Similarity of genes horizontally acquired by *Escherichia coli* and *Salmonella enterica* is evidence of a supraspecies pangenome

Katherine A. Karberg[a,1], Gary J. Olsen[a,b,c,2], and James J. Davis[a,b,1]

[a]Department of Microbiology, [b]Institute for Genomic Biology, and [c]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Edited by Norman R. Pace, University of Colorado, Boulder, CO, and approved November 2, 2011 (received for review June 14, 2011)

Most bacterial and archaeal genomes contain many genes with little or no similarity to other genes, a property that impedes identification of gene origins. By comparing the codon usage of genes shared among strains (primarily vertically inherited genes) and genes unique to one strain (primarily recently horizontally acquired genes), we found that the plurality of unique genes in *Escherichia coli* and *Salmonella enterica* are much more similar to each other than are their vertically inherited genes. We conclude that *E. coli* and *S. enterica* derive these unique genes from a common source, a supraspecies phylogenetic group that includes the organisms themselves. The phylogenetic range of the sharing appears to include other (but not all) members of the Enterobacteriaceae. We found evidence of similar gene sharing in other bacterial and archaeal taxa. Thus, we conclude that frequent gene exchange, particularly that of genetic novelties, extends well beyond accepted species boundaries.

genome evolution | horizontal gene transfer | microbiome

Microbiologists have long advocated the sequencing of diverse microbial genomes to enhance our understanding of physiology, phylogeny, and evolution. Genome sequences commonly reveal unique genes, even among close relatives. Although it is known that horizontal gene transfer contributes to species differences, strains of the same species can differ by as much 30% of the gene complement (1–3). This finding has led to a perspective in which microbial genomes are composed of a core set of vertically inherited genes that are common throughout the species and a set of variable genes that are acquired horizontally and can be unique to a given strain (4, 5).

Where do the unique genes come from? Two avenues of investigation have shaped our understanding of horizontal gene transfer: the genetic study of recombination of homologous genes among close relatives (bacterial genetics) and the phylogenetic study of nonhomologous transfer of genes from distant relatives (molecular phylogeny). Homologous recombination usually replaces existing genes with related sequences, and so is unlikely to introduce novel genes into a genome. Nonhomologous gene transfers can introduce novel genes, but phylogenetic analyses cannot reveal the sources of these genes when related genes have not been detected in other genomes.

In most genomes, the vertically inherited genes are adapted to codon usages characteristic of their genome and expression level (6, 7). In contrast, horizontally acquired genes often have distinctive base composition [guanosine + cytosine content (G+C)] and codon usage (8, 9), giving rise to an assumption that the transferred genes are from phylogenetically distant and disparate sources (10, 11). However, assuming disparate sources conflicts with the observation that many of the horizontally acquired genes in *E. coli* share a distinctive codon usage (9, 12, 13), suggesting that they come from a common source, possibly the host species themselves (13).

Results

Horizontally Acquired Genes Are Similar in Codon Usage. Seeking insight into the source(s) of horizontally acquired genes, we analyzed the codon usages of genes in five *E. coli* strains and five *S. enterica* strains (*SI Appendix*, *SI Materials and Methods*). Each strain has different pathogenic traits and host ranges, as well as distinctive unique genes. To compare the horizontally acquired and vertically inherited genes, we needed an impartial method for identifying these gene sets in each genome. Given our interest in codon usages, we avoided criteria based on G+C content and/or codon usage, choosing instead criteria based on phylogenetic distribution of orthologous genes. We took the genes shared by all 10 strains as those most likely to have been vertically inherited and the genes unique to a single strain as those most likely to have been recently acquired by horizontal transfer (14) (*SI Appendix*, *SI Text*). These criteria will miss some genes, but the number of false-positives will be very small, and the criteria are not biased by codon usage.

To characterize each set of genes, we computed modal codon usage, a metric less influenced by atypical genes than is the average (15). For shared (i.e., vertically inherited) genes, we used *E. coli* O157:H7 and *S. enterica* Typhimurium LT2 to represent their respective species. Because all unique genes are distinct, we represented each species by the pool of these genes across all five strains. Each modal codon usage (*SI Appendix*, Table S1) is a point in a 59-dimensional space, and the relationships among the codon usages can be characterized by the distances between them (Table 1, upper-right triangle). The distance between the shared gene codon usages of the two species (0.238) was 2.9 times larger than the distance between these species' unique gene codon usages (0.081). Most of this distance between the unique gene codon usages (0.051 ± 0.008) was due to finite sampling (*SI Appendix*, *SI Text*) (15).

