Horizontal gene transfer: A critical view

C. G. Kurland, B. Canback, and Otto G. Berg*
Department of Molecular Evolution, Evolutionary Biology Centre, University of Uppsala, S-75236 Uppsala, Sweden

It has been suggested that horizontal gene transfer (HGT) is the "essence of phylogeny." In contrast, much data suggest that this is an exaggeration resulting in part from a reliance on inadequate methods to identify HGT events. In addition, the assumption that HGT is a ubiquitous influence throughout evolution is questionable. Instead, rampant global HGT is likely to have been relevant only to primitive genomes. In modern organisms we suggest that both the range and frequencies of HGT are constrained most often by selective barriers. As a consequence those HGT events that do occur most often have little influence on genome phylogeny. Although HGT does occur with important evolutionary consequences, classical Darwinian lineages seem to be the dominant mode of evolution for modern organisms.

During the previous decade widespread interest in horizontal gene transfer (HGT) was stimulated by the acquisition of the first complete genome sequences (1–11). At that time a significant number of anomalies were found in the phylogenies of individual proteins, which were not in complete accord with ribosomal RNA (rRNA) phylogeny. The interpretation of these anomalies by Doolittle (7) was summarized in a new paradigm for evolution in which HGT was described as the "the essence of the phylogenetic process." More specifically, three testable assertions defined the new paradigm (7), which we refer to as the rampant HGT paradigm.

First, the eukaryotic nuclear genome was identified as a mosaic of prokaryotic sequences, with contributions from ancestral Archaea and bacteria (1–4). Second, acquisition of novel characters by HGT was thought to dominate adaptive sequence evolution, because the import of ready-made responses to environmental challenge was assumed to be much faster than tinkering with preexisting sequences (5–7). Finally, it was suggested that intensive gene transfer would replace the now classical tree based on rRNA sequences with a jumbled network, more thicket-like than tree-like (7, 8, 10).

There is no doubt that HGT occurs and that it has important evolutionary consequences. However, this is far from saying that HGT is the essence of modern genome phylogeny. We suggest that one reason that HGT has been ascribed such an inflated role is that its frequency has been overestimated by the failure to distinguish it from other phylogenetic anomalies. Although HGT may be frequent at the cellular level, the fixation of transferred sequences in modern global populations seems to be infrequent. Consequently, whole-genome phylogeny is remarkably similar to rRNA phylogeny. Finally, we have found no data suggesting that eukaryotic genomes have originated as fusions between the genomes of bacteria and Archaea. The data strongly suggest that Darwinian lineages are the essence of genome evolution for contemporary organisms and that the impact of HGT on genome phylogeny may be marginal.

Phylogenetic Anomalies

A common argument for the ubiquity of HGT has been that single-protein phylogenies are often at variance with phylogeny based on rRNA (7). However, there is a complex menu of potential anomalies that could generate such discrepancies. Furthermore, there are methods that identify or correct these artifacts of phylogenetic reconstruction (11–18).

Novel sequences may arise in genomes through sequence evolution of adaptive alleles or by the divergence of gene duplications (paralogs) or by the acquisition of alien sequences. Each of these may generate phylogenetic anomalies, because phylogenetic reconstruction normally rests on the assumptions that mutation rates are fairly constant and that inheritance is strictly vertical (11–18). However, gene transfers and segregating paralogs as well as gene loss may obscure strict vertical inheritance. Likewise, the rough constancy of sequence evolution is lost when there is particularly intense selection for mutant alleles or if a genome is evolving under the influence of biased mutation rates. Anomalous phylogenetic reconstructions also may be generated by improper clade selection, which often means examining too few clades, or relying on inadequate phylogenetic methods.

Recently, the thousands of coding sequences found in five taxa of photosynthetic bacteria (19) were reduced to a set of 188 orthologous lineages that could not be resolved into a single tree. From this failure alone it was concluded that HGT was responsible for the phylogenetic ambiguity of the selected data set. The complete absence in this study of any direct indication of HGT is remarkable. No attempts were made to identify the impact of segregating paralogs, gene loss, extreme mutation rates, or biased mutation rates on the reconstructions of this subset of orthologs. Accordingly, the claim from this study that genomic comparisons had revealed the unreliability of rRNA-based reconstruction is less than compelling.

