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

Alejandro Caro-Quintero and Konstantinos T. Konstantinidis

Supplementary Material

MATERIALS AND METHODS (DETAILED DESCRIPTION)

Amino acid and genome sequences used in this study. Predicted proteins from completed bacterial and archaeal genome projects were downloaded from NCBI on July 1, 2012 (2,001 genomes) to form an in-house searchable database. To avoid the effect of genome reduction in endosymbiotic organisms, which can bias comparisons of the magnitude of HGT across genomes, only free-living genomes with genome size larger than 2 Mbp were used in the analysis (1,356 genomes). The resulting set of genomes represented 28 phyla. Literature review was performed to identify physiological and ecological information for each genome, i.e., source of isolation, optimal growth temperature, and oxygen tolerance.

Homolog identification and database normalization. Orthologous genes for all possible pairs of genomes (1,838,736 pairs) were identified using the reciprocal best match approach (Wolf and Koonin 2012) and the USEARCH algorithm for its computational efficiency (Edgar 2010). Only best matches with identity higher than 40% and coverage of the query gene sequence higher than 70% were used in the analysis. For any pair of genomes, the gAAI value was calculated by averaging the identity of shared orthologs as suggested previously (Konstantinidis and Tiedje 2005). In order to reduce the redundancy (and thus, the size) of the database for faster computations, genomes were clustered in groups that shared higher than 95% gAAI, which corresponds to the frequently used standards to define bacterial species (Konstantinidis and Tiedje 2005). One genome from each of the resulting groups (n=879) was randomly selected to represent the group, and the gAAI values between the representative genomes were used to estimate the genetic divergence between the groups.

Quantifying HGT at the genome-level. For each genome triplet, two percentages were calculated (best-match ratios; see also main text and Fig. 1B): one for reference protein sequences that had a homolog in both the insider and outsider genomes (shared genes), and one for protein sequences with a homolog in either or both the insider and outsider genomes (all genes). The percentages were grouped by the genetic relatedness of the two genomes of the same phylum, i.e., triplets with gAAI values within ± 1% of a chosen gAAI value were grouped together. For each resulting group, the distribution of the percentages were normalized by standardization and fitted to a Gaussian distribution. P-values were calculated from this null distribution for each percentage, corrected for multiple testing using False Discovery Rate (FDR) statistic (Benjamini and Hochberg 1995, Shaffer 1995), and used to identify outliers (q-value threshold 0.005) in terms of high best-match ratio, representing cases of extreme inter-phylum HGT between the reference and the outsider genomes. Note that the HGT events detected by this analysis included both recent and ancestral events because all genes with a better match in the outsider relative to the insider (for shared genes) or no match in the insider were considered as horizontally transferred genes, independent of the identity of the match. Further, the effect of genetic divergence on the number of best-matches (e.g., Fig. 1B) and hence, on our conclusions, was presumably insignificant as only genomes with similar gAAI values were compared. All statistical analyses were performed using

MATLAB and the Statistics Toolbox, Release 2012b (MathWorks, Inc.; Natick, MA, USA).

Quantifying HGT at the gene level. Homologous protein-coding genes shared between the reference and outsider genomes were evaluated statistically to identify cases of HGT and determine the functional categories that are more commonly transferred across phyla. Two different statistical approaches were employed; one for homologs present in all genomes of the triplet (shared genes), and one for homologs only shared by the reference and outsider genomes (non-shared genes).

For shared genes, all homologs were grouped in sets based on the gAAI values (± 1%) of the corresponding triplets (gAAI between the reference and insider genomes; see above). For each set, the amino acid sequence identity between the reference and outsider homologs was subtracted from the identity between the reference and insider homolog (% identity with the insider - % identity with the outsider), and a distribution of the resulting numerical difference values was obtained. Therefore, one such distribution was calculated for triplets with the similar gAAI values based on all shared genes between the genomes in the triplets. Each distribution was fitted to a normal, polynomial, or gamma function and the function with the best fit to the observed distribution (Kruskal-Wallis test) was selected. The best function was the gamma for gAAI values ranging between 60-94% and the polynomial for 50-60% gAAI; hence, the gamma was preferred for the remaining analysis. The parameters of the gamma function were extracted and used to produce one general model for all genes and all gAAIs. This model described the expected probability of finding a homolog shared between the reference and outsider genomes with a specific amino acid sequence identity value. p-values were estimated from the cumulative density distribution of the model (1 – model; Fig. S2A) and the effect of multiple testing was accounted for using the FDR. HGT events were defined as cases where matches to the outsider had significantly higher identity compared to matches to the insider (p-value < 0.0001 and q-value <0.005).

For non-shared homologs, a different approach was used to distinguish cases of HGT from gene loss in the lineages of the insider genome. The approach was based on the assumption that the majority of fixed mutations among orthologs reflect vertical descent (Wolf and Koonin 2012), and therefore the variation in amino acid sequence identities among orthologs can be used as a null model to identify cases with sequence identity higher than expected due to HGT. Orthologs from different phyla were assigned, when possible, to the Cluster of Orthologous Groups (COGs), and the mean and standard deviation of the distribution of amino acid sequence identities between orthologs of the reference against the outsider (i.e., inter-phylum identity) were calculated for each COG. These values were used to statistically evaluate if the identity of the match(es) against the outsider genome was higher than expected by vertical descent and hence, attributable to HGT. A high stringency threshold (q-value threshold 0.005) was used to identify cases of HGT (Fig. S2B). Note that, contrary to shared-genes where a single model was used for all genes, the approach for non-shared genes employed a unique distribution (model) for each COG, accounting for the different degree of sequence conservation of genes (e.g., ribosomal protein-coding genes tend to be more conserved than metabolic ones). For genes with matches below the threshold in multiple genomes of the lineage that the outsider or insider genomes were assigned to, only the case with the highest identity was counted to avoid overestimating the frequency of the transferred function.

Networks of HGT. All pairs of genomes (donor and recipient) with significant signal of exchange were linked in networks that represented the extent of HGT. Networks were constructed using the Cytoscape V 2.8 software (Smoot et al 2011). Two networks were evaluated; one based on the whole-genome level analysis, to identify genomes that have undergone extreme HGT, and another based on the individual gene-level analysis, to find the most frequently transferred individual functions. Both HGT networks were analyzed using the Girvan-Newman greedy algorithm (Clauset et al 2004, Newman and Girvan 2004) as implemented in GLaY (Su et al 2010). This algorithm clusters the genomes into subnetworks that maximize the amount of connectivity (representing number of HGT events in this case).

To assess the level of enrichment of functional (e.g., type of respiration) or ecological (e.g., habitat of isolation) categories in the subnetworks against the expected distribution based on the abundance of each category in the complete dataset (i.e., prior to network analysis), the following approach was used. 1,000 replicate datasets were drawn, at random, from the complete dataset of all HGT events detected among all genomes from different phyla using the same number of genomes (556 and 810 genomes for the genome and gene level analyses, respectively) and the same number of HGT events as within each network. The abundance of categories in each replicate dataset (e.g., frequency of HGT events between a soil and a freshwater genome for assessing the effect of place of isolation) was quantified and a distribution was obtained based on these abundance values. The probability that the observed enrichment value from the subnetwork belonged to the latter distribution (p-value) was calculated by the Z-score test, using the one-tailed distribution.

Phylogenetic reconstruction. The phylogeny of 879 representative genomes was reconstructed using a similarity matrix built from the AAI values and the Neighbor Joining algorithm with 1000 bootstraps. Phylogenetic trees were visualized in Cytoscape V2.8 (Smoot et al 2011) and the putative partners of exchange were connected using an in-house Perl script. The resulting graph is shown in Figure 2, in which the lines representing the tree branches were removed for simplicity.

SUPPLEMENTARY FIGURES

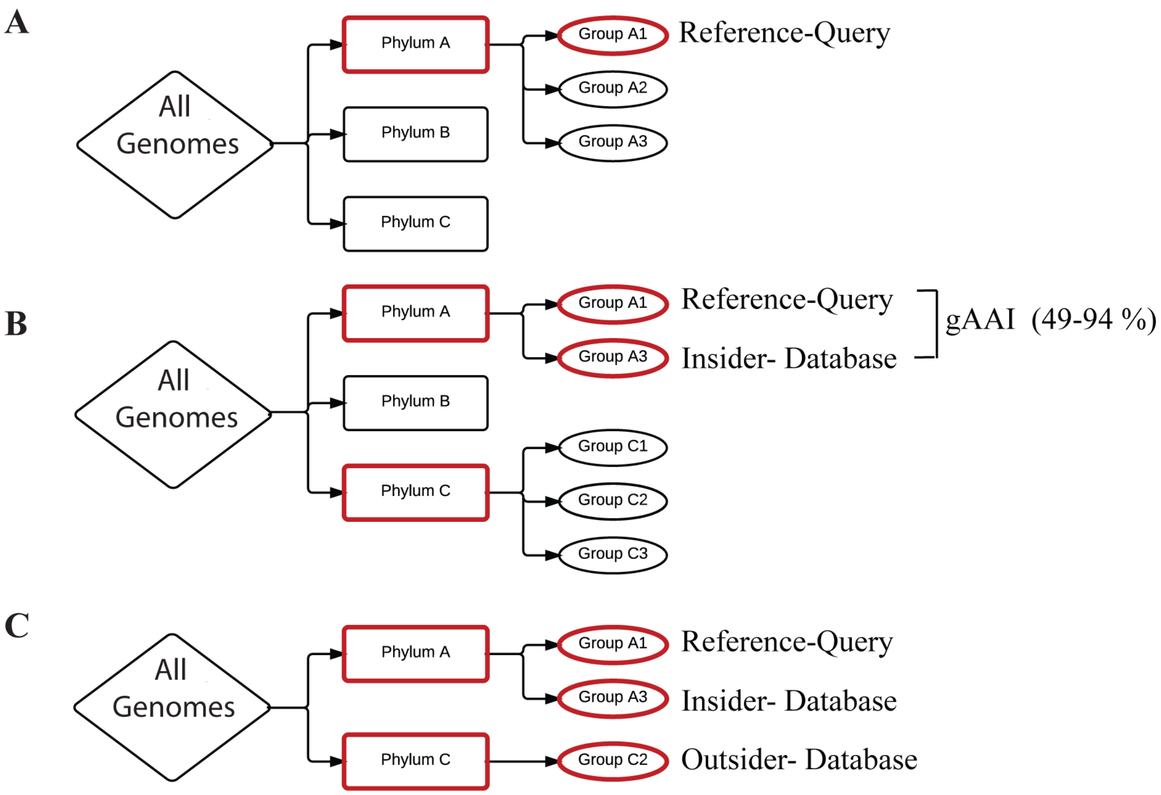

Figure S1. A schematic of the approach used to select genome triplets for assessing HGT between bacterial and archaeal phyla. The approach included the following steps: 1) randomly select a reference genome to begin to form a triplet of genomes **(Panel A)**; 2) select a second genome ("insider") representing the same phylum as the reference but from a different group based on gAAI **(Panel B)**; and finally, 3) select a genome representing a different phylum ("outsider") **(Panel C)**. The phylogenetic distance between the reference and insider genomes was measured by gAAI; all triplets characterized by similar gAAI values between the reference and insider genomes (-/+1% from the chosen gAAI values) formed a single set and were analyzed together (compared).

