How horizontal gene transfer builds prokaryote genomes?

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

* Both authors contributed equally to this manuscript

** Corresponding author

1. Department of Microbiology, Institute of Biomedical Sciences, University of Sao Paulo, Sao Paulo, Brazil.

2. Department of Statistics, Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil.

3. Center of Biotechnology, Department of Molecular Biology and Biotechnology, Federal University of Rio Grande do Sul, RS, Brazil

4. Department of Cell Physiology and Metabolism, Faculty of Medicine, University of Geneva, Geneva 1211, Switzerland

Corresponding author:
CFMM, Av. Prof. Lineu Prestes, 1374, Dept. of Microbiology, University of São Paulo, 05508-000, São Paulo, SP, Brazil.
Email for contact: cfmmenck@usp.br

Running title: Horizontal gene transfer in Prokaryotes

ABSTRACT

Horizontal gene transfer (HGT) is considered to be an innovative force in the genesis and evolution of species. However, the frequency of gene transfer varies when comparing different genomes and depends on the genes' roles in cell metabolism. In the present work, we developed a new large-scale analysis for the detection of genes potentially originated by HGT in 697 prokaryote genomes. Basically, this analysis is based on the similarity (BLAST) distance among homologs, in relation to the distance observed for the 16S rRNA gene on those organisms. When there is a discrepancy on these two-similarity distances, the genes were classified as HGT candidates. Based on this approach we estimated that approximately 15% of the genes of each prokaryote genome have an HGT origin. The methodology employed was strongly supported for evolution relationships among genes, tested by direct phylogenetic analysis of many of the HGT candidate genes. Moreover, interactome studies were performed for these genes considering the *Escherichia coli* (strain W3110) genome, where interaction data is well available. The results clearly show that the proteins encoded by HGT show much lower levels of interaction with other proteins, when compared to those that were detected as originated by vertical inheritance, confirming the complexity hypothesis. Therefore, the number of protein partners imposes limitations for this process. Moreover, a detailed function classification of the genes analyzed confirms that genes related to protein translation have a vertical inheritance. As expected, mobile genetic elements were confirmed as involved in HGT, but transport and binding proteins also have a strong bias for being acquired by horizontal transfer. As these genes are directly related to the cells exchange with environment, their transfer most likely contribute for their successful adaptation thorough evolution.

Keywords: Horizontal gene transfer, Evolution, Phylogenetic tree incongruence, Transport proteins.

INTRODUCTION

Horizontal (or lateral) gene transfer (HGT) is any process in which an organism incorporates genetic material from another organism of different species. This contrasts with the normal process of vertical inheritance where 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 (Lima et al., 2009; Ros and Hurst, 2009). The transfer of genetic material occurs more frequently from closely related species, although it also occurs between strains far outside the closed gene pool (Koonin and Wolf, 2008).

HGT has been shown to be widespread across the prokaryotic lineage, but the degree of the effect that it exerts in reconstructing the evolutionary history of organisms is still controversial (Boucher and Bapteste, 2009); (McInerney et al., 2008). Moreover, some researchers argue that HGT are constrained by important selective barriers, and have limited influence in the evolution of modern organisms (Kurland et al., 2003). On the other hand, (Dagan et al., 2008) have pointed out that it is entirely plausible that HGT has affected every single prokaryotic gene over the full span of evolutionary history and depending on the different approach for the gene choice, the rate of interest or the method used for reconstruction, a reticulated network may be generated rather than a vertical tree. The latter view is highly embraced by many researchers, who argue that a strictly vertical tree-like model can no longer explain life evolution.

There is increasing evidence that whenever HGT is frequent enough, different part of an organism's genome reflect different evolutionary histories (Doolittle and Bapteste, 2007); (Gogarten and Townsend, 2005). For instance, ribosomal components phylogenetically group Thermotogales with Aquificae while in whole-genome phylogenies, the Thermotogales clustered together with *clostridia* and *bacilli* (Gophna et al., 2005); (Zhaxybayeva et al., 2009). In addition, the increasing evidence of horizontal acquisition of genes based on discrepancies in tree topologies in prokaryotic genomes has been reported in several bacterial phyla (Qian and Parker, 2002);(Williams et al., 2010).

The function of genes prone to HGT is often reported as those related to operational (housekeeping) roles in the cell, while genes related to informational (DNA replication, RNA transcription and protein translation) probably impose evolutionary restrictions to HGT, and are

rarely found as transferred (Rivera et al., 1998). On the other hand, 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 complexity hypothesis (Jain et al., 1999). However, evidence exists that genes related to primary metabolism can be transferred among relatively distant clades (Lima et al., 2009). This is the case, for example, of arginine biosynthesis operon replacement (Lima and Menck, 2008) and NAD biosynthesis pathways (Lima et al., 2009), which are shared in some bacteria of the Xhantomonadales and Flavibacteriales groups and eukaryotes. The dramatic discordance within a set of taxa supports the notion that tree-like vertical inheritance-based view of evolution is insufficient while unfolding prokaryotic relationships.

