Codon Bias and Base Composition Are Poor Indicators of Horizontally Transferred Genes

Liisa B. Koski, Richard A. Morton, and G. Brian Golding

Department of Biology, McMaster University, Hamilton, Ontario, Canada

Horizontal gene transfer is now recognized as an important mechanism of evolution. Several methods to detect horizontally transferred genes have been suggested. These methods are based on either nucleotide composition or the failure to find a similar gene in closely related species. Genes that evolve vertically between closely related species can be divided into those that retain homologous chromosomal positions (positional orthologs) and those that do not. By comparing open reading frames in the *Escherichia coli* and *Salmonella typhi* genomes, we identified 2,728 positional orthologs since these species split 100 MYA. A group of 1,144 novel *E. coli* genes were unusually diverged from their *S. typhi* counterparts. These novel genes included those that had been horizontally transferred into *E. coli,* as well as members of gene pairs that had been rearranged or deleted. Positional orthologs were used to investigate compositional methods of identifying horizontally transferred genes. A large number of *E. coli* genes with normal nucleotide composition have no apparent ortholog in *S. typhi,* and many genes of atypical composition do, in fact, have positional orthologs. A phylogenetic approach was employed to confirm selected examples of horizontal transmission among the novel groups of genes. Our analysis of 80 *E. coli* genes determined that a number of genes previously classified as horizontally transferred based on base composition and codon bias were native, and genes previously classified as native appeared to be horizontally transferred. Hence, atypical nucleotide composition alone is not a reliable indicator of horizontal transmission.

Introduction

To understand how new genetic functions arise and to reconstruct the history of genome evolution, it will be necessary to unravel the mosaic nature of genomes. Several methods have been suggested to identify horizontally derived genes. Comparison of phylogenetic trees among individual genes allows those that have unusual origins to be recognized (Smith et al. 1992). Such an approach is very powerful but requires extensive sequence information for many closely related species. The genomes of most organisms, including bacteria, contain many paralogs, and unless sequences for entire genomes are available, incorrect conclusions may be made by comparing nonorthologous genes. Combining sequence similarity with chromosome position is another method for identifying horizontally introduced genes (Huynen and Bork 1998). Genes that diverged by vertical evolution will be found at homologous positions in closely related organisms, while those derived by introgression will lack a positional ortholog. Genetic rearrangements such as duplications, transpositions, and inversions complicate such comparisons.

DNA sequence analysis has also provided valuable clues about horizontal transfer events. Genome (A+T) content, dinucleotide frequencies, and synonymous codon usage (Lawrence and Ochman 1997) vary among organisms and are generally characteristic of evolutionary lineages. Several methods have been suggested to use these data to identify horizontally transferred genes. (Ochman and Lawrence 1996; Lawrence and Ochman 1998) used anomalous GC content at first and third codon positions together with synonymous codon usage, positional homology, and BLAST hits to analyze genes of the *Escherichia coli* strain MG1655 genome, suggesting that a minimum 17.6% of identified open reading frames (ORFs) had arisen via horizontal transfer since separation of the Escherichia and Salmonella lineages about 100 MYA. ORFs were initially identified as atypical if their GC contents at first and third codon positions were two or more standard errors (SE) higher or lower than the respective means for all genes in the genome. Chi-square (χ^2) of codon usage and the codon adaptation index (CAI) were also calculated for each gene. CAI is a measure of similarity of a gene's synonymous codon usage to that of a standard set of highly expressed genes for that organism (Sharp and Li 1987). Those genes with a high χ^2 and a low CAI were classified as atypical. From this list of atypical genes, known native genes that exhibit atypical base compositions were eliminated for other reasons, such as the amino acid content of the encoded protein. Lawrence and Ochman 1997 also estimated the time of introgression for each of the horizontally transferred genes found in *E. coli.* Transferred genes are subject to those mutational processes affecting the recipient genome. Amelioration is the process by which the acquired gene incurs substitutions and evolves to reflect the DNA composition of the new genome (Lawrence and Ochman 1998). Lawrence and Ochman estimated the time of introgression by examining the rate and extent of amelioration of the introgressed genes and determining how long each sequence had been subjected to the directional mutational pressures of the recipient genome. They estimated that many of the genes were introduced within the last 10 Myr, with the average time of introduction being 25.3 MYA. Médigue et al. 1991 used a chi-square distance measure to divide *E. coli* genes into three classes: genes of high expression, genes of low expression, and a third class containing horizontally transferred genes. Mrázek and Karlin 1999 used a measure assess-

Key words: horizontal gene transfer, *Escherichia coli,* codon bias

Address for correspondence and reprints: G. Brian Golding, Department of Biology, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4K1. E-mail: golding@mcmaster.ca.

Mol. Biol. Evol. 18(3):404–412. 2001

ing the bias of one group of genes against a second group in order to identify alien genes in the genomes of several bacteria.

It is important to assess the validity of nucleotide composition and synonymous codon usage as a measure to detect horizontally transferred genes. Introgressions may be undetected by these methods, since genes from closely related organisms may not have unusual nucleotide composition or codon bias. Also, genes that have been in the genome for a long period will have undergone amelioration. Therefore, the earliest genes introduced into the genome have probably fully ameliorated and will go undetected when using base composition and codon bias as a means of identification. Perhaps more importantly, the forces that shape normal compositional variation among genes within a genome are not well understood. *Escherichia coli* genes with a high CAI have a strong correlation to highly expressed genes (Sharp and Li 1987) and use a set of preferred codons that correspond to the tRNA molecules found in rapidly growing cells (Ikemura 1981). There may be other mutational or selectional forces that cause deviations of nucleotide composition from the genome average that might lead to a misclassification of genes as horizontally transferred.

A comparison of genes between closely related organisms can be used to evaluate methods of gene classification, as well as to identify genes that have atypical modes of evolution. The genomes of *E. coli* MG1655 (Blattner et al. 1997) and *S. typhi* (*Salmonella enterica* serovar *Typhi*) have been completely sequenced. These species diverged about 100 MYA (Doolittle et al. 1996) and have essentially colinear genetic maps (Krawiec and Riley 1990). We used protein similarity and conservation of local gene order to identify a group of 2728 *E. coli* genes that have positional orthologs in *S. typhi*. Phylogenetic analysis of selected examples of 1,144 ‘‘novel’’ *E. coli* genes suggested that on the order of 10%–15% of the *E. coli* genome may have been horizontally introduced. We also found that atypical nucleotide composition alone was not a reliable indicator of horizontal transmission.

