Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution

Tal Dagan* and William Martin

Institut für Botanik III, Heinrich-Heine Universität, Universitätsstrasse 1, 40225 Düsseldorf, Germany

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Nova Scotia, Canada, and approved November 21, 2006 (received for review July 25, 2006)

The amount of lateral gene transfer (LGT) that has occurred in microbial evolution is heavily debated. Efforts to quantify LGT through gene-tree comparisons have delivered estimates that between 2% and 60% of all prokaryotic genes have been affected by LGT, the 30-fold discrepancy reflecting differences among gene samples studied and uncertainties inherent in phylogenetic reconstruction. Here we present a simple method that is independent of gene-tree comparisons to estimate the LGT rate among sequenced prokaryotic genomes. If little or no LGT has occurred during evolution, ancestral genome sizes would become unrealistically large, whereas too much LGT would render them far too small. We determine the amount of LGT that is necessary and sufficient to bring the distribution of inferred ancestral genome sizes into agreement with that observed among modern microbes. Rather than testing for phylogenetic congruence or lack thereof across genes, we assume that all gene trees are compatible; hence, our method delivers very conservative lower-bound estimates of the average LGT rate. The results indicate that among 57,670 gene families distributed across 190 sequenced genomes, at least two-thirds and probably all, have been affected by LGT at some time in their evolutionary past. A component of common ancestry nonetheless remains detectable in gene distribution patterns. We estimate the minimum lower bound for the average LGT rate across all genes as 1.1 LGT events per gene family and gene family lifespan and this minimum rate increases sharply when genes present in only a few genomes are excluded from the analysis.

microbial evolution | phylogenomics | gene clusters

Few topics in evolutionary microbiology are as controversial as lateral gene transfer (LGT). Views on the issue span from one extreme that LGT exists but is insignificant in terms of its overall impact on the evolutionary process, such that a tree of microbial phylogeny can be reliably constructed (1–3), to the other extreme that LGT occurs in nature to such an extent that a simple bifurcating tree is an inadequate metaphor to represent the process of microbial evolution (4, 5). Efforts to resolve this debate have focused on attempts to quantify LGT frequency through evolutionary genome comparisons but are impaired by methodological issues.

There are currently three main approaches to quantifying LGT. The first involves identification of codon usage, GC content, or nucleotide-pattern properties within genomes that differ from the genomic norm and hence are likely to represent acquired sequences (6–8). This approach is powerful but can uncover only recent LGT events. The second approach involves gene-tree comparisons in search of incongruent branching patterns. This approach has delivered widely conflicting results, ranging from estimates that up to 60% of all genes are affected by LGT (9) to estimates that as few as 14% (10) or even only 2% are affected (11). The reason for such divergent quantitative estimates is primarily founded in the uncertainties inherent to phylogenetic reconstruction by using real data (12–14) and in differences among investigated gene and genome samples. A third approach entails inference of gene-gain and -loss events (15–18). Estimates using this approach, in which gene losses are weighted against gene acquisitions (LGTs) according to a predetermined loss-to-LGT ratio, suggest that between 40% (16) and 90% (17) of all gene families might be affected by LGT; these discrepancies are caused by different *a priori* specified gain/loss ratios and the genome samples studied.

An additional approach to inferring LGT but hitherto in a nonquantitative manner involves the identification of genes showing patchy distribution patterns across genomes (19, 20). Although differential gene loss can account for patchy distributions in individual instances, it cannot be invoked to account for all such patterns, because the inferred size of ancestral genomes would become unrealistically large. We reasoned that this phenomenon, which Doolittle *et al.* (21) have termed the "genome of Eden," could be used to estimate the rate of LGT. Given the current distribution of genes across genomes and a reference tree, one can calculate ancestral genome sizes under the assumption that all gene distributions are due to gene loss only. If ancient genome sizes become unrealistically large, incremental allowance of LGT should solve the genome-of-Eden problem, and the amount of LGT that causes inferred ancestral genome sizes to assume a size distribution similar to modern ones would be an estimator of the LGT rate.

However, what if a given gene tree is different from the reference tree? Here, we grant each gene the full benefit of all phylogenetic doubt; we assume (*i*) all gene trees are perfectly compatible with the same reference tree, (*ii*) gene loss is unpenalized, and (*iii*) no paralogy; that is, all within-genome duplications for each gene family are assumed to have occurred subsequent to the last divergence for each lineage. Taken together, these three assumptions mean we infer no LGTs from phylogenetic conflicts; hence, our approach delivers conservative lower-bound constraints for the minimum LGT rate during prokaryote genome evolution.

Results

The Distribution of Genes Across Genomes. Using the standard method (22), we clustered all 562,321 protein-coding genes present within 190 prokaryotic genomes [supporting information (SI) Table 4] into groups that we designate here as gene families. In the case of conspecific strains, the strain with the largest genome was used. Using a cutoff of 30% amino acid identity in the clustering procedure, these sequences fall into 57,670 families with two or more members in addition to 149,894 singletons excluded from pattern analysis. Each family constitutes a binary-coded presence/absence pattern (PAP) of shared genes across all genomes (Fig. 1*a*). In cases where a genome contains more than one member of a family, multiple copies condense to a single presence according to our full-benefit assumptions.

