J Mol Evol (1997) 44:383–397

JOURNAL OF MOLECULAR EVOLUTION

Amelioration of Bacterial Genomes: Rates of Change and Exchange

Jeffrey G. Lawrence,[1,]* Howard Ochman[2]

[1] Department of Biology, University of Utah, Salt Lake City, UT 84112, USA
[2] Department of Biology, University of Rochester, Rochester, NY 14627, USA

Received: 7 July 1996 / Accepted: 27 September 1996

Abstract. Although bacterial species display wide variation in their overall GC contents, the genes within a particular species' genome are relatively similar in base composition. As a result, sequences that are novel to a bacterial genome—i.e., DNA introduced through recent horizontal transfer—often bear unusual sequence characteristics and can be distinguished from ancestral DNA. At the time of introgression, horizontally transferred genes reflect the base composition of the donor genome; but, over time, these sequences will ameliorate to reflect the DNA composition of the new genome because the introgressed genes are subject to the same mutational processes affecting all genes in the recipient genome. This process of amelioration is evident in a large group of genes involved in host-cell invasion by enteric bacteria and can be modeled to predict the amount of time required after transfer for foreign DNA to resemble native DNA. Furthermore, models of amelioration can be used to estimate the time of introgression of foreign genes in a chromosome. Applying this approach to a 1.43-megabase continuous sequence, we have calculated that the entire *Escherichia coli* chromosome contains more than 600 kb of horizontally transferred, protein-coding DNA. Estimates of amelioration times indicate that this DNA has accumulated at a rate of 31 kb per million years, which is on the order of the amount of variant DNA introduced by point mutations. This rate predicts that the *E. coli* and *Salmonella enterica* lineages have each gained and lost more than 3 megabases of novel DNA since their divergence.

Key words: Genome amelioration — Directed mutation pressure — GC content — Horizontal transfer — Invasion genes — *Salmonella enterica* — *Escherichia coli* — Bacterial evolution

Introduction

Many of the initial characterizations of nucleic acids in bacterial genomes focused on the overall nucleotide composition (GC content) of each species. These analyses, as well as the recent accumulation of DNA sequence information, have resulted in several insights into the attributes, organization, and evolution of bacterial chromosomes. In general, these investigations have revealed four salient features of base composition in bacteria:

1. Base composition varies widely among bacterial species. Base compositions range from 25% GC in *Mycoplasma* to 75% GC in *Micrococcus*, which is much larger than the range in overall GC contents observed among animals or plants. The differences in base composition among bacterial species are largely due to biases in the mutation rates at each of the four bases—termed ''directional mutation pressure'' by Sueoka (1961, 1962, 1988, 1992, 1993, Sueoka et al. 1959)—which vary between species.
2. Base composition is related to phylogeny; i.e., GC contents of the chromosomes of closely related or-

* *Present address:* Department of Biological Sciences, 215 Clapp Hall, University of Pittsburgh, Pittsburgh, PA 15260, USA
Abbreviations: GC, guanine plus cytosine; MO, Muto and Osawa; kb, kilobase; mb, megabase; Myr, million years; CAI, codon adaptation index
Correspondence to: H. Ochman

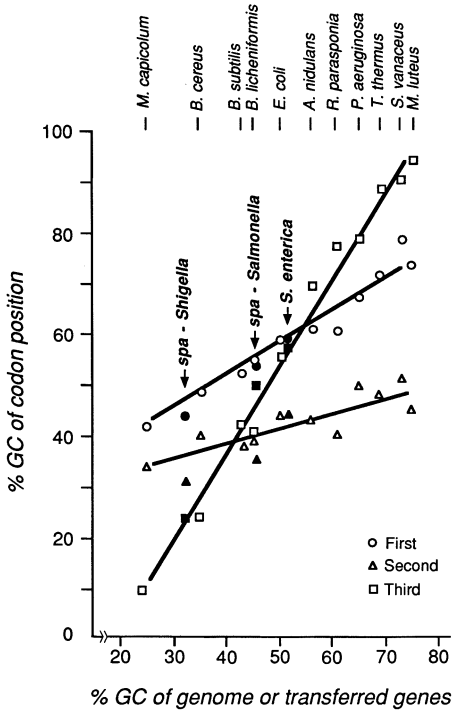

Fig. 1. Correlation between genomic GC content and GC content at each codon position within bacterial genomes [after Fig. 2 of Muto and Osawa (1987)]. Values for *E. coli* have been derived from data provided herein and values for *S. enterica* were obtained from Ochman and Lawrence (1996); values for all other species were obtained from Muto and Osawa (1987). GC contents of *spa* genes from *Salmonella* and *Shigella* are provided for comparison. Linear equations describing these relationships are presented in equations 13–15.

ganisms tend to be similar. Bacterial genera can be organized into groups of related lineages that can be subsequently characterized by their fairly similar nucleotide contents. This congruency indicates that base composition can be stable over long evolutionary periods (Ochman and Lawrence 1996).

3. Base composition is relatively homogeneous over the entire bacterial chromosome. This feature was evident from the initial characterizations of bacterial genomes using buoyant density gradients (Rolfe and Meselson 1959; Schildkraut et al. 1962; Sueoka et al. 1959) and has been confirmed with the accumulation of DNA sequence information (Fleischmann et al. 1995). Unlike the chromosomes of warm-blooded vertebrates, which are partitioned into large ''isochores'' of distinct base compositions (Bernardi 1989), there is comparatively little broad-scale spatial heterogeneity in GC content within bacterial genomes.
4. Within each species, the first, second, and third positions of codons, as well as the genes for structural RNAs, have characteristic base compositions (Muto and Osawa 1987). The differences in genomic GC contents are apparent in the base composition at each codon position of sequenced genes (Fig. 1). In that mutational biases are evident in the base composition of codons—particularly at third positions where most changes are synonymous—they can influence the choice among alternative synonymous codons.

As a consequence, regions having atypical base compositions or codon usage patterns have been cited as evidence of horizontal transfer: Since base composition is relatively uniform over the entire bacterial genome and is conserved within and among related lineages, genes with anomalous features are likely to have been acquired recently from a distantly related organism. This rationale has been applied to infer the ancestry of several chromosomally encoded genes in enteric bacteria. By examining the GC contents and codon usage patterns of sequenced genes from *Escherichia coli*, several groups have estimated the proportion of genes that were presumably introduced by horizontal transfer (Médigue et al. 1991; Ochman and Lawrence 1996; Whittam and Ake 1992).

While there is ample evidence that bacterial chromosomes contain sequences inherited by horizontal processes as well as by vertical transmission (Kidwell 1993; Syvanen 1994), the use of nucleotide composition or codon usage patterns to infer the ancestry of a region of the chromosome rests on an additional assumption—that the ''atypical'' features are not attributable to either stochastic variation or to selective factors specific to a sequence. Furthermore, if the base composition and codon usage pattern of a gene result from mutational biases, sequences introduced by horizontal transfer will incur substitutions and eventually manifest features typical of genes ancestral to chromosome. The process of amelioration, whereby a sequence adjusts to the base composition and codon usage of the resident genome, is a function of the relative rate of (A or T) $\rightarrow$ (G or C) mutations and should be most evident at sites with little or no functional constraints.

We have examined the rates and patterns of evolution of genes from the enteric bacteria *Escherichia coli* and *Salmonella enterica* to monitor change in base composition over evolutionary time, to estimate the total amount of acquired DNA within the *E. coli* genome, and to calculate the approximate time of introgression of genes gained through horizontal transfer. The broad similarity of these organisms allows us to identify those genes restricted to only one of the species and to compare the compositional structure, codon bias, and map positions of species-specific genes to those common to both species. From these measures, we have determined the rate of horizontal transfer over the long-term evolution of enteric bacteria.

Methods

DNA Sequences. A 1.43-megabase (mb) continuous sequence of the *E. coli* genome was constructed from the GenBank sequences U18997

(bases 1–372,438 of the 1.43-mb contig) and continuing through sequences U00039 [bases 366,476–591,894 (Sofia et al. 1994)], L10328 [bases 591,348–727,601 (Burland et al. 1993)], M87049 [bases 727,596–819,003 (Daniels et al. 1992)], L19201 [bases 819,001–915,484 (Plunkett et al. 1993)], U00006 (bases 915,932–1,091,586), and U14003 (bases 1,089,702–1,428,235). This sequence spans minutes 67 to 100 on the *E. coli* genetic map and constitutes approximately 30% of the chromosome. Protein-coding regions were identified using the annotation provided for the DNA sequences; open reading frames have been identified in a consistent manner across the length of the sequence by the single research group that deposited the sequences.

Identification of Horizontally Transferred DNA. To estimate the rate at which new genes have been acquired over the evolutionary history of *E. coli*, it is first necessary to identify the particular sequences that arose through horizontal transfer. While certain introgressed genes—such as those encoded by mobile genetic elements—can be identified on the basis of function, the ancestry of most introgressed genes can be inferred by one of two methods: (1) A gene may be confined to one taxon of a species group and be absent from closely related taxa. Because it is more likely for a gene to be introduced into one taxon, rather than deleted from all closely related taxa, this method of identifying horizontal transfer events is fairly robust. However, this phylogenetic approach requires a set of well-characterized species for comparison, and such information is not available for many groups. (2) In addition, genes that did not originate in their present genomic context (i.e., those that arose through horizontal transfer) often have properties that depart from the prevalent characteristics of the genome. For this approach, it is necessary establish a set of measures that characterize the ''typical'' genes from an organism in order to identify sequences suspected to have originated through horizontal transfer.

