Journal of Molecular Evolution (2018) 86:190–203
https://doi.org/10.1007/s00239-018-9836-x

ORIGINAL ARTICLE

Horizontal Gene Transfer Building Prokaryote Genomes: Genes Related to Exchange Between Cell and Environment are Frequently Transferred

Apuã C. M. Paquola[1,2] · Huma Asif[1,3] · Carlos Alberto de Bragança Pereira[4] · Bruno César Feltes[5,6] · Diego Bonatto[5] · Wanessa Cristina Lima[1,7] · Carlos Frederico Martins Menck[1]

Received: 14 December 2017 / Accepted: 15 March 2018 / Published online: 19 March 2018

Abstract
Horizontal gene transfer (HGT) has a major impact on the evolution of prokaryotic genomes, as it allows genes evolved in different contexts to be combined in a single genome, greatly enhancing the ways evolving organisms can explore the gene content space and adapt to the environment. A systematic analysis of HGT in a large number of genomes is of key importance in understanding the impact of HGT in the evolution of prokaryotes. We developed a method for the detection of genes that potentially originated by HGT based on the comparison of BLAST scores between homologous genes to 16S rRNA-based phylogenetic distances between the involved organisms. The approach was applied to 697 prokaryote genomes and estimated that in average approximately 15% of the genes in prokaryote genomes originated by HGT, with a clear correlation between the proportion of predicted HGT genes and the size of the genome. The methodology was strongly supported by evolutionary relationships, as tested by the direct phylogenetic reconstruction of many of the HGT candidates. Studies performed with *Escherichia coli* W3110 genome clearly show that HGT proteins have fewer interactions when compared to those predicted as vertical inherited, an indication that the number of protein partners imposes limitations to horizontal transfer. A detailed functional classification confirms that genes related to protein translation are vertically inherited, whereas interestingly, transport and binding proteins are strongly enriched among HGT genes. Because these genes are related to the cell exchange with their environment, their transfer most likely contributed to successful adaptation throughout evolution.

Keywords Horizontal gene transfer · Evolution · Phylogenetic tree incongruence · Transport proteins

Background

Horizontal (or lateral) gene transfer (HGT) is the process by which an organism incorporates genetic material from another organism of a different species. This process contrasts with the normal process of vertical inheritance in

Apuã C. M. Paquola and Huma Asif contributed equally to this manuscript.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00239-018-9836-x) contains supplementary material, which is available to authorized users.

✉ Carlos Frederico Martins Menck
cfmmenck@usp.br

[1] Department of Microbiology, Institute of Biomedical Sciences, University of Sao Paulo, Av. Prof. Lineu Prestes, 1374, São Paulo, SP 05508-000, Brazil

[2] Lieber Institute for Brain Development, Baltimore, MD, USA

[3] Department of Psychiatry and Behavioral Neurosciences, The University of Chicago, Chicago, IL, USA

[4] Department of Statistics, Institute of Mathematics and Statistics, University of Sao Paulo, São Paulo, Brazil

[5] Center of Biotechnology, Department of Molecular Biology and Biotechnology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[6] Institute of Informatics, Structural Bioinformatics and Computational Biology Laboratory, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil

[7] Department of Pharmacology, University of Heidelberg, Im Neuenheimer Feld 366, 69120 Heidelberg, Germany

which an organism receives genetic material directly from ascendants of the same species. HGT is particularly relevant in prokaryotic organisms, although gene transfer between eukaryotes and prokaryotes has also been described (Ros and Hurst 2009; Lima et al. 2009; Arama et al. 2015; Yin et al. 2016; Matassi 2017). The transfer of genetic material occurs more frequently from closely related species, although it also occurs between strains far outside the proximal gene pool (Koonin and Wolf 2008; Bosi et al. 2017).

HGT is widespread across the prokaryotic lineage, but to what degree it exerts an effect in reconstructing the evolutionary history of a given organism is still controversial (McInerney et al. 2008; Boucher and Bapteste 2009). Some researchers argue that HGT is constrained by important selective barriers, and has limited influence on the evolution of modern organisms (Kurland et al. 2003), while others report that HGT has affected every single prokaryotic gene over the full span of evolutionary history and that a reticulated network may be generated rather than a vertical tree (Dagan et al. 2008; Hily et al. 2017). The latter view is highly embraced by many researchers, who argue that a strictly vertical tree-like model can no longer explain the evolution of life.

There is increasing evidence that, whenever HGT is frequent enough, different parts of a genome reflect different evolutionary histories (Gogarten and Townsend 2005; Doolittle and Bapteste 2007). For example, the phylogeny of ribosomal components groups Thermotogales with Aquificales, while in whole-genome phylogenies, Thermotogales cluster together with *clostridia* and *bacilli* (Gophna et al. 2005; Zhaxybayeva et al. 2009). In addition, increasing evidence of horizontal acquisition based on discrepancies in tree topologies in prokaryotic genomes has been reported for several bacterial phyla (Comas et al. 2006; Williams et al. 2010; Cuecas et al. 2017).

Early work on HGT using four microbial genomes led to the hypothesis that genes related to operational (housekeeping) roles are more prone to HGT than genes related to information processing (replication, transcription, and translation) (Rivera et al. 1998; Mykowiecka et al. 2017). In a subsequent work with six genomes, the authors proposed that genes that encode products with complex interactions with many proteins in the cells (normally observed for informational genes) have fewer chances to be transferred, which is known as the complexity hypothesis (Jain et al. 1999). However, evidence exists that genes related to primary metabolism can be transferred among relatively distant clades: this is the case, for example, of the arginine biosynthesis operon (Martins-Pinheiro et al. 2016) and NAD biosynthesis pathways (Lima et al. 2009), which are shared by eukaryotes and bacteria from the Xanthomonadales and Flavobacteriales groups.

To further our understanding of HGT, we have developed a sequence similarity-based computational method for the detection of genes involved in HGT and a systematic study in 697 prokaryotic genomes. The method takes advantage of discrepancies between BLAST similarity scores and phylogenetic distances to predict HGT in a computationally efficient way. We use these predictions to perform a statistical analysis on the functional categories enriched or depleted in HGT, the relationship between genome size and HGT proportion, and the relationship between HGT and complexity of protein interaction networks.

Methods

Genome Dataset Processing

The complete sequences of 697 prokaryote genomes were retrieved from Omniome database version 22, which is part of the comprehensive microbial resource (CMR) (Peterson et al. 2001) and imported into the MySQL database manager. The pairwise 16S rRNA distances between the genomes were calculated as follows: (i) the 16S rRNA genes of the selected genomes were aligned with the program Kalign2 (Lassmann et al. 2009) using a gap penalty of 3.94; (ii) 100 bootstrap replicates were generated with the Seqboot program of the PHYLIP package; and (iii) for each replicate, the distance between the 16S rRNA sequences for each pair of genomes was calculated using the DNAdist program (F84 distance model) of the PHYLIP package (Felsenstein and Churchill 1996).

