The ISME Journal (2015) 9, 958–967

www.nature.com/ismej

ORIGINAL ARTICLE

Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria

Alejandro Caro-Quintero[1,4] and Konstantinos T Konstantinidis[1,2,3]
[1]*School of Biology, Georgia Institute of Technology, Atlanta, GA, USA;* [2]*School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA and* [3]*Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA, USA*

Genome sequencing has revealed that horizontal gene transfer (HGT) is a major evolutionary process in bacteria. Although it is generally assumed that closely related organisms engage in genetic exchange more frequently than distantly related ones, the frequency of HGT among distantly related organisms and the effect of ecological relatedness on the frequency has not been rigorously assessed. Here, we devised a novel bioinformatic pipeline, which minimized the effect of over-representation of specific taxa in the available databases and other limitations of homology-based approaches by analyzing genomes in standardized triplets, to quantify gene exchange between bacterial genomes representing different phyla. Our analysis revealed the existence of networks of genetic exchange between organisms with overlapping ecological niches, with mesophilic anaerobic organisms showing the highest frequency of exchange and engaging in HGT twice as frequently as their aerobic counterparts. Examination of individual cases suggested that inter-phylum HGT is more pronounced than previously thought, affecting up to ~16% of the total genes and ~35% of the metabolic genes in some genomes (conservative estimation). In contrast, ribosomal and other universal protein-coding genes were subjected to HGT at least 150 times less frequently than genes encoding the most promiscuous metabolic functions (for example, various dehydrogenases and ABC transport systems), suggesting that the species tree based on the former genes may be reliable. These results indicated that the metabolic diversity of microbial communities within most habitats has been largely assembled from preexisting genetic diversity through HGT and that HGT accounts for the functional redundancy among phyla.
The ISME Journal (2015) **9,** 958–967; doi:10.1038/ismej.2014.193; published online 14 October 2014

Introduction

Bacteria catalyze fundamental steps of the geochemical cycles in almost every habitat on Earth and are partners in important ecological relationships with many eukaryotic organisms (that is, symbiosis, protocooperation, competition). One of the key mechanisms that underlie this remarkable physiological diversity is horizontal gene transfer (HGT) (Ochman *et al.*, 2000, McDaniel *et al.*, 2010). In fact, recent analysis of protein families suggests that HGT, and not gene duplication, has driven protein expansion and functional novelty in bacteria, which is in contrast with most eukaryotic organisms (Treangen and Rocha, 2011). Sequencing of thousands of microbial genomes during the past two decades has allowed the identification of HGT events at different timescales, from ancestral to recent events, and between organisms of varied evolutionary relatedness, from closely related genomes to very distantly related ones (Gogarten *et al.*, 2002; Beiko *et al.*, 2005; Ochman *et al.*, 2005; Zhaxybayeva *et al.*, 2009a), revealing that HGT has affected the evolutionary history of most, if not all, bacterial lineages.

Genetic exchange between distantly related bacteria is generally thought to occur less frequently than between closely related organisms due to larger ecological differences and genetic mechanisms, such as lower recombination efficiency with higher sequence divergence, defense mechanisms against foreign DNA and incompatibility in gene regulation. Nevertheless, cases of extensive (that is, involving a couple of hundred genes or more) inter-phylum genetic exchange have been documented for organisms living under extreme environmental selection pressures, such as thermophilic (Zhaxybayeva *et al.*, 2009b) and halophilic organisms (Nelson-Sathi *et al.*, 2012). Recently, we reported extensive inter-phylum genetic exchange between mesophilic *Sphaerochaeta* (*Spirochaetes*) and *Clostridia* (*Firmicutes*) (Caro-Quintero *et al.*, 2012), which indicated that high levels of inter-phylum HGT might also occur among non-extremophilic

Correspondence: KT Konstantinidis, School of Civil and Environmental Engineering, Georgia Institute of Technology, 311 Ferst Dr., Atlanta, GA 30332-0512, USA.
E-mail: kostas@ce.gatech.edu
[4]Current address: Department of Integrative Biology, University of Texas, Austin, TX 78712, USA.
Received 4 April 2014; revised 12 July 2014; accepted 25 August 2014; published online 14 October 2014

organisms. Obtaining a more complete picture of the magnitude of HGT among distantly related organisms is important to better understand bacterial genome plasticity and fluidity, the limits in phylogenetic reconstruction, particularly at ancestral nodes, and the factors that facilitate such HGT events.

Quantification of HGT among distantly related organisms represents a challenging task, in part because of the lack of complete representation of the prokaryotic diversity and the low number of shared genes between such organisms (Ciccarelli *et al.*, 2006). There are currently two commonly used approaches to identify HGT, phylogenetic (that is, tree-based) and best-match analysis. Tree-based methods are powerful in detecting HGT and offer high sensitivity, but they are computationally intensive and therefore not suitable for whole-genome analysis of a large number of genomes. An alternative approach is the best-match analysis based on the Smith–Waterman algorithm or its variations (Smith and Waterman, 1981). In this approach, gene sequences or their translated peptides are searched against characterized genomes (database) and the identity of best-matches against distantly vs closely related genomes is examined to identify putative HGT events. These approaches are computationally less expensive and scalable to large data sets. However, the best-match approach may provide lower sensitivity compared with the tree-based methods, especially in cases where biases exist in the genome database used (for example, if certain taxa are under-represented) and/or gene loss (as opposed to HGT) has occurred (Koski and Golding, 2001).

Here we extended our previous best-match analysis (Caro-Quintero *et al.*, 2012) to all available complete genome sequences of free-living Archaea and Bacteria to quantitatively evaluate whether or not extensive inter-phylum HGT occurs within non-extreme environments and identify what environmental and ecological conditions favor such HGT events and which functional genes are more frequently transferred. Our bioinformatic pipeline minimized the effect of taxonomic classification and over-representation of specific phylogenetic groups to provide unbiased, quantitative estimates of HGT across all taxa evaluated.

