Gene Duplications Are At Least 50 Times Less Frequent than Gene Transfers in Prokaryotic Genomes

Fernando D. K. Tria * and William F. Martin

Department of Biology, Institute for Molecular Evolution, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

*Corresponding author: E-mail: tria@hhu.de.

Accepted: 22 September 2021

Abstract

The contribution of gene duplications to the evolution of eukaryotic genomes is well studied. By contrast, studies of gene duplications in prokaryotes are scarce and generally limited to a handful of genes or careful analysis of a few prokaryotic lineages. Systematic broad-scale studies of prokaryotic genomes that sample available data are lacking, leaving gaps in our understanding of the contribution of gene duplications as a source of genetic novelty in the prokaryotic world. Here, we report conservative and robust estimates for the frequency of recent gene duplications within prokaryotic genomes relative to recent lateral gene transfer (LGT), as mechanisms to generate multiple copies of related sequences in the same genome. We obtain our estimates by focusing on evolutionarily recent events among 5,655 prokaryotic genomes, thereby avoiding vagaries of deep phylogenetic inference and confounding effects of ancient events and differential loss. We find that recent, genome-specific gene duplications are at least 50 times less frequent and probably 100 times less frequent than recent, genome-specific, gene acquisitions via LGT. The frequency of gene duplications varies across lineages and functional categories. The findings improve our understanding of genome evolution in prokaryotes and have far-reaching implications for evolutionary models that entail LGT to gene duplications ratio as a parameter.

Key words: gene duplication, lateral gene transfer, frequency of events, prokaryote evolution.

Significance

Life is organized as cells, which come in two varieties: simple unicellular microbes, that are only visible under microscopes (prokaryotes), and larger cells with a nucleus that often organize into multicellular forms visible to the naked eye (eukaryotes). Evolutionary mechanisms that generate and maintain this grade of complexity that separates prokaryotes from eukaryotes are of interest. It has long been known that the main mechanism employed by eukaryotes to increase complexity is gene duplication: one gene diverges into two copies in the genome that can undergo independent evolution and thereby foster the origin of novel form and function. The role that gene duplications play in prokaryotes is less well understood, mainly because it is difficult to distinguish gene duplications from gene transfer events, which are the norm in prokaryotic genome evolution. We embarked upon a survey of thousands of prokaryotic genomes to determine the relative frequency of gene duplications and gene transfer in prokaryotic evolution. We found that gene duplications in prokaryotes are rare. By focusing on recent events where duplication and transfer can be unequivocally distinguished, we show that gene duplications are at least 50 times less frequent than lateral gene transfers in prokaryotic genomes.

Introduction

Prokaryotes and eukaryotes differ in their mode and mechanisms of genome evolution. In eukaryotes, gene duplication is a major factor generating multiple copies of genes per genome (Ohno 1970) with globins presenting the first well-studied example for the role of gene duplication in evolution (Zuckerkandl and Pauling 1962; Goodman 1981). In addition, eukaryotes also commonly undergo whole-genome

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

duplications, generating a duplication for every gene, followed by random processes of differential loss (Wolfe and Shields 1997). In prokaryotes, gene duplications are generally thought to be rare (Lerat et al. 2005) and whole-genome duplications of the type found in eukaryotes have so far not been reported, even though some prokaryotes can have very high numbers of genomes per cell (Soppa 2017). At the same time, lateral gene transfer (LGT) is very common in prokaryotes (Ochman et al. 2000), while the role of LGT in eukaryotic evolution is debated (Robinson et al. 2013; Arakawa 2016; Nagies et al. 2020; Tria et al. 2021). Conversely, the role of gene and genome duplications in eukaryotes is not debated (Li et al 2003; Lynch and Conery 2003; Tria et al. 2021), while the role of gene duplications in the evolution of prokaryotes is hard to quantify because of the confounding effects imposed by LGT. Acquisition of homologous gene copies by LGT can increase the number of related genes in a genome, making it difficult to compare the relative contribution of gene duplications and LGT (Coissac et al. 1997; Gevers et al. 2004).

Previous studies of prokaryote genomes that have aimed to distinguish gene duplications from LGT did so based on predefined sequence similarity cutoffs (Snel et al. 2002; Bratlie et al. 2010), sometimes coupled with positional information of genes within genomes (Treangen and Rocha 2011; Wang and Chen 2018) as a measure bearing on the relative likelihood of gene duplications versus LGT, rendering inferences of duplication frequency heavily contingent upon those parameters. In a well-cited study on the topic (Treangen and Rocha 2011), 110 prokaryotic genomes from eight distinct lineages were investigated for their relative frequency of duplications and LGT. Using sequence similarity and positional information of genes to disentangle gene duplications from LGT, they estimated that, depending upon the lineage, 80–98% of prokaryotic multicopy genes in their sample resulted from LGT rather than gene duplications. Those results translate to a minimum ratio of 4:1 and to a maximum ratio of 49:1 for the relative frequency of LGT to gene duplication in generating duplicate copies of a gene in a given prokaryotic genome. Since the study of Treangen and Rocha, thousands of prokaryotic genome sequences have become available, warranting a reinvestigation of the issue on the basis of a broader sample.

More recently, tree reconciliation approaches have emerged that simultaneously model the process of gene duplications, LGT, and gene loss along a reference species tree (Doyon et al. 2011). However, tree reconciliation approaches can hardly be used to infer the duplication to transfer ratio, because the method is dependent upon the input of prior rates of gene duplication, LGT, and gene loss, which are usually assumed to be equal at the outset of the calculations. Furthermore, reconciliation methods are strongly dependent on the fine topological details of phylogenetic trees, making it necessary to distinguish tree incongruences that are the result of methodological errors from tree incongruences that are founded in evolutionary processes. The distinction is challenging; however, especially for large data sets because topologies involving short internal branches in phylogenetic trees are generally hard to infer. Published studies using tree reconciliation models have returned results suggesting that rates of LGT and duplication in prokaryotes are roughly equal (Szöllősi et al. 2015), despite independent evidence to the contrary (Lerat et al. 2005; Treangen and Rocha 2011). For example, the average ratio of LGT per gene duplication ranged between 1 and 2 for Cyanobacteria (Szöllősi et al. 2015) and a ratio of about 0.3, that is, a higher rate of duplications than LGT was obtained for the Thaumarchaeota lineage of archaea (Sheridan et al. 2020). Reconciliation models are becoming more widely used, also as a means to simultaneously estimate the phylogeny upon which the duplications, losses, and LGT are inferred (Szöllosi et al. 2012; Williams et al. 2017; Coleman et al. 2021). At the same time, the reporting of assumptions about LGT to gene duplications ratios underlying such studies, as well insights about the impact of parameters choices upon the reconciliation results, are becoming increasingly opaque. Broadly based, robust estimates of the relative rates of duplications and LGT in prokaryotic genome evolution are thus needed, both to quantify the relative contribution of the processes in nature and to provide benchmarks for reconciliation methods.

Here, we quantify the relative contribution of gene duplication and LGT in prokaryotes using an approach that is independent of deep branching patterns in large phylogenetic trees—we focus our attention exclusively on evolutionarily recent gene duplications and LGT. By identifying recent, phylogenetically unambiguous events, we obtained direct and robust estimates for the relative frequency of gene duplications and LGT in prokaryotic genomes. We calculated the ratio of LGT to gene duplications termed transfer to duplication ratio, or *t/d*, across genes, functional categories, and lineages.