For the resulting 57,670 families, 20,680 different PAPs are observed, of which 17,583 (30%) are unique, and 3,097 are recurring. The most-frequent PAPs are typically shared by congeneric

Author contributions: T.D. and W.M. designed research; T.D. performed research; T.D. and W.M. analyzed data; and T.D. and W.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: LGT, lateral gene transfer; PAP, presence/absence pattern; SO, single origin.

*To whom correspondence should be addressed. E-mail: tal.dagan@uni-duesseldorf.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0606318104/DC1.

www.pnas.org/cgi/doi/10.1073/pnas.0606318104

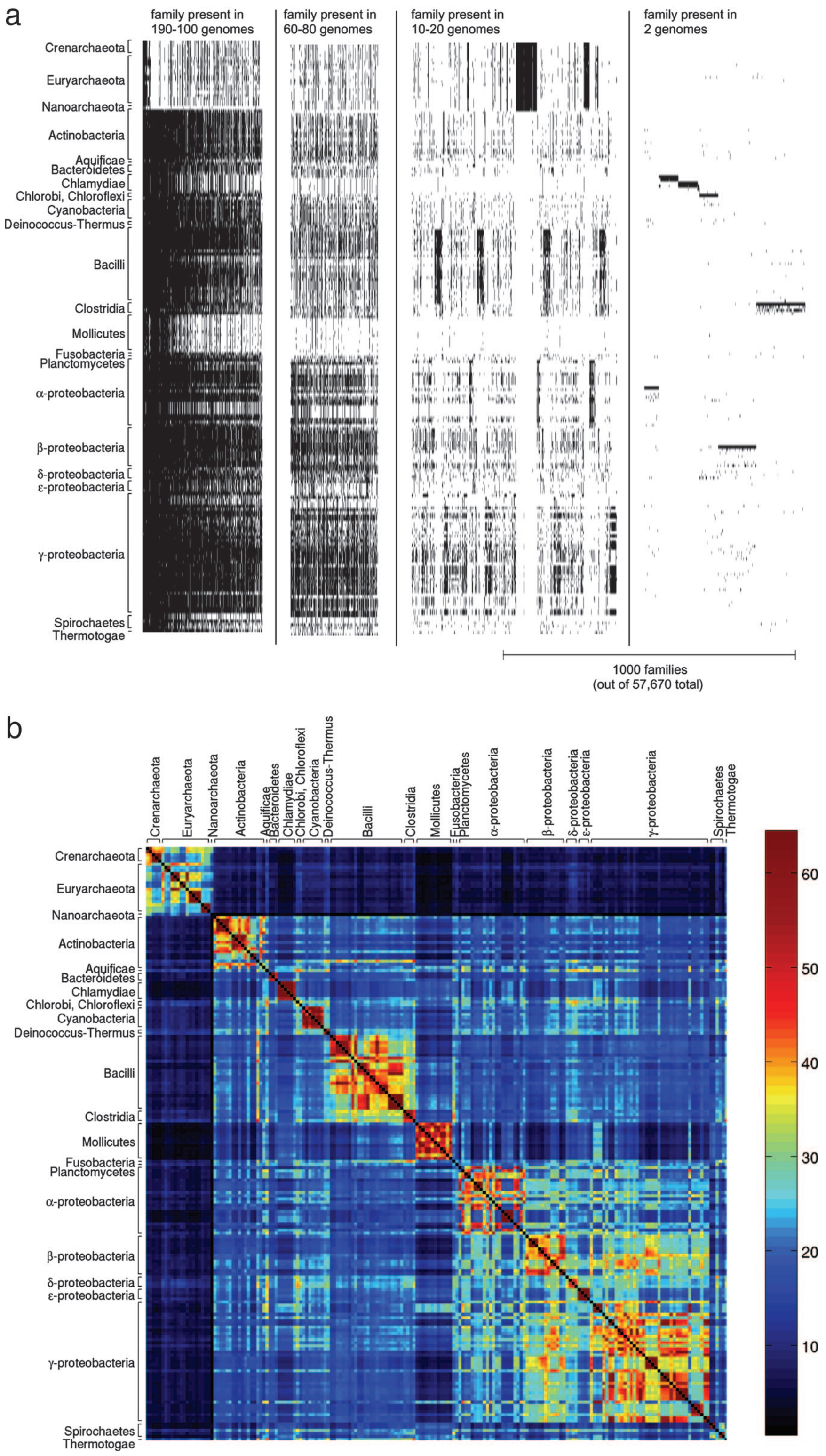

Fig. 1. The distribution of genes across genomes. (*a*) Presence (black) and absence (white) patterns for representative segments of the data comprising widely (present in 100–190 genomes), intermediately (60–80 and 10–20 genomes), and sparsely distributed genes (two genomes). (Note the scale bar.) (*b*) Color-coded matrix of the proportion of shared genes for all genome pairs, with genomes grouped by taxonomical classification. For the same matrix using random genome order, see SI Fig. 4. The proportion of shared genes for a genome pair *x,y* is calculated as the number of genes in genomes *x,y* that are found in shared clusters, divided by the total genes in genomes *x,y*. The color scale indicates the shared proportions of genes in percent. For example, archaebacterial genome pairs share 32 ± 16% (mean ± SD) families on average, whereas archaebacterial vs. eubacterial genome pairs share only 7 ± 3% of their families. For cyanobacteria, 61 ± 10% of each genome consists of families shared with another cyanobacterium, as opposed to 18 ± 5% in comparisons to noncyanobacteria. For proteobacteria, γ-proteobacteria share 38 ± 13% common families with other γ-proteobacteria, 26 ± 7% with other proteobacteria, and 18 ± 8% with nonproteobacteria.

genomes. Consistent with earlier findings (23), the proportion of shared families among genomes from different prokaryotic groups uncovers components of both vertical and horizontal inheritance (Fig. 1*b*).

Ancestral Genome Sizes Constrain the Average LGT Rate. To estimate the minimum amount of LGT in the present gene-distribution data, we first plotted the PAPs onto a reference tree for the rRNA operon (SI Fig. 5). We designate an evolutionary scenario that utilizes vertical inheritance and gene loss only as the loss-only model; gene distribution is governed solely by loss, each ancestral genome contains all families present in its descendants, and genomes hence become progressively larger back through time (Fig. 2*a*). With the present data and our reference tree, the prokaryotic ancestor would have had a genome encoding 57,670 families, exceeding the average genome size in our sample (2,198 families)

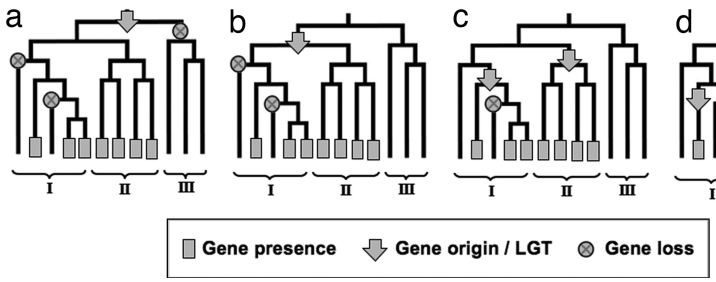