Materials and methods

To control for the effect of database representation in detecting HGT among distantly related genomes, all available genomes were compared in triplets (104 101 468 triplets). Each genome triplet included a reference genome (reference), a genome assigned to the same phylum as the reference (insider) and a genome assigned to a different phylum (outsider) (Supplementary Figure S1). For each triplet, all translated protein-coding genes of the reference genome (query) were searched against a database composed of the predicted proteins from the insider and the outsider genomes, and the percentage of best matches in the outsider was quantified (best-match ratio). In this way, phyla were compared based on the same number of genomes. Detection of horizontally transferred genes was performed based on the genome triplets as described in the Results section below. Further details about the bioinformatics procedures and statistics tests used are provided in the Supplementary Information.

Results

An approach to overcome the known limitations in detecting HGT

The analysis of the genome triplets showed that the more divergent the reference and insider genomes were, the larger the proportion of best matches of the reference to the outsider genome (Figure 1b, inset). The high proportion of best matches to the outsider were not attributable to gene loss in the insider because the same trend was observed when the analysis was restricted to genes shared by all three genomes in a triplet (Figure 1b). These results were presumably attributable to false-positive HGT detection, that is, when analyzing insider genomes that are distantly related to the reference genome it is likely to obtain a best match from a genome of a different phylum by chance alone due to the low signal-to-noise ratio between the insider and outsider genomes, rather than HGT. These interpretations were also consistent with previous studies showing low resolution of best-match approaches when analyzing distantly related genomes (Nelson *et al.*, 1999). This trend suggested that deep-branching genomes, for example, relatives from the same phylum with genome-aggregate average amino-acid identity (gAAI; Konstantinidis and Tiedje, 2005) $<60\%$ (Figure 1a), will always have a substantial amount of genes with best matches in a different phylum, irrespective of the occurrence or not of HGT. The results highlighted and quantified the limitation of best-match approaches with distantly related genomes; the quantification of the limitation provided the basis for an approach to overcome it.

To minimize false positives and identify genes and genomes that have undergone inter-phylum HGT with statistical confidence, approaches based on the distribution of best-match ratios (genome level) and sequence identities of orthologs (gene level) were used. At the genome level, the portion of total genes in the genome with best match(es) in the outsider was calculated for each triplet and compared with a distribution derived from all triplets with the same reference genome and genetic relatedness between the genomes in a triplet (measured by gAAI). Thus the effect of genetic divergence on the resulting data was presumably

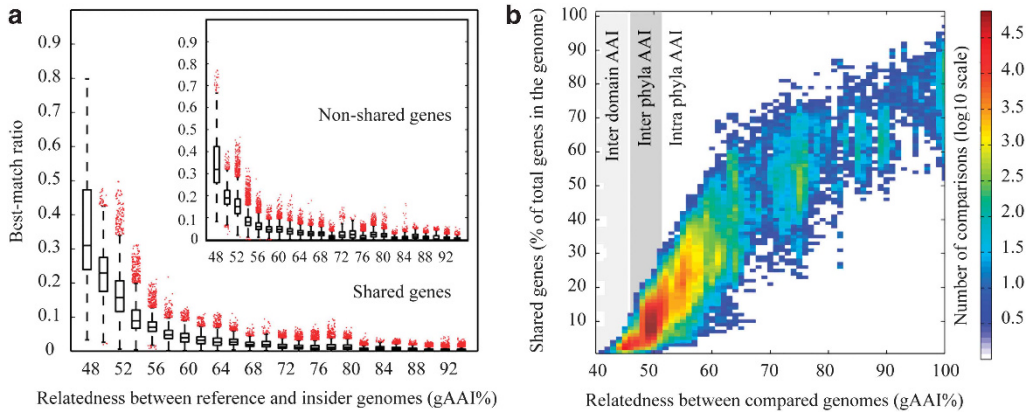

Figure 1 Dependence of the number of shared genes and intra- vs inter-phylum best match on the genetic divergence of the genomes compared. In all, 1 838 736 pairwise whole-genome comparisons were performed, and the relationship between genetic divergence, measured as gAAI, and the percentage of shared genes for these genomes is represented by a colored density plot (see scale). The shaded areas roughly correspond to the gAAI values between bacteria and archaea (inter-domain), between phyla and within phyla (**a**). The genomes were grouped in triplets, as described in the text, and the genes of the reference genome in the triplet were searched against the other two genomes, one representing the same phylum as the reference and the other representing a different phylum. The ratio of the number of best matches in the outsider vs all best matches (matches in insider and outsider) is plotted against the gAAI values between the two genomes of the same phylum in the triplet (boxplots in panel **b**). Each boxplot represents the distribution of ratios from 4000 randomly drawn triplets per unit of gAAI. Main graph shows the data for reference genes that had a match in both of the other two genomes in the triplet (shared genes); inset shows the genes that had a match in either (but not both) of the genomes. Red points represent the outliers. Note that the more divergent the genomes compared the higher, on average the ratio, which indicates that there is higher likelihood of detecting false-positive HGT events among more divergent genomes.

insignificant as only genomes with similar gAAI values were compared. At the gene level, the method evaluated individual genes by assessing how uncommon the sequence identity between the reference and the outsider genomes is compared with the average identity of all genes shared between genomes of triplets showing similar gAAI values, which represented the vertical inheritance (null model; see Supplementary Information online for more details on how this was done separately for shared and non-shared genes between the reference and the insider genomes). Note that the genome-level method included both ancestral and recent HGT events, because the identity of the match was not taken into account. In contrast, the gene-level approach primarily detected relatively recent HGT events, because it was based on detecting outliers in terms of sequence identity (Supplementary Figure S2). Using both approaches, significant inter-phylum HGT signal was detected in 811 out of the total 847 evaluated genomes, which suggests that inter-phylum HGT has been an important process in bacterial evolution.