Materials and Methods

Escherichia coli ORFs were obtained from the NCBI site (ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria/Ecoli/faa), along with the nucleotide coding sequences (/ffn). *Salmonella typhi* nucleotide sequences were produced by the *S. typhi* Sequencing Group at the Sanger Centre and obtained from ftp://ftp.sanger.ac.uk/pub/pathogens/st/ST.dbs (November 8, 1999). The single most similar *S. typhi* ORF for each *E. coli* ORF was obtained by TBLASTN (Altschul et al. 1990) using an expected value cutoff of 10.0. Each TBLASTN result produced an inferred *S. typhi* amino acid sequence and an alignment with the *E. coli* query. The length of the alignment generated by TBLASTN was noted, as it gave an indication of how extensive the similarity between the sequences might be. For example, a significant hit extending over just a few amino acids might indicate a similar motif but not the identification of an ortholog. The alignment lengths were used to further evaluate the similarity between genes. A protein distance between the two aligned genes was obtained from the PROTDIST program of PHYLIP, version 3.5 (Felsenstein 1994). PROTDIST distances estimate the expected fraction of amino acids changed between the two aligned amino acid sequences according to the PAM250 model of amino acid substitution. Each *E. coli* ORF was classified as ‘‘positionally conserved’’ as further described below. Since the *S. typhi* sequence used was not yet annotated, we used a neighbor method similar to that described by Huynen and Bork (1998) in which the nucleotide positions of adjacent *E. coli* ORFs and their *S. typhi* hits were compared in order to determine those (orphans) which were found in inconsistent chromosomal locations. When two adjacent *E. coli* ORFs hit the same *S. typhi* sequence, the most closely related was taken as the positional hit, and the other was taken as the orphan. Positional and orphan hits were further subdivided as described below. The list of *E. coli* MG1655 ORFs previously categorized as horizontally transferred (Lawrence and Ochman 1998) was obtained from ftp://ftp.pitt.edu/dept/biology/lawrence/ for comparison with this classification.

The fraction of (A+T) nucleotides in the first and third codon positions was calculated from the *E. coli* ORF’s coding sequence. Bivariate frequency distributions for all *E. coli* ORFs and those positionally conserved were compared using two-way classification and the odds ratio (fraction of positionally conserved ORFs in a cell divided by the fraction of all ORFs in a cell). The significance of this ratio was assessed with a *G*-test corrected for continuity and a Bonferroni correction for multiple tests.

CAI was calculated for each *E. coli* ORF according to the method of Sharp and Li (1987) using as a reference set of highly expressed *E. coli* genes the 27 genes used by Sharp and Li (1986). Since CAI depends on amino acid composition, CAI was also corrected by subtracting the value that would be obtained for a protein of the same composition which used synonymous codons according to the genome cumulative average. These deviations were divided by the genome standard deviation to give a Z_{CAI} score. Since these scores were distributed asymmetrically, extreme scores were assessed empirically.

An NCBI BLAST search was conducted on the amino acid sequence of each *E. coli* ORF. Sequences that hit to eight or more different species with blast expected values of less than 10^{-15} were used for phylogenetic analysis. Phylogenies were generated where possible for a number of novel *E. coli* ORFs, as well as for those genes previously classified as horizontally transferred by Lawrence and Ochman 1998 based on base composition and codon bias. Sequences were aligned using the CLUSTAL W algorithm (Thompson, Higgins, and Gibson 1994). For each gene, 100 bootstrap samples were constructed using SEQBOOT (Felsenstein 1994) and 100 phylogenetic trees were generated using the neighbor-joining algorithm (Saitou and Nei 1987). Us-