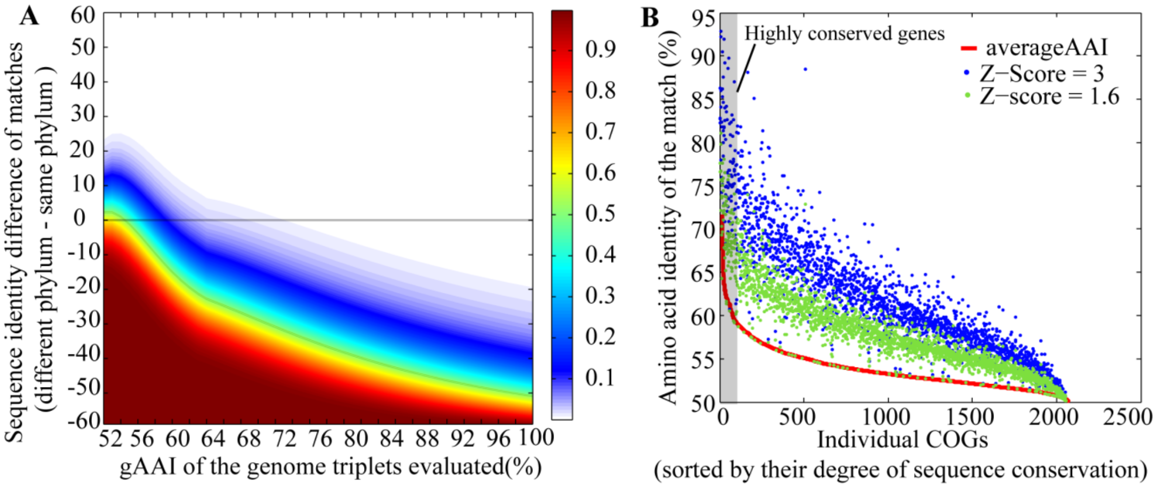

Figure S2. Identification of genes exchanged between bacterial and archaeal phyla with statistical confidence. Two different approaches were developed to evaluate the HGT signal for shared (reference gene has homologs in the two other genomes of a triplet) and non-shared genes (reference gene has homologs only in the outsider). For shared genes, a probabilistic model based on the distribution of amino acid sequence identity difference between the reference–insider match relative to the reference-outsider match was used to detect higher than expected identity of the reference genes with the outsider, which were identified as HGT events (see Material and Methods for details; **Panel A)**. For non-shared genes, the distribution of sequence identities was based on homologs from all genomes that showed similar gAAI to the reference-outsider pair assigned to the same (individual) COG (**Panel B)**. The plot shows the average amino acid identity between the homologs for each COG (red line), green dots represent 1.6 standard deviations from the average, and blue dots represent 3 standard deviations from the average. The latter threshold was used to identify HGT events (after correcting it for multiple testing).

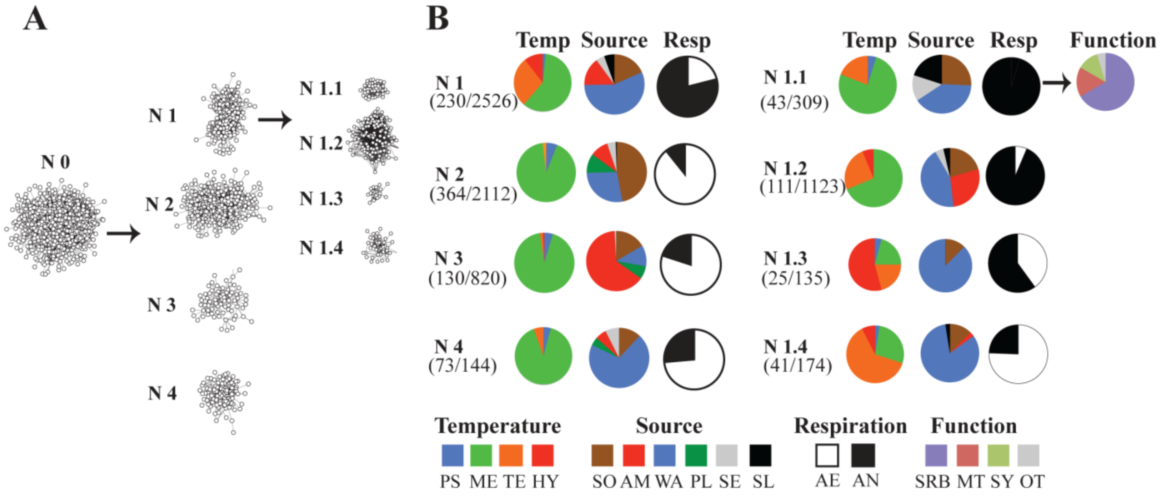

Figure S3. The effect of shared physiology and ecology on the structure of HGT networks. A network representing all inter-phylum HGT events was obtained as described in the main text and was divided into subnetworks using the community-clustering algorithm (GLaY) (Clauset et al 2004, Newman and Girvan 2004) that maximizes the connectivity between network nodes. Four subnetworks were obtained at the genome level analysis (N1, N2, N3, N4). Subnetwork N1 encompassed the highest number of anaerobic representatives and was further subdivided using GLaY. Four subnetworks were obtained (N1.1, N1.2, N1.3, N1.4**; Panel A)**. The optimal growth temperature (Temp), source of isolation (Source) and type of respiration (Resp), was extracted from the literature for all genomes in each subnetwork **(Panel B)** and categorized as follows: I) for optimal growth temperature category: psycrophilic (PS), mesophilic (ME), thermophilic (TE), and hyperthermophilic (HY). II) For source of isolation: soil (SO), animal associated (AM), aquatic (WA), plant (PL), sediment (SE), and sludge-bioreactor (SL); III) for respiration: aerobic (AE) and anaerobic (AN). The data revealed that the organisms grouped in network N1.1 had predominantly syntrophic interactions among themselves and were categorized further by their metabolic function (Function) to sulfate reducing bacteria (SRB), methanogens (MT), general syntrophic-secondary fermenting bacteria (SY) or other functions (OT). Note that respiration type separates more clearly subnetwork N1 from N2, N3 and N4 than the other categories evaluated. Also the subdivision of N1 creates two subnetworks that largely encompass syntrophic (N1.1) and fermentative organisms (N1.2).

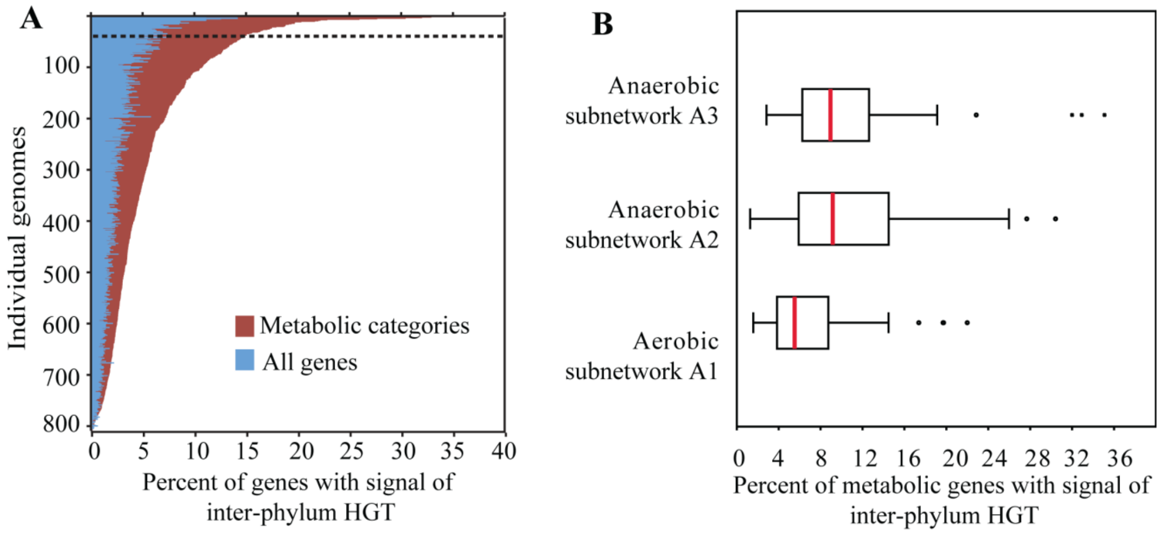

Figure S4. Frequency of inter-phylum HGT per genome and gene. Each bar represents one genome; the red portions of the bar represent the proportion of metabolic genes exchanged (i.e., the number of metabolic genes exchanged divided by the total number of metabolic genes in the genome); the blue portion represents the proportion of all genes exchanged (e.g., the number of genes exchanged divided by the total number of genes in the genome). Genomes are sorted by the percentage of genes exchanged. The dashed line represents the *Sphaerochaeta-Clostridia* case reported previously (Caro-Quintero et al 2012) **(Panel A**). The box plots represent the distribution of the percentages of metabolic genes exchanged for each genome triplet grouped by subnetwork (gene-level analysis; **Panel B**). The red line denotes the median, the left and right box boundaries represent the lower and upper quartiles and the whisker delimit the 97% percentile of the data, dots represent outliers. Note that the median of anaerobic networks A2 and A3 is almost twice as high as that of aerobic network A1.

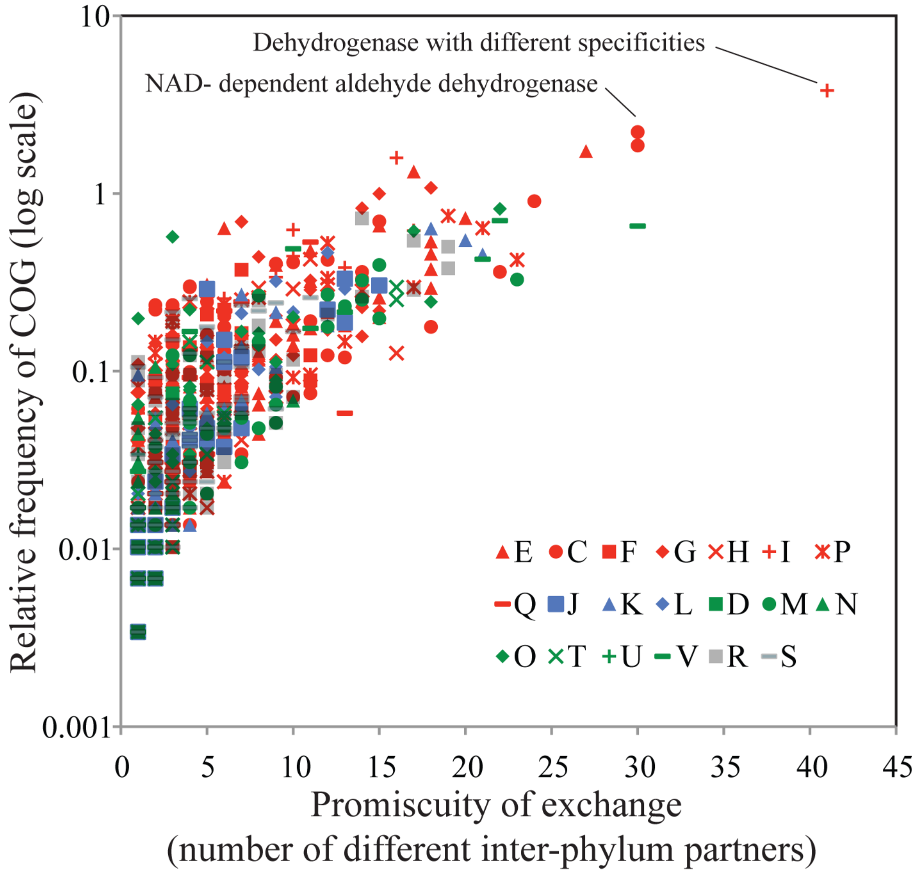

Figure S5. Relationship between frequency of HGT and promiscuity. All exchanged genes (q-value < 0.005) were assigned to an individual COG and the relative abundance of the COG (y-axis) is plotted against the number of different phyla partners (promiscuity) that exchanged the genes assigned to the COG (x-axis). Red symbols represent metabolic categories, green symbols represent cellular processes and signaling, blue symbols represent informational storage and processing, and gray symbols represent poorly characterized functions. Note that the higher the frequency of exchange, the higher usually the promiscuity of the exchanged (i.e., more different genomes exchanged the corresponding genes/COG). For instance, the "NAD-dependent aldehyde dehydrogenase" one of the most transferred categories has been exchanged across 30 different pairs of phyla. For the annotation of the letter of each COG functional category, please see Table S10.