Shared physiology and ecology underlie networks of high inter-phylum HGT

The influence of ecology and physiology on inter-phylum exchange was evaluated by generating networks that represented the cases of HGT. These networks were made by linking the donors and recipients with statistically significant signal of HGT (q-value threshold 0.005). Two networks were built, one for the genome-level analysis and one for the individual gene level. The subnetworks were named 'N' for genome-level (Supplementary Figure S3) and 'A' (Figure 2) for gene-level analysis. Within each network, a community-clustering algorithm (Clauset *et al.*, 2004; Newman and Girvan, 2004) was used to cluster the original network into subnetworks that maximized HGT connections among its members, that is, HGT to be more abundant among the genomes of the subnetwork compared with other genomes or subnetworks. Consistent with the expectations, inter-phylum exchange within a subnetwork involved 6 to 37 times more genes compared with between the subnetworks, depending on the subnetworks considered (Figure 2c). Ecological and physiological parameters were overlaid on the subnetworks to evaluate their correspondence with the clusters observed.

The analysis of the genome-level network revealed that HGT is strongly favored by (shared) ecology and oxygen tolerance. The genome-level network was split by the community-clustering algorithm into four subnetworks, N1, N2, N3 and N4 (Supplementary Figure S3). Subnetwork N3 was enriched in animal- or human-associated commensal and pathogenic genomes (64% of total genomes), which was significant based on the frequency of these genomes in the preclustered reference genome data set (P-value <0.001, Supplementary Table S1). The enriched genomes included members of the *Enterobacteriaceae* (*Proteobacteria*) and the *Streptococcaceae*, *Lactobacillales*, *Listeriaceae* and *Staphylococcacea* (*Firmicutes*). These findings agreed with a previous study that showed high levels of genetic exchange between human-associated bacteria (Smillie *et al.*, 2011); for example, $>70\%$ of the genes detected as exchanged by Smillie *et al.*

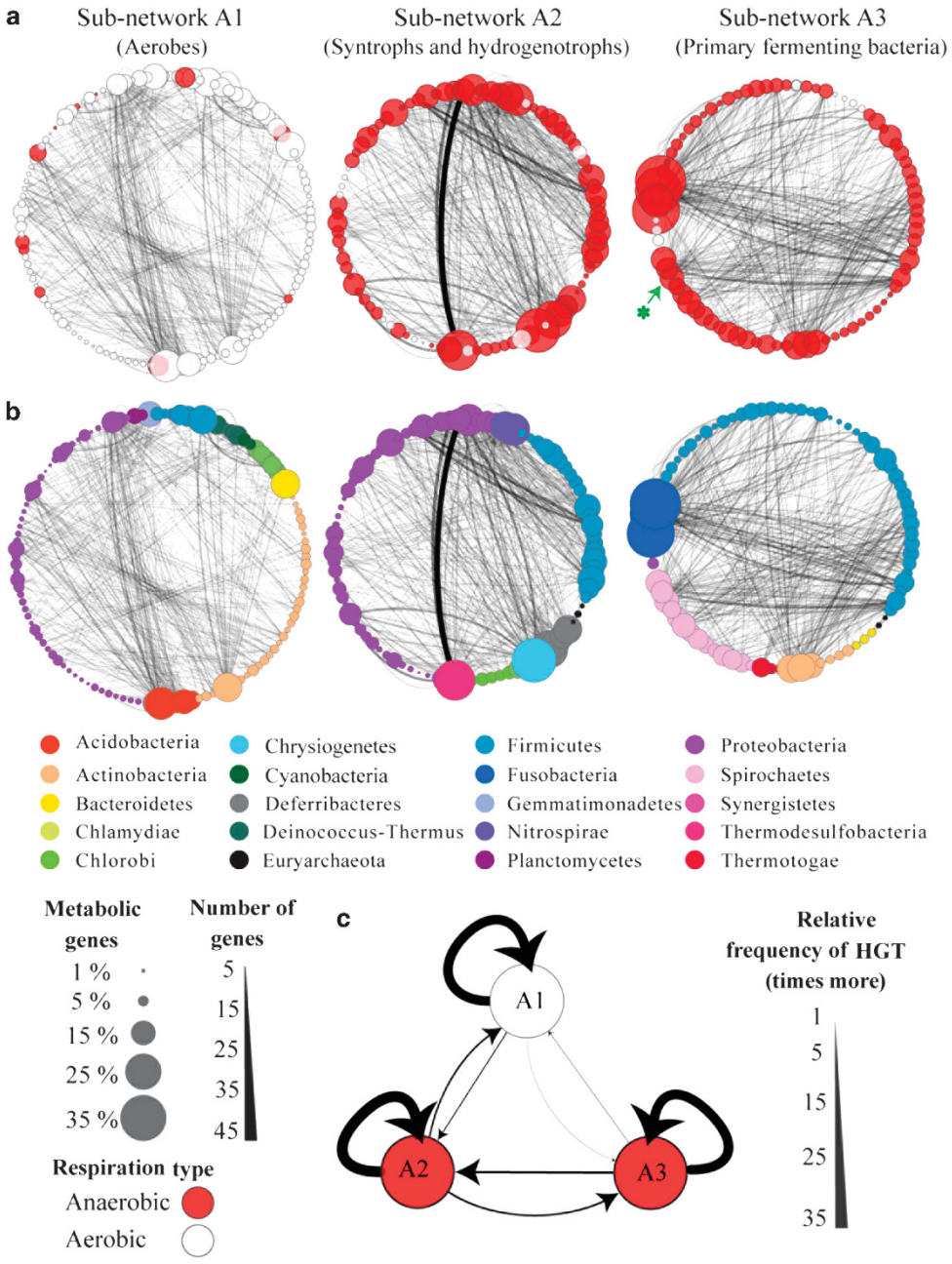

Figure 2 Networks of extensive inter-phylum HGT. A network representing all cases of HGT was constructed by linking genomes that had exchanged more than three genes and dividing the linkages into subnetworks that maximized the connectivity between nodes. Three subnetworks were obtained, A1, A2 and A3, in which nodes represent genomes and lines represent cases of HGT (**a**). The number of genes exchanged between the genomes is represented by the thickness of the lines (see scale at the bottom left). The percentage of the total metabolic genes in the genome transferred is represented by the size of each node, and the color of the node denotes aerobic (white) or anaerobic organisms (red). Panel (**b**) is identical to panel **a** with the exception that the colors of the taxa denote taxonomic affiliation instead of preference for oxygen (see figure key). The frequency of exchange within and between the networks was calculated by selecting randomly 40 genomes, with 1000 replicates, and taking the average of the number of exchanges detected in all replicates. The relative value was calculated by dividing all resulting average frequencies by the lowest inter-network frequency (see figure key; **c**). The green arrow denotes the *Sphaerochaeta–Clostridia* example discussed in the text.