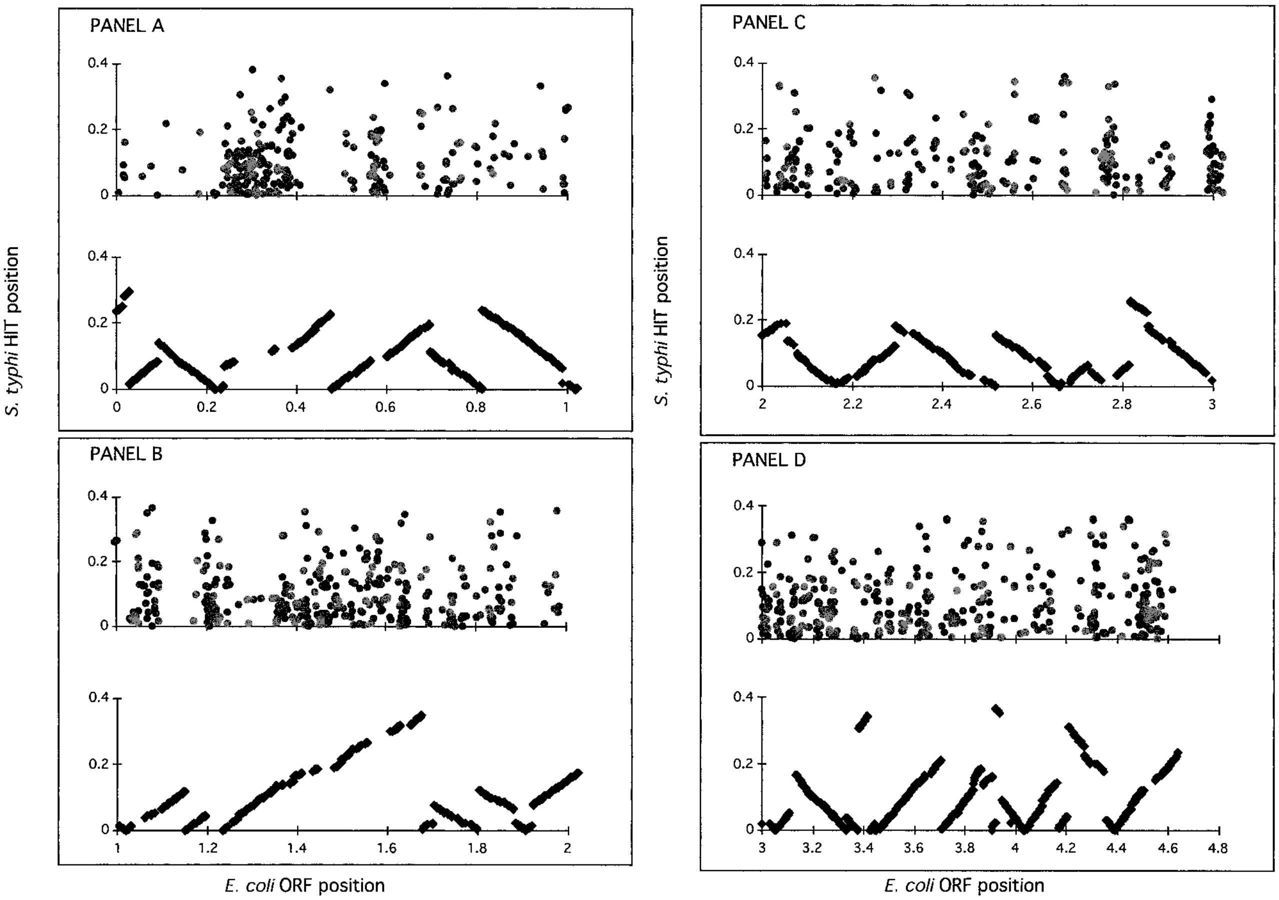


FIG. 1.—Locations of most similar TBLASTN hits on relative chromosomal maps of *Escherichia coli* and *Salmonella typhi.* The position of the *E. coli* ORF is the starting nucleotide from the Blattner et al. (1997) annotation. The *S. typhi* position is the first nucleotide of the TBLASTN hit. No correction has been made for contigs; therefore, *S. typhi* numbering is relative within each contig. *Escherichia coli*'s position is separated into four panels, 0–1, 1–2, 2–3, 3–4.6 million nucleotides, as given on the x-axis. The y-axis gives *S. typhi*'s position. *Escherichia coli* open reading frames classified as positionally conserved are at the bottom of each panel (squares), and orphans are at the top (circles).

ing the original sequence alignment, these 100 trees were then evaluated by the maximum-likelihood inference program PROTML from the MOLPHY package, version 2.2 (Adachi and Hasegawa 1992). If the maximum-likelihood tree showed evidence of horizontal transfer, a new tree was constructed. In this tree, *E. coli* was forced to be located next to Salmonella (or another closely related Gram-negative species). This rearranged tree represented a null hypothesis without horizontal transfer. If the likelihood of the rearranged tree was significantly below the likelihood of the original tree (Kishino and Hasegawa 1989), this hypothesis was rejected, and the gene was classified as ‘‘horizontally transferred.’’ If the best likelihood tree grouped *E. coli* with other closely related Gram-negative bacteria and the gene was a positional ortholog, we classified the gene as ‘‘native.’’

Results and Discussion

Identification of Vertically Transmitted Genes

The single most similar *S. typhi* ORF for each *E. coli* ORF was identified using TBLASTN (Altschul et al. 1990). These nearest-neighbor pairs were used to produce a dot plot showing the positional relationship between protein-coding genes in these two genomes (fig. 1). This figure is analogous to a protein sequence dot plot, but only the locations of those ORFs having the best TBLASTN scores are plotted. The majority of ORFs fall along continuous lines identifying genes that are candidates for vertical transmission since the divergence of *E. coli* and *S. typhi.* Such *E. coli* genes are termed ‘‘positionally conserved’’ in the *S. typhi* genome (fig. 1, bottom half of each panel; squares). The remaining *E. coli* genes are termed ‘‘orphans’’ (fig. 1, top half of each panel; circles).

Out of 4,289 *E. coli* ORFs, 2,951 (68.8%) were classified as positionally conserved. The advantage of determining positional conservation is that it enables identification of pairs of orthologous genes. The number of *E. coli* genes that have orthologs in *S. typhi* will be larger than the number that are positionally conserved, since the rearrangement of small groups of genes may result in their classification as orphans. There are a number of other reasons why an *E. coli* gene might be an orphan. The *S. typhi* ortholog may have been deleted, causing the *E. coli* gene to become paired with the next most similar *S. typhi* gene but located at a disparate position. Alternatively, the *E. coli* gene may have been

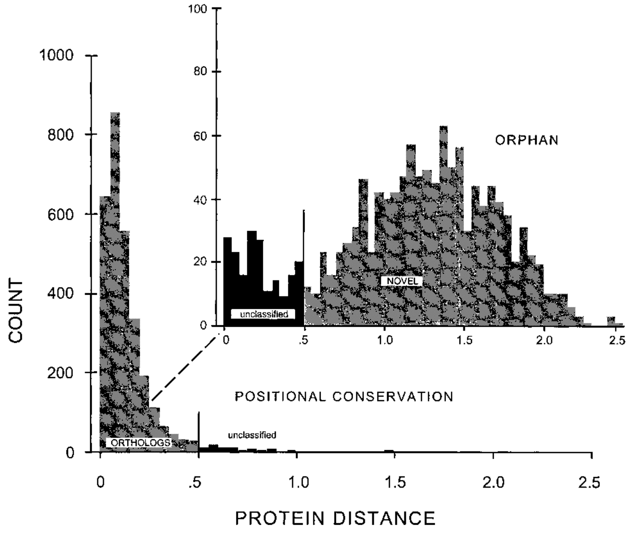


FIG. 2.—Classification of *Escherichia coli* protein-coding open reading frames (ORFs) by positional conservation and by protein distance from *Salmonella typhi.* The front panel shows the distribution of those ORFs with an *S. typhi* TBLASTN hit at a positionally consistent location ($N = 2{,}951$). The rear panel shows the orphan ORFs ($N = 1{,}309$). Both groups are arbitrarily divided at a protein distance of 0.5, forming a set of positional orthologs and novel ORFs without an *S. typhi* counterpart.

horizontally transferred into the genome, becoming paired with a related *S. typhi* gene. Only when a very similar gene replaces an *E. coli* gene at exactly the same position (homologous replacement) will an introgressed gene have a positionally conserved *S. typhi* hit. Orphans were not distributed randomly across the *E. coli* chromosome, indicating regions that have been active for introgression or deletion. For example, a large cluster of orphans was found between approximately 0.2×10^6 and 0.4×10^6 nucleotides corresponding to ‘‘loop1’’ as defined by Riley and Krawiec (1987), a region in the *E. coli* genome that is missing from *S. typhimurium.*