Results and Discussion

Inference of Recent Duplications, Recent Transfers and their Ratios for Prokaryotic Genes

To infer recent gene duplications in 5,655 prokaryotic genomes, with representatives from archaea and bacteria (supplementary table 1), we used clusters from ∼19 million protein-coding genes generated with the Markov clustering (MCL) algorithm. We obtained a total of 450,283 gene families (hereafter simply called families), out of which 260,972 were distributed across the genomes of at least two prokaryotic phyla. The 260,972 families found in at least two phyla were used to generate sequence alignments and reconstruct maximum-likelihood gene trees. Each gene tree (one per family) was then subjected to inferences of recent, genome-specific, gene duplications (fig. 1*b*). For each gene

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

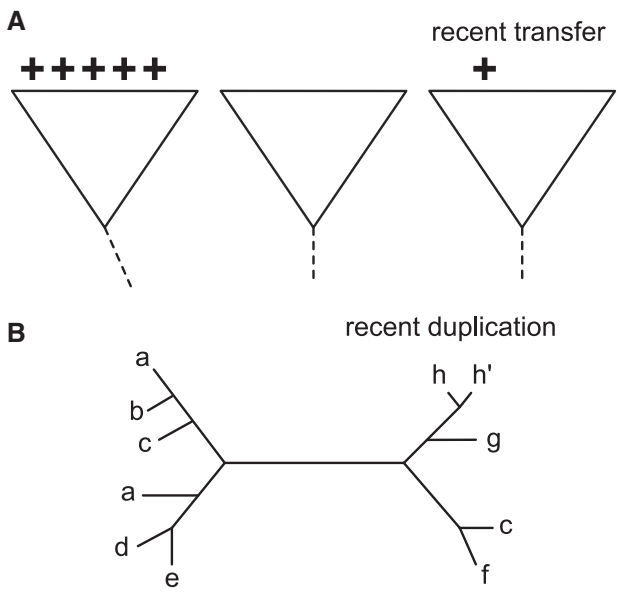

Fig. 1.—Schematic representation of the approaches used for inferences of recent gene transfers and recent gene duplications. (*a*) Recent gene transfers were inferred using the presence-absence distribution of genes (plus symbol) across prokaryotic genomes and the assignment of the genomes to taxa (triangles). A gene in a genome was considered to be the result of a gene transfer if no homologue was present in any other member from the same taxon. Genome-taxon assignments were performed using traditional prokaryotic classifications at different taxonomic levels: domain, phylum, class, order, family, genus, and species. (*b*) Recent gene duplications were inferred on the basis of gene trees and were identified as pairs of genes from the same genome (paralogs) that branch as sisters in the unrooted tree (*h and h′* leaves). Genes from the same genome that do not branch as sisters (for instance the *a* leaves) were not scored since they may be the result of either ancient gene duplication followed by differential gene loss or ancient gene transfer.

duplication, we scored the corresponding gene and genome where the gene duplication occurred. Across the 260,972 gene trees, our criteria identified 16,687 trees (6%) bearing recent duplications. That is, in a given tree, regardless of how many leaves (mean = 68.8, median = 10, SD = 309.7), at least one gene clade appeared in the tree in which two copies of the gene exist in the same genome and branch as sisters in the tree. This estimate is in contrast to eukaryotes, where 45% of all gene trees uncover recent, within genome duplications by the same criterion (Nagies et al. 2020; Tria et al. 2021). The 6% value for prokaryotes is a first indication of the relative paucity of recent gene duplication in prokaryotic genomes.

In our present sample involving thousands of sequenced genomes, the presence of (at least) two copies of a gene sequence in a given prokaryotic genome could result from a duplication event that occurred more recently than any speciation events for the corresponding lineage, or it could be the result of ancient duplication and extensive differential loss. We focused our attention on evolutionarily recent, genome-specific gene duplications because recent paralogs are easy to detect in gene trees (fig. 1*b*). In a large species sample, if a gene's nearest sister is encoded within the same genome it is likely a recent paralog, the result of a gene duplication within the lineage subsequent to any speciation that could be detected in the species sample. An alternative mechanism in generating gene sister pairs within a given genome is duplicative LGT, that is, independent acquisition of the same gene from the same donor twice. However, there is no a priori reason that such recursive targeted transfers involving the same donor–recipient pair should be frequent in prokaryotes, and we know of no reported cases that make such claims.

The decisive parameter for scoring a recent duplication in our present study is whether the closest phylogenetic sister of a gene resides within the same genome. This criterion benefits from the well-known circumstance that phylogenetic inference works better at the tips of trees than it does at their base. Hence, focusing on recent events mitigates false inferences that could arise from methodological error. We are fully aware, as will be the attentive reader, that sparse taxon sampling can generate trees in which a sequence pair residing in the same genome on a terminal branch could represent a very ancient duplication in the guise of a recent duplication. We will systematically investigate this effect of taxon sampling on duplication inferences in the present genome sample, whereby it will become evident that the more densely the taxon harboring the sister pair is sampled, the more reliable its inference as a recent duplication becomes.

Whereas recent duplications generate a sequence pair with a genome, recent LGT events generate a singleton within a taxonomic group. That is, genes present only in one genome of a given prokaryotic taxon (phylum for example), but present in any number of genomes of other taxa (fig. 1*a*), were scored as recent LGT. To quantify prokaryotic gene duplications relative to LGT, we inferred recent LGT across the same set of genes and genomes. To distinguish LGT across different taxonomic levels, we performed genome-taxon assignments for all genomes, from domain to species (supplementary table 1), and repeated the LGT inference procedure for each taxon level. This delivered estimates for the frequency of LGT across different taxonomic boundaries, recalling that the LGT can occur in at most one genome of the recipient taxon (fig. 1*a*), whether it be a sister species, order, or phylum.

The most common type of LGT was interspecies LGT, found in 227,974 genes (87%) whereas the least frequent was interdomain LGT, occurring in 5,338 genes (2%). Of course, an unknown number of inferred LGTs could be the result of differential loss, a possibility that becomes increasingly less likely with taxonomic distance, for example, interdomain LGT. We then used the total number of genes reflecting recent LGT relative to the total number of duplicated genes to derive an estimate for the ratio of LGT per gene duplication, the transfer to duplication ratio (*t*/*d*). As a first estimate, neglecting the effects of lineage sampling, we

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

Table 1.

Number of prokaryotic genes with recent gene duplication and recent LGT crossing different taxonomic boundaries (taxon level). The inferences were performed using all genomes (no filter) with taxonomic classifications available. To counter biases stemming from sparsely sample taxa, the analyses were repeated considering only genomes from taxa with ≥ 2 genomes and taxa with ≥ 6 genomes. Note that the genome set is variable at different taxonomic levels and only inferences of gene transfers are dependent upon taxonomic classifications. However, gene duplication inferences were performed on the same genome sets for comparisons. The number of genomes, number of taxa and the total number of genes distributed in the genome set are indicated.

Taxon Level	Transferred Genes (*t*)	Duplicated Genes (*d*)	*t*/*d*	No. of Genomes	No. of Taxa	No. of Genes
No filter						
Domain	5,338 (2.0%)	16,687 (6.4%)	0.32	5,655	2	260,972
Phylum	54,457 (20.9%)	16,643 (6.4%)	3.27	5,652	34	260,972
Class	78,813 (30.8%)	15,184 (5.9%)	5.19	5,543	65	255,886
Order	111,173 (42.9%)	16,369 (6.3%)	6.79	5,584	149	259,218
Family	134,535 (51.6%)	16,012 (6.1%)	8.40	5,567	310	260,738
Genus	165,738 (63.5%)	16,091 (6.2%)	10.30	5,608	871	260,972
Species	227,974 (87.4%)	16,687 (6.4%)	13.66	5,655	2,370	260,972
≥ 2 genomes						
Domain	5,338 (2.0%)	16,687 (6.4%)	0.32	5,655	2	260,972
Phylum	50,383 (19.3%)	16,563 (6.3%)	3.04	5,646	28	260,972
Class	73,108 (28.6%)	14,804 (5.8%)	4.94	5,530	52	255,847
Order	101,692 (39.2%)	15,401 (5.9%)	6.60	5,555	120	259,194
Family	115,729 (44.4%)	13,691 (5.3%)	8.45	5,479	222	260,692
Genus	100,264 (38.8%)	8,246 (3.2%)	12.16	5,118	381	258,383
Species	50,696 (22.5%)	741 (0.3%)	68.42	3,765	480	224,990
≥6 genomes						
Domain	5,338 (2.0%)	16,687 (6.4%)	0.32	5,655	2	260,972
Phylum	46,472 (17.8%)	16,393 (6.3%)	2.83	5,620	20	260,455
Class	65,051 (25.6%)	14,267 (5.6%)	4.56	5,482	36	254,103
Order	89,309 (34.8%)	13,795 (5.4%)	6.47	5,433	81	256,763
Family	86,784 (34.3%)	9,965 (3.9%)	8.71	5,203	128	253,356
Genus	53,245 (23.7%)	4,063 (1.8%)	13.10	4,417	138	224,678
Species	15,129 (10.5%)	146 (0.1%)	103.62	2,821	140	143,745

found that the *t*/*d* ratio across all genes ranged from 0.3 to 14 depending on the taxonomic level of the taxon considered, whereby the lower the taxonomic level, the more scare duplications became relative to LGTs (top rows in table 1). The first impression of *t*/*d* ratio of 14 for the species level reflects the raw numbers for the entire data, with no consideration of lineage sampling effects, an important factor that we investigated in greater detail below.