In this article, we present an *in silico* method developed for large-scale detection of HGT. The main operating principle in classifying a gene as involved in HGT is based on BLAST analyses, when the gene has high similarity hits to a phylogenetically distant taxon, identified by its 16S rRNA similarity distance. Interestingly, the analyses indicated that approximately 15% of the genes of each prokaryote genome are acquired by HGT, and the larger the genome the larger the number of HGT candidate genes. HGT candidates were further tested by phylogenetic validation purposes, for proteins involved in transport of macromolecules and protein synthesis in the bacterial phyla *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Spirochaetes*, *Cyanobacteria*, confirming the initial classification of these genes as potentially a result of HGT. Moreover, interactome analyses for genes from a well-studied genome (*Escherichia coli,* strain W3110) clearly indicate genes prone to HGT have less interaction than genes that keep vertical inheritance. All HGT candidate genes were mapped to the functional categories, and the results confirm that very low frequency of genes related to protein translation, DNA metabolism or RNA transcription are derived from HGT. But, interestingly among the other functional categories where genes are more prone to HGT there is a clear preference of mobile genetic elements (as expected) and transport and binding proteins, followed by the categories of pathogenicity and central intermediary metabolism. This work thus provides a large-scale support for the identification of genes that were potentially acquired by interspecies genetic exchange, and how these events add on the construction of prokaryotic genomes.

Comment [CM1]: Bruno and Diego, please check if you agree with this new sentence for Introduction.

Methods

Genomes dataset processing

The complete sequences of prokaryote 697 genomes were retrieved from Omniome database version 22, which is the part of Comprehensive Microbial Resource (CMR) (Peterson et al., 2001) and imported into the MYSQL database manager. The 16S rRNA distances of the genomes studied are calculated as follows (i) aligning 16S rRNA gene of the selected genomes with the program Kalign2 (Lassmann et al., 2009) using gap penalty of 3.94 (ii) bootstrapping (100 replicates) with Seqboot program of PHYLIP package (iii) for each replicate, the distances between the 16S rRNA are calculated using DNAdist program (F84 distance model) of PHYLIP package (Felsenstein and Churchill, 1996).

Prediction of genes involved in HGT

The genes under study were classified as vertically or horizontally transferred based on their target Blast-Extend-Repraze (BER) searches (in order of decreasing score) and two user defined parameters: d_{self} (0.105) and $d_{distant}$ (0.2) (Figure S1). The parameter d_{self} exclude the genomes that are very close to the genome under study whereas $d_{distant}$ is chosen by the user as the minimum distance of interest for evaluating horizontal transfer. Let X be the target BER of a gene under study, $\overline{d}_X$ average between the bootstrap replicates, the 16S rRNA distances from the organisms under study and the target X, and A is the first target "non-self" i.e. the first target which satisfies $\overline{d}_A > d_{self}$. The gene is considered: **Typical** (most probably of vertical origin) if $\overline{d}_A \leq d_{distant}$, **Atypical** (HGT candidate) if $\overline{d}_A > d_{distant}$, and **Undetermined** if there is no such target A with $\overline{d}_A > d_{self}$. The HGT genes were further classified into class 1 and class 2 (discussed later).

In order to reduce the computational cost of verification, we used an indirect approach which assumes that distances of bootstrap replicates is normally distributed, a condition that was subsequently verified with Shapiro-Wilk normality test for targets with $\overline{d}_X > 0.003$. Thus, the distance between the organisms under study and X range can be expressed by random variable d_X, with distribution: $d_X \sim N(\overline{d}_X, s^2_X)$ where s^2_X is the variance of bootstrap replicates. In the indirect approach, the requirement to assign a gene into subsequent class is verified if $P(d_B < d_A) \geq 0.99$. The probability $(d_B < d_A)$ is estimated based on the fact that: $d_A - d_B \sim N(\overline{d}_A - \overline{d}_B, s^2_A +$

s^2_B), according to the distribution property of the difference between two independent normally distributed random variables.

Enrichment analysis of functional gene categories potentially involved in HGT

In Omniome database, genes are annotated according to several functional and sub functional categories. We did a slight modification in these functional categories and carried out enrichment analysis for typical and atypical genes. Modifications made to the functional categories include creation of the category "Pathogenesis, toxin production and resistance", which was originally within the category "Cellular processes", creation of the category DNA restriction/Modification, which was originally in the category "DNA metabolism". The enrichment factor (or depletion) and statistical significance (p values) of atypical genes in each functional category was estimated using two-tailed Fisher's test (fisher.test function in R; with enrichment factor $\geq$ 1.5). The enrichment factor is the rate of change of the ratio between the annotated number of genes in the category X and non-annotated in the same category. In order to control the rate of false positive, we calculated the q values (< 0.01) from the p values obtained as proposed by Storey and Tibshirani (Storey and Tibshirani, 2003). The q value expresses the expected proportion of false positives among the results that were considered significant.

Phylogenetic validation

To further confirm the HGT candidates identified, we carried out phylogenetic validation for some genes. For phylogenetic reconstruction, sequence similarity analyses of the selected protein sequences were performed using BLASTP program from Expasy proteomics server (E-value < 10^{-04} and percent of similarity $\geq$ 30 %). Sequences were aligned using ClustalX 2.0 algorithm. The alignments produced were then refined in order to remove regions that are hyper variable or contain gaps with the Genedoc program (Nicholas et al., 1997). Evolutionary distances were computed using MEGA 4.0 with the Jones-Taylor-Thornton algorithm. Neighbor-Joining (NJ) algorithm was used to generate distance-based phylogenetic trees. Bootstrap test with 1,000 replicates is performed using MEGA 4.0. Bootstrap support larger than 50% is considered to identify supported nodes. Phylogenetic trees have been drawn with orthologs from the main representative groups, with at least five organisms from each group (except the cases

where sequence coverage is extremely low). Trees were visualized using MEGA 4.0 (Tamura et al., 2007).