(2011) were also identified by our methodology. Subnetwork N2 was significantly enriched in soil- and plant-associated bacteria (~50% of total) assigned to the *Rhizobiales*, *Bradyrhizobiaceae* and *Comamonadaceae* (phylum *Proteobacteria*) and *Streptomycetaceae* and *Micrococcaceae* (*Actinobacteria*) (P-value <0.001, Supplementary Table S1). On the other hand, subnetworks N1 and N4 were dominated by aquatic thermophilic and mesophilic organisms, respectively (~70% of total; P-value <0.001, Supplementary Tables S1 and S2). Mesophilic groups included organisms of the phylum *Chloroflexi*, *Chroorococcales* (*Cyanobacteria*), *Flavobacteriaceae* (*Bacteroidetes*) and *Alteromonadaceae*

(*Proteobacteria*). Meso- and hyper-thermophilic taxa included organisms of the phylum *Deinococcus–Thermus*, *Thermoanaerobacteriales* (*Firmicutes*), representatives of the phylum *Thermotogae*, and Achaea of the *Euryarchaeota* and *Crenarchaeota* phyla. Notably, among the evaluated parameters (16 parameters in total; see Supplementary Figure S3), oxygen tolerance appeared to correspond best with the subnetwork clustering. For instance, subnetwork N1 was mainly composed by anaerobic bacteria (80%), while N2, N3 and N4 were dominated by aerobic bacteria (89, 80 and 74%, respectively). These findings suggest that, among all evaluated environmental parameters, oxygen tolerance has the most important role in fostering HGT within aerobic and anaerobic environments.

Consistent with the above results, gene-level network analysis showed that oxygen tolerance best explained the clustering of genomes in the three largest subnetworks A1 (119 genomes), A2 (89 genomes) and A3 (82 genomes). For instance, subnetwork A1 was mainly composed of aerobic organisms while subnetworks A2 and A3 were mainly composed of anaerobic organisms. More specifically, subnetwork A2 was composed of hydrogenotrophic organisms (methanogenic achaea and sulfate-reducing bacteria) and syntrophic bacteria, while subnetwork A3 of primary fermenting bacteria (Figure 2a; enrichment P-value <0.001 for all subnetworks, Supplementary Tables S3 and S4). Analysis of the frequency of genes transferred showed that the metabolic functions in networks composed of (primarily) anaerobic bacteria, A2 and A3, have been exchanged twice as frequently compared with aerobic metabolic genes (subnetwork A1; Supplementary Figure S4B). Further, exchange between subnetworks A1 and A3 was the lowest, wherea A2 and A3 (both encompassing mostly anaerobic organisms) showed the highest frequency of inter-network exchange (Figure 2c).

The properties of the genomes in the four genome-level subnetworks (N1.1, N1.2, N1.3 and N1.4) were examined more closely. Subnetwork N1.2 was the most phylogenetically diverse, encompassing 11 different phyla, but strongly enriched in organisms of the *Firmicutes* phylum (57% of the total). Interestingly, elimination of *Firmicutes* from the network reduced the number of transfers (edges) by 97%, suggesting that organisms of this phylum are the most important partners in HGT for this subnetwork (see also subnetwork A3 in Figure 2b). Further analysis revealed two main physiological groups within N1.2. The first was composed of aquatic themophilic and hyperthermophilic bacteria (for example, *Thermoanaerobacterium xylanolyticum* and *Spirochaeta thermophila*) and the second of soil saprophytic fermenters (for example, *Sphaerochaeta* spp and *Clostridium cellulovorans*) and gut-associated bacteria from insects, humans and ruminants (for example, *Sphaerochaeta coccoides*, *Eubacterium rectale* and *Roseburia hominis*). Even though the latter organisms differ in their source of isolation and optimal growth temperature, they are all characterized by saccharolytic and fermentative lifestyles. Therefore the organisms grouped under subnetwork N1.2 suggested that organic matter degradation genes are relevant across several ecological niches that are rich in organic matter content and have been commonly transferred from/to *Firmicutes* multiple times.

Analysis of subnetworks N1.1 revealed the importance of strong ecological interactions (that is, protocooperation) in facilitating genetic exchange. The organism that were connected in this subnetwork were either syntrophs or had representatives reported to be partners of hydrogen-based syntrophic interactions. The three main phylogenetic groups in the network included syntrophic bacteria of the *Proteobacteria* phylum (for example, *Syntrophus* spp.), sulfate reducers of the *Firmicutes* phylum (for example, *Desulfotomaculum* spp.) and methanogenic archaea of *Euryarchaeota* phylum (for example, *Methanocella* spp.) (Schink and Stams, 2006). Therefore the high frequency of HGT between these groups indicated that syntrophic associations have a key role in facilitating inter-phylum HGT. These results were consistent with previous phylogenetic approaches that showed high gene sharing between syntrophic organisms (Cordero and Hogeweg, 2009). Additionally, it has been suggested that HGT is responsible for similar codon usage bias between *Pelotomaculum thermopropionicum* and other syntrophic organisms (Kosaka *et al*., 2008) and that syntrophic interactions between *Desulfovibrio vulgaris* and *Methanosarcina barkeri* had evolved as a result of ancestral HGT (Scholten *et al*., 2007). In conclusion, the previous syntrophic organisms represent examples of how tight ecological relationships (that is, physiological dependence and physical contact) have favored the transfer of genetic material between distantly related organisms.

