FUNCTIONAL ANALYSIS OF MICROBIAL GENOMES AND METAGENOMES

by

CHENGSHENG ZHU

A dissertation submitted to the

The Graduate School-New Brunswick

Rutgers, The State University of New Jersey

and

The Graduate School of Biomedical Science

For the degree of

Doctor of Philosophy

Graduate Program in Microbiology and Molecular Genetics

Written under the direction of

Dr. Yana Bromberg

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

MAY 2017

ABSTRACT OF THE DISSERTATION

FUNCTIONAL ANALYSIS OF MICROBIAL GENOMES AND METAGENOMES

by CHENGSHENG ZHU

Dissertation Director:

Dr. Yana Bromberg

Microorganisms are capable of carrying out molecular functionality relevant to a range of human interests, including health, industrial production, and bioremediation. Current microbial taxonomy is phylogeny-guided, *i.e.*, the organisms are grouped based on their evolutionary relationships. Due to horizontal gene transfer, evolutionary relatedness cannot guarantee genome-encoded molecular functional similarity. In this work, we establish a computational framework for comparison of microorganisms based on their molecular functionality. In the *fusion* (functional-repertoire similarity-based organism network) representation, organisms can be consistently assigned to groups based on a quantitative measure of their functional similarities. The results highlight the specific environmental factor(s) that explain the functional differences between groups of microorganism. We deposit the functional data in *fusion*DB, mapping bacteria and their functions to available metadata: habitat/niche, preferred temperature, and oxygen use. The web interface further

allows mapping new microbial genomes to the functional spectrum of reference bacteria. In the end, we describe *mi-faser* (microbiome functional annotation of sequencing reads), the meta-genomic/-transcriptomic analysis pipeline combining an algorithm that is optimised to map reads to molecular functions encoded by the read-correspondent genes, and a manually curated reference database of protein functions. With *mi-faser*, we identify previously unseen oil degradation-specific functions in BP oil-spill data, and reveal the role of gut microbiome in Crohn's disease pathogenicity, showing that the patient microbiomes are enriched in both the functions that promote inflammation and those that help bacteria survive it.

ACKNOWLEDGEMENT

I wish to thank my advisor, Dr. Yana Bromberg, to take me as her first student, patiently help me to develop my skills, correct my silly mistakes, encourage me when I was struggling, and generously offered me many invaluable advices on research, and sometimes on life itself. I would also like to thank my committee members, Dr. Max Haggblom, Dr. Jeff Boyd and Dr. Vikas Nanda, for all the useful discussions, kind suggestions, as well as reviewing the draft of this document. Special thanks to Dr. Tamar Barkay, who recommended me to Yana, and has always been really nice and helpful. I also want to thank my previous and current lab colleagues, who shared with me research and life. I am very grateful to Dr. Arik Harel, who helped me to learn the basics of bioinformatics, and welcomed me warmly to his house as a family member when I was down in my life. I would also like to thank Yannick Mahlich and Max Miller, who helped me to create these amazing online services of *fusion*DB and *mi-faser*. Many thanks to my collaborators who shared their data, and all the researchers who deposited data in public database. In the end, I really want to thank my parents, who supported me all these years.

TABLE OF CONTENTS

LIST OF TABLES

TABLES PAGE

LIST OF FIGURES

FIGURE PAGE

Chapter 1

My favourite definition of humour, as Bob Mankoff, the carton editor of *the New Yorkers* for the past two decades, masterfully put it, is the right amount of wrong. Being completely right is boring, while being completely wrong is meaningless. To me, keeping the balance between the fraction of right and wrong is the key to make a good joke, a joyful life, and, as I am about to start to describe, a PhD research project.

Anton Van Leeuwenhoek's first 1670's discovery of a microscopic organism, which he called "animalcules", has ushered in centuries of explosive growth of field of microbiology. Each day we find more and more impressive evidence of the importance of microbes in the natural environment and in human life. Microorganisms in the ocean and lakes are the basis of the global food chain, while gastrointestinal microbes help with digestion and offer otherwise inaccessible nutrition to their human and animal hosts. Microbes fix nitrogen, a necessary building block of organic compounds, break down dead matter, clean up waste, and otherwise contribute to biogeochemical flux cycles. Microorganisms are tremendously valuable in industrial and clinical applications. Thus, enormous efforts have been dedicated to answer two central questions:

who are these bacteria and *what do they do*?

The answers to these two questions have long been thought to be tightly correlated, where the latter defines the former. Unfortunately, we see that this

often not the case. In fact, bacterial taxonomy, which aims to answer the “who” question, is dynamic and controversial due to:

Difficulty cultivating organism mono-cultures in the lab. Classifying a new bacterial species requires obtaining its pure culture for direct study of its functionality. However, obtaining a pure culture is hard since most microbes naturally live within mutualistic communities. It is estimated that only 1% of the microbes are cultivable (Zengler, Toledo et al. 2002), as mimicking the natural environment of most microbes in the lab is difficult.

Horizontal gene transfer (HGT). Microbes are able to deliver genetic material across species (mostly via plasmid or phage). HGT makes the conventional biological species concept, where only the individuals of the same species are able to mate to make reproductively able progeny, consistently inapplicable to bacteria.

Difficulty measuring phenotypic traits. Even with the aid of advanced experimental techniques, bacterial phenotypic traits are much less clearly defined than those of animals or plants; *e.g.*, the colony/cell morphology and biochemical environments of bacteria *vs.* the presence of placenta or size/shape of leaves. Thus, a subjective set of poorly-defined features is often used to establish a new bacteria species (Vandamme, Pot et al. 1996). Additionally, as many as three hundred biochemical/physiological tests would only access 5-20% of the bacterial functional potential (Garrity GM 2001).

Since it is not easy to answer the question about what the bacteria do, the “who are these bacteria” question is currently answered by relying on incidental findings and often subjective judgment calls.

Current bacterial taxonomy relies on phylogeny. The development of the DNA-DNA hybridization technique in the 1960’s offered researchers an objective criterion for defining species (Brenner, Fanning et al. 1969). DNA-DNA hybridization entails extracting, denaturing, and mixing whole genome DNA from different organisms. The mixture is then incubated to form hybrid double-stranded DNA. The hybrid helix is more stable if the two organisms are closely related. The relative binding ratio (RBR), a measure for DNA-DNA similarity, of 70% is often used as the gold standard for identifying two organisms as coming from the same species (Wayne, et al., 1987). Note however, that there are cases where organisms of different species, or even different genera, exhibit ≥70% DNA-DNA similarity (Gevers, et al., 2005).

In the 1980’s, the discovery of 16S rRNA set a milestone in prokaryotic taxonomy. 16S rRNA is a component of the 30S subunit of the prokaryotic ribosome, universal in prokaryotes. Because of its correspondence with DNA-DNA hybridization (Figure 1), a new method for identifying species membership was adapted – 16S rRNA sequence similarity (Stackebrandt and Goebel, 1994). However, interpreting this criterion is not straightforward. As shown in Figure 1, organism pairs with <97% 16S rRNA similarity always share <70% RBR. However, ≥97% 16S rRNA similarity doesn’t guarantee ≥70% RBR. In other

words, <97% 16S rRNA similarity indicates different species, while ≥97% similarity could equally likely mean same or different species.

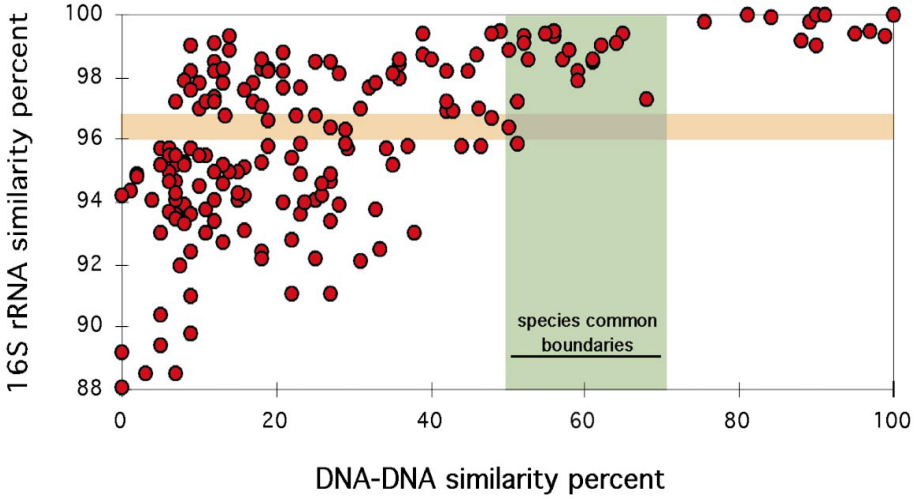


Figure 1-1 The correlation between 16S rRNA and DNA-DNA similarity (Rossello-Mora and Amann 2001). Each dot in the figure represents a pair of organisms. Organisms with 16S rRNA similarity above 97% share DNA-DNA similarity from ~4% to 100%. Organisms with 16S rRNA similarity below 97% have DNA-DNA similarity of less than 70%. The latter are classified as being of different species.

The 16S rRNA sequence has both conserved and variable regions (Neefs, Vandepeer et al. 1990). Conserved regions are used to design universal primers to amplify the variable regions (V1 to V9), which offer phylogenetic information; *i.e.* the accumulation of differences between the corresponding variable regions of 16S rRNA of different organisms can be thought of as representative of time since the split from the last common ancestor. Wide use of 16s rRNA as a phylogenetic marker has led to the development of many organism-sequence databases; *e.g.* GreenGenes, RDP, and Silva (DeSantis, Hugenholtz et al. 2006, Pruesse, Quast et al. 2007, Cole, Wang et al. 2009). Moreover, the current gold standard for bacterial taxonomy, Bergey's Manual, is a framework of prokaryotic taxonomy built around a backbone of 16S rRNA-derived phylogeny (Garrity GM 2001).

As mentioned above, however, 16S rRNA similarity is not sufficient to identify organism species. For example, Fox *et al*. found the 16S rRNA sequences of strains of *B. globisporus* and *B. psychrophilus* to be almost identical (>99.5%) (Fox, Wisotzkey et al. 1992), while the DNA-DNA hybridization experiment clearly showed that the organisms belonged to different species. Similar results were obtained in the study of members of the genera *Aeromonas* and *Plesiomonas* (Martinezmurcia, Benlloch et al. 1992). In addition to cross-species similarity 16S rRNA classification also suffers from intraspecies heterogeneity. For example, Case *et al.* (Case, Boucher et al. 2007) report 111 sequenced bacterial genomes and 460 corresponding 16S rRNA sequences – 4.6 16S rRNA copies per genome, on average. Though in most cases within species differences between 16S rRNA sequences are minor, some can reach 2%, and for *Thermoanaerobacteriales* the difference is over 10%. This ambiguity in inferring taxonomy with 16S rRNA sequences could lead to overestimation of microbial diversity in environmental samples. Finally, due to HGT, the current phylogeny-based taxonomic assignments cannot guarantee functional similarity; *i.e.* two microbes of the same taxonomic group inhabiting different environments may be functionally very different - just as your cousin may be more different from you than your unrelated best friend. Therefore, the current taxonomy's answer to the *who* question can incorrectly answer the *what* question, thus confusing researchers and hampering *en bulk* computational analysis.

Annotation of microbial molecular function potentials encoded in metagenomes. As the best available record of organism heritage and

functionality, the organism genome contains all the information necessary to define a species member (Boussau and Daubin, 2010). With the advent of high-throughput sequencing, the number of publicly available fully-sequenced bacterial genomes has drastically increased. Even the uncultivable microbes can now be accessed by metagenome sequencing, an approach that extracts DNA directly from the environment and sequences/reconstructs the fragments without regard for genome of origin. Metagenome sequencing has been applied to numerous environments, *e.g.* ocean, soil, and human gut (Venter, Remington et al. 2004, Simon, Wiezer et al. 2009, Qin, Li et al. 2010), and has vastly augmented the bacterial diversity in the tree of life (Hug, Baker et al. 2016). Metagenome can be annotated with or without read assembly. If the reads can be assembled into large contigs, existing pipelines, *e.g.*, RAST (Aziz, Bartels et al. 2008) and IMG (Markowitz, Chen et al. 2014), can be applied. However, assembly is often plagued by a large fraction of unassembled reads or short length contigs, which belong to the minor microbiome members, and by chimeric assemblies, which are especially common for complex and highly diverse samples. Downstream gene finding algorithms are further faced with incomplete and erroneously assembled sequences, complicating statistical model constructions. Read-based annotation, *e.g.*, using a platform such as MG-RAST (Aziz, Bartels et al. 2008), can access molecular functionality of the entire community. However, reads are usually annotated via function transfer by homology that, due to the short read length, is lacking in precision. This inaccuracy is additionally compounded by the erroneous computational

annotations of most genes in the reference databases (Schnoes, Brown et al. 2009).

Network analysis on large-scale biological data. Biological datasets usually encompass of a set of discrete objects (*e.g.* proteins or organisms) and their relationships (*e.g.* similarity). As such, they lend themselves naturally to network-based analyses, where connected objects can be organized into groups using clustering – a type of unsupervised learning. Clustering is completely data driven, requiring minimal (if any) prior knowledge of the generated data splits. Commonly used algorithms include K-means clustering and Hierarchical clustering. K-means aims to partition *n* objects into *k* clusters, so that the mean of within-cluster distance, *i.e.* the distance between each cluster member to the cluster center (the cluster representative), is minimal (Lloyd 1982). Hierarchical clustering, as the name suggests, seeks to build a hierarchy of clusters. It can be performed either “bottom up” (agglomerative) or “top down” (divisive) (Rokach 2005). These methods, however, are generally too computationally intensive for handling big data. They also often conspicuously lack the capability of taking the edge weight (*e.g.* level of similarity) into account.

Newer clustering techniques have been developed to deal with the influx of big data. For example, Markov Cluster (MCL) (Dongen 2000) is a fast and scalable clustering algorithm that is based on a very different paradigm. It simulates random walks along the nodes and edges of the graph (network) with two steps, expansion and inflation. Expansion computes random walks of higher length (more steps), linking the “start node” to the “end node”. It is expected that there

are more such higher length paths within cluster than across clusters. Therefore the more higher length paths are there between the two nodes, the more likely are they in the same cluster. The inflation step then selects for intra-cluster walks and against inter-cluster walks, based on the desired tightness of clusters. Iterations of expansion and inflation eventually result in the partition of the graph (Dongen 2000).

Another clustering method is the Louvain algorithm (Blondel 2008), which maps nodes in a network into clusters by considering both edge-weight and node connectivity. Note that when all nodes of a network are connected between themselves, edge-weight is the sole driver of partitioning. The ability to avoid thresholding of similarity (*i.e.* binary designation of similar vs. not similar), allows for better estimates representative of the continuous natural world. An adapted version of the Louvain method (Lambiotte 2008) allows users to further tune the cluster tightness, *i.e.*, the granularity, which offers flexible clustering levels.

This work focuses on using computational methods to answer the two key who and what questions phrased above. In Chapter 2, we describe *fusion* – a network-based analysis of molecular functions of all available bacteria, which allows grouping bacteria by functional similarity; we also compared *fusion* grouping to the current taxonomic assignments of bacteria. Chapter 3, introduces fusionDB, which contains the bacteria-to-function mapping (as described in Chapter 2) and allows users to map new bacterial genomes to their functional abilities and relative locations. In Chapter 4, we describe *mi-faser*, a pipeline (as well as online service) that offers fast and accurate functional annotation of

metagenomes, directly from read data. We conclude this work in Chapter 5 with a discussion of the potential future directions and applications of this study.

Chapter 2
Introduction

In biology, the field of taxonomy is tasked with describing, naming, and classifying organisms; the latter according to some metrics of similarity. Van Leeuwenhoek's observation of microscopic organisms launched centuries of classification based on morphology and physiology (Porter 1976). Since the 1960's, DNA-DNA hybridization (DDH) (Brenner, Fanning et al. 1969) has been the 'gold standard' for bacterial species demarcation. The current polyphasic species definition requires a DDH value >70%, as well as shared phenotypic characteristics, to assign two bacteria to the same species (Stackebrandt and Goebel 1994). Recent emergence of high-throughput genomic sequencing (Margulies, Egholm et al. 2005) highlighted the importance of genomic similarity in bacterial taxonomy. For example, studies have shown that the average genome nucleotide identity (ANI) classifies bacterial species as well as DDH values (Konstantinidis and Tiedje 2005). These new metrics also revealed previously unseen organismal relationships, highlighting the dynamic state of the prokaryotic taxonomy. As there is no one *true* taxonomy, subjectivity is a factor in comparing and contrasting conflicting classifications. Furthermore, special human interest, *e.g.* pathogenicity, and the desire to conserve existing naming conventions add to the inconsistency.

Today, prokaryotic taxonomy relies heavily on phylogenetics. However, there are non-phylogenetic alternatives for classification. Phenetics (Sokal 1973), for example, classifies organisms based on similarity regardless of shared ancestry.

The definition of the term “similarity” is fluid, but in its broadest sense implies a comparison of organism phenotypes, including their molecular functional capabilities. It is important to note that though both phylogeny-based taxonomy (cladistics) and phenetics can be used to investigate bacterial relationships, the questions that they try to answer are different. The task of phylogeny is reconstructing organismal evolutionary *history* – think *Tree of Life* (Woese and Fox 1977, Ciccarelli, Doerks et al. 2006) efforts. Phenetics, on the other hand, clusters organisms into *currently* consistent classes on the basis of observable traits. Closely related organisms are often phenotypically similar. However, the order of evolutionary descent does not directly translate to classification – just as whales are more related to cows than to fish, despite the obvious morphological, environmental, and functional similarities to the latter.

The current NCBI Taxonomy (Benson, Karsch-Mizrachi et al. 2009), a trusted computationally accessible resource, largely follows Bergey’s Manual of Systematic Bacteriology (Garrity GM 2001). Bergey’s Manual is a framework of prokaryotic taxonomy built around a backbone of 16S rRNA-derived phylogeny, which is used to find “unifying concepts of bacterial taxa [leading] to greater taxonomic stability and predictability.” However, as physiology and morphology are also relevant to classification, the boundaries between different taxa are often subjective and controversial (Garrity GM 2001). Additional techniques, *e.g.* multi-locus sequence analysis (MLSA) (Marrero, Schneider et al. 2013), are often used to compensate for the lack of 16S rRNA phylogeny resolution (Fox, Wisotzkey et al. 1992). For the (even highly accurate) computational organism classification

methods (Wu and Eisen 2008) this taxonomic flexibility contributes to inconsistent assignments.

Due to the absence of sexual reproduction and the presence of horizontal gene transfer (HGT), speciation is not strictly defined in prokaryotes. Therefore, the goal of greater classification *stability and predictability* could be better achieved via phenetically clustering organisms on the basis of quantifiable similarity of their molecular function capabilities. In early studies, Enterotubes, a one-stop shop for dozens of biochemical tests, were used to accurately classify *Enterobacteriaceae* (Titsworth, Grunberg et al. 1969); however, these could not be applied to other organisms. Gram staining, on the other hand, could broadly typify bacteria, but lacked in taxonomic resolution. In general, biochemical/physiological tests only reflect a small portion of bacterial functionality – as many as three hundred tests would only access 5-20% of the bacterial functional potential (Garrity GM 2001). Cheaper genome sequencing and advanced computational methods offer a different route for measuring bacterial functional capabilities.

Most of the molecular functionality of one bacterium, its functional repertoire, is carried by its proteome, the set of all proteins encoded by its genes. Note that while plasmid encoded proteins are also part of the proteome, for reasons discussed later in the manuscript, here we only focus on the proteins encoded on the bacterial chromosome. The current taxonomy usually reflects either the phenotypic manifestations of functional repertoire subsets (morphology, physiology) or high-level repertoire interpretations (*e.g.* DDH). Ideally, however, comparison between bacterial repertoires should offer a comprehensive metric

for clustering bacteria on the basis of their overall functional similarity – a combination of heritage and habitat impact.

We defined the functional repertoires of over 1,300 fully sequenced bacteria using protein clustering by HSSP (Homology-derived Secondary Structure of Proteins) distance (Rost 2002). HSSP techniques allow annotating two proteins as performing the same molecular function, without specifically defining the nature of this function. We also annotated our set of bacterial proteins via common function profiling tools: COG (Tatusov, Fedorova et al. 2003), Pfam (Punta, Coggill et al. 2012), and RAST (Aziz, Bartels et al. 2008). For the purposes of this work, we defined the similarity between any two organisms according to the percentage of functions they shared. We first validated the reliability of our functional similarity metric by using pairwise organism comparison to assign taxonomic ranks. Using the NCBI Taxonomy as a benchmark, we show that functional similarity, defined using any of the above-mentioned function annotation methods, is more descriptive of pairwise organism similarity than gene sequence identity – a novel finding. Additionally, our HSSP-based organism similarity metric was more accurate than metrics based on other function assignments evaluated in this study. Since HSSP is not limited by availability of annotations, our approach circumvents experimental limitations by including novel lesser-studied functions into organism classification.

We further identified natural clusters of bacteria in our functional-repertoire similarity-based organism network (FuSiON; flattened to *fusion*). Instead of assigning organisms into phylogeny-based classes, each of which may

encompass a wide range of environmentally, metabolically, and phenotypically diverse microbes, *fusion* groups them according to functional similarity. Our scheme allows for variability in the number of non-hierarchical organism modules, where the clustering resolution is adjustable to each specific application. Moreover, as *fusion* is inherently cut-off free, its clade assignments are largely independent of current database biases, *i.e.* our method will not tend to assign a novel microbe to *Proteobacteria* simply because a vastly larger and more diverse set of *Proteobacteria* genomes are available in our databases. We investigated the functional basis for some of the individual discrepancies between the current taxonomy and the *fusion* classification via case studies in *Cyanobacteria* and *Mycoplasma*. We describe how phylogenetically related bacteria can still be functionally very different, with the environment playing a key role in selecting for each organism's functional specificity. Our novel phenetic method for unambiguous and consistent classification of bacteria provides a complementary view to phylogenetic clade assignment. The dynamic nature of our network-based organism clustering provides an easy route for incorporation of additional organisms and organism features (*e.g.* plasmids) into the existing classification framework. *Fusion* is, thus, a more practical fit for biomedical, industrial, and ecological applications, *e.g.* (Glare, Caradus et al. 2012, Krishnan, Bharathiraja et al. 2014), as many of these rely on understanding the functional capabilities of the microbes in their environment, and are less concerned with phylogenetic descent.

Note that we are currently working on implementing a publicly available *fusion* work-bench, which will allow real-time assignment of novel organisms to *fusion* clades. For now, fusion development data, which are complete enough to reproduce our work, are available at http://bromberglab.org/?q=services. In the same place we also have a (rudimentary) database (fusionDB) that contains our computational results, implemented for any new requests on our network.

Methods

Datasets. We downloaded 1,374 bacterial proteomes from December 2011 NCBI GenBank release (Benson, Karsch-Mizrachi et al. 2009). Habitat information for these organisms was obtained from GOLD (Pagani, Liolios et al. 2012) and IMG (Markowitz, Chen et al. 2012).

Defining functional repertoires and their similarity. We defined the functional repertoire of a single microorganism to be the set of all molecular function capabilities carried by its proteome (excluding plasmids).

HSSP-based protein clustering. We performed an all-to-all PSI-BLAST (Altschul, Gish et al. 1990) of 4.2 million protein sequences in the 1,374 bacteria proteomes (parameters: e-value $1e^{-3}$; inclusion ethresh $1e^{-10}$; num iterations 3; max target seqs $1e^{9}$; num alignments $1e^{9}$). HSSP distances (Rost 2002) were calculated from the PSI-BLAST results (Eqn. 1), where *L* is the length of the alignment between two proteins and *Id* is the percentage of identical residues.

$$HSSP\ distance = \begin{cases} -99, L < 11 \\ Id - 480L^{-0.32\left(1+e^{-\frac{L}{1000}}\right)}, \ 11 < L \leq 450 \\ Id - 19.5, L > 450 \end{cases} \qquad \textbf{(Eqn. 1)}$$

The highest HSSP distance was selected for every pair of proteins when multiple alignments were possible. Note that here higher distance means higher similarity. A threshold of HSSP distance $\geq$10 was used to define two proteins as having similar function. At this threshold, the HSSP metric attains ~90% precision and ~40% recall in mapping functional identity of protein pairs (Rost 2002). We

further clustered these proteins into function groups using MCL (Markov Cluster Algorithm; parameter: -I 1.4) (Dongen 2000).

