# GENOMIC FLUX: GENOME EVOLUTION BY GENE LOSS AND ACQUISITION

*Jeffrey G. Lawrence and John R. Roth*

# 15

Genome evolution is the process by which the content and organization of a species' genetic information changes over time. This process involves four sorts of changes: (i) point mutations and gene conversion events gradually alter internal information; (ii) rearrangements (e.g., inversions, translocations, plasmid integration, and transpositions) alter chromosome topology with little change in information content; (iii) deletions cause irreversible loss of information; and (iv) insertions of foreign material can add novel information to a genome. Although the first two processes can create new genes, they act very slowly. Gene loss and acquisition are genomic changes that can radically and rapidly increase fitness or alter some aspect of lifestyle.

Most thought on genome evolution has focused on how the slow sequence changes can cause divergence of gene functions. This is understandable because available data suggest that horizontal genetic transfer has been a minor contributor to the evolution of eukaryotic lineages (with notable exceptions, such as the introduction of mitochondria and chloroplasts). In bacteria, however, both genetics and genome analysis provide extensive evidence for gene loss and horizontal genetic transfer. Analyses of these data suggest that gene loss and acquisition are likely to be the primary mechanisms by which bacteria adapt genetically to novel environments and by which bacterial populations diverge and form separate, evolutionarily distinct species. We suggest that bacterial adaptation and speciation are determined predominantly by acquisition of selectively valuable genes (by horizontal transfer) and by loss of weakly contributing genes (by mutation, deletion, and drift from the population) during periods of relaxed selection.

We propose that a limitation of genome expansion couples the rates of gene acquisition and loss. Genome size may be limited in part by population-based factors that limit the ability of cells to selectively maintain information; some limitation may also be imposed by physiological considerations. The balance between selective gene acquisition and secondarily imposed gene loss implies that addition of a foreign gene increases the probability of loss of some resident function of lower selective value. The interaction of these factors, we

*Jeffrey G. Lawrence,* Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260. *John R. Roth,* Department of Biology, University of Utah, Salt Lake City, UT 84112.

*Organization of the Prokaryotic Genome,* Edited by Robert L. Charlebois,
© 1999 American Society for Microbiology, Washington, D.C.

suggest, drives the divergence of bacterial types.

## DYNAMICS OF GENE LOSS

### Existence of a Gene Implies a Function

Traditional bacterial genetics allows identification of gene function by correlating mutant growth phenotypes with biochemical defects. One can demonstrate the functional importance of many DNA sequences by observing the consequences of their disruption. In contrast, genomic analyses identify genes solely as open reading frames, with possible similarity to genes of known function but without a direct tie to either phenotype or biochemical defect.

In identifying a gene by its sequence, rather than by mutations and phenotypes, one assumes implicitly that the very presence of a gene implies that it must confer a selectable function. That is, the gene could only have remained in the population if the encoded function is important; that is, mutants lacking the function show reduced fitness and are removed from the population by natural selection. This evolutionary argument implies that a mutant phenotype (perhaps difficult to demonstrate) will result if the gene is disrupted. In principle, a few genes might be encountered which either have just been introduced or have escaped selection, and the process of elimination has not yet run its course; data supporting such cases will be presented below. An important question arises when one tries to define the function of a gene (or determine whether it is one of the rare nonfunctional examples). That is, how important must a function be to assure the maintenance of its gene? How large a fitness contribution must a gene confer to remain in a genome?

### A Spectrum of Fitness Contributions

All bacterial genes are not equally important. Functions performed by bacterial cells are diverse, and while some are essential for life under all conditions (e.g., RNA polymerase), others may provide a benefit only under certain circumstances (e.g., TMAO reductase). In this way, one may sort bacterial genes into broad classes that reflect their average importance to the cell (Table 1). Mutations in genes making essential or very important contributions to the cell will be strongly counterselected in bacterial populations, since these mutants cannot compete effectively against otherwise uncompromised conspecific individuals. However, mutations in genes that do not contribute to cell fitness (e.g., selfish genes on transposons) will not be counterselected, since these neutral mutants do not put their bearers at a selective disadvantage.

Between these extremes lies the gray area of mutations that have subtle effects on fitness; we include two extreme classes of genes. Some genes may make a minimal contribution to fitness under all growth conditions; others may make a large contribution to fitness but do so only in a rare subset of environments. For either class of gene, the average selection coefficient is low. The fate of mutations in the most weakly selected genes is governed by a complex interaction of natural selection, random genetic drift, population size, population subdivision, and genetic exchange within the species.

For any one species, a different fraction of genes may make up each category in Table 1. In small genomes (e.g., that of *Mycoplasma genitalium*), a large fraction of genes are likely to be essential (42), while a smaller fraction of genes are likely to be essential in prokaryotes with larger genomes (e.g., *Escherichia coli*). Regardless of the distribution, some genes in any genome will fall at the bottom of the list and make a minimal contribution to cellular fitness (i.e., they are nearly neutral). These genes will be at greatest risk for functional loss by mutation, deletion, and genetic drift. This class of genes may comprise a large portion of genomic information.

In *E. coli*, few of the 4,286 protein-coding genes are essential. Isolation of temperature-sensitive lethal mutations suggests that only ~200 genes are essential on rich medium (55).

**TABLE 1** Fitness contributions of bacterial genes

| Class | Fitness contribution | Physiological role | Consequence of null mutation |
|---|---|---|---|
| Top | Large | Essential | Lethality |
| High | Large | Very important | Strong impairment |
| Middle | Moderate | Important | Obvious impairment |
| Low (I) | Very low | Minimally useful in all conditions | Subtle impairment; hard to detect experimentally |
| Low (II) | Very low | Important in rare conditions | Strong impairment in some environments |
| Neutral | None | None | None |

This estimate of the minimal number of essential genes in the *E. coli* chromosome (even when adjusted for failure to detect some genes by this method) is congruent with theoretical estimates of minimal gene number (256 genes) based on comparisons of several bacterial genomes (42). Moreover, a similar number of genes are detected by mutations that cause a nutritional requirement for growth. Together, these rather important gene classes constitute approximately 10% of the *E. coli* genome; the remaining 90% of genes must make smaller contributions to fitness or be needed only under particular conditions. This conclusion is supported by the analysis of *E. coli* mutants described above and by the observation that the genomes of free-living bacteria vary greatly in size. In addition, experimental approaches have revealed that lesions in a large proportion of genes in the yeast *Saccharomyces cerevisiae* have only minimal effects on fitness (65).

## A Minimal Fitness Contribution Is Required for Gene Maintenance

While a complete description of the processes governing the selective maintenance of bacterial genes is beyond the scope of this paper, some general guidelines can be described. If a gene is to avoid loss by mutation and genetic drift, it must provide some minimal average selective benefit to the cell; we call that value $s$. (This is the maintenance threshold in Fig. 1.) If mutation rates are low, genes with small fitness contributions can be maintained in a population. If mutation rates are high, a stronger selective coefficient would be required to maintain a gene in a population of the same size. Therefore, $s$ is proportional to the mutation rate, $\mu$ ($s \propto \mu$). As the mutation rate increases, null mutations in a gene under weak selection are more likely to drift to fixation, since defective alleles are created more rapidly than they can be removed by selection. As population size decreases, loss by drift becomes more likely and more genes become effectively neutral (49, 50). In these smaller populations, a larger selective value is required to maintain a gene in the population. That is, a gene must make a stronger contribution to fitness to be selectively maintained. Therefore, $s$ is inversely proportional to the effective population size, $N_e$ ($s \propto \mu/N_e$). The dynamics of how selection acts on mutant alleles is influenced by the rate of intraspecific recombination. As the recombination rate increases, selection can more effectively remove the steady accumulation of detrimental alleles from a population, and the species can avoid to some extent the fitness decline mandated by Muller's ratchet (41). As a result, the minimum fitness contribution required for selective maintenance in a haploid genome decreases as the recombination rate, $r$, increases:

$$s \propto \mu/rN_e \qquad (1)$$

Here, recombination allows selection to act more efficiently on individual alleles by plac-

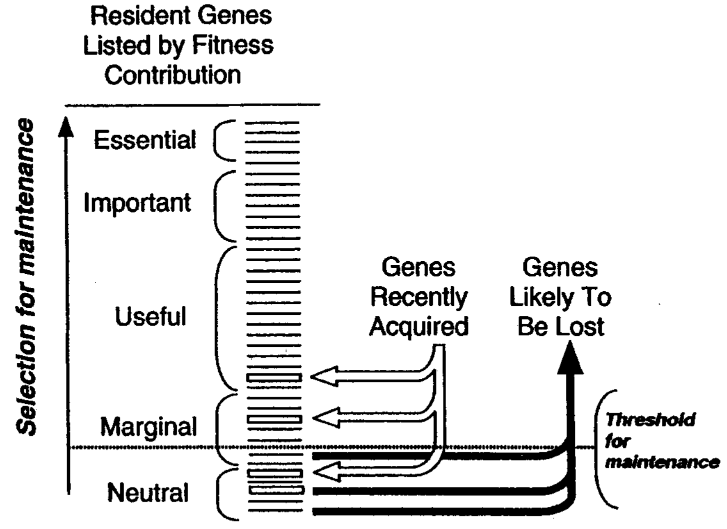
Resident Genes Listed by Fitness Contribution
Selection for maintenance
Essential
Important
Useful
Marginal
Neutral
Genes Recently Acquired
Genes Likely To Be Lost
Threshold for maintenance

**FIGURE 1** Genome evolution by genomic flux. A genome is depicted as a set of genes (lines and boxes) ranked by average selection coefficient; classes of genes discussed in Table 1 are noted on the left. Genes inherited vertically are represented by solid lines; foreign genes are indicated by open boxes. Genes below the threshold for maintenance cannot be maintained by natural selection and will be lost by mutation and deletion. Acquired genes introduced by horizontal transfer will ultimately be lost if they fail to confer sufficient fitness.

ing them in more genetic contexts and minimizing their ability to persist by association with valuable genes.

Thus, a gene must make a larger fitness contribution to be maintained as the mutation rate increases, as the population size decreases, or as the recombination rate decreases. Conversely, with lower mutation rates, bigger populations, or more recombination, genes making extremely subtle contributions to cellular fitness can maintain their positions. This relationship leads to two conclusions: (i) genes failing to make a minimal contribution to cellular fitness will be eliminated from the genome, and (ii) barring any change in mutation rate, recombination rate, or population size, the genome will have an upper size limit at which all resident genes can persist in a population of genomes by counterselection of less-fit mutants. Additional genes cannot be maintained by natural selection, and those genes making the smallest contribution to the organism's fitness will be lost.

Thus, the minimal selective value needed for gene maintenance is a variable dependent on several factors. These dependencies make it difficult to estimate the absolute values for the minimum required fitness contribution. Eukaryotic sexual species typically have high recombination frequencies, small population sizes, and mutation rates similar to or lower than those of bacteria. For such species, the estimate is made that a new mutation with no effect on fitness has a probability of $1/N$ of drifting to fixation, with $N$ being the population size. Therefore, in the absence of selection, every neutral allele in the gene pool has an equal probability of displacing all other alleles and drifting to fixation in the population. To prevent loss by null mutations and drift, a gene must provide a selection value great enough to oppose this drift, placing the required fitness contribution in the range of $1/N$.