Fig. 2. Gene loss and LGT can both account for patchy gene distributions. Schematic representation of four different LGT allowances. (*a*) In the loss-only model, all genes are assumed to have originated at the root of the tree; PAPs are attributed to gene loss only. (*b*) Introducing a gene origin in the SO model disperses gene origins over internal nodes of the tree according to their first occurrence. (*c*) In the $LGT_{\leq 1}$ model, each gene is allowed to have two origins, where one is an LGT. This model results in further dispersal of gene origins across the tree, hence smaller ancestral genomes. (*d*) Two additional LGTs are allowed in the $LGT_{\leq 3}$ model. Allowances of up to 7, 15, and 31 LGTs were also tested.

by 26-fold and the largest genome in our sample (8,317 families; *Bradyrhizobium japonicum*) by 7-fold (Table 1).

Such burgeoning genome sizes are indeed unrealistic, but so is the notion that new genes do not arise during evolution. Allowing new genes to arise over time according to the single-origin (SO) model (Fig. 2*b*) yields an ancestral prokaryote genome that contains 3,081 genes, those present in both archaebacteria and eubacteria (Fig. 3*a*). However, the SO model does not solve the genome-of-Eden problem; it merely transfers it into the middle ages of microbial evolution, where ancestral genome sizes soar once more to 12,000–14,000 genes, sizes that far exceed those observed among modern organisms (Table 1).

Thus, we either have to embrace the untenable assumption that microbial genome sizes were fundamentally different in the past than they are today or, preferably, we have to allow some amount of LGT. How much LGT is necessary to bring ancestral genome sizes into agreement with the observed contemporary range?

We started by allowing only one LGT event per family (Fig. 2*c*), the $LGT_{\leq 1}$ model. This model allows each gene to have two origins, one of which is an LGT. For 35% of our families, neither LGT nor loss is required; the remaining 65% accept one LGT. This average LGT rate of 0.65 LGT per family (Table 2) brings inferred ancestral genome sizes down to <8,000 genes (Fig. 3*b*), with a maximum of 7,607 and a mean of 2,858, closer to contemporary genomes (Table 1) but still a bit too large.

We tested additional evolutionary models allowing up to 3, 7, 15, or 31 LGTs per gene family. With increasing amounts of LGT allowed, inferred ancestral genome sizes shrink, as do the numbers of inferred gene losses per gene family (Fig. 2*d* and Table 2). Although most gene families do not require more than one LGT to map exactly onto the reference tree (Table 2), they are also quite small and offer little opportunity to observe LGT (SI Fig. 6). However, even in the extreme cases of the $LGT_{\leq 15}$ and $LGT_{\leq 31}$ models, no families accommodate the maximum number of LGTs allowed (Table 2). Because only a very small proportion of gene families require many LGTs to account for their phyletic distributions, allowing more LGTs hardly changes the average LGT frequency per gene family (Table 2).