Prediction of Genes Originated by HGT

For vertically inherited genes, protein similarity score to homologous genes tends to decrease as the phylogenetic distance to other organisms increases. Deviation from this trend is taken as indication for potential horizontal gene transfer. The protein similarity score used in this study is the BLAST-Extend-Repraze (BER), and it was adopted for two reasons: (1) it is precomputed in the CMR database for all pairs of protein-coding genes and (2) it makes protein sequence similarity more robust to sequencing errors by extending searches into nucleotide sequences flanking gene predictions. The BER tool extends the nucleotide query sequence 300 bp upstream and downstream of predicted genes and then performs a Smith–Waterman alignment to their top BLASTx protein matches. This extension procedure makes protein searches resilient to sequencing errors that could lead to truncated sequence in gene predictions. The phylogenetic distance measurement between two organisms used in this study is the 16S rRNA sequence distance and it was adopted as 16S rRNA is the prototypical vertical inheritance gene widely used to identify prokaryotes. The source of most of the 16S rRNA sequences was from Omniome database. When not available, the program RNAmmer was used to extract the 16S rRNA sequence from the genomic sequence (Lagesen et al. 2007). The classification of a gene as vertically inherited or horizontally transferred was based on their BER targets (in decreasing order of score) and two user-defined parameters: d_{self} (0.105) and d_{distant} (0.2) (Fig. S1). The parameter d_{self} excludes the genomes that are very close to the genome under study (close strains), whereas d_{distant} is chosen by the user as the minimum distance of interest for evaluating horizontal transfer. This particular choice of parameters is based on a taxonomical argument, so that most prokaryotic species from the same genus have pairwise distance $< d_{\text{self}}$. Conversely, the d_{distant} parameter reflects typical phylogenetic distances between different subdivisions of the proteobacteria group, the most represented in the database. Let X be the target BER of a gene under study, $\overline{d}_{\text{X}}$ be the 16S rRNA distance from the organism under study and the target X (averaged over bootstrap replicates), and A be the first "non-self" target, i.e., the first target that satisfies $\overline{d}_{\text{A}} > d_{\text{self}}$. The gene is considered: Typical (most likely of vertical origin) if $\overline{d}_{\text{A}} \leq d_{\text{distant}}$, Atypical (HGT candidate) if $\overline{d}_{\text{A}} > d_{\text{distant}}$, and undetermined if there is no such target A with $\overline{d}_{\text{A}} > d_{\text{self}}$ (meaning that there is no evidence of either HGT or VGT). The HGT genes were further classified into class 1 if there exists at least one BLAST hit B with lower score than A, such that $d_{\text{B}} < d_{\text{A}}$ with statistical significance (i.e., P($d_{\text{B}} > d_{\text{A}}$) > 0.99, estimated from the bootstrap replicates), or into class 2 if there is no such BLAST hit B.

Enrichment Analysis of the Functional Gene Categories Potentially Originated by HGT

An enrichment analysis for typical and atypical genes into functional gene categories was based on the classification system from the Omniome database, with some modifications. These modifications included the creation of the categories "Pathogenesis, toxin production, and resistance" (originally within the category "cellular processes") and "DNA restriction/modification" (originally in "DNA metabolism"). The enrichment factor (or depletion) and statistical significance (p values) of atypical genes in each functional category were estimated using the two-tailed Fisher's test (Fisher test function in R, with enrichment factor ≥ 1.5). The enrichment factor is the rate of change of the ratio between the annotated number of genes in category X and the number of non-annotated genes in the same category. To control the rate of false positives, we calculated the q values (< 0.01) from the obtained p values. The q value expresses the expected proportion of false positives among the results that are considered significant.

Phylogenetic Validation

To further validate the HGT candidates, we carried out phylogenetic analyses for transport and binding protein genes, mainly because this category appeared as the most HGT-enriched functional category. Orthologs were identified using the basic local alignment search tool (BLASTp) at ExPASy proteomics server (E value $< 10^{-04}$ and percent of similarity ≥ 30%). Multiple sequence alignments were performed with multiple sequence comparison by log-expectation (MUSCLE) algorithm (Edgar 2004). Phylogenies were constructed using Molecular Evolution Genetic Analysis, version 6.0 (Tamura et al. 2013; Iqbal et al. 2017). All trees were constructed using the maximum likelihood method with bootstrapping of 1000 iterations; only those bootstrap values larger than 50% were considered to identify supported nodes (Iqbal et al. 2017). Phylogenetic trees were drawn with orthologs from the main representative groups, with at least five organisms from each group (except for the cases in which sequence coverage is extremely low).

Interactome Data Mining and the Design of the Interactomes

To design the interactomes and to elucidate the interplay between the proteins that were encoded by genes with vertical inheritance or HGT candidates in a topological context, the metasearch engine STRING 9.1 was used. All of the genes that were classified as typical or atypical for the *E. coli* W3110 genome, as well as genes that are related to the transport of macromolecules and protein synthesis, were used as the initial seeds for network prospecting. Each connection (edge) possesses a degree of confidence between 0 and 1.0 (1.0 being the highest confidence). The parameters used were as follows: all prediction methods enabled, excluding text mining; degree of confidence, medium (0.400); and a network depth equal to 1. The results were analyzed with Cytoscape 2.8.2 (Shannon et al. 2003).

Results

Detecting Potential HGT Genes in Prokaryotic Genomes

A large-scale search of potential HGT genes was performed based on 697 microbial (Bacteria and Archaea) genomes (the list of genomes analyzed is provided in Table S1, Supplementary Material). The main assumption was that HGT genes would have BLAST results in which the order of similarity with orthologous proteins would differ from the 16S rRNA gene distances (which is a product of vertical inheritance). Genes with a typical distance (BLAST order similar to the 16S rRNA distance, $\overline{d}_{\mathrm{A}} \leq d_{\mathrm{distant}}$) were considered as from vertical inheritance. When the gene was observed only in closely related species (with no evidence for either HGT or vertical inheritance), it was considered undetermined. The atypical genes, which were potentially acquired by HGT, were those in which the 16S rRNA distance was higher than that expected for the order of orthologous genes ($\overline{d}_{\mathrm{A}} > d_{\mathrm{self}}$). These atypical genes were further classified into two different categories (Fig. S1): a gene that is assigned to class 1 is potentially involved in replacing the orthologous gene by horizontal transfer, while a gene that is assigned to class 2 may be involved in the acquisition of a novel biological function into the recipient genome.

Out of the 697 genomes, from more than 2 million analyzed genes, 18% were found to be atypical (potentially acquired by HGT, a list of these genes is provided in Table S2 Supplementary Material). As shown below, the number of phylogenetic neighbors (i.e., the number of genomes with a distance between d_{self} and d_{distant}) influences the determination of atypical genes (mainly due to the identification of class 2 genes). When considering only genomes with more than twenty neighbors, the fraction of atypical genes was on the same order of magnitude (15%) (Table 1).

The BLAST analysis for the search of HGT genes is sensitive to the number of phylogenetically closely related orthologs. This sensitivity is clearly shown when the number of atypical genes was plotted considering the number of phylogenetic neighbor genomes: when fewer neighbors were observed, there was a tendency for high variation in the number of atypical genes per genome (Fig. 1a). Not surprisingly, most of the atypical genes in less-represented genomes belong to class 2, and those assigned as class 1 increase with the number of neighbor genomes (Fig. 1b). However, when more than 20 neighbor genomes exist, the frequency of atypical genes stabilizes with an average of approximately 15% per genome (Fig. 1a), with most of them from class 1 (72%, Fig. 1b). From now on, the genes that were considered for this study refer only to those genomes with more than 20 phylogenetically related neighbors.

Table 1 Total number of genes that were classified as HGT (atypical), vertical inheritance (typical), or undetermined, for all 697 of the studied genomes and for the 491 genomes that have ≥ 20 phylogenetic neighbors

	All genomes		Genomes with ≥ 20 neighbors	
	Count	Fraction	Count	Fraction
Atypical	405,206	0.18	277,198	0.15
Typical	1,044,281	0.45	899,765	0.49
Undetermined	846,119	0.37	641,735	0.35
Total	2,295,606		1,818,698	

The proportion of HGT candidate genes was also assessed in relation to the number of genes in prokaryotic genomes (Fig. 1c). There is a clear positive correlation between the proportion of HGT genes and the number of genes in the genome, suggesting that organisms with larger genomes, once a core set of biological functions are encoded in the genome, can have a larger fraction of their genomes more susceptible to rapid evolution and gene exchange. Intracellular symbionts, which possess much reduced genome sizes, are among the organisms with fewer HGT genes. These organisms, which live in a hospitable environment provided by the host, have reduced opportunities for the uptake of new genes by HGT and strong selection for fast replication and a small genome.

Functional Category Profiling of HGT Genes

The functional identity of HGT genes is an essential question to understand how transfer contributes to evolution. The enrichment factor, which identifies whether the category presented a higher (or lower) proportion of genes that were atypical (or typical), was thus defined for each category. The categories of pathogenesis, toxin production, and resistance, DNA restriction/modification, transport proteins and binding, functions related to mobile elements, and intermediary metabolism are among those indicated as significantly enriched in HGT genes (> 30% of the genes in each category) (Table 2). However, protein synthesis, transcription, and DNA metabolism are the categories that were enriched in genes with vertical inheritance (< 15% of the atypical genes).