Interactome data mining and the design of the interactomes

To design the interactomes and to elucidate the interplay genes with vertical inheritance or HGT candidates in a topological context, the metasearch engine STRING 9.1 [http://string-db.org] (Jensen et al., 2009) ;(Snel et al., 2000) was used. All genes classified as typical or atypical for the *E.coli* (strain W3110) genome, as well genes related to transport of macromolecules and protein synthesis, were used as the initial seeds for network prospecting in STRING. Each connection (edge) possesses a degree of confidence between 0 and 1.0 (with 1.0 indicating the highest confidence). The parameters used to prospect the networks in STRING software 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 gathered were analyzed with Cytoscape 2.8.2 (Shannon et al., 2003).

Comment [CM2]: Bruno and Diego, VGT is not used in this manuscript for convenience.... So I will try to keep that out.

RESULTS

Detecting potential HGT genes in prokaryotic genomes

Large-scale search of potential HGT genes was performed based on 697 microbial (Bacteria and Archeae) genomes. The basic assumption of this search is that HGT genes would provide BLAST results in which the order of similarity with orthologous protein sequences differed from the 16S rRNA gene distances (which were considered basically as the product of vertical inheritance). Genes with **typical** distance relationship (BLAST order from target similar to 16S rRNA distance, $\overline{d}_A \leq d_{distant}$) were considered as derived from vertical inheritance. When the gene was observed only in closely related species they were considered **undetermined**. The **atypical** genes, potentially acquired by HGT, were those that the 16S rRNA distance was higher than expected for the order of orthologous genes, as defined by BLAST results ($\overline{d}_A > d_{self}$). Moreover, these were classified in two different categories: class 1, where there were other close species which also carry an orthologous gene with order related to 16S rRNA distance, and class 2, where the potentially HGT gene appears as a novel gene with the 16S rRNA distance above a certain threshold ($d_{distant}$ 0.2), corresponding for species of a different order. A gene assigned to

class 1 is potentially involved in replacing the orthologous gene by horizontal transfer, while a gene assigned to class 2 may be involved in the acquisition of a novel biological function into the recipient genome. Figure S1 (Additional file 1) illustrates each of these categories and how they were classified.

Table 1 shows the number of genes analyzed, out of 697 genomes, and the fraction (18%) that was indicated as atypical (potentially acquired by HGT). As shown below, the number of phylogenetic neighbors (i.e. the number of genomes with distance between d_{self} and $d_{distant}$) influences in the determination of atypical genes (mainly due to the identification of class 2 genes), thus Table 1 also indicates the number and the fraction of atypical genes (15%) where more than twenty neighbors were available.

Features of the HGT prokaryote genes

Formatted: Font: Bold, Italic

The BLAST analysis for the search of HGT genes is very sensitive to the number of phylogenetic closely related orthologous. This is clearly shown when the number of atypical genes was plotted considering the number of phylogenetic neighbor genomes: when few neighbors were observed, there is a tendency of high variation in the number of atypical genes per genome (Figure 1A). Not surprisingly, most of the atypical genes in less represented genomes are class 2, and those assigned as class 1 increase with the number of neighbor genomes (Figure 1B). On the other hand, when more than 20 neighbor genomes exist, the frequency of atypical genes stabilize with an average of approximately 15% per genome (Figure 1A), with most of them from class 1 (72%, Figure 1B). From now on, the genes considered for this study refer only to those genomes with more than 20 phylogenetically related neighbors.

The presence of HGT candidate genes was compared in relation to the number of genes in prokaryotic genomes, and the results are shown in Figure 1C. There is a clear correlation between the number of HGT genes and the number of genes in the genome. Thus, the larger the genome the higher the number of HGT genes. Curiously, a group of genomes have very few number of HGT genes. A closer look indicates these organisms correspond to intracellular symbionts, with a very reduced genome size. Thus, these results show that, as expected, under these conditions, evolution promotes reduced opportunities for the uptake of new genes by HGT and strong selection for a small genome.

Functions of proteins encoded by HGT prokaryote genes

The functional identity of HGT genes is an essential question to understand how transfer contributes to evolution. The categorization was performed based on the OMNIOME database, with few modifications. Only atypical and typical genes were considered for the analysis and in a defined category, the genes were checked for an enrichment factor, which identifies if it presented a higher (or lower) proportion of genes, in relation to the rate of genes in that category considering the global number of genes. The global findings are presented in Table 2, where 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 (more than 30% of the genes of that category) for HGT candidate genes. On the other hand, protein synthesis, transcription and DNA metabolism are the categories enriched in genes with vertical inheritance (typical genes) (less than 15% of atypical genes).