Comparison of the distributions of 190 modern genome sizes with 187 inferred ancestral genome sizes for differing LGT allowances using the Wilcoxon test (24) revealed that all models except $LGT_{\leq 3}$ are rejected at $\alpha = 0.02$ (Table 1). With no LGT, ancestral genome sizes are too large; however, with too much LGT, they become too small. Even for the $LGT_{\leq 3}$ model, only 8% of all families accept all three LGTs allowed, such that the average rate across all genes in the $LGT_{\leq 3}$ model is ≈1.1 LGT per gene family. This amount of LGT is sufficient to bring the distribution of ancestral genome sizes into congruence with that of modern genomes.

Too Much LGT Makes Ancestral Genomes Too Small. Allowance of many LGTs causes inferred ancestral genome sizes to become far too small in comparison to modern genomes (Fig. 3 *e* and *d*). The mean ancestral genome size in the models allowing seven or more LGTs is less than half the mean of modern genomes, and the size distributions of modern and ancestral genomes are different at $\alpha = 0.05$ using the Wilcoxon test (Tables 1 and 2). Furthermore, for genomes with ≤1,000 families, ancestral sizes are biased toward miniscule sizes with too much LGT allowance (Table 3). Thus, although the genome of Eden demands LGT to keep ancient genome sizes realistically small, too much LGT makes them unrealistically small.

Another problem with the models allowing many LGTs concerns the number of losses inferred per gene. In the $LGT_{\leq 31}$ model, 92% of all gene families are inferred to evolve without a single loss (Table 2), which is unrealistic, because loss events are abundant in bacterial evolution (25, 26). Hence, introducing too many LGTs turns gene loss, an important and common mechanism affecting genome size, into a rare event. The mean origin-to-loss ratio observed in the $LGT_{\leq 3}$ model is 1:1 (Table 2), twice the threshold value used for LGT inference in previous studies (16, 17) that constrained the LGT rate as opposed to estimating it.

Table 1. Modern and last-common ancestor (L-ca) genome sizes under different LGT allowances

		Inferred ancestral genome size (number of families)						
Genome	Modern genomes*	Loss only	Single origin	$LGT_{\leq 1}$	$LGT_{\leq 3}$	$LGT_{\leq 7}$	$LGT_{\leq 15}$	$LGT_{\leq 31}$
$L_{prokaryotic}$ca	2,198	57,670	3,081	2	2	2	2	2
$L_{archeabacterial}$ca	1,573	9,453	3,240	148	91	91	91	91
$L_{eubacterial}$ca	2,297	53,658	3,573	443	35	35	35	35
$L_{proteobacterial}$ca	2,690	35,903	13,652	5,872	3,303	2,119	1,517	1,147
$L_{cyanobacterial}$ca	2,187	5,526	10,509	3,598	1,938	1,306	1,014	886
$L_{actinobacterial}$ca	2,602	11,611	10,233	3,461	1,691	1,044	703	520
$L_{mollicute}$ca	432	1,714	660	557	485	415	352	300
Mean†	2,198	8,142	7,296	2,858	2,234	1,868	1,634	1,472
Ancestral vs. modern‡		<0.01	<0.01	<0.01	0.71	<0.02	<0.01	<0.01

*Average genome size for the group.

†For 190 modern genomes, for 187 ancestral genomes.

‡Probability that the two samples come from distributions of equal medians using the Wilcoxon Mann–Whitney test.

