The impact of long-distance horizontal gene transfer on prokaryotic genome size

Otto X. Cordero[1] and Paulien Hogeweg

Theoretical Biology and Bioinformatics, University of Utrecht, Padualaan 8 3584 CH, Utrecht, The Netherlands

Edited by James M. Tiedje, Michigan State University, East Lansing, MI, and approved October 30, 2009 (received for review July 11, 2009)

Horizontal gene transfer (HGT) is one of the most dominant forces molding prokaryotic gene repertoires. These repertoires can be as small as ≈200 genes in intracellular organisms or as large as ≈9,000 genes in large, free-living bacteria. In this article we ask what is the impact of HGT from phylogenetically distant sources, relative to the size of the gene repertoire. Using different approaches for HGT detection and focusing on both cumulative and recent evolutionary histories, we find a surprising pattern of nonlinear enrichment of long-distance transfers in large genomes. Moreover, we find a strong positive correlation between the sizes of the donor and recipient genomes. Our results also show that distant horizontal transfers are biased toward those functional groups that are enriched in large genomes, showing that the trends in functional gene content and the impact of distant transfers are interdependent. These results highlight the intimate relationship between environmental and genomic complexity in microbes and suggest that an ecological, as opposed to phylogenetic, signal in gene content gains relative importance in large-genomed bacteria.

functional gene content | microbial genomes | scaling | lateral gene transfer

In sequenced species, the size of the prokaryotic gene repertoire spans over an order of magnitude, from ≈200 genes in extreme endosymbionts like *Carsonella ruddii* (1) to >9,000 genes in soil-dwelling bacteria. The common view is that these differences reflect mainly the external demand for functional complexity that is imposed by the organism's lifestyle and environment (2–4). Indeed, the trends between genome size and functional gene content variation show that large genomes are enriched in functions like regulation, signaling, or secondary metabolism (2), which could allow organisms to reach a higher degree of ecological diversification.

Along bacterial lineages, gene repertoires are shaped by the dynamics of horizontal gene transfer (HGT), gene duplications, losses, and vertical inheritance (5, 6). Gene transfer is an important mechanism in prokaryotic genome evolution, both among early ancestors and in present-day ecosystems (7–9), with recent evidence showing it can lead to large variations in gene content and genome size in ecological time scales (10, 11). Virtually all prokaryotic genes may have been involved at least once in horizontal transfer (5, 7). Nevertheless, because most transfers take place between closely related organisms (12) with compatible gene content, it is unclear what is the contribution of those genes coming from distant lineages, possibly reflecting ecological rather than phylogenetic associations, to the different aspects of genome complexity.

In this article, we address that question by measuring both the cumulative impact of distant HGTs (dHGTs) along each lineage and its recent history on the leaves of the tree of life, as a function of genome size. Our study shows that large bacterial genomes tend to harbor a disproportionate amount of polyphyletic genes, often shared with other large genomes in distant lineages. An analysis of the functional bias of distant transfers reveals a strong connection between the incidence of dHGT across the different functional groups and their specific trends with respect to genome size. These results are important for the understanding of the processes underlying complexification in microbial genomes and highlight the interdependence between prokaryotic genome organization and environmental complexity.

Results and Discussion

Cumulative Impact of dHGT Increases Nonlinearly in Large Genomes. Our first approach to infer the contribution of dHGT to the size of bacterial gene repertoires builds on the reconciliation of the presence–absence distributions of gene families with the tree of life (TOL) (13). This method can detect transfers along the whole history of descent, as long as they explain patchy presence–absence distributions. However, transfers that do not disrupt the phylogenetic distribution of the family will remain undetected. Our method uses a maximum parsimony algorithm to reconstruct the evolutionary history of each gene family on the species tree (14, 15) (see *Materials and Methods*). Once ancestral gene contents have been inferred, we can measure the cumulative impact of dHGT on each genome by counting how many of those families that contribute to the species' gene content have been "created" multiple times along the evolutionary history. Multiple gene origins will be inferred in those families that show phylogenetically discordant presence–absence patterns. An illustration of what this means can be seen in Fig. 1: the red dashed lines follow the line of descent of a family that is created in two different lineages. This family will be thus added to the list of families involved in dHGT in all five descendant species that have kept the gene. The parsimony reconstruction relies on a predefined cutoff parameter, γ, which equals the number of independent gene losses that we are willing to accept without invoking multiple gene origins to explain the family distribution. In other words, γ defines the minimum topological distance, in number of ancestors, at which we detect transfer events.