Our approach was specifically designed to detect gene duplications and LGTs that are exclusive to a single genome. Note that our scoring criteria identify both duplications and LGT that give rise to a new gene in a given genome *before any speciation events have taken place*: single occurrence in the taxon for LGT and within-genome sisters for duplications. Therefore, for a given genome, the window of time within which an LGT or duplication can be observed is the same for both processes; hence, the frequencies observed are directly comparable, independent of whether or not the taxa assignments are natural or whether a given taxonomic rank such as "genus" is equivalent across different orders or phyla.

Independent of taxonomic level however, false inferences of recent LGT and gene duplications could arise due to sparsely sampled prokaryotic lineages since gene duplications followed by gene loss may mimic patterns of LGT, on the one hand, and apparent gene duplications could (though unlikely) result from duplicative LGT from closely related donors, on the other. To account for these possibilities, we repeated our analyses twice using the following quality-filters that improve the density of taxon sampling: 1) ignoring events (LGT and duplications) scored in genomes from taxa with only one member (lower rows, table 1) and 2) ignoring events in genomes from taxa with less than six members (lower rows, table 1). The effect of merely requiring two genomes from the given taxon to be present in the tree was substantial, increasing the value of the *t*/*d* ratio to 68. Adding further lineage sampling stringency, that is requiring at least six genomes to be present in the lineage used to infer recent LGT or genome-specific duplication, the value of the *t*/*d* ratio increased to over 100 (lower rows, table 1).

The values shown in table 1 concern the total number of genes subject to recent gene duplications and LGT, without consideration of the number of events these genes experienced during their evolution. Gene duplications and LGT can occur for the same gene multiple times during its history if independent events occurred in different genomes. To compare gene duplications and LGT in terms reoccurrence of

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

events, we counted the number of events across all genes for which at least one duplication and/or LGT was called and compared the distributions for duplications against LGT. We found that LGT not only affects a greater number of genes in comparison to gene duplication (table 1) but also that LGT is more recurrent than gene duplications in the genes where they occur, for most types of inter-taxa LGT (supplementary fig. 4, Supplementary Material online). In other words, prokaryotic genes have a higher tendency to be transferred in parallel by different donor–recipient genome pairs than their tendency to undergo independent duplications in different genomes.

Our results consistently demonstrate that LGT is far more frequent than gene duplication in prokaryotes. Taking the effect of lineage sampling into account, across all prokaryotic genes and lineages in our data, for each gene that undergoes a recent duplication there are 100 genes that are have undergone recent LGT. The ratio *t/d* ranged from 14 to 104 for interspecies LGT. The *t/d* was also high for LGT among higher taxa but smaller than the ratio obtained for interspecies LGT. Recent gene duplications were only more frequent than LGT in interdomain comparisons (table 1). Overall our estimates are consistent with high rates of LGT in prokaryotes (Ochman et al. 2000; Dagan et al. 2008), but markedly contrast with earlier reports of high rates of gene duplications relative to LGT based on tree reconciliation approaches (Szöllősi et al. 2015; Sheridan et al. 2020).

Recent Gene Transfers Vastly Outnumber Recent Gene Duplications in Prokaryotic Genomes and Lineages

The propensity for gene duplications and LGT may vary among prokaryotic lineages because of differences in population dynamics (selection regimes, population size, and evolutionary rates) and genetic makeup that permit, or preclude, duplication and/or uptake of external DNA. To disentangle potential lineage-specific trends underlying recent gene duplications and recent LGT, we first estimated the frequency of gene duplication and LGT for all genomes individually (supplementary table 2). For each genome, we quantified the fraction of duplicated genes and, similarly, the fraction of acquired genes via LGT. Filtering out taxa with only a single genome we found that, on average, 0.04–2.8% of the genes in a prokaryotic genome were inherited via recent LGT, with more frequent LGT among closely related lineages. For the same genome sets, the fraction of duplicated genes was on average 0.04%–0.18%. We then used the mean fraction of duplications and LGT across genomes to derive a *t/d* ratio, which ranged from 30–170 for interspecies LGT (fig. 2), 15–60 for inter-genus LGT, and smaller *t/d* ratios for LGT among higher taxa. The across-genome *t/d* ratio shows that interspecies LGT are, again, roughly two orders of magnitude more frequent than recent gene duplications.

Our analyses show that filtering out sparsely sampled taxa has some effect on the calculations of *t/d* values across genomes. Nevertheless, gene duplications are consistently less frequent than most types of LGT regardless of the taxon filter stringency applied (fig. 2). An additional and complementary filter concerns gene family size on which gene duplications and LGT were inferred. One possibility is that small gene families represent genes distributed at low frequency across the genomes. However, small gene families could also, in theory, be the result of incorrect gene predictions and artifactual sequence clustering, potentially biasing the *t/d* estimates across genomes presented here. To take that theoretical possibility into account, we repeated the analyses from figure 2 and calculated the frequency of gene duplications and LGT for each genome but discarding events inferred from small gene families that are distributed in less than ten genomes. The results show that the high *t/d* ratios across genomes are very robust and remain practically unchanged after filtering small gene families from our analyses (supplementary figs. 1–3, Supplementary Material online). Genes present in low frequency across genomes, that is small gene families, exert no influence on the estimates of LGT, gene duplications, and their ratios across genomes. All things being equal, small gene families are most likely either the result of recent gene origin or the result of high substitution rate within large gene families, or both. Such small families can also undergo gene flux across prokaryotic genomes or gene duplications (Nagies et al. 2020). That we see no effect of gene family size on *t/d* indicates that neither gene age (the time of gene family origin) nor substitution rate skew *t/d* ratios to higher or lower values. This generally uniform behavior also suggests that estimates of *t/d* based on recent events tend to be more or less constant across genes (yet see the case of transposons below).

To investigate variation in the frequency of LGT (*t*), duplications (*d*), and the *t/d* ratio for different prokaryotic lineages, we summarized the across-genome estimates for prokaryotic taxa. Gene duplications are slightly more frequent in archaea than in bacteria. The average fraction of recently duplicated genes was 0.06% for archaeal genomes, and 0.04% for bacterial genomes (supplementary table 3). The frequencies of LGT were more similar between archaea and bacteria than were gene duplication frequencies, for all intertaxa estimates of LGT except inter-domain LGT. For interspecies LGT, our estimates show that on average 1.8% of the genes in bacterial genomes were recent acquisitions, while the estimate was virtually the same for archaeal genomes (2%). The average fraction of gene acquisitions in bacteria and archaea is roughly the same for LGT across most of the taxonomic levels, but a clear difference emerges for inter-domain LGT which explains the origin of 1% of the genes in archaeal genomes while explaining the origin of only 0.01% of the genes in bacterial genomes. This observation can be readily explained since LGT between archaea and bacteria is highly asymmetric (Nelson-