For bacteria, which have much larger population sizes, the above estimate would suggest that an extremely small fitness contribution would be sufficient to allow maintenance in the population. In bacteria, mutation rate is likely to be a more prominent dictator of loss than drift; opposing this would require a fitness contribution in the realm of the likelihood of mutational loss ($10^{-5}$ per gene per generation), which is probably larger than $1/N$. Despite these reasonable calculations, the minimal fitness contribution could become much higher if many beneficial genes compete with each other to maintain a position in a genome of limited size. This possibility emerges from the model presented below.

Regardless of the numerical evaluation of the minimal fitness contribution, if genes are sorted in order of decreasing average fitness contribution (Table 1), some must necessarily fall at the end of the list and be prone to loss by mutation and drift. It is these genes that are always at risk for loss, especially when conditions change or newly acquired genes are inserted above them on the list.

### Evidence for Gene Loss

Genome sequence comparisons among enteric bacteria reveal genes that have been lost from certain lineages while being maintained in others. For example, the *phoA* gene, encoding alkaline phosphatase, has been lost from the *Salmonella* lineage but has been maintained in the genomes of virtually all other enteric bacteria (16). Such cases demonstrate that genes conferring a sufficient selectable phenotype in one ecological context may fail to provide such a benefit in another context and therefore be subject to loss by mutation and genetic drift.

## DYNAMICS OF GENE ACQUISITION

### Horizontal Transfer of Genetic Information

DNA may be introduced into bacterial genomes by many processes, including conjugation, bacteriophage-mediated transduction, and transformation. The general term *recombination* typically refers to the introduction of DNA from a conspecific cell, whereas *horizontal genetic transfer* or *lateral genetic transfer* refers to interspecific exchanges; we will use these definitions in the discussion that follows.

The insertion of a foreign gene into a bacterial genome does not guarantee its survival. Rather, the gene must confer a sufficiently large selective advantage to the cell to avoid loss by mutation and genetic drift; above, we have described this minimal fitness contribution ($s$). The fate of an introduced DNA sequence can be predicted by the model developed above for the loss of DNA sequences under weak selection.

One may infer that the vast majority of sequences introduced by horizontal transfer would fail to make a minimal contribution and would be lost. Several factors may explain the failure to make a contribution. (i) The introduced DNA does not encode a product. (ii) The acquired genes are not expressed in the new host. (iii) The acquired genes are expressed and could contribute to a valuable function, but they cannot provide that function without the help of other genes that were not cotransferred (see "Genomic Flux and the Evolution of Gene Clusters" below). (iv) The acquired gene produces a functional protein, but this function does not increase the fitness of the new host cell (either the cell already possessed the function, or the activity is not useful). Therefore, the introduction and incorporation of foreign DNA per se does not ensure its persistence in the recipient gene pool. We must distinguish between horizontal transfer of genetic information (the physical process of incorporating foreign DNA into one of a cell's replicons) and horizontal transfer of useful phenotypic information (foreign DNA sequences that can be maintained for long periods by selection for the encoded functions). Although cursory inspection of genomic sequences may reveal genes that were introduced by horizontal transfer (38, 47, 69), such analyses alone do not reveal whether these sequences are providing a useful function (i.e., have been maintained for long periods by selection). The acquisition of sequence that may or may not increase fitness is diagrammed (with loss) in Fig. 1.

### Horizontal Transfer of Useful Phenotypic Information

Horizontally acquired DNA can confer a selectable new function only if the genes are expressed and all of the new genes required for that function have been cotransferred. If the new function makes a sufficiently large contribution to fitness ($>s$, the threshold for maintenance), then the genes providing for this function may be maintained within the new host genome and escape loss from the population by mutation and genetic drift. In this way, the horizontal transfer of genetic information is successful. That is, long-term maintenance of the acquired DNA occurs only when that DNA provides phenotypic information that is useful to the recipient cell.

Some phenotypic capabilities may provide a selective advantage only to hosts growing in certain environmental regimes. Similarly, hosts growing in a particular ecological niche may find only certain kinds of functions useful. Therefore, although DNA may be transferred nonspecifically among a broad range of organisms, it persists in only a small subset of fortuitous cases, where the transferred DNA imparts a function of value to the host organism in its traditional, or a newly available, ecological context. Only in these rare cases does the horizontal-transfer event contribute to genome evolution (ignoring any mechanistic consequences of recombination). Below we describe methods for assessing the frequency of successful gene acquisition and the fraction of modern genomes that has been acquired by horizontal transfer of useful phenotypic information.

## Evidence for Gene Acquisition

The evidence for gene acquisition in many organisms has been detailed elsewhere (64). In bacteria, horizontally transferred genes frequently confer phenotypes that are characteristic of particular taxonomic groups. Bacterial taxonomists assign a new isolate to a particular species by scoring possession of particular phenotypes that are present in one lineage but not in another. Examples of group-specific traits among enteric bacteria are lactose utilization in *E. coli* and citrate transport in *Salmonella enterica*. *Salmonella*'s ability to synthesize cobalamin and use it to support propanediol degradation forms the basis of a test diagnostic for this bacterium (52). These species-specific abilities are frequently ones whose genes appear to have been acquired horizontally (based on nucleotide composition and codon usage bias) (37). These data suggest that horizontal transfer is an important aspect of the genomic changes that have caused the divergence of phylogenetic groups. Large-scale surveys which assess the numbers and ages of acquired genes in bacterial genomes by DNA sequence analysis are discussed below.

## Limits to Genome Expansion

The gene acquisition process outlined above implies that a genome might continuously accumulate genes that impart a minimal selective coefficient to the cell. However, it is clear that finite populations of cells cannot maintain an infinite number of genes. There is no evidence that bacterial genomes have been continuously increasing over evolutionary time. Rather, genome size is constrained by the population-genetic considerations outlined above. As detailed in equation 1, the minimal fitness contribution required to maintain a gene can be described as a function of the mutation rate, recombination rate, and population size ($s \propto \mu/rN_e$). If the selective benefit of an allele falls short of this level, it may be lost from the population by genetic drift. When applied to many genes in a genome, these forces limit genome expansion, because the minimum selective value required for each gene increases with the total number of genes under selection ($G$, the informative genome size).

The effect of genome size on the minimum selective value is a bit more difficult to explain but can be visualized in the following way. The overall fitness of a genome is a complex function of the combined fitness contributions of the many individual genes. As the number of genes increases, the target for deleterious mutations increases. Each deleterious mutation causes a fitness decrease for the organism and for alleles of other genes carried by that organism. Natural selection can reduce the frequency of deleterious mutations by favoring cells that do not carry deleterious mutations. However, as a genome increases in size, selection becomes less efficient at removing deleterious mutations from any one gene, since there are many more genes with potentially deleterious mutations.

First, consider a genome of sufficient size that, on average, one mutation occurs per genome per generation. Assuming that the number of individuals in the population exceeds the number of genes in a chromosome, every cell will acquire a potentially deleterious mutation every generation. If the recombination rate is sufficiently high, such mutations may be removed and Muller's ratchet may be avoided. As the number of genes increases, this selective cleansing becomes more difficult, until the number of mutations occurring every generation is more than can be counterselected. Eventually, null mutations will completely eliminate some genes from the population, specifically, those that made the smallest contribution to the organism's fitness. Thus, as genome size increases, the component genes effectively compete with one another for maintenance in the genome. The consequence of these effects is that maintenance of a gene in a large genome requires a higher selective value than maintenance of the same gene as part of a smaller genome.

The number of genes that can be simultaneously maintained by selection is limited by these population factors, which therefore limit genome expansion. The relationship between genome size (G) and these population factors is

$$G \propto rN_e/\mu \qquad (2)$$

Genome size, G, can increase only if recombination rates increase (increasing the efficiency with which selection can favor nonmutant alleles of each gene), population sizes increase (to decrease the rate of genetic drift), or mutation rates decrease (to minimize gene damage). The coupled gain and loss process is diagrammed in Fig. 1.

Empirical evidence supports this limitation on genome size. Despite high rates of horizontal genetic transfer among enteric bacterial species, the genomes of *E. coli*, *S. enterica*, and related organisms are notably uniform in size (2, 3, 45).

## Acquisition of Gene Clusters

For a horizontally transferred gene to provide a new, selectively valuable function, all other genes required for that function must be present or cointroduced into the new host genome and be appropriately expressed. Without simultaneous transfer of all required genes, a single acquired gene cannot provide a selective benefit and the gene will not remain in the new population.

In bacteria, genes for nonessential (but useful) metabolic functions are often found in operons or in closely linked clusters of independently transcribed genes. These operons and clusters frequently include all the genes needed to provide a selectable function. We have proposed that evolutionary formation of these clusters (selfish operons) can occur by stepwise aggregation of unlinked genes and that it is driven by selection for increased efficiency of horizontal transfer among prokaryotic genomes (36). Evidence for this model is detailed later.

# DYNAMICS OF GENOMIC FLUX

## Competition between Genes for Maintenance in a Genome

To this point, we have considered gene loss and acquisition as separate processes. Although both loss and acquisition strongly influence genome evolution, we suggest that the two processes are synergistic due to the limits on genome expansion. As detailed above, the minimal selective contribution required for gene maintenance, *s*, is a function of the mutation rate, the recombination rate, and the population size (equation 1). The number of genes, G, that can be maintained by selection is also a function of these parameters (equation 2). If mutation rate, recombination rate, and population size remain constant, it is clear that the minimum selective contribution required for maintenance of each individual gene is a function of the genome size:

$$s_i \propto \mu G/rN_e \qquad (3)$$

That is, as genome size approaches the maximum number of genes maintainable by natural selection, each gene must make a greater contribution to cellular fitness in order to persist. If a genome is small relative to the maximum maintainable size, a gene making only a small contribution to cellular fitness can be maintained, since mutations in this gene can be effectively counterselected. However, as the number of genes under simultaneous selection increases, mutations in this same gene may no longer be effectively counterselected.

Thus, genes in a genome are competing with each other for maintenance in the face of mutation, deletion, and genetic drift. Consequently, an increase in genome size caused by a valuable horizontally acquired gene may increase the likelihood that some less valuable gene (probably of totally unrelated function) will be lost. Since genome size is limited by selection, each gene within a genome competes with the new invading genes, and with other genes, for maintenance in the genome.

The competition among genes for selective maintenance complicates the process of gene acquisition and loss. Because of this competition, the selective hierarchy of genes outlined above (Table 1) changes when a gene for a new function is acquired. Consider genes for a weakly selected function (*wsf*) poised at the threshold of maintenance by natural selection, $s_i$. If the cell acquires genes for a novel function that provides a selective advantage greater than $s_i$, the *wsf* genes will be lost from the genome, because the minimum selective contribution required for maintenance has increased and the value provided by *wsf* is no longer sufficient for maintenance (equation 3). In this way, once genome size has reached its upper limit, the process of gene acquisition can indirectly cause loss of less valuable genes. Genomes that are continually acquiring useful functions probably persist at this upper size limit as genes are continually added and lost.

Acquired genes may also cause selective loss of previously resident sequences if the old functions conflict with the introduced ones. That is, an introduced function may disrupt metabolism or cellular processes and be unable to contribute maximally until some resident functions are eliminated by mutations. Thus, the innovations provided by horizontally acquired genes may drive selective elimination of some previous functions from the host genome. In this way, the deletion of ancestral sequences effectively increases the fitness impact of the acquired sequences.