Although similar in spirit to a method recently used in ref. 8 to infer the cumulative impact of horizontal transfers, our methods focus particularly on those presence–absence patterns that are highly discordant with the species tree, up to our cutoff γ. By increasing γ, we cannot only better discriminate horizontal transfers from gene losses, but, as shown in Fig. 2*A*, we focus on those transfers that take place between phylogenetically distant organisms. Fig. 2*A* shows that, for our tree with 333 species, at topological distances of three ancestors the estimated average percentage of 16S rRNA nucleotide identity between nodes is ≈87%, which in our species set normally crosses the boundary of taxonomic orders (see *Materials and Methods*). In this article we focus on the impact of theses type of transfers, occurring between relatively distant organisms and contributing to create

Author contributions: O.X.C. and P.H. designed research; O.X.C. performed research; O.X.C. analyzed data; and O.X.C. and P.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be sent at present address: Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. E-mail: ottoxcordero@gmail.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0907584106/DCSupplemental.

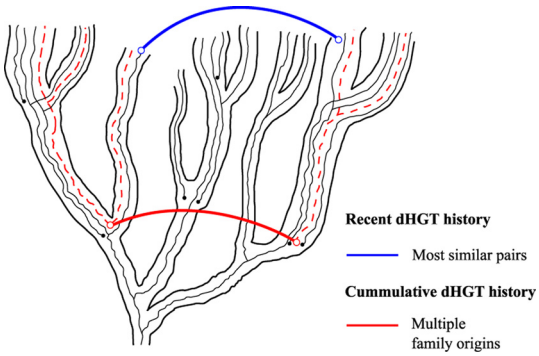

Fig. 1. Illustration of the ancestral and recent dHGT events on a TOL. Our analysis focus on the cumulative effect of ancestral dHGT and the recent history of dHGT in a separate manner. The dashed red lines follow the line of descent of a gene that has been transferred in an early ancestor, leaving a patchy presence–absence distribution as evidence of this event. The blue line shows a recent transfer event. The pairs of genes resulting from this recent transfer are closest to each other, despite the large distance between the species where they are found.

a strong polyphyletic signal in their gene content. We refer to them explicitly as dHGTs.

The estimation of dHGT based on presence–absence patterns is greatly affected by how stringent or inclusive the definition of a protein family is (16). For this reason, we use two extreme alternative approaches to define the protein families present in our set of 333 genomes annotated in the String v7.1 database (17). For the most stringent definition of a protein family, we use a database of small protein clusters (SPCs), that is cliques of reciprocal best hits (18), resulting in ≈140,000 families in our species set. In the most inclusive definition, we use clusters of orthologous groups (COGs) (19) and nonsupervised COGs (NOGs) as defined in String v7.1, which merge the SPCs into ≈39,000 COG/NOGs (hereafter referred to simply as COGs).

Fig. 2 *B–D* summarizes the results of the cumulative count of dHGT events with respect to genome size by using SPC data. Because of the uneven species sampling, with the stringent SPC family definition some genomes have an underrepresentation of families (many singletons), which is indicated in Fig. 2 *B–D* by the size and grayscale intensity of the data points. We see that in those species with a "healthy" percentage of proteins in SPCs (e.g., >65%, which holds for ≈68% of the species) the number of dHGT families increases nonlinearly with the total number of families. For those genomes, the relationship between number of SPCs and number of SPCs involved in dHGT can be described in log–log scale by a straight line with slope $\alpha > 1$. The greater α, the faster the increase in proportion of dHGT with respect to genome size. The linear regression fit in Fig. 2*C* (for species with a ratio of at least 65% SPCs/number of genes), gives $\alpha \approx 1.6$. These estimates of dHGT in protein clusters were obtained with $\gamma = 5$, which traces 350 families back to the last prokaryotic common ancestor (LPCA). Reconstructions with $\gamma = 3$ and $\gamma = 8$, which trace 152 and 605 families back to LPCA, result in $\alpha \approx 1.4$ and ≈ 2, respectively. The differences in α reflect the fact that the lower γ is the closer the number of dHGT families gets to saturation, in which case the trend in dHGT is the 1:1 line. This observation agrees with recent reports about the cumulative incidence of horizontal transfer in protein families, suggesting that most families could have been involved at least once in gene transfer (8). However, the fact that an increase in γ results in higher values of α indicates that larger genomes tend to harbor a larger proportion of highly polyphyletic families, which require multiple deletions to be reconciled with the species phylogeny.

The results described above for the case SPCs can be seen even more clearly with COGs. Because of the more inclusive nature of the COG data, all our species are well covered by gene families, and the superlinear trend between number of families and number of families with dHGT can be seen much more cleanly in Fig. 2*E*. Fig. 2*E* shows the results of reconstructions based on costs spanning from $\gamma = 1$ to $\gamma = 3$, which result in 728–2,834 families in the LPCA. Here, we use smaller values of γ because larger families have their origins closer to the root, biasing transfers toward nodes at higher taxonomic levels, that is, between shorter topological distances. The addition of these ancestral transfers, shared by more lineages, increases the cumulative count of dHGT and produces a more compact trend. In summary, also in the more inclusive COG family definition we see that the cumulative impact of phylogenetically incoherent families increases disproportionately with genome size.