The category enrichment was also analyzed considering individual genomes, and the results are shown in Figure 2, with each genome grouped by phylogenetically related clades. In this figure, the red color represents significant enrichment for HGT candidate genes (atypical genes) and blue color for vertical inheritance (typical genes). White labeling represents genomes where the category does not have any significant enrichment. Basically, the pattern is not very different from what is observed on the global analysis, except that in some categories the low number of genes implies loss of significance. This is the case of DNA restriction/modification genes, highly significant in the global analysis (Table 2), but not when individual genomes are analyzed. Interestingly, at specific categories (columns), in general, the same color (blue or red) is prevalent, reinforcing that these functions may be related either with HGT or vertical inheritance. These results show significant enrichment of many categories, but two of them call attention as a general phenomena in most of the genomes analyzed: protein synthesis, which is known to be related to vertical inheritance (blue) and transport and binding protein for potentially transferred genes (red).

Interactome analyses indicate proteins encoded by HGT genes have fewer interactions.

This data clearly confirms that genes related to informational metabolism have restricted possibilities to be transferred horizontally, and those related to more peripheric metabolism are prone to HGT. However, another interesting aspect is that HGT genes have been proposed as genes with lower number of interactions, a hypothesis commonly referred as the complexity

hypothesis (Jain et al., 1999). To investigate the topological aspects of connections between HGT and vertically inherited genes, we considered the *E.coli,* strain W3110 genome, where there is a good amount of information on the protein-protein interactions. Thus, genes classified as typical (vertical inheritance, yellow nodes) or atypical (HGT, blue nodes) were prospected by the STRING metasearch engine, and the results are shown in Figure 3A. When the genes were considered together, a strongly connected network is clearly observed, dominated by the high number of typical genes. However, if the groups were considered separately, the interaction reveals interesting aspects that differ them. First, it is clear that the number of connections is much higher for typical genes that show a total of 14,332 connections between 1,935 genes, a ratio of 7.4 connections per gene (Figure 3B). However, for those classified as atypical (HGT candidate genes), there is a total of 799 connections between 557 genes, a ratio of 1.4 (Figure 3C). This data strongly support the idea that the number of interactions a protein establishes in the cell restricts the chances of its gene to be transferred to and maintained in a different host genome during evolution. Moreover, the results with HGT genes also show that, although they have a lower number of interactions, it is clear that 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 (Figure 4) the higher level of connections for the genes with vertical inheritance is clearly observed when compared to the HGT candidate genes.

Phylogenetic validation

The data obtained in this work assumes BLAST searches can provide information on evolutionary distances. Although this is correct in many cases, protein domains may affect directly the BLAST analyses, yielding results that are biased and could promote false results. Thus, phylogenetic incongruence method was used to further test several genes detected as atypical. Six different bacterial groups were tested and the analyses were focused on genes related to transport and binding proteins (Table 3). Besides HGT candidate genes, as negative controls we retrieved some genes involved in protein synthesis that were identified as vertically inherited (data not shown). Figures 5 and 6 illustrate some of these phylogenies, and clearly confirm that by showing two different *Xanthomonas axonopodis pv. citri 306* transporter genes have eukaryote (Figure 5), or verrucomicrobia and bacteroidetes (Figure 6) as clade neighbors.

Comment [CM3]: Bruno, veja o comentário interessante da Wanessa: It'd be interesting to see this with the same N. If 557 VGT genes are picked at random, the number of connections would still be around 7, or it'd drop to around 2.0?

Comment [CM4]: Bruno, check if you think this sentence is more clear.

Comment [B.C.F.5]: Menck, should we add information about the clusters for Support Information?

Comment [CM6]: I guess we do not need to show clusters for the moment. If the reviewers ask for we can complemente in supplementary material.

Other transporter genes of the Xanthomonadales order were also investigated and the phylogenies indicate they group with evolutionarily unrelated bacterial genomes (some of these genes are shown in Additional file (Figures 2S and 3S). In addition, several other genes of the same category, but from different orders, were also tested and some are described and illustrated in the Additional file. Basically, the results confirm that genes classified as of potential HGT origin are grouped in clades that present phylogenetically distinct orthologs Figures 4S to Figure 16S (Additional file).

Formatted: English (U.K.)

In total, we have analyzed a total of 134 genes that were classified as atypical in the BLAST search, and 69 (51%) confirmed unambiguously the origin of these genes as acquired by HGT. The remaining 65 (49%) 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. On the other hand, the phylogenetic distribution for proteins related to protein synthesis confirmed (of 25 genes, 90-100%) unambiguously their vertical inheritance, as predicted by the first large scale screening. Therefore, the results support that the BLAST searches method was capable of successfully classify the genes that were acquired by horizontal transfer, or vertically inherited.

DISCUSSION

Analysis of complete genome sequences has enormously helped in addressing many issues concerning microbe evolution. One such area is the acquisition of new genes by HGT. To address that a particular gene owes its presence in a particular genome due to HGT, several methods have been proposed (Eisen, 2000);(Ragan, 2001). However, the variability in the outcome of these methods demands the need to develop new tools, which should evidence the gene evolution relationship to the host genome, and not only 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 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 megaplamids which can facilitate higher rate of transmembrane DNA translocation, providing the organisms with more flexibility to fit in different environments (Cordero and Hogeweg, 2009). Interestingly, intracellular symbiont

bacteria with very small genomes are those with lower frequency of HGT candidate genes. This is probably due to a dependence on the metabolic processes of the host cells, and a tendency to lose the extra genes.