Positionally conserved *E. coli* ORFs are generally less diverged than orphans. Over 90% have less than 0.3 expected amino acid substitutions per site (fig. 2) (mean = 0.142, SD = 0.166).These protein distances are comparable to the average of 0.04 nonsynonymous nucleotide substitutions per nonsynonymous site found by Sharp (1991) for a set of 67 orthologous *E. coli* and *S. typhi* genes. There are a few positionally conserved *E. coli* genes that are unusually distant (more than one amino acid substitution per site) from their *S. typhi* homologs. These may represent homologous replacements or a subset of genes that have diverged unusually rapidly. To distinguish between these possibilities, each gene must be examined individually by comparison with orthologs in closely related species. Positionally conserved genes that aligned over a length of more than 90% of their sequence and had a protein distance of less than 0.5 substitutions per site from their *S. typhi* hit were termed ‘‘positional orthologs’’ (fig. 2 and table 1). $D < 0.5$ was arbitrarily chosen as the cutoff because there were few positionally conserved genes with $D > 0.5$, and for the orphan genes, 0.5 fell between two modes of the protein distance distribution (fig. 2). This subset of 2,728 *E. coli* genes is a conservative estimate of the

Table 1
CAI and (A + T) Content of *Escherichia coli* Open Reading Frames (ORFs)

Classification Group	Number	Positionally Conserved	Fraction[a]	Odds Ratio[b]
All genes	4,289	2,951	0.688	1.0
$(A + T)_3 > 0.50$	976	437	0.448	0.65
$(A + T)_3 > 0.55$	529	163	0.308	0.45
$(A + T)_3 > 0.65$	127	18	0.142	0.21
$(A + T)_3 > 0.70$	56	5	0.089	0.13
$(A + T)_3 < 0.30$	70	26	0.371	0.54
$0.35 < (A + T)_3 < 0.50$	2,967	2,300	0.775	1.13
$(A + T)_1 > 0.50$	428	159	0.371	0.54
$(A + T)_1 > 0.55$	164	39	0.238	0.35
$(A + T)_1 > 0.65$	21	7	0.333	0.48
$(A + T)_1 < 0.30$	35	29	0.829	1.20
$(A + T)_1 < 0.35$	498	420	0.843	1.23
$(A + T)_1 > 0.50$ and $(A + T)_3 > 0.50$	298	76	0.255	0.37
$(A + T)_1 > 0.55$ and $(A + T)_3 > 0.65$	77	8	0.104	0.15
$0.35 < (A + T)_1 < 0.50$ and $0.30 < (A + T)_3 < 0.50$	2,632	2,001	0.760	1.11
CAI > 0.45	476	444	0.933	1.36
CAI > 0.725	31	31	1.000	1.45
CAI < 0.3	816	342	0.419	0.61
CAI < 0.175	73	14	0.192	0.28
0.30 < CAI < 0.45	1,895	1,465	0.773	1.12
$(A + T)_1 > 0.50$ and $(A + T)_3 > 0.50$ and CAI < 0.2	102	13	0.127	0.19
$(A + T)_1 > 0.55$ and $(A + T)_3 > 0.65$ and CAI < 0.2	49	2	0.041	0.06
$Z_{CAI} < -1.0325$	429	112	0.261	0.38
$Z_{CAI} < -1.476$	100	15	0.150	0.22

[a] Fraction of all *E. coli* genes in this class which are positionally conserved. This is an estimate of the probability of misclassification if all genes in this compositional range were horizontally transferred into *E. coli.*

[b] Fraction of positionally conserved ORFs in the class divided by fraction of all *E. coli* genes which are positionally conserved (0.688). Values greater than 1 indicate that there is an increased chance of finding a positionally conserved ORF in the class, while values less than 1 indicate that there is a decreased chance.

number of orthologs that have retained chromosomal order since *E. coli* and *S. typhi* separated. Of the *E. coli* genes previously classified as horizontally transferred (Lawrence and Ochman 1998), 18% (135/747) have a positional ortholog in *S. typhi*, an unexpected number if the average time of introgression is 25.3 MYA (Lawrence and Ochman 1997).

Identification of Unusually Diverged *E. coli* Genes

There were 1,309 *E. coli* ORFs for which the most similar *S. typhi* TBLASTN hit is at a disparate chromosomal position, and there were 29 which had no hit below the expected value cutoff of 10.0. In contrast to positionally conserved genes, orphans were clearly divided into several subgroups. Most *E. coli* orphans were distantly related ($D > 0.5$) to their *S. typhi* counterparts, indicating a mode of evolution different from that of positionally conserved genes (fig. 2). Many of these pairs were simply different members of the same protein family or shared functional motifs. These orphans could arise by two distinct processes: deletion from *S. typhi* or introgression into *E. coli.*

Another group of *E. coli* orphans were similar to their *S. typhi* neighbors. Their origin is probably complex. A small number have distances typical of positionally conserved genes and are likely the result of transposition events involving one or only a few genes. A second group had intermediate distances (0.2–0.5) larger than the average of positionally conserved genes. Some of these orphans could result from duplication events that connect two *E. coli* genes to the same *S. typhi* ORF. An example may be the two *E. coli* formate dehydrogenase operons (fdo and fdn), both of which hit the same group of *S. typhi* sequences. Only one is positionally conserved; the other is an orphan with a distance of 0.26–0.85, depending on the gene. Another cause of orphans at intermediate protein distance is deletion of a positional ortholog from the *S. typhi* chromosome, causing the *E. coli* ORF to hit a second *S. typhi* gene that is very similar to the one deleted. Only a thorough examination of individual genes can distinguish among these various possibilities.

In order to identify genes unique to the *E. coli* genome and not found in *S. typhi,* we removed 194 ORFs from the orphan group that were relatively close in protein distance. Since very few of the genes classified as positional orthologs had $D > 0.5$, we used this as a cutoff to divide orphans into a group of ‘‘novel’’ ORFs versus ‘‘unclassified’’ ORFs (fig. 2). A total of 1,144 *E. coli* ORFs were placed in the novel class, representing a subset of those genes which had no similar *S. typhi* neighbor at the expected chromosomal position. Novel orphans are caused by deletion from the *S. typhi* genome or by introgression into the *E. coli* genome. An example of the latter is the *E. coli* LacA protein, 88% of which aligns with an *S. typhi* ORF at a protein distance of 1.0. The *E. coli* lactose operon is thought to have been horizontally introduced into the *E. coli* genome (Buvinger et al. 1984). The same *S. typhi* ORF is also hit by *E. coli* YlaD, its positional ortholog, at a distance of 0.23, as well as by a second novel *E. coli* ORF, WbbJ, at a distance of 0.82. These *E. coli* genes are part of a group of acetyltransferases with sufficient structural similarity to account for their common *S. typhi* TBLASTN hit. Sixty-four percent (479/747) of the *E. coli* genes previously identified as horizontally transferred (Lawrence and Ochman 1998) belong to the group of novel genes.