Other function profiling tools. We obtained COG (Clusters of Orthologous Groups) (Tatusov, Fedorova et al. 2003) annotations for our dataset (personal communication with Dr. Yuri Wolf). We downloaded the Pfam database (release 27.0) (Punta, Coggill et al. 2012) and annotated all proteins using hmmscan (Eddy 2011) against both PfamA and PfamB with default settings. We kept the top hit for each protein with e-value < $1e^{-3}$. We used a local install of the RAST toolkit (myRAST) (Aziz, Bartels et al. 2008) to annotate the function of all proteins. Each annotation was made at the default reliability level (parameters: -reliability 3). All the proteins that were not annotated by COG, Pfam and RAST were counted as representing individual functions.

The *functional repertoire similarity* of two organisms was calculated as the number of shared functions in each functional repertoire (as defined by different tools above) divided by the bigger repertoire size. We assumed that similar organisms should have similar repertoire sizes, thus a vast difference indicates low similarity.

For comparison to gene content phylogenomic approaches, we also calculated the *whole-genome similarity* as the number of shared homologous proteins (homology inferred via 40% sequence identity) normalised by the bigger proteome size.

Annotation of function groups derived from HSSP-based protein clustering. We divided all 4.2 million proteins in our set into three categories based on their

RAST annotation: 1) *known*, sequences with available function annotation; 2) *hypothetical*, sequences with "hypothetical" or "putative" in their annotation, or annotated as "protein" or "Uncharacterized protein conserved in bacteria," and 3) *unknown*, sequences with no annotations at all. We further assigned all of our HSSP-based function groups to one of three categories; for a given function group: 1) *Kn* if it contains at least one sequence of the *known* category; 2) *Hy* if it contains no *known* sequences and at least one *hypothetical* sequence and 3) *Un* if it contains only unknown sequences. In addition, we also tagged our function groups as 1) *shared*, if they exist in more than one organism in the dataset or 2) *unique*, if they exist only in one organism.

Comparing the performance of the different pairwise similarity metrics to infer organism taxonomy. For every pair of organisms of known NCBI Taxonomy identity (Benson, Karsch-Mizrachi et al. 2009), functional repertoire similarities were computed based on COG, Pfam, RAST, and our HSSP-based method. Each method provided either (i) a correct assignment to the same taxon (true positive, TP), (ii) incorrect assignment to the same taxon (false positive, FP), (iii) correct assignment to different taxa (true negative, TN), or (iv) incorrect assignment to different taxa (false negative, FN). The accuracy (positive accuracy, precision; PA) and coverage (positive coverage, recall; PC) were computed for every metric at every threshold (Eqn. 2). We then compared the taxonomic classification performance of different functional similarity metrics and the proteome similarity.

$$PA = \frac{TP}{TP+FP} \qquad PC = \frac{TP}{TP+FN} \qquad \textbf{(Eqn. 2)}$$

Bootstrap analysis was performed by randomly sampling 10% of the data with replacement 100 times for each taxonomy level. AUC (Area Under the Curve) under the accuracy/coverage (precision/recall) curve was calculated (Eric Jones 2001 -) for every functional similarity metric and paired t-tests were performed for every pair of metrics.

Generating functional-repertoire similarity-based organism networks. ***Fusion*** and *fusion*+ networks were visualized using Gephi (Bastian, Heymann et al. 2009) OpenORD (Wu, Wu et al. 2011) and ForceAtlas2 (Jacomy, Venturini et al. 2014), respectively.

In *fusion* each 1,374 organisms (vertices/nodes) are connected by 943,251 edges whose weights reflect the pairwise organism functional repertoire similarities. In *fusion*+ vertices/nodes represent organisms and function groups. A (larger) organism node shares edges with its (smaller) function group nodes. Organism nodes are linked to each other only via function group nodes; *i.e.* there is no edge directly linking organism nodes. The common function group nodes are between organism nodes, while the unique function nodes tend localize near the edges of the network.

Calculating overall accuracy and coverage for singly linked networks. In single linkage clustering any two nodes that share an edge are assigned to a single cluster regardless of their similarity to other nodes in that cluster. The presence of an edge indicates similarity of organisms above a minimum cut-off, but the level of similarity is not further considered. Isolated organisms, with no connection to any other organism in our set, were not shown.

We measured the performance of single linkage clustering in identifying current taxonomic assignments for a series of similarity cut-offs (5%-100%, at step of 5%, Figure 1B,C). For each cut-off, we assigned all organisms in one single linkage cluster to the taxon of the most common organisms in that cluster; *e.g.* if a cluster of three organisms contained two organisms of taxon X, all three were assigned to the taxon X. The overall network accuracy was calculated as the sum of all the correctly assigned organisms divided by the total number of organisms (Eqn. 3).

$$OverallAcc = \frac{\sum_{i=1}^{n} correctly\ assigned\ organisms\ in\ cluster\ i}{total\ number\ of\ organisms} \qquad \textbf{(Eqn. 3)}$$

We also identified the organism clusters consistent with taxonomic assignments of their members; *e.g.* if 7 organisms are assigned to a taxon X, and 4 of them are in cluster A, then A is considered the *major* cluster of X. For each taxon, the coverage is the fraction of its members that are in the *major* cluster (Eqn. 4); *e.g.* for X in our example coverage is 57%. At 100% coverage all members of a taxon are in one cluster. For a given taxonomy level, the overall network coverage was calculated as the number of taxa with 100% coverage divided by the total number of taxa at this level (Eqn. 5). Note that taxa with only one member would contribute trivially to the performance, and thus were excluded for these calculations.

$$Cov = \frac{Organisms\ in\ the\ major\ cluster}{Total\ number\ of\ organisms\ in\ the\ taxon} \qquad \textbf{(Eqn. 4)}$$

$$OverallCov = \frac{Taxa\ with\ 100\%\ coverage}{Total\ number\ of\ taxa\ with\ more\ than\ one\ organism} \qquad \textbf{(Eqn. 5)}$$

Comparing single linkage functional network-based organism classification to the NCBI Taxonomy. The 100-layer network-derived hierarchy was built by starting at the threshold of 0% functional repertoire similarity, *i.e*. all 1,374 bacteria are in a single cluster, and moving outward in 1% increments until the 100% similarity threshold was reached. For a given cluster of organisms sharing at least X% similarity, we (i) clustered the organisms at (X+1)% similarity, (ii) calculated the distance between every two clusters by computing the average of all inter-cluster pairwise similarities of organisms and (iii) built a neighbor-joining tree (layer) of the clusters using PHYLIP (Felsenstein 2005). By combining all layers we obtained a 100-layer hierarchical tree-like structure. This hierarchical structure provides a compact visual representation of functional similarity of our large groups of microorganisms. Note, however, that it is not a phylogeny tree and does NOT directly convey organismal evolutionary relationships.

NCBI Taxonomy hierarchical tree-like structure was generated with iTOL (Letunic and Bork 2011) using the NCBI Taxonomy IDs (Benson, Karsch-Mizrachi et al. 2009). We then computed the correlation (ranged -1 to 1) between network and NCBI-derived hierarchical structure using Patristic (Fourment and Gibbs 2006). The hierarchical structures were first converted to distance matrices in which the distance between two organisms was calculated as the steps between them in the hierarchy. We also built 6 and 10 layer network-derived structures to show that the difference in the number of layers is not relevant to the comparison of the topological *relative* distances of any two organisms across hierarchies.

Detection of *fusion* modules and calculation of Jaccard index. We identified modules in the complete (no similarity cut-offs) *fusion* with Louvain method implemented in Gephi at a series of resolutions (0.05 to 1.2). We further calculated the Jaccard index to compare organism assignments from *fusion* modules to the NCBI Taxonomy. At a given resolution, the Jaccard index is calculated as the number of organism pairs assigned to both the same *fusion* module and the same NCBI Taxonomy bin, divided by the number of organism pairs assigned to either the same *fusion* module or the same NCBI Taxonomy bin.

Results and Discussion

HSSP-based functional repertoire similarity accurately measures pairwise bacterial relationships. We annotated functions of 4.2 million proteins, encoded in 1,374 fully sequenced bacterial genomes via COG, RAST, and Pfam. We also computed HSSP distances for every proteins pair (~1.6×10^{13} comparisons). The HSSP distance is a non-linear metric incorporating sequence identity and alignment length that has been parametrized to identify alignments of proteins of experimentally established identical functions (Rost 2002). Briefly, enzymes of experimentally defined identical function (defined by the Enzyme Commission (EC 1992)) were used to determine a threshold curve separating the alignment length *vs.* sequence identity space into regions of same *vs*. different functions; *i.e.* two proteins that fall above the curve share identical function, while those below the curve do not. The distance of every alignment along the sequence identity axis away from the curve (HSSP distance) reflects the reliability of these assignments of functional identity (Rost 2002).

We adopted an HSSP distance cut-off of 10, which annotates two proteins as sharing the same function with over 90% precision (accuracy/specificity, percentage of correct same-function predictions of all such predictions made), albeit at only ~40% recall (coverage/sensitivity, percentage of correct same-function predictions of all same-function pairs in the set) (Rost 2002). At this stringency, ~900,000 proteins (21% of 4.2 million in our set) were unique – one protein per functional group. The remaining 3.3 million clustered into ~335,000 functional groups (Table 1). Note that at lower HSSP cut-offs these groups can

be further consolidated, but at a significant loss to accuracy. We choose a more conservative, threshold to attain maximal resolution of assignment.

Table 2-1 Annotation status of HSSP-based function groups.

	Function groups (>1 sequence)	Function groups (1 sequence)	Total
Known (Kn)	190,272	245,430	435,702
Hypothetical (Hy)	119,825	267,160	386,985
Unknown (Un)	24,925	387,768	412,693
Total	335,022	900,358	1,235,380

We used RAST annotations to divide our HSSP-based functional groups into *Kn* (known; available annotation), *Hy* (hypothetical; likely protein existence, function not annotated) and *Un* (unknown; no annotation) sets (Table 1; Methods). We further confirmed that each HSSP-based function group contained proteins of similar RAST annotations (Table 2). Note that different function groups may contain proteins that carry out the same biochemical functions but in a different fashion, *e.g.* at different reaction rates. We found that many organisms contain proteins performing the *Kn* functions, while the *Hy* and *Un* functions tend to be organism specific, a conclusion that holds even if groups containing a single protein are excluded (Figure 1). As a corollary, proteins carrying functions that are more common across organisms are more likely to be annotated (Figure 1). Interestingly, we note that 26% (127,254 of 481,913) of the unannotated proteins in our set fall into the *Kn* (78%) and *Hy* (22%) HSSP-based function groups. We also show that for 71% of *Kn* groups (Table 2), 90-100% of annotated proteins in each group are functionally identical. Our protein clustering may thus help elucidate functions of tens of thousands of yet un-annotated proteins; we

anecdotally confirmed some of these via manual curation of new sequence annotations.

Table 2-2 Distribution of proteins of the same functional annotation among all the HSSP-based function groups.

Proteins with same annotation (% of all in a group)	# of groups (% of total)
0-50	12,777 (6.7)
50-90	41,634 (21.9)
90-100	135,861 (71.4)

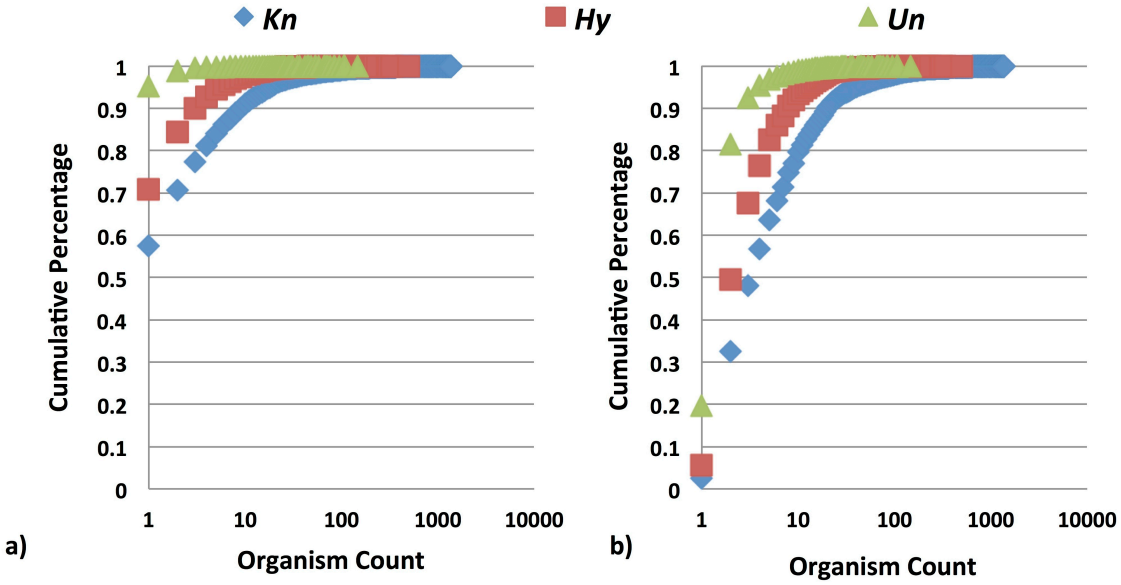


Figure 2-1 Function groups that are shared by many organisms are more likely to be experimentally annotated (*Kn>Hy>Un*). (a) all function groups, and **(b)** function groups containing at least two proteins.

We defined the functional repertoire of an organism as the set of all functional groups carried by the organism. The size of the repertoire is at most as large as the number of proteins in the proteome, but in-paralogs may fall into a single functional group. The functional similarity of two bacteria was calculated as the number of shared function groups normalized by the bigger repertoire size (Methods).

Our HSSP-based functional group comparison significantly (Wilcoxon rank-sum test, *p-value*<0.0001; Methods) more accurately recapitulates the NCBI taxonomic identity of organism pairs than using other function definitions (COG, RAST, and Pfam) at all taxonomy levels, except the genus and species, where RAST achieves comparable performance (Figure 2A-F) RAST's and HSSP's improved performance at these lower levels may be due to their "whole sequence"-based function annotation. Pfam works at the domain level, which is arguably too broad, including many proteins into one function class. COG is designed to detect orthology, *i.e.* evolutionary relationships, and thus its functional groups are likely too narrowly defined. HSSP's exemplary performance over all taxonomic levels is possibly due to the lack of dependence of its pairwise sequence comparisons on the external knowledge, *e.g.* Pfam domains, RAST functions, or COGs. Note that here we used COG instead of the more complete EggNOG (Powell, Forslund et al. 2014), as we felt that manual curation may carry more resolution. We obtained the latest set of COG annotations from its developers (2012 update, Yury Wolf personal communication). Here we show that *all* tested function-based metrics reflect the current taxonomic organism placement fairly well. We adopt HSSP for this work as it correlates best with the current taxonomy (Figure 2A-F), while circumventing limitation of available protein function annotations.

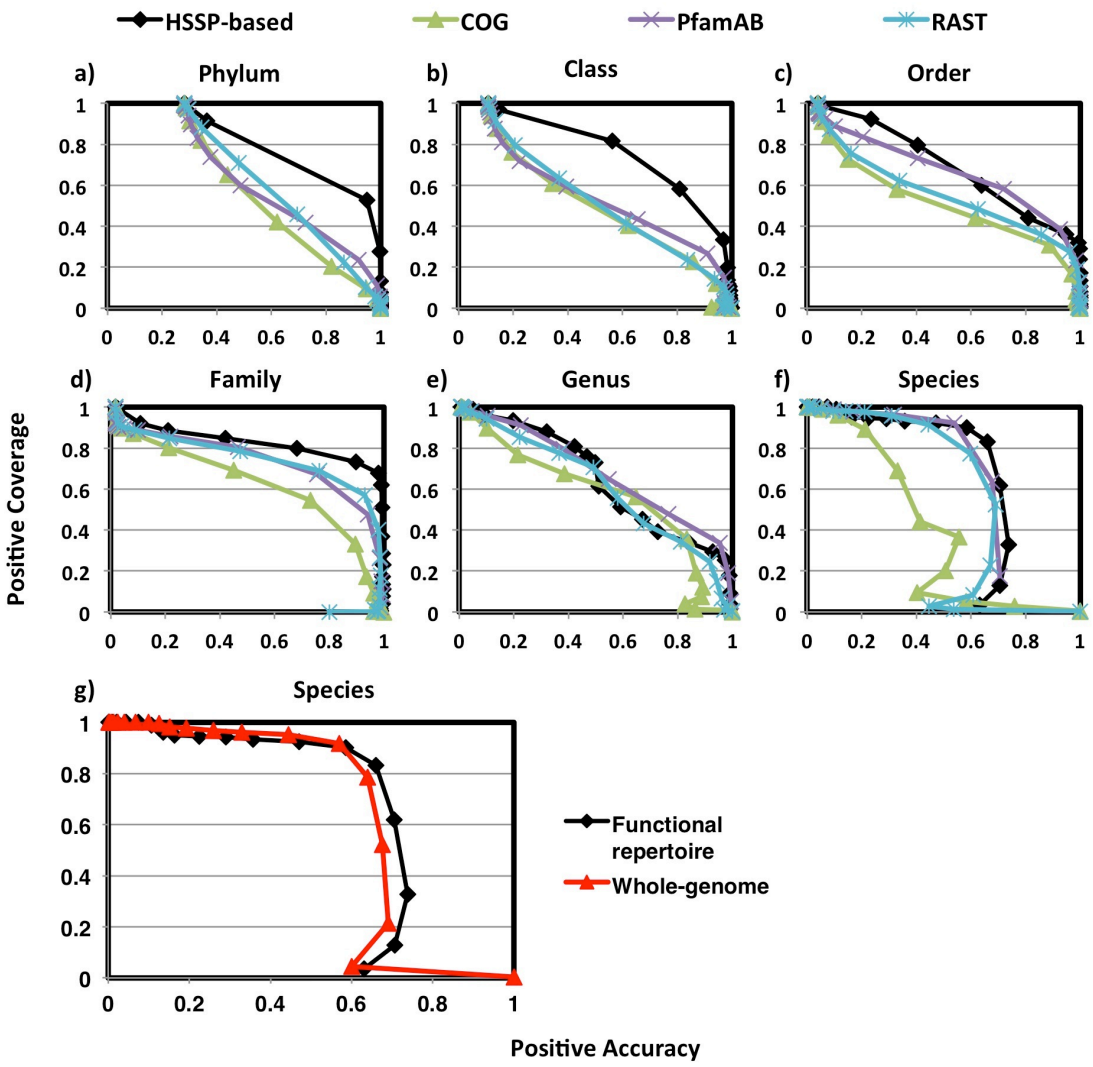


Figure 2-2 HSSP-based functional similarity correlates with the NCBI taxonomy better than other function definitions. Compared with COG, Pfam (both A and B model databases) and RAST -based functional repertoire similarity, HSSP shows better performance ($P<0.001$, Wilcoxon test) in assigning two organisms into the same **(a)** phylum, **(b)** class, **(c)** order, and **(d)** family. It is better ($P<0.001$) than Pfam and COG, but not significantly different from RAST in assigning **(e)** genus and **(f)** species. **(g)** HSSP-based functional repertoire similarity also classifies organisms into species significantly better ($P<0.001$) than a sequence identity-based metric.

As described above, the HSSP metric is more informative of function than protein sequence identity and alignment length alone (Rost 2002). Thus, although our method is mechanistically similar to sequence-based gene content phylogenomic approaches (Snel, Bork et al. 1999, Dagan, Artzy-Randrup et al. 2008), it is very different from the latter both (1) conceptually – we classify organisms based on

their current functional similarity rather than reconstructing their phylogeny and (2) practically – functional similarity significantly more accurately describes bacterial relationships than sequence identity-based methods (Wilcoxon rank-sum test *p-value*<0.0001; Figure 2G). The latter finding is intuitive, as function-based methods separate sequence-similar out-paralogs into different families, which sequence-based methods, by definition, cannot do. However, to the best of our knowledge the improvement of functional comparisons over gene content in classifying bacteria has not been experimentally shown before.

We find, perhaps unsurprisingly, that two nearly functionally identical (90% similarity) organisms belong to different species as often as a third of the time (Figure 2F). These functionally similar, yet taxonomically split organism pairs are not uniformly distributed throughout the taxonomy (Garrity GM 2001, Konstantinidis and Tiedje 2005). Here we show that most of these occur in three pathogenic genera: *Borrelia* (Lyme disease), *Brucella* (brucellosis), and *Mycobacterium* (leprosy, tuberculosis), suggesting possible bias of classification towards higher resolution for organisms of human interest. This preference is also evidenced by the relatively large number of experimental annotations of functions of the human-associated microbiome (Figure 3) Though such taxonomic resolution bias probably offers convenience in practice, it brings along an inconsistency that complicates *en bulk* analysis of microorganisms; *e.g.* computational methods cannot readily deal with the type of subjectivity that separates very similar organisms into different taxa (*e.g. Borrelia hermsii* and *Borrelia turicatae* share 99% functional similarity), while assigning different

organisms into the same taxon (*e.g. Clostridium botulinum* strains share less than 40% similarity). We argue that for practical use, it is often more important to know whether two organisms can perform the same molecular functions rather than if they share the same lineage.

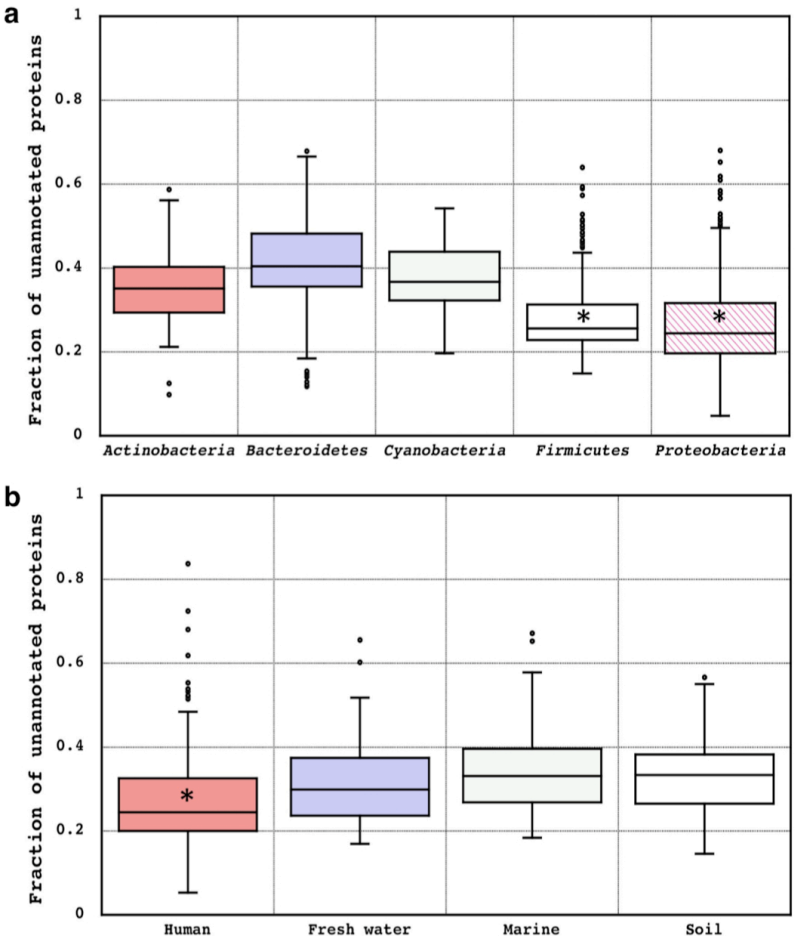


Figure 2-3 Bias in functional annotation of bacterial proteomes. Box-and-whisker plots representing the distributions of the unannotated fractions of bacterial proteomes among **(a)** major phyla and **(b)** different habitats. The upper/lower box bounds represent the corresponding quartiles with the median shown as the crossbar. The "whiskers" of each box are the set's minimum/maximum values with the outliers (individual points >1.5 times the interquartile distance away from the quartile bounds) not included in the calculations. The organisms from better/longer-studied phyla, or from the human microbiomes, are generally more functionally annotated. The * symbol indicates statistical significance for 1) Proteobacteria vs. Actinobacteria, Bacteroidetes and Cyanobacteria; 2) Firmicutes vs. Actinobacteria, Bacteroidetes and Cyanobacteria; 3) Human vs. Fresh water, Marine and Soil ($P<0.0001$, Mann-Whitney U-test). Note that increased "human interest" is indicated by the in-depth study of human symbiont organisms, as compared to inhabitants of other environments.