We used bootstrap resampling to assess whether this difference in shared gene versus unique gene modal codon usage could be due to statistical error (*SI Appendix*, *SI Text*). Even though such an analysis is expected to result in increased distances between codon usages (*SI Appendix*, *SI Text*), the unique gene codon usages were still 2.3-fold closer than are the shared gene codon usages (Table 1, upper-right triangle, values in parentheses). In all 10,000 resamplings, the unique gene codon usages were more similar than the shared gene codon usages (*SI Appendix*, Fig. S1*A*), and from the distribution of values, we

Author contributions: K.A.K., G.J.O., and J.J.D. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]K.A.K. and J.J.D. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: gary@life.illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1109451108/-/DCSupplemental.

Table 1. Distances between the codon usages of shared and combined unique gene sets for five *E. coli* and five *S. enterica* genomes

Gene set	No. of genes	Average G+C	Codon usage distance to			
			E. coli shared	*S. enterica* shared	*E. coli* unique	*S. enterica* unique
E. coli shared	2,040	0.530	—	**0.238** (0.251 ± 0.009)	0.543 (0.543 ± 0.012)	0.528 (0.529 ± 0.013)
S. enterica shared	2,040	0.547	***0.253*** *(0.255 ± 0.007)*	—	0.646 (0.649 ± 0.012)	0.607 (0.611 ± 0.014)
E. coli unique	4,001	0.486	*0.592 (0.593 ± 0.012)*	*0.670 (0.670 ± 0.013)*	—	**0.081** (0.108 ± 0.012)
S. enterica unique	1,903	0.500	*0.515 (0.516 ± 0.014)*	*0.556 (0.556 ± 0.017)*	***0.142*** *(0.146 ± 0.019)*	—

The shared gene codon usages are those of *E. coli* O157:H7 EDL933 and *S. enterica* subsp. *enterica* Typhimurium LT2. Distances between modal codon usages are in the upper-right triangle of the matrix. Distances between average codon usages are shown in italics in the lower-left triangle. Values in parentheses are the mean ± SD of the corresponding distance measurement for bootstrap resamplings of the gene sets (10,000 and 50,000 replicates for modal and average codon usages, respectively). The comparisons of shared genes to shared genes, and unique genes to unique genes, are shown in bold.

conclude that the difference is significantly greater than 0 ($P \sim 10^{-9}$) (*SI Appendix*, *SI Text*).

Although we consider modal codon usage a superior method for analyzing heterogeneous data, the similarity of the unique gene codon usages is also seen in the average codon usages of the gene sets (Table 1, lower-left triangle). Although the similarity of the unique gene codon usages is less dramatic, the unique genes of the species are significantly more similar in average codon usage compared with the shared genes ($P < 10^{-5}$) (*SI Appendix*, Fig. S1*B*). Thus, we conclude that the unique (presumably, the most recently acquired) genes of *E. coli* and *S. enterica* are more similar in codon usage than are these species' shared (i.e., vertically inherited) genes.

We used projections based on factorial correspondence analysis to display the codon usages of the individual genes in the *E. coli* O157:H7 and *S. enterica* LT2 genomes, as well as the modal usages of these genomes' shared genes and unique genes. The proximity of the unique gene modes relative to the separation of the shared gene modes is evident in the plots created (Fig. 1 and *SI Appendix*, Fig. S2). Qualitatively, the individual shared genes of the two species are offset in the plots, whereas the unique genes are more intermixed.