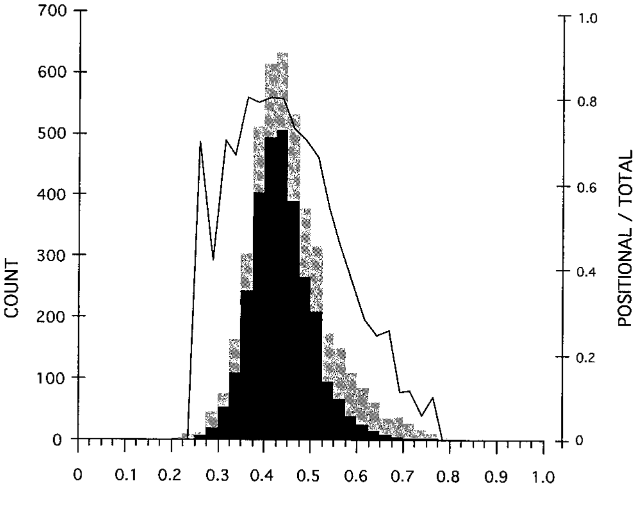
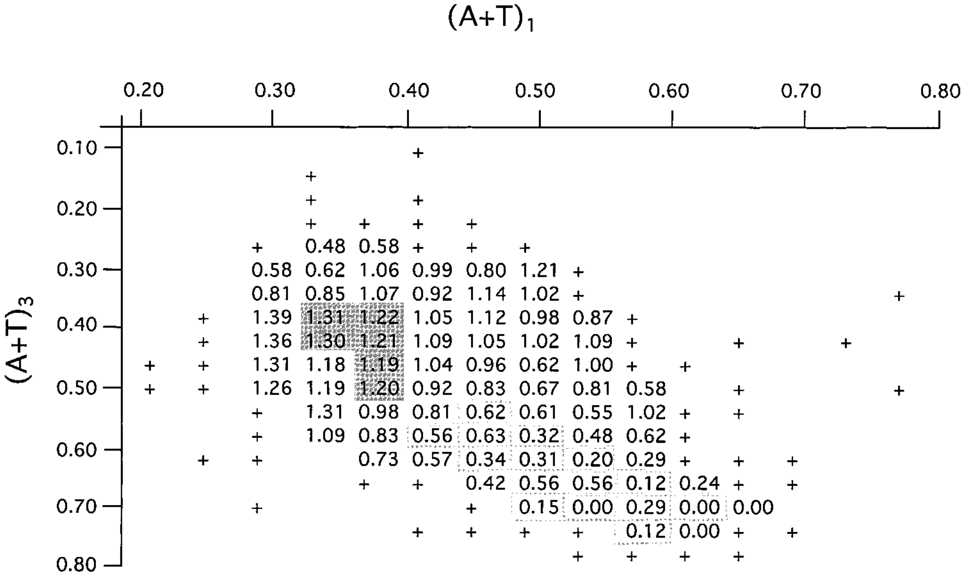


FIG. 3.—Distribution of the third-position (A+T) fraction for all *Escherichia coli* open reading frames (ORFs) ($N = 4{,}289$, gray bars) and positionally conserved ORFs ($N = 2{,}951$, black bars). The line represents the ratio of these two classes.

Codon Bias of *E. coli* ORFs

Escherichia coli genes identified as positionally conserved can be used to estimate the nucleotide composition of *E. coli*/*S. typhi* orthologs. This group of genes has undergone a conservative mode of evolution compared with orphan genes, as shown by their distinct protein distances (fig. 2). They are expected to have nucleotide compositions characteristic of genes resident in the *E. coli* genome since *E. coli* and *S. typhi* diverged. Novel genes that have been introduced into the *E. coli* genome should have compositions reflecting their origin. Positionally conserved genes can be used to test the hypothesis that extreme nucleotide composition is a reliable measure of horizontal transfer. If this is true, any group tentatively identified as foreign on the basis of extreme composition, especially at the third codon position, should contain few, if any, positionally conserved ORFs. Positionally conserved genes are less variable in their use of synonymous codons and tend to have fewer genes with high $(A+T)_3$ (fig. 3). Only 14% of ORFs with $(A+T)_3 > 0.65$ are positionally conserved, but this decreases to 9% for $(A+T)_3 > 0.7$ (table 1). The other extreme of $(A+T)_3$, corresponding to high (G+T) content, is less effective as a predictor of positional conservation, as 37% of ORFs with $(A+T) < 0.3$ are positionally conserved. Although extremes of $(A+T)_3$ are deficient in positionally conserved genes, the typical or modal range is not solely composed of this group. For example, out of 2,967 ORFs with (A+T) between 0.35



					+									
			+											
			+		+									
			+	+	+	+								
		+	0.48	0.58	+	+	+							
		0.58	0.62	1.06	0.99	0.80	1.21	+						
		0.81	0.85	1.07	0.92	1.14	1.02	+						+
	+	1.39	1.31	1.22	1.05	1.12	0.98	0.87	+					
	+	1.36	1.30	1.21	1.09	1.05	1.02	1.09	+		+		+	
+	+	1.31	1.18	1.19	1.04	0.96	0.62	1.00	+	+				
+	+	1.26	1.19	1.20	0.92	0.83	0.67	0.81	0.58		+			+
		+	1.31	0.98	0.81	0.62	0.61	0.55	1.02	+	+			
		+	1.09	0.83	0.56	0.63	0.32	0.48	0.62	+				
	+	+		0.73	0.57	0.34	0.31	0.20	0.29	+	+	+		
				+	+	0.42	0.56	0.56	0.12	0.24	+	+		
		+				+	0.15	0.00	0.29	0.00	0.00			
					+	+	+	+	0.12	0.00	+	+		
								+	+	+	+			

FIG. 4.—Table of odds ratios for the first- and third-position (A+T). The fraction of positionally conserved open reading frames (ORFs) in a cell is divided by the fraction of all ORFs in the cell (0.04 × 0.04). Plus signs indicate cells with less than five ORFs. Shaded cells are significant by a *G*-test, with $G > 11.6$.

and 0.50, fully 667 (22%) are not positionally conserved. Thus, an extreme base composition bias enhances detection but produces many false positives, while genes of intermediate base composition are mostly positional orthologs but may also include many genes that may have been horizontally transferred into *E. coli.*