Extensive inter-phylum HGT occurs often among mesophiles

To establish whether or not the extensive inter-phylum exchange previously observed in *Sphaerochaeta* (Caro-Quintero *et al*., 2012) represents a unique case, the proportion of genes in the genome that have a signal of inter-phylum exchange was quantified for every reference genome (genome-level analysis; Supplementary Figure S4A). The results showed that *Sphaerochaeta* ranked in the higher 97th percentile, with 6% of the total genes and 15% of all metabolic genes in the genome showing a signal of HGT. The analysis showed that *Sphaerochaeta* is not the only mesophile characterized by extensive genetic exchange; in fact, 24 out of the top 37 cases of extreme inter-phylum HGT also involved mesophiles (Supplementary Table S5). Collectively, these findings revealed that inter-phylum HGT is more pronounced than previously anticipated,

accounting for up to 16% of the total genes and 35% of the metabolic genes in some genomes. It should be mentioned that our method identified only HGT events with high confidence (*q*-value threshold 0.005); thus the previous results most likely represent an underestimation of the magnitude of HGT. For instance, using a less stringent cutoff (best match with >40% amino-acid identity over 70% length of the query protein) we have calculated previously that the same *Sphaerochaeta* genomes mentioned above have exchanged up to 40% of the total genes with *Firmicutes* (Caro-Quintero *et al.*, 2012).

Gene functional categories more frequently exchanged

Analysis of the predicted functions of genes that were transferred across phyla revealed that metabolic genes were among the most commonly exchanged genes, making up 60% of all detected HGT events and 70 of the top 100 most frequently exchanged individual functions (Figure 3a). The most frequently transferred genes were those related to lipid transport and metabolism, energy production and conversion, amino-acid transport and metabolism and carbohydrate transport and metabolism (Figure 3b). The specific functions most frequently exchanged included short dehydrogenases with different specificities (COG1028; 3.8% of all cases), NAD-dependent aldehyde dehydrogenases (COG1012; 2.2% of all cases), predicted oxidoreductases, ABC-type polar amino-acid transport system (COG1126; 1.8% of all cases) and acetyl-CoA acetyltransferase (COG0183; 1.7% of all cases). In contrast, informational functions were the least frequently transferred (12% of all cases); only four informational functions were found among the 100 most transferred functions (that is, peptide chain release factor RF-3, threonyl-tRNA synthetase, methionine aminopeptidase and methionyl-tRNA synthetase), and none of these categories were related to ribosomal proteins or DNA/RNA polymerases (Figure 3a).

The highly conserved informational genes frequently used to reconstruct the Tree of Life were transferred between phyla at extremely low frequencies. Only 6 genes out of the 36 described (Ciccarelli *et al.*, 2006) were detected as exchanged. These informational genes were transferred 151-fold less frequently, on average, than those that encode for the 6 most transferred functional categories (Figure 3a, inset; Supplementary Table S6). For example, arginyl-tRNA synthetase, a highly conserved gene, was transferred only once between *Salinispora tropica* and *Sorangium cellulosum* (all cases are provided in Supplementary Table S7). The detection of a high number of exchanged metabolic genes was not the result of over-representation of these categories in the reference genomes. For instance, the 6 most transferred metabolic functions were enriched 5–16-fold in the set of transferred genes compared with their average abundance in the genomes, whereas the 6 highly conserved genes were 5–20 times less abundant in the transferred gene set (Supplementary Table S6). The low frequency of exchange of informational genes is thought to be related to the high connectivity of their expressed proteins (Cohen *et al.*, 2011) and suggested that phylogenetic reconstruction based on these genes is largely impervious to HGT, at least for

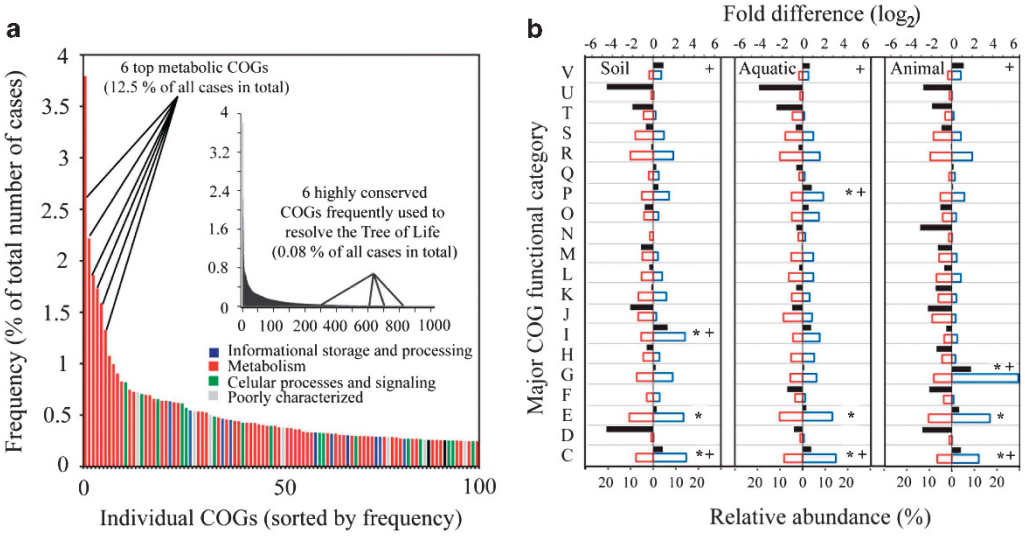

Figure 3 Functions transferred across bacterial and archaeal phyla. The top 100 COG proteins most frequently transferred across bacterial and archaeal phyla are shown (**a**). Individual COGs are colored based on the major functional category to which they are assigned (figure key). The genomes engaged in the HGT events detected were assigned to one of the three major habitats on Earth, and the functional enrichment of transferred genes within each habitat is also shown (**b**). Red bars represent the relative frequency of the COGs' major functional categories in the average genome (description of categories is provided in Supplementary Table S10). Blue bars represent the relative frequency based on genes exchanged. Black bars represent the fold difference between the previous two frequencies (that is, enrichment in HGT events). Symbols denote the categories most frequently exchanged (*) and with the higher fold increase (+).

the part of the Tree that can be robustly resolved by these genes (that is, within phylum but not phylum-level relationships).