## Genomic Flux and Ecological Differentiation

Gene acquisition, coupled with gene loss, is a dynamic process by which genetic material flows in and out of bacterial genomes. As noted above, long-term stability of transferred DNA must be associated with a selectable phenotype. Hence, the positively selected gain of valuable genes by a genome will be associated with loss of other genes, resulting in a constantly changing portfolio of phenotypic capabilities. The process of gene acquisition and loss will serve to create descendent populations that are genetically and phenotypically different from ancestral populations. This process is the inevitable outcome of the horizontal transfer of phenotypic information; the valuable acquired sequences will be maintained by selection, and insufficiently valuable native sequences will be displaced (Fig. 1).

Genomic flux can alter the phenotypic character of the recipient organism even when the acquired sequences encode functions that are highly similar to functions encoded by native sequences (e.g., *E. coli* maintains the horizontally acquired *argI* and *gapA* genes although each has a homologue performing a similar function). Unless the new and old versions have some sort of functional difference, natural selection cannot counterselect mutations in both sets of genes. An acquired sequence that provides precisely the same function as a native gene, and makes an identical, redundant contribution to cellular fitness, will have an equal chance of surviving mutational loss. This predicts that we may expect to see occasional examples of old functions being superseded by functionally equivalent genes acquired by horizontal transfer.

One can envision two outcomes of gene acquisition. First, the phenotypic alteration may allow the population to exploit its current ecological niche more effectively. This refines an existing species but does not create an ecologically distinct group of organisms. Alternatively, the acquired information may broaden the ecological niche of the descendent population, allowing it to exploit or survive new aspects of its surroundings. When this happens, the descendent population has novel capabilities that distinguish it from the ancestral population. This process of niche alteration defines bacterial speciation, whereby a group of organisms evolves and forms a distinct evolutionary lineage (67, 70).

The genotypic divergence of speciating lineages is increased by the inevitable loss of genes that accompanies gene acquisition. The losses make it unlikely that gene acquisition will simply broaden the ecological niche; rather, they mandate a qualitative change in niche definition. Lineages that prosper due to acquired traits will differ from the parent strain both by the newly acquired properties and by the gene losses that necessarily occur. As multiple genotypic differences that distinguish newly diverging species accumulate, the infrequent recombination among them will be less likely to cause coalescence of the lineages. To complete the speciation process, neutral point mutations lead to sequence differences between the diverging lineages; this reduces the homologous-recombination rate and imposes reproductive isolation. This drop in recombination is imposed by the sequence specificity of the recombination and mismatch repair (Mut) systems (53, 61, 68, 71).

Genomic flux makes bacterial divergence a natural outcome of the exploration of novel environments. The divergence of nascent species accelerates due to the combination of horizontal transfer and concomitant mutational loss of information. We think speciation would be less likely to occur (or would occur much more slowly) if it depended only on accumulation of internal point mutations. However, it is difficult to compare the two processes. It is unclear how frequently useful information is transferred or how efficiently novel metabolic capabilities can evolve by point-mutational change (internal gene duplication and divergence). It is also unclear how effectively internal gene formation could compete with acquisition of equivalent functions by horizontal transfer. Only by quantifying these two processes (discussed below) can we compare their importance. Below we provide evidence that in enteric bacteria, the magnitude of genomic flux is so large that it probably dominates the process of adding new functions to a genome.

## Limits to Applicability of Genomic Flux in Speciation

The genomic-flux model may be less important for some bacterial lineages. The genomes of enteric bacteria show properties that fit well with predictions of the model. These enteric bacteria have large (5 Mb) genomes that encode many nonessential (but presumably useful) functions; such modestly important functions are subject to loss and acquisition. Differences in these nonessential functions distinguish enteric bacterial species, allowing each lineage to effectively exploit a different ecological niche.

The genomic flux paradigm may apply less well to organisms that do not experience high rates of horizontal transfer or those unlikely to derive any selective benefit from acquired genes. For example, bacterial endosymbionts have little access to horizontal genetic transfer or intraspecific recombination and persist in the relatively constant environment of the host. One might expect that their smaller population sizes and lower recombination rates would serve to reduce the number of genes the organisms could maintain (equation 3). Not surprisingly, phylogenetic analyses show that the genomes of these organisms are becoming smaller (39). Genome size reduction is more dramatic in obligate endosymbionts, although these inferences are confounded by possible gene transfer from the endosymbiont genome to the host genome. In this lineage, acquisition of novel metabolic capabilities is unlikely to allow differentiation of this lineage into related, but unexplored, ecological niches. The minimal fitness contribution required to maintain a gene in such organisms would be expected to be high.

## MEASURING GENOMIC FLUX

### Assessing Gene Loss and Acquisition

To assess the contribution of genomic flux to genome evolution and speciation, one must measure rates of gene loss and acquisition. To do this, one must identify foreign genes and determine when each was acquired (and how many other genes were lost) since the divergence of sibling species from a common ancestor. One must then determine how long each acquired sequence has persisted in the genome. By comparing the complete genome sequences of two closely related organisms, one can easily identify the sequences that are unique to each species; these sequences have either been gained unilaterally by one taxon or lost unilaterally from the other. The next problem is to assess the arrival times of the foreign genes.

Complete genome sequences allow one to count genes unique to each member of a closely related species pair. *M. genitalium* has a 580-kb genome (19), while the closely related species *Mycoplasma pneumoniae* possesses an 816-kb genome (20). Although each gene in the *M. genitalium* genome has a homologue in the *M. pneumoniae* genome, the *M. pneumoniae* genome contains 209 genes not found in *M. genitalium* (21). Similarly, the genome of the sulfate-reducing archeon *Archaeoglobus* sp. (29) is 25% larger than that of the strict methanogen *Methanococcus janaschii* (11). One could speculate that the additional genes found in *Archaeoglobus* allow its heterotrophic lifestyle; such genes would not confer a selective advantage on autotrophs like *M. janaschii.*

These genome comparisons demonstrate that gene loss and acquisition contributed to the divergence of these bacterial species. However, these species pairs are difficult to analyze further because we lack robust phylogenies describing their relationships to each other and to similar bacteria. Without these phylogenies, we cannot easily identify which genes were lost by one species (and would be present in a closely related outgroup taxon) and which sequences were gained (and would be absent from other members of the phylogenetic group). More importantly, genome comparisons do not tell us whether an acquired foreign gene is providing a function that increases the fitness of the organism. This information would be provided if we knew how long each gene had persisted in the genome. A long persistence time would suggest that the sequence is valuable; that is, its mutant alleles are being removed from the population by selection.

An alternative tactic is to identify genes of foreign origin and determine how long ago each entered the genome. This method relies on intrinsic sequence characteristics rather than genome comparisons. From the entry times of foreign genes, one can estimate the age structure of the gene population and the overall rate of gene acquisition. If genomes are not increasing in size, one can conclude that the overall gene acquisition rate is accompanied by an approximately equal rate of gene loss. The lost genes would include some old ancestral genes and some of the recently added genes that failed to confer sufficient selective value.

The assumption of a constant genome size seems reasonable in cases where the members of a robust phylogeny possess similar genome sizes. In these cases one can compare the rate of information influx to the rate at which information is introduced by point-mutational change and thereby assess the relative contributions of these processes to genome evolution. We have explored this general strategy by examining the genomes of *Salmonella* and *E. coli.*

Enteric bacteria have a robust phylogeny (33) and do not vary substantially in genome size (45). Moreover, the various members of this clade inhabit diverse ecological niches—

they exploit soil and water regimes and the digestive tracts of insects, fish, amphibians, reptiles, birds, and mammals and show pathogenic lifestyles. Furthermore, they have obvious phenotypic characteristics that distinguish one species from another. The two enteric species *E. coli* and *S. enterica* are clearly distinguishable, and each is the closest major relative of the other. The sequence of the *E. coli* genome is available (4), and that of the *Salmonella* genome will be available shortly. Considerable sequence information is already available for *Salmonella*. Therefore, we can identify genes unique to *E. coli* and *S. enterica*, identify the foreign-looking genes, and analyze the age structure of each genome, making the assumption of no genome expansion. The genomic-flux model predicts that the differences between these taxa (including metabolic differences or degree of pathogenicity) arose by the process of gene loss and acquisition. The selfish operon model (see below) predicts that multigene phenotypes acquired by horizontal transfer will be found in gene clusters and operons.

## Identification of Foreign Genes Acquired by *E. coli*

Genes that are native to a genome show characteristic and species-specific nucleotide compositions, dinucleotide fingerprints (28), and codon usage biases (56, 57). These patterns emerge when genes experience the same directional mutation pressures for a long time (62, 63). The rate of mutations that convert an AT (or TA) pair to a GC (or CG) base pair is not necessarily the same as the rate of the reverse substitutions, GC (or CG) to AT (or TA). The relative rates of the two substitution types reflect intracellular deoxynucleoside triphosphate pools, the error frequency of DNA polymerases, and the error-specific effectiveness of mismatch repair systems. Organism-specific differences in these functions dictate the relative rates of AT and GC pair interconversion and establish the characteristic differences in nucleotide composition (percent G+C).

Codon usage bias measures the translational preferences among synonymous codons (some codons are translated more quickly, more efficiently, or more faithfully than others). These preferences reflect the relative abundance of different cognate tRNAs (23, 24) and the nature of the tRNA modifications that allow for discrimination among synonymous codons (29a). These factors modulate the effect of directional mutation pressures and are reflected in the relative percent G+C contents of the three codon positions in protein-coding sequences (43).

The effect of directional mutation pressure on position-specific nucleotide composition is shown in Fig. 2A. Muto and Osawa (43) correlated the overall nucleotide compositions of various bacterial genomes with the nucleotide composition of each of the three codon positions in coding sequences. They showed that nucleotides experiencing very weak selection for function (e.g., third codon positions, at which most substitutions fail to alter the encoded amino acid) show a dramatic increase in percent G+C content as genome percent G+C nucleotide composition increases. In contrast, nucleotides under strong natural selection (e.g., the second codon position, whose substitution alters the character of the encoded amino acid) show more modest changes. The behavior of the first position is intermediate, since a few changes there are synonymous. Although the original data of Muto and Osawa show substantial variance from a strict linear relationship between positional and overall percent G+C contents, their study was done at a time when rather little sequence data was available. When large sets of genes or whole genome sequences are added to the analysis, the relationships appear quite robust (linear equations are provided in Lawrence and Ochman [31]).