Paradoxically, it is the low incidence of long-distance transfers in most families (see Fig. S1) that helps explain our results. The fact that gene families tend to occur within closely related species implies that there are sets of core genes that percolate toward the leaves at different taxonomic levels. If a genome grows along a branch of the tree, it must do so by acquiring genes that are

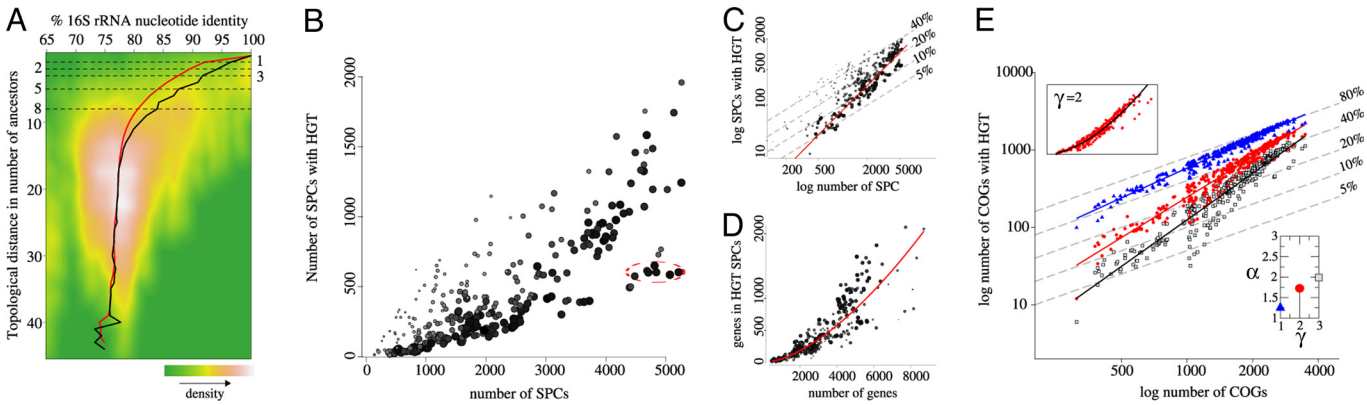

Fig. 2. dHGT and its cumulative impact. (*A*) Smoothed color density representation of the scatterplot between the topological distance and the percentage of 16S rRNA nucleotide identity between the leaves of the eubacteria subtree. The black line corresponds to the averaged 16S rRNA identities. The red line is the averaged trend when considering the estimated distances between internal nodes (see *Materials and Methods*). (*B*) Number of SPCs with dHGT vs. total number of SPCs in the genome, with $\gamma = 5$. (*B–D*) Shown are the percentages of genes in clusters with the size and color intensity of the data points. For most, well-covered species, there is a superlinear trend between the size of the gene repertoire and the contribution of dHGT. Encircled species correspond to *B. cereus/anthracis/thuringensis*. *C* shows the same data in log–log scale, and *D* shows the trend with respect to the total number of genes. The fitted lines have an $\alpha \approx 1.6$. (*E*) Number of dHGT COGs with respect to the total number of COGs in the genome for different values of γ.

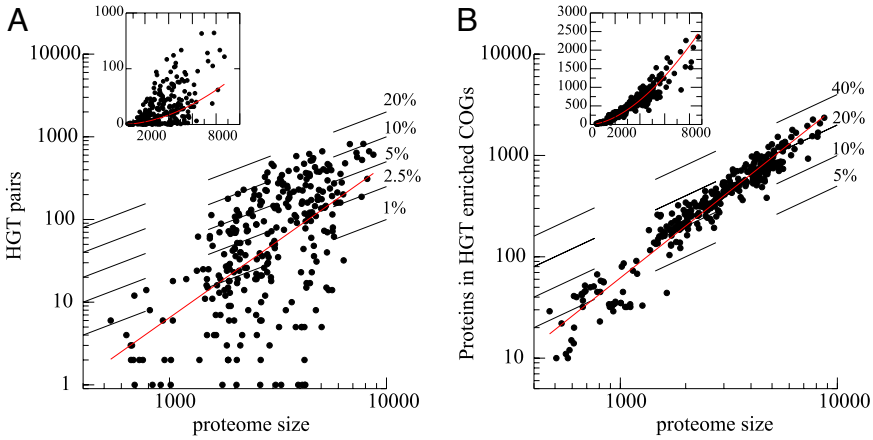

Fig. 3. Superlinear enrichment of recent transfers in large genomes (*A*) Relationship between genome size, in number of proteins, and number of proteins with bidirectional best hits in species at least branches apart. Despite the larger variation, the trend is consistent with the results shown in Fig. 2. The slope of the fit is $\alpha \approx 1.8$. (*B*) Extrapolation from HGT to gene families confirms superlinear relationship between genome size and dHGT and shows that the superlinear scaling of gene family groups is related to the dHGT enrichment. The log–log slope is $\alpha \approx 1.7$.