Notably, the functions with higher frequency of exchange (for example, NAD-dependent aldehyde dehydrogenases) were also those that have been transferred between organisms from a larger number of different phyla (Supplementary Figure S5) and did not typically represent exchange between only a few highly over-represented phyla. Thus it appears that genes assigned to these functional categories are more promiscuous and probably represent important adaptive functions in several different habitats.

The functional biases in exchanged genes within soil-, aquatic- and animal-associated organisms were examined more closely to elucidate what functions are selected within each corresponding environment (Supplementary Figures S3B and S9). Categories enriched in each environment included: lipid transport and metabolism (I) exchanged most abundantly in soil, inorganic ion transport and metabolism (P) in aquatic habitats, and carbohydrate transport and metabolism (G) among animal-associated bacteria. These results suggested that the functions exchanged across phyla do not represent random collections of genes but rather reflected the acquisition of ecologically important functions for the corresponding organisms within their habitat(s).

The role of inter-phylum HGT in bacterial adaptation

To examine the importance of genetic exchange between distantly related organisms for adaptation and ecology, the genome pairs with the highest number of exchanged genes were analyzed further, focusing on transferred regions with two or more syntenic genes (all data are provided in Supplementary Table S8). As expected, the analysis of exchanged genes between specific genomes reflected the general trends mentioned above for the complete genome set (for example, Figure 3). Here, three examples that clearly demonstrate the importance of inter-phylum HGT for acquiring metabolic capabilities essential for the ecological niches of the recipient organism are highlighted.

One of the most notable cases of inter-phylum HGT is between the syntrophic bacteria *P. thermopropionicum* (*Firmicutes*) and *Syntrophobacter fumaroxidans* (*Proteobacteria*). Three large, syntenic regions, encoding mostly genes involved in the electron transport chain for ATP production and active transport of nitrate or sulfonate, were identified between representatives of these taxa with significantly higher amino-acid identity than expected by vertical descent. The amino-acid identity of these regions ranged from 82% to 62% with an average of 67%; this level of identity is significantly higher than average identity of the ribosomal proteins, (61%). Further, the genes in the syntenic regions 2 and 3 appear to be involved in reverse electron transport during syntrophic propionate metabolism and be fundamental for the establishment of successful syntrophic relationships (Sieber *et al.*, 2012). Propionate is an important intermediate in the conversion of complex organic matter under anaerobic conditions and its oxidation to acetate requires the presence of a methanogenic partner to maintain low hydrogen partial pressure (de Bok *et al.,* 2004). These results not only show clear evidence of genetic exchange between distantly related organisms but also, more importantly, suggest that overlapping ecology within anoxic environments had favored the exchange of key adaptive genes.

Another extreme case of inter-phylum HGT was *Listeria ivanovii* (*Firmicutes*) and *Sebaldella termitidis* (*Fusobacteria*), where 11 syntenic regions were exchanged, encoding genes associated with carbohydrate metabolism and transport. The largest region, syntenic region 4, encodes for 16 genes involved in the propanediol utilization pathway. This represents a potentially important ecological function as *L. ivannovii* and *S. termitidis* have been associated with the ruminant and the termite guts, respectively, and propanediol is thought to be important in these anoxic environments (Obradors *et al.*, 1988). Propanediol is a major product of the anaerobic degradation of common plant sugars (for example, rhamnose and fucose); however, its degradation is highly toxic, and bacteria need microcompartments (carboxysomes) to enclose the highly reactive intermediates of the degradation (Sampson and Bobik, 2008). Consistent with this, several carboxysome structural proteins were also exchanged between these genomes (for example, GI numbers 347548556 and 269119660) relatively recently, as reflected by the high amino-acid identities, ranging from 57% to 85%. These findings suggest that the capabilities for degradation of plant sugars under anaerobic conditions were transferred between phyla multiple times and were likely fundamental for adaptation of the previous organisms to the animal gut environment.

Noteworthy cases of gene transfer between oral-associated bacteria, *Streptococcus gordonii* (*Firmicutes*) and *Leptotricha buccalis* (*Fusobacteria*), were also observed, where nine syntenic regions, mainly related to carbohydrate transport and metabolism, were exchanged. Among these regions, an operon of seven genes related to the degradation of lactose through the tagatose 6-phosphate pathway, with amino-acid identities ranging from 63% to 82%, was observed. Lactose is an important component of the human diet, and it has been suggested that lactose catabolism can influence the ecological balance of oral bacteria and colonization of oral cavities and soft tissues (Jagusztyn-Krynicka *et al.*, 1992; Chen *et al.*, 2002).

As expected, the main mechanism underlying these inter-phylum HGT events was likely non-homologous recombination. Several transferred genes were flanked by transposases and integrases

as exemplified by the HGT event between *Desulfurispirillum indicum* (*Chrysiogenetes*) and *Marinobacter aquaeolei* (*Proteobacteria*), where a cation efflux pump gene flanked by transposases and integrase genes (99.3% amino-acid identity) was recently exchanged (97% amino-acid identity). Additionally, syntenic phage-related proteins (~50 genes) were shared among aquatic bacteria, '*Candidatus* Nitrospira defluvii' (*Nitrospira*) and *Janthinobacterium* sp. strain Marseille (*Proteobacteria*), with high identity (85% average amino-acid identity), also indicating recent genetic exchange (Supplementary Table S8).