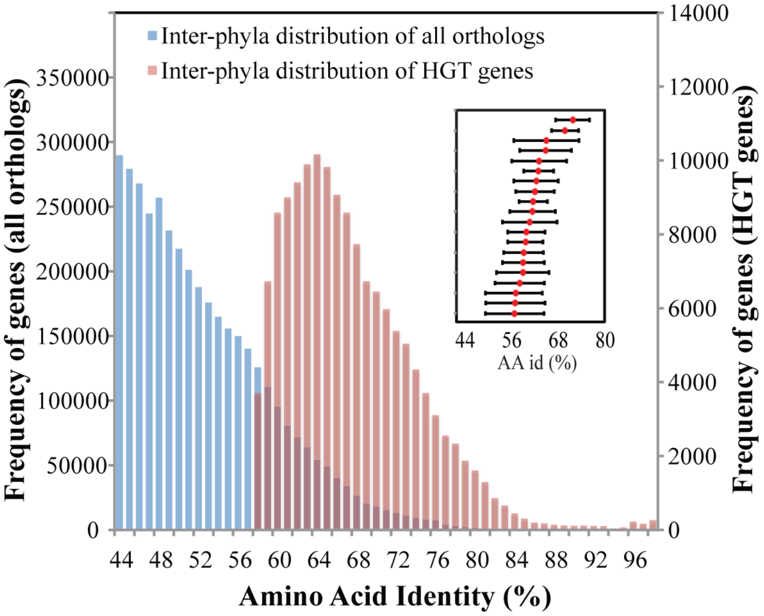

Figure S6. Assessing the effect of un-equal substitution rates in detecting HGT. The graph represents the distribution of amino acid identities of all shared orthologs not detected as horizontally transferred (blue columns) and all genes detected as transferred (red columns) between the reference and the outsider genomes for triplets with 58% gAAI between the reference and the insider genomes. The inset shows the inter-phyla average amino acid identity of the 20 most highly conserved orthologs in terms of their degree of sequence conservation and the error bars represent 1.6 standard deviations from the average. Note that the identity of transferred genes is equal or higher than that of highly conserved orthologs, which suggests that un-equal substitution rates in the insider genomes do not likely account for a substantial part of the HGT patterns observed (see also main text for more details).

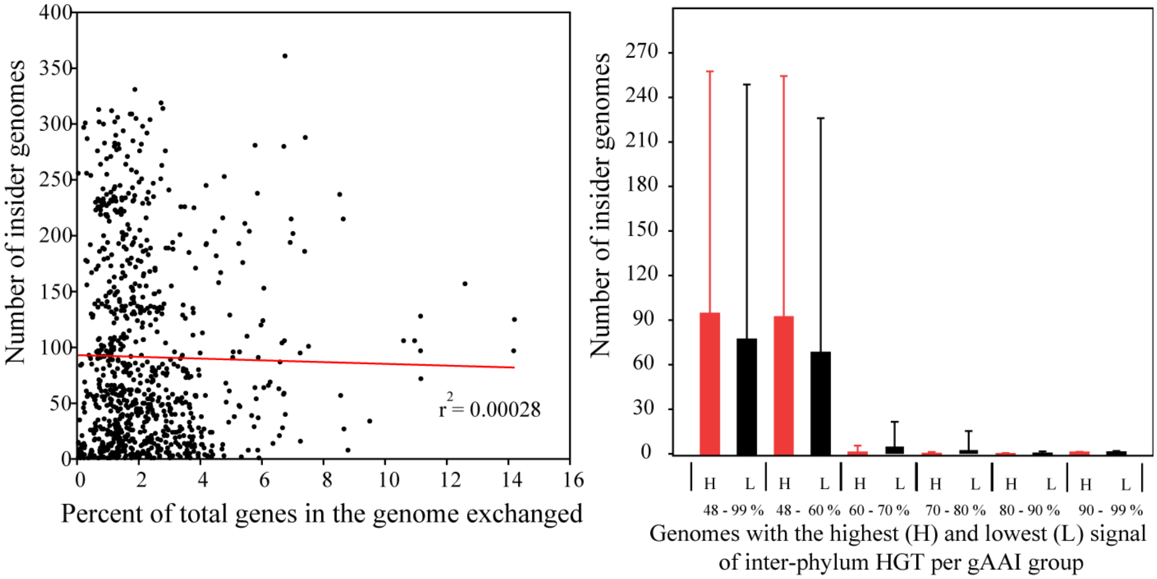

Figure S7. The influence of the number of insider genomes for each phylum on the number of genes exchanged between phyla. The correlation between the number of insider genomes and the percentage of the total genes in the reference genome with a signal of inter-phylum horizontal transfer was assessed based on all available genomes (Panel A). The statistical significance of the correlation, evaluated using parametric (Pearson's $r^2 = 0.00028$; shown on the graph) and non-parametric methods (Kendall's p-value= 0.293, Spearman's p-value = 0.331), was not significant. The average and standard deviation of the number of insider genomes for the 100 genomes with the highest (H) and lowest (L) percentage of exchanged genes were also not significantly different (p-value > 0.1; Panel B) for insider genomes of varied gAAI to the reference genome. These results suggested 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 but represent real HGT events. These results were consistent with our expectations given that the probabilistic models used in the detection of transferred genes (e.g., Fig. 1) are independent of the number of representatives (insider genomes used).

SUPPLEMENTARY TABLES

Table S1. The significance (or level of enrichment) of isolation source in the frequency of inter-phylum HGT as detected by the genome level approach. The genome-level network was based on a total of 11,808 HGT events detected among 556 genomes, and the observed values (5th column) were calculated using the frequency of the categories (1st and 2nd columns) that the genomes, which participated in the subset of these HGT events grouped under each subnetwork, were assigned to.

Reference	Outsider	Expected average	Expected standard deviation	Observed	Z-score	p-value
soil	soil	971.5	11.2	1165	17.2	***+
mammal	mammal	276.1	16.7	549	16.3	***+
insects	mammal	8.8	3.3	56	14.2	***+
sediment	sludge-waste_water-bioreactors	23.0	5.3	95	13.6	***+
sediment	fresh_water	200.9	14.3	388	13.1	***+
sludge-waste_water-bioreactors	sludge-waste_water-bioreactors	11.5	3.8	60	12.9	***+
sludge-waste_water-bioreactors	fresh_water	101.1	10.9	236	12.4	***+
sediment	sediment	44.2	7.0	119	10.6	***+
soil	plant	215.0	15.3	375	10.5	***+
soil	sludge-waste_water-bioreactors	105.6	11.2	209	9.2	***+
sludge-waste_water-bioreactors	sediment	22.3	4.9	67	9.1	***+
fresh_water	sludge-waste_water-bioreactors	92.6	10.2	172	7.8	***+
plant	plant	40.6	6.9	82	6.0	***+
insects	soil	14.9	4.1	36	5.2	***+
fresh_water	fresh_water	788.6	27.7	922	4.8	***+
marine	sludge-waste_water-bioreactors	46.7	7.1	79	4.5	***+
insects	sludge-waste_water-bioreactors	2.1	1.1	7	4.3	***+
sediment	marine	105.6	10.9	151	4.2	***+
insects	sediment	2.8	1.5	9	4.1	***+
marine	sediment	89.0	9.8	128	4.0	***+
sludge-waste_water-bioreactors	marine	53.9	7.8	79	3.2	**+
soil	mammal	538.5	23.0	611	3.2	**+
fresh_water	sediment	180.0	13.7	223	3.1	**+
soil	fresh_water	903.2	29.5	994	3.1	**+
mammal	sludge-waste_water-bioreactors	56.7	8.1	80	2.9	*+

soil	sediment	204.9	14.4	245	2.8	*+
soil	insects	10.0	3.3	19	2.7	*+
insects	fresh_water	13.4	4.0	24	2.7	*+
plant	mammal	151.8	13.3	184	2.4	*+
marine	fresh_water	384.1	20.2	429	2.2	NS
sediment	mammal	123.4	11.5	146	2.0	NS
sludge-waste_water-bioreactors	soil	111.7	11.4	134	2.0	NS
sediment	soil	219.0	15.5	245	1.7	NS
insects	plant	2.2	1.2	4	1.5	NS
marine	insects	4.5	2.2	7	1.1	NS
sludge-waste_water-bioreactors	mammal	62.5	8.5	72	1.1	NS
plant	soil	264.7	17.4	282	1.0	NS
marine	marine	189.3	14.8	198	0.6	NS
plant	sludge-waste_water-bioreactors	27.3	5.5	30	0.5	NS
mammal	insects	5.9	2.5	6	0.0	NS
insects	marine	7.3	2.9	6	-0.4	NS
sludge-waste_water-bioreactors	plant	23.0	4.9	20	-0.6	NS
sludge-waste_water-bioreactors	insects	1.6	0.8	1	-0.8	NS
fresh_water	insects	8.7	3.1	6	-0.9	NS
plant	insects	1.9	1.0	1	-0.9	NS
sediment	insects	2.3	1.2	1	-1.0	NS
fresh_water	marine	407.9	20.4	365	-2.1	NS
plant	sediment	50.8	7.3	33	-2.5	*-
soil	marine	469.4	22.1	374	-4.3	***-
sediment	plant	43.0	6.7	7	-5.4	***-
mammal	sediment	109.4	11.0	50	-5.4	***-
mammal	plant	118.5	11.1	54	-5.8	***-
plant	marine	127.1	12.2	56	-5.8	***-
fresh_water	plant	191.6	14.4	88	-7.2	***-
marine	plant	93.9	10.4	16	-7.5	***-
mammal	soil	511.0	23.5	327	-7.8	***-
fresh_water	mammal	482.5	22.9	298	-8.1	***-
plant	fresh_water	244.4	16.3	106	-8.5	***-
mammal	marine	241.4	16.4	84	-9.6	***-
marine	mammal	231.6	15.9	79	-9.6	***-
marine	soil	425.8	21.1	206	-10.4	***-
fresh_water	soil	871.2	30.4	536	-11.0	***-
mammal	fresh_water	473.6	22.4	177	-13.3	***-

*** P-value <0.0001, ** P-value <0.001, * P-value <0.01, + higher than expected, - lower than expected

Table S2. The significance (or level of enrichment) of oxygen tolerance and optimal growth temperature in the frequency of inter-phylum HGT as detected by the genome level approach. The genome-level network was based on a total of 6,668 HGT events detected among 556 genomes, and the observed values (5th column) were calculated using the frequency of the categories (1st and 2nd columns) that the genomes, which participated in the subset of these HGT events grouped under each subnetwork, were assigned to. The number of HGT events was higher for the isolation source categories (Table S1) compared to the categories shown here because organisms were frequently assigned to multiple isolation sources (i.e., they were not specific to one habitat).

Reference	Outsider	Expected average	Expected standard deviation	Observed	Z-score	p-value
anaerobic respiration hyperthermopilic	anaerobic respiration hyperthermopilic	1.4	0.7	62	91.3	***+
anaerobic respiration thermophilic	anaerobic respiration thermophilic	10.2	2.0	160	75.7	***+
anaerobic respiration hyperthermopilic	aerobic respiration hyperthermopilic	1.3	0.6	44	72.6	***+
anaerobic respiration thermophilic	anaerobic respiration hyperthermopilic	3.2	1.7	90	51.6	***+
aerobic respiration hyperthermopilic	anaerobic respiration hyperthermopilic	1.3	0.5	23	39.9	***+
anaerobic respiration mesophilic	anaerobic respiration thermophilic	82.6	8.8	392	35.3	***+
anaerobic respiration mesophilic	anaerobic respiration mesophilic	455.6	20.6	1174	34.9	***+
aerobic respiration thermophilic	aerobic respiration thermophilic	10.9	3.3	85	22.4	***+
anaerobic respiration hyperthermopilic	anaerobic respiration thermophilic	3.0	1.5	32	19.5	***+
aerobic respiration thermophilic	anaerobic respiration hyperthermopilic	3.4	1.7	31	16.1	***+
anaerobic respiration mesophilic	anaerobic respiration hyperthermopilic	24.5	5.1	102	15.3	***+
anaerobic respiration thermophilic	anaerobic respiration mesophilic	70.7	8.5	180	12.9	***+
aerobic respiration thermophilic	aerobic respiration hyperthermopilic	2.3	1.3	17	11.7	***+
aerobic respiration thermophilic	anaerobic respiration thermophilic	11.6	3.3	47	10.7	***+
anaerobic respiration thermophilic	aerobic respiration thermophilic	11.9	3.5	43	8.8	***+
anaerobic respiration thermophilic	aerobic respiration hyperthermopilic	2.5	1.3	14	8.8	***+
aerobic respiration hyperthermopilic	aerobic respiration thermophilic	2.2	1.2	12	8.4	***+
anaerobic respiration hyperthermopilic	aerobic respiration thermophilic	3.4	1.7	14	6.4	***+
aerobic respiration hyperthermopilic	aerobic respiration hyperthermopilic	1.1	0.3	3	6.2	***+