outside of this core, thus growing toward a set of potentially polyphyletic families. Without a core set of genes being conservatively transmitted along lineages, we would see that the number of dHGT families is at saturation (i.e., a linear trend with respect to genome size). In agreement, when we randomize the position of the leaves on the species tree we see that, independently of genome size, in average >75% of all SPCs in a genome have polyphyletic distributions at $\gamma = 5$, which results in a linear trend at a high offset.

Interestingly, in the case of the *Bacillus cereus/anthracis/thuringensis* (encircled points in Fig. 2*B*) we detect a much lower incidence of dHGT than in other genomes of similar size. In fact, we find that with SPCs the general superlinear trend between genome size and dHGT does not hold for the *Bacilli* group, which fits $\alpha \approx 1$. Even in the COG data, which focuses on more ancient dHGT, the α of *Bacilli* is ≈20% lower than in the whole species set. This observation shows that the nonlinear acquisition of polyphyletic families is not a trivial result of the reconstruction and suggests that there may be lineage-specific trends in the impact of dHGT. It this respect, it is important to notice that a large percentage of our species are within the proteobacteria (≈45%), and that the trends described in Fig. 2 *B–E* are therefore dominated by this large phylum (see Fig. S2).

In agreement with the view of prokaryotic genome size as an adaptation to changing environments, there is a strong correlation between the projected genome sizes of the ancestors involved in exchange of gene families. The Pearson correlation between genome size and the average size of their dHGT partners is $r = 0.48$ and 0.59 ($P < 0.001$) for small clusters ($\gamma = 5$) and COGs ($\gamma = 3$), respectively. Moreover, this correlation is lost in the control reconstructions with randomized leaves. The apparent size assortativeness of the dHGT partners makes sense if we think that genome size is determined by the complexity of the environment, because in such case organisms within the same ecological boundaries should tend to fall within relatively similar size classes.

Superlinear dHGT Trend Confirmed with Recent Transfers. In addition to the inference of cumulative dHGT, we followed a complementary dHGT detection approach by finding proteins whose closest homologs occur in distant species. Having studied the cumulative history of transfers with large and small protein clusters, this is the next and highest level of detail at which we can study phylogenetically incoherent protein distributions. We collected Smith-Waterman bidirectional best hits (SW BBH) between proteins present in species separated by at least eight branches in the TOL (see *SI Text*). Similarly as in the previous section, when we detect a pair of closest homologs in distant species we hypothesize that the gene has either diverged (or been lost) multiple times or that it has been horizontally transferred. In contrast to the presence–absence reconstruction, this method detects preferentially recent transfers that have not yet undergone vertical inheritance. Moreover, whereas the cumulative analysis focuses on shared histories of transfers, our recent transfer predictions reflect patterns of DNA exchange that are specific to the organism lifestyle and environment (9). Indeed, when we cluster organisms based on the number of recent transfers between them, we recover groups of species with a clear ecological signal, such as halophilic bacteria and archaea, or metabolically coupled species such as the methanogenic *archaea Methanosarcina barkeri*, the acetogenic bacteria *Moorella thermoacetica*, known to grow sustainedly on methanol only in coculture with a hydrogen-consuming methanogens (20), and the synthropic bacteria *Syntrophus aciditrophicus*, which can produce twice more acetate from benzoate in presence of methanogenic partners (21).

Fig. 3*A* shows the relationship between genome size and number of dHGT. The large spread in the data compared with Fig. 2 reflects a number of factors: a smaller set of proteins in the *y* axis, the large biases in species sampling, and cases of strong ecologically association like those mentioned in the previous paragraph. Still, the trend between genome size and proportion of dHGT is consistent with the results from the cumulative dHGT count and shows that, on average, larger genomes are composed of a larger proportion of recently transferred genes. Furthermore, in accordance with the reconstruction results, we observe that genome size is strongly correlated to the average genome size of their dHGT partners, $r = 0.50$ ($P < 0.001$).