Discussion

Genetic exchange between distantly related organisms representing different bacterial and/or archaeal phyla is thought to be very infrequent (Kurland *et al.*, 2003); however, our analysis revealed that inter-phylum exchanges had occurred in almost all of the evaluated genomes. Analysis of networks of HGT revealed that lifestyle and ecology drive most of the HGT events, especially the transfers involving a large number of metabolic genes, and that metabolic genes are exchanged twice as frequently among anaerobic organisms compared with aerobic ones and at least 150-fold more frequently, on average, than informational genes.

Extensive HGT among thermophiles, pathogens and cyanobacteria has been described previously, for example, 'highways' of HGT (Doolittle, 1998; Beiko *et al.*, 2005), and was attributed to substantial ecological overlap among the partners involved. Along the same lines, a recent study of intra-phylum HGT showed that very recent gene transfer events (reflected by >99% nucleotide sequence identity) are clearly structured by ecology, where the highest frequency of HGT was observed among organisms recovered from the same site of the human body (Smillie *et al.*, 2011). None of these previous studies, however, described cases of such extensive inter-phylum HGT as those described here or evaluated in detail the specific environmental and ecological parameters that underlie the 'highways' of inter-phylum HGT. In contrast to what was previously reported, our results showed that the most extensive genetic exchange occurs among mesophilic organisms with saccharolytic and fermenting metabolisms, mainly associated with anoxic environments characterized by high concentrations of plant organic matter such as those of the termite and ruminant guts. The differences between our findings and those reported previously might be related to the normalization of the database in terms of taxa representation and the fact that our method evaluated recent as well as more ancient HGT events.

Although the exact reasons for the higher frequency of HGT within anaerobic vs aerobic networks remain unclear, we hypothesize that more niche overlap and/or physical proximity among organisms is taking place within anaerobic environments, which favor HGT. For instance, aerobic microorganisms can frequently oxidize substrates to water and carbon dioxide without any significant cooperation with other organisms, whereas anaerobic microorganisms often depend to a greater extent on associations with different partners. For example, the complete anaerobic conversion of cellulose to methane and carbon dioxide requires the concerted action of at least four different metabolic groups of organisms, including primary fermenters, secondary fermenters and two types of methanogenic archaea (Schink and Stams, 2006).

To the best of our knowledge, extensive HGT within anaerobic mesophilic environments was first described between *Sphaerochaeta* spp and *Clostridia* (Caro-Quintero *et al.*, 2012). Thirty-seven cases with more extensive HGT than that observed in *Sphaerochaeta* were detected in the present study, 28 of which also involved anaerobic mesophilic organisms. Inspection of the individual genes exchanged suggested that the ability to engage in syntrophic metabolism, degrade toxic intermediates of plant organic matter decomposition and metabolize sugars in the oral cavity have been exchanged across phyla several times during the relatively recent evolutionary history. Thus, it appears that inter-phylum HGT has not only affected a substantial part of the genome in almost every bacterial lineage but was also fundamental for the adaptation of the organisms to their perspective ecological niche(s). In addition to individual organisms, broad phylogenetic taxa such as the *Firmicutes* participated in HGT much more frequently than expected by chance (for example, Figure 2b). The underlying mechanisms for the high promiscuity of *Firmicutes* in engaging in DNA exchange with multiple partners remain unclear but are likely related to genetic and/or physiological adaptations that facilitate genetic exchange. Alternatively, these organisms and their ancestors are the original innovators of anaerobic organic matter degradation and fermentation, and because of strong positive selection, the corresponding genes were frequently fixed after they had been horizontally transferred.

It is possible that other evolutionary processes accounted for some of the patterns observed, as opposed to real HGT between the reference and the outsider. For instance, faster sequence evolution (that is, higher substitution rates) of the insider gene relatively to its outsider homolog could result in higher amino-acid identity between the latter gene and the reference gene and, hence, (false) detection of gene exchange. Similarly, a triplet where the insider gene is a xenolog as opposed to an ortholog can result in higher identity between the reference and outsider genes. To assess the importance of these processes on our results, we evaluated the amino-acid identity of genes detected as horizontally transferred between the reference and the outsider genomes against the identity of all

homologs (most of which were orthologs) shared between the same genomes that were not detected as horizontally transferred. The two distributions differed dramatically, and transferred genes typically showed much higher identity compared with the non-transferred genes (Supplementary Figure S6). In fact, transferred genes showed higher identity even when compared with the most highly conserved, in terms of sequence conservation, housekeeping genes (for example, ribosomal proteins, DNA polymerase). If unequal substitution rates or xenologs were predominantly driving the HGT patterns observed and these affected many genes in the insider genome (not just a few), then the two distributions would have been more similar. Thus, although it is possible that a few instances might exist in our data set where unequal substitution rates or xenologs accounted for the HGT events detected, these should be much less important compared with real HGT events, especially in the cases of extensive inter-phylum HGT that we mostly focused on here, and the predominant signal is that our method is not affected by unequal substitution rates. We also confirmed that most cases of extensive inter-phylum HGT were not attributable to fewer or more divergent insider genomes in the phylum of the reference genome (hence, higher chance for having a best match in the outsider genome) relative to phyla/taxa with low frequency of HGT (Supplementary Figure S7). Therefore the cases of extensive HGT reported here are real and not an artifact of our methodology (for missed HGT events, see below).