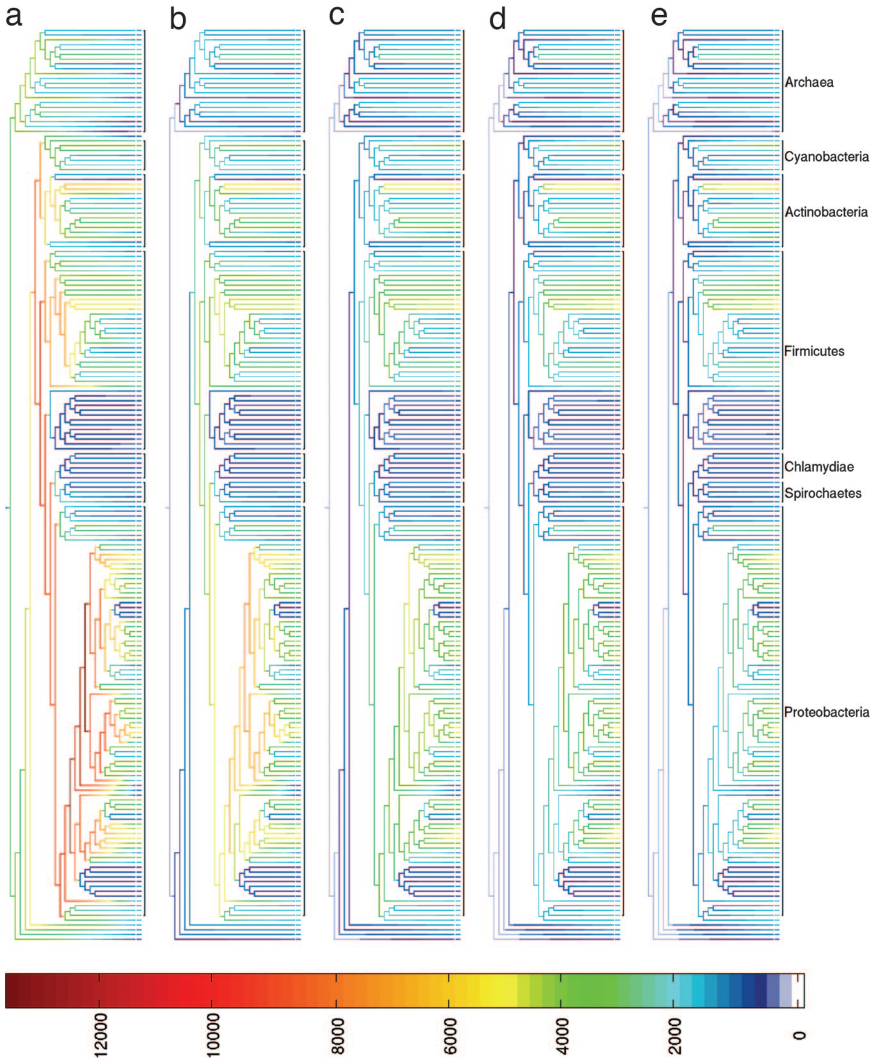

Fig. 3. Ancestral genome sizes reconstructed under the various reconstruction models. The colors of nodes and branches correspond to the inferred ancestral genome size, as indicated in the scale. *a–e* correspond to the SO, $\text{LGT}_{\leq 1}$, $\text{LGT}_{\leq 3}$, $\text{LGT}_{\leq 7}$, and $\text{LGT}_{\leq 15}$ models, respectively (see SI Figs. 7 and 8 for the same analysis using a reference tree reconstructed by neighbor joining and a random reference tree, respectively). To calculate the genome size in each hypothetical taxonomic unit, a binary recursive algorithm scans the reference tree from root to tips; the genome size of each hypothetical taxonomic unit is calculated as the cumulative sum of the origins minus the cumulative sum of losses inferred for previous nodes and the node itself.

The Tree Is Not Too Important, but Family Size Is. Neither different reference trees (using maximum likelihood with or without a gamma distribution of rate variation across sites or using ribosomal protein sequences) nor alternative rootings (within proteobacteria, actinobacteria, or mollicutes) affected the average LGT rate across all genes by >10% (SI Table 5). Changing the reference tree or rooting has little influence on the average LGT rate, because the majority of gene families are of small size and have a discrete taxonomic distribution; 50% of all genes fall into families that occur in ≤14 genomes, and 39% of all families occur in only two often congeneric (Fig. 1*a*) genomes and hence can exhibit one LGT at most.

The cutoff used to assort genes into gene families has a similarly small effect. Repeating the above analyses using cutoffs of 35% amino acid identity or the rather strict value of 40% in the clustering procedure, we found 61,981, and 66,118 families, respectively, and very similar distributions of family sizes (SI Fig. 6). Although the 35% and 40% cutoffs disrupt a few protein families that are united at the 30% cutoff, for example enolase or the ribosomal proteins S11 and L11 (data not shown), all models except $\text{LGT}_{\leq 3}$ are again excluded, and the average LGT rate drops slightly to 0.99 and 0.89, respectively (Table 2).

By contrast, small gene families exert a strong influence on the average LGT rate, because they are abundant and require little LGT to account for their distribution, regardless of the assumed tree. Accordingly, excluding small families from our analysis would deliver higher average rates. For example, if we use a very permissive cutoff of 25%, we obtain 53,349 families. Relative to the 30% cutoff, there are 4,724 fewer families of size <10, because small families are preferentially subsumed into larger families at this lower cutoff (SI Fig. 6). In this case, the average rate increases to 1.55 events per family (SI Table 5). This effect of small families can be further illustrated if we consider only families present in ≥10 genomes, where the $\text{LGT}_{\leq 15}$ model is preferred, and the average LGT rate jumps to 5.3 events per family (SI Table 5). Clearly, families present in ≥10 genomes can accept more LGT, but they speak for only 14% of all families in the data.

Indeed, the effect of excluding small families is even more dramatic than the use of a random tree. If we use a random tree (SI Fig. 4) to infer LGT rates as above, we are assuming there is virtually no vertical inheritance in the gene-distribution data. Accordingly, the random tree corresponds to an evolutionary scenario of LGT only. For the random tree, the Wilcoxon test excludes all models except $\text{LGT}_{\leq 31}$ and yields an average rate of 3.3 LGT events per gene family (Table 2). However, the random tree LGT rate is still lower than for families present in ≥10 genomes using the rRNA reference tree; gene family size bears upon the genome of Eden more heavily than does the assumed tree itself. Thus, although higher estimates for the lower-bound LGT rate can be obtained by disregarding small families (SI Fig. 9), small families contain the majority of genes. The genome-of-Eden constraint applies to all gene families, not just a select population thereof.