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

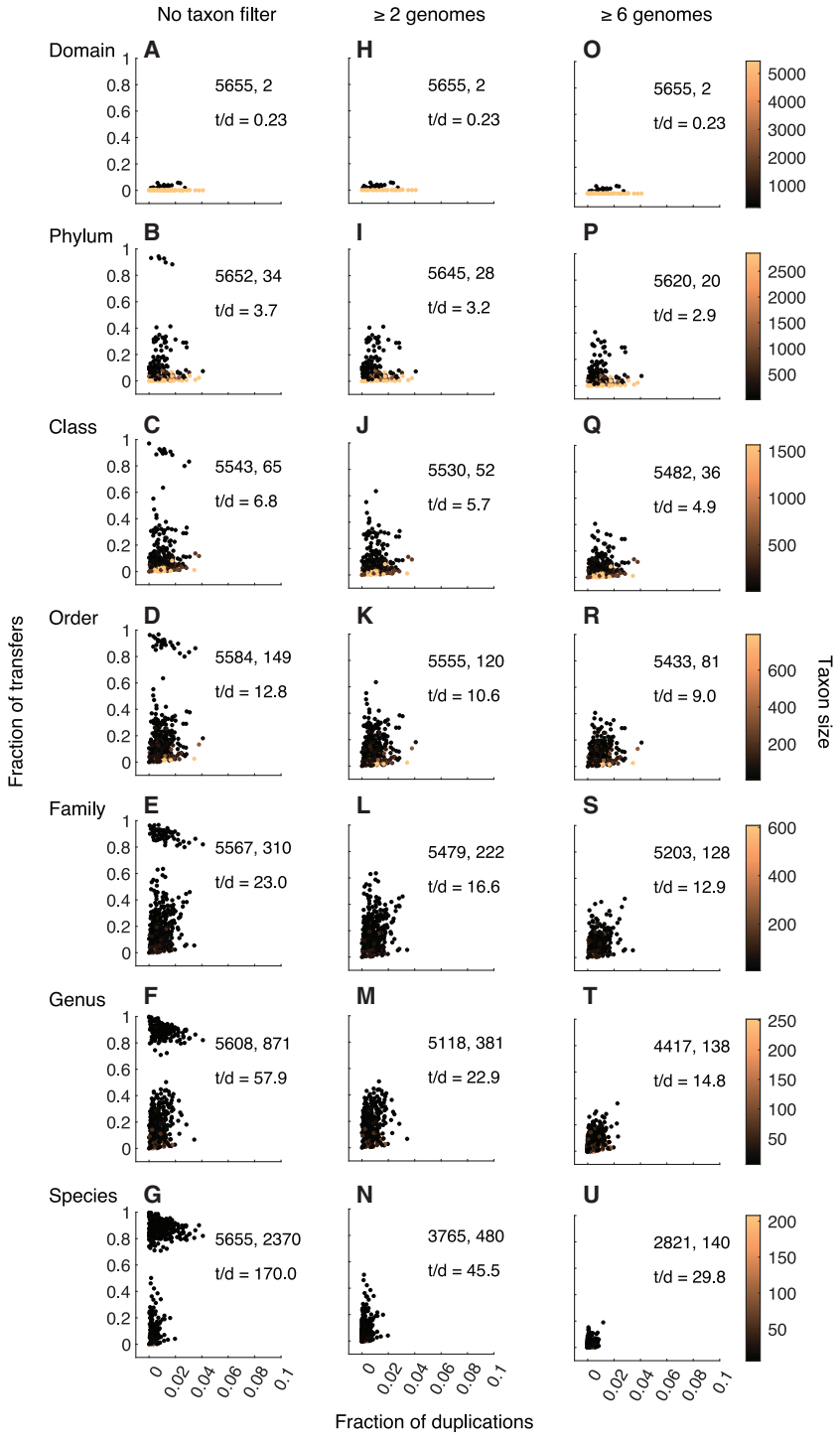

Fig. 2.—Quantification of recent gene transfers and recent gene duplications across 5,655 prokaryotic genomes. For each prokaryotic genome the number of gene duplications (horizontal axis) and gene transfers (vertical axis) are reported as fractions relative to the number of non-singleton genes (see Materials and Methods and fig. 1 for details on inferences). Recent gene transfers across different taxonomic ranges were distinguished: interdomain transfers (*a, h, o*), inter-phylum transfers (*b, i, p*), inter-class transfers (*c, j, q*), inter-order transfers (*d, k, r*), inter-family transfers (*e, l, s*), inter-genus transfers (*f, m, t*), and inter-species transfers (*g, n, u*). In (*a–g*), all genomes with taxonomic classifications were used. In (*h–n*), genomes belonging to taxa with less than one representative genome were discarded. In (*o–u*), genomes from taxa with less than five representative genomes were discarded. Inset upper numbers show the total number of genomes and taxa, respectively. *t/d* indicates the ratio of the mean fraction of transfers over the mean fraction of duplications. The color scale shows the number representative genomes affiliated to the same taxon (taxon size). See also supplementary figures 5–7, Supplementary Material online for the distribution plots in log scale.

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

Table 2.

Summary statistics for recent inter-species gene transfers and recent gene duplication in distinct bacterial (bottom) and archaeal (top) taxa with at least two representative genomes. The mean was taken across genomes for taxon and SD denotes the standard deviation. *t/d* is the ratio of mean fraction of LGT relative to the mean fraction gene duplications obtained for each taxon and a dash ('—') indicates lineages for which the ratio was not possible to estimate due to absence of detectable gene duplication in the genomes.

Domain	Class	Fraction of Transfers	Fraction of Duplications	No. of Nonsingleton Genes	No. of Genomes	
Archaea		Mean (SD)	Mean (SD)	Mean (SD)	Total	*t/d*
	Thermoprotei	7.19E−03 (1.17E−02)	3.91E−04 (5.91E−04)	2,480.9 (230.0)	25	18.4
	Archaeoglobi	4.31E−02 (1.05E−02)	1.07E−03 (2.77E−04)	2,334.0 (56.6)	2	40.4
	Halobacteria	2.42E−02 (2.99E−02)	3.93E−04 (8.37E−04)	2,687.0 (649.2)	8	61.7
	Methanobacteria	2.27E−02 (1.45E−02)	1.47E−04 (2.55E−04)	2,276.0 (13.5)	3	154.4
	Methanococci	2.21E−02 (6.49E−03)	1.03E−03 (1.18E−03)	1,736.6 (19.2)	5	21.3
	Methanomicrobia	2.57E−02 (2.96E−02)	9.84E−04 (1.85E−03)	3,440.0 (236.9)	14	26.1
	Thermococci	5.95E−02 (6.10E−02)	1.06E−03 (1.83E−03)	2,107.5 (194.1)	4	55.9
Bacteria	Actinobacteria	1.93E−02 (4.38E−02)	2.80E−04 (8.24E−04)	3,300.7 (1,761.1)	337	68.9
	Aquificae	5.36E−04 (7.58E−04)	0.00E+00 (0.00E+00)	1,864.5 (2.1)	2	
	Bacteroidia	2.43E−02 (2.05E−02)	2.72E−04 (3.76E−04)	2,886.2 (1.100.1)	20	89.3
	Flavobacteriia	2.48E−02 (4.21E−02)	7.26E−04 (2.05E−03)	2,683.1 (730.9)	37	34.1
	Chlamydiia	4.44E−04 (1.59E−03)	0.00E+00 (0.00E+00)	926.6 (45.4)	105	—
	Cyanobacteria	2.29E−02 (3.69E−02)	6.65E−04 (1.79E−03)	2,591.7 (810.6)	23	34.4
	Chlorobia	2.63E−01 (2.02E−02)	4.21E−03 (1.11E−03)	2,269.5 (96.9)	2	62.5
	Dehalococcoidia	1.25E−02 (7.07E−03)	2.19E−04 (4.50E−04)	1,402.7 (50.8)	13	57.3
	Deinococci	3.13E−02 (7.48E−03)	7.89E−04 (3.87E−04)	2,471.8 (416.6)	6	39.6
	Fibrobacteria	4.10E−03 (1.16E−03)	0.00E+00 (0.00E+00)	3,046.0 (2.8)	2	—
	Bacilli	1.57E−02 (2.74E−02)	3.61E−04 (9.00E−04)	2,998.2 (1.263.9)	833	43.6
	Clostridia	3.71E−02 (6.93E−02)	8.78E−04 (2.40E−03)	3,342.0 (767.1)	72	42.2
	Erysipelotrichia	1.70E−02 (1.75E−02)	0.00E+00 (0.00E+00)	1,596.0 (142.1)	4	—
	Fusobacteriia	2.90E−02 (1.81E−02)	8.38E−04 (6.87E−04)	2,116.6 (140.8)	13	34.6
	Nitrospira	6.74E−02 (6.18E−03)	9.24E−04 (6.60E−04)	2,172.0 (21.2)	2	73.0
	Acidithiobacillia	6.93E−02 (2.64E−02)	2.15E−03 (1.89E−03)	2,662.5 (53.9)	4	32.2
	Alphaproteobacteria	2.69E−02 (4.73E−02)	7.25E−04 (1.70E−03)	3,230.8 (1.907.9)	249	37.1
	Betaproteobacteria	2.15E−02 (4.78E−02)	3.96E−04 (9.68E−04)	4,262.0 (1.729.5)	353	54.4
	Deltaproteobacteria	1.00E−01 (1.28E−01)	1.34E−03 (1.79E−03)	3,989.6 (2.133.0)	20	74.9
	Epsilonproteobacteria	7.50E−03 (2.16E−02)	3.27E−04 (6.79E−04)	1,574.2 (144.3)	221	23.0
	Gammaproteobacteria	1.58E−02 (3.20E−02)	3.05E−04 (8.66E−04)	4,103.4 (1.249.2)	1,229	51.9
	Spirochaetia	2.35E−02 (3.12E−02)	5.03E−04 (8.64E−04)	2,075.6 (1.190.8)	39	46.8
	Mollicutes	1.08E−02 (2.17E−02)	7.01E−04 (2.48E−03)	714.6 (148.5)	102	15.3
	Thermotogae	1.67E−02 (3.03E−02)	1.57E−04 (2.54E−04)	1,897.6 (88.8)	10	106.5
	Verrucomicrobiae	7.94E−02 (6.90E−03)	6.59E−04 (9.32E−04)	1,523.5 (7.8)	2	120.6