The category enrichment was also analyzed considering individual genomes, with each genome grouped by phylogenetically related clades (Fig. 2). In this case, the pattern is not very different from that observed on the global analysis, except that in some categories the low number of genes implies a loss of significance. This is the case for DNA restriction/modification genes, which are highly significant in the global analysis (Table 2), but not when individual genomes are analyzed. These results indicate the significant enrichment of many categories, but two categories stand out: protein synthesis (for vertically inherited genes) and transport and binding proteins (for potentially transferred genes). Moreover, these results confirm that the genes that are related to informational metabolism have restricted possibilities to be transferred

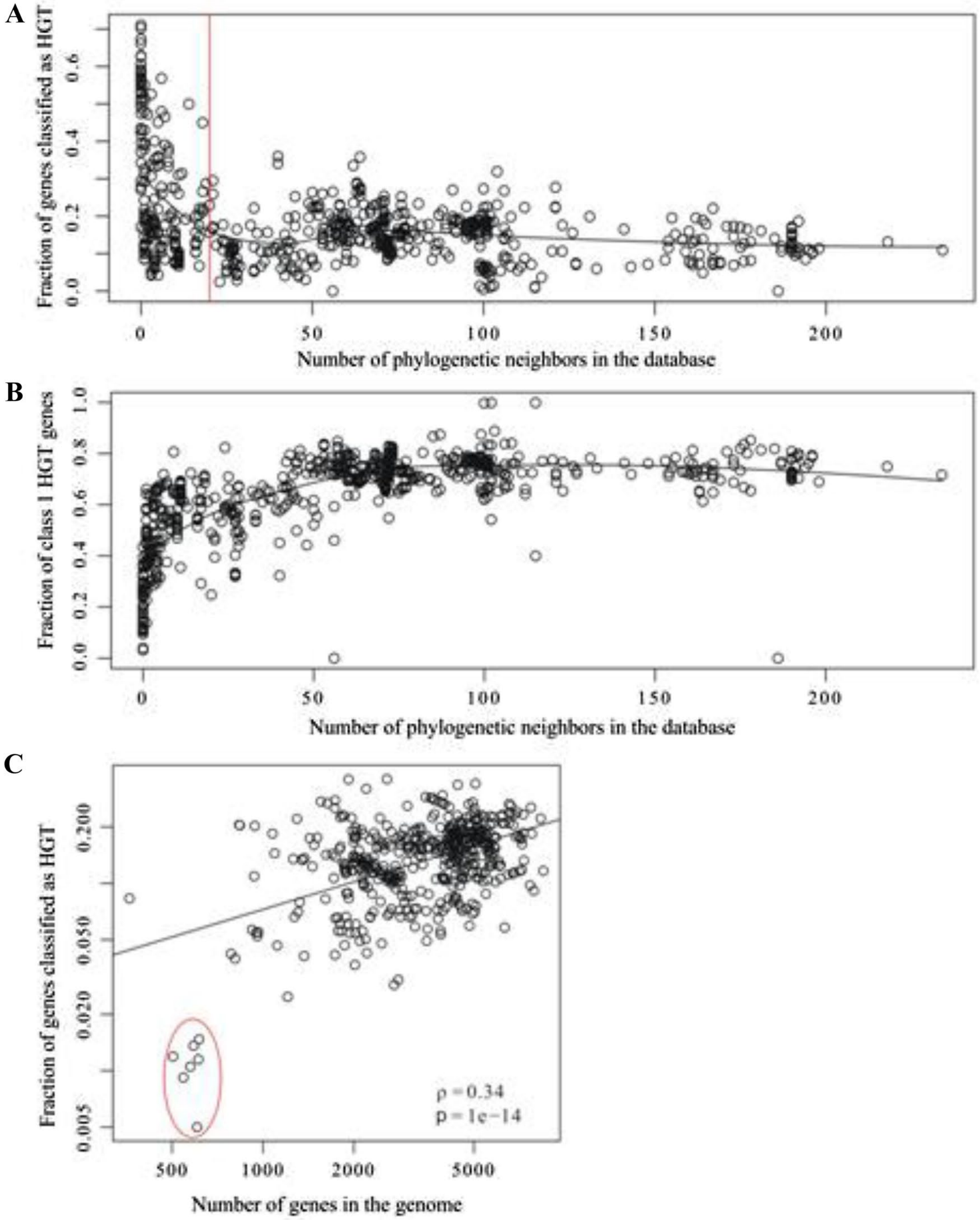


Fig. 1 Overview of HGT candidate genes in the prokaryote genomes. **a** Fraction of genes that were classified as HGT for each genome as a function of the number of phylogenetic neighbors, i.e., genomes with 16S rRNA distance between d_{self} and d_{distant} from the source genome. The black solid line is a LOESS fit to the data, and the red vertical line corresponds to the threshold of 20 neighbors that was used in this study. **b** Fraction of the HGT candidate genes that were assigned to class 1. The black solid line is a LOESS fit to the data. **c** Fraction of genes that were classified as HGT candidates as a function of the number of genes in the genome. ρ Spearman's rank correlation. The organisms that are annotated with the red ellipse are all insect intracellular endosymbionts

Table 2 Global proportion of HGT candidate genes as grouped by functional categories, for the genomes with ≥ 20 phylogenetic neighbors

Functional category	# Typical and atypical genes	% HGT candidate genes	Enrichment factor	$p<0.05$
Protein synthesis	67,687	8.48	0.36	*
Transcription	19,203	10.60	0.45	*
DNA metabolism	47,791	13.12	0.55	*
Purines, pyrimidines, nucleosides, and nucleotides	29,280	14.98	0.63	*
Biosynthesis of cofactors, prosthetic groups, carriers	52,210	15.31	0.65	*
Protein fate	60,149	15.96	0.67	*
Amino acid biosynthesis	42,650	16.11	0.68	*
Cellular processes	51,195	20.61	0.87	*
Cell envelope	84,649	23.73	1.00	*
Fatty acid and phospholipid metabolism	33,015	24.54	1.04	*
Regulatory functions	94,006	25.00	1.06	*
Signal transduction	15,865	25.47	1.08	*
Energy metabolism	152,019	25.49	1.08	*
Conserved hypothetical proteins	78,251	25.90	1.09	*
Unclassified	187,009	26.84	1.13	*
Non-conserved hypothetical proteins	25,800	27.79	1.17	*
Central intermediary metabolism	57,734	30.88	1.31	*
Pathogenesis, toxin production, and resistance	26,165	32.42	1.37	*
Transport and binding proteins	157,961	31.97	1.35	*
Mobile and extrachromosomal element functions	17,884	36.03	1.52	*
DNA restriction/modification	3,370	44.00	1.86	*
Total	1,176,963	23.55	NA	NA

The enrichment factor is the ratio of the per-category to the total HGT proportion

*The statistical significance of enrichment is calculated with Fisher's exact test

horizontally, and those that are related to more peripheral metabolism are prone to HGT.