Note that throughout this work, in order to compare our organism assignments to the current taxonomy, we conservatively excluded the plasmid proteomes. Plasmids contribute heavily to functional differentiation, as opposed to speciation, separating classes of microorganisms without explicit phylogenomic commitment. Moreover, plasmids follow independent evolutionary models (Sykora 1992, Halary, Leigh et al. 2010) and carry many of the environment-related functions (Lawrence 2002). We expect that including the plasmid genomes into our paradigm will show stronger impact of habitat and we intend to evaluate plasmid contribution in further work.

***Fusion* organism classification correlates with the NCBI Taxonomy.** We represented the functional similarity of our microorganisms as a network – *fusion* (functional-repertoire similarity-based organism network). In *fusion*, organisms are vertices (nodes), and edge lengths (weights) indicate pairwise functional repertoire similarities. Here all organisms (1,374 nodes) are at least somehow similar forming a fully connected network (943,251 edges). The minimum amount of similarity between two organisms is <1% -- these edges link the tiny *Candidatus* microbes (Table 3) to the much bigger organisms in our set. However, the most common level of similarity between two organisms is 7% (mean 7.7% and median 6%). These results indicate that our organisms are mostly functionally distant, but maintain a minimal set of identical, globally present, likely housekeeping, functions. In a representation that takes into account edge-weight and node density (Figure 4A; OpenORD layout (Wu, Wu et

al. 2011)), microorganisms cluster consistently within their NCBI Taxonomy groups.

Table 2-3 Six bacteria not matching any organisms in the functional repertoire-based network at 10% cutoff.

Organism	Functional Repertoire Size	Phylum/Class
Bdellovibrio bacteriovorus HD100 (uid61595)	3,426	Deltaproteobacteria
Candidatus Carsonella ruddii (uid58773)	181	Gammaproteobacteria
Candidatus Hodgkinia cicadicola Dsem (uid59311)	168	Alphaproteobacteria
Candidatus Tremblaya princeps PCIT (uid68741)	119	Betaproteobacteria
Fibrobacter succinogenes S85 (uid41169)	2,837	Fibrobacteres
Mycoplasma haemofelis Langford 1 (uid62461)	1126	Tenericutes

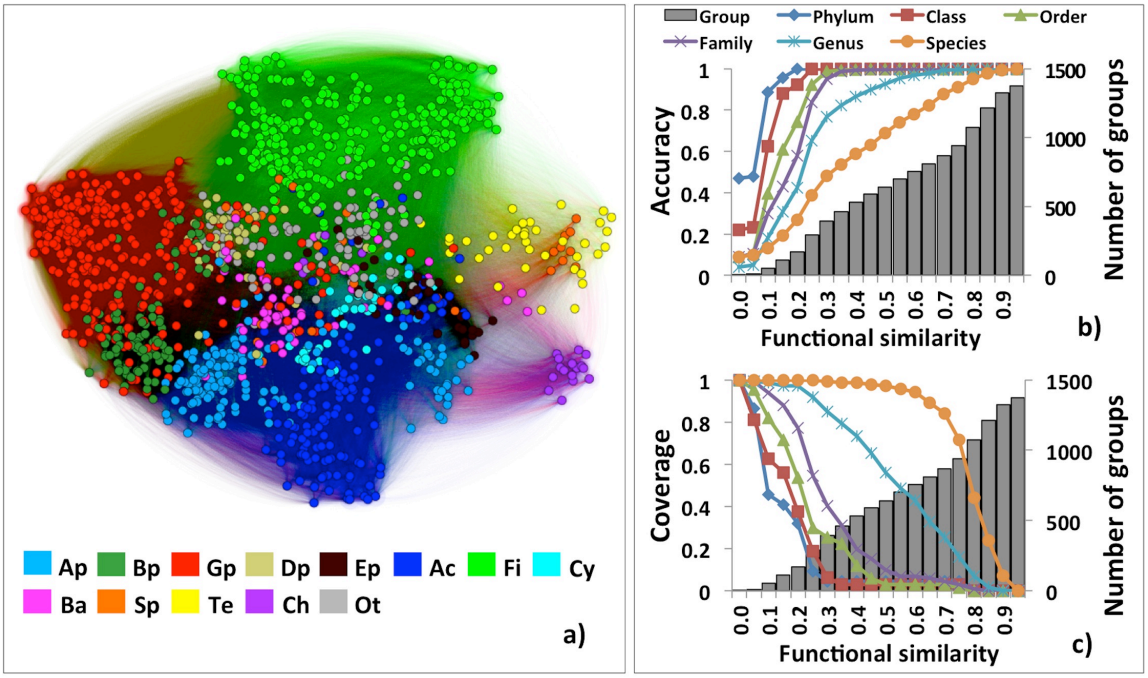


Figure 2-4 Fusion-based clustering correlates with NCBI Taxonomy. (a) fusion network colored by taxonomic rank. Ap-Alphaproteobacteria; Bp-Betaproteobacteria; Gp-Gammaproteobacteria; Dp-Deltaproteobacteria; Ep-Epsilonproteobacteria; Ac-Actinobacteria; Fi-Firmicutes; Cy-Cyanobacteria; Ba-Bacteroidetes; Sp-Spirochaetes; Te-Tenericutes; Ch-Chlamydiae; Ot-other minor phyla; **(b)** The overall accuracy of functional similarity networks at cut-offs from 5% to 100%, with step of 5%. The overall network accuracy is the fraction of correctly assigned organisms of the total number of organisms; i.e. overall accuracy of 100% indicates that all organisms in any one cluster are of the same taxon. The overall accuracy for each taxonomy level increases with the cut-off. Thus, lower taxonomy levels (e.g. genus, species) achieve 100% overall accuracy at higher cut-offs; **(c)** The overall coverage of the functional similarity networks at cut-offs from 5% to 100%, with step of 5%. The overall coverage is the percentage of taxa (excluding taxonomic singletons) with all members in one cluster at a given cut-off. Overall coverage of 100%, indicates no splitting of any of the taxa; i.e. one cluster per taxon. Lower taxonomy levels lose 100% overall coverage at higher cut-offs.

Earlier studies searched for natural discontinuity in the bacterial pairwise genome similarity space (Staley 2006, Goris, Konstantinidis et al. 2007), but found no unique break point that would reasonably assign taxa to large sets of organisms. To inspect for possible occurrence of these breakpoints in our network representation, we adopted a range of cut-offs in a single linkage clustering approach. In single linkage clustering any two nodes that share an edge are assigned to a single cluster regardless of their similarity to other nodes in that cluster. The presence of an edge indicates similarity of organisms above a

minimum cut-off, but the level of similarity (edge weight) is not further considered. Thus, a large and diverse set of organisms could form a single cluster at fairly high similarity cut-offs; *i.e.* if the similarity cut-off is 15% and organism A is 20% similar to organism B, while B is 20% similar to organism C, then all three organisms are assigned to the same cluster even if A and C share less than 10% similarity. At the 10% cut-off, *i.e.* when a minimum of 10% repertoire similarity creates an edge between two organisms, four clusters were formed, encompassing 1,368 (99% of all) organisms (Figure 5A). Note that at this 10% cut-off, we removed the majority of the edges in our network (~86%). As expected from a large and diverse group, 1,355 organisms fell into one cluster. The other three clusters contained a total of 13 organisms, including five of the *Planctomycetes* phylum in one cluster, six of the *Leptospira* genus in another, and two *Mycoplasma suis* species strains in the third. The separation of *Planctomycetes* can be explained by the uniqueness of this phylum (Fuerst and Sagulenko 2011). However, the split of *Leptospira* away from other genera of *Spirochaetes*, as well as the split of *Mycoplasma suis* and *Mycoplasma haemofelis* Langford 1 from each other and other *Mycoplasma,* highlight the (known) disagreements of the current taxonomic clade assignments with these organisms' functional abilities (Garrity GM 2001). Note, however, that *Spirochaetes* and *Tenericutes* (to which *Mycoplasma* belong) make up less than 2% of our set, each. Thus, their functional split could also suggest experimentally determined lack of similar genomes. The six singletons, *i.e.* organisms sharing less than 10% functional similarity with any other organisms in our dataset, are

summarized in Table 3. Individuality of some of these can be explained – *Fibrobacter succinogenes* S85 is the only *Fibrobacteres* member in our dataset, as may be the three *Candidatus* organisms of unusually small repertoire sizes. However, the reasons for differentiating *Bdellovibrio bacteriovorus* HD100 from its taxonomic neighbors must be rooted in the dissimilarity of functional annotations and taxonomic assignments.

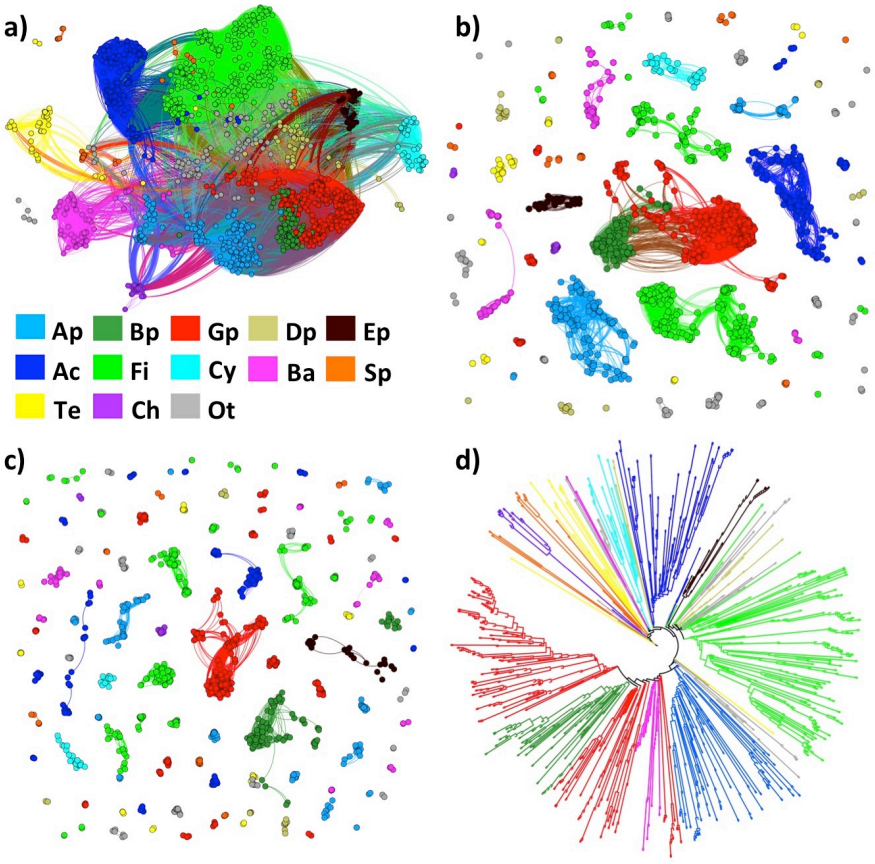


Figure 2-5 Functional network single linkage clustering correlates with NCBI taxonomy. Single linkage clusters based on minimum functional repertoire similarity cut-offs of **(a)** 10%, **(b)** 20% and **(c)** 30%. As the cut-off increases, smaller-size taxonomically consistent clusters are observed. The 100-layer hierarchy of the 1,374 bacteria in **(d)** is derived from cut-off-based single linkage clustering at 1% steps, from 0% to 100% cut-offs (Methods). Ap-Alphaproteobacteria; Bp-Betaproteobacteria; Gp-Gammaproteobacteria; Dp-Deltaproteobacteria; Ep-Epsilonproteobacteria; Ac-Actinobacteria; Fi-Firmicutes; Cy-Cyanobacteria; Ba-Bacteroidetes; Sp-Spirochaetes; Te-Tenericutes; Ch-Chlamydiae; Ot-other minor phyla.

With increasing cut-offs our network contained organism clusters that were progressively more taxonomically consistent at lower taxonomic ranks (Figure 5A-C). This split into clusters is informed by the variation in density of organisms across the network, *i.e.* the increased connectivity between nodes within one region as compared to outside the region. Note that density is artificially increased in regions of preferentially studied organisms (e.g. *Firmicutes* and *Proteobacteria*, Figure 3) To study the mapping of functional relationships to taxonomy we used 1% cut-off increments in the network to build a 100-layer hierarchical structure (Methods; Figure 5D). We found that this structure was somewhat topologically similar (corr=0.557) to the NCBI Taxonomy. However, the differences between the two indicated the absence of natural breakpoints correlating the current taxonomy to functional groupings of microorganisms.

To quantify the cluster-taxon consistency, we calculated the overall network accuracy and coverage at different cut-offs (Methods). With the cut-off increasing from 5% to 100%, the overall accuracy increases while the overall coverage decreases for each taxonomy level (Figure 5B and 5C). Note that the 100% overall accuracy for the species level is only attained at 100% cut-off, which results in one organism per cluster (Figure 5B); *i.e.* NCBI Taxonomy assigns highly similar organisms into different species. On the other hand, even 10% functional similarity does not guarantee 100% overall coverage for most (phylum to genus) taxonomic levels (Figure 5C). All strains of a single species consistently fall into a single cluster (100% overall coverage) only until the 30% cut-off; *i.e.* highly dissimilar organisms are classified into the same species.

The lowest cut-off resulting in 100% overall accuracy, along with the highest cut-off resulting in 100% overall coverage, define lower and upper bounds, respectively, of the functional repertoire similarity in assigning NCBI Taxonomy. Organisms in different clusters at cut-offs less than the lower bound are of different taxa, while organisms in the same cluster at cut-offs greater than the upper bound are of the same taxon. The ranges of uncertainty of taxonomic assignment (region between the lower and upper bound) are varied and often large, *e.g.* spanning cut-offs of 5-95% for genus-level classification (Figure 6A). Pairwise comparisons (Figure 6B) display similar behavior, highlighting inconsistencies in the prokaryotic taxonomy, previously quantified by *e.g. (Konstantinidis and Tiedje 2005)*. Arguably, even more disconcerting for pairwise comparisons is the fact that >90% of all organism pairs fall into this uncertainty range for all taxonomic ranks except for species (most organism pairs are of different species); *e.g.* for phylum level 97% of all organism pairs are in the uncertainty region. These results indicate that setting arbitrary cut-offs, whether network- or pairwise- comparison-based, in order to fit organisms into preset taxonomic bins, inevitably introduces unquantifiable and non-standardizable bias into annotations – a problem for large-scale organism and microbiome studies.

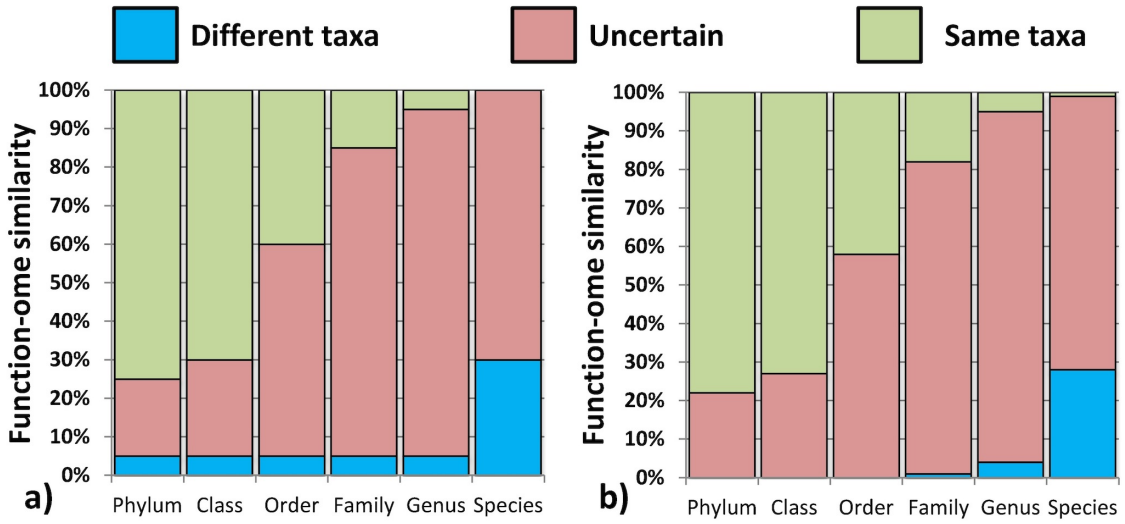


Figure 2-6 Both network-based single linkage clustering and pairwise functional repertoire similarity correlate poorly with NCBI taxonomy. Similarity range for each NCBI taxonomy level defined by **(a)** network and **(b)** pairwise functional repertoire similarity cut-offs. The green and blue parts of the column indicate the similarity ranges where organisms can be unequivocally assigned to the same or different taxa, respectively. The red part of the column indicates uncertainty where, such assignment cannot be conclusively made. Note that for organism pairs at all taxonomic levels, except species, **(b)**, the number of pairs in the red region exceeds 90%. Intuitively, for the species level, the majority of organism pairs (98%) fall into different species.

***Fusion* modules reflect non-hierarchical organism groupings.** State of the art in any field often concerns itself with describing available data points and extrapolation on the basis of observed trends. Current prokaryotic taxonomy is, thus, primarily defined on the basis of culturable and commonly studied microorganisms, *e.g. Proteobacteria* and *Firmicutes*, which make up 46.8% and 21.7% of our data set, respectively. Furthermore, the number of well studied organisms of a particular kind is often the driving force of taxonomic placement of newly discovered (sequenced) organisms; *i.e.* you could only compare a new organism to existing ones, so better represented clades are more likely to be populated with additional members. For example, when looking to classify a newly cultured microbe on the basis of 16S rRNA sequence similarity, one is

simply more likely to find a closer, even if not sufficiently close, sequence belonging to a well studied clade than to a poorly described one. Re-assignment of organisms to new clades on the basis of additional evidence is fairly common. However, follow-up studies are time consuming, limited to organisms of high interest, and, thus, unlikely to find all errors. High-throughput experimental methods (*e.g.* cheaper sequencing) and automated organism classification can contribute to further propagation of assignment errors. An unfortunate, but highly visible result of this state of the art is the significant difference in annotations of organism diversity of the same metagenomic sample using data provided by different 16S rRNA databases (Delmont, Prestat et al. 2012).

Network-based organism similarity representations can help alleviate issues of data availability bias. In a fully connected network of similarities non-overlapping modules, with denser (edge weight-wise) within-module and sparser across-module connectivity, imply natural organism grouping. The Louvain algorithm (Blondel 2008) maps nodes in a network into modules by considering both edge-weight (extent of similarity) and node connectivity. When all-to-all connectivity exists within a network, edge-weight is the sole driver of module detection; *i.e.* five very similar organisms can form a module of their own as well as ten or twenty organisms. In fact, a larger number of organisms is more likely to connect strongly outside the module and, thus, be subject to dispersion. A newly identified organism, placed into a fully connected network is then subject to forces (connections) pulling from all directions, to finally identify its placement. This placement is dynamic – as new organisms are added a network's

partitioning can change. As a result, this approach is more robust to dealing with natural organism diversity then static structures.

For our purposes, one big advantage of the Louvain algorithm is that it splits the fully connected *fusion* network into communities (modules) without a need for a set arbitrary similarity cut-off. However, a problem with this single best grouping of organisms is that when the global modularity function is optimized, there is a loss of resolution for smaller modules. An adapted version of the Louvain method (Lambiotte 2008), instead of modularity, aims to optimize stability of network partitions over time. Here, stability reflects flows of probability through the network, capturing important aspects of the global architecture and describing different optimal partitions of the network at different times. Simply put, a module is considered stable if random walkers (described by a particular Markov process (Lambiotte 2008)) do not escape from it within the set time limit. Thus, longer time limits (higher “resolution” parameter values (Figure 7) result in larger and coarser (more functionally diverse) modules. The size and diversity of organism modules can thus be optimized for each individual application.

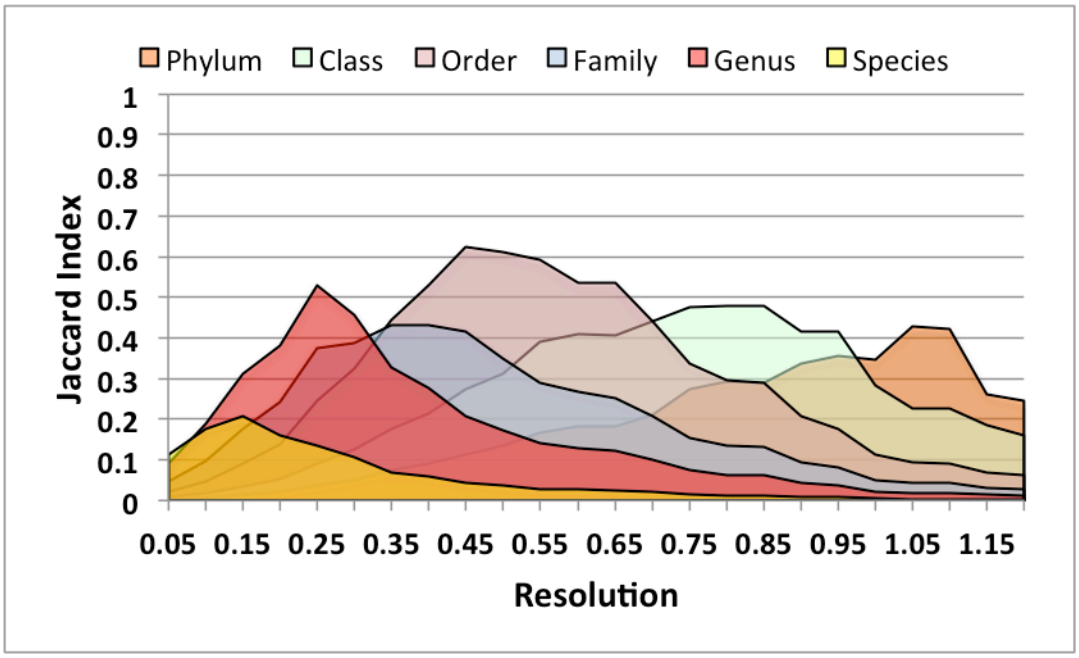


Figure 2-7 Organism pairs assigned to the same fusion module seldom overlap with pairs assigned to the same NCBI Taxonomy bin. With a given Louvain resolution and NCBI Taxonomy level, the Jaccard index is calculated as the number of organism pairs assigned to both the same fusion module and the same NCBI Taxonomy bin, divided by the number of organism pairs assigned to either the same fusion module or the same NCBI Taxonomy bin. Low Jaccard Index values highlight the inconsistency of functional microorganism abilities with the current taxonomic assignments.

While one may see the resolution parameter as cut-off equivalent, it is in fact quite different. In setting cut-offs on organism similarity we consistently group organisms within the same hierarchy – two organisms of the same species always belong to the same genus and the same phylum. On the other hand, tuning the stability of modules is a dynamic assignment. Thus, two organisms in a low-resolution module can belong to different modules at medium resolution and the same module again at high resolution. Note that this implementation of Louvain algorithm is not deterministic; that is two organisms (at the "edge" of similarity) can be sorted into different modules with two runs of the algorithm at the same resolution setting. Correspondence of partitions (estimated by *e.g.*

(Wommack, Bhavsar et al. 2008)) produced at the same resolution setting can thus be used to approximate meaningful partition points for growing *fusion* networks as new organisms are added. This option is not available for similarity cut-off-based schemes that are easily skewed by the availability of genomic data, which, for now, is heavily biased toward organisms of particular human interest (Figure 3B). Though *fusion* is also affected by genome availability, the effect is alleviated by all-to-all connectivity, which reduces the importance of node number in favor of edge weight for clustering purposes.

We detected the Louvain communities in the complete *fusion* network (no edges removed) using a set of resolution values. We compared organism pair assignments to the same Louvain community *vs*. the same NCBI taxonomic placement using the Jaccard index (species to phylum; resolution 0.05 to 1.2; Table 4; Figure 7). Here this metric (ranged [0,1], from no similarity and to identity, respectively) evaluates the percentage of organism pairs that is simultaneously assigned to the same module and the same taxonomic clade, of all same module or same clade assignments (Methods). For example, at the 0.8 resolution of *fusion* (Figure 8A; colors indicate modules) there are nine modules detected. The NCBI taxon (class for *Proteobacteria* and phylum for all others) of organisms in these modules varies (Figure 8B). Some modules demonstrate a highly homogeneous phylum/class distribution, while others are diverse. The Jaccard index of this resolution is 0.478 with NCBI class assignment and 0.294 for phylum assignment. This observation highlights the inconsistency of functional microorganism abilities with the current taxonomic assignments.

Table 2-4 Similarity of the NCBI Taxonomy assignments and fusion modules.