We have examined the features of genes from *Escherichia coli* and *Salmonella enterica* and have developed several criteria to identify horizontally transferred genes. First, the codon-position-specific GC contents have been determined for each gene. Genes native to the *E. coli* chromosome display characteristic values for these measures (gray bars in Fig. 2); on average, the first, second, and third codon positions for the genes of the region of the *E. coli* chromosome analyzed here are 59%, 43%, and 56% GC, respectively. Atypical protein coding regions (black bars in Fig. 2) were initially identified if the GC contents at first and third codon positions were 10% lower, or 8% higher, than their respective means. (The GC contents of second codon positions are typically too similar across species to be useful indicators.)

Some genes that are identified as ''atypical'' by this process may, in fact, be native to the *E. coli* chromosome: GC contents of certain genes are unusual due to selection on the encoded protein, which constrains the evolution of the nucleotide sequence by favoring certain codons. To distinguish native genes with unusual base compositions from horizontally transferred genes, the codon usage bias of each gene was examined. Genes encoding proteins with strong selective constraints, such as the *rplL* gene, can show a strong bias in codon usage; these codon usage biases typically differ among species. To measure bias in codon usage, the χ^2 of codon usage was determined (with expected values being the equivalent use of all synonymous codons). Because this χ^2 value detects only the presence of codon usage bias and does not indicate if codon usage is biased in any particular direction, we employed the Codon Adaptation Index [CAI, (Sharp and Li 1987)] to determine if codon preferences were biased toward the particular subset of codons employed by highly expressed genes in *E. coli*. If selection for preferred codons has resulted in an atypical GC content in a native *E. coli* gene, the gene would show both high χ^2 and high CAI values. For example, the *rplL* gene has an unusually low GC content at its third codon position due to the high lysine content of the RplL protein; lysine is preferentially encoded by AAA codon in highly expressed genes (Sharp and Li 1987). In such cases, these genes are discarded from the set of potentially horizontally transferred genes as originally identified on the basis of GC content. In contrast, genes of exogenous origin are often highly biased—exhibit a high χ^2 of codon usage—but show a low CAI value, indicating that they are not biased in the manner characteristic of *E. coli* genes.

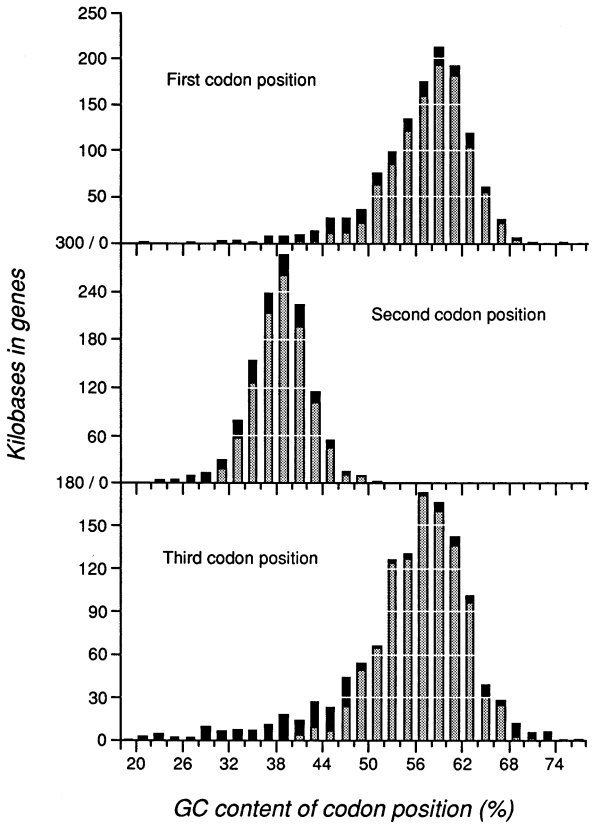

Fig. 2. GC contents of 1,294 *E. coli* genes. GC contents at each codon position are shown. *Gray bars* represent native (i.e., ancestral) genes; *black bars* represent genes inferred to have been acquired by horizontal transfer.

Computer Programs. Analyses of the nucleotide composition, codon usage bias, and dinucleotide frequencies of genes, as well as calculations of amelioration times and reverse amelioration kinetics employed programs available from the authors. DNA sequences were obtained from GenBank with the GCG program package (Devereux et al. 1984).

Results

Quantitation of Horizontally Transferred DNA in Escherichia coli

There have been several attempts to quantify the amount of horizontally transferred DNA in the *E. coli* and *S. enterica* genomes. Based on the GC contents of 500 genes, Whittam and Ake (1992) estimated that 6% of the *E. coli* chromosome arose through horizontal transfer; but, in an analysis of codon usage patterns, Médigue et al. (1991) concluded that as many as 16% of *E. coli* genes were of novel origin. Similarly, Ochman and Lawrence (1996) estimated that at least 10% of the genes

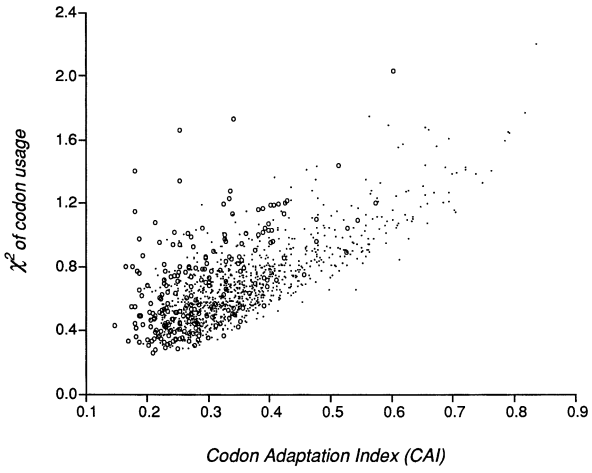

Fig. 3. Bivariate plot of the CAI vs χ^2 of codon usage for 1,189 *E. coli* genes. *Points* represent native genes ($n = 1{,}024$); *open circles* represent genes inferred to have been acquired by horizontal transfer ($n = 165$). Sequences less than 300 bp in length were not analyzed because small numbers of codons inflate χ^2 of codon usage.

in *S. enterica* sv. Typhimurium were acquired by horizontal transfer. However, these estimates of the amount of introgressed DNA may be biased since the studies focused on subsets of gene contributing to notable phenotypes in laboratory environments, whose DNA sequences had been determined and deposited in databanks.

To avoid any bias in sample selection, we analyzed 1.43 mb of continuous sequence from the *E. coli* chromosome. This region encompasses minutes 67 to 100 on the *E. coli* K-12 genetic map and contains 1,294 protein-coding regions constituting 1.23 mb of DNA; the 0.20 mb of non-protein-coding DNA was not used in this analysis. Using the methods described above, 200 of these protein-coding regions were identified as having been horizontally transferred into the *E. coli* genome. An additional 29 genes were deduced as having been horizontally transferred based on their function and/or chromosomal location. [For example, some genes of the *rfa* operon were not ostensibly atypical, but they were located within a large operon of known exogenous origin (Klena et al. 1993).] Therefore, a total of 229 genes were classified as having been acquired by horizontal transfer; the features of these genes had not fully ameliorated to resemble genes that are ancestral to the *E. coli* chromosome. Most of these genes show atypical codon usage bias (Fig. 3) as well as atypical nucleotide contents (Fig. 2). A list of genes within the 1.43-mb region identified as being introduced by horizontal transfer appears as Appendix I.

These 229 genes represent a total of 186 kilobases of DNA, or 17% of the protein-coding DNA within the region, which is very similar to the estimate (16%) obtained by Médigue et al. (1991). By extrapolation, we estimate that some 618 kb of the protein-coding sequences in the *E. coli* K-12 chromosome were acquired by horizontal transfer. These sequences represent a minimal estimate of the amount of introgressed DNA present in the *E. coli* chromosome for two reasons: (1) DNA that was introduced sufficiently long ago has since ameliorated to resemble fully the prevalent characteristics of the *E. coli* genome and will not be detected by our methods, and (2) DNA that was acquired from an organism whose base composition is similar that of *E. coli* will also remain undetected in our survey.

Some of the sequences identified as having been acquired by horizontal transfer have known phenotypes: The *rfa* genes encode proteins involved in outer membrane antigen synthesis (Klena et al. 1993), the *tdc* genes encode the proteins responsible for the degradation of threonine (Gross et al. 1988), and the *phn* genes encode functions for the conversion of phosphonates to phosphates (Metcalf and Wanner 1993; Wanner and Metcalf 1992). [Therefore, it is not surprising that the enzymes for phosphonate conversion by *S. enterica* perform mechanistically distinct biochemical reactions and are encoded by a set of genes that are not homologous to the *E. coli phn* genes (Jiang et al. 1995)]. As discussed below, most horizontally transferred genes serve nonessential role in *E. coli* metabolism (Lawrence and Roth 1996b), and many have no identified function.