In their review of the HGT data Gogarten *et al.* (10) invite the reader to contemplate a selected cohort of 30 instances of putative gene transfer. This list is used to support the notion that the genomes of bacteria that share a common environment may be mosaics created by the rampant exchange of sequence domains from both rRNA and proteins. We note that even if each of these had been rigorously identified as examples of HGT, the list is far too short to justify the conclusion that HGT is rampant. Thus, these examples were culled from nearly as many genomes, which means that they represent a minute fraction of the many thousands of proteins encoded by those genomes. Their impact on genome phylogeny would be correspondingly minute.

Indeed, protein-coding sequences from completely sequenced genomes can be used to reconstruct genome phylogenies that are remarkably similar to those obtained with rRNA (20–24). In one such study of 50 genomes (24), a simple distance measure is derived from the orthologous matches of all individual proteins in one genome tested against all the other proteins in the genomes of the cohort. These data then are used with a neighbor-joining algorithm to generate genome phylogeny. The congruence of this phylogenetic reconstruction with the rRNA phylogeny from the same genomes is striking. It strongly suggests that Darwinian descent is the dominant mode of genome evolution for these 50 genomes.

Snel *et al.* (24) estimate the frequencies of events such as gene loss, gene duplication, novel sequence genesis, and HGT in the cohort of fully sequenced genomes. For archaeal and bacterial ge-

*To whom correspondence should be addressed. E-mail: otto.berg@ebc.uu.se.

www.pnas.org/cgi/doi/10.1073/pnas.1632870100

nomes <15% of such phylogenetically troublesome events are ascribable to HGT. Unfortunately, an explicit estimate of the influences of biased and variable mutation rates was not presented in this study. Accordingly, even these modest estimates of HGT frequencies are likely to be inflated. Thus, a clear phylogenetic signal of vertical inheritance seems to dominate the "noise" generated by gene transfer as well as by the more frequent anomalies encountered in phylogenetic reconstruction of individual coding sequences (20–24).

BLAST-Based Phylogeny

Another systematic source of inflated HGT estimates has come from the uncritical use of BLAST searches to identify the most similar homologues in pairwise comparisons of archaeal, bacterial, and eukaryotic sequences (2, 25–28). The most closely related pairs in the different domains are identified then as members of orthologous lineages. For example, in this protocol a protein from the genome of a eukaryote that is found to be most similar to a protein from a bacterium is identified as a bacterial gene that has been transferred horizontally.

In this way eukaryotic proteins could be classified as more closely related to archaeal or bacterial homologues (2, 7, 25, 26). It was found that eukaryotic proteins involved in intermediary metabolism often were related more closely to bacterial homologues, whereas those with "informational" functions were identified more closely with archaeal sequences. This gave rise to the notion that the eukaryotic genome arose as some sort of fusion of ancestral archaeal and bacterial genomes (1–4). Similar BLAST studies also identified putative HGT events between bacteria and Archaea, particularly among thermophiles (27, 28). Most recently, the BLAST approach identified hundreds of putative bacterial homologues as interdomain gene transfers in the human genome (29). In contrast, phylogeny by more demanding methods casts doubt on such simplistic interpretations of the best-match searches for alien sequences.

For example, Canback *et al.* (30) have studied a large cohort of enzymes of intermediary metabolism, namely, the glycolytic enzymes from eukaryotes, bacteria, and Archaea. Six members of the Embden–Meyerhof and the Entner–Doudoroff pathways yield reconstructions that are relatively straightforward. The eukaryotic clades and bacterial clades cluster on separate branches that seem to emerge from a common ancestral node very much as observed in the unrooted rRNA tree (31). In no case are the eukaryotic clades rooted in any of the canonical bacterial phyla. In other words, there are no indications in global phylogenic reconstructions of these enzymes of intermediary metabolism suggesting that they were transferred from bacteria to eukaryotes.

Likewise, although there is evidence that HGT has trespassed into both archaeal and bacterial genomes, the two domains are well resolved in the reconstructions of Snel, Huynen, and their colleagues (23, 24, 32). Of course, there may be subtle distortions in the positioning of some of these clades due to exchanges particularly among the thermophiles. Furthermore, specialists such as the archaeal and bacterial thermophiles may contain genomic sequences that are convergent in the sense that they have been selected for the same strong compositional biases. For this reason, a rigorous estimate of the phylogenetic impact of biased mutation rates on the genomes of Archaea and bacteria is required before the influences of these putative instances of HGT on prokaryotic phylogeny can be evaluated in more detail. Here, it is worth emphasizing that, according to the phylogenetic reconstructions of Snel, Huynen, and their colleagues (23, 24, 32), the bacterial and archaeal thermophiles are well resolved.