	Modularity index	Number of *fusion* Modules	Number of NCBI clades	Jaccard index
Phylum	1.1	3	27	0.423
Class	0.8	9	43	0.416
Order	0.5	56	97	0.611
Family	0.4	99	204	0.433
Genus	0.3	170	493	0.458
Species	0.1	551	875	0.177

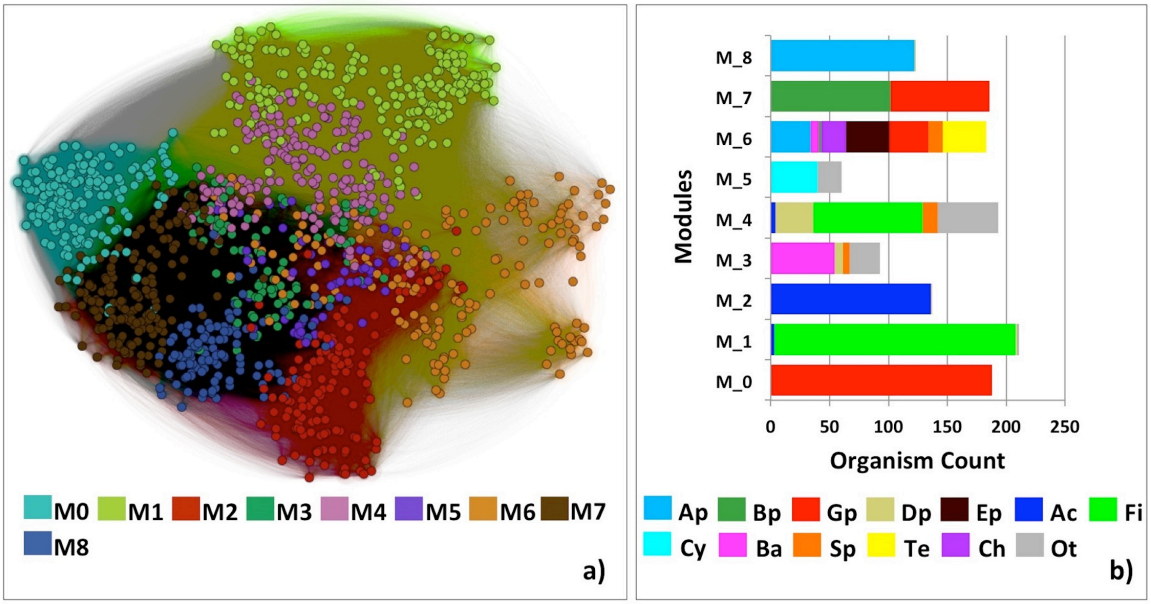


Figure 2-8 Fusion module detection reveals natural organism grouping. (a) Colors represent each of the nine fusion modules detected at resolution 0.8. **(b)** Organism diversity (NCBI Taxonomy) in each module is shown as: Ap-Alphaproteobacteria; Bp-Betaproteobacteria; Gp-Gammaproteobacteria; Dp-Deltaproteobacteria; Ep-Epsilonproteobacteria; Ac-Actinobacteria; Fi-Firmicutes; Cy-Cyanobacteria; Ba-Bacteroidetes; Sp-Spirochaetes; Te-Tenericutes; Ch-Chlamydiae; Ot-other minor phyla. The difference in diversity among the different modules reflects the inconsistencies of the current taxonomy.

We suggest that our novel network-based classification scheme reveals the natural grouping of organisms instead relying on arbitrary similarity cut-offs. Unlike classification based on *pairwise* organism similarity, *fusion* is more robust in handling microorganism diversity. It also alleviates the data availability (organism bias) problem and is a more practical fit for large-scale computational

analysis. In addition, without the limitation of preset discrete taxonomic bins, users can zoom in/out with different resolutions to find out the functional organism groups of their specific interest.

***Fusion*+ reveals functional basis of classification discrepancy.** To study the functional basis of taxonomic vs. functional discrepancies, we built, for several cases, a variant of the *fusion* network, *fusion*+. Our case studies were *Mycoplasma* and *Cyanobacteria* – organisms with well-known taxonomy assignment issues (Garrity GM 2001). *Fusion*+ has two types of nodes: organisms and functions that they perform. Organism nodes are connected by edges to their function nodes. Thus, while in *fusion* one edge connects each organism pair, in *fusion*+ the number of connecting edges is equal to the number of shared functions. Thus, *fusion* modules can be studied in depth in terms of specific functions or organism meta-data variables, *e.g.* salinity, temperature, or pH preferences.

***Mycoplasma* studies.** We created three *fusion*+ networks for 29 *Mycoplasma* strains, including (1) only their 1,848 *Kn* functions (Figure 9A), (2) 1,848 *Kn* and 1,347 *Hy* functions (3,195 total, Figure 9B), and (3) all 9,354 functions (Figure 9C). The shift of the *M. suis* and *M. haemofelis* Langford 1 away from other *Mycoplasma* between *Kn*-only (Figure 9A) and *Kn,Hy*-network (Figure 9B) illustrates the importance for classification of the yet unstudied (*Hy*) functions. Note that while adding the 1,518 *Un* (95% organism-unique) functions further increases the separation between all organisms in the network (Figure 9C) this effect can be largely attributed to the impact of repertoire size.

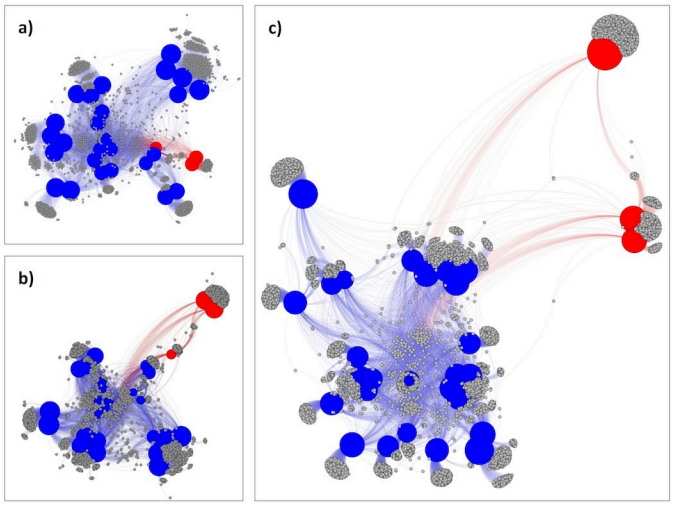


Figure 2-9 Mycoplasma fusion+ reveals the importance of Hy and Un functions in taxonomy assignment. The networks include **a)** Kn functions, **b)** Kn and Hy functions and **c)** all functions. Unique blood Mycoplasma organisms are indicated by red nodes, with the rest of Mycoplasma colored in blue. The length of edges represents the relative (not absolute) similarities between organisms. Note the resolution increases as Hy and Un functions added.

The separation of the two *M. suis* strains and *M. haemofelis* from other *Mycoplasma* is not surprising. As noted earlier, in the functional similarity network they form isolated clusters at a very low 10% cut-off (Figure 5A; Table 3). Previously known as *Eperythrozoon suis* and *Haemobartonella felis*, respectively, these three strains moved to the *Mycoplasma* genus on the basis of their 16S rRNA phylogeny (Neimark, Johansson et al. 2002, Krieg NR 2011). There are, however, ample biological differences of these strains as compared to other *Mycoplasma* (Uilenberg, Thiaucourt et al. 2004). Quantifying these differences is, however, very difficult – do they merit re-assignment to another clade or not? Our observations highlight the problem: these organism are assigned into a genus with less than 10% of common functionality – even organisms of different phyla are often more similar (Figure 5A). The structure of the *fusion* network, however, clearly groups them with other *Mycoplasma* all the way down to a resolution of 0.1. While the similarity of *fusion* modules and species assignments is fairly low

(Table 4), in this particular case the two metrics agree. Rooted in the same ancestor as other *Mycoplasma*, *M. suis* and *M. haemofelis* have evolved specific functional differences likely due to their unique epierythrocytic parasitic life styles (Neimark, Johansson et al. 2001). However, in the currently available microbial functional landscape, even these (very dramatic) in-clade differences do not make this set of organisms functionally different enough to merit complete clade dispersal. This example demonstrates the subjective (albeit successful, in this case) nature of current cladistic assignments when evolutionary relatedness does not equal functional similarity.

We further identified 26 (25 *Kn* and one *Hy*) functional groups shared between *M. suis* and *M. haemofelis* but not by other *Mycoplasma* (Table 5). Representative sequences from two of these groups are detected in a variety of other organisms from multiple phyla. The rest are exclusive to *M. suis* and *M. haemofelis*. Note that other organisms carry out the biochemical functions represented by these functional groups, but they do so using sufficiently different proteins from the ones specific to these *Mycoplasma* strains. These differences may include different protein stabilities, different rates of reaction, *etc*. For instance, many of these 25 *Kn* function groups are house-keeping; *e.g.* DNA polymerase subunits that are unlike others in our set, indicate a likely ancient split from other *Mycoplasma*.

Table 2-5 Blood Mycoplasma functional groups different from other Mycoplasma.

Cluster	#Seq	#Bacteria	Annotation
C_1	128	122	DNA-cytosine methyltransferase (EC 2.1.1.37)
C_2	82	77	Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1)

C_3	4	3	DNA polymerase III subunits gamma and tau (EC 2.7.7.7)
C_4	3	3	DNA polymerase III alpha subunit (EC 2.7.7.7)
C_5	3	3	Glucose-6-phosphate isomerase (EC 5.3.1.9)
C_6	3	3	Cardiolipin synthetase (EC 2.7.8.-)
C_7	3	3	RecA protein
C_8	3	3	SSU ribosomal protein S7p (S5e)
C_9	3	3	Inosine-5'-monophosphate dehydrogenase (EC 1.1.1.205)
C_10	3	3	Thioredoxin reductase (EC 1.8.1.9)
C_11	3	3	FIG006542: Phosphoesterase
C_12	3	3	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase (EC 2.1.1.61)
C_13	3	3	Phospholipid-lipopolysaccharide ABC transporter
C_14	3	3	Endonuclease IV (EC 3.1.21.2)
C_15	3	3	LSU ribosomal protein L13p (L13Ae)
C_16	3	3	Ferrichrome transport system permease protein FhuG
C_17	3	3	Transcription termination protein NusA
C_18	3	3	Preprotein translocase secY subunit (TC 3.A.5.1.1)
C_19	3	3	SSU ribosomal protein S5p (S2e)
C_20	3	3	SSU ribosomal protein S19p (S15e)
C_21	3	3	6-phosphofructokinase (EC 2.7.1.11)
C_22	3	3	Tryptophanyl-tRNA synthetase (EC 6.1.1.2)
C_23	3	3	Zn-dependent hydrolase (EC 3.-.-.-)
C_24	3	3	hypothetical protein
C_25	3	3	Adenylosuccinate lyase (EC 4.3.2.2)
C_26	3	3	ABC TRANSPORTER PERMEASE PROTEIN

One difference between *M suis* and *M. haemofelis* is their preferred hosts, swine and feline, respectively. The species differ from each other by 1,686 functions – 640 in *M. suis* (88% unique; remaining 79 genes shared with other *Mycoplasma*) and 1,046 in *M. haemofelis* (98% unique). This finding is in line with the fact that many hemotrophic *Mycoplasma* contain numerous paralogous gene families, which are thought to participate in antigenetic variation (Guimaraes, Santos et al.

2014). These functions are less annotated, but likely differentiate these organisms in ways necessary to evade specific host immune response.

***Cyanobacteria* studies.** We explored the *fusion*+ network of 40 *Cyanobacteria* (49,937 functions: 17,275 *Kn*, 21,465 *Hy*, 11,197 *Un*; 34,678 organism unique). Based on the 15,259 functions shared by at least two organisms, the *Cyanobacteria* separate into two clusters (Figure 10). In *fusion* this split is observed at resolution 0.3 – a genus equivalent. One cluster (Figure 10, top) contains 16 fresh-water *Cyanobacteria*, three symbionts (Peters 1991, Swingley, Chen et al. 2008, Thompson, Foster et al. 2012), two marine-water organisms and one isolated from marine mud. Note that the mud dweller, *Synechococcus* PCC 7002, is salt tolerant, but does not require salt for growth (Rippka, Deruelles et al. 1979). Another cluster (Figure 10, bottom) contains only marine *Cyanobacteria*. The *Synechococcus* genus members are found in both clusters with marine *Synechococcus* sharing more functionality with the marine *Prochlorococcus* than with the fresh water *Synechococcus*. The intra-genus diversity of *Synechococcus* (Rippka, Deruelles et al. 1979) suggests a division into five genera-equivalent subgroups (Garrity GM 2001). *Fusion*+ reveals that the fresh water and marine *Synechococcus* are significantly functionally different and should belong to different taxa, an unsurprising finding that is in line with both 16S rRNA-based phylogenetic (Schirrmeister, Anisimova et al. 2011) and phylogenomic (Criscuolo and Gribaldo 2011) studies. Bergey's Manual relies heavily on morphology for *Cyanobacteria* classification. However, for this specific example using phylogeny would produce more informative taxonomic

assignments. In other cases, phylogeny may be misleading. For example, according to evolutionary ancestry fresh-water *Synechococcus elongatus* strains should group together with the marine *Synechococcus* and *Prochlorococcus* (Criscuolo and Gribaldo 2011, Schirrmeister, Anisimova et al. 2011). However, *S. elongatus* is more functionally similar to fresh water *Synechococcus* (Figure 10) and should be grouped with them despite its evolutionary relationships to the marine subgroup.

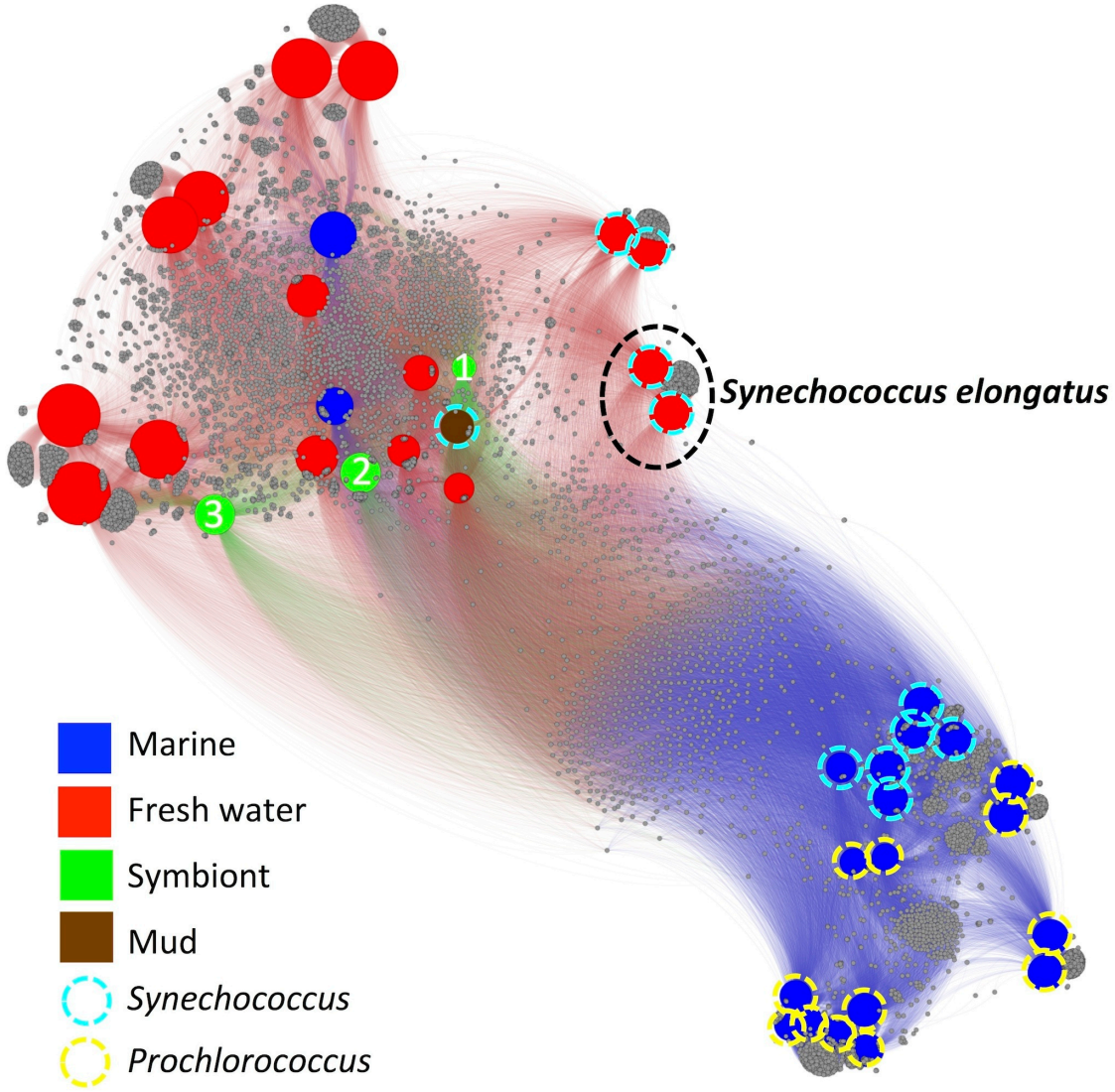


Figure 2-10 Fusion+ of 40 Cyanobacteria reveals environment impact on functions. The Cyanobacteria form one mostly fresh water cluster and one marine cluster. The members of Synechococcus exist in both clusters. The functions that are shared between marine Synechococcus and Prochlorococcus, yet not found in fresh water Cyanobacteria, are likely important in the marine environment. Symbiont1-cyanobacterium UCYN-A; Symbiont2-Acaryochloris marina MBIC11017; Symbiont3- Nostoc azollae 0708.

To further study salt tolerance, we identified 181 functional groups only shared by the marine *Synechococcus* and *Prochlorococcus* in our network. Of these, 15 groups include proteins from organisms of various phyla; *e.g.* one of these functions is present in *Allochromatium vinosum*, a halotolerant microbe surviving in both marine and freshwater environments (Weissgerber, Zigann et al. 2011).

This particular function is RAST annotated as a putative carboxysome peptide A, crucial in carbon fixation. We hypothesize that this *A. vinosum* version of the carboxysome subunit is either specific to salt adaptation or transferred together with other salt tolerance genes in an HGT event. We also identified 166 functions (including 21 *Hy* and one *Un* function) exclusive to and ubiquitous in the marine *Synechococcus* and *Prochlorococcus*. Of these, 34 were unique – not found in any other organisms (including the closest evolutionary neighbor, *S. elongatus)* in any other form (manual curation).

Functional similarity can standardize organism classification. *Fusion* offers a quantitative, objective, and consistent function-based measure of organism similarity. Its classifications correlate with the current taxonomy for many organisms, but not in cases where close phylogenetic relatives are functionally different. Our analysis supports previously reported trends of inconsistencies in the current taxonomy (Staley 2006, Goris, Konstantinidis et al. 2007). *fusion*'s functional repertoire definitions are more accurate for organism classification than sequence identity-based whole-genome comparisons. Moreover, our novel *network* scheme with module identification, to the best of our knowledge, is the first attempt to highlight naturally occurring clusters of organisms, without (arbitrary) pairwise similarity cut-offs. It is more robust than pairwise organism comparison in dealing with organism diversity, particularly since much of *fusion*'s resolution comes from using unstudied (or poorly studied) functions. Potentially, its use of functional similarities to identify organisms can facilitate organismal and functional diversity annotation of metagenomes and, under some circumstances,

even contamination detection in newly sequenced genomes. *fusion* reveals the significant roles that environmental factors play in determining functional abilities of organisms and highlights the key functions shared by different organisms in the same environment.

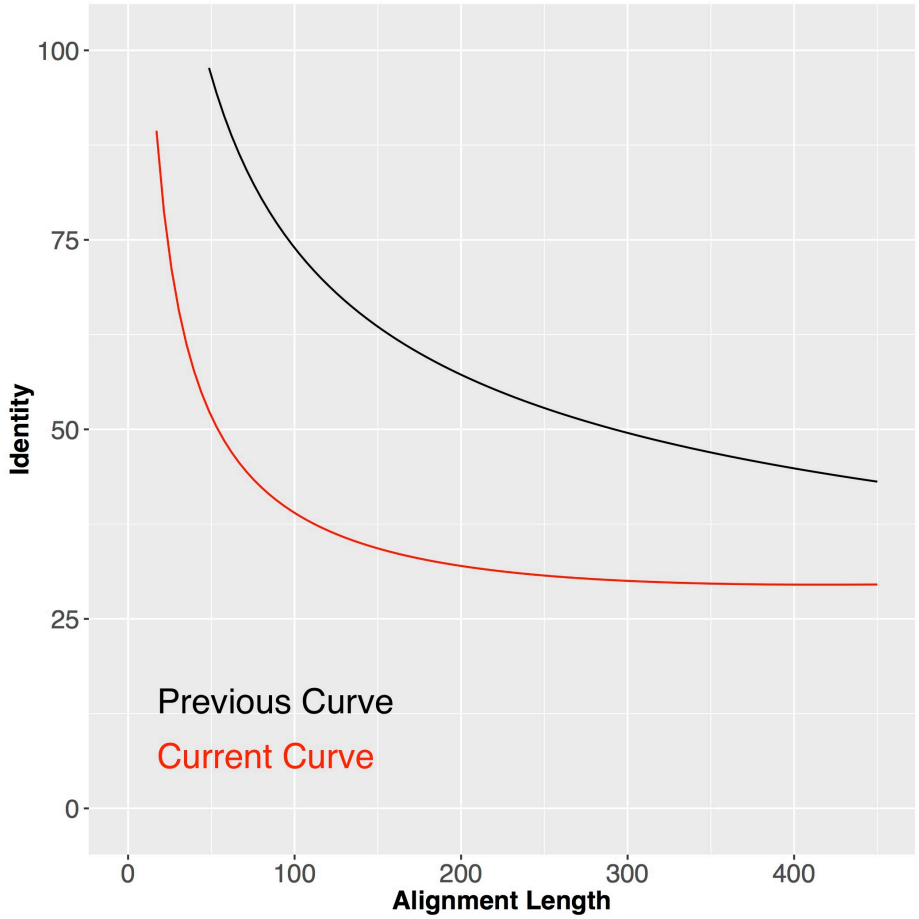
For large-scale analyses and practical applications requiring systematic organismal phenotype assessments, *e.g.* antibiotics development, bioremediation, and industrial uses, classification based on functional comparisons may carry more meaning than evolutionary relationships. *Fusion* is a novel framework for organism classification that (1) directly uses organism functional comparisons, eliminating the need to consider individual HGT events in addition to evolutionary lineage, (2) describes organismal diversity by identifying natural organism clusters in a similarity network instead of arbitrarily establishing cut-offs in levels of similarity per cluster, and (3) has an unlimited capacity to incorporate additional genetic data from plasmids and/or previously unseen organisms. At the very least, *fusion* offers a complementary view to the current taxonomy. Comparing the two classification schemes allows detection of functionally diversified strains – an ability that, potentially, has a wide range of applications, *e.g.* tracking and surveillance of bacterial pathogens.

Conclusion

Microorganism classification, like many other scientific strategies, is driven by expertise and available technology. Historically designed with more emphasis on the former the current taxonomy lacks consistency across assignments. Recent advances in sequencing abilities have created the possibility of exploiting entire organism functional pools for classification. Here we demonstrate *fusion* – a classification technique that compares molecular (genome encoded) functionality across microorganisms. *Fusion* can be used with a predictable consistency to classify newly sequenced organisms according to the current taxonomy. More importantly, it offers a novel and practical prokaryotic classification scheme, which is reflective of, but not dependent on, organism evolutionary history. *Fusion*'s ability to highlight functions key to particular environments will have great impact in industrial and clinical practices.

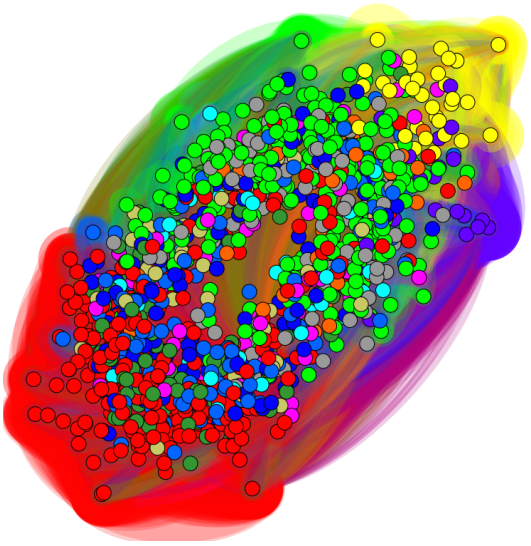
Change Note

In the next chapter I will describe a database built based upon the analysis presented here. However, we used a looser cutoff to define the function groups (See the picture below). With the current curve (red), there are more same-function alignments than previous (black), which ends up in function groups with larger sizes (See Table 2 in the next chapter).