Sathi et al. 2015; Méheust et al. 2018), the recipient genome is most often an archaeal lineage, and several bacteria-to-archaea LGT coincided with the origin of ancestral archaeal lineages. Our results show that bacteria-to-archaea LGT is still ongoing. The across-genome *t/d* ratios were 45 and 33 in bacteria and archaea, respectively, for interspecies LGT. For inter-domain LGT, the *t/d* ratios are 17 in archaea and 0.25 in bacteria.

For lineages at lower taxonomic levels, the results for inter-species LGT and gene duplications are summarized in table 2. Ranking the lineages by the mean fraction of duplicated genes rendered Achaeoglobi, Themococci, and Methanococci as the archaeal lineages with the highest incidence of gene duplications. The archaeal lineages with the lowest levels of duplications were Methanobacteria and Thermoprotei. For bacteria, the lineages with the highest frequencies of duplications were Chlorobi, Acidithiobacilla, and Deltaproteobacteria. The bacterial lineages with the lowest duplication frequencies were Aquificae, Chlamydia, Fibrobacteria, and Erysipelotrichia all of which showed no evidence of recent gene duplications in their genomes. The Chlamydia and Erysipelotrichia lineages are characterized by small genomes due to massive gene loss (Davis et al. 2013; Nunes and Gomes 2014).

Ranking lineages by the mean fraction of LGT instead results in Thermococci and Archaeoglobi as the top-ranking archaeal lineages, though the distribution is very narrow with the differences among archaeal lineages being very small. Thermoprotei and Methanococci were the archaeal lineages with the lowest fraction of LGT. For bacterial lineages, the distribution of LGT is clearly more variable than in archaea, with Deltaproteobacteria and Chlorobi genomes showing the highest fraction of recent gene acquisitions, 10% on average.

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

The bacterial lineages with the lowest LGT frequencies were Aquificae, Chlamydia, and Fibrobacteria.

Overall, our results show a correspondence between LGT and gene duplications because lineages with more recent LGT are often the lineages with more recent gene duplications, despite some exceptions (for instance, Methanococci). The average *t/d* ratio was 54 across all prokaryotic lineages, also for bacterial and archaeal lineages estimated separately. The *t/d* ranged between 18–154 in archaea and 15–121 in bacteria. We investigated whether the *t/d* ratio across lineages could be in part affected by the number of sampled genomes. The number of genomes alone cannot explain the distribution of *t/d* across lineages since the correlation of *t/d* with the number of genomes was small and non-significant ($P = 0.12$ and rho $= -0.3$, two-tailed Spearman correlation). Accordingly, Gammaproteobacteria, the best-sampled lineage with 1,229 representative genomes, attained a high *t/d* ratio of 52. Our results are in stark contrast with previous reports based on reconciliation models which estimated a *t/d* ratio of 1–2 in Cyanobacteria, depending on the cyanobacteria phylogeny used (Szöllősi et al. 2015). Here, we found the *t/d* ratio of 23 for Cyanobacteria. One could argue that the differences reported here differ from previous reports because we only deal with recent events. However, *t/d* ratios for terminal branches in Szöllősi et al. (2015) are yet smaller. Our results indicate that reconciliation models need to be used with caution and previous phylogenetic reconstructions obtained from reconciliation methods, that assume equal prior rates of LGT and duplications, may need reassessment [see for instance: (Coleman et al. 2021)].

The Contribution of Recent Gene Duplications and Recent LGT to Genome Size Expansion

The number of genes encoded in prokaryotic genomes is highly variable with small genomes harboring as few as 900 protein-coding genes as in the case of the intracellular pathogen *Chlamydia trachomatis* and over 10,000 protein-coding genes as in the multicellular-like *Kibdelosporangium phytohabitans* species. Genome size variation is not solely observed among distantly related lineages but also within species boundaries. For instance, a given *Escherichia coli* strain may have between 3,700–5,600 protein-coding genes in its genome.

Genome size variation results from the balance of gene gain and gene loss, which differ across lineages. Several factors are known to govern gain and loss rates across lineages including the ability to uptake external DNA via transformation, conjugation, and transduction (Vos et al. 2015); and population parameters such as selective regime and population size (McInerney et al. 2017). There is a theoretical upper bound for genome growth because of the energetic costs involved in chromosome maintenance and replication (Lane and Martin 2010). Notwithstanding energetic constraints, an open question is how fast genome expansion can occur. Are short-term effects of LGT and gene duplications relevant at all for genome plasticity?

We addressed this question by correlating the number of protein-coding genes in each genome against the estimates of gene duplications and LGT across genomes. The Spearman correlation analyses show that the fraction of gene duplications explains about 12% of the variation in the number of genes across genomes, whereas interspecies LGT explains 14% of the variation (fig. 3). That leaves a vast majority (~85%) of genome size variation that cannot be explained by recent events of gene gains (via duplication and LGT). One possibility is that the remaining 85% of genome size variation results from gene loss only. That explanation however carries a heavy corollary since a loss-only model would imply too large ancestral genomes (Dagan and Martin 2007). A more likely explanation is that the 85% genome size variation is in fact derived from gene loss and gene gains in ancestral lineages, mainly via LGT rather than duplication as our findings indicate.

Even though LGT may be as much as two orders of magnitude more frequent than gene duplications, as our results show, both processes contribute to a similar extent to genome size expansion. That is because gene duplications and LGT are themselves intercorrelated. One of the mechanisms that generates duplications is intrachromosomal recombination, usually facilitated by conserved DNA repeats in the genome (Reams and Roth 2015). Indeed, we noticed a significant correlation of LGT and gene duplications ($rho = 0.8$, $P < 0.01$, two-tailed Spearman correlation, data from table 2). Our analyses also indicate a nonsignificant association of LGT and genome size for interdomain and interphylum LGT (fig. 3). Only at the level of interclass does LGT have a significant, albeit small, contribution to genome size expansion in prokaryotes. Recent LGT plays an increasing role to genome size expansion for transfers among lower taxa, yielding a maximum contribution for inter-species LGT ($rho = 0.14$, $P < 0.01$, two-tailed Spearman correlation). Our results indicate that both recent gene duplications and recent LGT may generate observable genome size variation. Most of the variation, however, is a consequence of long-term balance of gene gains and losses (Ochman et al. 2000; Dagan and Martin 2007; Vos et al. 2015; Sela et al. 2016; McInerney et al. 2017).