Approximately 93% of the recent transfers shown in Fig. 3 take place between organisms with 16S rRNA nucleotide identities <90% (see *Materials and Methods*), which indicates that, despite the uneven species sampling, our topological distance cutoff does focus on distant transfers. Moreover, surrogate methods, which are completely independent of the species tree, confirm our results. These methods detect recently transferred genes originating from distant cellular sources based on the idea that the DNA composition of recently transferred genes should be more similar to the foreign source than to the rest of the genomic background (22). One such analysis of 116 prokaryotic genomes by Nakamura et al. (23) shows that, in clear agreement

with our results, the proportion of recent dHGT, detected by their aberrant DNA patterns, increases with genome size. We calculated $\alpha = 1.76$ based on their data, with a good log–log fit ($R^2 = 0.87$).

Based on the SW BBH identification of dHGT candidates, we tried to extrapolate from horizontally transferred genes back to gene families by detecting those COGs that are enriched in dHGT. The purpose of this exercise is 2-fold: we wondered whether by extrapolating to gene families we could obtain a less noisy estimate of the contribution of recent dHGT with respect to genome size, and we asked whether the incidence of dHGT on certain gene families implies their enrichment in larger genomes. We found a set of 4,250 COGs with significantly high incidence of dHGT ($P < 0.01$) and plotted the number of genes in these COGs vs. the total number of proteins in Fig. 3*B*. Fig. 3*B* shows that selecting COGs based only on their overrepresentation in dHGT, gives us a group of genes whose abundance scales superlinearly with genome size. In summary, the different approaches we have taken to quantify HGT as a function of genome size converge to point out that larger genomes have a larger percentage of polyphyletic genes.

dHGT and the Trends Between Genome Size and Functional Content. In recent years, it has been established that there are supralinear and superlinear trends between functional categories and genome size (24–26). Supralinear trends are observed in information processing functions like translation, which make a low percentage of the gene content of larger genomes, compared with their smaller counterparts. Superlinear trends are seen in functions like gene regulation, for which large genomes dedicate a large proportion of their gene content. Having seen the superlinear relationship between genome size and dHGT, we wondered whether these two observations are in fact connected, that is, whether dHGT is biased to occur more often in those functional groups that make a greater contribution to the gene content of large genomes.

To measure whether the incidence in dHGT correlates to the patterns of functional enrichment, Fig. 4*A* shows the relationship between the relative incidence of recent dHGT in functional groups and the trend of these groups vs. genome size. The ranking by relative enrichment in dHGT (Fig. 4*A*, *x* axis) is obtained by comparing the abundance of a functional group in the set of dHGT proteins against the expected frequency if one would sample the same number of proteins from each species (see *Materials and Methods*). In this way we can measure the dHGT enrichment of a functional group on top of their size-dependent biases. To get a more specific mapping between functional groups and dHGT, functional groups are here defined based on PFAM clans (27). The *y* axis in Fig. 4*A* shows how the number of genes in a PFAM clan scales with the total number of genes. Values <1 and >1 reflect supralinear and superlinear trends, respectively, and are obtained by fitting a straight line in log–log scale, in the same way we did when measuring the trends in dHGT. We observe a strong correlation between the level of dHGT enrichment and the nonlinearity of the functional trends ($r = 0.56$, $P < 0.001$), that is to say recent dHGT often plays a major role in the evolution of those functional categories that are enriched in large genomes. The clans in the upper right corner of Fig. 4*A* are substrate binding domains often seen in one-component transcriptional regulators [periplasmic binding protein (PBP)], generic membrane spanning transporters [drug/metabolite transporter (DMT) and multiple facilitator superfamily (MFS)], DNA binding domains [helix-turn-helix (HTH)], and catalytic domains often found in proteins involved in lipid and secondary metabolism (AB_hydrolase). Probably the most notable outliers are the CheY-like two-component systems. The low incidence of long-distance transfers in these proteins is consistent with the idea that lineage-specific expansions are the main source of new signal transduction genes in most signaling-enriched genomes (28). The discrepancy between the position of the HTH and CheY clans on the *x* coordinate in Fig. 4*A* shows that, although they scale in a similar way with respect to genome size, transcription regulation via phosphorylation cascades and direct substrate recognition in one-component systems (29) follows quite different evolutionary histories (see Fig. S3).