aerobic respiration mesophilic	aerobic respiration mesophilic	2112.3	37.8	2338	6.0	***+
anaerobic respiration mesophilic	aerobic respiration hyperthermopilic	15.6	3.9	27	2.9	*+
anaerobic respiration mesophilic	anaerobic respiration psychrophilic	21.5	4.6	29	1.6	NS
anaerobic respiration thermophilic	anaerobic respiration psychrophilic	4.1	2.0	6	0.9	NS
aerobic respiration thermophilic	aerobic respiration mesophilic	163.8	12.3	158	-0.5	NS
aerobic respiration hyperthermopilic	anaerobic respiration thermophilic	2.2	1.2	1	-1.0	NS
aerobic respiration psychrophilic	aerobic respiration hyperthermopilic	2.5	1.3	1	-1.1	NS
anaerobic respiration psychrophilic	aerobic respiration psychrophilic	3.4	1.7	1	-1.4	NS
aerobic respiration psychrophilic	aerobic respiration psychrophilic	8.1	2.8	4	-1.5	NS
anaerobic respiration thermophilic	aerobic respiration psychrophilic	9.9	3.2	5	-1.6	NS
anaerobic respiration psychrophilic	anaerobic respiration hyperthermopilic	2.2	1.2	0	-1.8	NS
anaerobic respiration hyperthermopilic	anaerobic respiration psychrophilic	1.8	1.0	0	-1.8	NS
aerobic respiration psychrophilic	anaerobic respiration psychrophilic	2.8	1.5	0	-1.8	NS
anaerobic respiration hyperthermopilic	aerobic respiration psychrophilic	3.1	1.6	0	-1.9	NS
aerobic respiration hyperthermopilic	aerobic respiration psychrophilic	2.0	1.0	0	-1.9	NS
anaerobic respiration psychrophilic	aerobic respiration hyperthermopilic	1.6	0.8	0	-1.9	NS
anaerobic respiration psychrophilic	aerobic respiration thermophilic	5.7	2.3	1	-2.0	NS
anaerobic respiration psychrophilic	anaerobic respiration psychrophilic	1.5	0.7	0	-2.0	NS
anaerobic respiration psychrophilic	anaerobic respiration thermophilic	5.8	2.3	1	-2.0	NS
aerobic respiration thermophilic	anaerobic respiration psychrophilic	3.9	1.9	0	-2.1	NS
aerobic respiration psychrophilic	anaerobic respiration hyperthermopilic	3.7	1.8	0	-2.1	NS
aerobic respiration hyperthermopilic	anaerobic respiration psychrophilic	1.4	0.6	0	-2.3	*-
aerobic respiration thermophilic	aerobic respiration psychrophilic	10.0	3.3	2	-2.5	*-
aerobic respiration mesophilic	anaerobic respiration psychrophilic	42.3	6.5	25	-2.7	*-
aerobic respiration mesophilic	aerobic respiration psychrophilic	139.0	11.4	108	-2.7	*-
aerobic respiration mesophilic	aerobic respiration hyperthermopilic	39.7	5.9	23	-2.8	*-
aerobic respiration psychrophilic	aerobic respiration thermophilic	12.5	3.7	1	-3.2	**-
anaerobic respiration hyperthermopilic	anaerobic respiration mesophilic	20.2	4.4	6	-3.2	**-

aerobic respiration psychrophilic	anaerobic respiration thermophilic	12.3	3.5	0	-3.5	**-
anaerobic respiration mesophilic	aerobic respiration thermophilic	83.6	9.2	51	-3.5	**-
aerobic respiration hyperthermopilic	anaerobic respiration mesophilic	12.2	3.4	0	-3.6	**-
aerobic respiration hyperthermopilic	aerobic respiration mesophilic	29.2	5.3	9	-3.8	***-
anaerobic respiration mesophilic	aerobic respiration psychrophilic	60.1	7.4	30	-4.1	***-
anaerobic respiration psychrophilic	anaerobic respiration mesophilic	27.5	5.2	2	-4.9	***-
aerobic respiration thermophilic	anaerobic respiration mesophilic	71.1	8.6	29	-4.9	***-
anaerobic respiration psychrophilic	aerobic respiration mesophilic	50.5	7.1	8	-6.0	***-
aerobic respiration mesophilic	anaerobic respiration hyperthermopilic	67.1	8.5	10	-6.7	***-
anaerobic respiration hyperthermopilic	aerobic respiration mesophilic	53.0	7.5	1	-7.0	***-
aerobic respiration psychrophilic	anaerobic respiration mesophilic	64.3	8.1	2	-7.7	***-
aerobic respiration mesophilic	aerobic respiration thermophilic	209.4	13.9	87	-8.8	***-
anaerobic respiration thermophilic	aerobic respiration mesophilic	176.5	13.1	53	-9.4	***-
aerobic respiration psychrophilic	aerobic respiration mesophilic	134.5	11.5	22	-9.8	***-
anaerobic respiration mesophilic	aerobic respiration mesophilic	988.4	28.9	661	-11.3	***-
aerobic respiration mesophilic	anaerobic respiration thermophilic	219.3	14.9	27	-12.9	***-
aerobic respiration mesophilic	anaerobic respiration mesophilic	1088.7	29.7	414	-22.7	***-

*** P-value <0.0001, ** P-value <0.001, * P-value <0.01, + higher than expected, - lower than expected

Table S3. The significance (or level of enrichment) of isolation source in the frequency of inter-phylum HGT as detected by the gene level approach. The gene-level network was based on a total of 119,635 HGT events detected among 810 genomes, and the observed values (5th column) were calculated using the frequency of the categories (1st and 2nd columns) that the genomes, which participated in the subset of these HGT events grouped under each subnetwork, were assigned to.

Reference	Outsider	Expected average	Expected standard deviation	Observed	Z-score	p-value
soil	soil	8647.9	49.9	11690	61.0	***+
soil	fresh_water	8020.8	82.7	11680	44.2	***+
sediment	marine	1091.4	33.2	2450	40.9	***+
plant	soil	2085.3	48.5	3521	29.6	***+
sediment	fresh_water	2014.9	49.8	3413	28.1	***+
soil	marine	4367.4	62.7	6010	26.2	***+
plant	fresh_water	1886.2	48.0	2823	19.5	***+
sediment	sediment	511.6	24.1	935	17.6	***+
fresh_water	fresh_water	6951.3	89.0	8500	17.4	***+
fresh_water	marine	3786.5	63.6	4814	16.2	***+
soil	insects	176.2	12.8	360	14.4	***+
sludge-waste_water-bioreactors	marine	963.4	31.0	1372	13.2	***+
sludge-waste_water-bioreactors	fresh_water	1766.4	48.9	2381	12.6	***+
mammal	insects	114.2	12.3	244	10.6	***+
soil	sediment	2113.3	48.0	2538	8.8	***+
sediment	sludge-waste_water-bioreactors	467.7	20.3	633	8.1	***+
marine	marine	1959.4	43.4	2266	7.1	***+
sludge-waste_water-bioreactors	sediment	457.2	20.7	591	6.5	***+
soil	sludge-waste_water-bioreactors	1873.7	49.9	2185	6.2	***+
sludge-waste_water-bioreactors	sludge-waste_water-bioreactors	401.9	22.9	537	5.9	***+
sludge-waste_water-bioreactors	insects	38.4	6.3	73	5.5	***+
fresh_water	sediment	1844.3	47.6	2065	4.6	***+
mammal	mammal	3185.1	61.5	3454	4.4	***+
fresh_water	sludge-waste_water-bioreactors	1643.8	43.1	1792	3.4	**+
plant	marine	1012.5	40.0	1134	3.0	*+
sediment	insects	41.6	7.0	56	2.1	NS
fresh_water	insects	149.5	14.8	178	1.9	NS
plant	sludge-waste_water-bioreactors	429.6	21.9	467	1.7	NS

sludge-waste_water-bioreactors	soil	1946.1	45.9	1973	0.6	NS
insects	soil	192.4	13.2	199	0.5	NS
insects	mammal	130.8	11.0	136	0.5	NS
plant	plant	292.8	17.1	292	0.0	NS
plant	insects	33.5	6.6	31	-0.4	NS
sediment	soil	2235.8	48.5	2212	-0.5	NS
plant	sediment	458.7	21.8	415	-2.0	NS
insects	sediment	43.6	6.9	26	-2.5	*-
insects	sludge-waste_water-bioreactors	40.7	6.5	24	-2.6	*-
marine	sediment	982.9	36.0	871	-3.1	**-
insects	plant	29.4	6.2	9	-3.3	**-
marine	insects	77.7	9.8	43	-3.5	**-
insects	fresh_water	171.6	15.1	118	-3.5	**-
insects	marine	90.9	9.9	55	-3.6	**-
fresh_water	soil	7799.2	95.1	7065	-7.7	***-
marine	fresh_water	3707.1	65.0	3194	-7.9	***-
sludge-waste_water-bioreactors	mammal	1244.5	33.5	935	-9.2	***-
mammal	sludge-waste_water-bioreactors	1134.4	39.2	750	-9.8	***-
marine	sludge-waste_water-bioreactors	881.0	27.3	576	-11.2	***-
mammal	sediment	1302.6	37.2	864	-11.8	***-
sludge-waste_water-bioreactors	plant	369.8	19.4	136	-12.1	***-
soil	plant	1716.4	43.5	1177	-12.4	***-
sediment	plant	393.3	20.4	135	-12.7	***-
sediment	mammal	1453.1	37.6	926	-14.0	***-
mammal	marine	2658.7	48.5	1789	-17.9	***-
plant	mammal	1439.9	39.6	663	-19.6	***-
mammal	soil	5235.2	78.6	3618	-20.6	***-
mammal	fresh_water	4833.3	75.7	3230	-21.2	***-
marine	plant	789.8	28.8	170	-21.5	***-
fresh_water	plant	1506.1	43.2	548	-22.2	***-
mammal	plant	1145.4	37.5	272	-23.3	***-
soil	mammal	5502.8	78.4	3592	-24.4	***-
marine	soil	4158.2	58.7	2310	-31.5	***-
fresh_water	mammal	4961.1	81.7	2310	-32.5	***-
marine	mammal	2674.2	55.1	809	-33.9	***-

*** P-value <0.0001, ** P-value <0.001, * P-value <0.01, + higher than expected, - lower than expected

Table S4. The significance (or level of enrichment) of oxygen tolerance and optimal growth temperature in the frequency of inter-phylum HGT as detected by the gene level approach. The gene-level network was based on a total of 67,951 HGT events detected among 810 genomes, and the observed values (5th column) were calculated using the frequency of the categories (1st and 2nd columns) that the genomes, which participated in the subset of these HGT events grouped under each subnetwork, were assigned to. The number of HGT events was higher for the isolation source categories (Table S1) compared to the categories shown here because organisms were frequently assigned to multiple isolation sources (i.e., they were not specific to one habitat).

Reference	Outsider	Expected average	Expected standard deviation	Observed	Z-score	p-value
anaerobic respiration mesophilic	anaerobic respiration thermophilic	972.0	32.3	3503	78.5	***+
anaerobic respiration mesophilic	anaerobic respiration mesophilic	4001.3	65.7	8480	68.2	***+
anaerobic respiration thermophilic	anaerobic respiration thermophilic	157.6	12.3	825	54.2	***+
aerobic respiration mesophilic	aerobic respiration thermophilic	2339.4	47.3	4325	42.0	***+
anaerobic respiration thermophilic	aerobic respiration thermophilic	179.7	11.2	566	34.4	***+
aerobic respiration thermophilic	aerobic respiration thermophilic	138.6	11.5	380	20.9	***+
anaerobic respiration thermophilic	anaerobic respiration hyperthermopilic	81.1	7.7	240	20.7	***+
aerobic respiration mesophilic	aerobic respiration mesophilic	20950.9	116.4	23266	19.9	***+
anaerobic respiration mesophilic	aerobic respiration thermophilic	902.7	28.9	1444	18.8	***+
anaerobic respiration thermophilic	anaerobic respiration mesophilic	856.9	29.7	1282	14.3	***+
anaerobic respiration hyperthermopilic	anaerobic respiration hyperthermopilic	27.2	5.4	99	13.3	***+
anaerobic respiration hyperthermopilic	anaerobic respiration thermophilic	72.7	8.7	175	11.7	***+
aerobic respiration hyperthermopilic	aerobic respiration thermophilic	29.1	5.6	90	10.8	***+
anaerobic respiration psychrophilic	anaerobic respiration thermophilic	57.0	7.9	88	3.9	***+
anaerobic respiration psychrophilic	anaerobic respiration mesophilic	199.0	16.0	245	2.9	*+
aerobic respiration thermophilic	anaerobic respiration thermophilic	167.7	13.9	198	2.2	NS
anaerobic respiration psychrophilic	aerobic respiration thermophilic	51.1	7.3	63	1.6	NS
anaerobic respiration hyperthermopilic	aerobic respiration thermophilic	70.7	8.2	82	1.4	NS
aerobic respiration psychrophilic	aerobic respiration thermophilic	137.8	11.2	144	0.5	NS
aerobic respiration	anaerobic respiration	152.1	13.1	156	0.3	NS