The method clearly is subject to the number of close evolutionary neighbors that are found on the genome bank analyzed, probably due to the low quality in the HGT prediction when few neighbors are found. However, when a genome has more than 20 neighbors in the database, then the frequency of HGT genes detected varies, mainly, in the range of 5 to 25%, with an average of 15% of the total number of genes. Most of them are class 1 (72% of 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 for this type of analysis) corresponds to novel biological functions acquired by the recipient organism.

Nakamura et al. (2004)(Nakamura et al., 2004), working with 116 prokaryotic genomes, reported several functional biases of HGT genes, identifying functions that are more prone to transfer. That work identified HGT genes based on gene composition, which detects mainly recent events, failing to detect HGT cases where sequences have completed the amelioration process. The functions mostly observed in genes detected as originated HGT were linked to DNA mobility, pathogenicity, DNA binding and cell surface. In this work, to map the HGT genes within functional groups, categories were based on the OMNIOME database. We found that the categories of DNA restriction/modification are those with the highest number of HGT genes. This result is in agreement with earlier study that implies that HGT has a greater influence on the distribution and evolution of DNA restriction/modification system, though the approach used for the analysis was based on biased codon usage (Jeltsch and Pingoud, 1996). The second highest category that was found enriched in HGT genes was transport and protein binding. This functional category includes well-characterized members of the super family of transporters, which are closely related in term of structure, function and evolutionary origin. These transporters generally known as ABC transporters can transport different substrates such as inorganic ions, amino acids, sugars, polysaccharides and even proteins (Higgins et al., 1990) and are mostly found in the cell surface. Recently, ABC transporter systems were also found as a main hotspot of gene transfer within the human gut microbiome (Meehan and Beiko, 2012). In fact, analysis of nickel and iron transport complex in many bacteria revealed several operons

associated with this function, each of which arose from separate HGT events (Mira et al., 2004); (Meehan and Beiko, 2012).

The enrichment of HGT genes in each category analysis were also identified for individual genomes. This provided us with a nice view of the categories that are prone to be transferred during evolution and those that are clearly limited for HGT and are found only as derived from the more common vertical inheritance. In this analysis, the less common genes, such as those in the DNA restriction/modification category, were not detected. However, the proteins of the category of transport were also clearly found as highly subject to transfer among different organisms, while proteins related to protein synthesis were observed to be refractory to HGT, as expected. Phylogenetic incongruence method validated most of the HGT candidates transport genes belonging to the bacterial phyla Proteobacteria, Firmicutes, Actinobacteria, Spirochaetes, and Cyanobacteria. As negative controls for HGT, genes from categories not linked to HGT (protein synthesis for example) confirmed their vertical inheritance. The phylogenetic trees obtained provide support to the model that transport proteins are in fact prone to HGT in prokarya, while those related to protein synthesis are very limited to vertical inheritance.

The finding that proteins related to transport are prone to HGT is interesting and goes in the same trend as previous findings of proteins in the cell surface. This is also consistent with the idea that proteins located on the periphery of metabolic network are prone to HGT (Pal et al., 2005). In that work, the proportion of *E.coli* HGT candidates (detected based on base composition) decreased from periphery to more central metabolic networks: transport, first reaction, intermediate reactions and biomass production. Therefore, the acquisition of genes related to transport occurs by horizontal transfer probably because they are directly responsible for the organism fitness in a certain environment. These proteins are those responsible for the exchange between the environment and the cytoplasm affecting process such as nutrient uptake, osmolarity, control of toxic substance entry, etc. 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, regarding the category of protein synthesis, there is probably a strong selective pressure for genes that would hamper them to be replaced, because of the essential roles these proteins play within the cell and the high levels of interactions these proteins are subject

performing their functions. This is in agreement with the idea that genes related to metabolic pathways related to information processing are more likely to follow vertical inheritance, and being restricted, and selected against, when they are transferred horizontally (Jain et al., 1999). The elevate number of connections of the genes classified as those vertically inherited, when compared to HGT genes, basically confirms the complexity hypothesis, where the role of proteins dependent on many interactions is probably a limitation for the gene to be kept when transferred to a different organism. Curiously, however, the interatoma data revealed that HGT genes have few connections, but they are not completely independent, as they can be connected among themselves. We believe this may be interpreted as evidence that the transfer of many of these genes do not occur individually, but within gene clusters with related functions. This could be the case of operons that would be transferred with selective advantages, what has been called selfish operons (Lawrence and Roth, 1996).

In summary, the HGT detection method developed in this work further recognizes that a substantial fraction of bacterial genomes are the result of horizontal transfer. The method can detect genes that have base composition similar to the whole host genome, thus including those that were incorporated a long time ago and have completed the amelioration process. Functional analysis confirm that HGT played an important role on the fitness to the surrounding it lives, providing the microbes with the tools necessary to face the adversities and survive in new environment.

Acknowledgements: This project received financial support from CAPES and CNPq (Brasília, DF Brazil) and FAPESP (São Paulo, Brazil).

References

Boucher, Y., and E. Bapteste, 2009, Revisiting the concept of lineage in prokaryotes: a phylogenetic perspective: Bioessays, v. 31, p. 526-36.

Cordero, O. X., and P. Hogeweg, 2009, The impact of long-distance horizontal gene transfer on prokaryotic genome size: Proc Natl Acad Sci U S A, v. 106, p. 21748-53.