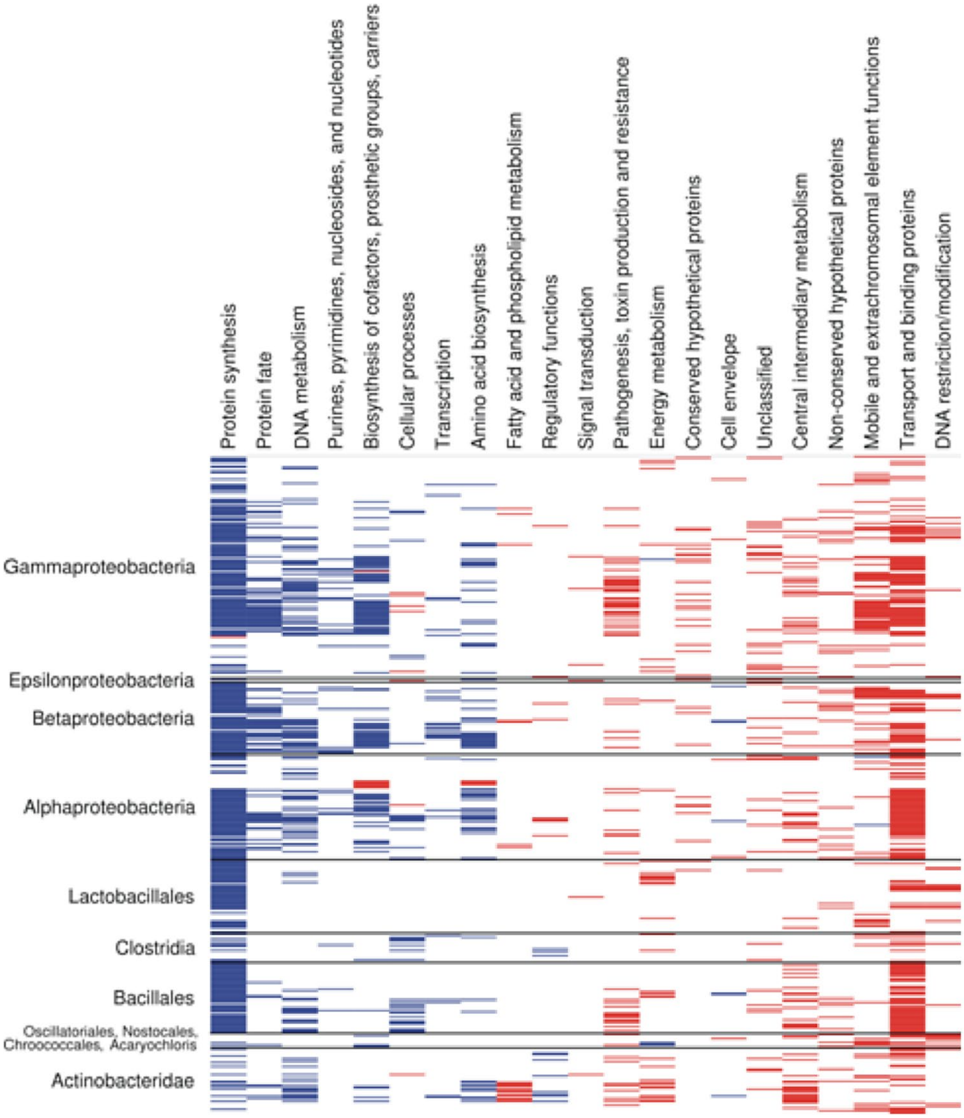
Interactome Analyses Indicate that the Proteins that are Encoded by HGT Genes Have Fewer Interactions

HGT genes have been proposed as genes with a lower number of interactions, a hypothesis that is commonly referred to as the complexity hypothesis (Jain et al. 1999; Yin et al. 2016; Matassi 2017). To investigate the topological aspects of the connections between HGT and vertically inherited genes, we considered the *E. coli* W3110 genome, for which a good amount of information on protein–protein interactions is available. The STRING metasearch engine prospected the typical and atypical genes (Fig. 3a). When all of the genes were considered, a strongly connected network is clearly observed, dominated by the high number of typical genes. However, if atypical and typical genes were considered separately, the interactome revealed different interesting aspects. First, the number of connections was much higher for typical genes (a total of 14,332 connections between 1935 genes, with a ratio of 7.4 connections per gene) than for atypical genes (799 connections between 557 genes, ratio of 1.4) (Fig. 3b, c). When 557 typical genes were chosen at random to test for interactions, the number of connections dropped to 5.4, but it was still much higher than that observed for HGT genes. These data strongly support the idea that the number of interactions that a protein establishes in the cell restricts the chances of its gene being transferred to and maintained in a different host genome during evolution. Moreover, these data also show that, although HGT genes have fewer interactions, they still have connections among themselves. When protein synthesis genes with vertical inheritance and HGT candidate genes of the transport and protein-binding category were considered separately (Fig. 4), a higher level of connections for the genes with vertical inheritance is clearly observed.

Phylogenetic Validation

The data presented assume that BLAST searches can provide information on evolutionary distances. Although this assumption is correct in many cases, protein domains may directly affect the BLAST analyses, yielding biased results and that could promote false results. Thus, the phylogenetic

Fig. 2 Functional enrichment matrix for genomes with ≥ 20 phylogenetic neighbors. Each row represents a genome and each column represents a functional category. A cell is marked red (or blue) if the HGT proportion in that category is significantly higher (or lower) than that for genes not in that category. Statistical significance is calculated using a chi-squared test with 10^6 Monte Carlo simulations and requiring a false discovery rate < 0.01 (Benjamini–Hochberg method) and an enrichment (or depletion) factor ≥ 1.5



incongruence method (i.e., incongruence between the phylogeny of the selected gene and the known species phylogeny for vertical inheritance) was used to further validate genes that are related to transport and binding proteins detected as atypical (Table 3; Figs. 5, 6). For example, two different *Xanthomonas axonopodis pv. citri 306* transporter genes have eukaryote (Fig. 5) or verrucomicrobia and bacteroidetes (Fig. 6) as clade neighbors. Several other transporter genes were also investigated, and the phylogenies confirm that the genes that were classified as of potential HGT origin are grouped in clades that present phylogenetically distant orthologs (Figs. S02–S16).

In total, from a total of 134 genes classified as atypical in the BLAST search (belonging to Proteobacteria, Firmicutes, Actinobacteria, Spirochaetes, and Cyanobacteria), 75 (56%) presented an incongruent phylogeny, confirming their origin by HGT. The remaining 59 (44%) were either weakly supported by the bootstrap values or had hits within the same group, but could not be excluded from the classification as potential HGT genes. However, the phylogenetic distribution of proteins that were related to protein synthesis confirmed (all of the 25 genes, analyzed) their vertical inheritance. Therefore, our phylogenetic analyses support the BLAST search method in successfully classifying genes as acquired either by horizontal transfer or by vertical inheritance.

Discussion

The analysis of complete genome sequences has helped enormously in addressing many issues concerning microbe evolution. One such area is the acquisition of new genes by horizontal gene transfer. To address the question of whether a particular gene owes its presence in a particular genome due to HGT, several methods have been proposed. These methods are broadly classified into sequence composition