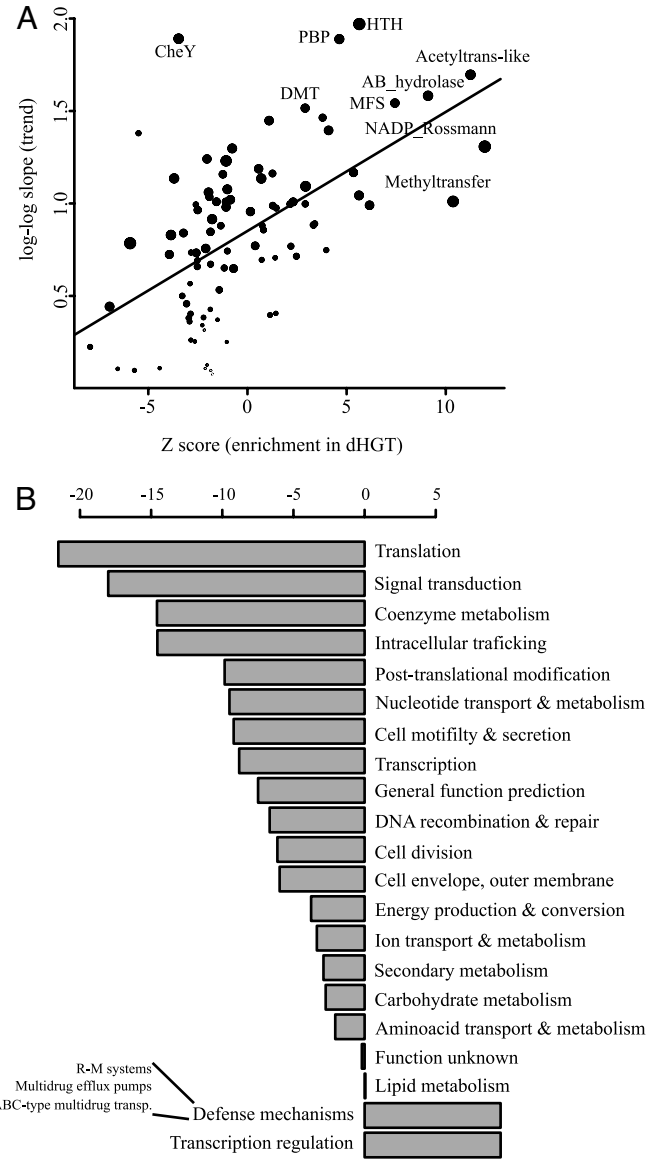

Fig. 4. Functional bias of dHGT. (*A*) dHGT depletion/enrichment vs. the scaling of PFAM clans with respect to genome size. The *y* axis corresponds to the scaling factor of the PFAM clans. Shown are the slopes for those clans present in at least 90% of the genomes (Fig. S3 shows the same picture for a cutoff of 70%). On the *x* axis, values >0 and <0 indicate enrichment or depletion in dHGT beyond the size-dependent expectation. The size of the points is scaled according to the goodness of the fit (R^2). DMT and MFS appear often in amino acid transporters. The NADP_Rossmann and methyltransferase clans cover large numbers of proteins, which could explain their log–log slope is close to 1. (*B*) Depletion/enrichment of dHGT across COG functional categories. For comparison with *A*, we have added the transcription regulation group, found under pathways or functional system (subcategory of transcription).

Fig. 4*B* provides a more global view of the functional bias of

the recent dHGT genes in terms of their COG annotation. Again, this ranking takes into account that the abundance of some functional groups depends on genome size. In this stringent test, only two categories are enriched in dHGT: defense mechanisms and transcription regulation. Within defense mechanisms the high incidence of dHGT corresponds specifically to transporters normally involved in extrusion of toxic compounds and to restriction—modification systems. The most dHGT-enriched functional category, transcription regulation, contains one-component systems, which are fusions of DNA binding and substrate recognition domains, linking in this way environmental stimuli to regulatory responses in a single protein (29).

Having established the relationship between incidence of dHGT and trends in functional groups, it is worth noting that the superlinear trend in dHGT genes or families is rather independent of individual functional categories, like transcription regulation. Removing this category from the data shown in Fig. 3*B* has only a marginal effect on slope α, which changes from $\approx$1.7 to $\approx$1.65. This finding suggests that the trends in functional content reflect the more general phenomenon of genome growth by dHGT.

How Can We Explain the Relationship Between Genome Size and dHGT? In recent years, the idea has grown that horizontal transfer is the dominant force in expanding the gene repertoires of bacterial genomes (30). In fact, an analysis of gene phylogenies in γ-proteobacteria suggests that only a small percentage of the genes found in extant species descend from the most recent common ancestor of this group (5). This idea implies that microbial genomes are highly dynamic entities, constantly acquiring and losing genes (30). Thus, as shown in ref. 8, for most organisms, the percentage of genes transferred at any point in the species' history could be close to 100%, independent of size. However, we have seen here that when we focus on those polyphyletic genes, which are distributed in a patchy manner among distant species, a superlinear trend with respect to the size of the genome emerges. This trend shows that whereas most genes may have originated from transfers, large genomes tend to have a greater contribution of phylogenetically discordant genes. Moreover, all our methods show a strong correlation between the size of the genomes where those polyphyletic genes are found. We envision two nonmutually exclusive explanations for these observations.