Extreme (A+T) fractions at first ($(A+T)_1$) and third ($(A+T)_3$) codon positions have been used as one of the signatures of horizontal transmission (Lawrence and Ochman 1997). The ability of first- and third-position nucleotide compositions together to predict positional conservation was investigated by two-way classification. An odds ratio of the fraction of positionally conserved ORFs in a 0.04 × 0.04 cell divided by the fraction of all ORFs in the cell indicates whether or not a compositional range is relatively enriched (>1) or depleted (<1) in positionally conserved genes (fig. 4). *Escherichia coli* ORFs near the modal rage (near 0.4 $(A+T)_1$ and 0.4 $(A+T)_3$) have a greater probability of belonging to the positionally conserved group. Out of 2,632 ORFs with $0.35 < (A+T)_1 < 0.50$ and $0.30 < (A+T)_3 < 0.50$, 2,001 (76%) are positionally conserved, compared with 69% (2,951/4,289) for all genes. In contrast, ORFs with extremely high $(A+T)_1$ and $(A+T)_3$ are underrepresented among positionally conserved genes (table 1). Most of these odds ratios were significantly different from 1.0 by a *G*-test (fig. 4). Thus, combining the first-position (A+T) fraction with the third-position fraction improves the ability of nucleotide composition to identify *E. coli* genes that are candidates for horizontal transfer.

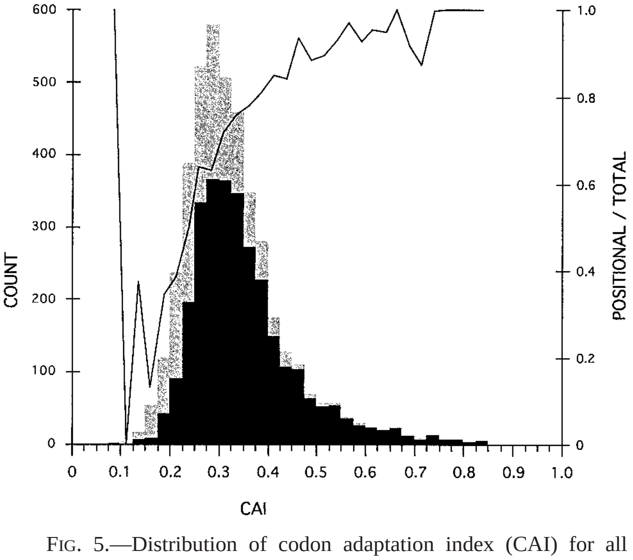


FIG. 5.—Distribution of codon adaptation index (CAI) for all *Escherichia coli* open reading frames (ORFs) (N = 4,289, gray bars) and positionally conserved ORFs (N = 2,951, black bars).The line represents the ratio of these two classes.

CAI was found to be an excellent indicator of positional conservation (fig. 5). Almost all *E. coli* ORFs with unusually large CAIs are positionally conserved (table 1). Out of 476 *E. coli* ORFs with CAI > 0.45, 444 (93%) are positionally conserved. This increases to 100% for the 31 ORFs with CAI > 0.725. Since CAI is correlated with expression level (Sharp and Li 1987), most high-expression genes are evolutionarily stable. On the other hand, about one third of the genes in the modal range of CAI are not positionally conserved. It is only when CAI is less than 0.175 that as much as 80% of *E. coli* ORFs are not positionally conserved. Thus, a low CAI is not as good as a high $(A+T)_1$ plus $(A+T)_3$ content for discriminating against positional conservation. One possible reason for this is that CAI is dependent on amino acid composition, and many positionally conserved genes with very low CAI may have atypical amino acids but not atypical codon usage. This possibility was examined by correcting the observed CAI by sub-

Table 2
Classification of *Escherichia coli* Protein-Coding Open Reading Frames (ORFs)

Group	*N*	Mean CAI	Mean $(A + T)_1$	Mean $(A + T)_3$	Native Trees	HT Trees
All ORFs	4,289	0.331	0.416	0.453		
Position Hits	2,951	0.352	0.403	0.438		
Orthologs	2,728	0.357	0.400	0.434	24	0
Unclassified	223	0.294	0.434	0.490	ND	ND
Orphans	1,338	0.282	0.445	0.486		
Novel	1,144	0.278	0.447	0.488	23	25
Unclassified	194	0.306	0.435	0.475	6	2

NOTE.—CAI = codon adaptation index; HT = horizontally transferred genes.

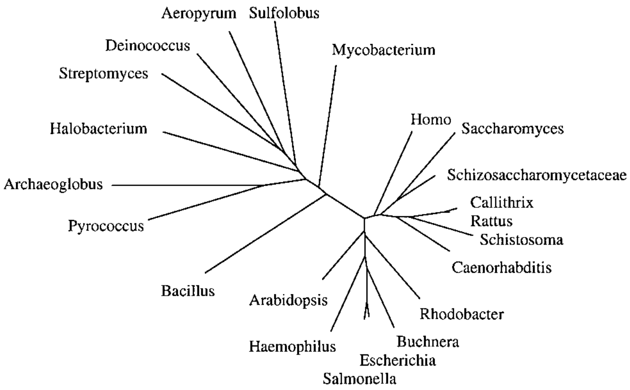


FIG. 6.—A maximum-likelihood gene phylogeny of GloB, a probable hydroxyacylglutathione hydrolase. The tree is unrooted with branch lengths proportional to the number of substitutions per site.

tracting the CAI expected for a protein with the same amino acids but using synonymous codons according to the frequencies of the cumulative genome average. Of the 429 ORFs below the 10th percentile score, 112 (26%) are positionally conserved, while of the lowest 100, only 15 are positionally conserved (table 1). Thus, this correction marginally improved the association of very low CAI with a lack of positional conservation. Even after correction, low CAI identified few genes as potential introgressions. Furthermore, many genes of intermediate CAI fail to have a positional ortholog in *S. typhi.*

It is possible that a significant fraction of those *E. coli* ORFs which are positionally conserved but have extreme nucleotide composition are homologous replacements. We removed those ORFs which were unusually distant from their *S. typhi* relatives, forming a subgroup of ‘‘positional orthologs’’ (fig. 2 and table 2). Far fewer ORFs with very low CAIs are positional orthologs. Of the 429 ORFs in the 10th percentile for Z_{CAI}, 64 were positional orthologs (compared with 112 positionally conserved), and of the lowest 100 ORFs, only 4 remained of the original 15 after removing distant *S. typhi* hits. Of the 21 positionally conserved ORFs with extreme $(A+T)_1$ (>0.64) and $(A+T)_3$ (>0.48), only 6 were positional orthologs (5% of all the genes in this compositional range). Thus, the tendency of evolutionarily stable *E. coli* genes to have higher GC contents and CAIs was confirmed by the separation of positional orthologs from positionally conserved genes.

