. *Trends Microbiol* 7:191–195.
7. Garcia-Vallve S, Romeu A, Palau J (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res* 10:1719–1725.
8. Rasko DA, et al. (2008) The pangenome structure of *Escherichia coli*: Comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* 190:6881–6893.
9. Welch RA, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* 99:17020–17024.
10. Brzuszkiewicz E, et al. (2006) How to become a uropathogen: Comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. *Proc Natl Acad Sci USA* 103:12879–12884.
11. Lloyd AL, Rasko DA, Mobley HL (2007) Defining genomic islands and uropathogen-specific genes in uropathogenic *Escherichia coli*. *J Bacteriol* 189:3532–3546.
12. Chen SL, et al. (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: A comparative genomics approach. *Proc Natl Acad Sci USA* 103:5977–5982.
13. Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R (2007) Genes under positive selection in *Escherichia coli*. *Genome Res* 17:1336–1343.
14. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
15. Fu YX, Li WH (1993) Statistical tests of neutrality of mutations. *Genetics* 133:693–709.
16. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426.
17. Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917.
18. Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22:2472–2479.
19. Weissman SJ, et al. (2007) Differential stability and trade-off effects of pathoadaptive mutations in the *Escherichia coli* FimH adhesion. *Infect Immun* 75:3548–3555.
20. Denamur E, Matic I (2006) Evolution of mutation rates in bacteria. *Mol Microbiol* 60:820–827.
21. Sokurenko EV, et al. (2004) Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in *Escherichia coli*. *Mol Biol Evol* 21:1373–1383.
22. Chattopadhyay S, et al. (2007) Haplotype diversity in ''source-sink'' dynamics of *Escherichia coli* urovirulence. *J Mol Evol* 64:204–214.
23. Chattopadhyay S, Dykhuizen DE, Sokurenko EV (2007) ZPS: Visualization of recent adaptive evolution of proteins. *BMC Bioinformatics* 8:187.
24. Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 89:167–170.
25. 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.
26. Wirth T, et al. (2006) Sex and virulence in *Escherichia coli*: An evolutionary perspective. *Mol Microbiol* 60:1136–1151.
27. Johnson JR, Owens KL, Clabots CR, Weissman SJ, Cannon SB (2006) Phylogenetic relationships among clonal groups of extraintestinal pathogenic *Escherichia coli* as assessed by multi-locus sequence analysis. *Microbes Infect* 8:1702–1713.
28. Suzuki Y, Gojobori T (2004) Analysis of coding sequences. *The Phylogenetic Handbook—A Practical Approach to DNA and Protein Phylogeny*, eds Salemi M, Vandamme A-M (Cambridge Univ Press, UK), p 298.
29. Kluge A, Farris JS (1969) Quantitative phyletics and the evolution of anurans. *Syst Zool* 18:1–32.
30. Bird A (1999) DNA methylation de novo. *Science* 286:2287–2288.
31. Deitsch KW, Moxon RE, Wellems TE (1997) Shared themes of antigenic variation in bacterial, protozoal, and fungal infections. *Microbiol Mol Biol Rev* 61:281–293.
32. Deszo EL, Steenbergen SM, Freedberg DI, Vimr ER (2005) *Escherichia coli* K1 polysialic acid *O*-acetyltransferase gene, *neuO*, and the mechanism of capsule form variation involving a mobile contingency locus. *Proc Natl Acad Sci USA* 102:5564–5569.
33. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc Natl Acad Sci USA* 105:7899–7906.
34. Wichman HA, Scott LA, Yarber CD, Bull JJ (2000) Experimental evolution recapitulates natural evolution. *Phil Trans R Soc London Ser B* 355:1677–1684.
35. Barlow M, Hall BG (2002) Predicting evolutionary potential: In vitro evolution accurately reproduces natural evolution of the TEM β-lactamase. *Genetics* 160:823–832.
36. Korotkova N, et al. (2007) Selection for functional diversity drives accumulation of point mutations in Dr adhesins of *Escherichia coli*. *Mol Microbiol* 64:180–194.
37. Sokurenko EV, Gomulkiewicz R, Dykhuizen DE (2006) Source-sink dynamics of virulence evolution. *Nat Rev Microbiol* 4:548–555.
38. Holt RD, Hochber ME (2002) Adaptive dynamics of infectious diseases. *Pursuit of Virulence Management*, eds Dieckmann U, Metz JAJ, Sabelis MW, Sigmund K (Cambridge Univ Press, UK), pp 104–119.
39. Anisimova M, Bielawski J, Dunn K, Yang Z (2007) Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes. *BMC Evol Biol* 7:154.
40. Lindgren PK, Karlson Å, Hughes D (2003) Mutation rate and evolution of fluoroquinolone resistance in *Escherichia coli* isolates from patients with urinary tract infections. *Antimicrob Agents Chemother* 47:3222–3232.
41. Roth JR, Kugelberg E, Reams AB, Kofoid E, Andersson DI (2006) Origin of mutations under selection: The adaptive mutation controversy. *Annu Rev Microbiol* 60:477–501.
42. Swofford DL (2000) PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods (Sinauer Associates, Sunderland, MA), Version 4.
43. 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.
44. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172:2665–2681.
45. Chan CX, Beiko RG, Ragan MA (2007) A two-phase strategy for detecting recombination in nucleotide sequences. *South African Comp J* 38:20–27.
46. Hall BG (2008) Simulating DNA coding sequence evolution with EvolveAGene 3. *Mol Biol Evol* 25:688–695.
47. Dennis G, Jr, et al. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 4:P3.