psychrophilic	thermophilic					
anaerobic respiration psychrophilic	anaerobic respiration psychrophilic	4.4	1.9	2	-1.3	NS
aerobic respiration psychrophilic	anaerobic respiration psychrophilic	17.5	4.7	9	-1.8	NS
anaerobic respiration psychrophilic	aerobic respiration psychrophilic	22.0	4.7	12	-2.1	NS
aerobic respiration psychrophilic	anaerobic respiration mesophilic	570.0	21.8	520	-2.3	NS
aerobic respiration thermophilic	aerobic respiration hyperthermopilic	31.3	5.1	19	-2.4	*-
anaerobic respiration psychrophilic	aerobic respiration hyperthermopilic	11.0	3.5	2	-2.6	**-
aerobic respiration hyperthermopilic	anaerobic respiration hyperthermopilic	12.5	3.9	2	-2.6	**-
anaerobic respiration hyperthermopilic	anaerobic respiration psychrophilic	17.1	4.6	3	-3.0	**-
aerobic respiration psychrophilic	aerobic respiration psychrophilic	67.0	8.2	39	-3.4	**-
anaerobic respiration psychrophilic	aerobic respiration mesophilic	382.2	19.6	310	-3.7	**-
aerobic respiration hyperthermopilic	anaerobic respiration thermophilic	33.8	5.7	12	-3.8	***-
anaerobic respiration thermophilic	anaerobic respiration psychrophilic	41.6	6.6	16	-3.9	***-
aerobic respiration hyperthermopilic	aerobic respiration psychrophilic	23.1	4.9	1	-4.5	***-
anaerobic respiration psychrophilic	anaerobic respiration hyperthermopilic	25.8	5.1	3	-4.5	***-
aerobic respiration psychrophilic	aerobic respiration hyperthermopilic	30.4	5.9	2	-4.8	***-
aerobic respiration mesophilic	anaerobic respiration thermophilic	2730.4	50.4	2464	-5.3	***-
anaerobic respiration thermophilic	aerobic respiration hyperthermopilic	38.3	6.4	3	-5.5	***-
aerobic respiration thermophilic	anaerobic respiration psychrophilic	36.3	5.8	4	-5.6	***-
aerobic respiration psychrophilic	aerobic respiration mesophilic	1217.2	35.7	1001	-6.1	***-
aerobic respiration thermophilic	anaerobic respiration hyperthermopilic	74.0	9.1	18	-6.2	***-
anaerobic respiration mesophilic	anaerobic respiration psychrophilic	156.4	13.0	74	-6.3	***-
anaerobic respiration hyperthermopilic	aerobic respiration psychrophilic	50.9	6.7	3	-7.2	***-
aerobic respiration psychrophilic	anaerobic respiration hyperthermopilic	68.1	8.1	3	-8.0	***-
aerobic respiration thermophilic	aerobic respiration psychrophilic	107.8	11.1	17	-8.2	***-
anaerobic respiration thermophilic	aerobic respiration psychrophilic	122.6	12.8	12	-8.6	***-
anaerobic respiration mesophilic	aerobic respiration hyperthermopilic	198.2	13.0	86	-8.7	***-
anaerobic respiration hyperthermopilic	anaerobic respiration mesophilic	357.7	20.9	150	-9.9	***-

anaerobic respiration mesophilic	anaerobic respiration hyperthermopilic	438.1	20.9	219	-10.5	***-
aerobic respiration hyperthermopilic	anaerobic respiration mesophilic	161.7	12.8	22	-10.9	***-
aerobic respiration mesophilic	anaerobic respiration psychrophilic	319.9	16.0	133	-11.7	***-
aerobic respiration mesophilic	aerobic respiration hyperthermopilic	541.5	21.8	263	-12.7	***-
aerobic respiration thermophilic	anaerobic respiration mesophilic	760.5	25.0	427	-13.3	***-
anaerobic respiration mesophilic	aerobic respiration psychrophilic	525.4	21.7	164	-16.7	***-
aerobic respiration mesophilic	anaerobic respiration mesophilic	10101.6	83.4	8708	-16.7	***-
aerobic respiration mesophilic	aerobic respiration psychrophilic	1226.1	35.7	598	-17.6	***-
aerobic respiration hyperthermopilic	aerobic respiration mesophilic	408.3	19.6	43	-18.7	***-
aerobic respiration thermophilic	aerobic respiration mesophilic	1822.7	43.7	630	-27.3	***-
aerobic respiration mesophilic	anaerobic respiration hyperthermopilic	1204.1	35.4	166	-29.4	***-
anaerobic respiration hyperthermopilic	aerobic respiration mesophilic	920.0	28.2	67	-30.3	***-
anaerobic respiration thermophilic	aerobic respiration mesophilic	2247.5	46.7	701	-33.1	***-
anaerobic respiration mesophilic	aerobic respiration mesophilic	9258.8	86.7	5372	-44.8	***-
aerobic respiration hyperthermopilic	aerobic respiration hyperthermopilic	2.4	1.3	0	-	NA
aerobic respiration hyperthermopilic	anaerobic respiration psychrophilic	6.3	2.3	0	-	NA
anaerobic respiration hyperthermopilic	aerobic respiration hyperthermopilic	10.2	3.1	0	-	NA

*** P-value <0.0001, ** P-value <0.001, * P-value <0.01, + higher than expected, - lower than expected

Table S5. Organisms with the highest percentage of genes acquired from organisms of different phyla. Organisms are ranked by the number of genes with signal of HGT, reported as the fraction of the total genes in the genome.

Genome name	Optimal growth temperature	Oxygen Tolerance	Metabolic categories (%)	Total genome (%)
Ilyobacter polytropus DSM 2926 uid59769	mesophilic	anaerobic	35.1	16.2
Leptotrichia buccalis C 1013 b uid59211	mesophilic	anaerobic	32.9	11.1
Sebaldella termitidis ATCC 33386 uid41865	mesophilic	Anae`robic	32.0	11.0
Desulfurispirillum indicum S5 uid45897	mesophilic	anaerobic	30.4	14.2
Thermodesulfatator indicus DSM 15286 uid68285	thermophilic	anaerobic	27.7	12.6
Deferribacter desulfuricans SSM1 uid46653	thermophilic	anaerobic	26.0	10.6
Fusobacterium nucleatum ATCC 25586 uid57885	mesophilic	aerobic	22.9	11.2
Thermodesulfovibrio yellowstonii DSM 11347 uid59257	thermophilic	aerobic	22.2	11.2
Candidatus Solibacter usitatus Ellin6076 uid58139	mesophilic	aerobic	22.0	5.9
Geobacter sulfurreducens KN400 uid161977	mesophilic	anaerobic	21.7	8.6
Candidatus Nitrospira defluvii uid51175	mesophilic	anaerobic	20.3	8.7
Thermaerobacter marianensis DSM 12885 uid61727	hyperthermopilic	aerobic	19.7	8.5
Rubrobacter xylanophilus DSM 9941 uid58057	thermophilic	aerobic	19.7	9.5
Rhodothermus marinus DSM 4252 uid41729	thermophilic	aerobic	19.7	7.2
Calditerrivibrio nitroreducens DSM 19672 uid60821	thermophilic	anaerobic	19.4	8.6
Eggerthella lenta DSM 2243 uid59079	mesophilic	anaerobic	19.1	6.7
Denitrovibrio acetiphilus DSM 12809 uid46657	mesophilic	anaerobic	18.8	6.8
Geobacter uraniireducens Rf4 uid58475	mesophilic	anaerobic	18.8	7.0
Slackia heliotrinireducens DSM 20476 uid59051	mesophilic	anaerobic	18.8	6.6
Desulfotomaculum kuznetsovii DSM 6115 uid67357	mesophilic	anaerobic	18.7	7.5
Heliobacterium modesticaldum Ice1 uid58279	thermophilic	anaerobic	17.9	6.1
Ammonifex degensii KC4 uid41053	thermophilic	anaerobic	17.5	7.2
Anaerobaculum mobile DSM 13181 uid168323	thermophilic	anaerobic	17.5	8.8
Gemmatimonas aurantiaca T 27 uid58813	mesophilic	aerobic	17.4	5.9
Treponema primitia ZAS 2 uid67367	mesophilic	anaerobic	17.3	5.1
Eggerthella YY7918 uid68707	mesophilic	anaerobic	17.1	6.2
Treponema brennaborense DSM 12168 uid66607	mesophilic	anaerobic	16.7	6.3
Granulicella mallensis MP5ACTX8 uid49957	mesophilic	aerobic	16.6	6.4
Treponema succinifaciens DSM 2489 uid65781	mesophilic	anaerobic	16.4	5.1
Flexistipes sinusarabici DSM 4947 uid68147	thermophilic	anaerobic	16.4	6.7
Geobacter metallireducens GS 15 uid57731	mesophilic	anaerobic	16.2	6.9
Clostridium clariflavum DSM 19732 uid82345	thermophilic	anaerobic	15.6	4.8
Desulfurivibrio alkaliphilus AHT2 uid49487	mesophilic	anaerobic	15.5	6.1
Thermosediminibacter oceani DSM 16646 uid51421	thermophilic	anaerobic	15.3	7.4
Desulfobulbus propionicus DSM 2032 uid62265	mesophilic	anaerobic	15.1	5.3
Pelobacter carbinolicus DSM 2380 uid58241	mesophilic	anaerobic	15.1	6.7
Sphaerochaeta pleomorpha Grapes uid82365 **	mesophilic	anaerobic	15.0	5.9

Table S6. Comparison of the frequency of inter-phylum HGT between the most transferred metabolic and informational genes used to resolved the Tree of Life.

Functional group (COGs)	Functional Classification	Frequency in HGT genes (%)	Frequency in the genome (%)	Ratio (In HGT / In genome)
COG0080	Informational	0.039	0.059	0.654
COG0012	Informational	0.013	0.059	0.219
COG0018	Informational	0.010	0.056	0.173
COG0172	Informational	0.010	0.061	0.160
COG0522	Informational	0.006	0.061	0.105
COG0495	Informational	0.003	0.055	0.058
Total		0.081	0.351	
COG1028	Metabolic	3.793	0.731	5.185
COG1012	Metabolic	2.215	0.361	6.139
COG0667	Metabolic	1.860	0.147	12.681
COG1126	Metabolic	1.731	0.145	11.965
COG0183	Metabolic	1.587	0.186	8.557
COG0129	Metabolic	1.328	0.082	16.213
Total		12.515	1.651	
Ratio (Metabolic/Informational)		155.4	4.7	

Table S7. Detected cases of inter-phyla HGT of highly conserved housekeeping genes.

Functional category (COGs)	Accession number (gi)	Detected partners of exchange
Predicted GTPase (COG0012)	114331141	*Nitrosomonas eutropha <-> Candidatus Nitrospira defluvii*
	30249777	*Nitrosomonas europaea <-> Candidatus Nitrospira defluvii*
	134299143	*Desulfotomaculum reducens <-> Rhodopseudomonas palustris*
	225873673	*Acidobacterium capsulatum <-> Bdellovibrio bacteriovorus*
Arnyl-tRNA synthetase (COG0018)	145592712	*Salinispora tropica<-> Soranum cellulosum*
	159035826	*Salinispora arenicola <-> Soranum cellulosum*
	302870428	*Micromonospora aurantiaca <-> Soranum cellulosum*
Ribosomal protein L11 (COG0080)	386357197	*Streptomyces cattleya <-> Thermosynechococcus elongatus*
	145596448	*Salinispora tropica <-> Thiomicrospira crunogena*
	159039848	*Salinispora arenicola <-> Thiomicrospira crunogena*
	331699209	*Pseudonocardia dioxanivorans<-> Staphylococcus haemolyticus*
	302869987	*Micromonospora aurantiaca <-> Thermosynechococcus elongatus*
	284992891	*Geodermatophilus obscurus<-> Staphylococcus haemolyticus*
	148263130	*Geobacter uraniireducens<-> Clostridium acetobutylicum*
	253701933	*Geobacter M21 <-> Eubacterium rectale*
	322418353	*Geobacter M18 <-> Eubacterium rectale*
	197117312	*Geobacter bemidjiensis<-> Eubacterium rectale*
	86739282	*Frankia CcI3 <-> Anabaena variabilis*
	117927504	*Acidothermus cellulolyticus <-> Synechococcus JA 3 3Ab*
Seryl-tRNA synthetase (COG0172) **	386356659	*Streptomyces cattleya<-> Streptococcus suis*
	111221587	*Frankia alni<-> Streptococcus suis*
	392413758	*Desulfomonile tiedjei <-> Streptococcus suis*
Leucyl-tRNA synthetase (COG0495) **	241205056	*Rhizobium leguminosarum<-> Bacillus cereus*
Ribosomal protein S4 (COG0522)	253998040	*Methylovorus glucosetrophus<-> Clostridium lentocellum*
	253995743	*Methylotenera mobilis <-> Clostridium lentocellum*

Table S8. Most extensive cases of genetic exchange across phyla observed.