Recent Gene Duplications Are Largely Associated with Transposons

Although it is likely that no prokaryotic gene is immune to LGT, some genes have a higher tendency to be transferred than others because of their functions. For instance, genes offering a clear selective advantage to the cell, like antibiotic resistance and toxin-antitoxin genes, are more likely to spread in nature via LGT (Bennett 2008; Ramisetty and Santhosh

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

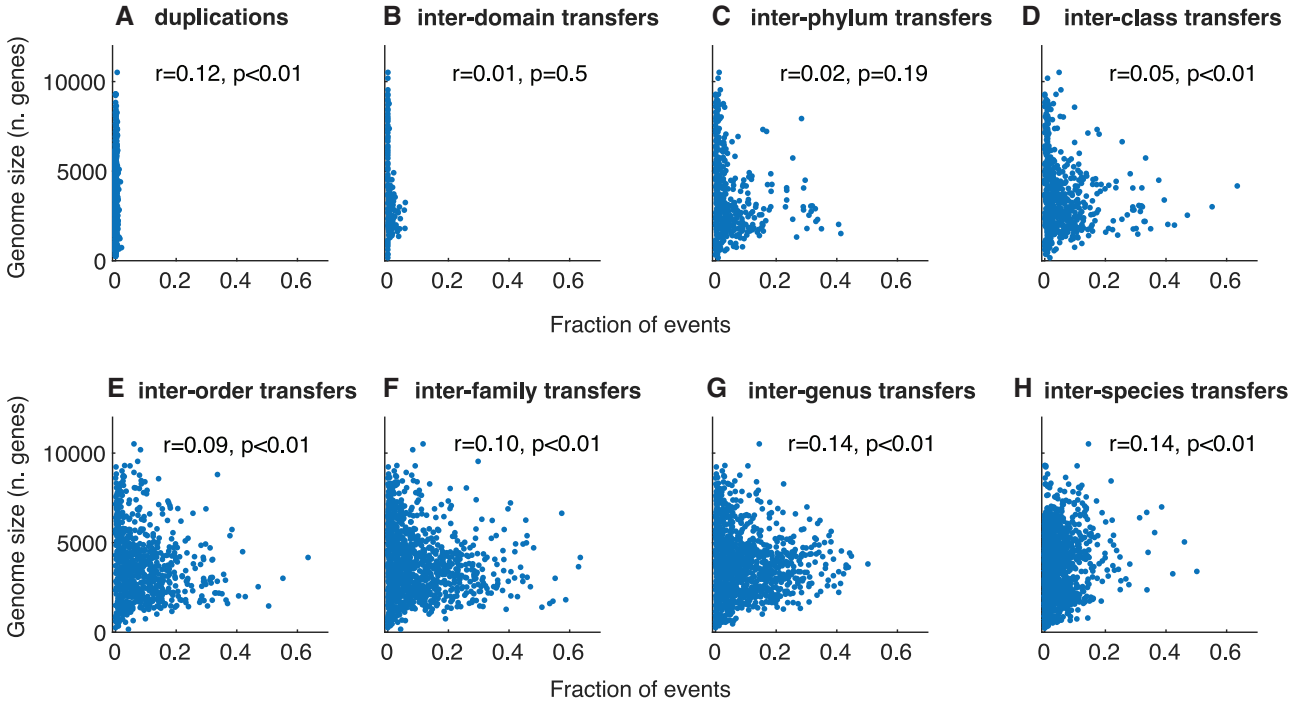

Fig. 3.—Effect of recent gene duplications and recent gene transfers to genome-size expansion in prokaryotes. The plot shows genome size, measured as the number of protein-coding genes (vertical axis), against the fraction (horizontal axis) of recent gene duplications (*a*) and recent gene transfers (*b*–*h*), using genomes affiliated to taxa with more than one representative genome (see panels *h*–*n* in fig. 2 for sample sizes). Insets: *r* denotes the Spearman correlation coefficients, and *p* denotes the FDR adjusted *P*-values from the two-tailed tests (see Materials and Methods).

2016). A more general observation is that widely distributed and conserved genes, playing a central role in the metabolism, are more resistant to LGT. Thus, there is a whole spectrum of transferability, or verticality, across gene functions in prokaryotes (Cohen et al. 2011; Nagies et al. 2020). The tendency for a gene to duplicate depending on its function is somewhat less well characterized. It has been reported that selfish genetic elements like transposons are duplicated more often (Treangen and Rocha 2011) because of their ability to "copy-and-paste" (Curcio and Derbyshire 2003). We therefore asked: Does *t*/*d* vary across gene functions in prokaryotes and if so, how and by how much?

To answer these questions, we annotated the genes in our data according to 26 functional categories from the KEGG database and tabulated the distribution of LGT (*t*), gene duplications (*d*), and the *t*/*d* ratios across the functional categories. We then performed statistical tests to assess whether gene duplications and/or LGT were enriched in some of the 26 functional categories (see Materials and Methods). Out of 260,972 genes in our data, we were able to annotate 89,496 for which homologs were found in the KEGG database. The proportions of genes with duplications and LGT in the subset of genes with annotations are larger than the proportions obtained for all genes, with 2% of the annotated containing duplications and 46% of the genes containing interspecies LGT.

Gene functions are not mutually exclusive and a gene in the KEGG database may be assigned to more than one category so the total number of annotated genes is smaller than the sum of assignments across all functional categories. Despite this, variation in the frequency of gene duplications across functional categories is small with 2% of the genes in a category, on average, having experienced a gene duplication. "Genetic information processing" stands out as the functional category with the highest incidence of gene duplications with 7% of the genes within the category harboring duplications (table 3). Looking at the specific function of the duplicated genes related to "genetic information processing," we found that most of them are in fact transposons or retrotransposons, as from NCBI annotations (supplementary table 4). Other functional categories enriched with gene duplications are: "replication and repair," "carbohydrate metabolism," and "transport and catabolism." Notably, all of the functional categories enriched with gene duplications are also enriched with LGT, indicating that genes undergoing increases in copy-number are subject to the same functional and selective constraints.

The *t*/*d* ratio has a variable distribution across functional categories between 7 and 66. The functional category with the smallest ratio is "genetic information processing." The category with the largest ratio is "signal transduction" because gene duplications are extremely rare in this category

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

Table 3.

Functional distribution of the genes analyzed in this study. All genes show the total number of annotated genes for each functional category. Genes with inter-species LGT (*t*) and genes with duplications (*d*) were scored only for species with at least 2 members. Functional annotations were performed using the KEGG database (see Methods for details). *t/d* denotes the ratio of transfers relative to duplications and FDR denotes the adjusted p-value from the one-tailed binomial test (enrichment test).

KEGG Category (B Level)	All Genes	Transferred Genes (*t*)	Duplicated Genes (*d*)	*t/d*	FDR (*t*)	FDR (*d*)
Genetic information processing	4,836	2,622 (54%)	357 (7%)	7.3	0.000	0.000
Membrane transport	19,982	9,283 (46%)	325 (2%)	28.6	0.509	1.000
Carbohydrate metabolism	4,831	2,435 (50%)	130 (3%)	18.7	0.000	0.001
Replication and repair	3,497	1,702 (49%)	116 (3%)	14.7	0.005	0.000
Transcription	7,244	3,932 (54%)	105 (1%)	37.4	0.000	1.000
Poorly characterized	6,211	2,560 (41%)	94 (2%)	27.2	1.000	1.000
Amino acid metabolism	3,772	2,054 (54%)	85 (2%)	24.2	0.000	0.212
Metabolism	4,257	2,106 (49%)	75 (2%)	28.1	0.000	1.000
Transport and catabolism	2,842	1,605 (56%)	74 (3%)	21.7	0.000	0.026
Cellular community—prokaryotes	3,985	1,771 (44%)	62 (2%)	28.6	1.000	1.000
Cellular processes and signaling	3,900	1,597 (41%)	52 (1%)	30.7	1.000	1.000
Energy metabolism	2,701	1,026 (38%)	48 (2%)	21.4	1.000	1.000
Enzyme families	3,732	1,387 (37%)	44 (1%)	31.5	1.000	1.000
Cell motility	3,619	1,383 (38%)	41 (1%)	33.7	1.000	1.000
Glycan biosynthesis and metabolism	3,348	1,513 (45%)	41 (1%)	36.9	1.000	1.000
Metabolism of cofactors and vitamins	2,440	1,122 (46%)	41 (2%)	27.4	1.000	1.000
Xenobiotics biodegradation and metabolism	1,602	941 (59%)	41 (3%)	23.0	0.000	0.136
Signal transduction	6,709	2,654 (40%)	40 (1%)	66.4	1.000	1.000
Lipid metabolism	2,859	1,402 (49%)	37 (1%)	37.9	0.004	1.000
Nucleotide metabolism	1,417	608 (43%)	26 (2%)	23.4	1.000	1.000
Translation	2,413	915 (38%)	24 (1%)	38.1	1.000	1.000
Metabolism of terpenoids and polyketides	1,472	734 (50%)	24 (2%)	30.6	0.006	1.000
Folding, sorting, and degradation	1,872	662 (35%)	23 (1%)	28.8	1.000	1.000
Drug resistance	1,754	778 (44%)	23 (1%)	33.8	1.000	1.000
Metabolism of other amino acids	744	357 (48%)	21 (3%)	17.0	0.345	0.136
Biosynthesis of other secondary metabolites	506	254 (50%)	8 (2%)	31.8	0.079	1.000
Total	102,545	47,403 (46%)	1,957 (2%)	24.2	—	—

Note.—Significantly enriched, FDR < 0.05.

with only 40 genes having duplications out of 6,709 genes within this category. Despite the variation across gene functions, the *t/d* ratio is on average 29 across functional categories. Hence, the excess of LGT over gene duplications is robust and applies equally to prokaryotic genes of different functions.