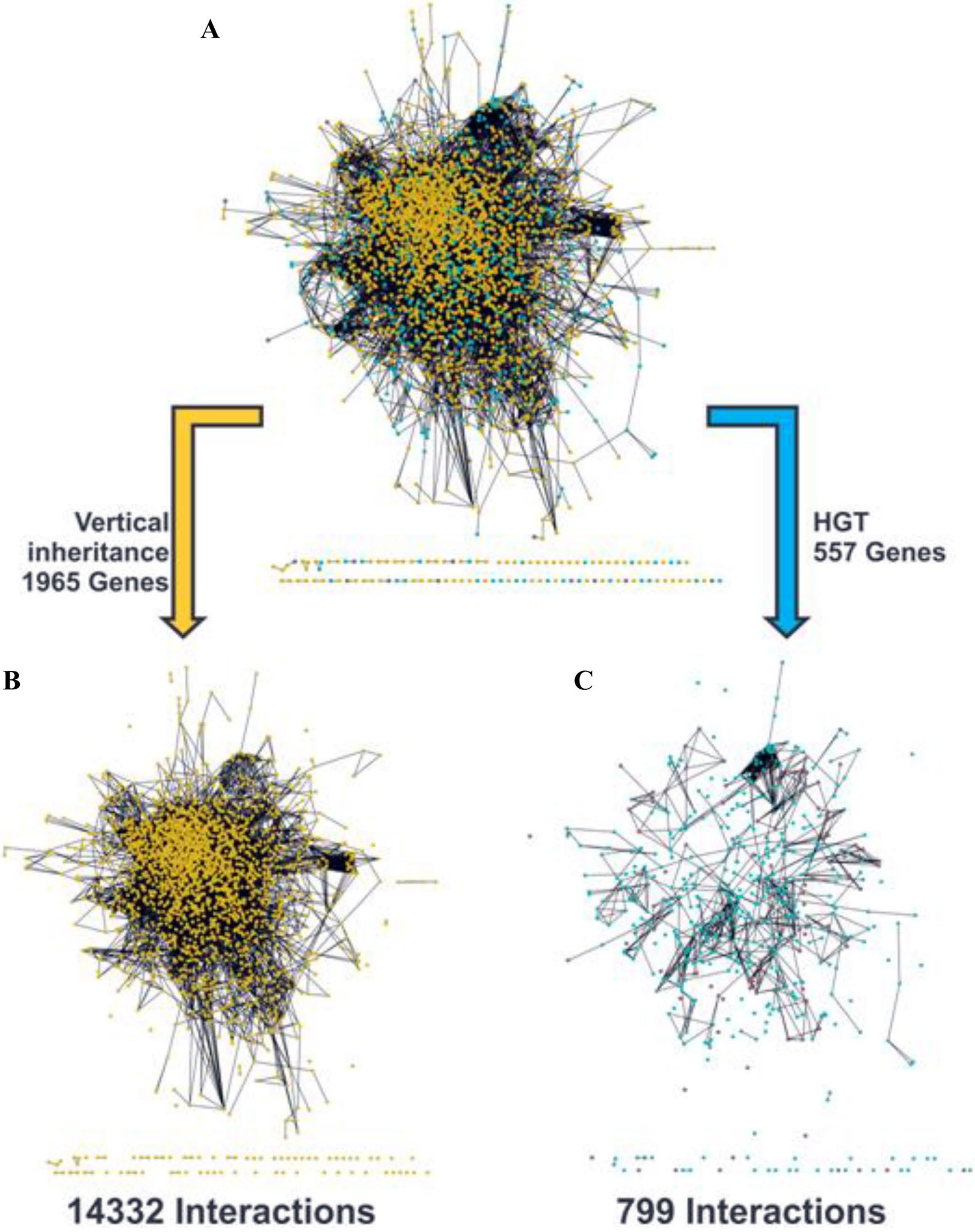


Fig. 3 Vertical inheritance and HGT networks showing the interplay between typical and atypical genes of the *E. coli* strain W3110 genome; vertically inherited genes are depicted in yellow and HGT genes in blue. **a** Global network, containing 2522 connected genes and 111 not connected genes. This global network was subdivided into two networks: **b** for those of vertical inheritance, containing 1965 genes, and **c** HGT, with 557 genes

methods or the phylogeny-based methods (Ragan 2001). For example, HGT genes can be distinguished by identifying genes with unusual features (e.g., nucleotide composition or codon usage) when compared to other genes from its genome, before it ameliorates (Mrázek and Karlin 1999). These methods only require the sequence of the host genome, precluding the need of comparison to other genomes. The main drawback is that unusual compositions

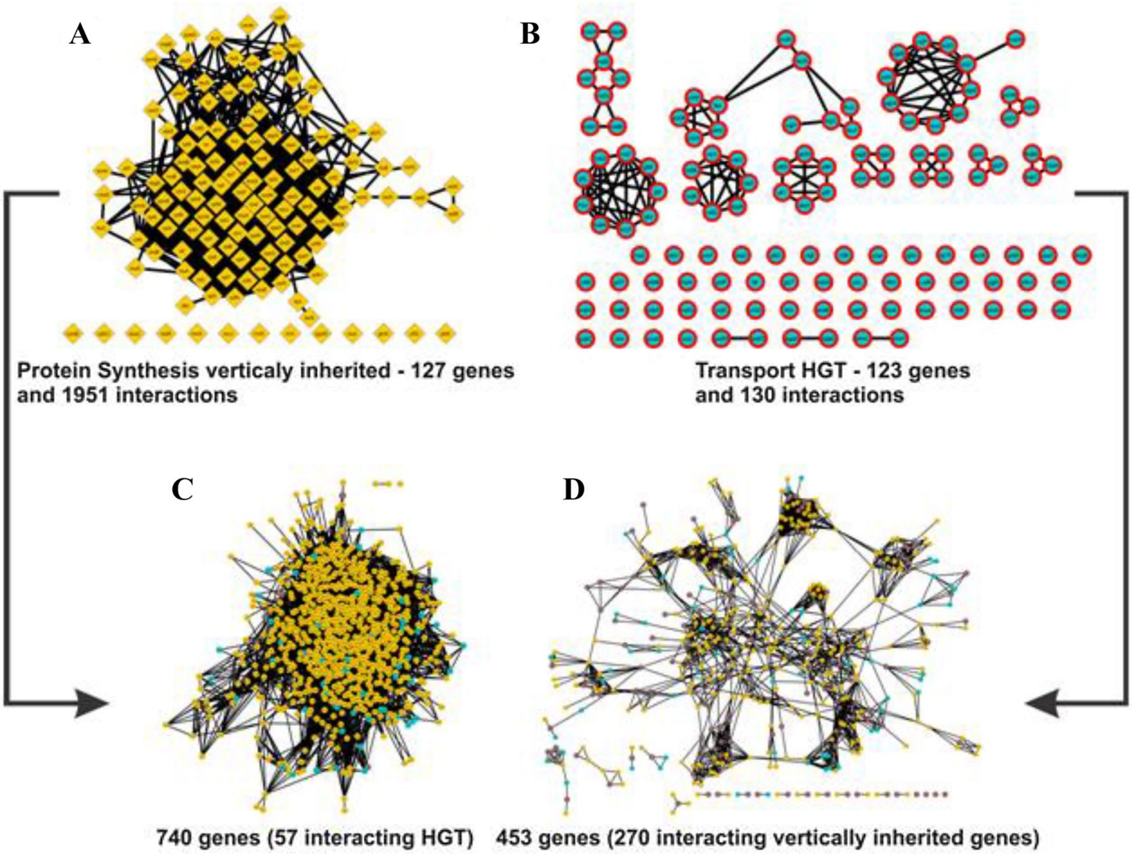

A
B
Protein Synthesis verticaly inherited - 127 genes and 1951 interactions
Transport HGT - 123 genes and 130 interactions
C
D
740 genes (57 interacting HGT)
453 genes (270 interacting vertically inherited genes)

Fig. 4 Interactome of the proteins encoded by genes related to protein synthesis or transport and binding protein in *E. coli* strain W3110. **a** All of the vertically inherited genes (yellow) that are related to protein synthesis (diamond shaped); **b** all of the HGT genes (blue) that are related to transport and binding function (red bordered nodes); **c** the genes in the subnetwork in (**a**) were expanded by all of the immediately connected genes, indicating that vertically inherited genes have a preference for their own class; and **d** the genes in subnetwork in (**b**) were also expanded by all of the immediately connected genes, indicating that HGT genes also have a preference for vertically inherited genes

can also be caused by factors other than HGT, such as selection, mutation bias. Moreover, this method is unable to detect neither transfers between species with similar compositions nor those that have occurred a long enough time ago and amelioration has been completed (Lawrence and Ochman 1997). Phylogeny-based detection of HGT is one of the most commonly used approaches for detecting HGT. It is based on the fact that HGTs can cause incongruence in the gene tree as well as create conflict with the species phylogeny (i.e., there is no grouping of taxa in common). However, the variability in the outcome of these methods demands the development of new tools that consider evolutionarily related genomes and not only the genome composition.

In this work, we present a method that can detect, in large scale, HGT candidate genes through BLAST similarity. Clearly, the variation in genome size and the number of phylogenetically close neighbors can considerably affect the results. The results indicate that the larger the genome size, the larger the number of HGT genes, in agreement with previous studies (Cordero and Hogeweg 2009). A possible explanation for this finding is that larger genomes tend to be composed of multiple plasmids and megaplasmids that can facilitate higher rates of transmembrane DNA translocation, providing the organisms with more flexibility to survive in different environments (Cordero and Hogeweg 2009). Interestingly, intracellular symbiotic bacteria with very small genomes are those with a lower percentage of HGT genes. This is most likely due to a dependence on the metabolic processes of the host cells, and a tendency to lose the extra genes.