A Model for the Amelioration of Bacterial DNA

To determine the rate at which transferred genes have been incorporated into the *E. coli* chromosome, it is necessary to estimate the time of introduction of the DNA fragments acquired through horizontal transfer. To estimate these times, we examined the rate and extent of amelioration of the 229 introgressed genes and determined how long each sequence has been subjected to the directional mutational pressures of the recipient genome. To perform this analysis, the following mathematical model of DNA amelioration was derived.

The nucleotide composition of a DNA sequence typically reflects an equilibrium between selection and directional mutation pressures (Sueoka 1962; 1988; 1992). When DNA is transferred to a novel genome—one experiencing different directional mutation pressures—its base composition will change to reach a new equilibrium. Mathematical models describing these changes have been developed (Sueoka 1962) and express the changes in DNA composition in terms of direction mutation pressures. We have derived a model of amelioration to examine this process from an evolutionary standpoint. This model does not quantify directional mutational pressures; rather, net changes in DNA composition are expressed as fractions of the nucleotide substitution rates. Furthermore, all parameters for this model can be derived from existing nucleotide sequence information.

The rate of amelioration can be expressed as a func-

tion of the rates of evolutionary change, or substitution rates (S in equation 1); substitution rates at both synonymous and nonsynonymous sites have already been estimated for *E. coli* and *S. enterica* (Doolittle et al. 1996; Ochman and Wilson 1987, 1988). An empirically determined substitution rate, S, can be viewed as the sum of the rate of change at guanine or cytosine nucleotides (R_{GC}) and the rate at adenine or thymidine nucleotides (R_{AT}):

$$S = R_{GC} + R_{AT} \tag{1}$$

In turn, the rate of change at GC nucleotides (R_{GC}) can be expressed as the sum of two rates, those that change G or C nucleotides to A or T nucleotides ($R_{GC \to AT}$ includes G → T, G → A, C → T, and C → A mutations) and those that do not ($R_{G,C \to C,G}$ specifies G → C and C → G mutations):

$$R_{GC} = R_{GC \to AT} + R_{G,C \to C,G} \tag{2a}$$

$$R_{AT} = R_{AT \to GC} + R_{A,T \to T,A} \tag{2b}$$

Similarly, the rate of change at AT nucleotides (R_{AT}) can be expressed as two analogous classes (equation 2b). Since all transition mutations, and half of the transversions, change the GC content of DNA, we can simplify these equations by expressing equation (2a) solely in terms of one rate of change and the transition/transversion ratio (IV ratio). We assume that the two transversion mutations are equally frequent; reasonable deviations from this assumption do not substantially alter the model:

$$R_{GC} = R_{GC \to AT} + \left[\frac{1}{2 \times \text{IV ratio} + 1} \times R_{GC \to AT}\right] = \frac{\text{IV ratio} + 1}{\text{IV ratio} + 1/2} \times R_{GC \to AT} \tag{3}$$

An analogous equation can be derived for the R_{AT} parameter. Substituting these values into equation (1) yields the following relationships:

$$S = \left[\frac{\text{IV ratio} + 1}{\text{IV ratio} + 1/2} \times R_{GC \to AT}\right] + \left[\frac{\text{IV ratio} + 1}{\text{IV ratio} + 1/2} \times R_{AT \to GC}\right] = \frac{\text{IV ratio} + 1}{\text{IV ratio} + 1/2} \times [R_{GC \to AT} + R_{AT \to GC}] \tag{4}$$

Equation (4) expresses the total substitution rate (S) as a function of the two directional mutation rates and the transition/transversion ratio. As proposed by Sueoka (1988, 1992, 1993), the combined action of these two rates will result in an equilibrium GC content; if the two rates are not equal, the GC content of the DNA will deviate from 0.5, or 50% GC. Therefore, the equilibrium GC content of the DNA (GC^{EQ}) can be expressed in terms of the ratio of the directional mutation rates:

$$GC^{EQ} = \frac{R_{AT \to GC}}{R_{AT \to GC} + R_{GC \to AT}} \tag{5a}$$

$$AT^{EQ} = \frac{R_{GC \to AT}}{R_{AT \to GC} + R_{GC \to AT}} \tag{5b}$$

and thus

$$R_{AT \to GC} + R_{GC \to AT} = \frac{R_{AT \to GC}}{GC^{EQ}} = \frac{R_{GC \to AT}}{AT^{EQ}} \tag{6}$$

Substituting these values into equation (4) yields the following two corresponding relationships:

$$S = \frac{\text{IV ratio} + 1}{\text{IV ratio} + 1/2} \times \frac{R_{AT \to GC}}{GC^{EQ}} \tag{7a}$$

$$S = \frac{\text{IV ratio} + 1}{\text{IV ratio} + 1/2} \times \frac{R_{GC \to AT}}{AT^{EQ}} \tag{7b}$$

Equations (7a) and (7b) can be solved to obtain a directional mutation rate (*e.g.*, $R_{AT \to GC}$) in terms of the substitution rate (S), the equilibrium GC or AT content, and the transition/transversion ratio:

$$R_{AT \to GC} = S \times \frac{\text{IV ratio} + 1/2}{\text{IV ratio} + 1} \times GC^{EQ} \tag{8a}$$

$$R_{GC \to AT} = S \times \frac{\text{IV ratio} + 1/2}{\text{IV ratio} + 1} \times AT^{EQ} \tag{8b}$$

Further, the change in the GC content over time can be expressed as the gain in GC nucleotides minus the loss of GC nucleotides, letting AT^{HT} and GC^{HT} denote the base composition of horizontally transferred DNA:

$$\Delta GC^{HT} = [AT^{HT} \times R_{AT \to GC}] - [GC^{HT} \times R_{GC \to AT}] \tag{9}$$

Substituting the rate values with expressions derived from equations (8a) and (8b), we obtain the following:

$$\Delta GC^{HT} = S \times \frac{\text{IV ratio} + 1/2}{\text{IV ratio} + 1} \times [AT^{HT} \times GC^{EQ} - GC^{HT} \times AT^{EQ}] \tag{10}$$

Table 1. Amelioration parameters used for *E. coli* and *S. enterica*

Class	Proportion of sites		Substitution rate[a] (% Divergence per Myr)
	Synonymous	Nonsynonymous	
Synonymous	1	0	0.910
Nonsynonymous	0	1	0.045
First codon position	0.09	0.91	0.123
Second codon position	0	1	0.045
Third codon position	0.72	0.28	0.668

[a] Based on a divergence time between *E. coli* and *S. enterica* of 100 million years (Doolittle et al. 1996; Ochman and Wilson 1987, 1988)

Because AT = 1 − GC, we can substitute for the AT^{HT} and AT^{EQ} in equation (10):

$$\Delta GC^{HT} = S \times \frac{\text{IV ratio} + 1/2}{\text{IV ratio} + 1} \times [(1 - GC^{HT}) \times GC^{EQ} - GC^{HT} \times (1 - GC^{EQ})] \quad (11)$$

This reduces to:

$$\Delta GC^{HT} = S \times \frac{\text{IV ratio} + 1/2}{\text{IV ratio} + 1} \times [GC^{EQ} - GC^{HT}] \quad (12)$$

As expected, the rate of change in GC content of horizontally transferred DNA (ΔGC) is proportional both to the substitution rate (S) and to the magnitude of the difference in nucleotide composition between horizontally transferred DNA and equilibrium values. Cumulatively, this model allows the change in GC content to be predicted by three parameters—the transition/transversion ratio, the rate of evolutionary change for the class of nucleotides being examined, and the equilibrium GC content of the DNA—all of which can be obtained from comparative studies of nucleotide sequences. In addition, different classes of nucleotides—e.g., those at different codon positions—experience different selective pressures and evolve at different rates. Therefore, different classes of nucleotides sites can be analyzed independently to provide additional information about the degree of amelioration of a particular sequence.

In the following sections, we have applied these principles to examine the rates and patterns of amelioration in sequences from the enteric bacteria *Escherichia coli* and *Salmonella enterica*. For this pair of species, all three parameters required for equation (12) have been determined. The transition/transversion ratio is taken to be 2:1 [inspection of equation (12) reveals that reasonable variation in this parameter has little impact on ΔGC], the divergence of *E. coli* and *S. enterica* from their common ancestor is estimated to have occurred approximately 100 million years ago (Doolittle et al. 1996; Ochman and Wilson 1987; 1988), and the average divergence at nonsynonymous and synonymous codon positions has been calculated as 4.5% and 91%, respectively (Sharp 1991). Assuming a molecular clock, these divergence values provide rates of divergence of 0.045% and 0.91% per million years at nonsynonymous and synonymous sites (Table 1). The substitution rate is one-half of this value (assuming equal rates of evolution along the *E. coli* and *S. enterica* lineages), yielding the substitution rates of 0.0245% and 0.455% per Myr per lineage. Because most codons bear both synonymous and nonsynonymous sites, the rates of substitution at each codon position can be calculated as weighted averages of the overall substitution rates at nonsynonymous and synonymous sites (see Table 1).