One type of explanation is that the intrinsic rate of illegitimate recombination per gene increases in large genomes as a result of the larger number of transposons, integrases, or phage elements that may facilitate the integration of foreign DNA (31–33). In addition, larger genomes tend to be composed of multiple plasmids or megaplasmids, which suggest a higher rate of transmembrane DNA translocation. Another factor that may increase the intrinsic rate of HGT is the positive correlation observed between genome size and the modularity of the biochemical networks (34). In line with the complexity hypothesis (35), horizontal transfers could be more successful in more modular genome architectures because of the lower, potentially deleterious, pleotropic effects.

The second type of explanation for the higher effective rate of distant transfers is that the extrinsic rate of dHGT, driven by environmental factors, increases in large genomes. A comparison between metagenomics of farm soil and Sargasso sea samples shows that soil bacteria not only have larger genomes but also live in communities with higher cell density and taxonomic diversity (36, 37). If indeed genome size is correlated to biodiversity and cell density, one can expect that, in average, the amount of DNA available for uptake during the lifetime of a bacterium is larger in large genomes. An attractive hypothesis is that the increase in phylogenetic diversity in the environment and the broadening of gene repertoires in the genome could be a circular process: more complex ecological interactions in species-rich communities could increase the demand for larger gene repertoires, which are expanded by accepting genes from those phylogenetically distant organisms in the environment.

Horizontal transfer detection is an elusive problem, and often there is little overlap between different approaches (22). We therefore rely on independent dHGT detection techniques, which focus on transfer events that occur on different evolutionary time scales. These methods converge in showing that large bacterial genomes tend to be disproportionately enriched in transfers from distant lineages. Moreover, all of our methods reveal a positive correlation between the sizes of the donor and recipient genomes. These results highlight the intimate relationship between genomic and environmental complexity in microbes.

The clear relationship between the incidence of dHGT and the scaling factors of functional groups (Fig. 4*A*) indicates that the trends in functional categories may be directly linked to the incidence of dHGT. It is suggestive that some agreement can be seen as well at the level of specific lineages. In parallel to what we saw in the cumulative impact of dHGT, in *Bacilli*, one-component regulatory systems, which as we saw are often transferred, scale near-linearly with respect to the whole protein set (26). Further studies are needed to elucidate to what extent there are lineage-specific trends in genome complexification and clarify the relationship between the incidence of polyphyly and the ecological dynamics of microbial communities.

Materials and Methods

Inference of Ancestral dHGT Events. The inference of ancestral dHGT events is based on a gene content reconstruction along a TOL. As a reference TOL we use the species tree available in the MicrobesOnline server (www.microbesonline.org) (38). This tree was pruned down to the 333 prokaryotic species present in the String 7.1 database (17). The algorithm (detailed in ref. 15) assigns creation, deletion, and duplication events to the nodes of the TOL for each gene family (see *SI Text*). As shown in Fig. S4, this method results in ancestral genome size distributions that closely resemble that of extant species. We can find families with dHGT by detecting cases in which the most parsimonious reconstruction produces gene creations on multiple ancestors. The cumulative dHG data with SPCs and COGs are shown in Tables S1 and S2, respectively. Table S3 shows the recent dHGT data.

Measuring 16S rRNA Similarities. Prealigned 16S rRNA sequences were obtained from the Ribosomal Database Project (rdp.cme.msu.edu) (39). Positions with >5% gaps where removed from the alignments, leaving blocks of 1,375 nt for bacteria and 1,424 nt for archaea, which were used to calculate all pairwise nucleotide identity percentages within each kingdom. For the purpose of our analysis, 16S rRNA identities between bacteria and archaea were assumed to be below threshold in all cases. To estimate 16S rRNA distances involving ancestral nodes, we averaged the identity percentages found between the groups of extant leaves under the internal nodes in question. We also checked this simple approach by using the ancestral sequences reconstructed by an unweighted parsimony algorithm. We find that both methods give consistent estimates. Based on these estimates, we found that across all our horizontal transfer detection methods, less than $\approx$10% of all transfers are found at 16S rRNA nucleotide identities >90%, and less than $\approx$15% >87%, which normally crosses the border of taxonomic orders. Moreover, removal of those transfers between nodes with 16S rRNA nucleotide identities >90% or even 85%, does not affect the superlinearities (α remains within 90% confidence interval).

Mapping of Functional Categories. We detected the enrichment in dHGT in a functional group by comparing the observed fraction of dHGT in the group to the expected frequency if the same number of genes per species were sampled. In this way we control for the biases of dHGT and functional groups toward large genomes. We calculate the *z*-score, $Z = (\text{observed} - E(x))/\sigma(x)$, relative to the binomial sampling expectation simply to obtain a ranking of underrepresented and overrepresented categories. The binomial expectation was also used to detect COGs enriched in dHGT at $P < 0.01$ in Fig. 3. Because

of the dependency between family size and number of expected transfers (Fig. S1) and the fact that average family size can strongly differ between categories, we prefer the SW BBH data for our analysis.