When the modal codon usages are compared across all strains, the distances between them (*SI Appendix*, Table S2) yield a tree (Fig. 2) that clearly separates unique genes and shared genes. The shared genes are further separated by species, consistent with independent divergence. In contrast, the unique gene codon usages for strains of the two species are interspersed. This interdigitation is not statistical noise due to smaller numbers of genes; when the numbers of shared genes were reduced to match the numbers of unique genes, the shared gene modes always

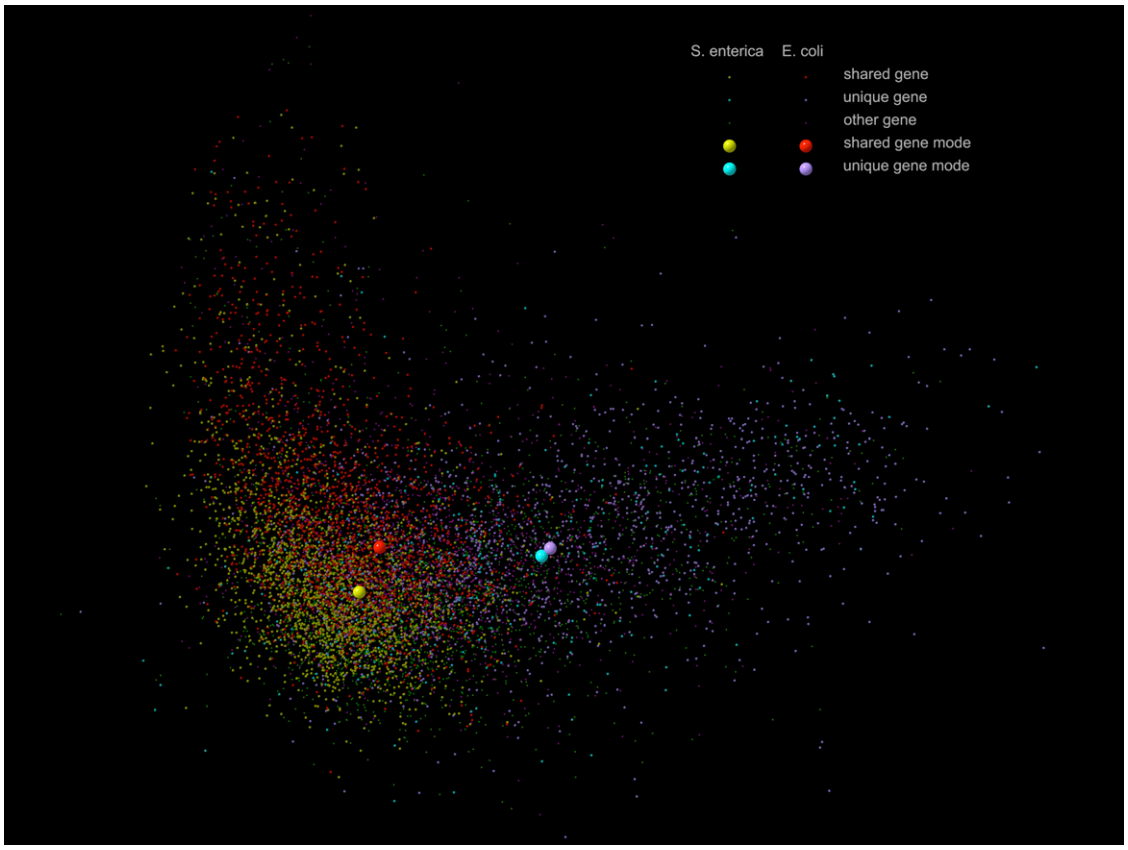

Fig. 1. First two axes of a factorial correspondence analysis of the codon usages of *E. coli* O157:H7 and *S. enterica* LT2 genes. For each species, colors distinguish shared genes, unique genes, and genes with other distributions (i.e., found in between two and nine strains). Also shown are the modal codon usages of the shared and unique genes of the species.