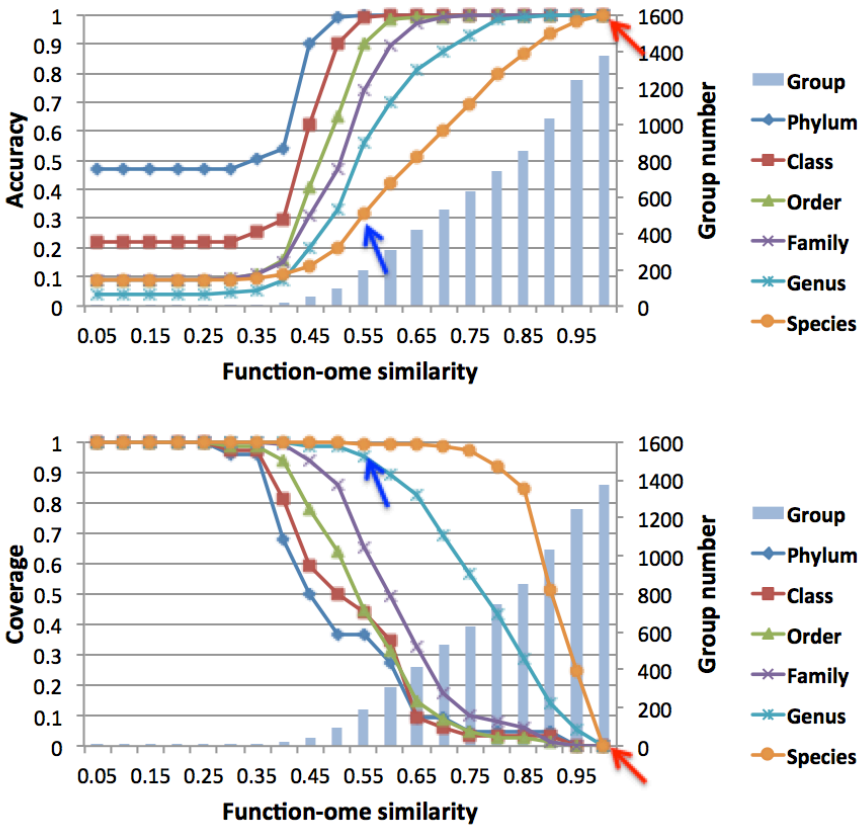


Note that this modification doesn't change the main conclusions we made above. For comparison, below I listed some key figures generated from the new calculation.

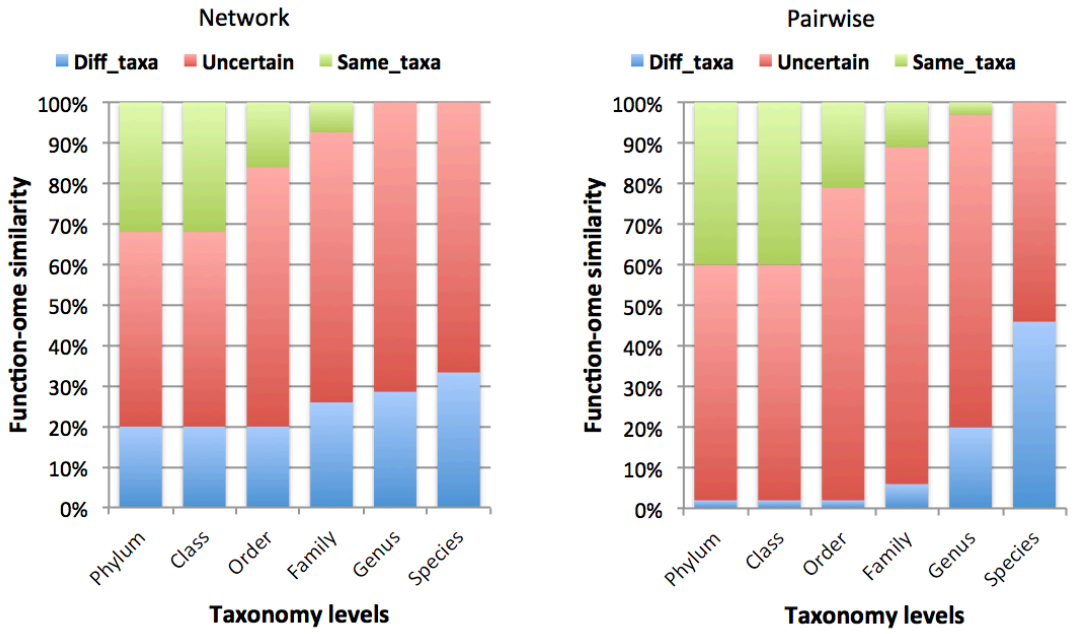
Correspond to Figure 4a



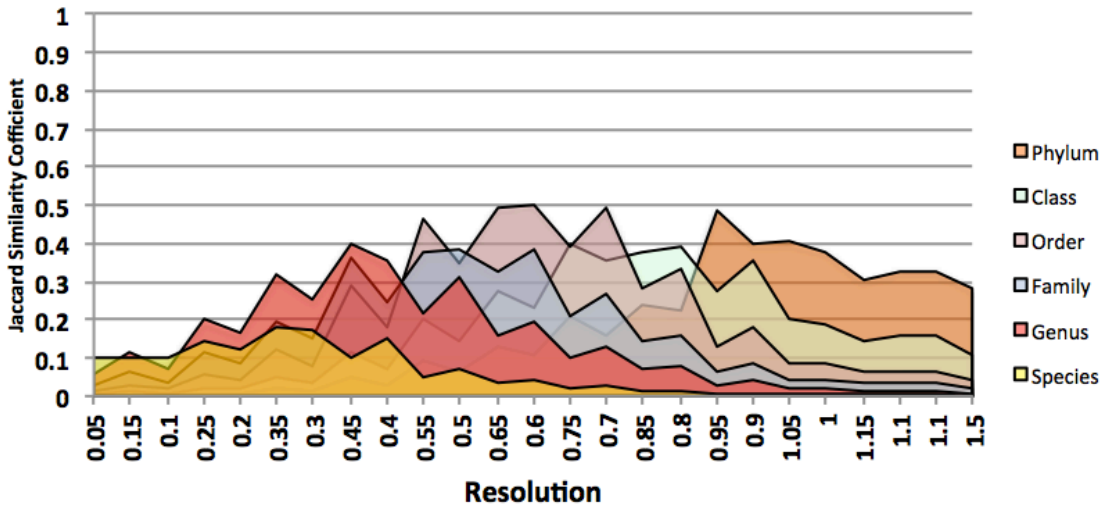
Correspond to Figure 4b and 4c



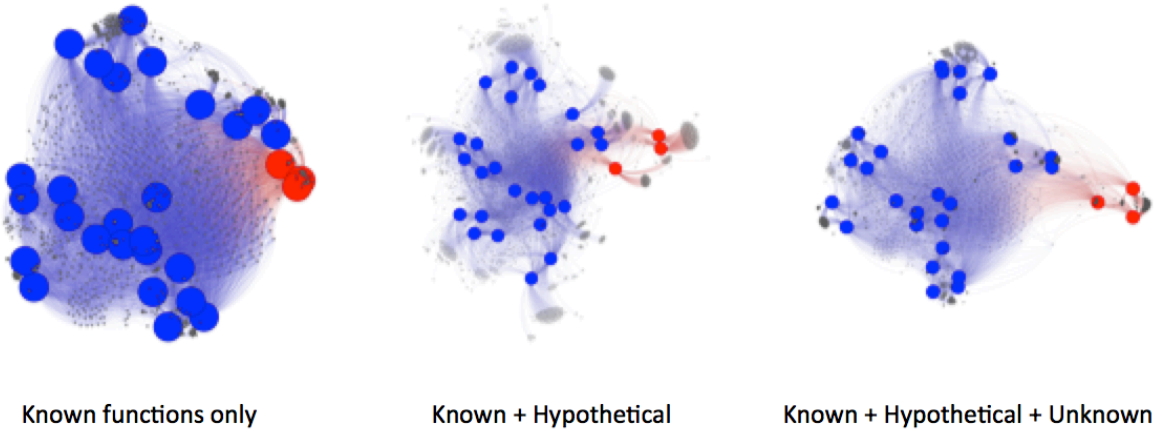
Correspond to Figure 6



Correspond to Figure 7

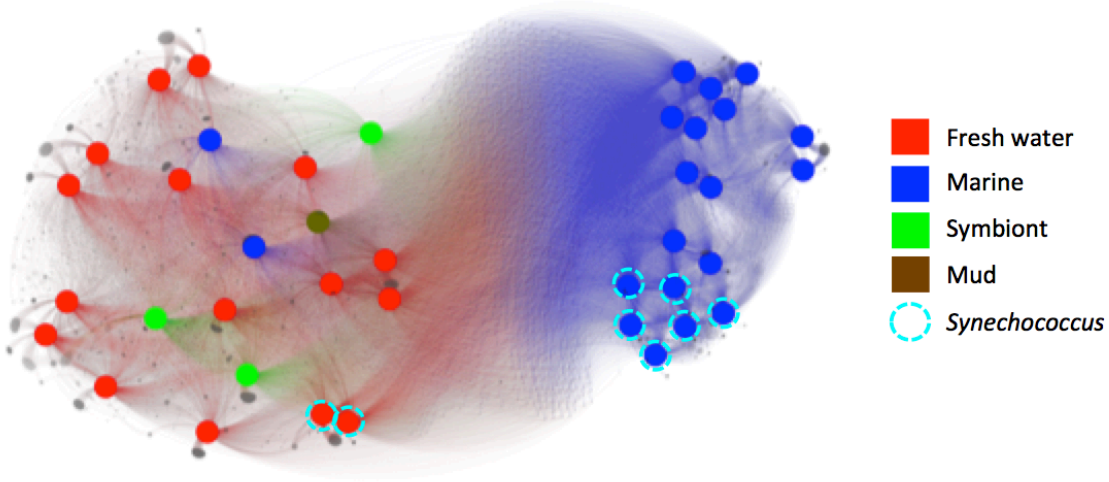


Correspond to Figure 9



Known functions only | Known + Hypothetical | Known + Hypothetical + Unknown

Correspond to Figure 10



Chapter 3

Introduction

Microorganisms are capable of carrying out much of molecular functionality relevant to a range of human interests, including health, industrial production, and bioremediation. Experimental study of these microbes to optimize their uses is expensive and time-consuming; *e.g.* as many as three hundred biochemical/physiological tests only reflect 5-20% of the bacterial functional potential (Garrity GM 2001). The recent drastic increase in the number of sequenced microbial genomes has facilitated access to microbial molecular functionality from the gene/protein sequence side, via databases like Pfam (Benson, Karsch-Mizrachi et al. 2009), COG (Tatusov, Fedorova et al. 2003), TIGRfam (Haft, Selengut et al. 2003), RAST (Aziz, Bartels et al. 2008) and others. Note that the relatively low number of available experimental functional annotations limits the power of these databases in recognizing microbial proteins that provide novel functionality. Additional information about microbial environmental preferences can be found, *e.g.* in GOLD (Pagani, Liolios et al. 2012). While it is well known that environmental factors play an important role in microbial functionality (Cohan 2001), none of the existing resources directly link environmental data to microbial function.

We mapped bacterial proteins to molecular functions and studied the functional relationships between bacteria in the light of their chosen habitats. We previously developed *fusion* (Sun, Yu et al. 2015), an organism functional similarity network, which can be used to broadly summarize the environmental

factors driving microbial functional diversification. Here we describe *fusion*DB – a database relating bacterial *fusion* functional repertoires to the corresponding environmental niches. *fusion*DB is explorable via a web-interface by querying for combinations of organism names and environments. Users can also map new organism proteomes to the functional repertoires of the reference organisms in *fusion*DB; including, notably, matching proteins of yet unannotated function across organisms. The submitted organisms are visualized, and can be further explored, interactively as *fusion* networks in the context of selected reference genomes. Additionally, the web interface generates *fusion*+ networks, *i.e.* views that explicitly indicate shared microbial functions.

Our overall analyses of the *fusion*DB data for the first time give quantitative support for the fact that environmental factors driving microbial functional diversification. To demonstrate *fusionDB* functionality, for individual organisms we mapped a recently sequenced genome of a freshwater *Synechococcus* bacterium to *fusion*DB. In line with our previous findings (Sun, Yu et al. 2015), we demonstrate that this microorganism is more functionally related to other fresh water Cyanobacteria than to the marine *Synechococcus*. In a case study on *Bacillus* microbes we use *fusionDB* to track organism-unique functions and illustrate the detection of core-function repertoires that capture traces of environmentally driven horizontal gene transfer (HGT). *fusionDB* is a unique tool that provides an easy way of analysing the, often unannotated, molecular function spectrum of a given microbe. It further places this microbe into a context of other reference organisms and relates the identified microbial function to the

preferred environmental conditions. Our approach allows for detection of microbial functional similarities, often mediated via horizontal gene transfer, that are difficult to recover via phylogenetic analysis. We note that *fusionDB* may also be useful for the analysis of functional potentials encoded in microbiome metagenomes. We expect that *fusionDB* will facilitate the study of environment-specific microbial molecular functionalities, leading to improved understanding of microbial lifestyles and to an increased number of applied bacterial uses.

Methods

Database setup. *fusion*DB is based on alignments of 4,284,540 proteins from 1,374 bacterial genomes (Dec. 2011 NCBI GenBank (Benson, Karsch-Mizrachi et al. 2009). For each bacterium, we store its (1) NCBI taxonomic information (Benson, Karsch-Mizrachi et al. 2009) and, where available, (2) environmental metadata (temperature, oxygen requirements, and habitat; GOLD (Pagani, Liolios et al. 2012). The environments are generalized, *e.g. thermophiles* include hyper-thermophiles. "*No data*" is used to indicate missing annotations (Table 1). The general *fusion* (functional repertoire similarity-based organism network) protocol is described in (Sun, Yu et al. 2015). Briefly, all proteins in our database are aligned against each other using three iterations of PSI-BLAST (Altschul, Gish et al. 1990) and the alignment length and sequence identity are used to compute HSSP (Rost 2002). A network of protein similarities is then clustered using MCL (Dongen 2000) clustering. For fusionDB the original *fusion* algorithm was modified to use less stringent protein functional similarity criteria (with HSSP distance cutoff = 10), which resulted in 457,576 functions (protein clusters; Table 2). Each bacterium was thus mapped to a set of functions, its functional repertoire. Therefore our functional repertoires include all the bacterial functions, regardless of annotation. We are thus able to make function predictions, even the functions that have not been annotated before, for proteins in new bacteria.

Table 3-1 Taxonomic composition of environmentally distinct groups.

		# phylum	# family	# organisms
Temperature	Mesophile	23	166	1083
	Thermophile	16	42	115
	Psychrophile	3	16	33
	*No Data	14	68	143
Oxygen Use	Facultative	10	51	207
	Aerobe	14	115	481
	Anaerobe	20	71	245
	*No Data	15	88	232
Habitat**	Soil	8 (11)	43 (75)	78 (279)
	Host	11 (15)	59 (94)	329 (706)
	Marine	10 (15)	24 (49)	61 (116)
	Fresh water	10 (18)	37 (84)	69 (271)
	*No Data	15	88	206

*No Data indicates missing annotations.
** One organism can be annotated with multiple habitats (*e.g.* both soil and host). The first number includes only organisms with one annotation, whereas the number in parenthesis includes organisms with multiple habitats.

Table 3-2 Annotation status of HSSP-based function groups.

	Function groups (>1 sequence)	Function groups (1 sequence)	Total
Known (Kn)	54,522	15,738	70,260
Hypothetical (Hy)	85,252	89,282	174,534
Unknown (Un)	22,802	189,980	212,782
Total	162,576	295,000	457,576

Web interface. *fusion*DB web interface has two functions: *explore* and *map new organisms*. The *explore* section contains access to all the 1,374 bacteria and their metadata. Users can search these with (combinations of) organism names and environmental preferences by using text box input or built in filters. User-selected organism set is then used to create a *fusion* network, in which organism nodes are connected by functional similarity edges. The *fusion* network can be viewed in an interactive display, as well as downloaded as network data files or static images. The user-defined color labels of the organism nodes reflect

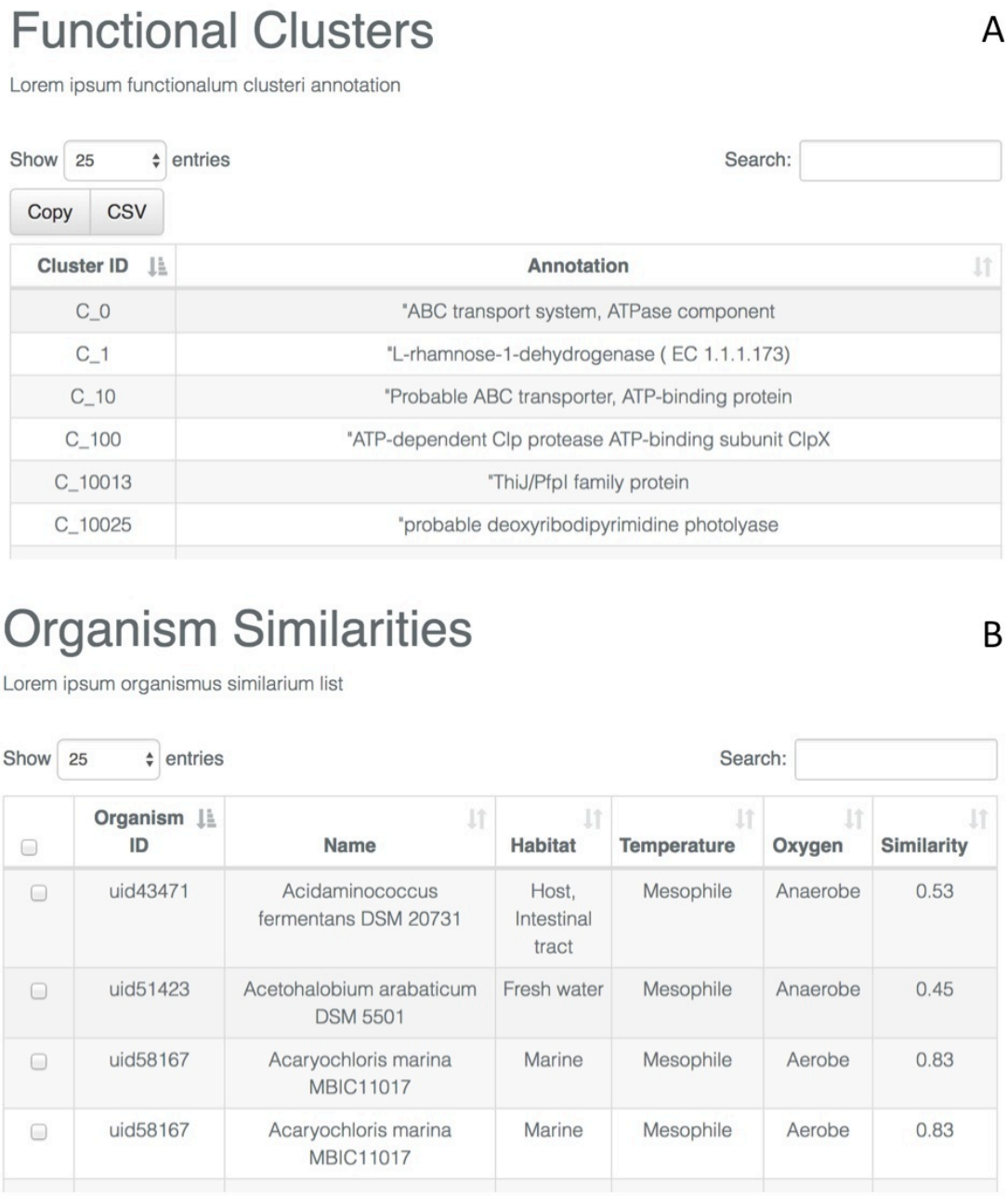
microbial taxonomy or environment. In the interactive display clicking an organism node reveals its taxonomic information and environmental preferences, while clicking an edge between two organisms yields a list of their shared functions. A *fusion*+ network can further be generated from the same list of organisms. There are two types of vertices (nodes) in *fusion*+: organism nodes and function nodes. Organism nodes are connected to each other only through the function nodes they share. The number of edges (degree) of an organism node represents the total number of functions of the organism; the relative position of each organism node is determined by the pull *towards* other organisms via the common functions and *away* from others via unique functions (Sun, Yu et al. 2015). Like *fusion*, *fusion*+ can be interactively displayed, downloaded, and colored by the users' choices. For both network types, users can further retrieve the functions shared by the selected organisms - the core-functional repertoire of the set. Note that the function annotation is from myRAST (Aziz, Bartels et al. 2008). This feature is an efficient tool for investigating functions underlying organism diversification, particularly within different environment conditions.

In the *map* section, users can submit their own new organism proteomes (in fasta format) to our server. The submitted proteins are PSI-BLASTed against *fusion*DB and assigned to stored functions using the HSSP distance cutoff = 10. Note that novel proteins that can't be assigned to existing functional groups are reported as functional singletons. Additionally, protein alignments that exceed 12 CPU hours of run-time are eliminated from future consideration. In testing, we found that no

more than 0.1% of the proteins fall into this category. Although long run-times usually indicate that query proteins likely align to many others in our database, they contribute only a small fraction to the overall bacterial similarity and are eliminated for the sake of a faster result turn-around. The server sends out emails to users when mapping is finished. The *map* result page contains two tables: one is the list of functions of the submitted bacterium, while the other contains pairwise functional similarities (Eqn. 1) between the submitted bacterium and the reference proteomes in *fusion*DB (Figure 1).

$$\text{similarity} = \frac{\text{shared functions}}{\text{the larger functional repertoire size}} \qquad \textbf{Eqn. 1}$$

Both tables can be easily sorted, searched and exported as comma-separated files. The submitted proteome is further mapped to user-selected reference organisms with *fusion* and/or *fusion*+ as describe above.



Functional Clusters

A

Lorem ipsum functionalum clusteri annotation

Show 25 entries

Search:

Copy CSV

Cluster ID	Annotation
C_0	"ABC transport system, ATPase component
C_1	"L-rhamnose-1-dehydrogenase (EC 1.1.1.173)
C_10	"Probable ABC transporter, ATP-binding protein
C_100	"ATP-dependent Clp protease ATP-binding subunit ClpX
C_10013	"ThiJ/PfpI family protein
C_10025	"probable deoxyribodipyrimidine photolyase

Organism Similarities

B

Lorem ipsum organismus similarium list

Show 25 entries

Search:

Organism ID	Name	Habitat	Temperature	Oxygen	Similarity
uid43471	Acidaminococcus fermentans DSM 20731	Host, Intestinal tract	Mesophile	Anaerobe	0.53
uid51423	Acetohalobium arabaticum DSM 5501	Fresh water	Mesophile	Anaerobe	0.45
uid58167	Acaryochloris marina MBIC11017	Marine	Mesophile	Aerobe	0.83
uid58167	Acaryochloris marina MBIC11017	Marine	Mesophile	Aerobe	0.83

Figure 3-1 Example of fusionDB map result page. (a) The functional clusters table contains all the fusionDB functions that the submitted genome mapping to. The search box allows search for certain functions with partial match. **(b)** The organism similarities table contains all the 1374 bacteria in fusionDB, with their ID, name, metadata and functional similarity to the submitted genome. Users can easily sort the table by desired column. The search box allows search in multiple column with space-separated search terms, e.g., “Acidaminococcus Host Mesophile”.

Analysis of environment-driven organism similarity. For each environmental condition in *fusionDB*, we sampled organism pairs where organisms were from (1) the same condition (SC, *e.g.* both mesophiles) and (2) different conditions (DC, *e.g.* thermophile *vs*. mesophile). To alleviate the effects of data bias, the organisms in one pair were always selected from different taxonomic groups (different Families). The smallest available set of pairs, SC-psychrophile contained 33 organisms from 17 Families (Table 1; 136 pairs – 48 same phylum, 88 different phyla; due to high functional diversity of *Proteobacteria,* its classes were considered independent phyla). For all other environment factors we sampled, 100 bootstrap times, 136 organism pairs for both SC and DC sets, covering this same minimum taxonomic diversity. We calculated the pairwise functional similarity (Eqn. 1) distributions and discarded organism pairs with less than 5% similarity.