Directional mutation pressures are evident, in that every long-time resident gene of a genome shows an extremely predictable and characteristic nucleotide composition for the first, second, and third codon positions (Fig. 2A). In contrast, nonnative genes are identi-

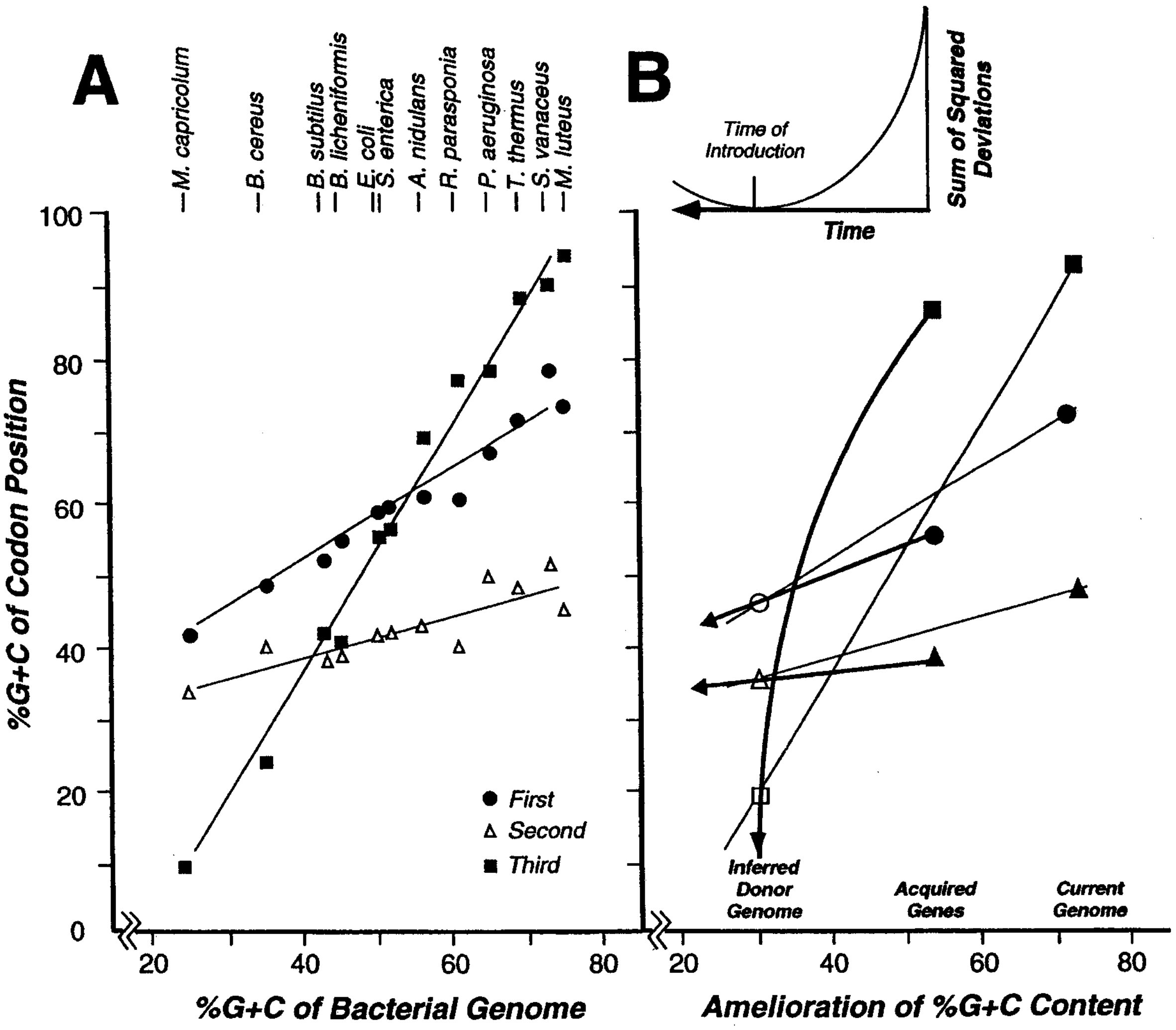

FIGURE 2 (A) Relationships between overall nucleotide composition of a bacterial genome and nucleotide compositions of the three codon positions (first to third); after Muto and Osawa (43) and Lawrence and Ochman (31). The organisms providing the data are shown at the top. The data for *E. coli* and *S. enterica* were calculated from 100 and 25% of the genome sequences, respectively, after known horizontally transferred sequences were removed. (B) Process of amelioration used to infer the time of introduction of acquired genes (31). The acquired genes (shaded symbols) are atypical for the genome (solid symbols) in which they are found. The codon position-specific nucleotide compositions of acquired genes are back-ameliorated (equation 4) until the minimum deviation (by least-squares analysis) from the Muto and Osawa relationships (open symbols) are obtained. The heavy lines indicate the codon position-specific nucleotide compositions during back-amelioration. The arrows indicate the calculated back-amelioration process used to estimate the elapsed time since the sequence showed the pattern of the donor. The inset graph shows the deviation of the curves from the Muto and Osawa relationships as a function of time rather than overall percent G+C.

fiably different, because they still show the sequence patterns imposed by the directional mutation pressures of their donor organisms. For example, a gene from *Bacillus cereus*, with 28% G+C at the third codon position, would be readily detectable as unusual in a background of *E. coli* genes with 58% G+C content at this position. Moreover, the codon usage bias and dinucleotide frequencies of *B. cereus* genes would be strikingly different from the major patterns found in the *E. coli* genome. These patterns have been used by many workers to identify foreign genes in partially sequenced bacterial genomes (28, 31, 38, 47, 69).

Lawrence and Ochman (32) have used these criteria to identify all of the horizontally transferred genes in the complete *E. coli* genome sequence. They used the nucleotide composition of codon positions, patterns of codon usage bias, and dinucleotide frequencies to determine that 15% of the *E. coli* genome (755 of 4,288 genes) was made up of genes identifiably introduced by horizontal transfer. This figure agrees well with previous estimates based on subsets of *E. coli* genes (38, 69). It should be noted that these methods provide a minimum estimate of the numbers of acquired genes, since they would not detect foreign genes donated by an organism with directional mutation pressures similar to those of *E. coli.* The 755 foreign genes (547.8 kb) were introduced into the *E. coli* genome in at least 234 lateral-transfer events since this species diverged from the *Salmonella* lineage 100 million years ago.

The large number of acquired genes supports the magnitude of horizontal transfer but does not reveal how many of the acquired genes contribute a useful function. Some of these genes, like those found on mobile genetic elements and prophages, may have been introduced recently into the *E. coli* genome and may not contribute to the fitness of the organism. To assess the role of gene acquisition in selective divergence, we must determine how many of these foreign genes have remained in the genome for sufficient time to assure us that they are maintained by selection. That is, we must estimate the entry time of genes that have arrived since the divergence of the *E. coli* and *Salmonella* lineages ~100 million years ago (40, 48).

## Estimating the Introduction Time of Acquired Genes

The same characteristics of genes that suggest a foreign origin can help assess their time of introduction. As described above (see also Fig. 2), all genes that have been long-term residents of a bacterial genome experience the same directional mutation pressures and consequently evolve to exhibit relatively uniform patterns of nucleotide composition, codon usage bias, and dinucleotide frequencies. These genes are readily identified as foreign when encountered, following horizontal transfer, in a recipient genome which experiences different directional mutation pressures.

As these foreign genes are selectively maintained in the new genome, they accumulate substitutions that occur under the new set of directional mutation pressures. Over time, their nucleotide compositions will evolve to resemble patterns reflected by native genes (31). During this period of amelioration, however, the nucleotide composition of ameliorating genes will reflect a combination of directional mutation pressures: those imparted by their donor organism and the modifications imposed since entry into the new host genome. A gene caught in the act of amelioration is between the equilibrium states of the donor and recipient genomes and will show a sequence pattern that is unlike that of either donor or recipient (or any other well-ameliorated genome). The unique properties of nonequilibrium genes not only show their foreign origin but also permit an estimation of how long the amelioration process that has brought them from the well-ameliorated donor condition to their present (nonequilibrium) intermediate state has been under way. One can estimate the elapsed time since transfer by assessing the degree to which patterns deviate from the Muto and Osawa relationships (31).

The degree to which the compositional patterns of ameliorating genes depart from the relationships defined by typical genomes allows quantification of the time these genes have experienced the directional mutation pressures of their recipient genome. Lawrence and Ochman (31) described the rates at which the nucleotide composition of each codon position will change over time. At any moment, the amelioration rate for each codon position can be expressed as a function of the overall substitution rate, $R$ (a function of the synonymous and nonsynonymous substitution rates), the current nucleotide composition of the horizontally transferred sequence, $GC^{HT}$, the nucleotide composition of the recipient genome, $GC^{Native}$, and the transition/transversion ratio (IV):

$$\Delta GC^{HT} = [(IV + 1/2)/(IV + 1)] \cdot R \cdot (GC^{Native} - GC^{HT}) \quad (4)$$

Three equations, each derived from equation 4, describe the rate of amelioration for each codon position. If the positional nucleotide composition of an acquired gene deviates from the Muto and Osawa equilibrium relationships, each codon position may have been ameliorating, as described in equation 4. These amelioration equations can be used to determine the length of time that would be required to cause the observed deviation of the three codon positions from Muto and Osawa equilibria. By computational "back-amelioration" of the sequence, one can estimate the introduction time of an acquired gene. This is diagrammed in Fig. 2B. Normal amelioration in this example is proceeding leftward toward the host pattern.

## The Age Structure of the *E. coli* Genome

Each of the 750 foreign genes in the *E. coli* genome was subjected to amelioration analysis (32) to estimate the time at which each entered the genome. Introduction times ranged from 0 to 100 million years ago. As seen in Fig. 3, about a third (225) of the foreign genes do not show any sign of amelioration, since their codon-positional percent G+C contents conform to the Muto and Osawa relationships expected for a sequence of their overall percent G+C content. These genes were likely introduced very recently and still reflect the directional mutation pressures of their donor genomes. Without direct genetic evidence, we cannot assess whether any of these newly acquired genes influences the fitness of *E. coli* and will be maintained in the genome.

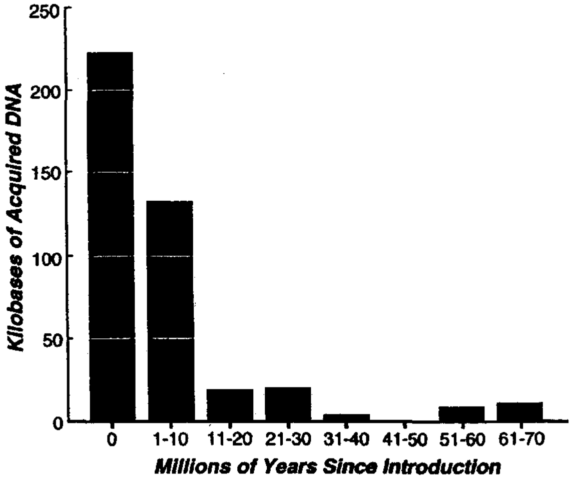

**FIGURE 3** The distribution of times of introduction for horizontally acquired genes in *E. coli*; after Lawrence and Ochman (32). All foreign genes that could be ameliorated successfully are included. Roughly one-third of the foreign genes are not included because their positional percent G+C contents did not converge to fit the Muto and Osawa relationships (see the text).

Another third of the foreign genes have been impossible to date by the amelioration method. They are abnormal with respect to the Muto and Osawa relationships, but the back-amelioration process does not converge on a pattern that fits these relationships, so they are not represented in Fig. 3. Since this group includes transposable elements, prophages, and other selfish elements, we suspect that the odd nature of their sequences may result from their having moved repeatedly from one genome to another with long periods in which their sequences ameliorated without selection. Thus, two-thirds of the foreign genes are either newly acquired or from selfish elements; neither class of genes is likely to impart a selectable phenotype and contribute to species divergence. We expect that these genes will ultimately be deleted from the chromosome; selfish elements will continue to propagate and to persist, but older copies will eventually be removed (34, 44).

Another third of the acquired sequences shows signs of amelioration since entering the *E. coli* genome (Fig. 3). In these genes, mutations that abolish gene function appear to have been counterselected (the reading frames are not interrupted by nonsense codons), while mutations that alter nucleotide composition but do not abolish function have accumulated. These data suggest that these acquired genes have improved the average fitness of *E. coli* and have been maintained by natural selection. The age distribution of these useful foreign genes shows that fairly recent additions are more common than later arrivals, suggesting that many useful genes are held by selection for considerable time but are ultimately lost because of insufficient selective value. We assume that the rate of introduction of DNA by horizontal transfer has remained constant since the divergence of the *E. coli* and *Salmonella* lineages, and the vast majority of horizontally acquired genes have failed to confer a sufficiently useful function to become permanent residents.