It is also important to point out that, due to the still limited representation of the total natural microbial diversity by genome sequences, many more cases of extensive inter-phylum HGT currently evade detection. Shortage of available sequenced relatives, as opposed to physiological or ecological barriers to HGT, largely accounted for the lack of HGT signal in 36 of the total 847 genomes evaluated (Supplementary Table S9, which shows rates of HGT for all genomes). Consistent with the latter interpretations, the 36 genomes were over-represented in deep-branching (basal) lineages such as methanogenic *Archaea* and did not include human- or animal-associated organisms (which are well represented in the databases). Advances in DNA sequencing and single-cell technologies have exponentially lowered the cost of genome sequencing and, as a consequence, the pace at which natural diversity is being characterized is continuously increasing (Rinke *et al.*, 2013). To keep up with this trend, faster methods for HGT detection are needed. The simple strategy presented here, which is based on comparisons of genomes in triplets and the statistical evaluation of the sequence identity of homologs, provides means for fast HGT detection. In addition, our strategy provides a standardized framework to compare rates of HGT between organisms, identify the putative partners of exchange and assess the functions exchanged.

Collectively, our results suggest that members of some microbial communities essentially share their metabolism through a network of HGT, while preserving phylogenetic distinctiveness at housekeeping genes and that barriers to genetic exchange among distantly related organisms may not be as strong as previously thought. Therefore, although members of microbial communities appear to share metabolic genes and pathways as a somewhat 'common good' (McInerney *et al.*, 2011), highly conserved genes remain phylogenetically informative.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Frank Loeffler, King Jordan, Janet Hatt and two anonymous reviewers for helpful discussions and suggestions regarding the manuscript. This work was supported by the US National Science Foundation (NSF) award nos. 0919251 and 1241046.

References

Beiko RG, Harlow TJ, Ragan MA. (2005). Highways of gene sharing in prokaryotes. *Proc Natl Acad Sci USA* **102**: 14332–14337.

Caro-Quintero A, Ritalahti KM, Cusick KD, Loffler FE, Konstantinidis KT. (2012). The chimeric genome of *Sphaerochaeta*: nonspiral spirochetes that break with the prevalent dogma in spirochete biology. *MBio* **3**: e00025-12.

Chen YY, Betzenhauser MJ, Snyder JA, Burne RA. (2002). Pathways for lactose/galactose catabolism by *Streptococcus salivarius*. *FEMS Microbiol Lett* **209**: 75–79.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**: 1283–1287.

Clauset A, Newman MEJ, Moore C. (2004). Finding community structure in very large networks. *Phys Rev E* **70**(6 Pt 2): 066111.

Cohen O, Gophna U, Pupko T. (2011). The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol* **28**: 1481–1489.

Cordero OX, Hogeweg P. (2009). The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci USA* **106**: 21748–21753.

de Bok FAM, Plugge CM, Stams AJM. (2004). Interspecies electron transfer in methanogenic propionate degrading consortia. *Water Res* **38**: 1368–1375.

Doolittle WF. (1998). You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet* **14**: 307–311.

Gogarten JP, Doolittle WF, Lawrence JG. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**: 2226–2238.

Jagusztyn-Krynicka EK, Hansen JB, Crow VL, Thomas TD, Honeyman AL, Curtiss R 3rd. (1992). *Streptococcus mutans* serotype c tagatose 6-phosphate pathway gene cluster. *J Bacteriol* **174**: 6152–6158.

Konstantinidis KT, Tiedje JM. (2005). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**: 6258–6264.

Kosaka T, Kato S, Shimoyama T, Ishii S, Abe T, Watanabe K. (2008). The genome of *Pelotomaculum thermopropionicum* reveals niche-associated evolution in anaerobic microbiota. *Genome Res* **18**: 442–448.

Koski LB, Golding GB. (2001). The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52**: 540–542.

Kurland CG, Canback B, Berg OG. (2003). Horizontal gene transfer: a critical view. *Proc Natl Acad Sci USA* **100**: 9658–9662.

McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. (2010). High frequency of horizontal gene transfer in the oceans. *Science* **330**: 50.

McInerney JO, Pisani D, Bapteste E, O'Connell MJ. (2011). The Public Goods Hypothesis for the evolution of life on Earth. *Biol Direct* **6**: 41.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH *et al.* (1999). Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399**: 323–329.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO *et al.* (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of *Haloarchaea*. *Proc Natl Acad Sci USA* **109**: 20537–20542.

Newman MEJ, Girvan M. (2004). Finding and evaluating community structure in networks. *Phys Rev E* **69**(2 Pt 2): 026113.

Obradors N, Badia J, Baldoma L, Aguilar J. (1988). Anaerobic metabolism of the L-rhamnose fermentation product 1,2-propanediol in *Salmonella typhimurium*. *J Bacteriol* **170**: 2159–2162.

Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.

Ochman H, Lerat E, Daubin V. (2005). Examining bacterial species under the specter of gene transfer and exchange. *Proc Natl Acad Sci USA* **102**(Suppl 1): 6595–6599.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Sampson EM, Bobik TA. (2008). Microcompartments for B-12-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. *J Bacteriol* **190**: 2966–2971.

Schink B, Stams AM. (2006). Syntrophism among prokaryotes. In: Dworkin M, Falkow S, Rosenberg E, Schleifer K-H, Stackebrandt E (eds). *The Prokaryotes*. Springer: New York, NY, USA, pp 309–335.

Scholten JC, Culley DE, Brockman FJ, Wu G, Zhang W. (2007). Evolution of the syntrophic interaction between *Desulfovibrio vulgaris* and *Methanosarcina barkeri*: Involvement of an ancient horizontal gene transfer. *Biochem Biophys Res Commun* **352**: 48–54.

Sieber JR, McInerney MJ, Gunsalus RP. (2012). Genomic insights into syntrophy: the paradigm for anaerobic metabolic cooperation. *Annu Rev Microbiol* **66**: 429–452.

Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**: 241–244.

Smith TF, Waterman MS. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**: 195–197.

Treangen TJ, Rocha EP. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* **7**: e1001284.

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. (2009a). Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. *Genome Biol Evol* **1**: 325–339.

Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT *et al.* (2009b). On the chimeric nature, thermophilic origin, and phylogenetic placement of the *Thermotogales*. *Proc Natl Acad Sci USA* **106**: 5865–5870.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)