Table 2. LGT and gene-loss statistics under different LGT allowances

		LGT allowances				
	Tree	LGT$_{\le 1}$	LGT$_{\le 3}$	LGT$_{\le 7}$	LGT$_{\le 15}$	LGT$_{\le 31}$
Average per-family LGT rate	ML	0.65	1.08*	1.41	1.65	1.83
	NJ	0.65	1.06*	1.36	1.60	1.77
	ML$_{35\%}$†	0.61	0.99*	1.26	1.46	1.60
	ML$_{40\%}$†	0.56	0.89*	1.11	1.26	1.37
	Random	1.0	1.79	2.37	2.85	3.27*
Percent of families accepting LGT$_{max}$	ML	65	8.3	0.2	0	0
	NJ	65	6.8	0.1	0	0
	Random	99.9	18.2	0.3	0	0
Average losses per family	ML	4 ± 9	2 ± 7	2 ± 6	1 ± 5	1 ± 4
	NJ	4 ± 9	2 ± 7	1 ± 5	1 ± 4	1 ± 3
	Random	17 ± 21	12 ± 20	9 ± 18	6 ± 15	4 ± 13
Percent of families with no losses	ML	65	77	85	89	92
	NJ	66	78	85	90	93
	Random	39	61	72	79	84
Average origin/loss ratio	ML	1/2	1/1	1/1	3/1	3/1
	NJ	1/2	1/1	2/1	3/1	3/1
	Random	2/17	1/4	1/3	2/3	1/1

ML, maximum likelihood; NJ, neighbor joining.

*The corresponding ancestral genome-size distribution was not excluded with the Wilcoxon test. All other LGT allowances per row were excluded at $\alpha \le 0.05$ (see also SI Table 5).

†Values in the rows ML$_{35\%}$ and ML$_{40\%}$ were inferred from the dnaml tree by using the clusters obtained for the amino acid identity cutoff indicated. All other values correspond to the clusters obtained with the 30% identity cutoff.

Discussion

The frequency of LGT affects inferred ancestral genome size. No LGT results in untenably large ancestral genomes, whereas too many LGTs result in untenably small ancestral genomes. With the present sample, an average LGT rate on the order of ≈1.1 events per family per family life span (Fig. 3 and Table 1) provides the best fit of inferred ancestral genome sizes to those currently observed in real microbes. This average LGT rate is a very conservative lower bound, because it is based on the assumptions that all families investigated contain orthologs only, and that all gene trees are compatible.

One could argue that ancient genomes were bigger than those of today, and that the amount of LGT inferred here is still not necessary. Indeed, it has been suggested that the vast majority of all LGT occurred before the origin of cells, and that little or none has occurred since (1). However, this suggestion cannot be true, because nucleotide-pattern comparisons indicate that LGT is still an ongoing process today (6–8). One could also argue that ancient genomes were much skimpier than those of today, and that they inflated only recently in a case of evolutionary last-minute shopping, such that higher average LGT rates than those inferred here would be tenable. However, in the absence of evidence to the contrary, Occam's razor would prefer the simpler premise that genome sizes, rates of loss, and rates of LGT in the past, on average, were not fundamentally different from those of today. In the LGT$_{\le 3}$ model, genome size is not only similar to the values currently observed among prokaryotes (Tables 1 and 3); it is also far more constant across time than in the other models (Fig. 3). The same is true for gene-origin and -loss frequencies (SI Fig. 10). By allowing the frequencies of LGT, gene origin, and gene loss to vary freely, we obtain a picture of genome evolution that is marked by uniformity of all three parameters over time and lineages, but only if genome-size distributions remain uniform as well.

This observation and the comparison of modern and inferred genome-size distributions (Fig. 3 and Table 1) indicate that average LGT rates of ≈1.1 LGT per family are necessary and sufficient to account for the present distribution of genes across 190 prokaryotic genomes. This conservative lower-bound estimate stands against and is irreconcilable with recent inferences from gene-tree comparisons that as much as 86% (10) or even 98% (11) of all genes are related by vertical inheritance only. The burgeoning effects that so much vertical inheritance would have upon ancestral genome sizes were not considered in those studies. Rampant vertical inheritance leads to the genome of Eden, and a modest amount of LGT offers remedy.

Table 3. Number of genomes ≤1,000 families in size under different LGT allowances