The BLAST method also depends on the number of close evolutionary neighbors found in the genome analyzed, most likely due to the low reliability in HGT prediction when few neighbors are found. However, when a genome has more than 20 neighbors in the database, then the frequency of detected HGT genes varies, mainly, in the range of 5–25%, with an average of 15% of the total number of genes. Most of these genes are class 1 (72% of

Table 3 Genes that were classified as class 1 HGT were selected from different bacterial groups including Proteobacteria, Firmicutes, Actinobacteria, Spirochaetes, and Cyanobacteria

Bacterial groups	Gene locus tag	Gene name	Nearest neighbor	NJ bootstrap (%)	Sequence length (A.A)
Xanthomonadales	XAC0860	ABC transporter ATP-binding protein	Eukaryota	99	480
	XC_2737	ABC transporter ATP-binding protein	Acidobacteria	74	291
	XAC0179	ABC transporter ATP-binding protein	Cyanobacteria	95	281
	XAC0819	Nucleoside transporter	Verrucomicrobia and Bacteroidetes	67	432
Alpha proteobacteria	CC_1204	AcrB/AcrD/AcrF family protein	Gamma Proteobacteria	62	1047
Firmicutes	RBAM_004660	*mntH*	Gamma Proteobacteria	99	424
	RBAM_003770	*yckJ* (amino acid ABC transporter permease)	Actinobacteria	84	229
	BA_0232	Oligopeptide ABC transporter	Actinobacteria	91	318
	BAA_0983	Sulfate permease family protein	Actinobacteria	100	492
Actinobacteria	Aaur_0057	Arsenite efflux pump	Deinococcus	87	331
	Aaur_0347	D-ribose transport system ATP-binding protein	Chloroflexi	96	514
	CE0611	*gntP* (gluconate permease protein)	Beta Proteobacteria	100	464
	Rv0362	Magnesium transporter (*mgtE*)	Gamma Proteobacteria	100	460
Spirochaetes	TDE_0130	sodium/dicarboxylate symporter family protein	Firmicutes	95	457
	LBF_0368	Periplasmic binding protein of an ABC transporter complex	Aquificae	97	422
Cyanobacteria	AM1_0014	Uracil–xanthine permease	Deferribacteres	60	416
	Slr0982	ABC transporter (*rfbB*)	Archaea	99	430

the detected HGT genes), suggesting that only a minor fraction (28% belonging to HGT class 2, and this number may decrease with the number of complete genomes available) corresponds to novel biological functions that are acquired by the recipient organism.

An interesting genome wide algorithm to identify HGT genes, known as DarkHorse, was proposed before (Podell and Gaasterland 2007; Podell et al. 2008). Similarly to this work, gene alignment (BLAST) was also used, but HGT inference was based on a species distance metric calculated from the species taxonomy, providing lists of genes that can be potentially transferred among different organisms. A recent work has performed genome wide search of HGT genes in prokaryotes using orthologous genes tree and reconciliation with 16S RNA reference trees (Jeong and Nasir 2017), providing an important preliminary list of potential HGT candidate genes.

Nakamura et al. (2004), working with 116 prokaryotic genomes, reported several functional biases of HGT genes, identifying functions that are more prone to transfer (such as DNA mobility, pathogenicity, DNA binding, and cell surface) (Nakamura et al. 2004). That study identified HGT genes based on gene composition, which detects mainly recent events, failing to detect HGT cases in which sequences have completed the amelioration process. In this work, we found that DNA restriction/modification had the highest number of HGT genes, in agreement with an earlier study that reported that HGT has a greater influence on the distribution and evolution of DNA restriction/modification systems, although the approach that was used for the analysis was based on biased codon usage (Jeltsch and Pingoud 1996). The second highest enriched category was transport and protein binding. This functional category includes well-characterized members of the superfamily of transporters, which are closely related in terms of structure, function, and evolutionary origin. ABC transporters can transport different substrates such as inorganic ions, amino acids, sugars, polysaccharides, and even proteins (Higgins et al. 1990), and have been recently found as a main hotspot for gene transfer within the human gut microbiome (Meehan and Beiko 2012). In fact, an analysis of the nickel and iron transport complex in many bacteria revealed several operons that are associated with this function, each of which arose from separate HGT events.

The phylogenetic incongruence method validated most of the HGT transport genes belonging to the bacterial phyla Proteobacteria, Firmicutes, Actinobacteria, Spirochaetes, and Cyanobacteria. In contrast, the genes that are involved

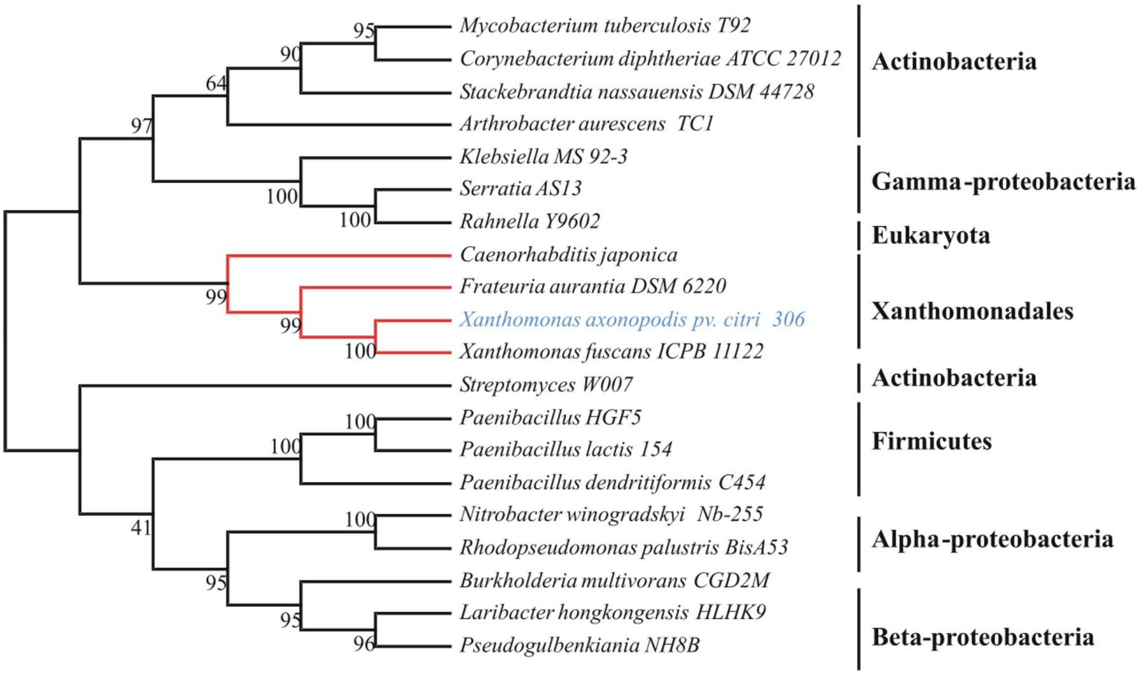

95
90
64
97
Mycobacterium tuberculosis T92
Corynebacterium diphtheriae ATCC 27012
Stackebrandtia nassauensis DSM 44728
Arthrobacter aurescens TC1
Actinobacteria
Klebsiella MS 92-3
100
Serratia AS13
100
Rahnella Y9602
Gamma-proteobacteria
Caenorhabditis japonica
Eukaryota
99
Frateuria aurantia DSM 6220
99
Xanthomonas axonopodis pv. citri 306
100
Xanthomonas fuscans ICPB 11122
Xanthomonadales
Streptomyces W007
Actinobacteria
100
Paenibacillus HGF5
100
Paenibacillus lactis 154
Paenibacillus dendritiformis C454
Firmicutes
41
100
Nitrobacter winogradskyi Nb-255
Rhodopseudomonas palustris BisA53
Alpha-proteobacteria
95
Burkholderia multivorans CGD2M
95
Laribacter hongkongensis HLHK9
96
Pseudogulbenkiania NH8B
Beta-proteobacteria

Fig. 5 Phylogenetic evidence for HGT of a *Xanthomonas axonopodis pv. citri 306* transporter gene (XAC0860: ABC transporter ATP-binding protein). Taxonomic associations are shown after genus names, with the selected strain highlighted in blue. The evolutionary history of the selected class 1 HGT gene was inferred by using the maximum likelihood method based on the JTT matrix-based model using 1000 bootstraps. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches. The tree shows the phylogenetic placement (in red) of Xanthomonadales with Eukaryota (99% bootstrap)

in protein synthesis have phylogenies similar to 16S rRNA, confirming their vertical inheritance. The finding that proteins that are related to transport are more prone to HGT is interesting and agrees with previous findings of cell surface proteins. This result is also consistent with the idea that the proteins that are located on the periphery of a metabolic network are prone to HGT (Pál et al. 2005). In that study, the proportion of *E. coli* HGT candidates (detected based on base composition) decreased from peripheral to more central metabolic networks: transport, first reaction, intermediate reactions, and biomass production. The acquisition of genes that are related to transport by horizontal transfer might directly impact the organism fitness in a certain environment, as these proteins are responsible for the exchange between the environment and the cytoplasm, affecting process such as nutrient uptake, osmolarity, and control of toxic substance entry. Thus, obtaining these functions directly from an organism already living in a certain environment would promote a selective advantage that would be kept in the genome of those individuals. On the other hand, there is most likely a strong selective pressure for protein synthesis genes that would prevent them from being replaced, given the essential roles that these proteins play in the cell and their high levels of interactions. This result agrees with the idea that the genes that are involved to metabolic pathways related to information processing are more likely to follow vertical inheritance, and are restricted and selected against when they are transferred horizontally (Jain et al. 1999).