Applying the equations derived above, we have modeled the amelioration of horizontally transferred DNA (Fig. 4). First, ΔGC values are calculated for each codon position using both the codon-position-specific rate of change (Table 1) and the codon-position-specific GC content (Fig. 2) of the native genome. Second, the GC content of the horizontally transferred DNA is incrementally adjusted by the ΔGC values to predict the composition of the DNA in 1 million years—after which, some nucleotide substitutions could have altered the nucleotide content of the DNA. This process is repeated successively to estimate the change in nucleotide composition at each codon position of a gene over an evolutionary time scale. As is evident from inspection of Table 1, the third codon positions—those experiencing the weakest selection—will ameliorate quickly, while first and second codon positions will ameliorate more slowly (Fig. 4).

Acquisition and Amelioration of spa *Genes by Enteric Pathogens*

As shown above, it is possible to model the process of amelioration and predict the rate and patterns of compositional change in sequences acquired through horizontal transfer. However, observation of how these processes have influenced the evolution of genes in natural populations requires knowledge of two variables—the original base composition of an introgressed sequence and its

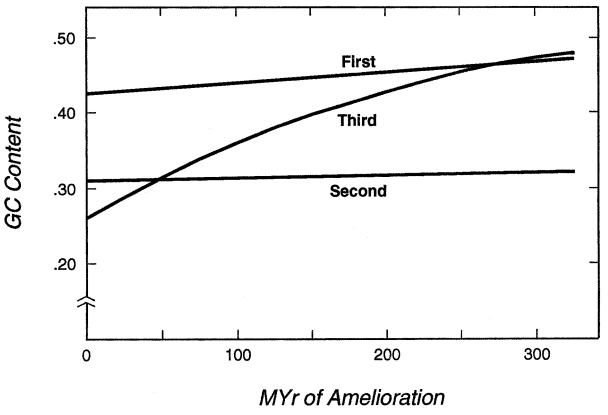

Fig. 4. The GC contents of the three codon positions of the *Shigella flexneri spa* genes are plotted for 325 iterations of 1 Myr of amelioration toward the base composition of the *Salmonella enterica* genome.

duration in the new host. The age and origin of a sequence can be traced by analyzing the occurrence of homologous regions among lineages with known phylogenetic relationships and divergence times. However, the original base composition of an introgressed sequence is more difficult to determine. The *spa* genes of enteric bacteria may provide both of these parameters.

Several bacterial pathogens utilize a common pathway to export the proteins required for host cell entry [see review by Barinaga (1996)]. The size, organization, and nucleotide sequences of the *inv/spa* genes of *Salmonella enterica*—which are required for bacterial entry into mammalian intestinal cells—correspond to those of the *mxi/spa* complex from the *Shigella flexneri* virulence plasmid (Fig. 5). Many of the proteins encoded by the corresponding gene clusters have been shown to perform analogous functions (Ginocchio and Galán 1995; Groisman and Ochman 1993; Maurelli 1994). But despite broad-scale similarities in the structure and function of these gene clusters, their physical properties—as well as their genomic locations and phylogenetic distribution—indicate that these clusters were acquired independently by *Shigella* and by *Salmonella* (Groisman and Ochman 1993; Ochman and Groisman 1995). The *spa* region is ancestral to all serovars of *Salmonella enterica* (Li et al. 1995) and, therefore, must have been incorporated into the *S. enterica* genome prior the appearance of *Shigella flexneri,* which is much more closely related to *E. coli* than is *S. enterica* (Ochman et al. 1983); the placement of *Shigella* species into a separate genus is based solely on the pathogenic character of these strains, not on their relationship with other entire species. Additional features preclude *S. enterica* as the source of the *Shigella flexneri* invasion genes: the *mxi/spa* region of *S. flexneri* has a base composition of only 33% GC (compared to 47% GC in *S. enterica*) and resides on a large plasmid containing several additional virulence loci with no known homologues in *S. enterica.*

This unusual evolutionary scenario, whereby two species have each independently acquired the same set of genes, allows us to examine how sequences evolve following incorporation into distinct genetic backgrounds. The GC contents of the *mxi/spa* genes of *Shigella* are fairly homogeneous, as expected for genes that were acquired relatively recently from a single source. Therefore, the nucleotide compositions of these genes are likely to be similar to the base composition of this region in its original host (Fig. 5). In contrast, the overall GC content and the range of variation in GC contents are much higher in the *inv/spa* genes of *Salmonella* than in the corresponding genes from *Shigella.* These differences reflect the longer residence of this region in the *Salmonella* genome. Each *inv/spa* gene in *Salmonella* has ameliorated to a higher GC content than its *Shigella* homologue (Fig. 5); the degree of amelioration is a function of both the amino acid composition of the encoded protein and the intensity of selection.

Because the process of amelioration is caused principally by mutational biases, the effects of this process are most obvious at sites having little or no selective constraints. And indeed, the most pronounced differences in GC contents between the *spa* gene clusters of *Shigella* and *Salmonella* are seen at third codon positions, where more than 70% of possible substitutions are synonymous (Table 2). Moreover, when one examines the change in base composition at each codon position, the degree of variation in base composition is lowest at third codon positions; the lower variance in base composition is expected if sites are free from selective constraints and substitutions are neutral. In contrast, variable degrees of selection for individual protein sequences constrain substitutions at first and second codons positions, and these varying constraints lead to greater variation in base composition at these codon positions.

Based on phylogenetic information, it is possible to estimate the times when the *spa* gene clusters were acquired by *S. enterica* and *S. flexneri.* Because the *spa* genes are present in all subspecies of *S. enterica,* but absent from other closely related enteric species, these genes were acquired after *Salmonella* diverged from *E. coli* some 100 Myr ago (Li et al. 1995). Furthermore, virulence in all species of *Shigella* depends upon the presence of an invasion plasmid (which contains the *spa* genes), and the emergence of virulent strains of *Shigella*—from its nonvirulent parent species, *E. coli*—has been dated to 25 million years ago based on the extent of sequence divergence (Ochman et al. 1983). However, it is clear from the levels of similarity between homologues—encoded proteins range from 19% to 64% identical (Fig. 5)—that the *spa* genes of *Shigella* and *Salmonella* probably diverged more than 100 Myr ago; typical homologous proteins from *E. coli* and *S. enterica* are, on average, 93% identical (Sharp 1991).

If the base compositions of the *spa* genes in *Shigella* are similar to those in their previous host, analysis reveals that the *Salmonella spa* genes would require on

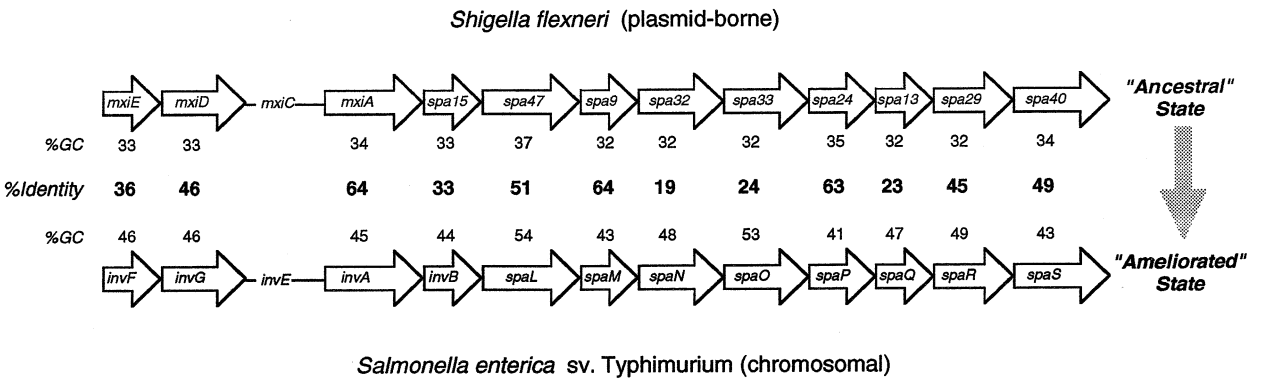

Fig. 5. Genes of the *mxi/spa* gene cluster in *Shigella flexneri* and the *inv/spa* gene cluster in *Salmonella enterica* sv. Typhimurium. Pairs of homologous genes are displayed in corresponding positions in the two gene clusters; the directions of the *arrows* denote the direction of transcription. Overall GC content of each gene and percent identity between the encoded pairs of proteins are noted. (Orientation and complete sequence information is not available for the *mxiC* and *invE* genes.)