Finally, when the putative bacterial transfers to the human genome identified by BLAST searches were subjected to more comprehensive analyses (32, 33) it was found that most of these had homologues that are also present in invertebrate ancestors of vertebrates. In other words, most of these human sequences seem to be inherited vertically, although the lineages were obscured initially by gene loss when the cohort of clades examined was too small. Further analysis is required to determine whether the residue of 40–80 putative gene transfers are genuine examples of HGT or whether they are due to other phylogenetic anomalies such as segregating paralogs.

In summary, the best-match BLAST search is a deceptively simple method for phylogenetic reconstruction. Regrettably, it does not distinguish the different phylogenetic anomalies described in the previous section from genuine HGT events. It seems that although the BLAST protocol is a useful beginning to phylogenetic analysis, it is not reliable as a self-contained phylogenetic method precisely because it is too simple.

Transient Sequences

Because base composition, codon usage, and oligonucleotide frequencies vary in characteristic ways from genome to genome, it is possible to identify alien sequences as deviants from such genome-specific characteristics (5, 6, 34, 35). Often these sequence characteristics are statistically distributed rather than uniform within a genome, which may lead to incorrect estimates of alien sequence content (35–38). Nevertheless, there is no doubt that a substantial fraction of genome sequences have been acquired by HGT. Indeed, the contribution of HGT has been estimated to vary between 0% and 17% with a mean of 6% in bacterial genomes (5).

Alien imports are expected to respond to their new genomic environment by taking up its sequence characteristics. That is to say, the deviant alien sequences are progressively "ameliorated" by mutation (35). By studying amelioration rates of the second and third codon positions in the imported sequences Lawrence and Ochman (6) estimate the ages of the imports. They conclude that most are relatively recent, that is to say, imported less than a few million years ago. This suggests that older imports have been purged from the genome (6, 37, 39, 40). This instability implies that most alien sequences do not improve the fitness of their hosts. By the same logic, the very small fraction of alien sequences that are perpetuated in genomes must be adaptive.

Selective Barriers

The ease of transferring alien sequences under laboratory conditions with the aid of plasmids, transposons, and the like is seductive. In fact, the detection and stability of such transfers strictly depends on selective characteristics that may or may not be associated with functions of the transferred gene. In the absence of a selective marker, the transgenic cell most often will simply disappear from the culture in a matter of days because the alien transfer is toxic to its new host (8, 11, 41–45). Even if it is not toxic, a transferred gene that is not performing some adaptive function will be lost by random mutation (39). It seems that an alien sequence is stable only as long as it is selected.

When the cellular machinery is in an early phase of its evolution, sequence evolution is most intense. The probabilities that mutant variants confer selective advantages to their new hosts will be relatively high in this phase (46–49), and there is a relatively good chance that alien sequences will be fixed in their new hosts, at least until a better variant comes along. This conjectured phase of cellular evolution is called the progenote (46–49). A defining characteristic of the progenote population is that global HGT and freely segregating paralogs preclude the delineation of ver-

tical progenote lineages. For this reason the progenote population is thought of as a reticulate network of genomes rather than a collection of tree-like Darwinian lineages (46–49).

In time, the progenote is expected to evolve to a radically slower tempo of sequence evolution simply because selection will have improved the performance characteristics of individual cellular components. The replacement by mutant variants or alien components will become progressively less likely as the workings of cellular components improve. In addition, intensive selection of mutant variants is expected to improve not just the functional efficiencies of individual proteins but also the coherence of cooperative interactions between the members of integrated biochemical systems or cellular networks.

Once the integration of a biochemical network has gone far enough, alien homologues will have little chance to improve the fitness of that network. Indeed, as cellular networks become more coherent and idiosyncratic, alien components are expected to be detrimental most often to the physiological output of the networks and alien transfer becomes increasingly uncommon. Woese (48, 49) has described the evolution from a cellular network that is amenable to HGT to one that is not as a transition through a Darwinian barrier. After this transition, a network is inherited predominantly in vertical lineages.