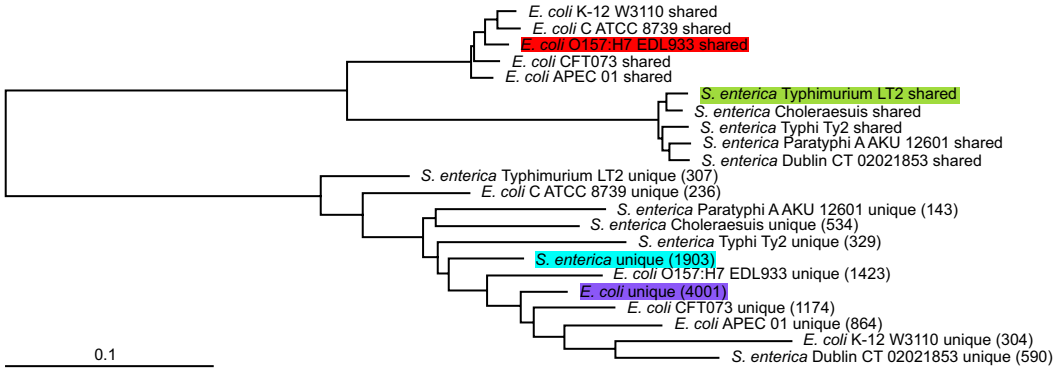

Fig. 2. Tree of *E. coli* and *S. enterica* modal codon usages. The distances between modal codon usages of the shared and unique genes from all five strains of each species (*SI Appendix*, Table S2) were used to construct the tree. There are 2,040 shared genes; the number of unique genes follows each genome name. Background colors are the same as in Fig. 1.

(1,000 replicates) separated by species in the resulting trees (*SI Appendix*, *SI Text*).

The unique genes are dispersed around their respective chromosomes and plasmids (Fig. 3). When the separation of unique genes by at least one intervening shared gene is taken as evidence of independent acquisition, the unique genes of *E. coli* O157:H7 and *S. enterica* LT2 are divided into 172 and 71 distinct regions, respectively (*SI Appendix*, Table S3). Thus, the similarity of the unique genes is not an artifact of a small number of transfers that happen to have sampled similar sources.

Why Are the Unique Gene Codon Usages So Similar? We consider three categories of possible explanations for the similarity in unique gene codon usages: convergence by random drift, acquisition from a common source of genes, and intraorganismal selection. Is it plausible that the unique genes are threefold more similar in codon usage compared with the shared genes because they have independently drifted to a common value? We address this question from two perspectives: whether or not the shared gene codon usages fit a simple drift model, and the statistical difficulty of getting threefold closer.

To postulate that independent drift makes the unique genes much more similar compared with the shared genes, we need to define the endpoint of the presumed drift. Two obvious trends are the drift toward lower G+C content, particularly at third position of the codons, and a more equal use of synonymous codons. These trends have been attributed to drift accompanying the relaxation of translational selection (16–18). However, neither effect is reflected equally across sets of synonymous codons (*SI Appendix*, Tables S1 and S4). The amount of drift toward a presumed unselected codon usage at third codon positions differs greatly among amino acids, and the use of silent site purines and silent site pyrimidines do not extrapolate from one amino acid to another (*SI Appendix*, Table S4). These data are not consistent with a uniform relaxation of codon bias in the unique genes. The complexity of the observed pattern of codon usage leads us to conclude that random drift to some equilibrium value could not have caused the similarity seen in the unique genes of *E. coli* and *S. enterica* (*SI Appendix*, *SI Text*).

Although the foregoing codon usages do not appear to be the result of a simple drift model, we also must consider the possibility that the unique gene modal codon usages are threefold closer by chance. However, this is threefold closer in a multidimensional space, in our case, a space with 41 degrees for freedom. At first approximation, there are 3^{-41} ($\sim 3 \times 10^{-20}$) times as many possible codon usages within a distance of 0.08 (the distance between unique gene modes) as within a distance of 0.24 (the distance between shared gene modes). That is, assuming that all codons are influenced independently by drift, the probability of the unique genes drifting from the shared gene codon usage and converging on codon usages that are threefold closer is less than 1 in 1 quadrillion. Even if the drift were biased, the biases would need to be more similar for the horizontally acquired genes than for the vertically inherited genes (*SI Appendix*, *SI Text*).

We conclude that these data are not consistent with a uniform relaxation of codon bias in the unique genes, or with a random accumulation of neutral mutations. That is, the unique genes are

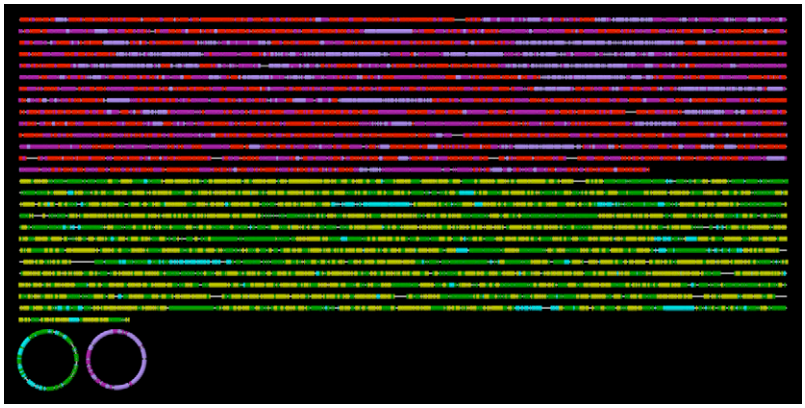

Fig. 3. Interspersion of shared and unique genes on the *E. coli* O157:H7 and *S. enterica* LT2 replicons. Each protein coding sequence is colored by its category (shared, unique, or other) and organism, as in Fig. 1.