Table 2. Compositional structure of the *spa* operons of *Salmonella enterica* and *Shigella flexneri*

	Base composition (% GC)[a]			
	Overall	1[b]	2[b]	3[b]
Salmonella spa	46.6	53.8 ± 4.2	36.8 ± 4.5	48.7 ± 4.1
Shigella spa (native)	33.3	42.8 ± 3.2	30.9 ± 4.1	25.9 ± 2.5
Difference[c]	13.3	10.9 ± 4.9	5.9 ± 5.0	22.8 ± 6.3
Shigella spa (ameliorated)[d]	42.5	47.1 ± 2.6	31.7 ± 3.9	48.7 ± 3.6
Difference[e]	41.1 ± 3.7	6.6 ± 4.5	5.1 ± 4.9	0.0 ± 2.2
Salmonella genome	52.1	58.8	38.1	57.4

[a] Unweighted averages for 12 pairs of homologous genes present in both *Salmonella enterica* and *Shigella flexneri* (see Fig. 5)
[b] Base composition at first, second, and third codon positions
[c] Difference between *Salmonella spa* and *Shigella spa* (native)
[d] Predicted nucleotide composition of *Shigella spa* genes after 325 years of amelioration
[e] Difference between *Salmonella spa* and *Shigella spa* (ameliorated)

average more than 300 Myr of amelioration (in a *Salmonella*-like genome) to reach their present GC contents (see Fig. 4). After this time, the third codon positions of the *spa* genes of *Shigella* would change from 25.9% to 47.8% GC (Table 2, Fig. 4) and resemble the GC contents of the corresponding genes from *Salmonella.* However, even after 325 Myr of amelioration toward the base composition of the *Salmonella* genome, the first and second codons positions of these genes would still manifest anomalous GC contents. The discrepancies in the amelioration times necessary to transform the base compositions at each codon position suggest that the base compositions of the first and second codon positions of the *Salmonella spa* genes have changed much more quickly than expected. This rapid amelioration was most likely due to higher rates of nucleotide substitution at these sites than experienced by typical *S. enterica* genes (Table 1) and may have resulted from either relaxed, or potentially diversifying, selection. The amelioration analyses suggest that the base compositions of the first and second codon positions of the *S. enterica spa* genes reflect substitution rates that are at least tenfold higher than that seen for typical chromosomal genes. It is not surprising that the *spa* genes may have experienced diversifying selection since they transport antigenic determinants that influence pathogenicity. Alternatively, the evolutionary history of the *spa* genes may be far more complicated, involving multiple transfer events among many host genomes; however, such evolutionary scenarios do not provide directly testable hypotheses regarding the evolution of base composition.

Amelioration in Reverse

The analysis of the *spa* genes in *Shigella* and *Salmonella* was possible because their original GC contents, as well as their probable times of introduction, could be inferred; however, this information is not readily available for most horizontally transferred genes. Because sequences transferred to a genome of distinct base composition will ameliorate with time, it should be possible to ''back-ameliorate'' a sequence to determine when it was introduced into the new host genome. Yet, this is not simply a matter of applying the equations derived above to obtain the amount of time necessary to reach a certain base

composition because the original GC content of the transferred fragment is rarely known. However, the codon-position-specific base compositions reflect the history of a gene within the genome; because each codon position has a characteristic rate of substitution, comparisons of the extent of change at each position can reveal the amount of time that a particular sequence has been subjected to the directional mutation pressures of the resident genome.

By analyzing the characteristics of the sequences within the 1.43-mb contiguous sequence from *E. coli*, we estimated that 618 kb of DNA has been introduced into the *E. coli* genome by horizontal transfer (see above). To determine the rate at which novel DNA is incorporated by *E. coli*, we must estimate the average length of time these horizontally transferred sequences have been present in the *E. coli* chromosome. Based on the relationships shown in Fig. 1, we have derived the following linear equations—whose predictive value extend only over the biologically significant range of genomic base compositions (20% to 80% GC)—to approximate the patterns of nucleotide composition across codon positions:

$$GC_{1st} = 0.615 \times GC_{Genome} + 26.9 \qquad (13)$$

$$GC_{2nd} = 0.270 \times GC_{Genome} + 26.7 \qquad (14)$$

$$GC_{3rd} = 1.692 \times GC_{Genome} - 32.3 \qquad (15)$$

The nucleotide compositions of protein-coding regions within a given bacterial species should conform to the values predicted by these Muto and Osawa (MO) relationships; however, horizontally transferred genes that are undergoing the process of amelioration will not. Each codon position is subject to different selective constraints (Tables 1 and 2) which cause the nucleotide compositions of each codon position of horizontally transferred genes to ameliorate at different rates.

These differences in the rates of amelioration at each codon position furnish a property unique to horizontally transferred genes and allow us to estimate the amount of time the gene has been residing in a genome. Recently transferred genes show the patterns of nucleotide composition typical of the donor genome, and fully ameliorated genes show the nucleotide compositions of the recipient genome; however, the nucleotide compositions of genes during the process of amelioration do not resemble those of either the donor or recipient genomes. But most importantly, the base compositions of genes in the process of amelioration do not reflect any of the patterns predicted by the curves in Fig. 1; this feature is evident in the *S. enterica spa* genes. The amelioration equations derived above predict how nucleotide composition changes over time when sequences are placed in novel genomes. These equations can be employed to estimate the original nucleotide composition of horizontally transferred genes that do not conform to the MO relationships due to the process of amelioration. In this case, GC contents of the three codon positions can each be ''back-ameliorated'' until the sequence conforms to the MO relationships, thereby providing estimates both of the amount of time that an introgressed gene has been ameliorating—and, therefore, of the likely time of introduction into the recipient genome—and of the nucleotide composition of the donor genome (Fig. 6).

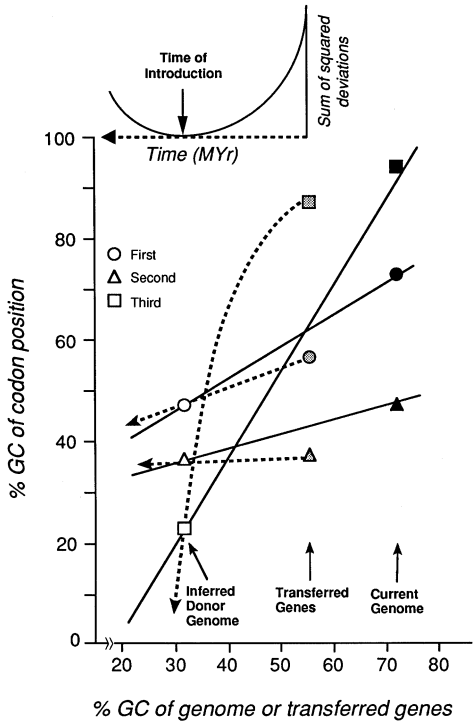

Fig. 6. Application of reverse amelioration to estimate the duration of introgressed genes in a novel genetic background. GC contents of the hypothetical recipient genome, and of genes suspected of being acquired by horizontal transfer, are superimposed on the curves (equations 13–15) derived by Muto and Osawa (1987) and displayed in Fig. 1. *Dashed lines* depict the projected GC contents of each codon position predicted by reverse-amelioration algorithms. The graph in the upper left displays the least-squares analysis of the deviation of the values predicted by amelioration from the MO relationships. Time of introgression into the recipient genome is taken as the point of smallest least-squares deviation from the MO relationships.

To estimate the time of introgression of a DNA sequence, the amelioration equations are applied in reverse to estimate the previous GC content of each codon position. In this procedure, the change in GC content, ΔGC, is subtracted from the GC content of the horizontally transferred gene to predict what the nucleotide composition at each codon position had been prior to amelioration. Calculations are performed iteratively, at 1-million-year intervals, until the predicted codon-specific GC contents yield a minimum least-squares difference from the MO relationships (Fig. 6). The time of introgression, expressed as millions of years before present, is assigned to that point.

Time of Introduction of the Salmonella enterica cob *operon*

To test the model of reverse amelioration, we first analyzed the sequence of the *cob* operon of *Salmonella enterica* sv. Typhimurium. The *cob* operon encodes the proteins responsible for cobalamin (vitamin B_{12}) biosynthesis, and previous studies have shown that the *cob* operon was obtained by the *Salmonella* lineage following the divergence of this species group from *E. coli* (Lawrence and Roth 1995, 1996a). As for the *inv/spa* gene cluster described above, these data place an upper bound on the time of introduction of the *cob* operon into the *Salmonella* lineage; the introgression of the *cob* operon must have occurred less than 100 Myr ago. Furthermore, the *cob* operon is almost universally conserved among *Salmonella* species, placing the likely time of its introduction into *Salmonella* prior to the radiation of the present-day lineages, approximately 50 Myr ago. Therefore, the distribution of the *cob* operon among enteric bacteria places both upper and lower bounds on the time of introgression of the *cob* operon into the ancestor of *Salmonella enterica.* When the amelioration equations were applied to the 4,988 codons of the *cob* operon of *S. enterica* sv. Typhimurium (Chen et al. 1995; Roth et al. 1993), the least-squares analysis reveals that the *cob* operon was introduced into the *Salmonella* lineage 71 Myr ago, an estimate which falls within the time span predicted by genetic and evolutionary analyses.

To obtain confidence intervals for these estimates, and to test the behavior of the reverse amelioration model, random subsets of codons of the *cob* operon were sampled and analyzed by the reverse amelioration algorithms. Two types of analyses were performed. First, ($N - 1$) codons were selected at random with replacement—a process termed ''bootstrapping'' (Efron and Gong 1983)—and nucleotide compositions were determined from the resampled pool of codons. Then, the reverse-amelioration process was repeated on the bootstrapped subsample to yield an estimate of the hypothetical time of introduction of the resampled data set. [The use of the bootstrap for establishing the confidence intervals of complex measures has been well established in the biological literature (e.g., Felsenstein 1985; Lawrence and Hartl 1992).] This process was performed on 1,000 bootstrapped replicates to generate a distribution of hypothetical times of introduction. Confidence intervals were derived from the distribution of these values, as the variance of this distribution is equivalent to the variance of the time of introduction calculated for the entire *cob* operon. The analysis of the *cob* operon sequence predicts a time of introgression 71 Myr, with a 90% confidence interval of 62 to 89 Myr ago, showing that the divergence of the *S. enterica* and *E. coli* lineages (100 Myr ago) significantly preceeded the time of introduction of the *cob* operon into the *S. enterica* genome. The use of a 90% confidence interval reflects the distribution of amelioration times generated by the bootstrapped data sets. Since amelioration times are bounded on one side (0 Myr ago) but not bounded on the other, the distribution has one extremely long tail, where an occasional dataset failed to ameliorate (leading to an artificial 2-billion-year amelioration time). To eliminate these aberrant points, 90% confidence intervals were employed in subsequent analyses.