Pelotomaculum thermopropionicum and Syntrophobacter fumaroxidans

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_147676911	gi_116751364	YP 001211126.1 ABC-type nitrate/sulfonate/bicarbonate transport system, ATPase component	62
1	gi_147676910	gi_116751363	YP 001211125.1 ABC-type nitrate/sulfonate/bicarbonate transport system, periplasmic components	61.9
1	gi_147676909	gi_116751362	YP 001211124.1 ABC-type nitrate/sulfonate/bicarbonate transport system, permease component	65.6
1	gi_147676908	gi_116751361	YP 001211123.1 hypothetical protein PTH 0573	40.8
1	gi_147676907	gi_116751360	YP 001211122.1 ABC-type nitrate/sulfonate/bicarbonate transport system, periplasmic components	54.4
1	gi_147676906	gi_116751359	YP 001211121.1 hypothetical protein PTH 0571	49.3
1	gi_147676905	gi_116751358	YP 001211120.1 permease	64.2
1	gi_147676904	gi_116751357	YP 001211119.1 hypothetical protein PTH 0569	64.6
2	gi_147678107	gi_116751351	YP 001212322.1 transcriptional regulator	68.5
2	gi_147678106	gi_116751350	YP 001212321.1 acyl CoA:acetate/3-ketoacid CoA transferase	79.2
2	gi_147678105	gi_116751349	YP 001212320.1 aromatic ring hydroxylase	81.5
2	gi_147676350	gi_116751348	YP 001212319.1 acyl-CoA dehydrogenases	81.5
2	gi_147676849	gi_116751347	YP 001211064.1 electron transfer flavoprotein	68.3
2	gi_147676352	gi_116751346	YP 001210567.1 electron transfer flavoprotein, alpha subunit	61
2	gi_147676353	gi_116751344	YP 001210568.1 dehydrogenases	67.1
2	gi_147678100	gi_116751343	YP 001212315.1 ferredoxin-like protein	72.9
2	gi_147676354	gi_116751343	YP 001210569.1 ferredoxin-like protein	71.9
2	gi_147678099	gi_116751342	YP 001212314.1 sugar phosphate permease	72.8
3	gi_147678347	gi_116748291	YP 001212562.1 NADH:ubiquinone oxidoreductase, 24 kD subunit	75.2
3	gi_147678346	gi_116748290	YP 001212561.1 NADH:ubiquinone oxidoreductase, NADH-binding 51 kD subunit	81.9

3	gi_147678345	gi_116748289	YP 001212560.1 hydrogenase subunit	81.8
3	gi_147677315	gi_116748288	YP 001211530.1 hypothetical protein PTH 0980	70.9
3	gi_147677319	gi_116748287	YP 001211534.1 thiamine biosynthesis protein ThiH	73.8

Desulfurivibrio alkaliphilus and Thermodesulfatator indicus

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_297569850	gi_337286693	YP 003691194.1 ATP synthase F1, epsilon subunit	64.1
1	gi_297569851	gi_337286692	YP 003691195.1 ATP synthase F1, beta subunit	81.1
1	gi_297569852	gi_337286691	YP 003691196.1 ATP synthase F1, gamma subunit	53.6
1	gi_297569853	gi_337286690	YP 003691197.1 ATP synthase F1, alpha subunit	71.1
2	gi_297568705	gi_337287265	YP 003690049.1 acetolactate synthase, small subunit	66
2	gi_297568704	gi_337287264	YP 003690048.1 acetolactate synthase, large subunit, biosynthetic type	66.1
3	gi_297570015	gi_337287397	YP 003691359.1 flavodoxin/nitric oxide synthase	64.3
3	gi_297570016	gi_337287396	YP 003691360.1 desulfoferrodoxin	75
4	gi_297568804	gi_337287522	YP 003690148.1 CO dehydrogenase/acetyl-CoA synthase complex, beta subunit	67.7
4	gi_297568803	gi_337287521	YP 003690147.1 CO dehydrogenase/acetyl-CoA synthase delta subunit, TIM barrel	65.2
5	gi_297569271	gi_337286233	YP 003690615.1 ATP-dependent protease La	61
5	gi_297569272	gi_337286232	YP 003690616.1 ATP-dependent Clp protease, ATP-binding subunit ClpX	66.7
5	gi_297569273	gi_337286231	YP 003690617.1 ATP-dependent Clp protease, proteolytic subunit ClpP	69.4
6	gi_297569689	gi_337285563	YP 003691033.1 flagellar biosynthesis protein FlhA	61.4
6	gi_297569688	gi_337285562	YP 003691032.1 flagellar biosynthetic protein FlhB	45.9
6	gi_297569686	gi_337285560	YP 003691030.1 flagellar biosynthetic protein FliQ	50.6

6	gi_297569685	gi_337285559	YP 003691029.1 flagellar biosynthetic protein FliP	60.5
7	gi_297568282	gi_337285778	YP 003689626.1 sulfite reductase, dissimilatory-type alpha subunit	65.3
7	gi_297568283	gi_337285777	YP 003689627.1 sulfite reductase, dissimilatory-type beta subunit	67.7
8	gi_297569325	gi_337286362	YP 003690669.1 ATP phosphoribosyltransferase	70.1
8	gi_297569326	gi_337286361	YP 003690670.1 Phosphoribosyl-AMP cyclohydrolase	67.5
8	gi_297568921	gi_337286359	YP 003690265.1 3-deoxy-D-manno-octulosonate cytidylyltransferase	55.2
9	gi_297570151	gi_337286467	YP 003691495.1 ornithine carbamoyltransferase	62.8
9	gi_297569521	gi_337286466	YP 003690865.1 thiamine biosynthesis protein ThiC	64.6

Streptococcus gordonii Challis substr CH1 and Leptotrichia buccalis C 1013 b

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_157149908	gi_257125329	YP 001450422.1 acetoin dehydrogenase	72.1
1	gi_157151664	gi_257125330	YP 001450421.1 acetoin dehydrogenase	78.2
1	gi_157151137	gi_257125331	YP 001450420.1 dihydrolipoamide acetyltransferase	62.5
1	gi_157150243	gi_257125332	YP 001450419.1 dihydrolipoamide dehydrogenase	65.3
1	gi_157149679	gi_257125333	YP 001450418.1 lipoate protein ligase A	65
2	gi_157150143	gi_257125371	YP 001450805.1 galactose-6-phosphate isomerase subunit LacA	66
2	gi_157149701	gi_257125372	YP 001450797.1 galactose-6-phosphate isomerase subunit LacB	78.9
2	gi_157151561	gi_257125373	YP 001450796.1 tagatose-6-phosphate kinase	62.8
2	gi_157151000	gi_257125374	YP 001450795.1 tagatose 1,6-diphosphate aldolase	71.4
2	gi_157150563	gi_257125375	YP 001450793.1 PTS system lactose-specific transporter subunit IIA	65.7
2	gi_157151244	gi_257125376	YP 001450792.1 PTS system lactose-specific transporter subunit IIBC	80.5
2	gi_157150880	gi_257125377	YP 001450791.1 6-phospho-beta-galactosidase	82

3	gi_157151415	gi_257125430	YP 001450823.1	F0F1 ATP synthase subunit alpha	60
3	gi_157151073	gi_257125432	YP 001450821.1	F0F1 ATP synthase subunit beta	70.4
4	gi_157150337	gi_257125543	YP 001449457.1	V-type ATP synthase subunit A	66.6
4	gi_157149878	gi_257125544	YP 001449458.1	V-type ATP synthase subunit B	73.2
5	gi_157150912	gi_257125927	YP 001449690.1	malate dehydrogenase	68.4
5	gi_157150902	gi_257125929	YP 001449344.1	tRNA-specific 2-thiouridylase MnmA	64.6
7	gi_157150310	gi_257126555	YP 001450452.1	putative lipoprotein	68.9
7	gi_157150275	gi_257126556	YP 001450451.1	tat translocated dye-type peroxidase family protein	64.2
7	gi_157149693	gi_257126557	YP 001450450.1	FTR1 family iron permease	52
7	gi_157150071	gi_257126558	YP 001450449.1	Sec-independent protein translocase TatC	59.4
7	gi_157151040	gi_257126559	YP 001450448.1	twin arginine-targeting protein translocase	62.5
8	gi_157149993	gi_257126077	YP 001450429.1	ATP-dependent protease ATP-binding subunit ClpX	60.6
8	gi_157151545	gi_257126078	YP 001450909.1	ATP-dependent Clp protease proteolytic subunit	59.6
8	gi_157149990	gi_257126963	YP 001449596.1	dihydroorotate dehydrogenase 1A	78.1
8	gi_157149754	gi_257126964	YP 001450542.1	NAD-dependent deacetylase	62.7
9	gi_157151254	gi_257125263	YP 001451012.1	integral membrane protein	78.2
9	gi_157151094	gi_257125264	YP 001449935.1	glycerol kinase	59
10	gi_157150100	gi_257125243	YP 001450958.1	PTS system mannose/fructose/sorbose family transporter subunit IID	68
10	gi_157150304	gi_257125244	YP 001450957.1	phosphotransferase system enzyme II	63.1
10	gi_157151038	gi_257125245	YP 001450956.1	phosphotransferase system enzyme II	61.8

Desulfurispirillum indicum S5 and Marinobacter aquaeolei VT8

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_317050217	gi_120553820	YP 004111333.1 transposase IS204/IS1001/IS1096/IS1165 family protein	99.3
1	gi_317050216	gi_120553821	YP 004111332.1 lipoprotein signal peptidase	98.8
1	gi_317050206	gi_120553822	YP 004111322.1 cation efflux protein	97
1	gi_317050205	gi_120553826	YP 004111321.1 Cd(II)/Pb(II)-responsive transcriptional regulator	90.4
1	gi_317050214	gi_120553826	YP 004111330.1 Cd(II)/Pb(II)-responsive transcriptional regulator	97.8
1	gi_317050213	gi_120553909	YP 004111329.1 integron integrase	52.5
1	gi_317050211	gi_120553989	YP 004111327.1 small multidrug resistance protein	68
2	gi_317050253	gi_120553460	YP 004111369.1 nitrogen regulatory protein P-II	64.3
2	gi_317050254	gi_120554275	YP 004111370.1 general secretion pathway protein G	65.2
3	gi_317051135	gi_120555535	YP 004112251.1 sulfate adenylyltransferase small subunit	77.7
3	gi_317051136	gi_120555646	YP 004112252.1 sulfate adenylyltransferase large subunit	63.9
4	gi_317051301	gi_120553293	YP 004112417.1 TRAP dicarboxylate transporter subunit DctM	80
4	gi_317051300	gi_120553294	YP 004112416.1 tripartite ATP-independent periplasmic transporter subunit DctQ	61.3
4	gi_317051299	gi_120553295	YP 004112415.1 family 7 extracellular solute-binding protein	69.3
4	gi_317051296	gi_120553973	YP 004112412.1 ABC transporter-like protein	60.5
4	gi_317051303	gi_120554460	YP 004112419.1 binding-protein-dependent transporter inner membrane component	68
4	gi_317051304	gi_120554461	YP 004112420.1 ABC transporter-like protein	55.6
5	gi_317051351	gi_120554670	YP 004112467.1 Agmatine deiminase	52.1
5	gi_317051350	gi_120554671	YP 004112466.1 nitrilase/cyanide hydratase and apolipoprotein N-acyltransferase	62
5	gi_317051352	gi_120554979	YP 004112468.1 TRAP transporter, 4TM/12TM fusion protein	62.8
5	gi_317051353	gi_120554980	YP 004112469.1 TAXI family TRAP transporter solute receptor	63.1