The biochemical equivalent of a Darwinian network has been referred to as a kinetically optimized system (50, 51). Examples of such systems are the networks responsible for chemotaxis, sporulation, stress responses, pathogenesis, respiration, photosynthesis, and so on. The point here is that the many components of physiological networks are tuned to each other in the sense that the concentrations of network components, their binding interactions, as well their catalytic rates have coevolved to provide an overall optimal physiological output for each organism in its particular environment.

The kinetic optimization of the network involved in the growth response of bacteria is reasonably well characterized (50–52). This network corresponds to the better part of the mass of a modern bacterium under most of its physiological states. Most important, ribosomes play a commanding role in this network, the study of which provides experimental access to the constraints on the import of alien rRNA sequences. The kinetically optimized growth network has evolved such that each of its components has a structure that supports a rate of function that, when normalized by the total mass of the network, is maximized (50, 51). In brief, fast and small are beautiful. If a cell acquires a variant component that lowers the functional rate or raises the mass investment of the growth network, that mutant will lower the fitness of the cell. There are innumerable examples of simple mutations that deoptimize translation and lower growth rates in the absence of a compensating selective condition (51). What about gene transfers?

Because rRNA evolves more slowly than protein-coding genes, it might be anticipated that the consequences of exchanging closely related rRNA sequences would be negligible for the fitness of bacteria (10). Indeed, exacting genetic reconstructions of *Escherichia coli* with rRNA from its very close relative *Salmonella typhimurium* indicate that the transgenic bacteria can grow at rates that are similar to those of the homologous reconstruction, i.e., at doubling times of 52.6 vs. 47.9 min in broth, respectively (53). Not surprisingly, the growth-rate differences between the homologous and the transgenic reconstructions increase as the phylogenetic distance between the donor and recipient increases (53). The question is whether these results support the interpretation that horizontal transfer of rRNA is frequent (10).

In our view, a 9% loss of growth rate is catastrophic, and in nature it would most certainly lead to the rapid extinction of a transgenic bacterium. For example, the take-over probability is $<10^{-40}$ for a single transgenic bacterium with a 9% growth disadvantage that appears in a steady-state population of 1,000 normal *E. coli* cells (39, 54). In larger populations, the take-over probability is even closer to zero. Obviously even a fraction of a 1% difference in growth rate is a definitive handicap from an evolutionary perspective. Indeed, the deceptively small growth-rate defects of bacteria with transgenic ribosomes (53) go a long way to explain the rarity of rRNA transfers observed in nature.

Furthermore, those few rRNA transfers detected thus far are found in genomes containing both the orthologous rRNA of the host as well as the alien rRNA sequences (55, 56). The rRNA operons from many organisms have been sequenced, and thousands of 16S rRNA as well as hundreds of 23S rRNA sequences have been recorded in the databases. Comparison of these figures with the two anomalous ones (55, 56) probably gives a reliable measure of the frequency with which rRNA is partially transferred between different microorganisms. As emphasized previously there are no examples of complete replacements of rRNA from one organism by the rRNA of another (11). In other words, rRNA operons are reasonably stable phylogenetic markers.

Rao and Varshney (57) provide a good example of how alien protein transfers impact the translation system of kinetically optimized cells. They have studied *in vivo* the functions of a ribosome release factor from *Mycobacterium tuberculosis* in the translation cycle of *E. coli*. They find that, by itself, the release factor from *M. tuberculosis* will not rescue *E. coli* with defective factor, but if elongation factor G and release factor from *M. tuberculosis* are introduced together, the defective *E. coli* can be rescued. But even an alien factor that rescues the defective *E. coli* at a given expression level will kill its new host if it is expressed at inappropriate levels (58). Evidently, both dissonant molecular interactions and inappropriate expression levels limit the adaptive value of transferred sequences in transgenic cells.

Patchy Populations

Implicit in our view of the origin of Darwinian lineages is a notion of environmental "normality," of recurrent (not constant) environmental challenge to which cell lineages have adapted. Of course, organisms will also encounter novel situations that require novel adaptations including sequence evolution and sequence acquisition. It is here that HGT is undeniably relevant to the genomes of modern cells. However, we suggest that cell populations exploit gene transfers sparingly both with respect to the range and the frequencies with which they fix such sequences. The reasons for this may be found in the interplay of two factors, the selective barriers in highly integrated cellular networks and the large population size characteristic of most organisms, i.e., characteristic of microorganisms.