The increased number of connections of vertically inherited genes compared to HGT genes confirms the complexity hypothesis, in which the role of proteins that depend on many interactions is most likely a limitation for the gene to be kept when transferred to a different organism. The interactome data revealed that HGT genes have few connections but, interestingly, they are not completely independent, as they can be connected among themselves. This result may be interpreted as evidence that the transfer of many of these genes does not occur individually but within gene clusters with related functions. Genes in operons could, for example, be transferred bringing selective advantages to the receptor genome, the so-called selfish operon hypothesis (Lawrence and Roth 1996; Lawrence and Ochman 1997).

Conclusion

The HGT detection method that developed in this work further recognizes that a substantial fraction of bacterial genomes is the result of horizontal transfer. This method

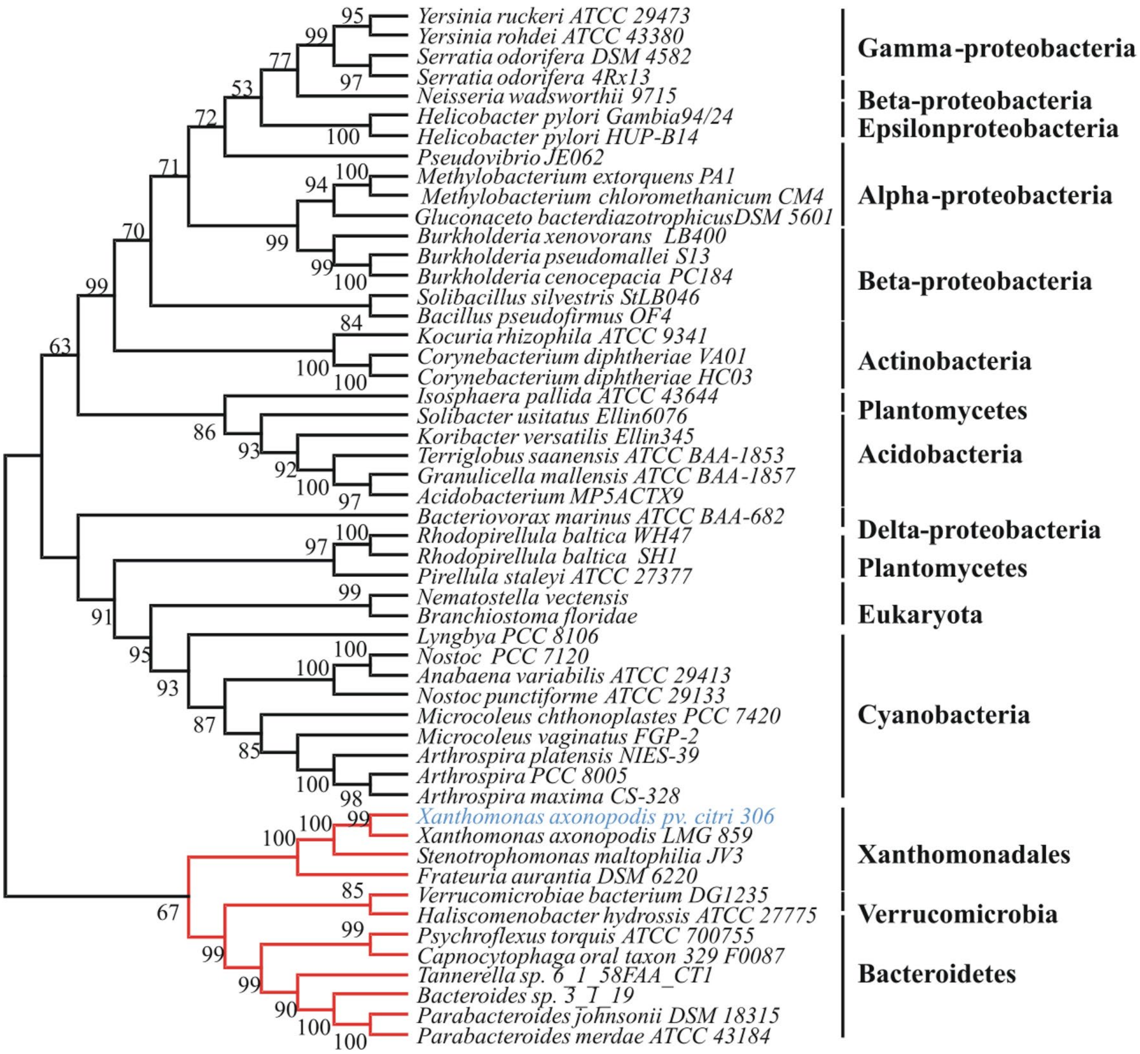

95
99
77
97
53
72
100
71
100
94
99
99
100
70
99
63
84
100
100
86
93
92
100
97
100
97
99
91
95
100
100
93
87
85
100
98
100
99
100
85
67
99
99
99
90
100
100
Yersinia ruckeri ATCC 29473
Yersinia rohdei ATCC 43380
Serratia odorifera DSM 4582
Serratia odorifera 4Rx13
Neisseria wadsworthii 9715
Helicobacter pylori Gambia94/24
Helicobacter pylori HUP-B14
Pseudovibrio JE062
Methylobacterium extorquens PA1
Methylobacterium chloromethanicum CM4
Gluconaceto bacterdiazotrophicusDSM 5601
Burkholderia xenovorans LB400
Burkholderia pseudomallei S13
Burkholderia cenocepacia PC184
Solibacillus silvestris StLB046
Bacillus pseudofirmus OF4
Kocuria rhizophila ATCC 9341
Corynebacterium diphtheriae VA01
Corynebacterium diphtheriae HC03
Isosphaera pallida ATCC 43644
Solibacter usitatus Ellin6076
Koribacter versatilis Ellin345
Terriglobus saanensis ATCC BAA-1853
Granulicella mallensis ATCC BAA-1857
Acidobacterium MP5ACTX9
Bacteriovorax marinus ATCC BAA-682
Rhodopirellula baltica WH47
Rhodopirellula baltica SH1
Pirellula staleyi ATCC 27377
Nematostella vectensis
Branchiostoma floridae
Lyngbya PCC 8106
Nostoc PCC 7120
Anabaena variabilis ATCC 29413
Nostoc punctiforme ATCC 29133
Microcoleus chthonoplastes PCC 7420
Microcoleus vaginatus FGP-2
Arthrospira platensis NIES-39
Arthrospira PCC 8005
Arthrospira maxima CS-328
Xanthomonas axonopodis pv. citri 306
Xanthomonas axonopodis LMG 859
Stenotrophomonas maltophilia JV3
Frateuria aurantia DSM 6220
Verrucomicrobiae bacterium DG1235
Haliscomenobacter hydrossis ATCC 27775
Psychroflexus torquis ATCC 700755
Capnocytophaga oral taxon 329 F0087
Tannerella sp. 6_1_58FAA_CT1
Bacteroides sp. 3_1_19
Parabacteroides johnsonii DSM 18315
Parabacteroides merdae ATCC 43184
Gamma-proteobacteria
Beta-proteobacteria
Epsilonproteobacteria
Alpha-proteobacteria
Beta-proteobacteria
Actinobacteria
Plantomycetes
Acidobacteria
Delta-proteobacteria
Plantomycetes
Eukaryota
Cyanobacteria
Xanthomonadales
Verrucomicrobia
Bacteroidetes

Fig. 6 Phylogenetic evidence for HGT of a *Xanthomonas axonopodis pv. citri 306* transporter gene (XAC0819: nucleoside transporter). The tree shows the phylogenetic placement (in red) of Xanthomonadales with Verrucomicrobia and Bacteroidetes (67% bootstrap). Details are as shown in Fig. 5

can detect genes that have a base composition similar to that of the host genome, thus including those that were incorporated a long time ago and have completed the amelioration process. Functional analyses confirm that HGT plays an important role in the environmental fitness, providing the microbes with the tools that are necessary to face the adversity and survive in a new environment.