not accommodated by the selection-mutation-drift theory of codon usage evolution (17, 18). Any theory that proposes that similarities in *E. coli* and *S. enterica* unique gene codon usages are due to random drift must explain how the unique gene modal codon usages can independently converge on the same tiny subset of codon usage space while maintaining a complex pattern of third codon position base preferences.

The possibility of a common source of unique genes raises the question of where the donor pool resides. It is hard to avoid the conclusion that the reservoir of unique genes is cellular life; although they are essential vectors of transfer, plasmids, phage and naked DNAs do not replicate without a host cell.

Given the dramatic codon usage differences between shared and unique genes in a genome, it has been appealing to suggest that the unique genes come from a phylogenetically distinct source (10, 11). We searched for potential donor source(s) of genes among the complete microbial genomes and sequenced human gut microbiome isolates. Although this search was limited to a methodology that is applicable to all genomes (*SI Appendix*, *SI Text*), the best potential sources of these genes appear to be the nonnative genes of *E. coli*, *S. enterica*, and other Enterobacteriaceae, including *Citrobacter*, *Cronobacter*, *Enterobacter*, *Pectobacterium*, and *Shigella* (*SI Appendix*, Tables S5 and S6 and Figs. S3 and S4). This codon usage similarity does not span all Enterobacteriaceae, however. In particular, the distances from the *Yersinia* unique gene modal codon usage to those of *S. enterica* and *E. coli* unique genes (0.307 and 0.329, respectively) are fourfold greater than the distance between the *E. coli* and *S. enterica* unique genes (0.081) (*SI Appendix*, Fig. S5), and nonnative gene codon usages of *Yersinia* spp. do not match as many *E. coli* and *S. enterica* unique genes as do the nonnative gene usages of the other aforementioned Enterobacteriaceae (*SI Appendix*, Tables S5 and S6 and Figs. S3 and S4).

It would be difficult to explain the similar codon usages by selection in gene transfer or maintenance. The unique genes in these genomes are products of phage-, plasmid-, and transposon-mediated transfers. We are unaware of any evidence indicating that these gene transfer mechanisms select for a specific codon usage. Indeed, the mosaic nature of many mobile elements demonstrates that these mechanisms tolerate different codon usages (15, 19). Similarly, we are unaware of any integration mechanism that selects for a particular codon usage.

To persist, an integrated gene must not be harmful. Accordingly, a striking property of the unique genes in *E. coli* and *S. enterica* is that they are not random samples of an organismal genome; they almost entirely lack paralogs of universal and highly conserved genes (*SI Appendix*, *SI Text*), a property that previous studies have attributed to toxicity in a recipient (20). The nearly complete absence of paralogs of core genes also may suggest that some form of punctuation (e.g., unidentified recombination sites) distinguishes DNA regions that are most successfully transferred from those that are less successfully transferred. Avoiding toxicity may select for lower G+C content via proteins like H-NS, which nonspecifically repress expression of low G+C genes (21, 22), but there is no evidence suggesting that this is related to codon selection per se (21, 22). Any resulting reduction in G+C content minimally constrains codon usage; we found comparable codon usage diversity within genera spanning 35–65% genomic G+C content (*SI Appendix*, Table S7 and Fig. S6).

We looked for a possible codon usage convergence of *E. coli* and *S. enterica* gene sequences with lower G+C content. When genes are drawn from a common pool, G+C content has a small (but finite) influence on codon usage distances, but the divergence between *E. coli* and *S. enterica* is much larger (*SI Appendix*, Fig. S7), particularly at higher G+C content. We also repeated the analyses presented in Table 1, but limited to genes with 52% ± 2% G+C content (*SI Appendix*, Table S8). The unique genes were twice as close in codon usage compared with the shared genes.