Dagan, T., Y. Artzy-Randrup, and W. Martin, 2008, Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution: Proc Natl Acad Sci U S A, v. 105, p. 10039-44.

Doolittle, W. F., and E. Bapteste, 2007, Pattern pluralism and the Tree of Life hypothesis: Proc Natl Acad Sci U S A, v. 104, p. 2043-9.

Eisen, J. A., 2000, Horizontal gene transfer among microbial genomes: new insights from complete genome analysis: Curr Opin Genet Dev, v. 10, p. 606-11.

Felsenstein, J., and G. A. Churchill, 1996, A Hidden Markov Model approach to variation among sites in rate of evolution: Mol Biol Evol, v. 13, p. 93-104.

Gogarten, J. P., and J. P. Townsend, 2005, Horizontal gene transfer, genome innovation and evolution: Nat Rev Microbiol, v. 3, p. 679-87.

Gophna, U., W. F. Doolittle, and R. L. Charlebois, 2005, Weighted genome trees: refinements and applications: J Bacteriol, v. 187, p. 1305-16.

Higgins, C. F., S. C. Hyde, M. M. Mimmack, U. Gileadi, D. R. Gill, and M. P. Gallagher, 1990, BINDING PROTEIN-DEPENDENT TRANSPORT-SYSTEMS: Journal of Bioenergetics and Biomembranes, v. 22, p. 571-592.

Jain, R., M. C. Rivera, and J. A. Lake, 1999, Horizontal gene transfer among genomes: The complexity hypothesis: Proceedings of the National Academy of Sciences of the United States of America, v. 96, p. 3801-3806.

Jeltsch, A., and A. Pingoud, 1996, Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems: Journal of Molecular Evolution, v. 42, p. 91-96.

Jensen, L. J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, 2009, STRING 8--a global view on

proteins and their functional interactions in 630 organisms: Nucleic Acids Res, v. 37, p. D412-6.

Koonin, E. V., and Y. I. Wolf, 2008, Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world: Nucleic Acids Research, v. 36, p. 6688-6719.

Kurland, C. G., B. Canback, and O. G. Berg, 2003, Horizontal gene transfer: A critical view: Proceedings of the National Academy of Sciences of the United States of America, v. 100, p. 9658-9662.

Lassmann, T., O. Frings, and E. L. L. Sonnhammer, 2009, Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features: Nucleic Acids Research, v. 37, p. 858-865.

Lawrence, J. G., and J. R. Roth, 1996, Selfish operons: horizontal transfer may drive the evolution of gene clusters: Genetics, v. 143, p. 1843-60.

Lima, W. C., and C. F. Menck, 2008, Replacement of the arginine biosynthesis operon in Xanthomonadales by lateral gene transfer: J Mol Evol, v. 66, p. 266-75.

Lima, W. C., A. M. Varani, and C. F. M. Menck, 2009, NAD Biosynthesis Evolution in Bacteria: Lateral Gene Transfer of Kynurenine Pathway in Xanthomonadales and Flavobacteriales: Molecular Biology and Evolution, v. 26, p. 399-406.

McInerney, J. O., J. A. Cotton, and D. Pisani, 2008, The prokaryotic tree of life: past, present... and future?: Trends Ecol Evol, v. 23, p. 276-81.

Meehan, C. J., and R. G. Beiko, 2012, Lateral gene transfer of an ABC transporter complex between major constituents of the human gut microbiome: Bmc Microbiology, v. 12.

Mira, A., R. Pushker, B. A. Legault, D. Moreira, and F. Rodriguez-Valera, 2004, Evolutionary relationships of Fusobacterium nucleatum based on phylogenetic analysis and comparative genomics: Bmc Evolutionary Biology, v. 4.

Nakamura, Y., T. Itoh, H. Matsuda, and T. Gojobori, 2004, Biased biological functions of horizontally transferred genes in prokaryotic genomes: Nature Genetics, v. 36, p. 760-766.

Nicholas, K., H. Nicholas, and D. Deerfield, 1997, GeneDoc: analysis and visualization of genetic variation, *in* H. B. J. Nicholas, ed.

Pal, C., B. Papp, and M. J. Lercher, 2005, Adaptive evolution of bacterial metabolic networks by horizontal gene transfer: Nature Genetics, v. 37, p. 1372-1375.

Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White, 2001, The Comprehensive Microbial Resource: Nucleic Acids Research, v. 29, p. 123-125.

Qian, J. H., and M. A. Parker, 2002, Contrasting nifD and ribosomal gene relationships among Mesorhizobium from Lotus oroboides in Northern Mexico: Systematic and Applied Microbiology, v. 25, p. 68-73.

Ragan, M. A., 2001, Detection of lateral gene transfer among microbial genomes: Current Opinion in Genetics & Development, v. 11, p. 620-626.

Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake, 1998, Genomic evidence for two functionally distinct gene classes: Proceedings of the National Academy of Sciences of the United States of America, v. 95, p. 6239-6244.

Ros, V. I. D., and G. D. D. Hurst, 2009, Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant?: Bmc Biology, v. 7.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, 2003, Cytoscape: a software environment for integrated models of biomolecular interaction networks: Genome Res, v. 13, p. 2498-504.

Snel, B., G. Lehmann, P. Bork, and M. A. Huynen, 2000, STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene: Nucleic Acids Res, v. 28, p. 3442-4.