After correcting for this inevitable deletion of ameliorated foreign selected genes, we estimate the rate of horizontal transfer of genes conferring selectable phenotypes (horizontal transfer of phenotypic information) into *E. coli* to be ~16 kb/million years (32). The age structure of this group of genes suggests that despite being held by selection for some time, most of these genes are ultimately eliminated (a small proportion of genes has ameliorated beyond our ability to detect them, but this process is much slower than the apparent deletion rate). As noted above, we assume that this rate of acquisition is balanced by a comparable rate of deletion of ancestral genes that could no longer be maintained by natural selection. The overall rate of introduction of DNA is much higher (in excess of 64 kb/million years) and includes sequences which fail to make a contribution to the fitness of the cell and have remained only long enough to be detected in the genome of *E. coli* K-12. We suspect that this vastly underestimates the true flux of total sequence. Our interpretation of the genome structure is diagrammed in Fig. 4.

The identities of the horizontally acquired sequences support the genomic-flux model of speciation. Horizontally acquired genes encode many of the functions by which *E. coli* and *Salmonella* differ. These include *E. coli* genes for lactose utilization (*lac*), phosphonate utilization (*phn*), iron citrate transport (*fec*), and tryptophan degradation (*tna*) and *Salmonella* genes for cobalamin biosynthesis (*cob*), propanediol degradation (*pdu*), citrate utilization (*tct*), and host invasion (*spa*). Other functions present in only one taxon (like alkaline phosphatase [PhoA] in *E. coli* [16]) are unique because the corresponding genes were deleted from the other taxon. We know of no phenotypic property that discriminates between these taxa and is not correlated with a gene loss or acquisition event. No taxonomically useful function has been identified in either taxon that has arisen by internal duplication and divergence of an ancestral gene to allow evolution of a new function by point mutation.

## Genomic Flux versus Point-Mutational Change

The overall rate of horizontal transfer of selectively valuable phenotypic information (16 kb/million years) is lower than the total rate of sequence introduction, since much information does not contribute to fitness. Based on the substitution rates of *E. coli* genes (56), scattered point mutations are estimated to alter the information of the *E. coli* genome at a rate equivalent to 22 kb/million years (31). However, unlike the overall rate of DNA introduction by horizontal transfer, very few of these changes are likely to contribute to cellular fitness; the bulk (~90%) are at synonymous codon positions and cause no change in the nature of the encoded protein. Therefore, while gene acquisition and point mutations both introduce change into bacterial genomes, the qualitative nature of the information they furnish is quite different. Acquired sequences must provide a selectable function in order to be maintained by natural selection and show signs of amelioration; very few point muta-

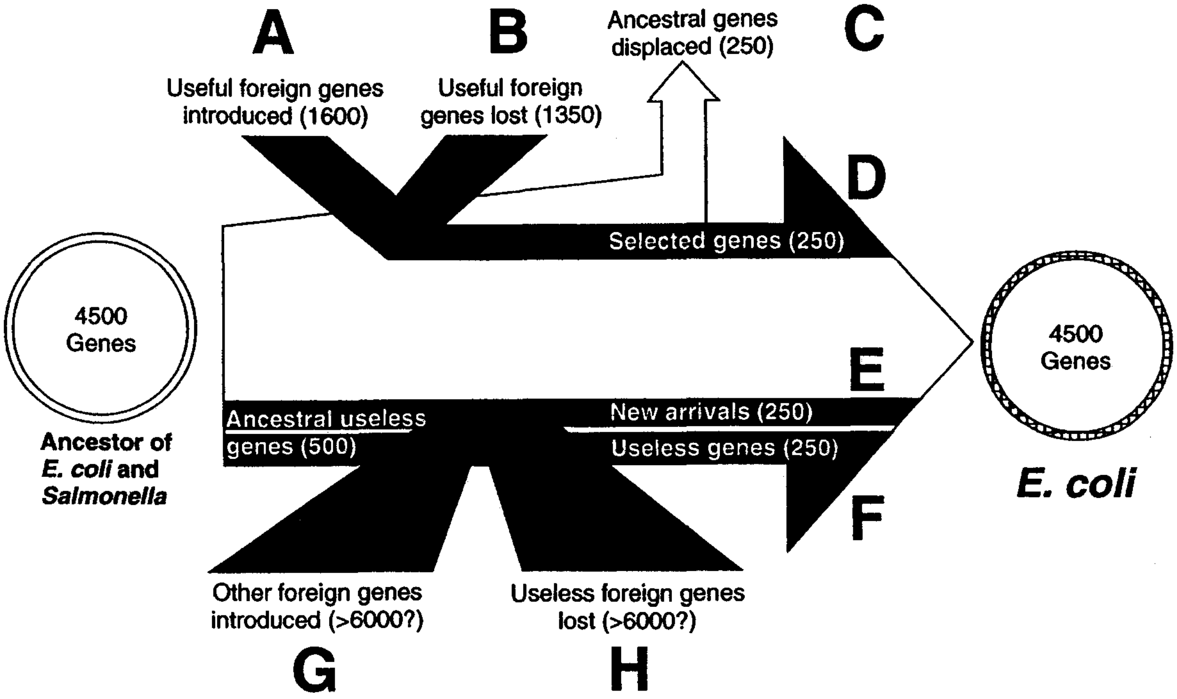
A
B
Ancestral genes displaced (250)
C
Useful foreign genes introduced (1600)
Useful foreign genes lost (1350)
D
Selected genes (250)
4500 Genes
Ancestor of E. coli and Salmonella
E
4500 Genes
E. coli
Ancestral useless genes (500)
New arrivals (250)
Useless genes (250)
F
Other foreign genes introduced (>6000?)
Useless foreign genes lost (>6000?)
G
H

**FIGURE 4** Schematic representation of the effects of genomic flux on the *E. coli* genome. The events depicted occurred since divergence of *E. coli* and *Salmonella* from the common ancestor. The useful gene flux (ABD) depicted above the major arrow describes genes that are kept in the genome long enough to show measurable amelioration. Although large numbers of potentially useful genes enter the chromosome (A) and may be maintained for some time, most of these are ultimately lost (B); however, about 250 remain (D). The number of selectively maintained foreign genes (D) is matched by a loss of ancestral genes (C). The flux below the major arrow (EFGH) represents sequences that provide no selective advantage. Large amounts of DNA with no value are likely introduced into the chromosome (G), but the vast majority are removed by deletion (H). About 500 of these genes are still in the genome but will eventually be removed by mutation and drift; these include those known to be useless (F), like mobile genetic elements, and those merely too recently arrived to have been subject to deletion (E). The newly arrived genes have not been tested by selection, but very few are likely to remain. As shown by the black bars, the ancestral chromosome also contained a portion of selectively useless genes that would have been deleted.

tions are likely to improve cellular fitness. For this reason, we maintain that new genes contributing to the long-term evolution of bacterial species are more likely to have been obtained by horizontal transfer than by internal duplication and divergence (point mutations).

## Salmonella and Its Divergence from *E. coli*

Analysis of the *E. coli* and *Salmonella* genomes by using the principles outlined above allows several conclusions regarding the divergence of *Salmonella* and *E. coli* from their common ancestor. Many of these predictions will be better tested when the complete sequence of the *Salmonella* genome is available. We predict that 15 to 30% of each genome will comprise sequences absent from the other taxon; DNA hybridization data supports this conclusion. Whole-genome DNA-DNA hybridization studies showed that the *E. coli* and *S. enterica* genomes are 45% "related" (7–10). This estimate reflects two processes: (i) more than half of each genome is comprised of shared sequences that are ~85% identical (48, 56), and (ii) the remaining portion of each genome (between 25 and 45%) is unique (8, 10). As detailed above, the sequences unique to each genome include both acquired sequences and genes that have been lost from the other species' genome (Fig. 4).

Lateral genetic transfer contributed many of the features used to identify strains of *Salmonella*. A notable example is $B_{12}$ metabolism, the basis of the Rambach test (52) for *Salmonella* identification, which scores the ability to synthesize cobalamin (*cob*) and perform cobalamin-dependent degradation of propanediol (*pdu*). *Salmonella* acquired these functions in a single horizontal-transfer event that added a block of over 40 contiguous genes, nearly 1% of the *Salmonella* genome. These abilities are shared by virtually every isolate of *S. enterica* (35). Amelioration analysis suggests that the genes were acquired 71 million years ago (31), after the divergence of the *E. coli* and *Salmonella* lineages (~100 million years ago) but before the radiation of the salmonellae (~50 million years ago). The phylogenetic distribution of these functions supports the amelioration method for estimating transfer time.

The gene arrangement within the acquired sequence block (*pdu-cob*) supports the importance of horizontal transfer. This block includes about 40 genes that act together to provide a single selectable phenotype—the ability to degrade a carbon source and synthesize the cofactor needed for this degradation. The block of genes includes four independently transcribed units: the *pdu* operon for degradation of propanediol (15 to 20 genes), the *pduF* gene for importing propanediol, the *cob* operon for synthesis of $B_{12}$ (20 genes), and the *pocR* gene for regulation of all three transcripts (54). There is no known reason that these transcription units need to be close together in order to perform their functions. The fact that all were acquired in a single transfer event suggests why they happen to be close together—their proximity in some donor organism made it possible for *Salmonella* to acquire a complex metabolic capability and a selectable phenotype. Below, we will propose that such gene clusters and operons are formed in a process driven by the horizontal-transfer process discussed above.

A notable feature of the *Salmonella* lineage is its widespread exploitation of pathogenic lifestyles. Strains of *S. enterica* are pathogens of many organisms, including humans, mice, cows, reptiles, amphibians, and poultry (22). Many genes facilitating this pathogenic lifestyle are found in clusters known as pathogenicity islands. Analysis of these regions in *Salmonella* has indicated that many essential virulence factors, including those for attachment (*Salmonella* pathogenicity island [SPI-1]), invasion (SPI-2), and macrophage survival (SPI-3), are found on segments of DNA acquired by horizontal transfer (46, 58). The occurrence of horizontally acquired pathogenicity islands in many pathogens (1) suggests that the genomic-flux model may provide a general framework for describing the evolution of many bacterial lineages. A more complete evaluation of the impact of genomic flux in *Salmonella* evolution can be performed when the genomic sequence becomes available.

Examination of available *Salmonella* sequence data suggests that the rate of foreign-sequence acquisition has been similar to that estimated (see above) for *E. coli*. However, this equality of acquisition rates need not be true and may not be seen for all pairs of sister species. One species may have continued to live in a manner similar to that of the common ancestor, while the sister species diverged to explore some novel environmental situation. The exploratory species may have acquired more foreign genes that gave it the capabilities needed to support its divergence, while the conservative species did not experience significant selection for acquisition of new functions.

If *Salmonella* acquires useful foreign sequence at the rate of 16 kb per million years estimated for *E. coli*, we assume (as we did for *E. coli*) that it will lose ancestral sequence at a similar rate. However, gain and loss of sequences will occur independently in the two lineages, and their genomes will diverge to the extent that they gain and lose different information. By this analysis alone (without comparison to an outgroup taxon), we cannot estimate directly how much sequence was lost unilaterally from the *Salmonella* or *E. coli* lineage or how much ancestral sequence was lost from both genomes independently. However, even without this information, we can use the above considerations to predict what will be seen when the two genomes are compared after 100 million years of divergence.