Phylogenetic analysis. Genes homologous to pyruvate, phosphate dikinase (PPDK) were extracted from selected organisms via the best hit from BLASTP at E value cut-off of 1e-3. We performed multiple sequence alignment and reconstructed Neighbor-joining tree with the online version of Mafft (Katoh, Misawa et al. 2002). The phylogenetic tree was later visualized via FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

Results and Discussion

Map new Synechococcus genomes to fusionDB. We downloaded the full genome of *Synechococcus sp.* PCC 7502 (GCA_000317085.1) as translated protein sequence fasta (.faa file) from the NCBI Genbank (Benson, Karsch-Mizrachi et al. 2009) and submitted it to our web interface. This 3,318 protein fresh water Cyanobacteria is isolated from a Sphagnum (peat moss) bog (Pagani, Liolios et al. 2012). 2,889 (87%) of the bacterial proteins mapped to 2,206 *fusion*DB functions and 426 (13%) were functional singletons; three proteins exceeded runtime and were excluded, Methods). The whole process from submission to receiving a results notification e-mail took a little under three and a half hours. The mapping indicates that *Synechococcus sp.* PCC 7502 is functionally most similar (56%) to *Synechocystis* PCC 6803, a fresh water organism closely related to *Synechococcus*. It also shares a high functional similarity with a mud *Synechococcus* (*S.sp.* PCC 7002; 53%) and with other fresh water *Synechococcus* (*S.elongatus* PCC 7942 and *S.elongatus* PCC 6301; 52%). Notably, but not surprisingly, *Synechococcus sp.* PCC 7502 shares much less functional similarity (40-42%) with the marine *Synechococcus* bacteria. This relationship is clearly demonstrated by the *fusion*+ networks (Figure 2). There are 874 functions shared by all the twelve *Synechococcus*, the core-function repertoire for this genus, and 1,128 functions shared among the fresh water *Synechococcus*. These additional 254 functions are likely important for surviving in the fresh water, as opposed to the marine, environment, *e.g.* low salinity and low osmotic pressure. See http://bromberglab.org/node/32 for these functions.

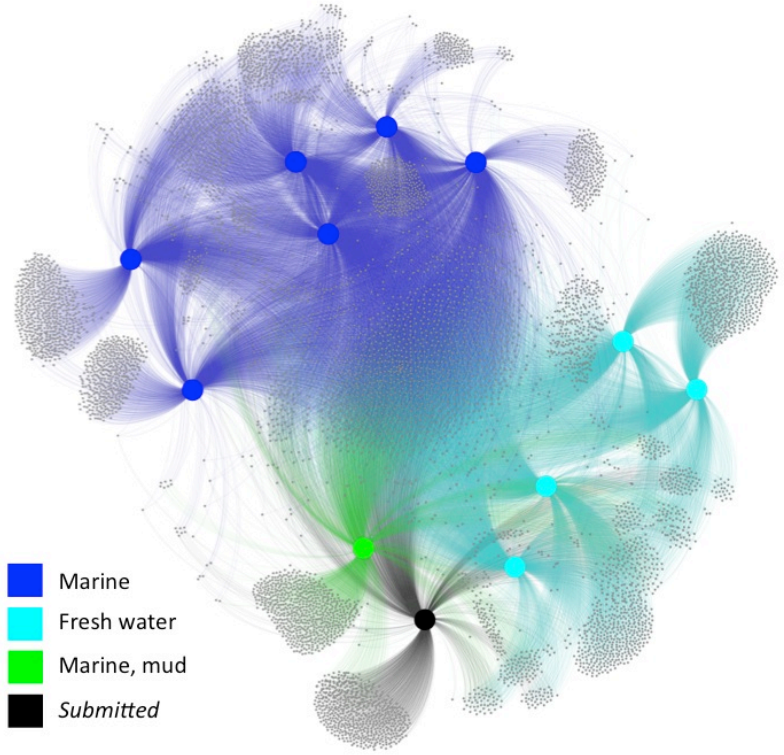


Figure 3-2 The fusion+ view of all Synechoccocus genomes. The submitted Synechococcus sp. PCC 7502 (black) cluster with the fresh water Synechococcus organisms (light blue). Note that the Synechococcus sp. PCC 7002 (green), which is isolated from marine mud, is salt tolerant but does not require salt for growth (see (Zhu, Delmont et al. 2015)).

Environment significantly affects microbial function. Not surprisingly, the SC-thermophile and SC-psychrophile pairs demonstrate significantly higher similarities comparing to all DC pairs (Figure 3A). Notably, the higher functional similarity between thermophiles than between psychrophiles suggests that protein functional adaptation to low temperature is less drastic than to high temperature – an interesting finding itself. Contrast to the extremophiles, mesophile organisms seem to have huge functional diversity as the SC-mesophile similarities are comparable to those the DC pairs (Figure 3A).

Different molecular pathways of aerobic-respiration and anaerobic-respiration/fermentation explain the highest dissimilarity between the aerobes and anaerobes (DC-anaerobe-aerobe; Figure 3B). Interestingly, the SC-anaerobe similarities are higher than the SC-aerobe similarities, probably because the more ancient anaerobic-respiration/fermentation machinery is more simple and conserved.

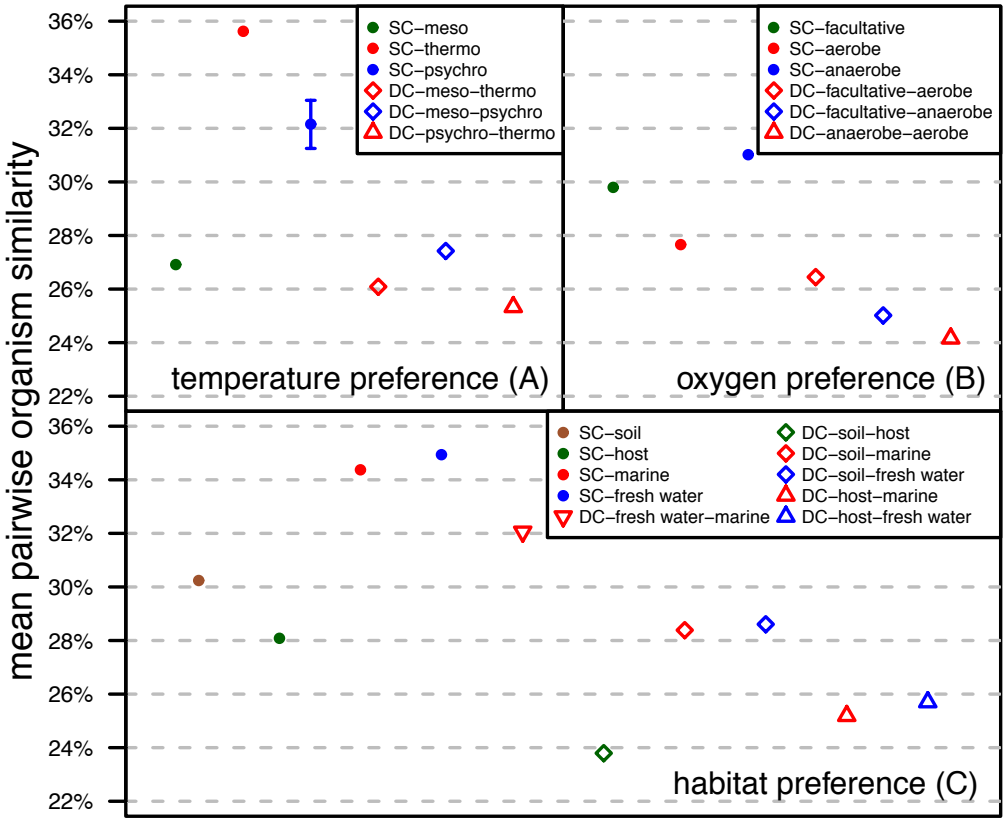


Figure 3-3 Organism pairwise similarity is higher among organisms living in the same environmental conditions. The mean pairwise similarity for same (SC) and different (DC) condition organisms according to **(a)** temperature preference, **(b)** oxygen requirement, and **(c)** habitat. For all points without error bars, the standard errors are vanishingly small.

Habitat-based DC samples show lower pairwise organism similarity than SC samples as well (Figure 3C), except for DC-fresh water-marine, which is not surprising due the same aquatic condition. SC-host displays the lowest mean

organism similarity of the habitat SC samples. We speculate it is the result from the evolutionary pressure to deal with diverse host defence mechanisms (Hornef, Wick et al. 2002). The soil organisms also share low functional similarity, which is likely due to soil's heterogeneity at physical, chemical, and biological levels, from nano- to landscape scale (Bastian, Heymann et al. 2009).

In general, SC organisms across all environmental factors are more functionally similar than DC organisms (Figure 3; with exceptions mentioned above; Kolmogorov-Smirnov test p-val<2.5e-6). In other words, organisms in the same environment are generally more similar than organisms from different environments. This finding is intuitive and many studies have shown HGT within environment-specific microbiomes (Saye, Ogunseitan et al. 1987, Kim, Moon et al. 2012, Liu, Chen et al. 2012). Our results, however, for the first time quantify on a broad scale the environmental impact on microorganism function diversification.

Case study of a temperature driven HGT event. In *fusion*DB *explore*, we extracted thermophilic, mesophilic, and psychrophilic species representatives (one per species) of the *Bacillus* genus from *fusion*DB. We also added two other thermophilic organisms, *D. carboxydivorans* CO-1-SRB and *S. acidophilus* TPY*,* to generate a *fusion*+ network (Table 3, Figure 4). The non-*Bacillus* thermophiles were more closely related to the thermophilic *Bacilli*. All five thermophiles exclusively share three functions. One is a likely pyruvate phosphate dikinase (PPDK) that, in extremophiles, works as a primary glycolysis enzyme (Chastain, Failing et al. 2011). Phylogenetic analysis suggests an HGT event between

thermophilic organisms or a differential gene-loss in Bacilli that no longer live under high temperature (Figure 5). The other two shared functions are carried out by proteins translated from mobile genetic elements (MGEs) that mediate the movement of DNA within genomes or between bacteria (Frost, Leplae et al. 2005). Shared closely related MGEs in distant organisms imply HGT (Krupovic, Gonnet et al. 2013). We thus suggest that *fusion*DB offers a fast and easy way to trace functionally-necessary HGT within niche-specific microbial communities.

Table 3-3 Temperature preferences of organisms used in the case study.

Id	Temperature
Bacillus amyloliquefaciens DSM 7	Mesophile
Bacillus anthracis A0248	Mesophile
Bacillus cereus ATCC 14579	Mesophile
Bacillus coagulans 2 6	Thermophile
Bacillus coagulans 36D1	Thermophile
Bacillus licheniformis ATCC 14580	Mesophile
Bacillus megaterium DSM319	Mesophile
Bacillus subtilis 168	Mesophile
Bacillus thuringiensis Al Hakam	Mesophile
Bacillus tusciae DSM 2912	Thermophile
Bacillus weihenstephanensis KBAB4	Psychrotolerant
Desulfotomaculum carboxydivorans CO 1 SRB	Thermophile
Sulfobacillus acidophilus TPY	Thermophile

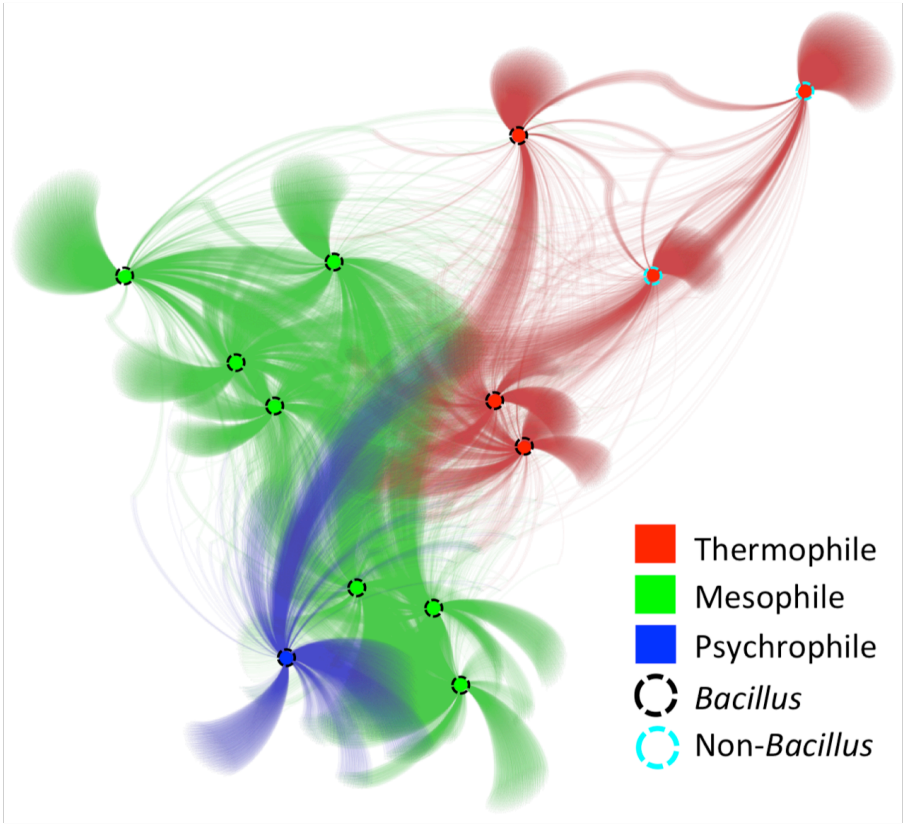


Figure 3-4 fusion+ visualization of Bacillus and thermophilc Clostridia organisms. Large organism nodes are connected via the small function nodes. The two thermophilic Clostridia are connected to the thermophilic Bacilli via functions that are likely to be horizontally transferred.

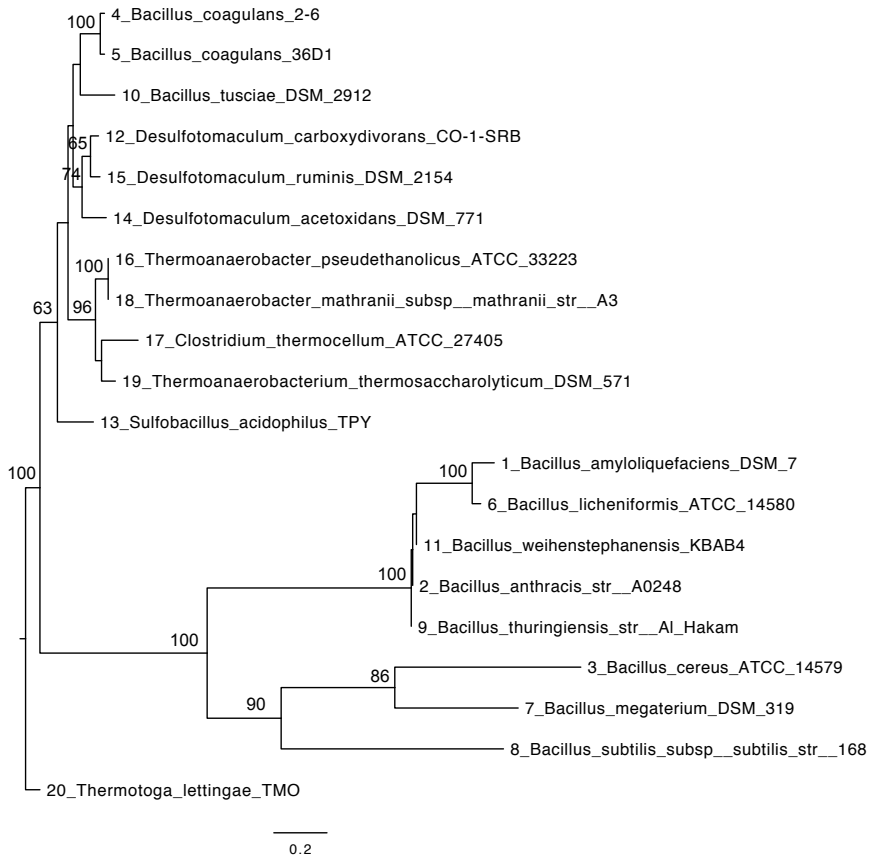


Figure 3-5 Phylogenetic analysis of pyruvate, phosphate dikinase (PPDK) gene suggests horizontal gene transfer between thermophilc Bacilli and Clostridia, or a differential gene loss in non-thermophilic Bacilli. The three thermophilic Bacilli reside within the theromophilc Clostridia clade.

We have highlighted the importance of environmental factors for microbial function, and demonstrated the capability of *fusion*DB to not only annotate functions, but also directly link function to environment. Although it was developed for mapping new microbial genomes, *fusion*DB also has the potential for microbiome annotation. By mapping the proteins translated from metagenomes assembly to *fusion*DB, both the functional and taxonomical can be obtained. We look forward to making fusionDB more useful in this direction.

Conclusion

*fusion*DB links microbial functional similarities and environmental preferences. Our data analysis reveals environmental factors driving microbial functional diversification Mapping new genomes to the reference genomes, it offers a novel, fast, and simple way to detect core-function repertoires, unique functions, as well as traces of HGT. With more microbial genome sequencing and further manual curation of environmental metadata, we expect that *fusion*DB will become an integral part of microbial functional analysis protocols in the near future.

Chapter 4

Introduction

Microorganisms inhabit every available niche of our planet, and our bodies are no exception. Microbes that survive and thrive in the environments at the extremes of temperature, pH, and chemical or radiation contamination possess unique molecular functions of high industrial, clinical, and bioremediation value. Specifically, in the human body, the microbiome critically impacts our health. For example, Crohn's Disease (CD) is a multifactorial disorder resulting from the interplay of individual genetic susceptibility, the gastrointestinal (GI) microbiome and other environmental factors. Taxonomic surveys of the GI microbiome have revealed microbial community features that are unique to CD patients, *e.g.* overall loss of microbial diversity (Manichanh, Rigottier−Gois et al. 2006, Dicksved, Halfvarson et al. 2008), as well as depletion and enrichment of certain bacterial taxa (Frank, St. Amand et al. 2007, Martinez-Medina, Aldeguer et al. 2009, Sokol, Seksik et al. 2009, Frank, Robertson et al. 2011). Establishing whether these observed microbial community shifts contribute to pathogenesis or, instead, correlate with or result from the disease onset, requires understanding not only what are the microbes involved, but also what they do. Earlier studies indicate that in association with CD, the microbiome molecular function potential is more consistently disturbed than taxonomic makeup (Morgan, Tickle et al. 2012). More thorough functional analyses, *e.g.* based on deep metagenomic sequencing, are necessary to elucidate these findings.

Metagenome functional annotation can be performed with or without genome assembly. If the reads can be assembled into large contigs, existing annotation pipelines, such as RAST (Aziz, Bartels et al. 2008) and IMG (Markowitz, Chen et al. 2014), can be applied. However, assembly is difficult and often plagued by a large fraction of unassembled reads or short length contigs, which belong to the minor microbiome members, and by chimeric assemblies, which are especially common for complex and highly diverse samples. Downstream gene finding algorithms are further faced with incomplete and erroneously assembled sequences, complicating statistical model constructions. Read-based annotation, *e.g.*, using a platform such as MG-RAST (Aziz, Bartels et al. 2008), can access molecular functionality of the entire community. However, reads are usually annotated via function transfer by homology that, due to the short read length, is lacking in precision. This inaccuracy is additionally compounded by the erroneous computational annotations of most genes in the reference databases (Schnoes, Brown et al. 2009).

Here, we compiled a gold standard set of reference proteins (GS), with experimentally annotated molecular functions. We further developed *faser* (functional annotation of sequencing reads), an algorithm that uses alignments of translated sequencing reads to full-length proteins to annotate read-"parent protein" molecular functionality. *faser* annotates reads with higher precision at higher resolution, *i.e.* more specific functionality, than PSI-BLAST. In a benchmark test, the functional annotations produced by MG-RAST *vs.* the combination of the *faser* algorithm with the GS database were orthogonal.

Furthermore, when GS was replaced with md5nr, MG-RAST's reference database, *faser* annotated 20% more reads than MG-RAST at a comparable precision level.

Our *mi-faser* pipeline implementation (Figure 1), combining *faser* and GS, is highly parallelized, making use of all available compute cores and processing a (~10GB/70M read) meta-genomic/-transcriptomic file in under half an hour (using 400 compute cores, on average). We applied our *mi-faser* to metagenomic data collected from beach sands in different stages of oil contamination (Rodriguez-R, Overholt et al. 2015). Here, *mi-faser* was able to identify oil degradation functionality that was missed by MG-RAST. We also analyzed the GI tract microbiome data from CD patients and their relatives. We found the microbiome functional profiles were similar between healthy individuals but different across patients and between patients and their healthy relatives. Particularly, our analysis revealed that CD patients' microbiomes were enriched in functions that help bacteria survive inflammation, *i.e.* glutathione metabolism and ribosomal RNA methyltransfer, and in functions that cause inflammation, *i.e.* lipopolysaccharide and acetaldehyde production. These results suggest the microbiome's role in CD-associated pathogenicity.

mi-faser results of the metagenomes analysed in this manuscript are available at http://services.bromberglab.org/mifaser/results/example.

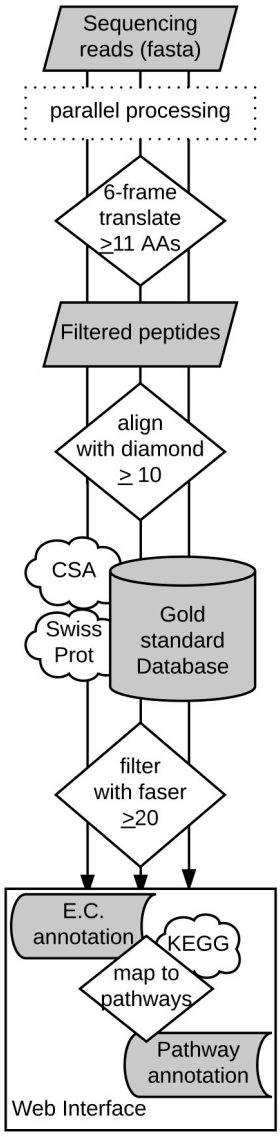


Figure 4-1 mi-faser pipeline. mi-faser is parallelized and runs a load balancer to submit jobs to available [1-2000] compute cores. Under normal functioning conditions (~400 available cores, onaverage), it takes ~30 minutes to process a single (10G/70M read) meta-genome/-transcriptome.

Methods

Datasets. To compile the *PE1-set*, we extracted from SwissProt (Oct. 2015) (Bairoch, Boeckmann et al. 2004) proteins that are 1) bacterial, 2) with evidence of existence, i.e. SwissProt protein evidence is 1, and 3) explicitly assigned an E.C. (Enzyme Commission) number (EC 1992); note that we excluded proteins with incomplete annotations, *e.g.*, 1.1.1.-, as well as those with multiple annotations. From the *PE1-set*, we further extracted proteins whose functions are experimentally verified (Evidence="any experimental assertion*"; EXP-set*). We also identified the overlap between the *PE1-set* and the proteins in the Catalytic Site Atlas database (*CSA-set*) (Furnham, Holliday et al. 2014). We defined our gold-standard dataset (*GS-set*) as the combination of *CSA-set* and *EXP-set*, with 100% identical sequences removed.

For each protein of the *PE1-set* and *GS-set*, we extracted the corresponding gene from ENA (European Nucleotide Archive) (Leinonen, Akhtar et al. 2010) (including 5' UTR and 3' UTR) and randomly generated 10 DNA reads (50-250 nucleotides) that overlap at least one nucleotide of the coding region. We further performed 6-frame translations of the reads and excluded peptides shorter than 11 amino acids. We defined the corresponding peptide collection as *rPE1-set* and *rGS-set*.

We downloaded from MG-RAST the *md5nr* database and defined its proteins as the *md5nr-set*.

We obtained six beach sand metagenomes from a previous study of the Deepwater Horizon oil spill (Rodriguez-R, Overholt et al. 2015). In this study,

metagenomic DNA was sequenced using Illumina MiSeq with paired-end strategy to produce 151bp reads. The samples reside in NCBI (BioProject PRJNA260285), including 1) pre-oil phase samples, OS-S1 (SRX692936) and OS-S2 (SRX695904), 2) oil phase samples, OS-A (SRX696142) and OS-B (SRX696240), and 3) post-oil recovered phase samples, OS-I600 (SRX696250) and OS-I606 (SRX696254).