Storey, J. D., and R. Tibshirani, 2003, Statistical significance for genomewide studies: Proceedings of the National Academy of Sciences of the United States of America, v. 100, p. 9440-9445.

Tamura, K., J. Dudley, M. Nei, and S. Kumar, 2007, MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0: Molecular Biology and Evolution, v. 24, p. 1596-1599.

Williams, K. P., J. J. Gillespie, B. W. S. Sobral, E. K. Nordberg, E. E. Snyder, J. M. Shallom, and A. W. Dickerman, 2010, Phylogeny of Gammaproteobacteria: Journal of Bacteriology, v. 192, p. 2305-2314.

Zhaxybayeva, O., K. S. Swithers, P. Lapierre, G. P. Fournier, D. M. Bickhart, R. T. DeBoy, K. E. Nelson, C. L. Nesbo, W. F. Doolittle, J. P. Gogarten, and K. M. Noll, 2009, On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales:

Proceedings of the National Academy of Sciences of the United States of America, v. 106, p. 5865-5870.

FIGURE LEGENDS

Figure 1. Overview of HGT candidate genes in the prokaryote genomes (A) Fraction of genes 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 used in this study. (B) Fraction of HGT candidate genes assigned to the class 1. The black solid line is a LOESS fit to the data. (C) Fraction of genes classified as HGT candidate genes as a function of the number of genes in the genome. rho: Spearman rank correlation. The organisms annotated with the red ellipse are all insect intracellular endosymbionts.

Figure 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 (blue) of the HGT proportion in that category is significantly higher (lower) than for genes not in that category. Statistical significance is calculated using a chi-squared test with 10^6 Monte Carlo simulations, and requiring false discovery rate < 0.01 (Benjamini-Hochberg method) and enrichment (or depletion) factor >= 1.5.

Figure 3. Vertical inheritance and HGT networks showing the interplay between typical and atypical genes of *E.coli* strain W3110 genome; genes detected as with vertical inheritance are illustrated in yellow and genes detected as HGT were colored blue. A) Vertical inheritance - HGT global network, containing 2522 connected genes and 111 not connected. This global network was subdivided in two networks B) for those of vertical inheritance, which contains 1965 genes, and C) HGT, containing 557 genes.

Figure 4. Genes related to protein synthesis and transport and binding protein in *E. coli* strain W3110. A) All vertically inherited genes (yellow) related to protein synthesis (diamond shaped); B) All HGT genes related (blue) to transport and binding protein function (red bordered nodes); C) The genes in the subnetwork in (A) were expanded by all immediately connected genes, showing that vertically inherited genes have a preference for its own class and; D) The genes in subnetwork in B were also expanded by all immediately connected genes, showing that HGT genes have preference for vertically inherited genes as well.

Figure 5. Phylogenetic evidence for HGT of the *Xanthomonas axonopodis pv. citri 306* transporter gene (XAC0860: ABC transporter ATP-binding protein). Taxonomic associations are shown after genus names, with selected strain highlighted in blue. The distance tree of selected

class1-HGT gene was computed by Neighbor-Joining (NJ) method. Numbers on the node correspond to bootstrap replicates. The phylogenetic tree showed a phylogenetic placement (in red) of *Xanthomonas axonopodis pv. citri 306* with eukaryota (with 100% of bootstrap).

Figure 6. Phylogenetic evidence for HGT of the *Xanthomonas axonopodis pv. citri 306* transporter gene (XAC0819: Nucleoside transporter). The phylogenetic tree showed a phylogenetic placement (in red) of *Xanthomonas axonopodis pv. citri 306* with verrucomicrobia and bacteroidetes (with 73% of bootstrap). Details as in Figure 5.

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

Table 2. Global proportion of HGT candidate genes grouped by functional categories, for the genomes with ≥ 20 phylogenetic neighbors. Enrichment factor is the ratio of the per-category to the total HGT proportion. Statistical significance of enrichment is calculated with Fisher's exact test.

Table 3. Genes classified as class 1 HGT selected from different bacterial groups including Proteobacteria, Firmicutes, Actinobacteria, Spirochaetes and Cyanobacteria. All genes shown here are presented in phylogenetic trees.

Additional file 1

Figure S1. HGT detection method illustrated with four examples of vertically and horizontally transferred genes. For each gene from the source organism, a list of BLAST hits is compiled, along with the 16S rRNA distances between the source and each target organism. Each plot shows the 16S rRNA distance as a function of BLAST hit rank. The blue and red horizontal lines correspond to the parameters d_self and d_distant, respectively. The plots show examples of (A) a vertically transferred gene; (B) a class 1 horizontally transferred gene; (C) a class 2 horizontally transferred gene; (D) a gene with undetermined HGT status.

Figure S2. Phylogenetic evidence for HGT of the *Xanthomonas campestris pv. Campestris 8004* transporter gene (XC_2737: ABC transporter ATP-binding protein). Taxonomic associations are shown after genus names, with selected strain highlighted in blue. The distance tree of selected class1-HGT gene was computed by Neighbor-Joining (NJ) method. Numbers on the node correspond to bootstrap replicates. The phylogenetic tree showed a phylogenetic placement (in

red) of *Xanthomonas campestris pv. Campestris 8004* with acidobacteria (with 60% of bootstrap).