The influence of large population size is seen in the fact that strongly selected alien gene transfers such as those encoding antibiotic-resistance phenotypes or pathogenesis islands are often found in patches, i.e., in small subpopulations of a global population (11, 39). Indeed, there is a long tradition of population genetics and ecological genetics that analyze so-called metapopulations with discontinuous distributions of genes (59–62). Here, the metapopulation refers to a population of patches rather than to a population of individuals. More recently, we (39) studied the dynamics of patches generated by limited HGT and subject to destruction by random mutations and selective sweeps. The focus of this analysis is on very

large populations of the sort characteristic of microorganisms, for which HGT seems most relevant.

In a stable situation, mutations in cells are likely to be either destructive or neutral (54). That is to say, once a cell lineage has evolved optimized integrated components, mutant or alien components are not likely to improve the fitness of the lineage. However, a novel environmental challenge, such as a new antibiotic, demands a novel adaptation, such as the expression of antibiotic resistance. But, the antibiotic will be found most often only within a restricted part of the range of the bacterium. Where it is absent, selection for resistance is absent. Now, our notion of the optimized cell implies that in the absence of antibiotic, mutant or novel sequences supporting resistance will be debilitating or, less frequently, neutral, as indeed observed for antibiotic resistance (11, 41–45, 51).

Furthermore, the *a priori* probability of fixation for a neutral sequence is determined by the reciprocal of the population size (39, 54). The actual population size is normally a very large number for microorganisms, perhaps as large as 10^{20} (63). Neutral imports will diffuse through the population, but they will also be the targets of random inactivating mutation and deletion. For very large populations, such as those of microorganisms, the very small *a priori* fixation probability (e.g., 10^{-20}) for a neutral sequence is reduced to virtually nothing when the large mutational target size for inactivation is accounted for (39). Of course, debilitating active genes will be purged even more rapidly.

Accordingly, in our example of adaptation to antibiotic, we expect sequences supporting the resistance phenotype to be residents of only the small patches where the antibiotic is found, stabilized by contingent selection. In contrast, antibiotic-resistance sequences diffusing through the antibiotic-free range of the population, in the absence of selection, will be purged by random mutation (11, 39). Here is the explanation for the transient character of most alien imports in bacteria as well as the fact that they are discontinuously distributed in patches (5, 6, 37, 39, 64). The alien imports that are stabilized in patches by contingent selection will be present in only a small fraction of the genomes of a microorganism. On average, such alien imports have little influence on global genome phylogeny because of their limited range.

Even if a nominally equivalent functional homologue transferred from one organism to another is not directly damaging to its new host, i.e., it is a neutral replacement, the probability that it will be fixed within the new global population is negligible [i.e., one half the reciprocal of the population size (39, 65)]. That is to say, the "nonorthologous homologs" of Gogarten *et al.* (10) are not likely to be fixed in a global population. However, the fate of an imported replacement gene is somewhat different than that of a novel one, because inactivating mutations have a limited impact in this case (39). If import of neutral replacement genes is recurrent throughout the whole range of the organism, the rate of fixation of such a gene becomes independent of population size, just like the rate of neutral mutation fixation in standard theory (54). The question then is how frequent such imports are and how likely they are to be neutral. As discussed above, the selective barriers in an optimized system makes neutrality unlikely. In *E. coli* the overall rate of import has been estimated at 10^{-6} genes per cell per generation (6, 37, 39). If one in a thousand is a possible replacement and one in a hundred of these is neutral, the rate of gene replacement in *E. coli* would be one per 10^9 years.

The residence time of a neutral coding sequence in *E. coli* seems to be <1 million years (6, 37, 39). This turnover of what may be a significant fraction of genomic sequences is reflected in an all-pervasive heterozygosity of genome sequences within a global population. This arises from the transient diffusion of unselected sequences through the population in various states of mutational disarray (39). The implication of this heterozygosity is that we can never point to anything but a consensus sequence when we refer to a global genome population. Identification of an alien sequence in one isolate is no guarantee that the same alien sequence will be present in another isolate from the same global population (6, 37, 39).