<table>
<tr><th></th><th colspan="8">Number of genomes</th></tr>
<tr><th>Genome size (number of families)</th><th>Modern genomes</th><th>Loss only</th><th>Single origin</th><th>LGT≤1</th><th>LGT≤3</th><th>LGT≤7</th><th>LGT≤15</th><th>LGT≤31</th></tr>
<tr><td>0–100</td><td></td><td></td><td></td><td></td><td>||</td><td>||||</td><td>||||</td><td>||||||</td></tr>
<tr><td>101–200</td><td></td><td></td><td></td><td></td><td></td><td>||</td><td>||||</td><td>||||</td></tr>
<tr><td>201–300</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td>||</td><td>||</td></tr>
<tr><td>301–400</td><td>||||</td><td></td><td></td><td></td><td>||||</td><td>||||||||</td><td>||||||||||</td><td>||||||||||||</td></tr>
<tr><td>401–500</td><td>||||||</td><td></td><td>||</td><td>||||||</td><td>||||||||</td><td>||||</td><td>||</td><td>||||||||</td></tr>
<tr><td>501–600</td><td>||||||</td><td></td><td>||||||</td><td>||||</td><td>||</td><td>||</td><td>||||||||</td><td>||||</td></tr>
<tr><td>601–700</td><td>||</td><td></td><td>||</td><td></td><td>||||||</td><td>||||||</td><td>||||||</td><td>||||||||||</td></tr>
<tr><td>701–800</td><td>||</td><td></td><td>||||</td><td>||||</td><td>||||</td><td>||||||</td><td>||||||||</td><td>||||||||</td></tr>
<tr><td>801–900</td><td>||||||||</td><td>||||</td><td>||||||</td><td>||||||||</td><td>||||||||||</td><td>||||||||</td><td>||||||||</td><td>||||||||||</td></tr>
<tr><td>901–1,000</td><td>||||||</td><td>||||</td><td>||</td><td>||</td><td>||</td><td>||||</td><td>||||||||</td><td>||||||</td></tr>
<tr><td>Sum</td><td>37</td><td>8</td><td>24</td><td>31</td><td>43</td><td>49</td><td>66</td><td>73</td></tr>
</table>

Above and beyond our full-benefit assumptions, the lower-bound nature of our ≈1.1 LGT per family estimate has two further caveats. First, it is possible that the first origin we infer for each gene is not a birth event, but itself is an LGT from an unsampled genome. Although no genome sample size would exclude that possibility, if we assume that all families were born outside rather than within the lineages sampled, our estimate for the average rate would increase only to ≈2.1. Second, our methods count only observed events; unobserved gene families or events (27) are disregarded.

Our findings indicate that LGT occurs very frequently among prokaryotes in terms of having impact upon individual gene family distributions, in that at least 65% of all families (and given the ultraconservative nature of our full benefit assumptions, probably all) have been affected during the course of evolution. These results can be taken as support for the view that a core set of genes that has remained immune to LGT throughout all of evolution is unlikely to exist (28, 29). The estimates of the average LGT rate reported here represent solely the amounts required to keep ancestral genome-size distributions within realistic bounds; additional contributions from gene-tree comparisons or nucleotide-pattern analyses were not considered. However, despite much LGT, gene-distribution patterns are still nonrandom, as Fig. 1*b* attests. Further specification of the extent of this important process of natural variation among prokaryotes is germane to understanding the evolutionary mechanisms that govern the distribution of genes across genomes.

Materials and Methods

Data. Completely sequenced prokaryotic genomes were downloaded from the National Center for Biotechnology Information (NCBI) web site (www.ncbi.nlm.nih.gov; August 2005 version). For each species, only the strain with the largest number of genes was used. Of 190 genomes (562,321 proteins) in the data, 22 are archaebacterial, and 168 are eubacterial.

Gene Families (PAPs). All proteins in the 190 genomes were clustered by similarity into gene families by using the reciprocal best BLAST hit (BBH) approach. Each protein was BLASTed against each of the genomes. Pairs of proteins that resulted as reciprocal BBHs of E value $< 1^{-10}$ were aligned by using ClustalW (30) to obtain amino acid identities. Using a cutoff of 30% amino acid identity in the clustering procedure (22), the proteins fell into 57,670 families with two or more members, in addition to 149,894 singletons that were excluded from pattern analysis. The resulting gene families represent a PAP of gene distribution across the prokaryotic genomes. A PAP includes 190 digits; if a gene family includes one or more genes from genome i, then digit x_i in its corresponding pattern is "1"; otherwise, it is "0."

Reconstruction of Phylogenetic Trees. For the reference tree, the sequences of the rRNA operon (16S, 23S, and 5S) from all 190 genomes were aligned by using ClustalW (30) and concatenated, and gapped sites were removed. This alignment was used for phylogenetic reconstruction by maximum likelihood (ML) with and without rate variation using dnaml (31), PhyML (32), and Neighbor Joining (33). The tree of concatenated L11 and S11 ribosomal protein sequences was inferred with PhyML (32). Trees were rooted between archaebacteria and eubacteria; additional roots of the dnaml tree (between proteobacteria, actinobacteria, or mollicutes and other genomes) were tested. The random tree was obtained by shuffling species names in the ML tree and rooting on the longest internal edge. All Newick format trees are provided in SI Table 6.

Evolutionary Model Reconstruction and Calculation of Ancestral Genome Size. In the loss-only model, all gene families (57,670) are assumed to have originated at the root. The loss events for each gene are estimated by using a binary recursive algorithm that scans the tree and infers the minimum number of losses. When a gene is absent in a whole clade, a single loss event is inferred in the common ancestor of that clade (e.g., Fig. 2*a*, clade III). In the SO model, each gene family is assumed to have originated at its first occurrence on the reference tree. A binary recursive algorithm scans the tree root to tips to identify the first hypothetical taxonomic unit (ancestral node) that is the common ancestor of all gene "present" cases (e.g., the common ancestor of clades I + II in Fig. 2*b*).