The general expectations are diagrammed in Fig. 5. Assuming similar events in each lineage and a random loss of ancestral sequences, we predict that the roughly 4,000 genes under selection in both *Salmonella* and *E. coli* will prove to be distributed in the following way. The two genomes will each contain roughly 3,000 shared genes inherited vertically from their common ancestor. Each will contain an added 250 ancestral genes that have been lost from the other species by deletion. Each genome will contain about 750 foreign genes, a different set in each organism, that were acquired by horizontal transfer; of these, about 250 are under selection and contributing to the fitness of the organism. The rest contribute no valuable function: they are either new arrivals that have yet to be eliminated (250) or they are selfish elements that are decaying without selection (250). These expectations are generally borne out by the available sequence data and by DNA hybridization tests done many years ago (7, 10).

## IMPACT OF GENOMIC FLUX

### Mechanisms of Gene Acquisition

The high rate of introduction of DNA into the *E. coli* genome, in excess of 64 kb/million

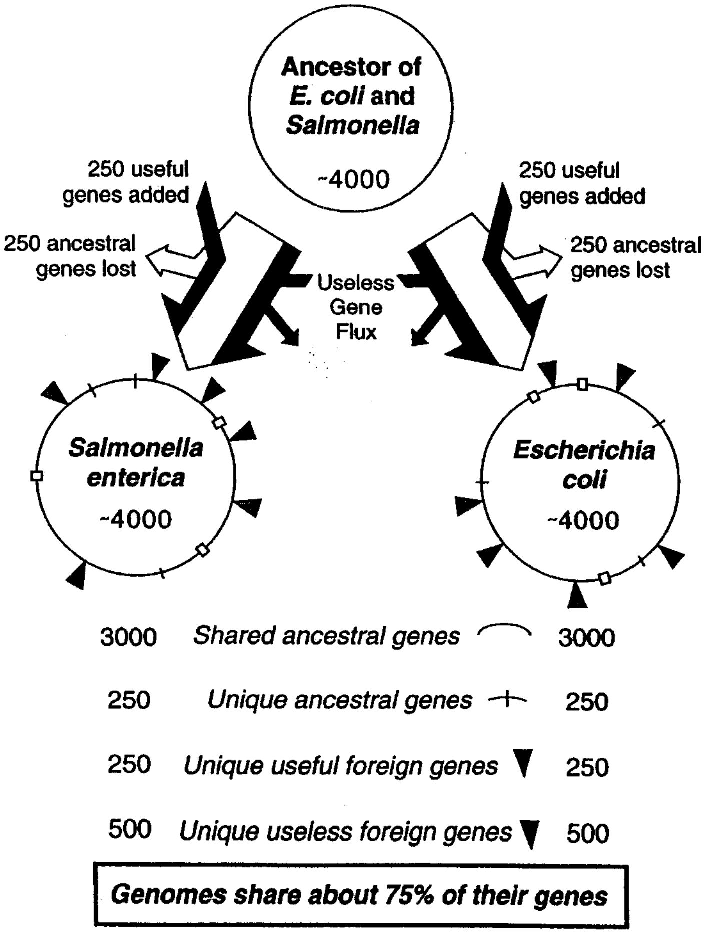
Ancestor of *E. coli* and *Salmonella*
~4000
250 useful genes added
250 ancestral genes lost
Useless Gene Flux
250 useful genes added
250 ancestral genes lost
*Salmonella enterica*
~4000
*Escherichia coli*
~4000
3000 Shared ancestral genes 3000
250 Unique ancestral genes 250
250 Unique useful foreign genes 250
500 Unique useless foreign genes 500
Genomes share about 75% of their genes

**FIGURE 5** Divergence of *Salmonella* and *E. coli* genomes. Based on the model presented in the text, the likely events and final genomic consequences of this act of speciation are portrayed. On genomes, triangles indicate acquired foreign material. Open boxes indicate lost ancestral material.

years, requires plausible mechanisms for the introduction and stable chromosomal incorporation of heterologous genes. It may seem unlikely that sequences could be transferred in natural environments at the rates inferred here. Several facets of this process should be appreciated. First, even an extremely low rate of heterologous recombination will ensure a high rate of gene acquisition by a species if the population size is large and the transferred gene provides a selective advantage. This combination allows extremely rare transfer events to introduce genes that will rise to high frequency in the population. Second, genes may initially be introduced into the cytoplasm on independent replicons (plasmids or bacteriophages) that can provide a long period of replication and selective maintenance prior to integration into the bacterial chromosome. Lastly, transposons are site-specific recombination mechanisms that may mediate insertion of foreign genes; these mechanisms circumvent the need for homologous recombination with its demand for close sequence similarity.

There is evidence for transposon- or bacteriophage-mediated acquisition of foreign genes in the *E. coli* chromosome (32). First, many acquired genes lie adjacent to tRNA genes; bacteriophages are known to use tRNA genes as chromosomal integration sites (14, 25, 51), and many acquired gene blocks adjacent to tRNA genes include homologues of bacteriophage genes. We suspect that other acquired sequences adjacent to tRNA genes may be remnants of prophages from which identifiable bacteriophage genes have been deleted; far too many more foreign genes are found adjacent to tRNA genes than would be expected at random.

A significant fraction (68%) of the insertion sequences found in the *E. coli* chromosome are associated with horizontally acquired genes, often lying at the boundary between native and acquired sequences (32). A foreign segment would initially be flanked by direct-order copies of the insertion sequence (IS) element if an introduced circular DNA fragment were integrated by replicative transposition. (If replicative transposition occurs between a linear fragment and the chromosome, a second exchange would be required—transposition, recombination, or an illegitimate event—to recircularize the recipient chromosome ([59]). The IS element mediating the transposition probably resided on the *E. coli* chromosome and not on the acquired fragment, since IS elements of each class within *E. coli* are nearly identical in DNA sequence.

One might argue that the association of IS elements with acquired genes is not due to a role in integration but rather reflects the dispensability of foreign sequences, which allows them to be used as targets with minimal selective consequences. This alternative would predict that all IS elements would be found more often with foreign sequences. The mechanistic involvement in incorporation seems more likely to us because certain IS elements show a much higher likelihood of association than others (six of seven for IS*2* but zero of three for IS*186*). A notable example of apparent IS-mediated gene acquisition is *E. coli*'s G+C-rich *argI* region at min 7, which is flanked by IS*1* elements.

## Genomic Flux and Speciation

Gene loss and acquisition represent a powerful mechanism by which genomes adjust to major changes in organism lifestyle. Such adjustments are implicit in bacterial speciation. Speciation is the process by which an ancestral population of organisms—defined by their exploitation of a particular ecological niche—evolves to form two populations, each exploiting a separate ecological niche and recombining more often within their own group than with the sister species. As detailed above, genomic flux appears to have played a big role in the speciation event by which *E. coli* and *S. enterica* diverged. The distribution of these organisms in natural environments (18, 66) and their distinctive behavior in clinical situations (22) suggest that they enjoy significantly different lifestyles and inhabit substantially different ecological niches.

We postulate that such diversification into ecologically distinct organisms would not be feasible (in the 100-million-year period) if the genomes of these organisms diverged solely by point mutation or internal rearrangement. We estimate (see above) that about 80% of the genomes of *E. coli* and *Salmonella* are derived from their common ancestor and that each species possesses about 1,000 genes that are absent from the other's genome. The observed differences appear to be the result of unilateral gene acquisition and unilateral gene loss. None of the features now used to distinguish these organisms taxonomically can be attributed to an accumulation of point mutations and internal functional divergence. Rather, gene acquisition (and attendant gene loss) has provided each organism with distinct selectable phenotypes.

## Evolution of Metabolic Novelty

Even if horizontal transfer of phenotypic information is a major factor in bacterial evolution, metabolic novelties must ultimately evolve by stepwise mutational changes that cause functional alteration of preexisting genes. We argue that, when the possibility of horizontal transfer exists, the slow process of functional modification cannot provide for efficient competitive exploitation of an environment. We suggest that the obviously homologous genes with distinct functions seen in a single bacterial genome (apparent paralogues) probably arose initially in distinct lineages as the sole member of the gene family (orthologues). After the orthologues diverged in different genomes, they were brought into a common lineage by horizontal transfer. Thus, we suggest that most of the apparent paralogues seen in bacterial genomes are actually horizontally transferred orthologues. After the microbial world came to include a wide variety of highly adapted genes, assorting these information bits by horizontal transfer became much more likely than reinvention of functions by internal duplication and divergence. In this sense, we suspect there will be few examples of true paralogues in bacterial genomes.

The evolution of metabolic novelty by duplication and divergence is almost always a slow, stepwise, and inherently inefficient process. For example, the evolution of a catabolic pathway for utilization of a hexose may require the following:

1. alteration of binding specificities for several enzymes to accommodate binding of new substrates and intermediates
2. alteration of binding sites of these enzymes to minimize binding of their original substrates
3. alteration of the active sites of these enzymes to allow them to perform their catalytic functions efficiently and effectively on the new substrates and intermediates
4. alteration of release activities of the new products to optimize enzyme turnover
5. alteration of regulatory proteins to discriminate between old and new substrates
6. alteration of regulatory interactions to allow gene expression under potentially different growth conditions
7. coordinate evolution of multiple enzymes to accommodate new substrates and intermediates

We maintain that all of these evolutionary steps are required to provide a function that allows efficient, competitive exploitation of novel environments. Organisms at any stage in the evolutionary process outlined above will not effectively compete with those that have acquired a preformed highly evolved gene complex which performs the same task.

We propose that when horizontal acquisition is possible and appropriate information modules preexist, internal evolution of metabolic novelty (reinvention of the wheel by duplication and divergence) cannot occur in the context of competition. When bacterial speciation entails the competitive invasion of an ecological niche, the evolution of metabolic novelty cannot be correlated with bacterial speciation. The analysis of *E. coli* and *S. enterica* detailed above supports this hypothesis; none of the characteristics that discriminate between these taxa can be attributed to the evolution of metabolic novelty by intragenomic duplication and divergence. Rather, all of these differences, which we posit were intimately involved in the divergence of the lineages and adaptation to their individual niches, can be attributed to gene loss and gene acquisition. All novel phenotypes were conferred by horizontally transferred genes. Therefore, genes for the diverse metabolic pathways have formed slowly at earlier times and not during the course of competitive invasion of novel ecological niches. Regardless of the relative rates of these two processes, it is clear that gene loss and acquisition have facilitated exploration of novel environments and allowed more rapid divergence of bacterial types in competitive situations.

## Genomic Flux and the Evolution of Gene Clusters

The genomic-flux model predicts that horizontal gene transfer mediates the acquisition of novel phenotypic capabilities. If multiple gene products are required to confer a selectable phenotype, the phenotype cannot be transmitted unless all of the required genes are simultaneously mobilized. A subset of the necessary genes will not provide a selectable phenotype and cannot be maintained by natural selection. Therefore, the applicability of the genomic-flux model beyond simple functions requiring the product of only a single gene is contingent upon the physical clustering of genes that provide for a single function.

Bacterial genomes are notable for having clusters of cotranscribed genes, usually contributing to a single selectable function (26). These clusters include genes from a variety of gene families that must have been brought together after their initial formation. As described previously (36) and outlined below, we believe that formation of these clusters was driven by the selective advantage conferred on the clustered alleles themselves (in comparison to the same genes in an unclustered state). The clustered alleles are fitter in a global sense because they can transfer more widely and (like transposable elements) have a selfish advantage. This advantage need not be associated with any physiological improvement of host phenotype. We propose that the prevalence of gene clusters stands as evidence that genomic flux has historically been a primary contributor to genome evolution.