Long-term persistence of a gene requires replication, repair, and occasional usefulness. Several authors have concluded that recently acquired genes have higher substitution rates as they adapt to their host genome (23–25); however, such selection would be expected to increase the variation in synonymous codon usage, not to lead to convergence on a common value in distinct species. The best-documented phenomenon affecting codon usage of acquired genes is amelioration (26); however, amelioration does not explain the distinctiveness of unique genes from the host codon usage or the extreme similarity of these genes between species.

Thus, we propose that a plurality of the unique genes in *E. coli* and *S. enterica* genomes have similar codon usages, because they are drawn from a common biological reservoir that extends further than previously suggested (4, 27), going far beyond the phylogenetic range of homologous recombination (28, 29) and crossing species boundaries. Although alternatives are possible, they would require selecting the same codon usage for the genes acquired by two species, while allowing the codon usages of their vertically inherited genes to diverge (*SI Appendix*, *SI Text*).

Related Observations in Other Taxa. This phenomenon, the similar codon usages of horizontally acquired genes in related species, may be phylogenetically widespread. A comparison of *Agrobacterium* species revealed that the modal codon usages of their plasmids, which have very different gene contents, are more similar (distances of 0.061–0.139) than the modal codon usages of their chromosomes (distances of 0.209–0.392) (*SI Appendix*, Table S9). Moreover, a comparison of the Archaea *Methanosarcina acetivorans* and *Methanosarcina mazei* found that the unique gene modes are closer than the shared gene modes (*SI Appendix*, Table S10). However, this trend is less consistent with the more distant species *M. barkeri* (*SI Appendix*, Table S10), and we did not find similarity in the unique gene codon usages among strains of *Bacillus* (*cereus* subgroup), *Streptococcus*, and *Sulfolobus*. Whether this finding is related to limitations in strain sampling, to noise due to small numbers of unique genes, or to a true lack of codon usage similarity in the unique genes is unclear.

Discussion

Our data indicating that a plurality of unique genes in *E. coli* and *S. enterica* are nearly indistinguishable in codon usage are not easily reconciled with random drift or uptake from distant phylogenetic sources (10, 11), but are more consistent with the concept of drawing on a common gene pool. These findings call into question both traditional and contemporary ideas of microbial species. For example, the pangenome concept posits that members of a species are composed of a shared set of core genes and a collection of variable genes (the pangenome) present in some, but not all, members of the species (4, 27). This concept does not exclude DNA acquisition from more distantly related donors, but does propose that exchanges of a phylogenetically circumscribed gene pool are the primary basis of diversity in a species. Although conceptually in accordance with this pangenome concept, our data suggest that frequent exchange extends beyond a biologically meaningful definition of species. The distinctive codon usage of the exchanged genes presumably results from a complex history of genomic environments during passage through a series of hosts, none of which retain the genes long enough for them to ameliorate to an individual host codon usage (11, 26). Although homologous recombination has a profound influence in close relatives, and some genes are transferred across vast phylogenetic distances, we are now defining an intermediate range over which transfer appears to be rampant, creating a superspecies pangenome.

Materials and Methods

Sequence Data. The genomes analyzed and the steps in their retrieval are described in detail in *SI Appendix, SI Materials and Methods*.

Identification of Shared and Unique Genes. Genes in two genomes were considered orthologous (and hence shared) if they were found to be bidirectional best hits using BLASTP (30). Two genes were considered bidirectional best hits if they were each other's best match between the two genomes being compared, had at least 80% amino acid sequence identity, and matched over at least 80% of the protein length.

For each *E. coli* and *S. enterica* genome, the shared gene set comprised the 2,040 genes for which bidirectional best hits identified presumed orthologs in all 10 genomes. Genes were defined as unique if they did not have a bidirectional best hit in any of the nine other genomes. The numbers of unique genes in each genome are shown in Fig. 2. Because the unique genes of each strain are distinct, the unique gene modal codon usages (below) of each species were defined by combining the unique genes from all five strains, giving a total of 4,001 *E. coli* unique genes and 1,903 *S. enterica* unique genes. For *Yersinia*, shared genes were defined as those linked by bidirectional best hits across the 10 strains of *Y. pestis* and *Y. pseudotuberculosis* analyzed, and unique genes as those lacking bidirectional best hits in any of the other nine *Yersinia* genomes. Shared and unique genes among the three *Methanosarcina* genes were identified as described for *E. coli* and *S. enterica*, except here the BLASTP matches required at least 70% amino acid sequence identity and covered at least 70% of the protein length.