Combining low CAI and high (A+T) at first and third positions identifies almost no positionally conserved genes. Out of 49 genes with $(A+T)_1 > 0.55$ and $(A+T)_3 > 0.65$ and CAI < 0.2, only 2 were positionally conserved (table 1). While this seems to be an excellent predictor of nonorthologous transmissions, it identifies very few such genes.

Phylogenies Can Identify Horizontally Transferred Genes

Horizontal transfer produces chromosomes containing genes with different ancestries and durations in the genome (Lawrence and Ochman 1998) Introgression can alter the topologies of gene trees; therefore, a phylogenetic approach can be employed to detect horizontally transferred genes. If a gene is confined to one taxon or species, it is more likely to have been acquired through horizontal transfer than to have been lost independently from multiple lineages (Smith, Feng, and Doolittle 1992). If it can be shown that many different genes from the same organism have statistically different phylogenies, this would indicate that horizontal gene transfer is playing a major role in the evolution of the species.

A total of 80/102 statistically significant protein trees were generated (see http://life.biology.mcmaster.ca/~liisa/appendix.html) and used to explore how accurately our classification can identify potential horizontally transferred genes (table 2). A large number of genes cannot be used to generate phylogenies, as homologs have not yet been identified in a sufficient number of species. Out of 24 protein trees tested for the group of positional orthologs, all generated phylogenies consistent with vertical evolution. We expect all of the genes in this category to be native to *E. coli* since its divergence from Salmonella. An example of an *E. coli* gene of unusual nucleotide composition which has evolved normally is *gloB*, coding for a probable hydroxyacylglutathione hydrolase. It is one of the four positional orthologs among the 100 *E. coli* genes with the lowest Z_{CAI}. However, its $(A+T)_1$ (0.51) and $(A+T)_3$ (0.35) do not place it among the ORFs that are extreme in this measure. *GloB* was categorized by Lawrence and Ochman 1998 as horizontally transferred. However, its tree and homologous chromosomal position is consistent with a vertically evolving gene within *E. coli* and Salmonella (fig. 6). Of the 24 positional orthologs whose trees do not show evidence of horizontal transfer, Lawrence and Ochman 1998 classified 15 as horizontally transferred based on codon bias and base composition. This does not represent an estimate of the frequency of misclassification, however, as we specifically analyzed a subsample of positional orthologs which had been previously classified as horizontally transferred. A lack of phylogenetic evidence for horizontal transfer is also not proof that it did not occur, since introgressions between closely related organisms would go unnoticed by this technique. Our phylogenetic analysis simply fails

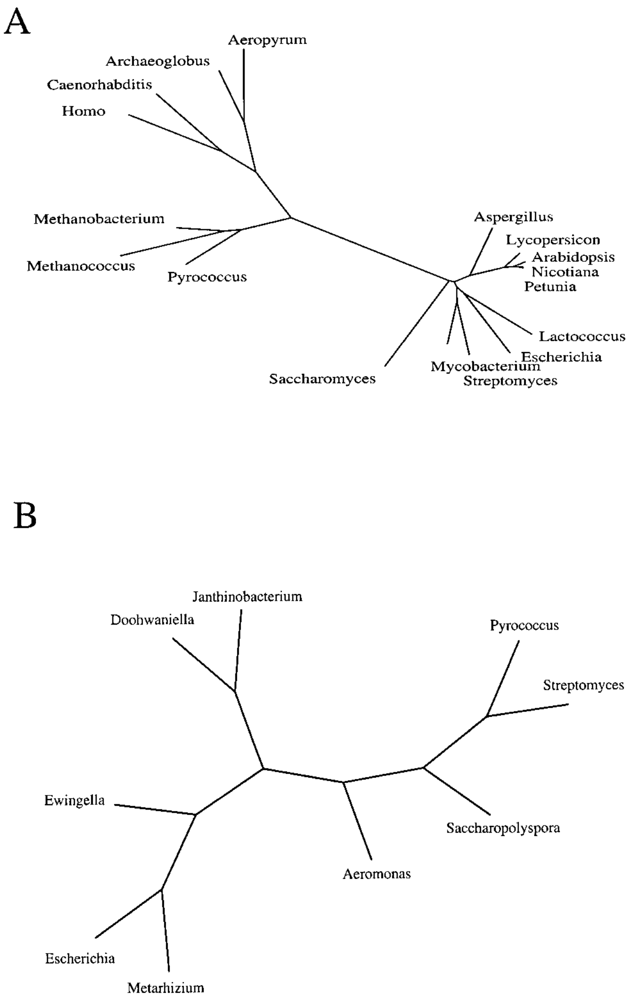


FIG. 7.—Maximum-likelihood gene phylogenies of (*A*) GadB, a glutamatedecarboxylase isozyme, and (*B*) YheB, a probable endochitinase. GadB has no known *Salmonella typhi* or other Gram-negative homologs. The trees are unrooted with branch lengths proportional to the number of substitutions per site.

to find evidence that horizontal transfer has taken place between distantly related species.

In contrast to *E. coli* genes with a positional ortholog, we expect, and indeed do find, a large number of horizontally transferred genes in the novel category. A total of 48 gene trees were generated for *E. coli* ORFs classified as novel and having no obvious ortholog in *S. typhi*. Many of these have typical nucleotide compositions; two are shown in figure 7. GadB is one of two very similar *E. coli* glutamate decarboxylase isozymes that have been reported to be absent from *S. typhimurium* (Smith et al. 1992). In fact, *gadB* has a CAI (0.501) that falls in the region otherwise indicative of highly expressed, conserved genes. YheB is an endochitinase with typical base composition and CAI (0.436). Both *gadB* (fig. 7*A*) and *yheB* (fig. 7*B*) were classified by Lawrence and Ochman 1998 as native to *E. coli*. According to the gene trees, both appear to be involved in horizontal transfer with *E. coli* and a distantly related species (*gadB* has no other Gram-negative homolog, while *yheB* is located next to *Metarhizium anisopliae*, a eukaryotic species). Of 25 novel ORFs for which gene trees give evidence of horizontal transfer, Lawrence and Ochman 1998 classified 12 as native to the *E. coli* genome using codon bias and base composition. Again, this does not represent a random sample of native genes. It is also possible that a fraction of the 25 horizontally transferred genes are incorrect, as systematic errors could lead to an incorrect tree topology. About half of the gene trees generated for novel ORFs are consistent with vertical evolution. These may be genes that have been deleted in Salmonella or genes originating from horizontal transfer events between *E. coli* and a closely related Gram-negative species.