Authors' Contribution AP, WL, and CM designed and developed the method. CM defined the strategy and supervised. HA, CP, BC, and DB analyzed the data. HA, AP, and CM wrote, discussed, and edited the manuscript. AP, HA, and BC prepared the figures. WL discussed and edited the manuscript.

Funding This project received financial support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brasília, DF, Brazil) and Comissão de Aperfeiçoamento do Ensino Superior (CAPES, Brasília, DF, Brazil) and Foundation for Research Support of the State of Sao Paulo (FAPESP, São Paulo, SP, Brazil).

Availability of Data and Materials All data generated or analyzed during this study are included in this published article and its additional files.

Compliance with Ethical Standards

Competing interests The authors declare that they have no competing interests.

References

Arama DP, Soualmia F, Lisowski V, Longevial J-F, Bosc E, Maillard LT, Martinez J, Masurier N, El Amri C (2015)

Pyrido-imidazodiazepinones as a new class of reversible inhibitors of human kallikrein 7. Eur J Med Chem 93:202–213. https://doi.org/10.1016/j.ejmech.2015.02.008

Bosi E, Fondi M, Orlandini V, Perrin E, Maida I, de Pascale D, Tutino ML, Parrilli E, Lo Giudice A, Filloux A, Fani R (2017) The pangenome of (Antarctic) *Pseudoalteromonas bacteria*: evolutionary and functional insights. BMC Genom 18:93. https://doi.org/10.1186/s12864-016-3382-y

Boucher Y, Bapteste E (2009) Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective. Bioessays 31:526–536. https://doi.org/10.1002/bies.200800216

Comas I, Moya A, Azad RK, Lawrence JG, Gonzalez-Candelas F (2006) The evolutionary origin of Xanthomonadales genomes and the nature of the horizontal gene transfer process. Mol Biol Evol 23:2049–2057. https://doi.org/10.1093/molbev/msl075

Cordero OX, Hogeweg P (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. PNAS 106:21748–21753. https://doi.org/10.1073/pnas.0907584106

Cuecas A, Kanoksilapatham W, Gonzalez JM (2017) Evidence of horizontal gene transfer by transposase gene analyses in Fervidobacterium species. PLoS ONE 12:e0173961. https://doi.org/10.1371/journal.pone.0173961

Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. PNAS 105:10039–10044. https://doi.org/10.1073/pnas.0800679105

Doolittle WF, Bapteste E (2007) Pattern pluralism and the tree of life hypothesis. PNAS 104:2043–2049. https://doi.org/10.1073/pnas.0610699104

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Felsenstein J, Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104. https://doi.org/10.1093/oxfordjournals.molbev.a025575

Gogarten JP, Townsend JP (2005) Horizontal gene transfer, genome innovation and evolution. Nat Rev Microbiol 3:679–687. https://doi.org/10.1038/nrmicro1204

Gophna U, Doolittle WF, Charlebois RL (2005) Weighted genome trees: refinements and applications. J Bacteriol 187:1305–1316. https://doi.org/10.1128/JB.187.4.1305-1316.2005

Higgins CF, Hyde SC, Mimmack MM, Gileadi U, Gill DR, Gallagher MP (1990) Binding protein-dependent transport systems. J Bioenerg Biomembr 22:571–592. https://doi.org/10.1007/BF00762962

Hily J-M, Demanèche S, Poulicard N, Tannières M, Djennane S, Beuve M, Vigne E, Demangeat G, Komar V, Gertz C, Marmonier A, Hemmer C, Vigneron S, Marais A, Candresse T, Simonet P, Lemaire O (2017) Metagenomic-based impact study of transgenic grapevine rootstock on its associated virome and soil bacteriome. Plant Biotechnol J. https://doi.org/10.1111/pbi.12761

Iqbal A, Goldfeder MB, Marques-Porto R, Asif H, Souza JG de, Faria F, Chudzinski-Tavassi AM (2017) Revisiting antithrombotic therapeutics; sculptin, a novel specific, competitive, reversible, scissile and tight binding inhibitor of thrombin. Sci Rep 7:1431. https://doi.org/10.1038/s41598-017-01486-w

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. PNAS 96:3801–3806. https://doi.org/10.1073/pnas.96.7.3801

Jeltsch A, Pingoud A (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. J Mol Evol 42:91–96

Jeong H, Nasir A (2017) A preliminary list of horizontally transferred genes in prokaryotes determined by tree reconstruction and reconciliation. Front Genet 8:112. https://doi.org/10.3389/fgene.2017.00112

Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Res 36:6688–6719. https://doi.org/10.1093/nar/gkn668

Kurland CG, Canback B, Berg OG (2003) Horizontal gene transfer: a critical view. PNAS 100:9658–9662. https://doi.org/10.1073/pnas.1632870100

Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35:3100–3108. https://doi.org/10.1093/nar/gkm160

Lassmann T, Frings O, Sonnhammer ELL (2009) Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 37:858–865. https://doi.org/10.1093/nar/gkn1006

Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44:383–397. https://doi.org/10.1007/PL00006158

Lawrence JG, Roth JR (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. Genetics 143:1843–1860

Lima WC, Varani AM, Menck CFM (2009) NAD biosynthesis evolution in bacteria: lateral gene transfer of kynurenine pathway in Xanthomonadales and Flavobacteriales. Mol Biol Evol 26:399–406. https://doi.org/10.1093/molbev/msn261

Martins-Pinheiro M, Lima WC, Asif H, Oller CA, Menck CFM (2016) Evolutionary and functional relationships of the dha regulon by genomic context analysis. PLoS ONE 11:e0150772. https://doi.org/10.1371/journal.pone.0150772

Matassi G (2017) Horizontal gene transfer drives the evolution of Rh50 permeases in prokaryotes. BMC Evol Biol 17:2. https://doi.org/10.1186/s12862-016-0850-6

McInerney JO, Cotton JA, Pisani D (2008) The prokaryotic tree of life: past, present… and future? Trends Ecol Evol 23:276–281. https://doi.org/10.1016/j.tree.2008.01.008

Meehan CJ, Beiko RG (2012) Lateral gene transfer of an ABC transporter complex between major constituents of the human gut microbiome. BMC Microbiol 12:248. https://doi.org/10.1186/1471-2180-12-248

Mrázek J, Karlin S (1999) Detecting alien genes in bacterial genomes. Ann NY Acad Sci 870:314–329

Mykowiecka A, Szczesny P, Gorecki P (2017) Inferring gene-species assignments in the presence of horizontal gene transfer. IEEE/ACM Trans Comput Biol Bioinform. https://doi.org/10.1109/TCBB.2017.2707083

Nakamura Y, Itoh T, Matsuda H, Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–766. https://doi.org/10.1038/ng1381

Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37:1372–1375. https://doi.org/10.1038/ng1686

Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The comprehensive microbial resource. Nucleic Acids Res 29:123–125

Podell S, Gaasterland T (2007) DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol 8:R16. https://doi.org/10.1186/gb-2007-8-2-r16

Podell S, Gaasterland T, Allen EE (2008) A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. BMC Bioinformatics 9:419. https://doi.org/10.1186/1471-2105-9-419

Ragan MA (2001) Detection of lateral gene transfer among microbial genomes. Curr Opin Genet Dev 11:620–626

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA 95:6239–6244

Ros VI, Hurst GD (2009) Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant? BMC Biol 7:20. https://doi.org/10.1186/1741-7007-7-20

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729. https://doi.org/10.1093/molbev/mst197

Williams KP, Gillespie JJ, Sobral BWS, Nordberg EK, Snyder EE, Shallom JM, Dickerman AW (2010) Phylogeny of gammaproteobacteria. J Bacteriol 192:2305–2314. https://doi.org/10.1128/JB.01480-09

Yin Z, Zhu B, Feng H, Huang L (2016) Horizontal gene transfer drives adaptive colonization of apple trees by the fungal pathogen *Valsa mali*. Sci Rep 6:33129. https://doi.org/10.1038/srep33129

Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, Nelson KE, Nesbø CL, Doolittle WF, Gogarten JP, Noll KM (2009) On the chimeric nature, thermophilic origin, and phylogenetic placement of the thermotogales. Proc Natl Acad Sci USA 106:5865–5870. https://doi.org/10.1073/pnas.0901260106