The cotranscription of many clustered genes (operons) is a further refinement of this process and has allowed subsequent coregulation of gene clusters whose products participate in a single metabolic process. Since the elucidation of gene clusters in the 1950s (15) and that of operon structure in the 1960s (26, 27), four theories have been offered to explain the evolution of bacterial gene clusters; these are reviewed in reference 36. The natal theory proposes that gene clusters resulted from the tandem duplication and divergence of parental genes that occurred during initial evolution of metabolic pathways. While this process may explain clusters of homologous genes (e.g., mammalian globin gene clusters), bacterial operons typically contain genes whose products belong to distinct gene families (e.g., kinases, methylases, and dehydrogenases) and were likely assembled from previously existing unlinked genes.

The Fisher theory postulates that coadapted genes—those whose products have been selected to work together efficiently—will appear to be more tightly linked than expected, since recombination may disrupt particularly advantageous combinations of coadapted alleles (17). Consider two genes, each with two alleles, *A* and *a* at one locus and *B* and *b* at another locus. If natural selection favors organisms bearing a coadapted combination of alleles at these loci (*AB* or *ab*) and counterselects organisms bearing more poorly cooperating combinations (*Ab* or *aB*), linkage disequilibrium will occur among the alleles at these loci, simply because the good combinations confer greater fitness. Apparent disequilibrium among alleles would be expected, even for genes on different chromosomes, because the alternative combinations have lower fitness. This model was extended (5, 60) to explain the origins of clusters of genes in haploid organisms, especially in bacteriophage genomes (6, 12, 13). It was postulated that selection favors the assembly of genes into clusters, so that unfavorable recombination between coadapted genes occurs less frequently. This model is restricted to the assembly of operons bearing coadapted alleles (probably encoding products that interact physically) in a freely recombining, variable population. The model is not supported by the fact that organisms with the highest recombination rates (obligatory sexual species) show little evidence of clustering related genes. Conversely, the largely asexual bacterial lineages show a strong tendency to cluster genes with related functions.

The coregulation model (probably the most widely accepted) postulates that genes are found in operons because coordinate expression of the constituent genes is beneficial to the cell. While the benefits of coregulation may contribute to selection for maintenance of a gene cluster, this selection cannot drive formation of the clusters, since coregulation cannot provide a strong selective advantage for the intermediate states that must exist prior to transcriptional fusion. Moreover, coregulation is seen for many bacterial genes (e.g., the *E. coli arg*, *nad*, *pur*, and *pyr* genes) that are not assembled into operons, making it clear that clustering is not essential to coregulation. We have proposed a different model to account for the evolution of operons.

The selfish operon model posits that gene clusters were assembled as a natural consequence of frequent horizontal transfer and that they serve to increase the fitness of the constituent genes by facilitating their distribution to a wide variety of genomes (36). As detailed below, this model provides a way of selecting for progressive clustering of genes and for the maintenance of gene clusters once formed.

## Genomic Flux Facilitates Gene Clustering

The assembly of genes into clusters must entail a series of intermediate steps, during which not all of the members of the gene clusters are physically close. More precise positioning of genes in cotranscribable groups is likely to require additional steps. The most effective means of juxtaposing genes (deletion of intervening genetic material) will not be useful when the intervening material is selectively valuable. For this reason, we discount coregulation as a plausible selective force to drive the evolution of operons.

Horizontal transfer of genes to new genomes provides a selective force that is not subject to the above limitations (36). As detailed above, the introduction of a foreign DNA sequence to a genome does not ensure its persistence. Sequences that do not contribute to fitness are subject to deletion. Therefore, when a large fragment is introgressed, those functions contributing a valuable phenotype will be selectively maintained while intervening noncontributing material will be deleted. In this way, transferred genes that act together to confer a valuable phenotype will be maintained and brought closer together. Although deletion of intervening genes was not possible in the context of the donor genome (where the deletions were deleterious), deletion is inevitable in the context of a new recipient genome. Therefore, any loosely linked collection of cooperating genes, just close enough to be transferred between organisms, will become progressively more tightly linked with time in the new host. As the cluster of related genes becomes tighter, it will transfer more efficiently to new hosts. It should be noted that the base sequence of genes in a cluster may be identical before and after clustering has occurred. Thus, clustering is not expected to provide any phenotypic improvement to the host (compared to the same genes in an unclustered state) but rather provides a benefit to the clustered genes themselves (wider distribution). We termed this the selfish operon model because we consider physical proximity a selfish property of the constituent genes, allowing more frequent transfer to naive genomes. The process for evolution of gene clusters is diagrammed in Fig. 6.

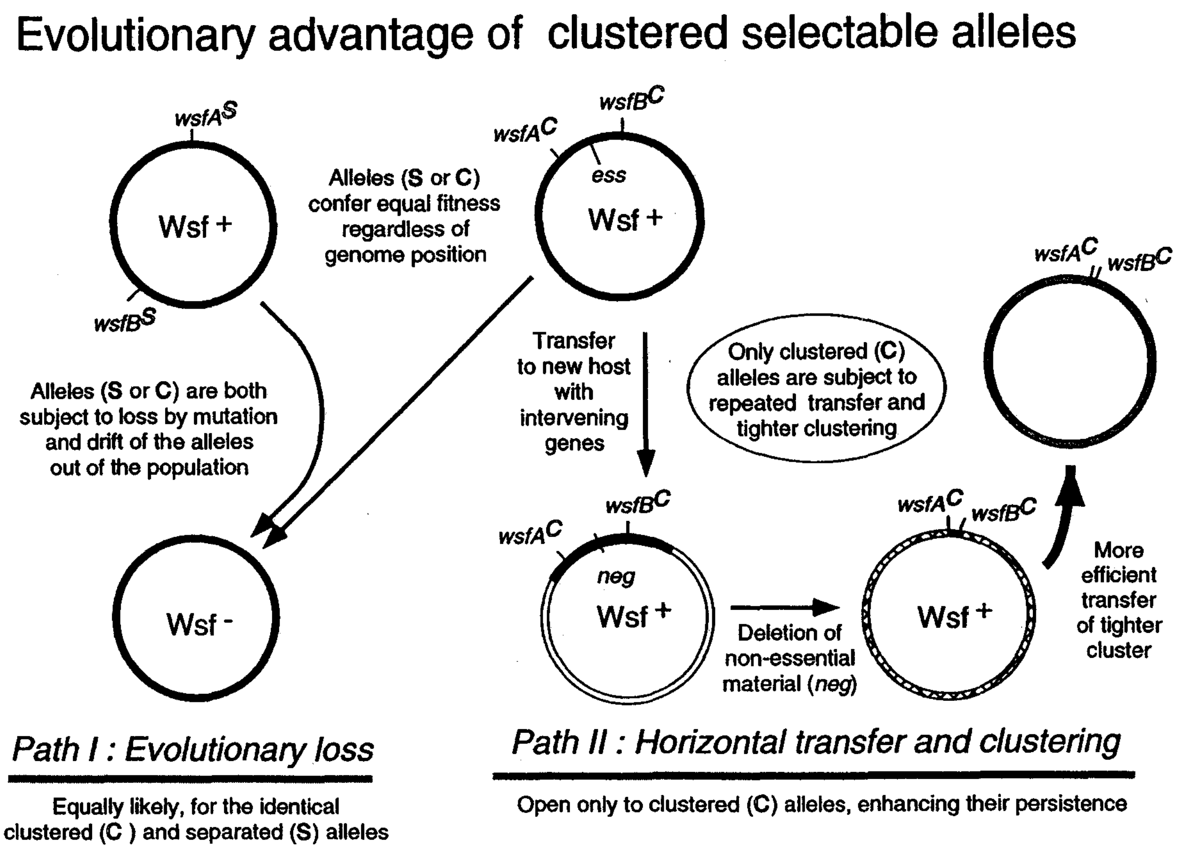
Evolutionary advantage of clustered selectable alleles
wsfA^S
Wsf +
wsfB^S
Alleles (S or C) confer equal fitness regardless of genome position
wsfA^C
wsfB^C
ess
Wsf +
Alleles (S or C) are both subject to loss by mutation and drift of the alleles out of the population
Transfer to new host with intervening genes
Only clustered (C) alleles are subject to repeated transfer and tighter clustering
wsfA^C wsfB^C
Wsf -
wsfB^C
wsfA^C
neg
Wsf +
Deletion of non-essential material (neg)
wsfA^C wsfB^C
Wsf +
More efficient transfer of tighter cluster
Path I : Evolutionary loss
Equally likely, for the identical clustered (C ) and separated (S) alleles
Path II : Horizontal transfer and clustering
Open only to clustered (C) alleles, enhancing their persistence

**FIGURE 6** Mechanism for clustering of genes by horizontal transfer. The diagram compares the fates of identical alleles of two genes, *wsf*, that together confer a weakly selectable function. In one organism these genes are clustered (C), and in the other they are separated (S). Both sets are equally subject to loss by mutation and drift. However, clustered *wsf* genes can spread horizontally to new genomes. Following transfer, genes (*ess*) that were essential in the donor become nonessential (*neg*) in the new species. While the selected *wsf* genes are selectively maintained, the *neg* genes are deleted. This tightens the *wsf* cluster and enhances the likelihood of its further transfer.

### Tighter Gene Clustering and Cotranscription Facilitate Horizontal Transfer

As described above, assembly of related genes into clusters is driven because the clustered state improves the efficiency with which those genes spread between and among species. As gene clusters become tighter, their ability to be mobilized increases, making the model described here operate with ever-increasing efficiency. All known mechanisms of gene transfer increase in efficiency with decreasing size of the fragment that must be transferred. Genes contributing to a single function have a very low probability of cotransfer if they are dispersed on a bacterial chromosome. However, once even loosely aggregated genes have been assembled into a cluster following a single horizontal transfer (see above), these genes have a higher probability of successful cotransfer (and cooperative provision of a phenotype) to new genomes.

Cotranscription of gene clusters (organization of gene clusters into operons) facilitates horizontal transfer by minimizing the number of promoters that must function in the new context. Gene clusters may frequently be transferred into recipient cells with a transcription apparatus different from that of the donor; no phenotype can be conferred unless all promoters function in the new host. This problem is minimized as the number of required promoters is reduced and is eliminated if a block of cotranscribable genes integrates near a host promoter. Furthermore, operons of translationally coupled genes gain the additional advantage of not requiring de novo translation start signals in the new host. Therefore, operons of cotranscribed, translationally coupled genes are highly portable packages of genetic information that function in the widest variety of organisms. For this reason, the cotranscription can be considered an additional selfish property of the constituent genes that extends the benefits accrued by proximity. Once an operon has formed, for purely selfish reasons, a regulatory mechanism may provide additional benefits; in this way, coregulation may contribute to selection for the maintenance of a gene cluster but it cannot provide selection for the initial assembly, or cotranscription, of that cluster.