Codon Usage Analyses. Most of the analyses in this study are based on modal codon usage (15). Analogous to a mode in statistics, modal codon usage is the expected codon usage frequencies that match the largest number of genes in a set of genes (with matching meaning that the gene is not significantly different; $P < 0.1$) (15). Relative to average codon usage, modal codon usage minimizes the effects of genes with aberrant codon usages. Native codon usage uses an axis to accommodate expression-related variation in codon usage (31). A gene that is significantly different ($P < 0.1$) from all points on the native codon usage axis is classified as nonnative.

Distances between codon usages were calculated as described previously (15). The uncertainty in distances and the significance of differences in distances were evaluated using bootstrap analyses (32) in which one replicate is composed of a resampling of the 4,001 *E. coli* unique genes, a resampling of the 1,903 *S. enterica* unique genes, and a resampling of the 2,040 orthologous pairs of *E. coli* O157:H7 and *S. enterica* LT2 shared genes. To test whether the distance between shared gene codon usages is significantly greater than the distance between unique gene codon usages, the distribution of the difference in distances among the bootstrap samples was examined (*SI Appendix, SI Text*).

A Monte Carlo simulation was used to assess the expected difference in codon usage of samples from a common pool. The two gene sets compared randomly redistributed into groups of the same size as the original sets, the modal codon usages were computed for the new groups, and the distances between the modes were computed. Values reported are mean ± SD of results from 10 randomizations.

For trees of codon usages, pairwise distances were converted to a corresponding tree using the neighbor-joining method (33), as implemented in the neighbor program of the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html) (34).

Factorial correspondence analysis of relative synonymous codon usage (i.e., codon usage normalized per amino acid) was computed using CODONW (35). Factorial correspondence analysis and genome drawings were rendered using POV-Ray (http://www.povray.org/). All symbols are represented as spheres at a common depth, so that gene symbols can overlap without fully obscuring one another.

Interspersion of Unique and Shared Genes. For the genomes of *E. coli* O157:H7 and *S. enterica* Typhimurium LT2, the number of distinct regions with one or more unique genes separated by a minimum number of shared genes was tabulated, with the required number of shared genes varying from 1 to 10. To qualify as a delimiter, the shared genes could have any unique genes interspersed.

Possible Source(s) of Unique Genes. For each of (*i*) the complete bacterial and archaeal genome in the SEED database (36) (accessed using the Web services API; ref. 37), (*ii*) the genomes from the human gut microbiome project (http://genome.wustl.edu/pub/organism/Microbes/Human_Microbiome_Project/GI_Tract/) (38), and (*iii*) 17 additional enterobacterial genomes from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) (39), the modal codon usages of the genome and of its nonnative genes were determined. For each of these codon usages, the unique genes in *E. coli* O157:H7 and *S. enterica* Typhimurium LT2 that are not significantly different ($P < 0.1$) were identified, and the fraction of all unique gene codons in the matching sets (i.e., the number of matching genes weighted by their length) was calculated.

Effect of G+C Content on Codon Usage Divergence. All of the genomes from the SEED database (36) that have more than one species within the same genus were analyzed. Then the modal codon usage for the genomes of each individual species was calculated, and the distances between the modal codon usages of each species within the genus were measured. In cases where multiple strains of the same species were available, the median distance of all pairs of strains is reported. In cases where more than two species of a genus were available, the distance between all pairs of species was measured, and the median distance, average distance, and rms distance are reported.

***Agrobacterium* Chromosomes and Plasmids.** For comparisons of *Agrobacterium* species, the chromosomes of each species were used to represent vertically inherited genes, and the plasmids were used to represent recently horizontally acquired genes. For each genome, the chromosomal genes and the plasmid genes were pooled separately, and the modal codon usage for each set was computed. For this, the 2.65-Mbp replicon of *A. radiobacter* K84 was considered a chromosome.