ACKNOWLEDGMENTS. We thank Gabino Sanchez-Perez for guiding discussions. This work was supported by Netherlands Organization for Scientific Research Grant 635.100.001.

1. Nakabachi A, et al. (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. *Science* 314:267.
2. Konstantinidis KT, Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci USA* 101:3160–3165.
3. Dobrindt U, Hacker J (2001) Whole genome plasticity in pathogenic bacteria. *Curr Opin Microbiol* 45:550–557.
4. Daubin V, Moran NA (2004) Comment on ''The origins of genome complexity.'' *Science* 306:978.
5. Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* 3:e130.
6. Woese CR (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci USA* 97:8392–8396.
7. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 19:2226–2238.
8. Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 105:10039–10044.
9. Beiko RG, Harlow TJ, Ragan MA (2005) Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* 102:14332–14337.
10. Tettelin H, et al. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial ''pan-genome.'' *Proc Natl Acad Sci USA* 102:13950–13955.
11. Thompson JR, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science* 307:1311–1313.
12. Lawrence JG, Hendrickson H (2003) Lateral gene transfer: When will adolescence end? *Mol Microbiol* 50:739–749.
13. Ragan MA, Charlebois RL (2002) Distributional profiles of homologous open reading frames among bacterial phyla: Implications for vertical and lateral transmission. *Int J Syst Evol Microbiol* 52:777–787.
14. Boussau B, Karlberg EO, Frank AC, Legault BA, Andersson SG (2004) Computational inference of scenarios for α-proteobacterial genome evolution. *Proc Natl Acad Sci USA* 101:9722–9727.
15. Cordero OX, Snel B, Hogeweg P (2008) Coevolution of gene families in prokaryotes. *Genome Res* 18:462–468.
16. Dagan T, Martin W (2007) Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci USA* 104:870–875.
17. Von Mering C, et al. (2007) STRING 7: Recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* 35:D358–D362.
18. Klimke W, et al. (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res* 37:D216–D223.
19. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637.
20. Jiang B, et al. (2009) Atypical one-carbon metabolism of an acetogenic and hydrogenogenic *Moorella thermoacetica* strain. *Arch Microbiol* 191:123–131.
21. Elshahed MS, McInerney MJ (2001) Benzoate fermentation by the anaerobic bacterium *Syntrophus aciditrophicus* in the absence of hydrogen-using microorganisms. *Appl Environ Microbiol* 67:5520–5525.
22. Ragan MA, Harlow TJ, Beiko RG (2006) Do different surrogate methods detect lateral genetic transfer events of different relative ages? *Trends Microbiol* 14:4–8.
23. Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* 36:760–766.
24. Cases I, De Lorenzo V, Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* 11:248–253.
25. Van Nimwegen E (2003) Scaling laws in the functional content of genomes. *Trends Genet* 19:479–484.
26. Cordero OX, Hogeweg P (2007) Large changes in regulome size herald the main prokaryotic lineages. *Trends Genet* 23:488–493.
27. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36:D281–D288.
28. Alm E, Huang K, Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* 2:e143.
29. Ulrich LE, Koonin EV, Zhulin IB (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol* 13:52–56.
30. Kuo CH, Ochman H (2009) The fate of new bacterial genes. *FEMS Microbiol Rev* 33:38–43.
31. Danilevich VN, Stepanshin YG, Volozhanstev NV, Golub EI (1978) Transposon-mediated insertion of R factor into bacterial chromosome. *Mol Gen Genet* 161:337–339.
32. Salyers AA, Shoemaker NB, Stevens AM, Li LY (1995) Conjugative transposons: An unusual and diverse set of integrated gene transfer elements. *Microbiol Rev* 59:579–590.
33. Rocha EP (2008) Evolutionary patterns in prokaryotic genomes. *Curr Opin Microbiol* 11:454–460.
34. Kreimer A, Borenstein E, Gophna U, Ruppin E (2008) The evolution of modularity in bacterial metabolic networks. *Proc Natl Acad Sci USA* 105:6976–6981.
35. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806.
36. Raes J, Korbel JO, Lercher MJ, Von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* 8:R10.
37. Tringe SG, et al. (2005) Comparative metagenomics of microbial communities. *Science* 308:554–557.
38. Alm EJ, et al. (2005) The MicrobesOnline web site for comparative genomics. *Genome Res* 15:1015–1022.
39. Cole JR, et al. (2009) The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 37:D141–D145.