A second analysis was performed to determine the minimum number of codons required to accurately predict the time of introgression of the *cob* operon. The *cob* operon's 4,988 codons can provide an accurate estimate of codon-specific GC contents. Fewer codons are likely to yield less reliable estimates of GC contents and, in turn, times of introgression. To test the effect of sample size on the estimates of introgression times, small subsets of codons were selected from the *cob* operon and analyzed by reverse amelioration. We found that samples of fewer than 1,500 codons (4.5 kb of protein-coding DNA) provide unreliable estimates of nucleotide composition, which will lead to less accurate estimates of the time of introgression. For example, subsamples of fewer than 1,500 codons of the *cob* operon provided estimates of the time of introduction whose 90% confidence intervals exceed the earliest acquisition date of 100 Myr. We conclude that these subsamples provided too few codons to accurately estimate base compositions.

Horizontal Transfer and Genomic Flux

We identified 186 kb of horizontally transferred, protein-coding DNA in a 1.43-mb continuous DNA sequence from the *E. coli* chromosome. Ideally, we would estimate the time of introduction of each segment of DNA inferred to be of exogenous origin; however, the nature of the reverse-amelioration of algorithms requires knowledge of the GC contents of each codon position. As detailed above, these parameters are difficult to estimate for DNA sequences fewer than 4,500 nucleotides, a length well above that of the typical bacterial gene. Therefore, to estimate times of introduction, the horizontally transferred genes were sorted into 33 groups, each comprising genes that bear similar nucleotide compositions at each codon position (Table 3); a caveat to this method is discussed below.

To obtain an error estimate for the times of introduction, we performed a bootstrap analysis—as described above for the *cob* operon—on the pools of DNA sequences to obtain confidence intervals for the times of introduction (Table 3). The nucleotide compositions of the groups of pooled genes were analyzed by the reverse-amelioration algorithms to estimate the time of introduction for each set of genes (Table 3). The times of introduction for the 22 of the 33 pools were accurately estimated by the reverse-amelioration algorithms (Fig.

Table 3. Pools of horizontally transferred genes found in the *E. coli* genome

Pool	Genes[a]	Bases[b]	Average GC content				Time ameliorating[d] (Myr)	Range[e] (Myr)
			ORF	1[c]	2[c]	3[c]		
1	4	2682	29.1	37.1	27.3	23.0	17	12–24
2	9	7479	33.8	41.9	33.3	26.3	26	24–29
3	2	879	32.8	34.8	31.4	32.1	22	14–34
4	6	6135	33.7	39.5	30.2	31.3	na[f]	
5	5	6768	36.7	44.8	32.4	33.1	93	85–109
6	4	4470	39.7	49.6	37.7	31.7	0	0–4
7	4	2586	35.2	35.0	26.8	39.4	na	
8	7	4446	38.2	44.1	34.2	36.4	na	
9	8	6414	41.6	48.9	37.6	38.2	17	12–22
10	7	5595	44.2	54.3	40.3	38.0	0	0–4
11	5	3480	45.6	59.7	39.6	37.4	na	
12	11	6243	41.9	45.0	37.9	42.9	na	
13	8	7620	43.3	48.7	38.7	42.5	37	29–48
14	7	6492	44.5	51.6	40.4	41.6	22	17–26
15	9	6660	45.0	55.5	36.8	42.8	15	9–24
16	11	5118	46.6	59.8	37.0	42.8	na	
17	7	4056	49.6	65.1	39.8	43.9	na	
18	9	8367	44.4	45.9	39.0	48.2	54	42–69
19	6	4893	45.9	49.4	40.4	47.9	na	
20	10	8943	46.5	52.5	39.0	47.9	3	0–12
21	12	9712	47.7	56.0	39.8	47.4	2	0–7
22	10	6471	49.4	60.2	39.9	48.0	0	0–5
23	6	5097	50.4	64.3	38.6	48.3	na	
24	7	3894	48.3	50.4	42.2	52.5	63	58–72
25	8	5304	53.1	62.6	40.8	55.9	98	91–112
26	4	4281	51.0	53.4	37.5	62.2	48	42–57
27	6	7533	54.1	60.9	38.6	62.8	0	0–3
28	4	7350	57.3	61.3	43.8	66.9	26	24–29
29	8	6777	58.2	65.5	42.5	66.7	4	0–13
30	6	6687	57.1	59.0	41.3	71.2	na	
31	9	5802	60.2	65.4	43.7	71.4	0	0–4
32	4	2706	58.4	56.6	42.0	76.6	na	
33	5	4521	61.6	67.9	41.6	75.4	1	0–5
AVE[g]	229	185461					25.2	

[a] Number of genes in each pool
[b] Number of nucleotides in each pool
[c] Base composition at first, second, and third codon positions
[d] Time of introgression of the genes in each pool calculated by reverse amelioration
[e] 90% confidence intervals on the time of introgression calculated by bootstrap analysis of nucleotide sequences in each pool
[f] na—not applicable. (Reverse amelioration process did not provide a significant least-squares estimate of the time of introgression.)
[g] The sum of ''Genes'' and ''Bases'' columns and the weighted average of the ''Time Ameliorating'' column

7). (In each case, the fit of the ameliorated base compositions to the Muto and Osawa relationships was better—that is, showed a smaller sum of squared deviations—than the average fit of the base compositions defining equations 13–15.) Many genes have been introduced recently, that is, within the last 10 million years; the average time of introduction for all pools, calculated as a weighted mean of the individual estimates, is 25.3 million years ago (Table 3). The paucity of genes identified as having been introduced more than 10 million years ago can be attributed to two factors: (1) the amelioration of genes beyond our ability to detect them as aberrant and (2) the deletion of the more ancient genes from the chromosome.

In 11 cases, the analysis of pooled sequences was discarded as unreliable. In these cases, the fit of the ameliorated base compositions to the Muto and Osawa relationship was far poorer—that is, showed a larger sum of squared deviations—than the average fit of the base compositions defining equations (13–15). For these pools, the reverse-amelioration equations could not find a reasonable fit to the MO relationships due to the anomalous nucleotide compositions.

Some introgressed genes show high degrees of codon usage bias, as seen in high χ^2 of codon usage (see above and Fig. 3). Since these genes do not accurately reflect the base composition of their native genome—such as the *rplL* gene, which does not reflect the base composition of *E. coli*—amelioration analysis cannot be performed. The 229 *E. coli* genes were partitioned into 33 narrowly defined pools, each contrived to have low variance in base composition. This process allowed for ac-

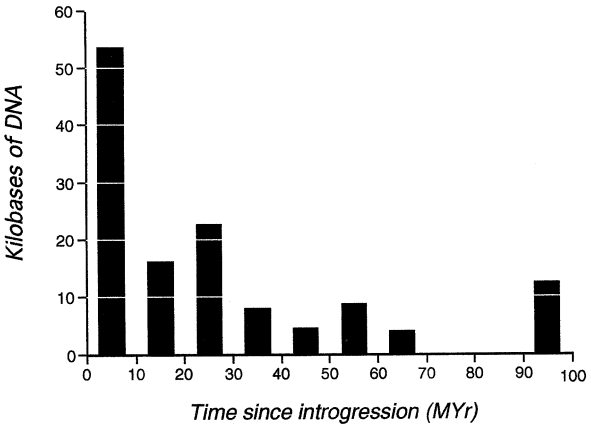

Fig. 7. Times of introgression of DNA fragments into the *E. coli* chromosome. The amount of DNA reflects cumulative pooled sequences (see Table 3).

curate back-amelioration of 22 pools, yielding amelioration times with relatively narrow 90% confidence intervals. The excluded genes include those whose original base compositions did not reflect their donor genomes. Because the exclusion of these genes modifies the variance—but not the mean—of the GC contents of a subsample, the lack of these genes from the remaining pools does not affect the overall results of the reverse-amelioration analysis. When these genes are included in the 22 remaining pools, the resulting amelioration times have larger variances.

Although genome size may vary within an enteric species (Bergthorsson and Ochman 1995; Harsono et al. 1993), the average sizes of enteric bacterial chromosomes are fairly congruent, suggesting that the bacterial chromosomes are not constantly increasing in size. Therefore, the gain of 618 kb of DNA by horizontal transfer must be offset by an approximately equal loss of DNA over the long-term evolution of these species. Because some of the DNA that is lost will be of exogenous origin, the 618 kb is a minimum estimate of amount of DNA that has been introduced into the *E. coli* chromosome. Correcting for the deletion of some acquired sequences—assuming that all horizontally transferred sequences and 50% of the existing chromosomal sequences are equally likely to be deleted—we estimate that 803 kb of DNA has been introduced into the *E. coli* chromosome; 185 kb of horizontally transferred DNA has been lost by deletion and 618 kb remains.