Figure S3. Phylogenetic evidence for HGT of the *Xanthomonas axonopodis pv. citri 306* transporter gene (XAC0179: ABC transporter ATP-binding protein). The phylogenetic tree showed a phylogenetic placement (in red) of *Xanthomonas axonopodis pv. citri 306* (in red) with cyanobacteria (with 98% of bootstrap). Details as in Figure S2.

Figure S4. Phylogenetic evidence for HGT of the *Caulobacter crescentus CB15* transporter gene (CC_1204: AcrB/ AcrD/AcrF family protein). The phylogenetic tree showed a phylogenetic placement (in red) of *Caulobacter crescentus CB15* with gamma proteobacteria (with 64% of bootstrap). Details as in Figure S2.

Figure S5. Phylogenetic evidence for HGT of the *Bacillus amyloliquefaciens strain FZB42* transporter gene (RBAM_004660: mntH). The phylogenetic tree showed a phylogenetic placement (in red) of *Bacillus amyloliquefaciens strainFZB42* with *gamma-proteobacteria* (with 98% of bootstrap) but far from other firmicutes, which suggested a possible HGT event. Details as in Figure S2.

Figure S6. Phylogenetic evidence for HGT of the *Bacillus amyloliquefaciens (strain FZB42)* transporter gene (RBAM_003770: *yckJ* it is a part of a binding-protein-dependent transport system. Probably responsible for the translocation of the substrate across the membrane). The phylogenetic tree showed a phylogenetic placement (in red) of *Bacillus amyloliquefaciens (strain FZB42)* with actinobacteria (with 91% of bootstrap). Details as in Figure S2.

Figure S7. Phylogenetic evidence for HGT of the *Bacillus anthracis* transporter gene (BA_0232: oligopeptide ABC transporter). The phylogenetic tree showed a phylogenetic placement (in red) of *Bacillus anthracis* with actinobacteria (with 94% of bootstrap). Details as in Figure S2.

Figure S8. Phylogenetic evidence for HGT of the *Bacillus anthracis* transporter gene (BAA_0983: Sulfate permease family protein). The phylogenetic tree showed a phylogenetic placement (in red) of *Bacillus anthracis* with actinobacteria (with 100% of bootstrap). Details as in Figure S2.

Figure S9. Phylogenetic evidence for HGT of the *Mycobacterium tuberculosis str H37Rv* transporter gene (Rv0362: Magnesium transporter). The phylogenetic tree showed a phylogenetic placement (in red) of *Mycobacterium tuberculosis str H37Rv* with gamma-

proteobacteria (with 100% of bootstrap) but far from other Actinobacteria, which suggested a possible HGT event. Details as in Figure S2.

Figure S10. Phylogenetic evidence for HGT of the *Arthrobacter aurescens TC1* transporter gene (Aaur_0057: arsenite efflux pump). The phylogenetic tree showed a phylogenetic placement (in red) of *Arthrobacter aurescens TC1* with Deinococcus (with 90% of bootstrap). Details as in Figure S2.

Figure S11. Phylogenetic evidence for HGT of the *Arthrobacter aurescens TC1* transporter gene (Aaur_0347: D-ribose transport system ATP-binding protein). The phylogenetic tree showed a phylogenetic placement (in red) of *Arthrobacter aurescens TC1* with chloroflexi and beta-proteobacteria (with 98% of bootstrap). Details as in Figure S2.

Figure S12. Phylogenetic evidence for HGT of the *Corynebacterium efficiens strain DSM 44549* transporter gene (CE0611: gluconate permease protein). The phylogenetic tree showed a phylogenetic placement (in red) of *Corynebacterium efficiens strain DSM 44549* with proteobacteria (with 100% of bootstrap). Details as in Figure S2.

Figure S13. Phylogenetic evidence for HGTof *Treponema denticola strain ATCC 35405* transporter gene (TDE_0130: sodium/dicarboxylate symporter family protein). The phylogenetic tree showed a phylogenetic placement (in red) of *Treponema denticola strain ATCC 35405* with firmicutes (with 97% of bootstrap). Details as in Figure S2.

Figure S14. Phylogenetic evidence for HGT of *Leptospira biflexa serovar Patoc strain Patoc 1 / Ames* transporter gene (LBF_0368: Periplasmic binding protein of an ABC transporter complex). The phylogenetic tree showed a phylogenetic placement (in red) of *Leptospira biflexa serovar Patoc strain Patoc 1 / Ames* with aquificae and bacteroidetes (with 99% of bootstrap). This gene was only found in leptospira with no other spirochaetes. Details as in Figure S2.

Figure S15: Phylogenetic evidence for HGT of *Synechocystis sp. PCC 6803* transporter gene (Slr0982: ABC transporter). The phylogenetic tree showed a phylogenetic placement (in red) of *Synechocystis sp. PCC 6803* with Archaea (with 100% of bootstrap). Details as in Figure S2.

Figure S16: Phylogenetic evidence for HGT of *Acaryochloris marina (strain MBIC 11017)* transporter gene (AM1_0014: Uracil-xanthine permease). The phylogenetic tree showed a phylogenetic placement (in red) of *Acaryochloris marina (strain MBIC 11017)* with Deferribacteres (with 80% of bootstrap). Details as in Figure S2.