## Summary

We have outlined a model for the evolution of bacterial genomes through the synergistic processes of gene acquisition and gene loss. From analysis of the *E. coli* genome, we estimate that genes are lost and gained at a rate of 16 kb/million years. The information gained by this process allows exploration of novel ecological niches under competitive conditions; the coupled loss of genes cuts the new organism off from its old lifestyle and contributes to its divergence from the ancestral population. We suggest that this process has driven the organization of bacterial genes into clusters and cotranscribed operons, which allow a single horizontally transferred fragment to confer novel phenotypic capabilities on recipient cells. The organization of genes into operons reflects the important role in bacterial evolution and speciation played by genomic flux—the development of bacterial genomes by gene loss and gene acquisition.

## REFERENCES

1. **Barinaga, M.** 1996. A shared strategy for virulence. *Science* **272:**1261–1263.
2. **Bergthorsson, U., and H. Ochman.** 1995. Heterogeneity of genome sizes among natural isolates of *Escherichia coli. J. Bacteriol.* **177:**5784–5789.
3. **Bergthorsson, U., and H. Ochman.** 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli. Mol. Biol. Evol.* **15:**6–16.
4. **Blattner, F. R., G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277:**1453–1474.
5. **Bodmer, W. F., and P. A. Parsons.** 1962. Linkage and recombination in evolution. *Adv. Genet.* **11:**1–100.
6. **Botstein, D.** 1980. A theory of modular evolution for bacteriophages. *Ann. N.Y. Acad. Sci.* **354:**484–491.
7. **Brenner, D. J., and D. B. Cowie.** 1968. Thermal stability of *Escherichia coli-Salmonella typhimurium* deoxyribonucleic acid duplexes. *J. Bacteriol.* **95:**2258–2262.
8. **Brenner, D. J., and S. Falkow.** 1971. Molecular relationships among members of the enterobacteriaceae. *Adv. Genet.* **16:**81–118.
9. **Brenner, D. J., G. R. Fanning, K. E. Johnson, R. V. Citarella, and S. Falkow.** 1969. Polynucleotide sequence relationships among members of the *Enterobacteriaceae. J. Bacteriol.* **98:** 637–650.
10. **Brenner, D. J., G. R. Fanning, F. J. Skerman, and S. Falkow.** 1972. Polynucleotide sequence divergence among strains of *Escherichia coli* and closely related organisms. *J. Bacteriol.* **109:** 953–965.
11. **Bult, C. J., O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H.-P. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter.** 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273:** 1058–1073.

12. **Campbell, A., and D. Botstein.** 1983. Evolution of lambdoid phages, p. 365–380. *In* R. W. Hendrix, J. W. Roberts, F. W. Stahl, and R. A. Weisberg (ed.), *Lambda II.* Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
13. **Casjens, S., G. Hatfull, and R. Hendrix.** 1992. Evolution of dsDNA tailed-bacteriophage genomes. *Virology* **3**:383–397.
14. **Cheetham, B. F., and M. E. Katz.** 1995. A role for bacteriophages in the evolution and transfer of bacterial virulence determinants. *Mol. Microbiol.* **18**:201–208.
15. **Demerec, M., and P. Hartman.** 1959. Complex loci in microorganisms. *Annu. Rev. Microbiol.* **13**:377–406.
16. **DuBose, R. F., and D. L. Hartl.** 1990. The molecular evolution of alkaline phosphatase: correlating variation among enteric bacteria to experimental manipulations of the protein. *Mol. Biol. Evol.* **7**:547–577.
17. **Fisher, R. A.** 1930. *The Genetical Theory of Natural Selection.* Oxford University Press, Oxford, United Kingdom.
18. **Foltz, V. D.** 1969. *Salmonella* ecology. *J. Am. Oil Chem. Soc.* **46**:222–224.
19. **Fraser, C. M., J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, J. L. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. L. Fuhrmann, D. T. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J.-F. Tomb, B. A. Dougherty, K. F. Bott, P.-C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. I. Hutchison, and J. C. Venter.** 1995. The minimal gene complement of *Mycoplasma genitalium. Science* **270**:397–403.
20. **Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann.** 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.* **24**:4420–4449.
21. **Himmelreich, R., H. Plagens, H. Hilbert, B. Reiner, and R. Herrmann.** 1996. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium. Nucleic Acids Res.* **25**:701–712.
22. **Hoff, G. L., and D. M. Hoff.** 1984. Salmonella and Arizona, p. 69–82. *In* G. L. Hoff, F. L. Frye and E. R. Jacobson (ed.), *Diseases of Amphibians and Reptiles.* Plenum Press, New York, N.Y.
23. **Ikemura, T.** 1980. The frequency of codon usage in *E. coli* genes: correlation with abundance of cognate tRNA, p. 519–523. *In* S. Osawa, H. Ozeki, H. Uchida, and T. Yura (ed.), *Genetics and Evolution of RNA Polymerase, tRNA and Ribosomes.* University of Tokyo Press, Tokyo, Japan.
24. **Ikemura, T.** 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. *J. Mol. Biol.* **158**:573–597.
25. **Inouye, S., M. G. Sunshine, E. W. Six, and M. Inouye.** 1991. Retrophage phi R73: an *E. coli* phage that contains a retroelement and integrates into a tRNA gene. *Science* **252**:969–971.
26. **Jacob, F., and J. Monod.** 1962. On the regulation of gene activity. *Cold Spring Harbor Symp. Quant. Biol.* **26**:193–211.
27. **Jacob, F., D. Perrin, C. Sanchez, and J. Monod.** 1960. L'opéron: groupe de gènes à expression coordonée par un opérateur. *C. R. Acad. Sci.* **250**:1727–1729.
28. **Karlin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.* **11**:283–290.
29. **Klenk, H. P., R. A. Clayton, J. F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, and J. C. Venter.** 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature* **390**:364–370.

29a. **Lawrence, J.** Unpublished results.

30. **Lawrence, J. G.** 1997. Selfish operons and speciation by gene transfer. *Trends Microbiol.* **5**:355–359.
31. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* **44**:383–397.
32. **Lawrence, J. G., and H. Ochman.** 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* **95**:9413–9417.
33. **Lawrence, J. G., H. Ochman, and D. L. Hartl.** 1991. Molecular and evolutionary relationships among enteric bacteria. *J. Gen. Microbiol.* **137**:1911–1921.
34. **Lawrence, J. G., H. Ochman, and D. L. Hartl.** 1992. The evolution of insertion sequences within enteric bacteria. *Genetics* **131**:9–20.
35. **Lawrence, J. G., and J. R. Roth.** 1996. Evolution of coenzyme B12 among enteric bacteria: evidence for loss and reacquisition of a multigene complex. *Genetics* **142**:11–24.
36. **Lawrence, J. G., and J. R. Roth.** 1996. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**:1843–1860.

37. **Lawrence, J. G., and J. R. Roth.** 1997. Roles of horizontal transfer in bacterial evolution. *In* M. Syvanen and C. Kado (ed.), *Horizontal Gene Transfer.* Chapman and Hall, London, United Kingdom.

38. **Medigue, 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.

39. **Moran, N. A.** 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc. Natl. Acad. Sci. USA* **93:**2873–2878.

40. **Moran, N. A., M. A. Munson, P. Baumann, and H. Ishikawa.** 1993. A molecular clock in endosymbiotic bacteria is calibrated using insect hosts. *Proc. R. Soc. Lond. B* **253:**167–171.

41. **Muller, H.** 1932. Some genetic aspects of sex. *Amer. Nat.* **66:**118–138.

42. **Mushegian, A. R., and E. V. Koonin.** 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93:**10268–10273.

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

44. **Naas, T., M. Blot, W. M. Fitch, and W. Arber.** 1994. Insertion sequence-related genetic variation in resting *Escherichia coli. Genetics* **136:** 721–730.

45. **Ochman, H., and U. Bergthorsson.** 1995. Genome evolution in enteric bacteria. *Curr. Opin. Genet. Dev.* **5:**734–738.

46. **Ochman, H., and E. A. Groisman.** 1996. Distribution of pathogenicity islands in *Salmonella* spp. *Infect. Immun.* **64:**5410–5412.

47. **Ochman, H., and J. G. Lawrence.** 1996. Phylogenetics and the amelioration of bacterial genomes, p. 2627–2637. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology*, 2nd ed. American Society for Microbiology, Washington, D.C.

48. **Ochman, H., and A. C. Wilson.** 1988. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26:**74–86.

49. **Ohta, T.** 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **264:**96–98.

50. **Ohta, T.** 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Popul. Biol.* **10:**254–275.

51. **Pierson, L. S. D., and M. L. Kahn.** 1987. Integration of satellite bacteriophage P4 in *Escherichia coli.* DNA sequences of the phage and host regions involved in site-specific recombination. *J. Mol. Biol.* **196:**487–496.

52. **Rambach, A.** 1990. New plate medium for facilitated differentiation of *Salmonella* spp. from *Proteus* spp. and other enteric bacteria. *Appl. Environ. Microbiol.* **56:**301–303.

53. **Rayssiguier, C., D. S. Thaler, and M. Radman.** 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature* **342:**396–401.

54. **Roth, J. R., J. G. Lawrence, and T. A. Bobik.** 1996. Cobalamin (coenzyme B12): synthesis and biological significance. *Annu. Rev. Microbiol.* **50:**137–181.

55. **Schmid, M. B., N. Kapur, D. R. Isaacson, P. Lindroos, and C. Sharpe.** 1989. Genetic analysis of temperature-sensitive lethal mutants of Salmonella typhimurium. *Genetics* **123:**625–633.

56. **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.

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

58. **Shea, J. E., M. Hensel, C. Gleeson, and D. W. Holden.** 1996. Identification of a virulence locus encoding a second type III secretion system in *Salmonella typhimurium. Proc. Natl. Acad. Sci. USA* **93:**2593–2597.

59. **Sonti, R. V., D. H. Keating, and J. R. Roth.** 1992. Lethal transposition of Mu*d* phages in Rec− strains of *Salmonella typhimurium. Genetics* **133:**17–28.

60. **Stahl, F. W., and N. E. Murray.** 1966. The evolution of gene clusters and genetic circularity in microorganisms. *Genetics* **53:**569–576.

61. **Stambuk, S., and M. Radman.** 1998. Mechanism and control of interspecies recombination in *Escherichia coli.* I. Mismatch repair, methylation, recombination and replication functions. *Genetics* **150:**533–542.

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

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

64. **Syvanen, M., and C. Kado.** 1998. *Horizontal Gene Transfer.* Chapman & Hall, Ltd., London, England.

65. **Thatcher, J. W., J. M. Shaw, and W. J. Dickinson.** 1998. Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl. Acad. Sci. USA* **95:**253–257.

66. **Thomason, B. M., J. W. Biddle, and W. B. Cherry.** 1975. Detection of salmonellae in the environment. *Appl. Microbiol.* **30:**764–767.

67. **Van Valen, L.** 1976. Ecological species, multispecies, and oaks. *Taxon* **25:**223–239.

68. **Vulic, M., F. Dionisio, F. Taddei, and M. Radman.** 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in Enterobacteria. *Proc. Natl. Acad. Sci. USA* **94:**9763–9767.

69. **Whittam, T. S., and S. Ake.** 1992. Genetic polymorphisms and recombination in natural populations of *Escherichia coli*, p. 223–246. *In* N. Takahata and A. G. Clark (ed.), *Mechanisms of Molecular Evolution*. Japan Scientific Society Press, Tokyo, Japan.

70. **Wiley, E. O.** 1978. The evolutionary species concept reconsidered. *Syst. Zool.* **27:**17–26.

71. **Zahrt, T. C., and S. Malot.** 1997. Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. *Proc. Natl. Acad. Sci. USA* **94:**9786–9791.