Conclusions

Here, we have shown that for a sample consisting of thousands of prokaryotic genomes, the frequency of gene transfer is at least one order of magnitude and in many cases two orders of magnitude higher than the frequency of gene duplication as a mechanism for generating new genes (or additional copies of preexisting genes) within the same genome. A critic might interject that we have not considered the effects of gene loss within the same genome, but loss of a recent duplicate will erase the evidence for its existence as will loss of a recent LGT. In that sense, loss affects both processes equally and its unlikely to bias one estimate over the other for these recent, genome-specific, events. Considering all factors investigated here, it seems that the best (most accurate) estimate for the LGT to duplication ratio, *t/d*, is obtained at the species level because LGT is more common within species boundaries (Popa and Dagan 2011).

At the species level, the effect of LGT in increasing prokaryotic genome size is, perhaps surprisingly, not strong (Gautreau et al. 2020). This is because the average frequency of recent LGT estimated here only concerns genes that occur at most once in the entire pangenome of the recipient taxon, and present at any number of genomes in the pangenomes of other taxa. Genes found to be present only in a single genome were not counted as acquisitions via LGT, because of the lack of evidence for the presence of the gene in a putative donor taxon, and the possibility of *de novo* gene evolution that could also explain the origin of singleton genes. However, it is likely that a considerable number of singletons are in fact the result of recent LGT and these could potentially account for the larger contribution of LGT to genome size expansion in prokaryotes.

The frequency of recent duplications is lowest for the species level, indicating that tip sister relationships for two copies occurring within the same genome are very rare in

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

comparison to LGT (table 1). The average *t/d* ratio across different prokaryotic phyla (Table 2) is 55. Lineage sampling (fig. 2) is much more important than the variation of *t/d* across functional categories (Table 3). Taken together, this indicates that for prokaryotes, LGT introduces genes into a genome at least 50 times more frequently than within genome duplications do, whereby the effect of lineage sampling is substantial, such that the best estimate might be that duplications occur at only 1% the frequency of LGTs in prokaryotes (Table 1). Our range for the relative frequency of LGT to duplications (50–100) includes the highest estimate obtained by Treangen and Rocha (2011) based on a genome sample 50 times smaller than ours. The average *t/d* ratios we obtained for archaea, 54, and bacteria, 55, are almost identical indicating that the natural tendency for duplication is low and the natural tendency to undergo LGT is high in both groups.

Our study provides estimates for the *t/d* ratio that contrast with those obtained from reconciliation methods, which often conclude similar rates of LGT and gene duplications in prokaryotes [see for instance (Szöllősi et al. 2015)]. The reasons for these differences are 1) the estimates obtained from reconciliation studies were derived from species and gene samples orders of magnitude smaller than ours; 2) reconciliation models require input of prior rates for gene duplications and LGT, which in current implementations of reconciliation methods are assumed to be equal; and 3) the conflation gene tree-species tree incongruences arising from random phylogenetic and methodological errors versus the workings of gene duplications and/or LGT. In contrast, our study provides estimates for a very large sample of genes and genomes. By focusing on recent events, where phylogeny is least ambiguous, and avoiding assumptions about duplication and LGT rates a priori, we were able to minimize many false inferences tallied in previous reconciliation studies.

Though prokaryotes can attain very high ploidy levels (Soppa 2017), they do not undergo whole-genome duplications (Wolfe and Shields 1997). We observed that individual gene duplications are also rare in prokaryotic genomes, likely due to "gene dosage" effects whereby the benefits in harboring multiple copies of the same gene (higher gene expression, for instance) is offset by extra energetic costs in the form of chromosome maintenance and replication (Lane and Martin 2010; Andersson et al. 2015; Wein et al. 2021). Hence, paralogues generated by intrachromosomal recombination are quickly lost for most of the prokaryotic gene functions. Mobile genetic elements such as transposons are notable exceptions and accumulate more gene duplications than any other gene function, because of their inherent ability to multiply within the chromosome. Overall, the low gene duplication frequencies that we observe for prokaryotes make sense. Realistic and phylogeny-independent estimates of the relative rates of LGT and gene duplication in prokaryotes are crucial for understanding and modeling prokaryotic genome evolution. Having taken many factors into account, the data indicate that the value of 54:1 obtained across lineages, rounded to 50:1, is a conservative lower bound estimate for the relative frequency of LGT over gene duplications in an average prokaryotic genome at any given point in time, using the present as the point of reference.

Materials and Methods

Data Set Preparation

Chromosome-encoded protein sequences for 5,655 prokaryotic genomes were downloaded from NCBI (Pruitt et al. 2007), version September 2016 (see supplementary table 1 for detailed species composition). The genomes were selected based on assembly to chromosome-level, and RefSeq status (O'Leary et al. 2016), while discarding metagenome-assembled genomes to avoid phylogenetic errors resulting from sequence contaminations (Garg et al. 2021). We then performed all-vs-all BLAST searches (Altschul et al. 1990) with BlastP, version 2.5.0, using default parameters and selected all reciprocal best hits with e-value $\leq 10^{-10}$. The selected protein pairs were aligned with the Needleman–Wunsch algorithm using EMBOSS needle (Chojnacki et al. 2017), and the pairs with global identity values $< 25\%$ were discarded. The retained global identity pairs were used for gene clustering using the Markov clustering algorithm (Enright et al. 2002) (MCL) version 12–068, with the following parameters for pruning: -P 180000, -S 19800, -R 25200. 260,972 gene families spanning at least two prokaryotic phyla were retained. Sequence alignments for each individual family were generated using MAFFT (Katoh 2002), with the iterative refinement method that incorporates local pairwise alignment information (L-INS-i; version 7.130). The resulting alignments were used to reconstruct maximum-likelihood trees with RAxML version 8.2.8 (Stamatakis 2014), using the WAG model of protein evolution, allowing for rate-heterogeneity across sites, and a predefined random seed for reproducibility (input parameters: -m PROTCATWAG -p 12345).

Inference of Recent Gene Duplications and Recent LGT

Gene Duplications

Recent, within-genome, gene duplications were inferred for genes sampled from the same genome that branched as sisters in the gene tree (fig. 1*b*). Gene duplications were scored for the genes fulfilling the criterion and the genomes harboring them. The frequency of duplications on a given genome was calculated as the number of duplicated genes divided by the total number of clustered genes (nonsingletons).

Lateral Gene Transfer

Recent LGTs were inferred based on gene distribution across genomes. A gene present in a single genome from a prokaryotic taxon was considered as a recent gene acquisition (fig. 1*a*). To detect LGT at different taxonomic ranges we considered taxonomic classifications at all levels (from domain to species)

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

and repeated the LGT inference procedure for each taxonomic level (supplementary table 1), discarding genomes without taxonomic assignments. LGT was scored for the genes fulfilling the criterion and the genomes harboring them. The frequency of LGT on a given genome was calculated as the number of acquired genes divided by the total number of clustered genes (nonsingletons).

Filter

To account for the possibility of false inferences arising from sparsely sampled prokaryotic lineages we considered two independent quality-filters: i) discarding genomes from taxa with only one member; ii) discarding genomes from taxa with less than six members.

Note that for LGT the various combinations of LGT types and filters resulted in a total of 21 LGT inferences, each based on a genome set. To allow for direct comparisons between LGT and gene duplications, gene duplication inferences were performed on the same genome sets and the frequencies of gene duplications and LGT were always compared using estimates derived from the same genome set.