ACKNOWLEDGMENTS. We thank Dr. Claudia Reich for her helpful suggestions. Portions of this work were supported by National Aeronautics and Space Administration Grant NAG 5-12334 (issued through the Exobiology Program), Department of Energy Grant FG02-01ER63146, and National Institutes of Health Contract HHSN266200400042C (via a subcontract from the University of Chicago). J.J.D. acknowledges support from the Institute for Genomic Biology Fellows Program.

1. Perna NT, et al. (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409:529–533.
2. Siew N, Fischer D (2003) Twenty thousand ORFan microbial protein families for the biologist? *Structure* 11:7–9.
3. Fischer D, Eisenberg D (1999) Finding families for genomic ORFans. *Bioinformatics* 15: 759–762.
4. Tettelin H, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." *Proc Natl Acad Sci USA* 102:13950–13955.
5. 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.
6. Grantham R, Gautier C, Gouy M, Mercier R, Pavé A (1980) Codon catalog usage and the genome hypothesis. *Nucleic Acids Res* 8:r49–r62.
7. Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15: 1281–1295.
8. Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: The genealogy of ORFans in *E. coli*. *Genome Res* 14:1036–1042.
9. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* 222:851–856.
10. Ochman H, Lerat E, Daubin V (2005) Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci USA* 102(Suppl 1):6595–6599.
11. van Passel MWJ, Marri PR, Ochman H (2008) The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLOS Comput Biol* 4:e1000059.
12. Badger JH (1999) Exploration of microbial genomic sequences via comparative analysis. PhD dissertation (Univ of Illinois at Urbana-Champaign, Urbana, IL), pp 45–92.
13. Daubin V, Lerat E, Perrière G (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol* 4:R57.
14. Smith MW, Feng DF, Doolittle RF (1992) Evolution by acquisition: The case for horizontal gene transfers. *Trends Biochem Sci* 17:489–493.
15. Davis JJ, Olsen GJ (2010) Modal codon usage: Assessing the typical codon usage of a genome. *Mol Biol Evol* 27:800–810.
16. Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol* 146:1–21.
17. Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. *J Mol Evol* 24:28–38.
18. Sharp PM, Li W (1986) Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for "rare" codons. *Nucleic Acids Res* 19:7737–7749.
19. Schlesinger DJ, Shoemaker NB, Salyers AA (2007) Integration and excision of a *Bacteroides* conjugative transposon, CTnDOT. *Appl Environ Microbiol* 73:4226–4233.

20. Sorek R, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318:1449–1452.
21. Navarre WW, et al. (2006) Selective silencing of foreign DNA with low GC content by the H-NS protein in S*almonella*. *Science* 313:236–238.
22. Dorman CJ (2007) H-NS, the genome sentinel. *Nat Rev Microbiol* 5:157–161.
23. Hao W, Golding GB (2006) The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome Res* 16:636–643.
24. Kuo CH, Ochman H (2009) The fate of new bacterial genes. *FEMS Microbiol Rev* 33: 38–43.
25. Davids W, Zhang Z (2008) The impact of horizontal gene transfer in shaping operons and protein interaction networks—direct evidence of preferential attachment. *BMC Evol Biol* 8:23.
26. Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: Rates of change and exchange. *J Mol Evol* 44:383–397.
27. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pangenome. *Curr Opin Genet Dev* 15:589–594.
28. Baron LS, Gemski P, Jr., Johnson EM, Wohlhieter JA (1968) Intergeneric bacterial matings. *Bacteriol Rev* 32:362–369.
29. Rayssiguier C, Thaler DS, Radman M (1989) The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* 342:396–401.
30. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
31. Davis JJ, Olsen GJ (2010) Characterizing the native codon usage of a genome: An axis projection approach. *Mol Biol Evol* 28:211–221.
32. Efron B (1979) Bootstrap methods: Another look at the jackknife. *Ann Stat* 7:1–26.
33. Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425.
34. Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* 5: 164–166.
35. Peden JF (1999) Analysis of codon usage. PhD dissertation (Univ of Nottingham, Nottingham, UK), pp 50–102.
36. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702.
37. Disz T, et al. (2010) Accessing the SEED genome databases via Web services API: Tools for programmers. *BMC Bioinformatics* 11:319.
38. Nelson KE, et al.; Human Microbiome Jumpstart Reference Strains Consortium (2010) A catalog of reference genomes from the human microbiome. *Science* 328: 994–999.
39. Wheeler DL, et al. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 35(Database issue):D5–D12.