This transient patch model was developed in the context of populations that maintain a degree of genomic coherence by selective genetic sweeps (39, 66–70). Here, sequences in one patch may spread to other patches by either gene transfer or the selective growth of invasive genomic variants. In addition, neutral or weakly selected sequences may "hitchhike" on sweeps of selected sequences. However, in the absence of selection these neutral sequences will be purged eventually by random mutation.

Discussion

The suggestion (5, 7, 8, 10) that HGT is the preferred vehicle for novel sequence evolution for modern cells is contradicted by the observations briefly surveyed here. Apparently, populations of organisms prefer to take advantage of a dynamic genetic variability to provide mechanisms for rapid adaptive sequence evolution. Anyone who has selected antibiotic-resistant mutants from a pure culture of bacteria knows that the heterozygosity of bacterial genomes is often more than adequate to support immediate adaptations to environmental challenge. Likewise, hypervariable mutants rather than average members within a population are more likely to be selected for an adaptive response to a novel environmental challenge (71).

HGT certainly occurs, but it seems to be in the minority of anomalous phylogenetic events observed in fully sequenced genomes (20–24). Other discontinuous genome events such as gene loss, gene duplication, and the segregation of paralogs as well as the generation of orphan sequences provide much more frequent challenges to genome phylogeny than does HGT (11).

Once the data based on the best-match BLAST protocol are set aside, there seems to be no phylogenetic data available to support the idea that the eukaryotic genome originated as a fusion of bacterial and archaeal genomes. Rather, there are phylogenetic data such as that for the translation apparatus, the transcription apparatus, and glycolysis, suggesting that all three domains are vertical descendents of a common ancient ancestor (30, 48, 49, 72, 73). Furthermore, ≈1,000 proteins, mostly from eukaryotes, have been identified that have no orthologs in more than one domain (74–76). The existence of such domain-specific signature proteins is consistent with the vertical descent of all three domains from a common ancestor, which we identify with the progenote population.

It seems likely that HGT was a powerful evolutionary force in the era of the progenote (48, 49, 72). In contrast, modern cells seem to be intolerant of even minor mutant variants, not to mention alien transfers (41–45, 50, 51, 53, 57, 58). Of course, founder effects in the divergence of global populations as well as uncommon, strongly selected transfers may lead to the fixation of HGT in global populations of modern cells (74, 77, 78). For example, although mitochondria undoubtedly are descendants of α-proteobacteria (79–81), more α-proteobacterial sequences encoding mitochondrial proteins are found in nuclear genomes than in mitochondrial ones (74, 77). There are other well documented examples of transfer such as those among the aminoacyl-tRNA synthetases (74, 77, 78). Nevertheless, the sampling of a very large number of genomic events has uncovered a modest number of persistent

gene transfers (see, for example, refs. 10 and 20–24). Presumably these have been fixed in global populations under unusual circumstances.

Only two partial transfers of rRNA are known among the thousands of examples of organisms for which rRNA sequences are available (55, 56). This, along with the observation that transgenic bacteria containing rRNA operons from very closely related organisms are debilitated (53), suggests that rRNA is an unusually stable phylogenetic marker. Indeed, there are no indications that transgenic exchanges of rRNA domains are as common in nature as suggested by Gogarten *et al.* (10). All in all, the available data suggest that rRNA-based phylogeny is robust and that Darwinian lineages are the essence of phylogeny.

We are grateful to Irmgard Winkler for help with the manuscript. We thank Siv Andersson, Dan Dykhuizen, Adam Wilkins, and Carl Woese for helpful criticism of versions of this text. We also thank Mark Achtman, Martin Embley, Jeffrey Lawrence, and Richard Moxon for stimulating conversations and help with the literature. Finally, we acknowledge several anonymous referees for their comments. The Swedish Research Council supports O.G.B.