If the average time of introduction of the 803 kb of horizontally transferred genes identified in the *E. coli* chromosome is 25.3 million years ago (Table 3), the rate of introduction can be estimated at 31 kb per Myr. Given a divergence time between *E. coli* and *S. enterica* of 100 million years ago, this rate predicts that each lineage has gained and lost approximately 3 mb of protein-coding DNA since these species diverged. We detect only 618 kb of this DNA because the majority of the horizontally transferred DNA either has ameliorated or has been lost by deletion. Because the chromosomes of these species contain, on average, 4.8 mb of DNA, each lineage has gained and lost a total amount of DNA corresponding to 60% of the size of its current genome.

Discussion

Based on an analysis of a sequenced region encompassing more than 30% of the *E. coli* chromosome, we estimate that 17% of the protein-coding sequences were obtained through horizontal transfer. Although this value is very similar to the amount of acquired DNA reported by Médigue et al. (1991), it is probably an underestimate of the amount of horizontally transferred DNA present in the *E. coli* chromosome because our analysis would detect neither sequences obtained from genomes that are similar to *E. coli* nor fully ameliorated genes whose features are indistinguishable from ancestral sequences. The estimate of 17% horizontally transferred genes is higher than the amount of unique DNA observed in alignments of the genetic maps of *E. coli* K-12 and *S. enterica* sv. Typhimurium LT2. The lineage maps of these enteric species are nearly 90% identical, suggesting that roughly 10% of the chromosome arose through either horizontal transfer of DNA into, or deletion of DNA from, either lineage. However, the linkage maps of these two species are incomplete and incongruities of less than 10 kb are rarely detected in these comparisons. Furthermore, the phenotypes of most horizontally transferred genes have not been specified and relatively few of these genes have been mapped.

Although little of the DNA introduced by horizontal transfer remains in the extent *E. coli* chromosome, sequences acquired through horizontal transfer can have considerable impact of the evolution of a bacterial species. Many features distinguishing *E. coli* and *S. enterica* can be attributed to horizontally transferred genes. For example, the *lac* operon, allowing for the degradation of β-galactosides, was introduced into the ancestor of *E. coli* by horizontal transfer (Buvinger et al. 1984), and the *cob* and *pdu* operons, providing for vitamin B_{12} biosynthesis and the B_{12}-dependent degradation of propanediol by the *S. enterica*, also arose through horizontal transfer (Lawrence and Roth 1995, 1996a). Moreover, many of the genes implicated in *Salmonella* pathogenesis—such as those required for host-cell invasion (Groisman and Ochman 1993) and intramacrophage survival (Ochman et al. 1996)—show several characteristics of horizontally acquired sequences.

By calculating the temporal duration of horizontally transferred genes in the *E. coli* chromosome, we estimate that sequences are acquired and lost at a rate of 31 kb per million years. Based on this rate, the *E. coli* chromosome has gained and lost over 3 mb of DNA—roughly 60% of the size of its current genome—since diverging from

Salmonella. In comparison to the amount of DNA gained through horizontal transfer, applying a substitution rate of 0.455% per million years per lineage—i.e., the synonymous substitution rate which should be similar to the actual rate of mutation across the *E. coli* genome—yields approximately 22,000 point mutations, or 22 kb of variant DNA, occurring in the *E. coli* genome per million years. Although quantitatively similar amounts of variation are introduced through mutation and through horizontal transfer, the types of genetic information furnished by these processes are qualitatively very different. Nucleotide substitutions arising by mutation—or by intragenic recombination among strains, which is thought to generate about the same level of *genic* diversity as point mutations (Guttman and Dykhuizen 1994a,b; Milkman and Bridges 1990, 1993)—would provide modest changes in the encoded phenotype and would only rarely confer novel characteristics to the organism.

Although many horizontally transferred sequences may have little impact on the organism, acquired sequences can provide a novel function immediately upon introgression and, in effect, change the character of the species. The introduction of operons conferring novel metabolic capabilities (e.g., the *lac*, *spa*, *cob*, and *pdu* operons) may have allowed the ancestors of *E. coli* and *S. enterica* to explore novel ecological niches that were previously unavailable. If bacterial diversification and speciation is dependent upon the competitive invasion and exploitation of ecological niches, genes gained through horizontal transfer would be more likely to promote this process than would ancestral sequences altered by mutation and selection.

Acknowledgments. We thank D. Dykhuizen, E. Kofoid, J. Roth, and P. Sharp for enlightening discussions and G. Fox, R. Milkman, and an anonymous reviewer for helpful comments on the manuscript. This work was supported by grants GM-15868 (J.G.L.) and GM-48407 (H.O.) from the National Institutes of Health.

References

Barinaga M (1996) A shared strategy for virulence. Science 272:1261–1263

Bergthorsson U, Ochman H (1995) Heterogeneity of genome size among natural isolates of *Escherichia coli.* J Bacteriol 177:5784–5789

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Burland V, Plunkett III G, Daniels DL, Blattner FR (1993) DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. Genomics 16:551–561

Buvinger WE, Lampel KA, Bojanowski RJ, Riley M (1984) Location and analysis of nucleotide sequences at one end of a putative *lac* transposon in the *Escherichia coli* chromosome. J Bacteriol 159: 618–623

Chen P, Ailion M, Weyland N, Roth J (1995) The end of the *cob* operon: evidence that the last gene (*cobT*) catalyzes synthesis of the lower ligand of vitamin B_{12}, dimethylbenzimidazole. J Bacteriol 177:1461–1469

Daniels DL, Plunkett III G, Burland V, Blattner FR (1992) Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. Science 257:771–778

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12: 387–395

Doolittle RF, Feng DF, Tsang S, Cho G, Little E (1996) Determining the divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Efron B, Gong G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. Am Stat 37:36–48

Felsenstein J (1985) Confidence intervals on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Ginocchio CC, Galàn JE (1995) Functional conservation among members of the *Salmonella typhimurium* InvA family of proteins. Infect Immun 63:729–732

Groisman EA, Ochman H (1993) Cognate genes govern invasion of host epithelial cells by *Salmonella typhimurium* and *Shigella flexneri.* EMBO J 12:3779–3787

Gross TJ, Schweizer HP, Datta P (1988) Molecular characterization of the *tdc* operon of *Escherichia coli* K-12. J Bacteriol 170:5352–5359

Guttman DS, Dykhuizen DE (1994a) Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380–1383

Guttman DS, Dykhuizen DE (1994b) Detecting selective sweeps in naturally occurring *Escherichia coli.* Genetics 138:993–1003

Harsono KD, Caspar CW, Luchansky JB (1993) Comparison and genomic sizing of *Escherichia coli* O157:H7 isolates by pulsed-field gel electrophoresis. Appl Environ Microbiol 59:3141–3144

Jiang W, Metcalf WW, Lee KS, Wanner BL (1995) Molecular cloning, mapping, and regulation of Pho regulon genes for phosphonate breakdown by the phosphonatase pathway of *Salmonella typhimurium* LT2. J Bacteriol 177:6411–6421

Kidwell M (1993) Lateral transfer in natural populations of eukaryotes. Annu Rev Genet 27:235–256

Klena JD, Pradel E, Schnaitman CA (1993) The *rfaS* gene, which is involved in production of a rough form of lipopolysaccharide core in *Escherichia coli* K-12, is not present in the *rfa* cluster of *Salmonella typhimurium* LT2. J Bacteriol 175:1524–1527

Lawrence JG, Hartl DL (1992) Inference of horizontal genetic transfer: an approach using the bootstrap. Genetics 131:753–760

Lawrence JG, Roth JR (1995) The cobalamin (coenzyme B_{12}) biosynthetic genes of *Escherichia coli.* J Bacteriol 177:6371–6380

Lawrence JG, Roth JR (1996a) Evolution of coenzyme B_{12} synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. Genetics 142:11–24

Lawrence JG, Roth JR (1996b) Selfish operons—horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860

Li J, Ochman H, Groisman EA, Boyd EF, Nelson K, Solomon F, Selander RK (1995) Relationship between evolutionary rate and cellular location among the Inv/Spa invasion proteins of *Salmonella enterica.* Proc Natl Acad Sci USA 92:7252–7256

Maurelli AT (1994) Virulence protein export systems in *Salmonella* and *Shigella*: a new family or lost relatives. Trends Cell Biol 4: 240–242

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

Metcalf WW, Wanner BL (1993) Evidence for a fourteen-gene, *phnC* to *phnP* locus for phosphonate metabolism in *Escherichia coli.* Gene 129:27–32

Milkman R, Bridges MM (1990) Molecular evolution of the *E. coli* chromosome. III. Clonal frames. Genetics 126:505–517

Milkman R, Bridges MM (1993) Molecular evolution of the *E. coli* chromosome. IV. Sequence comparisons. Genetics 133:455–468

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Ochman H, Groisman EA (1995) The evolution of invasion in enteric bacteria. Can J Microbiol 41:555–561

Ochman H, Lawrence JG (1996) Phylogenetics and the amelioration of bacterial genomes. In: Neidhardt FC, Curtiss III R, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella*: cellular and molecular biology. 2nd ed. American Society for Microbiology, Washington, DC, pp 2627–2637