In the $\text{LGT}_{\leq 1}$ model, each gene family is allowed to have two gene origins, where one is an LGT. The first origin is inferred as in the SO model, followed by researching for a gene origin in each of the two clades branching from the first-origin node (e.g., Fig. 2*c*). If the hypothetical taxonomic unit that was inferred as the first origin has no gene "absent" descendants, the gene family is inferred to have a single origin. Once the nodes of the two origins are set, losses are inferred as in the loss-only model.

We tested additional models allowing 4, 8, 16, and 32 origins, where one is an origin, and the rest are LGTs. These are implemented in the same way as in the $\text{LGT}_{\leq 1}$ model, except that the origin search is iterated. For example, a search for origins under the $\text{LGT}_{\leq 4}$ model entails (*i*) a search for the first origin (as in the SO model); (*ii*) a search for the next origin in descendants (as in the $\text{LGT}_{\leq 1}$ model); and (*iii*) for each next origin, another search. If an origin has no gene "absent" descendants, the number of origins inferred is smaller than the maximum allowed (e.g., Fig. 2*d*, clade II, where three origins are inferred under the $\text{LGT}_{\leq 4}$ model). The distributions of ancestral and modern genome sizes were compared by using the Wilcoxon Mann–Whitney nonparametric test (24). For cutoffs other than 30%, the inferred distributions were compared with the modern distribution for clusters at the respective cutoff.

We thank Giddy Landan, Einat Hazkani-Covo, Mike Steel, Andrew Roger, and Martin Embley for comments and discussions. Financial support was provided by Deutsche Forschungsgemeinschaft.

1. Woese CR (2000) *Proc Natl Acad Sci USA* 97:8392–8396.
2. Kurland CG, Canback B, Berg OG (2003) *Proc Natl Acad Sci USA* 100:9658–9662.
3. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P (2006) *Science* 311:1283–1287.
4. Doolittle WF (2004) in *Microbial Phylogeny and Evolution: Concepts and Controversies*, ed Sapp J (Oxford Univ Press, New York), pp 119–133.
5. Gogarten JP, Doolittle WF, Lawrence JG (2002) *Mol Biol Evol* 19:2226–2238.
6. Moszer I, Rocha EP, Danchin A (1999) *Curr Opin Microbiol* 2:524–528.
7. Mrazek J, Karlin S (1999) *Ann NY Acad Sci* 870:314–329.
8. Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) *Nat Genet* 36:760–766.
9. Lerat E, Daubin V, Ochman H, Moran NA (2005) *PloS Biol* 3:e130.
10. Beiko RG, Harlow TJ, Ragan MA (2005) *Proc Natl Acad Sci USA* 102:14332–14337.
11. Ge F, Wang LS, Kim J (2005) *PLoS Biol* 3:e316.
12. Penny D, Hendy MD, Steel M (1992) *Trends Ecol Evol* 7:73–79.
13. Hillis DM (1995) *Syst Biol* 44:3–16.
14. Lopez P, Casane D, Philippe H (2002) *Mol Biol Evol* 19:1–7.
15. Snel B, Bork P, Huynen MA (2002) *Genome Res* 12:17–25.
16. Kunin V, Ouzounis CA (2003) *Genome Res* 13:1589–1594.
17. Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) *BMC Evol Biol* 3:2.
18. Hao W, Golding GB (2006) *Genome Res* 16:636–643.
19. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, *et al.* (1999) *Nature* 399:323–329.
20. Rosas-Magallanes V, Deschavanne P, Quintana-Murci L, Brosch R, Gicquel B, Neyrolles O (2006) *Mol Biol Evol* 23:1129–1135.
21. Doolittle WF, Boucher Y, Nesbo CL, Douady CJ, Andersson JO, Roger AJ (2003) *Philos Trans R Soc London B* 358:39–57.
22. Enright AJ, Van Dongen S, Ouzounis CA (2002) *Nucleic Acids Res* 30:1575–1584.
23. Rivera MC, Lake JA (2004) *Nature* 431:152–155.
24. Zar JH (1999) *Biostatistical Analysis* (Prentice–Hall, Upper Saddle River, NJ).
25. Moran NA (2002) *Cell* 108:583–586.
26. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, *et al.* (2001) *Nature* 409:1007–1011.
27. Spencer M, Susko E, Roger AJ (2006) *Evol Bioinformat* 2:165–186.
28. Susko E, Leigh J, Doolittle WF, Bapteste E (2006) *Mol Biol Evol* 23:1019–1030.
29. Doolittle WF (1999) *Science* 284:2124–2128.
30. Thompson JD, Higgins DG, TJ Gibson (1994) *Nucleic Acids Res* 22:4673–4680.
31. Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) (Department of Genome Sciences, University of Washington, Seattle).
32. Guindon S, Gascuel O (2003) *Syst Biol* 52:696–704.
33. Saitou N, Nei M (1987) *Mol Biol Evol* 4:406–425.