1. Gogarten, J. P., Olendzenski, L., Hilario, E., Simon, C. & Holsinger, K. E. (1996) *Science* **274,** 1750–1751.
2. Feng, D. F., Cho, G. & Doolittle, R. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 13028–13033.
3. Martin, W. & Müller, M. (1998) *Nature* **392,** 37–41.
4. Lopez-Garcia, P. & Moreira, D. (1999) *Trends Biotechnol.* **24,** 88–93.
5. Ochman, H., Lawrence, J. G. & Groisman, E. A. (2000) *Nature* **405,** 299–304.
6. Lawrence, J. G. & Ochman, H. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 9413–9417.
7. Doolittle, W. F. (1999) *Science* **284,** 2124–2129.
8. de la Cruz, F. & Davies, J. (2000) *Trends Microbiol.* **8,** 128–133.
9. Smith, M. W., Feng, D. F. & Doolittle, R. F. (1992) *Trends Biochem. Sci.* **17,** 489–493.
10. Gogarten, J. P., Doolittle, W. F. & Lawrence, J. G. (2002) *Mol. Biol. Evol.* **19,** 2226–2238.
11. Kurland, C. G. (2000) *EMBO Rep.* **1,** 92–95.
12. Galtier, N. & Guoy, M. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 11317–11321.
13. Galtier, N., Tourasse, N. & Gouy, M. (1999) *Science* **238,** 220–221.
14. Philippe, H. & Forterre, P. (1999) *J. Mol. Evol.* **49,** 509–523.
15. Yang, Z. (1995) *Genetics* **139,** 993–1005.
16. Felsenstein, J. (2001) *J. Mol. Evol.* **53,** 447–455.
17. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) *Mol. Biol. Evol.* **11,** 605–612.
18. Lopez, P., Forterre, P. & Philippe, H. (1999) *J. Mol. Evol.* **49,** 496–508.
19. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. (2002) *Science* **298,** 1616–1619.
20. Fitzgibbon, S. T. & House, C. H. (1999) *Nucleic Acids Res.* **27,** 4218–4222.
21. Tekaia, F., Lazcano, A. & Dujon, B. (1999) *Genome Res.* **9,** 550–557.
22. Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001) *Nat. Genet.* **28,** 281–285.
23. Korbel, J. O., Snel, B., Huynen, M. A. & Bork, P. (2002) *Trends Genet.* **18,** 158–162.
24. Snel, B., Bork, P. & Huynen, M. (2002) *Genome Res.* **12,** 17–25.
25. Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G. & Little, E. (1996) *Science* **271,** 470–477.
26. Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6239–6244.
27. Aravind, L., Tatusov, R. L., Wolf, Y. I., Walker, D. R. & Koonin, E. V. (1998) *Trends Genet.* **14,** 442–444.
28. Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., *et al.* (1999) *Nature* **399,** 323–329.
29. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature* **409,** 860–921.
30. Canback, B., Andersson, S. G. E. & Kurland, C. G. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6097–6102.
31. Woese, C. R. (1983) In *Evolution from Molecules to Men*, ed. Bendall, D. S. (Cambridge Univ. Press, Cambridge, U.K.), pp. 209–233.
32. Huynen, M. A., Snel, B. & Bork, P. (1999) *Science* **286,** 1443.
33. Salzburg, S. L., White, O., Peterson, J. & Eisen, J. A. (2001) *Science* **292,** 1903–1906.
34. Medigue, C., Rouxel, T., Vigier, P., Henaut, A. & Danchin, A. (1991) *J. Mol. Biol.* **222,** 851–856.
35. Lawrence, J. G. & Ochman, H. (1997) *J. Mol. Evol.* **44,** 383–397.
36. Koski, L. B. & Golding, G. B. (2001) *Mol. Biol. Evol.* **18,** 404–412.
37. Hooper, S. D. & Berg, O. G. (2002) *J. Mol. Evol.* **54,** 734–744.
38. Guindon, S. & Perriere, G. (2001) *Mol. Biol. Evol.* **18,** 1838–1840.
39. Berg, O. G. & Kurland, C. G. (2002) *Mol. Biol. Evol.* **19,** 2265–2276.
40. Mira, A., Ochman, H. & Moran, N. A. (2001) *Trends Genet.* **17,** 589–596.
41. Andersson, D. I. & Levin, B. R. (1999) *Curr. Opin. Microbiol.* **2,** 487–491.
42. Bergstrom, C. T., Lipsitch, M. & Levin, B. R. (2000) *Genetics* **155,** 1505–1519.