Conclusions

It has been suggested by Lawrence and Ochman 1998 that a minimum 18% of the *E. coli* genome has arisen through horizontal transfer. This was estimated by analyzing the base composition and codon bias of the complete *E. coli* strain MG1655 sequence. The fact that 48% (23/48) of the trees for novel genes indicated horizontal transmission suggests that about 10%–15% of the total *E. coli* genome is the result of introgression. It is clear that horizontal transfer has had a major role in the evolution of this species.

We confirmed that a majority of the *E. coli* ORFs with extreme nucleotide compositions do not have positional orthologs in the *S. typhi* genome.$(A+T)_1$ and $(A+T)_3$ were found to be a better indicator of nonorthologous genes than $(A+T)_3$ alone or CAI, while a combination of CAI and (A+T) at first and third positions was found to be an even better indicator. An unusually large CAI, on the other hand, is a good indicator of orthologous transmission. Over 90% of those *E. coli* ORFs with CAI $>$ 0.45 were found to have an *S. typhi* ortholog, indicating that high-expression genes have been retained in both genomes since their divergence.

While extreme nucleotide composition is a good predictor of nonorthologous transmission, it identifies only a small fraction of such genes. Approximately 20% of *E. coli* ORFs with typical nucleotide composition and codon usage, representing over 600 genes, have no positional ortholog in *S. typhi*. Most of these are genes that have been either deleted from the *S. typhi* genome or horizontally introduced into the *E. coli* genome. Gene trees for those ‘‘novel’’ ORFs for which sufficient data were available (table 2) suggest that approximately 52% of these are horizontal introgressions. Therefore, using nucleotide composition alone to identify horizontal introgressions would miss approximately 300 genes. It is perhaps not surprising that many horizontal introgressions would have nucleotide compositions typical of the *E. coli* genome, since gene transfer is expected to be more likely between closely related bacteria. The process of amelioration would also account for the typical composition of some horizontally transferred genes.

Even though there have been numerous additions and deletions to the *E. coli* and *S. typhi* genomes, our analysis indicates that these two species have retained a common core of approximately 2,700 genes. Many of these are highly expressed *E. coli* genes, as identified by CAI. These conserved genes account for the colinearity of the Escherichia and Salmonella genetic maps (Riley and Krawiec 1987; Krawiec and Riley 1990). This conserved core must provide for the common properties of Escherichia and Salmonella, while those added by introgression determine distinctions among subgroups and adaptations to novel environments (Ochman, Lawrence, and Groisman 2000).

Phylogenetic analysis has confirmed that a substantial fraction of novel *E. coli* ORFs originated from horizontal transfer since this species diverged from Salmonella. Even the Escherichia and Salmonella genera themselves may have split because of introgressions that provided distinctive functions (Groisman, Saier, and Ochman 1992; Groisman et al. 1993). It is clear from our analysis of gene trees and chromosomal position that base composition and codon usage patterns should not be used without additional support to identify horizontally transferred genes. Combining measures based on base composition with a phylogenetic approach is required to eliminate vertically evolved genes with atypical composition. Although the analysis of gene trees is time-consuming and requires a number of well-characterized species, it is an important tool in the analysis of horizontal gene transfer and should be employed whenever possible.

LITERATURE CITED

Adachi, J., and M. Hasegawa 1992. Molphy: programs for molecular phylogenetics, I. Protml: maximum likelihood inference of protein phylogeny. Computer Science Monographs.

Altschul, S. F., W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman. 1990. Basic alignment search tool. J. Mol. Biol. **215**:403–410.

Blattner, F. R., G. Plunkett III, C. A. Bloch et al. (17 coauthors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1474.

Buvinger, W. E., K. A. Lampel, R. J. Bojanowski, and M. Riley. 1984. Location and analysis of nucleotide sequences at one end of a putative lac transposon in the *Escherichia coli* chromosome. J. Bacteriol. **159**:618–622.

Doolittle, R. F., D. F. Feng, S. Tsang, G. Cho, and E. Little. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. Science **271**: 470–477.

Felsenstein, J. 1994. PHYLIP (phylogeny inference package). Version 3.5. Distributed by the author, University of Washington, Seattle.

Groisman, E. A.., M. H. Saier Jr., and H. Ochman. 1992. Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the *Salmonella* genome. EMBO J. **11**: 1309–1316.

Groisman, E. A., M. A. Sturmoski, F. R. Solomon, R. Lin, and H. Ochman. 1993. Molecular, functional, and evolutionary analysis of sequences specific to *Salmonella*. Proc. Natl. Acad. Sci. USA **90**:1033–1037.

Huynen, M. A., and P. Bork. 1998. Measuring genome evolution. Proc. Natl. Acad. Sci. USA **95**:5849–5856.

Ikemura, T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J. Mol. Biol. **151**:389–409.

Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. J. Mol. Evol. **29**:170–179.

Krawiec, S., and M. Riley. 1990. Organization of the bacterial chromosome. Microbiol. Rev. **54**:502–533.

Lawrence, J. G., and H. Ochman. 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44**:383–397.

———. 1998. Molecular archaeology of the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95**:9413–9417.

Médigue, C., T. Rouxel, P. Vigier, A. Henaut, and A. Danchin. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. J. Mol. Biol. **222**:851–856.

Mrázek, J., and S. Karlin. 1999. Detecting alien genes in bacterial genomes. Ann. N.Y. Acad. Sci. **870**:314–329.

Ochman, H., and J. G. Lawrence. 1996. *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. Pp. 2627–2637 *in* American Society for Microbiology, Washington, D.C.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

Riley, M., and S. Krawiec. 1987. *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. Pp. 967–981 *in* American Society for Microbiology, Washington, D.C.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sharp, P. M. 1991. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J. Mol. Evol. **33**:23–33.

Sharp, P. M., and W. H. Li. 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for ‘rare’ codons. Nucleic Acids Res. **14**:7737–7749.

———. 1987. The Codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

Smith, D. K., T. Kassam, B. Singh, and J. F. Elliott. 1992. *Escherichia coli* has two homologous glutamate decarboxylase genes that map to distinct loci. J. Bacteriol. **174**:5820–5826.

Smith, M. W., D. F. Feng, and R. F. Doolittle. 1992. Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem. Sci. **17**:489–493.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Wolfgang Stephan, reviewing editor

Accepted November 22, 2000