Ochman H, Wilson AC (1987) Evolutionary history of enteric bacteria. In: Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaechter M, Umbarger HE (eds) *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society for Microbiology, Washington, DC, pp 1649–1654

Ochman H, Wilson AC (1988) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol 26:74–86

Ochman H, Whittam TS, Caugant DA, Selander RK (1983) Enzyme polymorphism and genetic population structure in *Escherichia coli* and *Shigella.* J Gen Microbiol 129:2715–2726

Ochman H, Soncini FC, Solomon F, Groisman EA (1996) Identification of a pathogenicity island required for *Salmonella* survival in host cells. Proc Natl Acad Sci USA 93:7800–7804

Plunkett G, Burland V, Daniels D, Blattner F (1993) Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes. Nucleic Acids Res 21:3391–3398

Rolfe R, Meselson M (1959) The relative homogeneity of microbial DNA. Proc Natl Acad Sci USA 45:1039–1042

Roth JR, Lawrence JG, Rubenfield M, Kieffer-Higgins S, Church GM (1993) Characterization of the cobalamin (vitamin B_{12}) biosynthetic genes of *Salmonella typhimurium.* J Bacteriol 175:3303–3316

Schildkraut CL, Marmur J, Doty P (1962) Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J Mol Biol 4:430–443

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium.* J Mol Evol 33:23–33

Sharp PM, Li W-H (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sofia HJ, Burland V, Daniels DL, Plunkett G, Blattner FR (1994) Analysis of the *Escherichia coli* genome. V. DNA sequence of the region from 76.0 to 81.5 minutes. Nucleic Acids Res 22:2576–2586

Sueoka N (1961) Variation and heterogeneity of base composition of deoxy-ribonucleic acids: a compilation of old and new data. J Mol Biol 3:31–40

Sueoka N (1962) On the genetic basis of variation and heterogeneity in base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114

Sueoka N (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J Mol Evol 37:137–153

Sueoka N, Marmur J, Doty P (1959) Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. Nature 183:1429–1431

Syvanen M (1994) Horizontal gene flow: evidence and possible consequences. Annu Rev Genet 28:237–261

Wanner BL, Metcalf WW (1992) Molecular genetic studies of a 10.9-kb operon in *Escherichia coli* for phosphonate uptake and biodegradation. FEMS Microbiol Lett 79:133–139

Whittam TS, Ake S (1992) Genetic polymorphisms and recombination in natural populations of *Escherichia coli.* In: Takahata N, Clark AG (eds) Mechanisms of molecular evolution. Japan Scientific Society Press, Tokyo, pp 223–246

Appendix I. Genes and open reading frames in a 1.43-mb region of the *E. coli* genome that were identified as horizontally transferred

Gene	Position	Length	Gene	Position	Length	Gene	Position	Length
f138	14,479	417	o224	142,857	675	*yhdM*	219,453	426
f104	14,892	315	o793	143,552	2382	*yhdN*	219,889	369
o118	33,664	357	o159	145,930	480	f139	234,256	420
f406	43,206	1,221	f338	146,447	1017	f489	234,667	1,480
tdcC	44,440	1,332	o238	147,671	717	o271	236,326	816
tdcB	45,793	990	o104	166,232	315	*bfr*	246,997	477
tdcA	46,881	939	f90	166,602	273	f199	255,721	360
tdcR	48,008	345	o59	193,367	180	o55	280,195	168
o186	48,587	561	f220	193,549	663	f146	337,150	441
o395	49,169	1,188	o252	203,181	759	o392	362,600	1,179
o54	50,417	165	o256	209,766	771	o138	363,775	417
sohA	57,755	336	f85	210,512	258	*insA*	364,220	276
o154	58,090	465	*aroE*	210,766	819	*insB*	364,414	504
o158	64,922	477	f176	211,589	531	o94	365,496	285
o194	68,178	585	f169	212,167	510	*nikA*	394,404	1,575
o231	68,842	696	*smg*	212,739	474	*nikB*	395,978	945
o863	69,491	2,592	f102	213,184	309	*nikC*	396,919	834
o429	72,093	1,092	*smf*	213,546	762	*nikD*	397,752	765
f167	81,007	504	*fms*	214,437	510	*nikE*	398,513	807
f547	100,740	1,644	*fmt*	214,961	948	*yhhG*	399,325	402
o69	123,381	210	*fmu*	215,954	1,161	*rhsB*	399,929	4,236
gltB	135,377	4,554	*trkA*	217,266	1,377	*yhhH*	404,151	369
gltF	141,921	765	*mscL*	218,772	411	*yhhI*	405,115	1,137

Appendix I. Continued

Gene	Position	Length	Gene	Position	Length	Gene	Position	Length
f231	409,528	696	o155	678,247	468	o98	1,133,113	297
f540	411,703	1,623	o195	678,763	588	o133a	1,190,662	402
f123	413,587	372	f538	681,342	1,617	o228a	1,204,233	687
f409	413,963	1,230	o161	774,743	297	f91	1,205,049	276
o383	415,476	1,152	f125	780,996	378	f264	1,207,161	795
f111	420,119	336	f126	783,120	381	o165	1,218,240	501
o260	432,025	783	f138	783,514	417	o110	1,253,043	333
orf	432,916	1,017	f170	821,546	513	f131	1,253,344	396
slp	434,662	600	*mob*	822,055	585	o164	1,254,759	495
yhiF	435,417	531	o89	822,709	270	o616	1,255,937	1,851
f215	435,989	648	o328	823,056	987	o105	1,260,345	318
f112	436,700	339	*dsbA*	824,059	627	o204b	1,260,845	615
f110	437,142	333	o490	824,797	1,473	o396	1,277,286	1,191
o190	437,729	573	f310	826,310	933	yi21	1,278,762	411
o175	439,100	528	o326	841,805	981	yi22	1,279,130	906
f242a	444,624	729	o421	842,887	1,266	f442	1,282,638	1,329
f274a	445,720	825	f723	844,243	2,172	o425	1,284,533	1,278
f62	476,725	189	o81	859,903	246	o377	1,285,807	1,134
dppF	482,592	1,005	o80	860,366	243	o100	1,287,696	303
dppD	483,593	984	o351	867,641	1,056	*ins30*	1,288,001	1,152
dppC	484,587	903	o99	893,600	300	*fecE*	1,291,225	768
dppB	485,499	1,020	o142	893,926	429	*fecD*	1,291,993	957
dppA	486,826	1,608	*murI*	946,049	870	*fecC*	1,292,946	999
f276	499,060	831	*murB*	952,690	1,029	*fecB*	1,293,941	909
o96	500,204	291	*birA*	953,715	966	*fecA*	1,294,888	2,325
o173	501,406	522	*coaA*	954,709	951	*fecR*	1,297,299	954
o283	501,924	852	f51	955,846	156	*fecI*	1,298,249	522
yiaA	507,650	441	*htrC*	970,418	540	*insA*	1,299,062	276
yiaB	508,133	354	*yjaC*	1,000,934	2,187	*insB*	1,299,256	294
xylF	511,857	993	o442	1,031,588	1,329	f262a	1,305,548	789
xylG	512,927	1,542	o69	1,039,870	210	f404	1,318,192	1,215
o155	524,470	468	o84	1,040,956	255	f241	1,319,318	726
9157a	525,055	474	o235	1,041,234	708	*fimB*	1,321,490	603
yibA	547,064	843	o174	1,050,234	525	*fimE*	1,322,570	597
yibG	548,904	462	o528	1,056,102	1,587	*fimA*	1,323,648	549
yibD	569,772	1,035	f430	1,062,414	1,293	*fimI*	1,324,153	648
yibB	573,551	873	*phnQ*	1,094,440	408	*fimC*	1,324,837	726
rfaD	574,726	933	*phnP*	1,094,973	759	*fimD*	1,325,630	2,637
rfaF	575,668	1,047	*phnO*	1,095,733	435	*fimF*	1,328,276	531
rfaC	576,718	960	*phnN*	1,096,154	558	*fimG*	1,328,819	504
rfaL	577,687	1,260	*phnM*	1,096,711	1,137	*fimH*	1,329,342	903
rfaK	578,978	1,074	*phnL*	1,097,844	681	f276	1,336,023	831
rfaZ	580,084	852	*phnK*	1,098,635	759	f323	1,343,276	972
rfaY	581,006	699	*phnJ*	1,099,390	846	o521	1,352,899	1,566
rfaJ	581,722	1,017	*phnI*	1,100,228	1,065	o109b	1,354,668	330
rfaI	582,778	1,020	*phnH*	1,101,292	585	*mcrC*	1,357,446	1,047
rfaB	583,797	1,110	*phnG*	1,101,873	453	*mcrB*	1,358,492	1,398
rfaS	584,920	936	*phnF*	1,102,326	726	*hsdS*	1,360,602	1,395
rfaP	585,892	798	f73	1,103,072	222	*hsdM*	1,361,993	1,590
rfaG	586,682	1,125	*phnE*	1,103,290	621	*hsdR*	1,363,783	3,567
rfaQ	587,803	1,035	*phnD*	1,103,965	1,017	*mrr*	1,367,483	915
o394	617,692	1,185	*phnC*	1,105,006	789	o241	1,384,010	726
o307	640,994	924	*adiY*	1,117,697	762	o225a	1,384,693	678
o197	656,237	594	o90	1,132,614	273			