Functional Annotation of Genes

The BRITE (Biomolecular Reaction pathways for Information Transfer and Expression) database was downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG, version September 2017) (Kanehisa et al. 2016), including protein sequences and their assigned functions according to the KO identifiers. The protein sequences from the 5,655 prokaryotic genomes were mapped to the KEGG database using "BlastP." Only the best hits with an e-value $\leq 10^{-10}$ and local coverage $\geq$ 80% were selected. After assigning a functional category for each protein sequence, at the "B" level from KEGG, a majority rule was used to assign a functional category for the gene families. For genes for which the majority rule rendered tied annotations, all equally supported functional categories were assigned to the family.

Statistical Analyses

Correlation of Gene Duplications and LGT with Genome Size

The frequency of recent gene duplications and LGT was estimated for each genome. For LGT at different taxonomic levels (from domain to species), only genomes from taxa with more than one member were used. For gene duplications, only genomes from species with more than one member were considered. For each considered genome, the frequency of events was paired with genome size, measured as the total number of protein-coding genes (supplementary table 2). The correlation between recent events and genome size was assessed with the two-tailed Spearman correlation test. The resulting *P*-values were adjusted with the Benjamin–Hochberg procedure (Benjamini and Hochberg 1995) and considered significant at FDR < 0.05.

Enrichment Tests across Functional Categories

To evaluate whether gene duplications and LGT happened more frequently in some gene functions than the theoretical expectation, we performed enrichment tests across the KEGG categories for each type of event separately—gene duplications and LGT—as follows.

For each KEGG category, we performed the one-tailed binomial test such that

$$H_0 : f \leq s$$
$$H_1 : f > s,$$

where f denotes the observed frequency of the event within the functional category (data in table 3) and s denotes the theoretical expectation, calculated as the frequency of the event across all functional categories ($s = 46\%$ for LGT and $s = 2\%$ for gene duplications). The resulting *P*-values were adjusted with the Benjamin–Hochberg procedure (Benjamini and Hochberg 1995) and considered significant at FDR < 0.05.

Code

All data analyses performed in this study were performed with custom MATLAB scripts, available upon request.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Data Availability

Sequence alignments, phylogenetic trees, and supplementary tables are available under: https://figshare.com/s/262ae6a8b2c4a281cf74.

Acknowledgments

We thank the European Research Council (advanced grants 666053 and 101018894), the Volkswagen Foundation (grant 93 046) and the Moore Simons Initiative on the Origin of the Eukaryotic Cell (grant 9743) for financial support.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403–410.

Andersson DI, Jerlström-Hultqvist J, Näsvall J. 2015. Evolution of new functions de novo and from preexisting genes. Cold Spring Harb Perspect Biol. 7(6):a017996.

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

Arakawa K. 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. Proc Natl Acad Sci U S A. 113(22):E3057.
Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 57(1):289–300.
Bennett PM. 2008. Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. Br J Pharmacol. 153(S1):S347–357.
Bratlie MS, et al. 2010. Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics 11(1):588.
Chojnacki S, Cowley A, Lee J, Foix A, Lopez R. 2017. Programmatic access to bioinformatics tools from EMBL-EBI update: 2017. Nucleic Acids Res. 45(W1):W550–W553.
Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol. 28(4):1481–1489.
Coissac E, Maillier E, Netter P. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. Mol Biol Evol. 14(10):1062–1074.
Coleman GA, et al. 2021. A rooted phylogeny resolves early bacterial evolution. Science 372(6542):eabe0511.
Curcio MJ, Derbyshire KM. 2003. The outs and ins of transposition: from MU to kangaroo. Nat Rev Mol Cell Biol. 4(11):865–877.
Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105(29):10039–10044.
Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc Natl Acad Sci U S A. 104(3):870–875.
Davis JJ, Xia F, Overbeek RA, Olsen GJ. 2013. Genomes of the class Erysipelotrichia clarify the firmicute origin of the class Mollicutes. Int J Syst Evol Microbiol. 63(Pt 7):2727–2741.
Doyon JP, Ranwez V, Daubin V, Berry V. 2011. Models, algorithms and programs for phylogeny reconciliation. Brief Bioinform. 12(5):392–400.
Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575–1584.
Garg SG, et al. 2021. Anomalous phylogenetic behaviour of ribosomal proteins in metagenome-assembled asgard archaea. Genome Biol Evol. 13(1):evaa238.
Gautreau G, et al. 2020. PPanGGOLiN: depitcting microbial diversity via partioned pangenome graph. PLoS Comput Biol. 16(3):e1007732.
Gevers D, Vandepoele K, Simillion C, Van De Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. Trends Microbiol. 12(4):148–154.
Goodman M. 1981. Globin evolution was apparently very rapid in early vertebrates: a reasonable case against the rate-constancy hypothesis. J Mol Evol. 17(2):114–120.
Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44(D1):D457–462.
Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.
Lane N, Martin W. 2010. The energetics of genome complexity. Nature 467(7318):929–934.
Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3(5):e130.
Li WH, Gu Z, Cavalcanti AR, Nekrutenko A. 2003. Detection of gene duplications and block duplications in eukaryotic genomes. J Struct Funct Genomics. 3(1-4):27–34.
Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 3(1-4):35–44.
McInerney JO, McNally A, O'Connell MJ. 2017. Why prokaryotes have pangenomes. Nat Microbiol. 2:17040.
Méheust R, et al. 2018. Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. Genome Biol. 19(1):75.
Nagies FSP, Brueckner J, Tria FDK, Martin WF. 2020. A spectrum of verticality across genes. PLoS Genet. 16(11):e1009200.
Nelson-Sathi S, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517(7532):77–80.
Nunes A, Gomes JP. 2014. Evolution, phylogeny, and molecular epidemiology of Chlamydia. Infect Genet Evol. 23:49–64.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405(6784):299–304.
Ohno S. 1970. Evolution by gene duplication. Heidelberg (Berlin): Springer.
O'Leary NA, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44(D1):D733–D745.
Popa O, Dagan T. 2011. Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol. 14(5):615–623.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35(Database issue):D61–5.
Ramisetty BCM, Santhosh RS. 2016. Horizontal gene transfer of chromosomal type II toxin–antitoxin systems of *Escherichia coli*. FEMS Microbiol Lett. 363(3):fnv238.
Reams AB, Roth JR. 2015. Mechanisms of gene duplication and amplification. Cold Spring Harb Perspect Biol. 7(2):a016592.
Robinson KM, Sieber KB, Dunning Hotopp JC. 2013. A review of bacteria-animal lateral gene transfer may inform our understanding of diseases like cancer. PLoS Genet. 9(10):e1003877.
Sela I, Wolf YI, Koonin EV. 2016. Theory of prokaryotic genome evolution. Proc Natl Acad Sci U S A. 113(41):11399–11407.
Sheridan PO, Thames Consortium, et al. 2020. Gene duplication drives genome expansion in a major lineage of Thaumarchaeota. Nat Commun. 11(1):5494.
Snel B, Bork P, Huynen MA. 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res. 12(1):17–25.
Soppa J. 2017. Polyploidy and community structure. Nature Microbiol. 2:16261.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30(9):1312–1313.
Szöllosi GJ, Boussau B, Abby SS, Tannier E, Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci U S A. 109(43):17513–17518.
Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. Philos Trans R Soc Lond B Biol Sci. 370(1678):20140335.
Treangen TJ, Rocha EPC. 2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet. 7(1):e1001284.
Tria FDK, et al. 2021. Gene duplications trace mitochondria to the onset of eukaryote complexity. Genome Biol Evol. 13(5):evab055.
Vos M, Hesselman MC, te Beek TA, van Passel MWJ, Eyre-Walker A. 2015. Rates of lateral gene transfer in prokaryotes: high but why? Trends Microbiol. 23(10):598–605.
Wang S, Chen Y. 2018. Phylogenomic analysis demonstrates a pattern of rare and long-lasting concerted evolution in prokaryotes. Commun Biol. 1(1):12.
Wein T, et al. 2021. Essential gene acquisition destabilizes plasmid inheritance. PLoS Genet. 17(7):e1009656.

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021

Williams TA, et al. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc Natl Acad Sci U S A. 114(23):E4602–4611.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387(6634):708–713.

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution and genetic heterogeneity. In: Horizons in biochemistry. New York: Academic Press. p. 189–225.

Associate editor: Tal Dagan

Downloaded from https://academic.oup.com/gbe/article/13/10/evab224/6380140 by Universitaetsbibliothek Duesseldorf user on 18 November 2021