43. Björkman, J., Nagaev, I., Berg, O. G., Hughes, D. & Andersson, D. I. (2000) *Science* **287,** 1479–1482.
44. Borman, A. M., Paulous, S. & Clavel, F. (1996) *J. Gen. Virol.* **77,** 419–426.
45. Cowen, L. E., Kohn, L. M. & Anderson, J. B. (2001) *J. Bacteriol.* **183,** 2971–2978.
46. Woese, C. R. (1965) *Proc. Natl. Acad. Sci. USA* **54,** 1546–1552.
47. Woese, C. R. & Fox, G. E. (1977) *J. Mol. Evol.* **10,** 1–6.
48. Woese, C. R. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 6854–6859.
49. Woese, C. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 8392–8396.
50. Ehrenberg, M. & Kurland, C. G. (1984) *Q. Rev. Biophys.* **17,** 45–82.
51. Kurland, C. G. (1992) *Annu. Rev. Genet.* **26,** 29–50.
52. Dong, H., Nilsson, L. & Kurland, C. G. (1996) *J. Mol. Biol.* **260,** 649–663.
53. Asai, T., Zaporojets, D., Squires, C. & Squires, C. L. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 1971–1976.
54. Kimura, M. (1983) *The Neutral Theory of Evolution* (Cambridge Univ. Press, Cambridge, U.K.).
55. Mylvaganam, S. & Dennis, P. P. (1992) *Genetics* **130,** 399–410.
56. Yap, W. H., Zhang, Z. & Wang, Y. (1999) *J. Bacteriol.* **181,** 5201–5209.
57. Rao, A. R. & Varshney, U. (2001) *EMBO J.* **20,** 2977–2986.
58. Ohnishi, M., Janosi, L., Shuda, M., Matsumoto, H., Terawaki, T. & Kaji, A. (1999) *J. Bacteriol.* **181,** 1281–1291.
59. Wright, S. (1951) *Ann. Eugenics* **15,** 323–354.
60. Maruyama, T. & Kimura, M. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 6710–6714.
61. Whitlock, M. C. & Barton, N. H. (1997) *Genetics* **146,** 427–441.
62. Hanski, I. (1998) *Nature* **396,** 41–49.
63. Ochman, H. & Wilson, A. C. (1987) in *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, ed. Neidhardt, F. C. (Am. Soc. Microbiol., Washington, DC), pp. 1649–1654.
64. Ochman, H. & Jones, I. B. (2000) *EMBO J.* **19,** 6637–6643.
65. Lynch, M., O'Hely, M., Walsh, B. & Force, A. (2001) *Genetics* **159,** 1789–1804.
66. Linz, B., Schenker, M., Zhu, P. & Achtman, M. (2000) *Mol. Microbiol.* **36,** 1049–1058.
67. Falush, D., Kraft, K., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M. & Suerbaum, S. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 15056–15061.
68. Maynard Smith, J. & Haigh, J. (1974) *Genet. Res.* **23,** 23–35.
69. Dykhuizen, D. E. & Green, L. (1991) *J. Bacteriol.* **173,** 7257–7268.
70. Berg, O. G. (1995) *J. Theor. Biol.* **173,** 307–320.
71. Tadei, F., Radman, M., Smith, J. M., Touponce, B., Gouyon, P. H. & Godelle, B. (1997) *Nature* **387,** 700–702.
72. Woese, C. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8742–8747.
73. Olsen, G. J. & Woese, C. R. (1996) *Trends Genet.* **12,** 377–379.
74. Karlberg, O., Canback, B., Kurland, C. G. & Andersson, S. G. E. (2000) *Yeast* **17,** 170–187.
75. Graham, D. E., Overbeek, R., Olsen, G. J. & Woese, C. R. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 3304–3308.
76. Hartman, H. & Fedorov, A. (2001) *Proc. Natl. Acad. Sci. USA* **99,** 1420–1425.
77. Macrotte, E. M. M., Xenarios, I, van der Bliek, A. M. & Eisenberg, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12115–12120.
78. Woese, C. R., Olsen, G. J., Ibba, M. & Soell, D. (2000) *Microbiol. Mol. Biol. Rev.* **64,** 202–236.
79. Andersson, S. G. E., Zomorodipour, A., Andersson, J. O., Sicheritz-Ponten, T., Alsmark, U. C. M., Podowski, R. M., Naslund, A. K., Eriksson, A. S., Winkler, H. H. & Kurland, C. G. (1998) *Nature* **396,** 133–140.
80. Gray, M. W., Burger, G. & Lang, B. F. (1999) *Nature* **283,** 1476–1481.
81. Thorsness, P. E. & Fox, T. D. (1993) *Genetics* **134,** 21–28.