We additionally obtained 11 human gut (fecal) microbiome samples from a family affected by CD from the PopGen biobank (Schleswig-Holstein, Germany). Of these, nine members were self-reported as healthy and two were affected. Metagenomic data were generated using the Illumina Nextera DNA Library Prep Kit and sequenced 2x125bp on an Illumina HiSeq2500. In total, 424.8 million paired-end reads were generated with a median number of 38.9 million read pairs per sample. Adapter trimming was performed using Trimmomatic (Bolger, Lohse et al. 2014) in paired-end mode, discarding reads shorter than 60 bp. Quality filtering was done using Sickle (Joshi and Fass 2011) run in paired-end mode, with a quality threshold of 20 and a minimum length of 60bp. To remove contaminating host sequences from the dataset, DeconSeq (v0.4.3) (Schmieder and Edwards 2011) was run with the human reference genome (GRCh38) as database. Only read-pairs where both sequences survived quality control were retained. On average 11.76 % of raw reads were discarded, leaving 374.8 million read pairs for downstream analysis.

***faser* curve optimization.** We PSI-BLASTed the *rGS-set* against the *GS-set* (parameters: evalue $1e^{-3}$; inclusion ethresh $1e^{-10}$; num iterations 3;

max_target_seqs 1,000,000), excluding self-hits, *i.e.* peptide hits of their "parent" proteins. For any peptide, functional annotation (E.C. number) was inherited from the "parent" protein; one nucleotide overlap required to transfer annotation. A peptide-protein alignment is considered positive if the functional annotations of the peptide and the aligned protein match exactly at the selected number of E.C. digits, and negative otherwise. Any given alignment can be plotted in an L (alignment length) *vs.* Id (alignment sequence identity) two-dimensional space. Further, an exponential decay curve (as for HSSP calculations, (Rost 2002)) can be used to identify the alignments in this space as true positives (alignments of peptides to proteins of identical function that fall above or on the curve), false positives (different functions above or on the curve), true negatives (different functions below the curve) and false negatives (identical functions below the curve). From these values we calculated precision (positive accuracy; Eqn. 1) and recall (positive coverage; Eqn. 2) for different curve parameters (*a* and *b* in Eqn. 4), optimizing the latter to fit a curve best separating positive from negative alignments in terms of the highest F-measure (Eqn. 3).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad \textbf{(1)}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad \textbf{(2)}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \textbf{(3)}$$

$$b \times L^{-a \times (1 + e^{-\frac{L}{1000}})} \quad \textbf{(4)}$$

To avoid overestimating performance of *faser*, we clustered the *GS-set* with CD-hit at 40% sequence identity and split the clusters into ten subsets. We further optimised *faser* curve parameters in 10-fold cross-validation, *i.e.* we iteratively optimised the curve on nine subsets and tested it on the remaining one, repeating this process ten times for a different subset as the test set. We evaluated the performance reported here by summing the numbers of true and false positives and negatives in each test set. As all ten curves were very similar in parameters, we took the average of these to establish the final *faser* curve.

To summarise, the *faser* curve is meant to predict from a peptide-protein alignment, whether the "parent" protein of the peptide and the aligned protein share the same function (E.C. annotation). Additionally, the distance of the alignment point to the curve along the sequence identity (Id) axis indicates the reliability of the prediction.

Evaluating *faser* using DIAMOND results. We extracted the proteins from the *GS-set* and *md5nr-set* that had identical UniProt IDs. We performed searches against the *md5nr* database using PSI-BLAST (parameters: evalue $1e^{-3}$; inclusion ethresh $1e^{-10}$; num iterations 3; max_target_seqs 1,000,000) and DIAMOND (parameters: min-score 10; k 1,000,000). We further excluded from the results the alignments to subject proteins that were not in the overlap set. We compared the *faser* values calculated from the results of different alignment algorithms by performing a 100-fold bootstrap, sampling ~20% of the results at each iteration. Note that we used the bootstrap approach to assess the consistency of the observed performance differences.

Comparison to MG-RAST. We submitted the artificial metagenome as well as the six sand metagenomes for processing to MG-RAST via its website and downloaded the resulting function annotations via MG-RAST API (Wilke, Bischof et al. 2015). We used the KEGG (Kanehisa, Sato et al. 2016) annotations from the *md5nr* database to establish the annotated E.C.s. Note that although proteins can carry out multiple functions, in this study we, conservatively, only included proteins with unique and complete E.C. annotations; i.e. we excluded proteins with incomplete or multiple E.C. annotations.

We compared different database/algorithm combinations for the annotation of the same sample (SOM Figure 3). The Venn diagrams of the numbers of E.C.s annotated by different such combinations were generated by Venny (Oliveros 2007). When comparing across sand metagenome samples from different phases, sample-specific E.C.s were removed as uninformative (<1% of total E.C.s in both cases). The correlation between samples was calculated with Spearman's rho, ρ, offered in the R package, Hmisc (Frank E Harrell Jr 2016).

Functional analysis of CD metagenomes. NMDS (Non-metric multidimensional scaling) (Kruskal 1964) analysis (Shepard plot in Figure 14), along with the subsequent permanova test was carried out using the Vegan R package (Jari Oksanen 2016). From the distributions of E.C.s in the microbiomes of healthy individuals, we calculated the "confidence range" for each E.C. as Q1 – 3*IQR (three interquartile ranges below the first quartile) to Q3 + 3*IQR (three interquartile ranges above the third quartile). Patient E.C.s that fell outside this range were identified as significantly depleted or enriched, respectively. Pathway

analysis was performed with the KEGG Mapper tool (Kanehisa, Sato et al. 2016). Jaccard Index was calculated as the size of intersection divided by the size of union of the two sample sets.

Results and Discussion

Few proteins have experimentally verified function annotation. Among the 332,193 bacterial proteins in SwissProt (Oct. 2015) (Bairoch, Boeckmann et al. 2004, Boutet, Lieberherr et al. 2016), only 18,240 (~5%) have been experimentally shown to *exist*. Of these, we extracted 5,965 that have unique (one per protein) and explicit (all four digits) Enzyme Commission (E.C.) annotations (*PE1-set*; Methods). From this *PE1-set*, we further selected proteins whose *functions* were experimentally verified, as noted in the Catalytic Site Atlas (*CSA-set*) (Furnham, Holliday et al. 2014) or SwissProt (*EXP-set*) (Bairoch, Boeckmann et al. 2004, Boutet, Lieberherr et al. 2016). After filtering, our set contained 2,848 (2,810 non-redundant at 100% sequence identity; *GS-set*) bacterial proteins of experimentally verified function. Note that this is the cleanest available dataset of functional annotations; *i.e.* functional annotations in public databases are usually based on (many rounds of) function transfer by homology and are, as such, often questionable.

***faser* is more accurate for function transfer by homology than PSI-BLAST.** We created artificial reads from the gene nucleotide sequences corresponding to the proteins in *GS-set* and *PE1-set* (6-frame translated to peptides, *rGS-set* and *rPE1-set*, Methods). We further PSI-BLASTed (Altschul, Gish et al. 1990) the *rGS-set* against itself*,* excluding self-hits, to determine the equation of the curve (Eqn. 1) separating the correct alignments (same function) from the incorrect ones (different functions) in the *L* (alignment length) *vs. Id* (sequence identity) space. Our approach was modeled after the HSSP metric for function transfer

between full-length proteins (Schneider, de Daruvar et al. 1997, Rost 2002). We optimised the curve parameters to maximise the F measure (Methods), representative of best separation of peptide-protein alignments of the same function (E.C. annotation) from those of different functions (Methods). Thus, if a given alignment is above the curve, the "parent protein" of the peptide and the aligned reference protein are predicted to share function. The *faser* score (the distance from the curve along the *Id* axis) indicates the reliability of such predictions. This measure clearly outperforms PSI-BLAST e-value in annotating function (AUC of 0.78 vs. 0.62, respectively; recall calculated with the background of all PSI-BLAST results at e-value $=10^{-3}$; Figure 2). For example, at recall levels of ~50%, the *faser* score (= 20) is nearly 90% accurate, which is >30% more than e-value ($=10^{-18}$; Figure 2). E-value reaches ~90% precision at cut-offs $<10^{-36}$, which corresponds to recall of less than 7% (Figure 2).

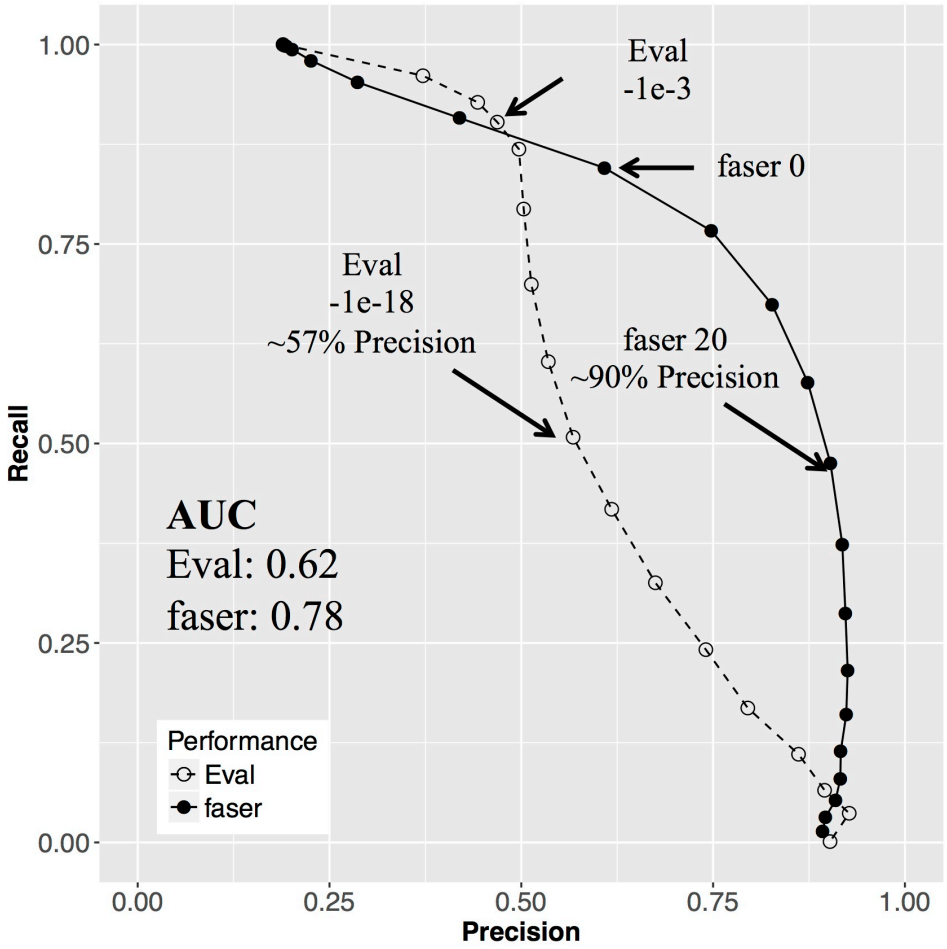


Figure 4-2 faser outperforms PSI-BLAST in annotating read functions. At most cutoffs, faser (filled circles) is more precise than BLAST (empty circles). For example, for nearly half the reads, it provides as much as 90% annotation accuracy as compared to 57% attained by PSI-BLAST (arrows at faser score=20 and e-value=e-18). At the default cutoff of 0, faser attains similar accuracy as PSI-BLAST at e-value=e-18, but for ~35% more reads.

The number of matching E.C. digits reflects the level of resolution of function annotation; *i.e.* proteins that share only the first three E.C. digits have similar functions with slight differences. For example, both 1.1.1.1 and 1.1.1.2 are alcohol dehydrogenases, but with different electron acceptors: NAD+ and NADP+, respectively. PSI-BLAST exhibits comparable performance to *faser* when matching the first three E.C. digits (Figure 3), but fails to differentiate functions at the fourth digit resolution level, producing a large number of false positives (Figure 2). *faser* resolves the fourth E.C. digit at >90% precision with

>40% recall. At all cut-offs, when compared to PSI-BLAST, *faser* consistently offers as much as ~50% higher recall at same precision level and up to ~25% higher precision at same recall level (Figure 2).

$$faser\ score = \begin{cases} -100, L < 11 \\ Id - 352.3L^{-0.302\times(1+e^{-\frac{L}{1000}})}, L \geq 11 \end{cases} \quad \textbf{(1)}$$

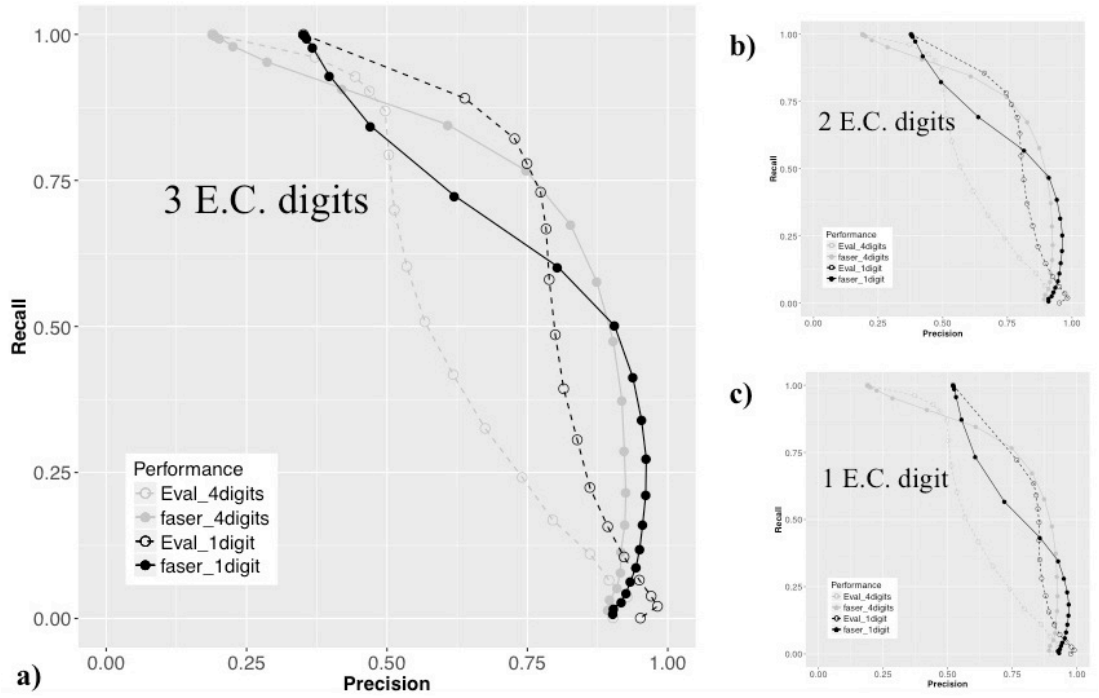


Figure 4-3 PSI-BLAST performance is comparable to faser when the "same function" definition is loose. Matching only the first **a)** three E.C. digits, **b)** two E.C. digits, **c)** one E.C. digit. Note that at high precision (Methods) faser still outperforms PSI-BLAST with significantly higher recall.Figure 3. PSI-BLAST offers overall performance comparable to *faser* when the same function requires match only the first a) three E.C. digits; b) two E.C. digits; c) one E.C. digit. Note that at high precision level, *faser* still outperforms PSI-BLAST with significantly higher recalls.

Note that although *faser* was developed using PSI-BLAST, it can also be calculated via other alignment mechanisms. In a benchmark bootstrap test (Methods), we compared the performance of the two alignment mechanisms in computing *faser* score. DIAMOND runs significantly faster (~30,000 times), yet produces similar *faser* scores when compared to PSI-BLAST (Pearson

correlation coefficient of 0.99±0.001; Methods). At *faser* score =20, DIAMOND missed 5.2±0.5% (287±32 peptide to protein matches) of the 5,490±146 PSI-BLAST-identified matches, but gained an additional 7.6±0.4% (418±21 matches). Interestingly, among the matches identified by PSI-BLAST but missed by DIAMOND, only two thirds (69.0±3%) were correct, *i.e.* matching the right peptide to protein. On the other hand, DIAMOND correctly identified matches that were missed by PSI-BLAST almost all (98±0.8%) the time. Note that these results suggest that *faser* can potentially be used with a range of alignment mechanisms. To alleviate the long alignment runtimes, we exhaustively tested the option and switched to DIAMOND (Buchfink, Xie et al. 2015).

***faser* offers complementary function annotations to MG-RAST**. We compared *faser* performance to that of MG-RAST (Aziz, Bartels et al. 2008), one of the most popular public metagenome annotation platforms. We considered both algorithm and database levels using the: 1) *faser* algorithm with the *GS-set* database (F_G, the *mi-faser* pipeline); 2) *faser* algorithm with the *md5nr* database (Wilke, Harrison et al. 2012) (F_M; *faser*-md5nr), 3) MG-RAST algorithm with *md5nr* database (M_M, the MG-RAST pipeline) (Figure 4; Methods). Note that we could not run the MG-RAST algorithm with the *GS-set* database because the MG-RAST developers advised against it, citing complicated installation.

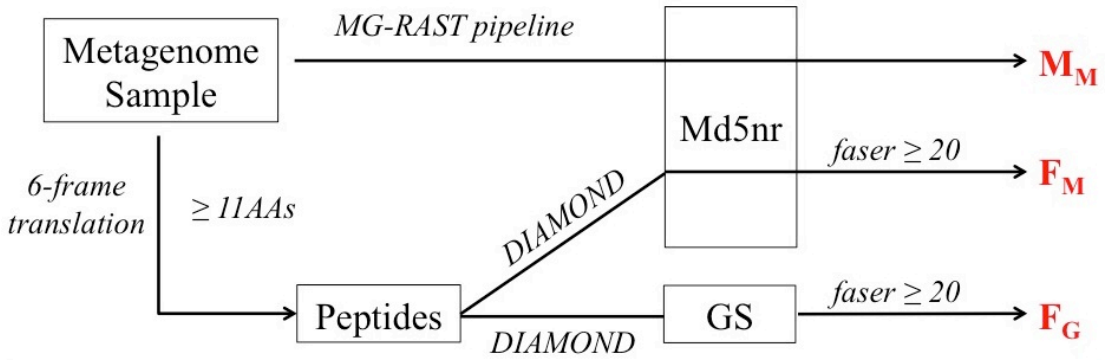


Figure 4-4 Algorithm and database comparisons. Note that the combination of the MG-RAST algorithm with the GS database is missing because the MG-RAST developers advised against local installation of their software.

When the *rPE1-set* is used as the artificial metagenome, the F_G and M_M annotations are significantly different (Table 1), although both pipelines annotate a similar number of reads (Figure 5A). This variation in performance is not biased toward any specific E.C. class (Figure 6). Note that the *rPE1-set* is a superset of *GS-set*, which likely contributes to the improved performance of the F_G pipeline. The differences between F_G and M_M annotations (Figure 5B, first column) stem from the differences between the databases (*GS-set* vs. *md5nr*) and/or algorithms (*faser vs.* MG-RAST). The divergence between F_G and F_M annotations (Figure 5B, second column) indicates that the database differences contribute significantly to the F_G/M_M variation. Note that this difference is not surprising as the *GS-set* and *md5nr* share only 779 E.C.s (62% and 29%, respectively).

Table 4-1 Artificial metagenome (rPE1) annotation by F_G, F_M and M_M.

	F_G	F_M	M_M
Annotated reads	35,119	48,481	30,800
Multi-E.C. reads*	819	11,373	200
Erroneously annotated reads	1,436	5,705	4,237
Correctly annotated reads	33,683	31,103	26,363
Precision	98%	85%	86%

* Reads with multiple E.C. annotations were excluded from the analysis.

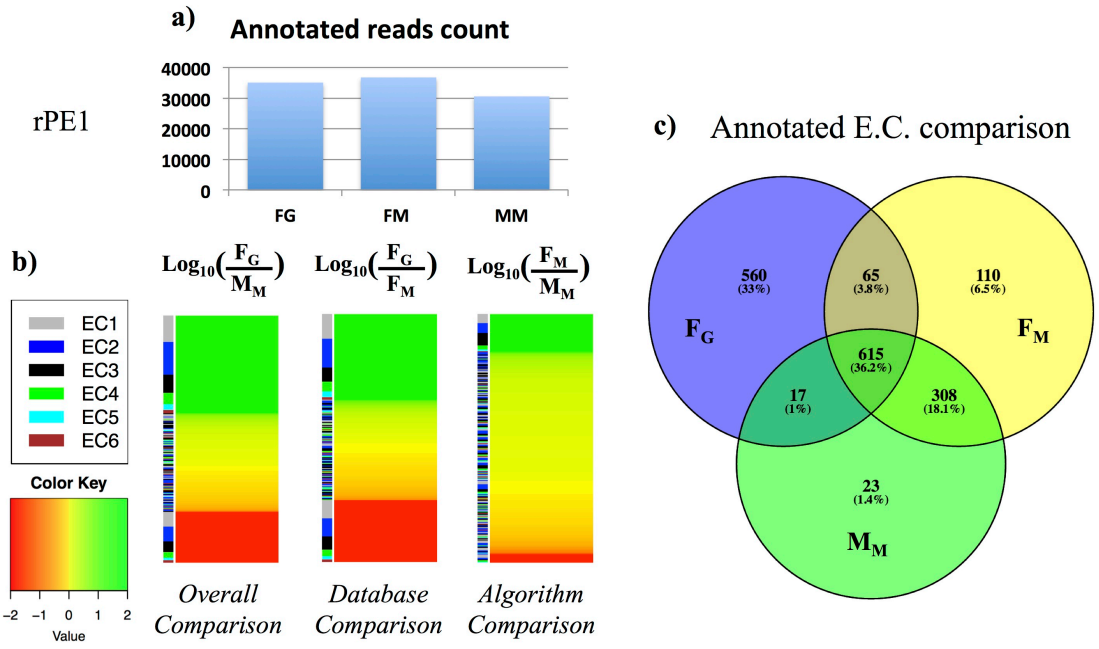


Figure 4-5 The faser algorithm in combination with the GS database annotates the artificial metagenome functions in a manner complementary to MG-RAST. a) The number of reads annotated by each combination of algorithms and databases; **b)** the read abundance by E.C. annotated via each combination of algorithm/database; **c)** the total E.C. count annotated via each combination of algorithm/database.

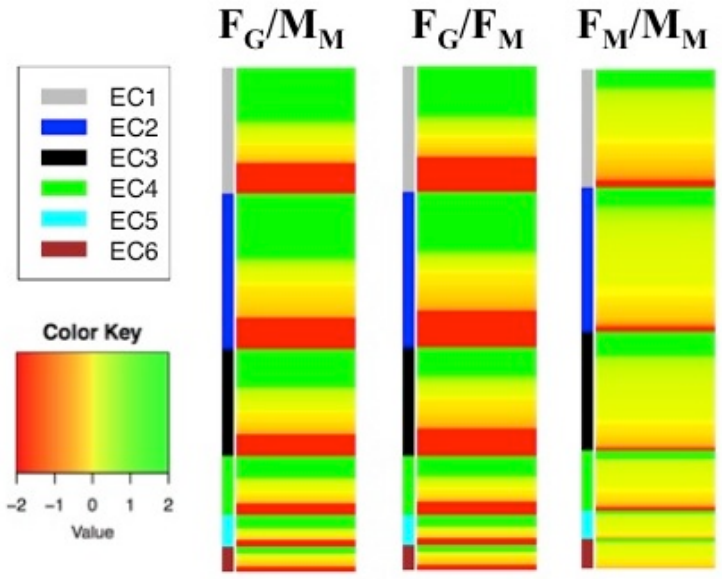


Figure 4-6 The annotation differences are not biased towards specific E.C.s. Note that this is the illustration corresponding to Figure 4-5B with different order of E.C.s.

The comparison between F_M and M_M results is more interesting (Figure 5B, third column), as it highlights the differences between the *faser* and MG-RAST

algorithms. Using the same *md5nr* database, *faser* (F_M) annotated ~20% more reads than MG-RAST (M_M, Figure 5A) with comparable precision (Table 1). Note that the precision reported in these comparisons is affected by the misannotation (~14%), *i.e.* UniProt proteins in both the *GS-set* and *md5nr* annotated with different E.C. numbers – a finding, which is in line with a previous study (Schnoes, Brown et al. 2009). F_M and M_M identified 923 E.C.s in common, while 175 and 40 E.C.s were uniquely identified by *faser* and MG-RAST, respectively (Figure 5C). After exclusion of the database-specific E.C.s, the database impact was reduced (F_G/F_M, Figure 7), yet we still observed substantial F_G/M_M differences largely due to the pipeline algorithms (Figure 7).