7	gi_317052328	gi_120556164	YP 004113444.1 phosphonate ABC transporter periplasmic phosphonate-binding protein	64.2
7	gi_317052327	gi_120556165	YP 004113443.1 phosphonate ABC transporter ATPase subunit	69.8
7	gi_317052326	gi_120556166	YP 004113442.1 phosphonate ABC transporter inner membrane subunit	66.9
7	gi_317052325	gi_120556167	YP 004113441.1 phosphonate ABC transporter inner membrane subunit	65.8

Caldicellulosiruptor hydrothermalis 108 and Thermotoga thermarum DSM 5069

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_312128371	gi_338730006	YP 003991766.1 3-isopropylmalate dehydrogenase	72.8
1	gi_312128370	gi_338730005	YP 003991765.1 3-isopropylmalate dehydratase, small subunit	74.4
1	gi_312128369	gi_338730004	YP 003991764.1 3-isopropylmalate dehydratase, large subunit	83.6
1	gi_312128368	gi_338730295	YP 003991973.1 pyridoxine biosynthesis protein	64.6
2	gi_312128334	gi_338730008	YP 003991968.1 oligopeptide/dipeptide ABC transporter ATPase subunit	68.8
2	gi_312128333	gi_338730007	YP 003992157.1 tryptophan synthase subunit alpha	60.6
3	gi_312128165	gi_338729930	YP 003992156.1 tryptophan synthase subunit beta	74.5
3	gi_312128164	gi_338730159	YP 003992155.1 phosphoribosylanthranilate isomerase	62
4	gi_312127781	gi_338731040	YP 003992154.1 indole-3-glycerol-phosphate synthase	70.9
4	gi_312127780	gi_338730055	YP 003992153.1 anthranilate phosphoribosyltransferase	81.4
5	gi_312127526	gi_338730088	YP 003992152.1 glutamine amidotransferase of anthranilate synthase	74.7
5	gi_312127524	gi_338730086	YP 003992151.1 chorismate binding-like protein	67.9
6	gi_312127472	gi_338730808	YP 003992346.1 histidinol dehydrogenase	62.8
6	gi_312127471	gi_338730807	YP 003992345.1 ATP phosphoribosyltransferase	65.4

7	gi_312127099	gi_338731576	YP 003993207.1 isocitrate dehydrogenase (nad(+))	69.8
7	gi_312127094	gi_338731090	YP 003993245.1 acetolactate synthase, large subunit, biosynthetic type	63.3
8	gi_312126892	gi_338730292	YP 003993244.1 acetolactate synthase, small subunit	60.5
8	gi_312126891	gi_338730293	YP 003993243.1 ketol-acid reductoisomerase	66.3
8	gi_312126890	gi_338730294	YP 003993242.1 2-isopropylmalate synthase	64.6

Candidatus Nitrospira defluvii and Janthinobacterium Marseille

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_302035457	gi_152981820	YP 003795779.1 hypothetical protein NIDE0063	61.5
1	gi_302035458	gi_152981934	YP 003795780.1 mercuric resistance operon regulatory protein	67.7
1	gi_302035459	gi_152982220	YP 003795781.1 mercury ion transport protein	69.2
1	gi_302035460	gi_152982938	YP 003795782.1 periplasmic mercury ion binding protein	71.9
1	gi_302035462	gi_152982873	YP 003795784.1 hypothetical protein NIDE0068	74.2
1	gi_302035463	gi_152982221	YP 003795785.1 hypothetical protein NIDE0069	95.8
1	gi_302035464	gi_152981666	YP 003795786.1 putative site-specific recombinase, resolvase family (phage related)	94.4
1	gi_302035465	gi_152982797	YP 003795787.1 hypothetical protein NIDE0071	91.3
1	gi_302035466	gi_152983289	YP 003795788.1 hypothetical protein NIDE0072	82.4
1	gi_302035471	gi_152983290	YP 003795793.1 hypothetical protein NIDE0079	71.3
1	gi_302035472	gi_152982677	YP 003795794.1 hypothetical protein NIDE0080	80.9
1	gi_302035473	gi_152982207	YP 003795795.1 hypothetical protein NIDE0081	90.4
1	gi_302035474	gi_152982323	YP 003795796.1 hypothetical protein NIDE0082	84.4
1	gi_302035475	gi_152982824	YP 003795797.1 hypothetical protein NIDE0083	85.2
1	gi_302035476	gi_152982461	YP 003795798.1 hypothetical protein NIDE0084	75.5
1	gi_302035477	gi_152982378	YP 003795799.1 hypothetical protein NIDE0085	83.3
1	gi_302035478	gi_152982760	YP 003795800.1 hypothetical protein NIDE0086	67.4
1	gi_302035479	gi_152981706	YP 003795801.1 putative DNA primase'	87.9
1	gi_302035480	gi_152982142	YP 003795802.1 putative polynucleotidyl transferase	90.2
1	gi_302035481	gi_152983291	YP 003795803.1 hypothetical protein NIDE0090	79.6

1	gi_302035483	gi_152982005	YP 003795805.1 site-specific DNA-methyltransferase N-4/N-6 (phage related)	85.1
1	gi_302035484	gi_152981982	YP 003795806.1 site-specific DNA-methyltransferase N-4/N-6 (phage related)	92
1	gi_302035485	gi_152983294	YP 003795807.1 hypothetical protein NIDE0094	82.4
1	gi_302035486	gi_152982304	YP 003795808.1 hypothetical protein NIDE0095	85.7
1	gi_302035487	gi_152982162	YP 003795809.1 hypothetical protein NIDE0097	68.2
1	gi_302035489	gi_152982161	YP 003795811.1 hypothetical protein NIDE0099	96.6
1	gi_302035490	gi_152981093	YP 003795812.1 phage terminase large subunit	95.3
1	gi_302035491	gi_152982062	YP 003795813.1 hypothetical protein NIDE0101	95.7
1	gi_302035492	gi_152982876	YP 003795814.1 hypothetical protein NIDE0102	93.1
1	gi_302035493	gi_152982972	YP 003795815.1 hypothetical protein NIDE0103	93.2
1	gi_302035494	gi_152981081	YP 003795816.1 phage portal protein, lambda family	87.3
1	gi_302035495	gi_152981533	YP 003795817.1 putative phage minor capsid protein C	73.8
1	gi_302035496	gi_152982282	YP 003795818.1 hypothetical protein NIDE0106	80
1	gi_302035497	gi_152982830	YP 003795819.1 hypothetical protein NIDE0107	91
1	gi_302035498	gi_152982165	YP 003795820.1 hypothetical protein NIDE0108	80
1	gi_302035499	gi_152982164	YP 003795821.1 hypothetical protein NIDE0109	92.9
1	gi_302035500	gi_152982980	YP 003795822.1 hypothetical protein NIDE0110	71.8
1	gi_302035501	gi_152982163	YP 003795823.1 hypothetical protein NIDE0111	98.4
1	gi_302035502	gi_152982160	YP 003795824.1 hypothetical protein NIDE0112	95.5
1	gi_302035503	gi_152982159	YP 003795825.1 hypothetical protein NIDE0113	98.1
1	gi_302035504	gi_152982158	YP 003795826.1 putative phage tail length tape measure protein	91.3
1	gi_302035505	gi_152982157	YP 003795827.1 hypothetical protein NIDE0115	96.9
1	gi_302035506	gi_152982156	YP 003795828.1 hypothetical protein NIDE0116	87.8
1	gi_302035507	gi_152982127	YP 003795829.1 hypothetical protein NIDE0117	87.8
1	gi_302035509	gi_152982262	YP 003795831.1 hypothetical protein NIDE0119	89
1	gi_302035510	gi_152982615	YP 003795832.1 hypothetical protein NIDE0120	97.5
1	gi_302035511	gi_152982541	YP 003795833.1 hypothetical protein NIDE0121	87.1
1	gi_302035512	gi_152982982	YP 003795834.1 hypothetical protein NIDE0122	91

2	gi_302036778	gi_152979893	YP 003797100.1 chorismate synthase	74.2
2	gi_302036779	gi_152980654	YP 003797101.1 ribonuclease H	68.8
3	gi_302038815	gi_152981067	YP 003799137.1 multidrug efflux system subunit C	60.7
3	gi_302038816	gi_152981117	YP 003799138.1 multidrug efflux system subunit B	63.2

Clostridium saccharolyticum WM1 and Sphaerochaeta pleomorpha Grapes

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_302385696	gi_374314595	YP 003821518.1 binding-protein-dependent transport system inner membrane protein	72.3
1	gi_302385695	gi_374314596	YP 003821517.1 binding-protein-dependent transport system inner membrane protein	69.1
1	gi_302385694	gi_374314597	YP 003821516.1 extracellular solute-binding protein	64.4
2	gi_302386292	gi_374314822	YP 003822114.1 ABC transporter	72.1
2	gi_302386293	gi_374314823	YP 003822115.1 inner-membrane translocator	75.9
2	gi_302386294	gi_374314824	YP 003822116.1 LacI family transcriptional regulator	72.2
3	gi_302387219	gi_374314977	YP 003823041.1 short-chain dehydrogenase/reductase SDR	76
3	gi_302387813	gi_374314978	YP 003823635.1 4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase	59.6
4	gi_302385761	gi_374315043	YP 003821583.1 L-fucose isomerase-like protein	63.5
4	gi_302385109	gi_374315044	YP 003820931.1 class II aldolase/adducin family protein	62.8
5	gi_302387893	gi_374315132	YP 003823715.1 protein-tyrosine phosphatase	76.4
5	gi_302387095	gi_374315133	YP 003822917.1 redox-active disulfide protein 2	50.4
5	gi_302387097	gi_374315134	YP 003822919.1 permease	69.9
6	gi_302388266	gi_374315140	YP 003824088.1 ABC transporter	56.6
6	gi_302386838	gi_374315141	YP 003822660.1 inner-membrane translocator	70.4

6	gi_302386840	gi_374315143	YP 003822662.1 ABC transporter	63.4
6	gi_302386841	gi_374315144	YP 003822663.1 basic membrane lipoprotein	64.6
7	gi_302384518	gi_374315235	YP 003820340.1 flavodoxin/nitric oxide synthase	75.8
7	gi_302387044	gi_374315237	YP 003822866.1 arsenical-resistance protein	69.4
7	gi_302387889	gi_374315238	YP 003823711.1 ArsR family transcriptional regulator	60
8	gi_302387949	gi_374315291	YP 003823771.1 tryptophan synthase subunit beta	77.2
8	gi_302387950	gi_374315292	YP 003823772.1 tryptophan synthase subunit alpha	60.8
9	gi_302385599	gi_374315380	YP 003821421.1 binding-protein-dependent transport system inner membrane protein	66.9
9	gi_302385598	gi_374315381	YP 003821420.1 extracellular solute-binding protein	64.2
10	gi_302387979	gi_374315440	YP 003823801.1 dihydroxy-acid dehydratase	65.7
10	gi_302387980	gi_374315441	YP 003823802.1 3-isopropylmalate dehydrogenase	62.8
10	gi_302386582	gi_374315442	YP 003822404.1 3-isopropylmalate dehydratase small subunit	70.2
10	gi_302386583	gi_374315443	YP 003822405.1 3-isopropylmalate dehydratase large subunit	71.1
10	gi_302386585	gi_374315446	YP 003822407.1 ketol-acid reductoisomerase	67
11	gi_302386734	gi_374315727	YP 003822556.1 polar amino acid ABC transporter inner membrane subunit	71.9
11	gi_302386735	gi_374315728	YP 003822557.1 family 3 extracellular solute-binding protein	61.5
12	gi_302387418	gi_374315759	YP 003823240.1 malate/L-lactate dehydrogenase	62.8
12	gi_302387311	gi_374315763	YP 003823133.1 ABC transporter	66.1
12	gi_302387310	gi_374315764	YP 003823132.1 ABC transporter	59.1
12	gi_302387309	gi_374315765	YP 003823131.1 inner-membrane translocator	59.7
12	gi_302387308	gi_374315766	YP 003823130.1 inner-membrane translocator	78.5
12	gi_302387307	gi_374315767	YP 003823129.1 extracellular ligand-binding receptor	73.8
12	gi_302385731	gi_374315788	YP 003821553.1 sodium ion-translocating decarboxylase subunit	60.9

			beta	
12	gi_302384784	gi_374315790	YP 003820606.1 dCMP deaminase	62.3
13	gi_302384774	gi_374315940	YP 003820596.1 xylose isomerase domain-containing protein TIM barrel	65.6
13	gi_302384775	gi_374315941	YP 003820597.1 binding-protein-dependent transport system inner membrane protein	72.4
13	gi_302384776	gi_374315942	YP 003820598.1 binding-protein-dependent transport system inner membrane protein	67.6
13	gi_302384777	gi_374315943	YP 003820599.1 extracellular solute-binding protein	68.6
14	gi_302384523	gi_374316702	YP 003820345.1 ABC transporter	67.1
14	gi_302384524	gi_374316703	YP 003820346.1 inner-membrane translocator	67.2
14	gi_302384525	gi_374316704	YP 003820347.1 LacI family transcriptional regulator	72.6
15	gi_302385244	gi_374317120	YP 003821066.1 extracellular solute-binding protein	62.7
15	gi_302385245	gi_374317121	YP 003821067.1 tripartite AtP-independent periplasmic transporter subunit DctQ	68.2
15	gi_302385246	gi_374317122	YP 003821068.1 TRAP dicarboxylate transporter subunit DctM	81.9
16	gi_302386148	gi_374317162	YP 003821970.1 phage major capsid protein, HK97 family	62.8
16	gi_302386147	gi_374317163	YP 003821969.1 peptidase S14 ClpP	53.7
16	gi_302386146	gi_374317164	YP 003821968.1 phage portal protein, HK97 family	66.2

Deferribacter desulfuricans SSM1 and Geobacter uraniireducens Rf4

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_291280213	gi_148265082	YP 003497048.1 acetyl-CoA C-acetyltransferase	66.8
1	gi_291280212	gi_148265081	YP 003497047.1 3-hydroxybutyryl-CoA dehydrogenase	65.3
1	gi_291280211	gi_148265080	YP 003497046.1 3-hydroxybutyryl-CoA dehydratase	62.8
1	gi_291280210	gi_148265079	YP 003497045.1 butyryl-CoA dehydrogenase	74.3
1	gi_291280209	gi_148263663	YP 003497044.1 iron-sulfur cluster-binding protein	65.8

1	gi_291280208	gi_148265077	YP 003497043.1	electron transfer flavoprotein subunit beta	69.3
1	gi_291280207	gi_148265076	YP 003497042.1	electron transfer flavoprotein subunit alpha	72.5
1	gi_291280192	gi_148265419	YP 003497027.1	acetate kinase	70.2
2	gi_291279999	gi_148264216	YP 003496834.1	cytochrome bd oxidase subunit II	65.7
2	gi_291279998	gi_148264217	YP 003496833.1	cytochrome bd oxidase subunit I	69.8
3	gi_291279856	gi_148265390	YP 003496691.1	nitrogen regulatory protein P-II	72.3
3	gi_291279855	gi_148264278	YP 003496690.1	glutamine synthetase type I	70.4
4	gi_291279849	gi_148263653	YP 003496684.1	long-chain fatty-acid-CoA ligase	65.4
4	gi_291279848	gi_148263654	YP 003496683.1	3-hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase	66.8
4	gi_291279847	gi_148263655	YP 003496682.1	3-ketoacyl-CoA thiolase	74.6
4	gi_291279846	gi_148263656	YP 003496681.1	acyl-CoA dehydrogenase	76.2
5	gi_291279843	gi_148264890	YP 003496678.1	HNH endonuclease	69.2
5	gi_291279842	gi_148262944	YP 003496677.1	phosphoenolpyruvate carboxykinase (ATP)	62
6	gi_291279569	gi_148264363	YP 003496404.1	2-isopropylmalate synthase	64.5
6	gi_291279568	gi_148264364	YP 003496403.1	aspartate kinase monofunctional class	63.8
7	gi_291279489	gi_148264234	YP 003496324.1	riboflavin synthase beta chain	61.8
7	gi_291279488	gi_148264235	YP 003496323.1	riboflavin biosynthesis bifunctional protein RibBA	67.6
8	gi_291279312	gi_148264247	YP 003496147.1	malate dehydrogenase	75
8	gi_291279311	gi_148264248	YP 003496146.1	isocitrate dehydrogenase NADP-dependent	67.2
8	gi_291279310	gi_148263996	YP 003496145.1	aconitate hydratase	71.6
9	gi_291279213	gi_148263639	YP 003496048.1	citrate synthase	65.3

9	gi_291279211	gi_148262430	YP 003496046.1 porphobilinogen synthase	70.8
10	gi_291278972	gi_148263636	YP 003495807.1 acyl-CoA synthase	61.7
10	gi_291278971	gi_148266340	YP 003495806.1 pyruvate:ferredoxin oxidoreductase	66.5
11	gi_291278510	gi_148262626	YP 003495345.1 Ni-Fe hydrogenase small subunit	66.9
11	gi_291278509	gi_148262625	YP 003495344.1 Ni-Fe hydrogenase large subunit	73.4

Listeria ivanovii PAM 55 and Sebaldella termitidis ATCC 33386

Synthenic Region	Gi 1	Gi 2	Annotation	a.a. identity (%)
1	gi_347547968	gi_269118910	YP 004854296.1 putative NADP-specific glutamate dehydrogenase	65.1
1	gi_347549798	gi_269118929	YP 004856126.1 putative phosphate ABC transporter ATP binding protein	64
2	gi_347548523	gi_269119662	YP 004854851.1 putative PduU protein	60.5
2	gi_347548524	gi_269119663	YP 004854852.1 putative PduV protein	44.1
2	gi_347548529	gi_269119665	YP 004854857.1 putative propanediol utilization protein PduA	76.5
2	gi_347548530	gi_269119652	YP 004854858.1 putative propanediol utilization protein PduB	75.9
2	gi_347548531	gi_269119653	YP 004854859.1 putative propanediol dehydratase subunit alpha	76.7
2	gi_347548532	gi_269119654	YP 004854860.1 putative diol dehydrase subunit gamma	58.5
2	gi_347548533	gi_269119655	YP 004854861.1 putative diol dehydrase subunit gamma PddC	54.7
2	gi_347548534	gi_269119656	YP 004854862.1 putative diol dehydratase-reactivating factor large subunit	67.3
2	gi_347548535	gi_269119657	YP 004854863.1 putative diol dehydratase-reactivating factor small chain	41.7
2	gi_347548537	gi_269119665	YP 004854865.1 putative carboxysome structural protein	82.8
2	gi_347548543	gi_269119659	YP 004854871.1 putative ethanolamine utilization protein EutE	55.4
2	gi_347548556	gi_269119660	YP 004854884.1 putative carboxysome structural protein	56.6
2	gi_347548557	gi_269119666	YP 004854885.1 putative acetaldehyde dehydrogenase / alcohol dehydrogenase	60.7
2	gi_347548558	gi_269119661	YP 004854886.1 putative carboxysome structural protein	85.7

2	gi_347548560	gi_269119668	YP 004854888.1	putative PduL protein	51.5
2	gi_347548562	gi_269119670	YP 004854890.1	putative carbon dioxide concentrating mechanism protein	62.8
3	gi_347547746	gi_269121938	YP 004854074.1	putative phospho-beta-glucosidase	67.2
3	gi_347547927	gi_269121939	YP 004854255.1	putative 6-phospho-beta-glucosidase	61.1
3	gi_347547940	gi_269121939	YP 004854268.1	putative 6-phospho-beta-glucosidase	67.3
3	gi_347550094	gi_269121938	YP 004856422.1	putative beta-glucosidase	68.4
4	gi_347547782	gi_269121842	YP 004854110.1	putative oxidoreductase	71.3
4	gi_347549403	gi_269121832	YP 004855731.1	putative oxidoreductase	70.9
5	gi_347547708	gi_269121624	YP 004854036.1	DeoR family transcriptional regulator	69
5	gi_347547709	gi_269121623	YP 004854037.1	putative N-acetylmannosamine-6-phosphate epimerase	80.7
5	gi_347547710	gi_269121621	YP 004854038.1	putative mannose-specific PTS system enzyme IIB	64.7
5	gi_347547711	gi_269121620	YP 004854039.1	putative mannose-specific PTS system enzyme IIC	84.3
5	gi_347547712	gi_269121619	YP 004854040.1	putative mannose-specific PTS system enzyme IID	78.3
5	gi_347547713	gi_269121618	YP 004854041.1	putative mannose-specific PTS system enzyme IIA	61.1
6	gi_347549949	gi_269121095	YP 004856277.1	putative phosphotriesterase	70.2
6	gi_347549950	gi_269121096	YP 004856278.1	putative PTS enzyme IIC component	67.9
7	gi_347548252	gi_269120483	YP 004854580.1	putative amino acid ABC transporter ATP-binding protein	66.1
7	gi_347549641	gi_269120483	YP 004855969.1	putative amino acid ABC transporter ATP binding protein	61.2
8	gi_347550146	gi_269120141	YP 004856474.1	hypothetical protein	61.7

8	gi_347550147	gi_269120140	YP 004856475.1	putative alcohol dehydrogenase	74.9
8	gi_347550148	gi_269120139	YP 004856476.1	putative sugar ABC transporter permease	69.6
8	gi_347550149	gi_269120138	YP 004856477.1	putative sugar ABC transporter permease	65.1
9	gi_347548281	gi_269119824	YP 004854609.1	putative PTS system, beta-glucoside enzyme IIB component	67.9
9	gi_347548282	gi_269119823	YP 004854610.1	putative PTS system, Lichenan-specific enzyme IIC component	71.8
9	gi_347548284	gi_269119821	YP 004854612.1	putative oxidoreductase	62.3
10	gi_347548555	gi_269119678	YP 004854883.1	putative carboxysome structural protein EutL	70
10	gi_347548564	gi_269119679	YP 004854892.1	putative ethanolamine utilization protein EutH	73.8
11	gi_347548553	gi_269119676	YP 004854881.1	eutB gene product	71.8
11	gi_347548552	gi_269119675	YP 004854880.1	eutA gene product	51.1

Table S9. The genomes used in the study and the fraction of their genes detected to be horizontally transferred from/to other phyla. The data shown are based on the gene-based analysis (see main text for details), and the genomes are shown in order of increasing rate of HGT. The Table is provided as a separate Microsoft Excel file.

Table S10. Description of the COG general functional categories.

Adapted from the COG website: http://www.ncbi.nlm.nih.gov/COG/

Category	Description	General category
A	RNA processing and modification	Information processes and signaling
B	Chromatin Structure and dynamics	Information processes and signaling
C	Energy production and conversion	Metabolism
D	Cell cycle control and mitosis	Cellular processes and signaling
E	Amino Acid metabolism and transport	Metabolism
F	Nucleotide metabolism and transport	Metabolism
G	Carbohydrate metabolism and transport	Metabolism
H	Coenzyme metabolism	Metabolism
I	Lipid metabolism	Metabolism
J	Translation	Information processes and signaling
K	Transcription	Information processes and signaling
L	Replication and repair	Information processes and signaling
M	Cell wall/membrane/envelop biogenesis	Cellular processes and signaling
N	Cell motility	Cellular processes and signaling
O	Post-translational modification, protein turnover	Cellular processes and signaling
P	Inorganic ion transport and metabolism	Metabolism
Q	Secondary Structure	Metabolism
T	Signal Transduction	Cellular processes and signaling
U	Intracellular trafficking and secretion	Cellular processes and signaling
Y	Nuclear structure	Cellular processes and signaling
Z	Cytoskeleton	Cellular processes and signaling
R	General Functional Prediction only	Poorly Characterized
S	Function Unknown	Poorly Characterized

REFERENCES

Benjamini Y, Hochberg Y (1995). Controlling the False Discovery Rate - a practical and powerful approach to multiple testing. *J Roy Stat Soc B Met* **57:** 289-300.

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**.

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

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461.

Konstantinidis KT, Tiedje JM (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102:** 2567-2572.

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

Shaffer JP (1995). Multiple Hypothesis-Testing. *Annu Rev Psychol* **46:** 561-584.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011). Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **27:** 431-432.

Su G, Kuchinsky A, Morris JH, States DJ, Meng F (2010). GLay: community structure analysis of biological networks. *Bioinformatics* **26:** 3135-3137.

Wolf YI, Koonin EV (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol* **4:** 1286-1294.