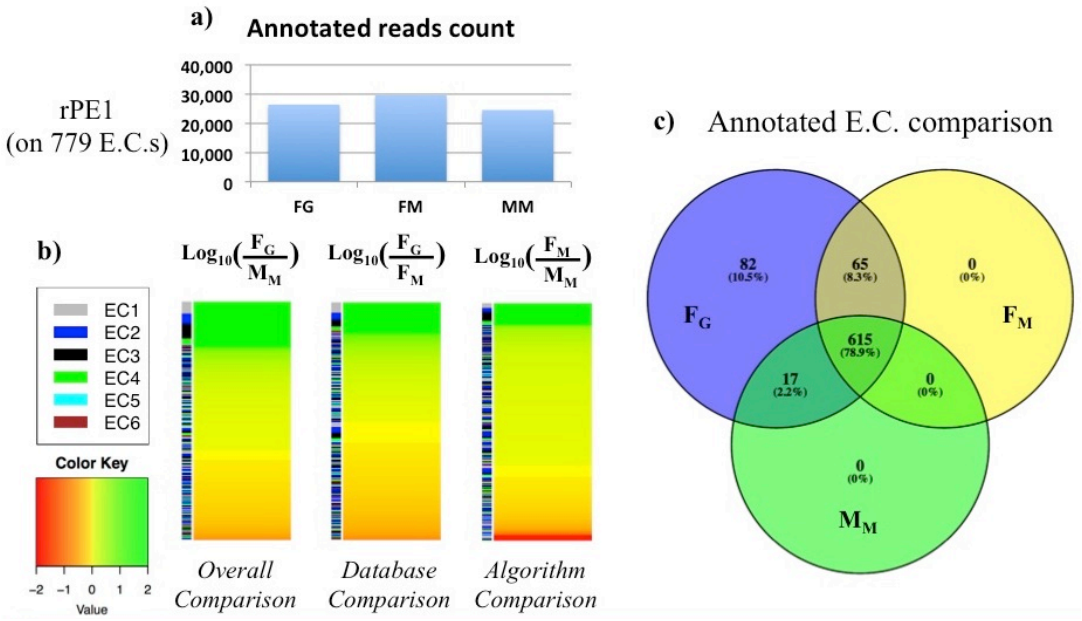


Figure 4-7 The faser algorithm and the GS database (mi-faser) annotate artificial metagenome (rPE1 set) functions better than MG-RAST. This Figure represents only the E.C.s shared between the GS and md5nr database. **a)** the number of reads annotated by each combination of algorithms and databases; **b)** the E.C. abundance annotated differently between combinations of algorithms and databases; **c)** the E.C. count of annotations in various combinations of algorithms and databases.

We further extended the comparison of the annotation methods to six metagenomic samples from the Deepwater Horizon oil spill beach sand study (Rodriguez-R, Overholt et al. 2015) (Methods). Note that in this real-life case, there was no "correct" annotation to use for comparing annotation results. However, it appears that F_M and M_M results are orthogonal. For example, for OS-A (oil phase) F_M annotated >50% more reads than M_M (Figure 8A); moreover, there were 220 E.C.s unique to F_M and 42 E.C.s unique to M_M (Figure 8C). Annotation of other samples followed a similar pattern. Database differences resulted in a significant disparity between the number of reads annotated in each sample by F_G and M_M (*e.g.* Figure 8B). However, both pipelines agreed that: (1) samples taken in the same phase were highly functionally correlated (Table 2 and 3), (2) samples in oil phase were functionally more correlated with samples in recovered phase than pre-oil phase (Table 2 and 3, which may indicate that the environment has fully recovered from the contamination), and (3) ~20% of reads in all samples mapped to housekeeping functions (housekeeping E.C.s complied from (Gil, Silva et al. 2004)). This agreement across methods suggests that F_G reflects true variation in functionality between samples from a perspective complimentary to M_M.

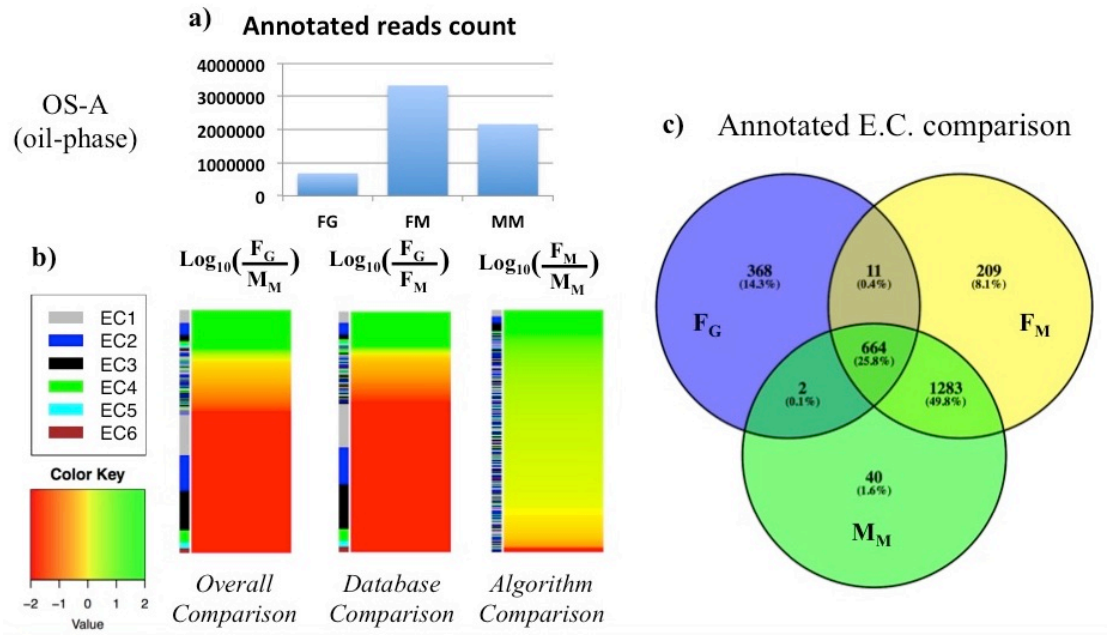


Figure 4-8 The faser algorithm and GS database (mi-faser) annotate BP-oil-spill metagenome functions differently from MG-RAST. a) the number of reads annotated by each combination of algorithms and databases; **b)** the E.C. abundance annotated differently between combinations of algorithms and databases; **c)** the E.C. count of annotations in various combinations of algorithms and databases.

Table 4-2 Spearman correlation between sand samples annotated by F_G.

		Oil		Pre-oil		Recovered	
		OS_A	***OS_B***	***OS_S1***	***OS_S2***	***OS_I600***	***OS_I606***
Oil	***OS_A***	1	-	-	-	-	-
	OS_B	0.98	1	-	-	-	-
Pre-oil	***OS_S1***	0.89	0.89	1	-	-	-
	OS_S2	0.89	0.89	0.98	1	-	-
Recovered	***OS_I600***	0.93	0.95	0.93	0.93	1	-
	OS_I606	0.93	0.94	0.93	0.93	0.99	1

Table 4-3 Spearman correlation between sand samples annotated by M_M.

		Oil		Pre-oil		Recovered	
		OS_A	***OS_B***	***OS_S1***	***OS_S2***	***OS_I600***	***OS_I606***
Oil	***OS_A***	1	-	-	-	-	-
	OS_B	0.98	1	-	-	-	-
Pre-oil	***OS_S1***	0.92	0.93	1	-	-	-
	OS_S2	0.93	0.93	0.96	1	-	-
Recovered	***OS_I600***	0.94	0.95	0.94	0.94	1	-
	OS_I606	0.94	0.95	0.94	0.94	0.99	1

We further searched for functions enriched in oil phase metagenomes as compared to either pre-oil or recovered phases. F_G returned 909 E.C.s (65%, 588 E.C.s, are *GS-set* specific), while M_M returned 1,627 E.C.s (65%, 1062 E.C.s, are *md5nr* specific). Even for the E.C.s existing in both databases, F_G and M_M revealed considerable discrepancies in across-phase abundance fold-changes; ρ=0.46 (Spearman's rho) for oil-to-recovered phase and only ρ=0.09 for oil-to-pre-oil phase (Figure 9). We explored E.C.s annotated as highly enriched (≥ 5 times) in oil-phase as compared to other phases by F_G, yet unchanged or even decreased by M_M. There are nine of these E.C.s in oil-to-pre-oil comparison and ten in oil-to-recovered comparison, with three E.C.s overlapping across comparisons; *i.e.* enriched in the oil phase as compared to either pre-oil or recovered phases (Table 4 and 5). Of the three overlapping E.C.s, two are particularly notable: 1.3.11.1 (catechol 1,2-dioxygenase) directly associates with BTEX (Benzene, Toluene, Ethylbenzene and Xylenes) degradation, while 1.8.99.1 (assimilatory sulfite reductase) is essential for sulfur reducing bacteria, known to degrade BTEX.

Table 4-4 F_G-unique functions annotated as enriched in Oil phase compared to Pre-oil phase (annotated as unchanged or decreased by M_M).

EC	Fold increase	Anno
1.1.1.169	6	2-dehydropantoate 2-reductase
1.13.11.1*	8	catechol 1,2-dioxygenase
1.18.1.1	8	rubredoxin---NAD+ reductase
1.8.99.1*	15	assimilatory sulfite reductase
2.1.1.61	7	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
2.7.1.45	6	2-dehydro-3-deoxygluconokinase
3.1.3.18	7	phosphoglycolate phosphatase
3.1.4.1	10	phosphodiesterase I
4.2.1.109*	5	methylthioribulose 1-phosphate dehydratase

*Functions also enriched comparing to Recovered phase.

Table 4-5 F_G-unique functions annotated as enriched in Oil phase compared to Recovered phase (annotated as unchanged or decreased by M_M).

EC	Fold increase	Anno
1.1.1.290	6	4-phosphoerythronate dehydrogenase
1.11.1.1	6	NADH peroxidase
1.13.11.1*	8	catechol 1,2-dioxygenase
1.6.5.2	11	NAD(P)H dehydrogenase (quinone)
1.8.99.1*	15	assimilatory sulfite reductase
2.7.1.107	6	diacylglycerol kinase (ATP)
3.1.1.73	10	feruloyl esterase
3.1.21.2	6	deoxyribonuclease IV
4.2.1.109*	16	methylthioribulose 1-phosphate dehydratase
4.2.1.83	13	4-oxalomesaconate hydratase

*Functions also enriched comparing to Pre-oil phase.

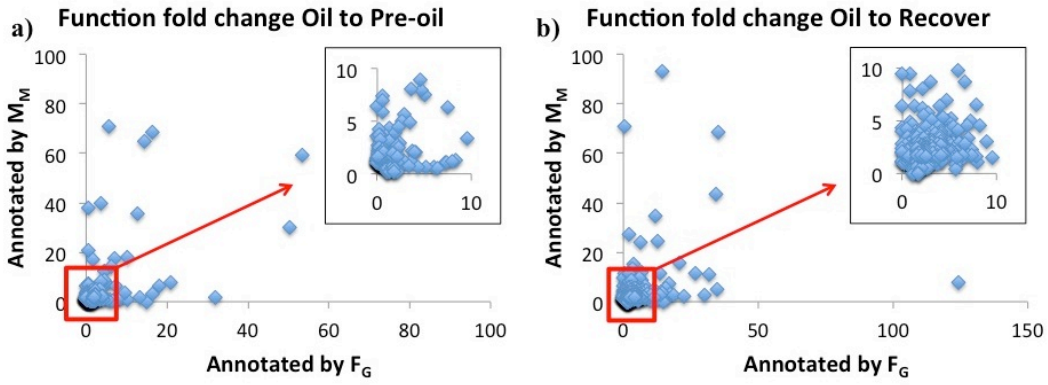


Figure 4-9 F_G and M_M annotations reveal different fold-changes of E.C. functions across phases. In oil phase as compared to a) pre-oil phase (ρ=0.09, Spearman's rho) **and b)** recovered phase (ρ=0.46).

mi-faser reveals microbial functions associated with Crohn's Disease (CD).

We used our *mi-faser* pipeline (Figure 1) to analyse 11 microbiomes from individuals of the same extended family – two CD affected patients and nine first-degree relatives (Figure 10A). The members of this family live in three households that are no more than 32km apart from each other, with the CD affected individuals living in households 17km away. No statistically significant distinction between functional profiles of individuals in the study was observed on the basis of generational or household differences (Figure 10B; p-value =0.48 and =0.51 respectively, permanova test (Anderson 2001)). The nine healthy individuals share highly similar microbiome functional profiles (rho, ρ=0.93±0.03; Figure 10B; Table 6). This finding is in line with previous studies that show that microbiome functional profiles across healthy individuals are more consistently maintained than bacterial species profiles (Morgan, Tickle et al. 2012). On the other hand, the microbiome functional profiles of the two CD patients are not only distinct from those of their healthy relatives (Figure 4B; ρ=0.75±0.11; p-

value=0.013, permanova test), but also between themselves (ρ=0.72; Figure 10B; Table 6). Note that the former holds true even within the same household. In concert, these finding indicates that either there are different microbiome pathogenesis mechanisms of CD or that CD has a diverse impact on microbiome functionality.

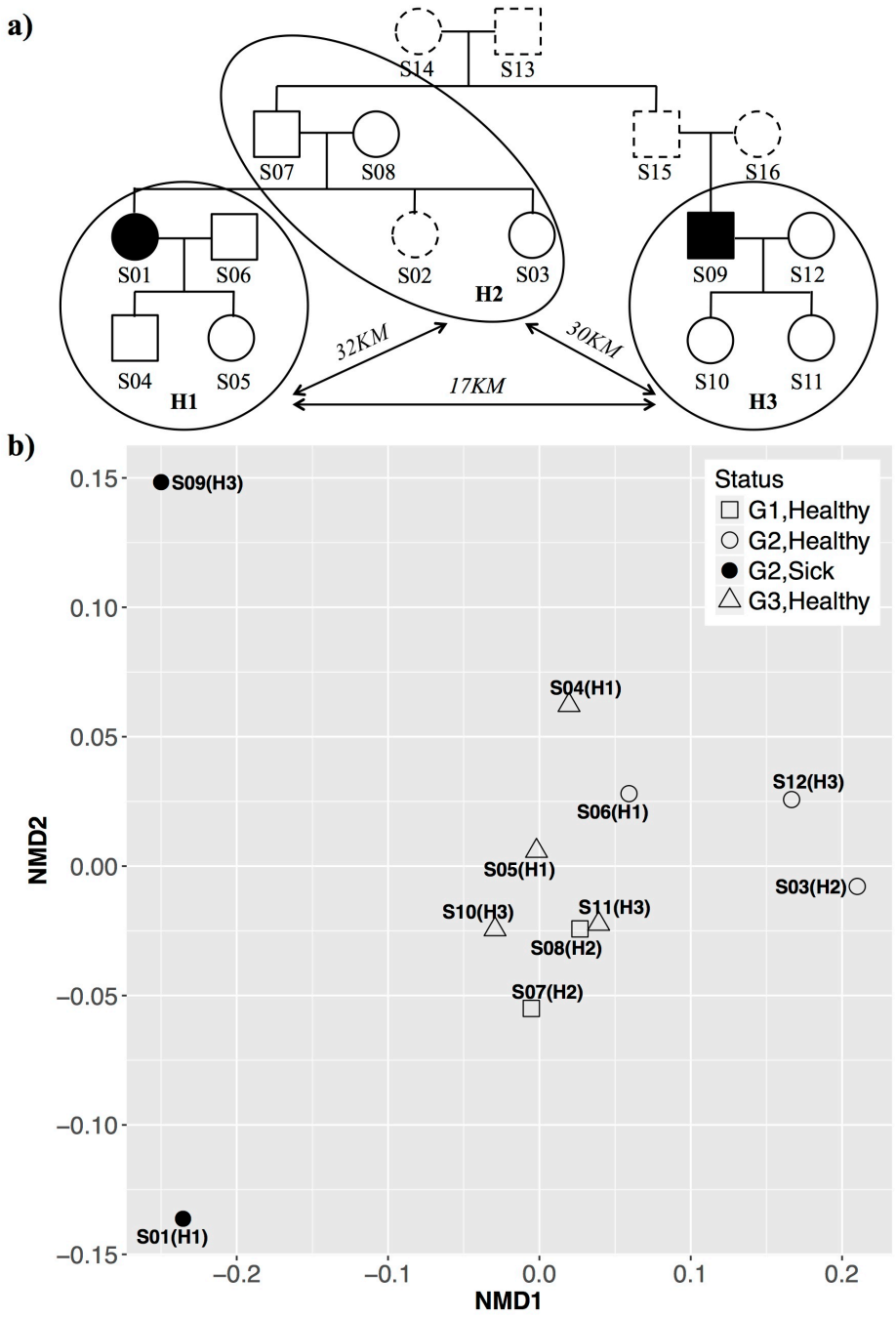


Figure 4-10 Functional capabilities of microbiomes of CD-affected individuals differ from healthy individuals and from each other. a) The pedigree of the family in our study. Filled markers indicate CD affected individuals and empty markers are healthy individuals; dashed outline markers indicate individuals not included in this study. Individuals grouped by circles live in the same household. **b)** The non-metric multidimensional scaling (NMDS) graph represents the distribution of individual microbiome functional profiles. Samples are labeled with identifiers (S1-S11) and household numbers (H1, H2, or H3, in parenthesis). Legend marker numbers (G1 - grandparents, G2 - parents, G3 - children) represent generations, while marker shapes relate generations and CD status. Sick individuals (filled markers) localize separately from each other and from the cluster of healthy individuals (empty markers).

Table 4-6 Spearman correlation between microbiome samples.

	S01*	S03	S04	S05	S06	S07	S08	S09*	S10	S11	S12
S01*	1	-	-	-	-	-	-	-	-	-	-
S03	0.58	1	-	-	-	-	-	-	-	-	-
S04	0.70	0.88	1	-	-	-	-	-	-	-	-
S05	0.77	0.89	0.95	1	-	-	-	-	-	-	-
S06	0.65	0.92	0.93	0.94	1	-	-	-	-	-	-
S07	0.81	0.87	0.91	0.96	0.90	1	-	-	-	-	-
S08	0.74	0.90	0.93	0.97	0.95	0.96	1	-	-	-	-
S09*	***0.72***	0.60	0.69	0.69	0.63	0.67	0.66	1	-	-	-
S10	0.74	0.86	0.91	0.93	0.90	0.91	0.93	0.68	1	-	-
S11	0.74	0.91	0.92	0.96	0.93	0.96	0.96	0.65	0.92	1	-
S12	0.59	0.94	0.91	0.89	0.93	0.86	0.90	0.63	0.87	0.91	1

* Samples from CD patients.

We identified those E.C.s in our microbiomes whose abundance significantly changed in each patient compared to healthy individuals (Methods). S01 and S09 both have a large fraction of such E.C.s (44% and 32% respectively, sum of enriched and depleted, Table 7). For example, ten E.C.s enriched in both S01 and S09 are annotated as rRNA methyltransferases (Table 8), which are known to be essential for microbial response to environmental stresses (Baldridge and Contreras 2014). We further explored these E.C.s to identify pathways uniquely altered in each patient; *e.g.* more than half of Biotin metabolism pathway E.C.s are altered in S01, while Xylene degradation is enriched only in S09 (Figure 11). There are also pathways that are similarly changed in both patients, *i.e.* they are enriched in the same E.C.s; for example, glutathione metabolism and lipopolysaccharide biosynthesis (Figure 11 & 12; Jaccard index =0.5 and =0.73 respectively). Given the distant microbiome functional profiles between S01 and S09 (Figure 10B), these similarities are unlikely to occur by chance. Glutathione is known to help bacteria survive oxidative stress, thus the enriched glutathione

pathway could be a response to inflammation (Masip, Veeravalli et al. 2006); a previous study has reported enrichment in abundance of genes associated with glutathione transportation in CD patients (Morgan, Tickle et al. 2012). However, the latter study (Morgan, Tickle et al. 2012) also suggested a decrease in propanoate and butanoate metabolism, both of which showed overall enrichment in S01 and S09 (Figure 11). Finally, to the best of our knowledge, the role of the lipopolysaccharide (LPS) biosynthesis pathway in CD patient microbiomes has not yet been reported. However, bacterial LPS is previously reported to increase intestinal tight junction permeability in mouse modules (Guo, Al-Sadi et al. 2013). Tight junctions normally form a selective seal between adjacent intestinal epithelial cells. Its increased permeability induces luminal pro-inflammatory molecules, resulting in sustained inflammation and tissue damage (Lee 2015). Additionally, we also observed differences within individual pathway changes between patients. For example in the glycolysis/gluconeogenesis pathway, S01 is depleted in proteins necessary to convert glucose to pyruvate, while the pyruvate metabolism pathways are enriched (Figure 13A). S09 shows a similar pattern, while enriching an alternative route from glyceraldehyde-3P to glycerate-3P (Figure 13B). Interestingly, in both patients, most enriched E.C.s in pyruvate metabolism lead to acetaldehyde production (Figure 13), a metabolite also known to induce tight junction disruption in intestinal epithelial cells (Atkinson and Rao 2001). Thus, our result indicates the microbiome function shift in CD patients contributes to pathogenicity, while helps the bacteria survive host inflammation (Figure 14).

Table 4-7 Number of enriched and depleted E.C.s of the two CD patients.

	Total E.C.	Enriched E.C.				Depleted E.C.			
		pathway	non pathway	sum	%	pathway	non pathway	sum	%
S01	945	246	132	378	40	32	10	42	4
S09	902	190	75	265	29	19	5	24	3

Table 4-8 Significantly altered E.C.s from patient microbiomes that are not assigned to any pathways. Shading indicates E.C.s with decreases abundance.

E.C.	Annotation
1.14.12.17	nitric oxide dioxygenase
1.16.3.2	bacterial non-heme ferritin
1.17.1.2	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase
1.3.1.91	tRNA-dihydrouridine20 synthase [NAD(P)+]
1.8.1.8	protein-disulfide reductase
2.1.1.166*	23S rRNA (uridine2552-2'-O)-methyltransferase
2.1.1.176*	16S rRNA (cytosine967-C5)-methyltransferase
2.1.1.177*	23S rRNA (pseudouridine1915-N3)-methyltransferase
2.1.1.182*	16S rRNA (adenine1518-N6/adenine1519-N6)-dimethyltransferase
2.1.1.186*	23S rRNA (cytidine2498-2'-O)-methyltransferase
2.1.1.189*	23S rRNA (uracil747-C5)-methyltransferase
2.1.1.190*	23S rRNA (uracil1939-C5)-methyltransferase
2.1.1.191*	23S rRNA (cytosine1962-C5)-methyltransferase
2.1.1.242*	16S rRNA (guanine1516-N2)-methyltransferase
2.1.1.266*	23S rRNA (adenine2030-N6)-methyltransferase
2.1.1.61	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
2.10.1.1	molybdopterin molybdotransferase
2.3.1.118	N-hydroxyarylamine O-acetyltransferase
2.4.1.180	lipopolysaccharide N-acetylmannosaminouronosyltransferase
2.4.2.43	lipid IVA 4-amino-4-deoxy-L-arabinosyltransferase
2.7.1.170	anhydro-N-acetylmuramic acid kinase
2.7.7.19	polynucleotide adenylyltransferase
2.7.7.42	[glutamine synthetase] adenylyltransferase
2.7.7.59	[protein-PII] uridylyltransferase
2.7.7.72	CCA tRNA nucleotidyltransferase
2.7.7.75	molybdopterin adenylyltransferase
2.7.8.33	UDP-N-acetylglucosamine---undecaprenyl-phosphate N-acetylglucosaminephosphotransferase
2.8.1.4	tRNA sulfurtransferase
2.8.3.16	formyl-CoA transferase
3.1.1.29	aminoacyl-tRNA hydrolase

3.1.11.1	exodeoxyribonuclease I
3.1.11.2	exodeoxyribonuclease III
3.1.13.1	exoribonuclease II
3.1.13.5	RNase D
3.1.21.7	deoxyribonuclease V
3.1.26.11	tRNase Z
3.1.26.12	RNase E
3.1.3.23	sugar-phosphatase
3.2.1.17	lysozyme
3.2.2.28	double-stranded uracil-DNA glycosylase
3.4.13.9	Xaa-Pro dipeptidase
3.4.21.105	rhomboid protease
3.4.23.36	signal peptidase II
3.5.1.105	chitin disaccharide deacetylase
3.5.1.28	N-acetylmuramoyl-L-alanine amidase
3.6.3.34	iron-chelate-transporting ATPase
4.3.1.15	diaminopropionate ammonia-lyase
5.1.3.32	L-rhamnose mutarotase
5.2.1.8	peptidylprolyl isomerase
5.4.99.12	tRNA pseudouridine38-40 synthase
5.4.99.19	16S rRNA pseudouridine516 synthase
5.4.99.22	23S rRNA pseudouridine2605 synthase
5.4.99.27	tRNA pseudouridine13 synthase
1.11.1.1	NADH peroxidase
3.4.11.4	tripeptide aminopeptidase

* Enriched rRNA methyltransferase

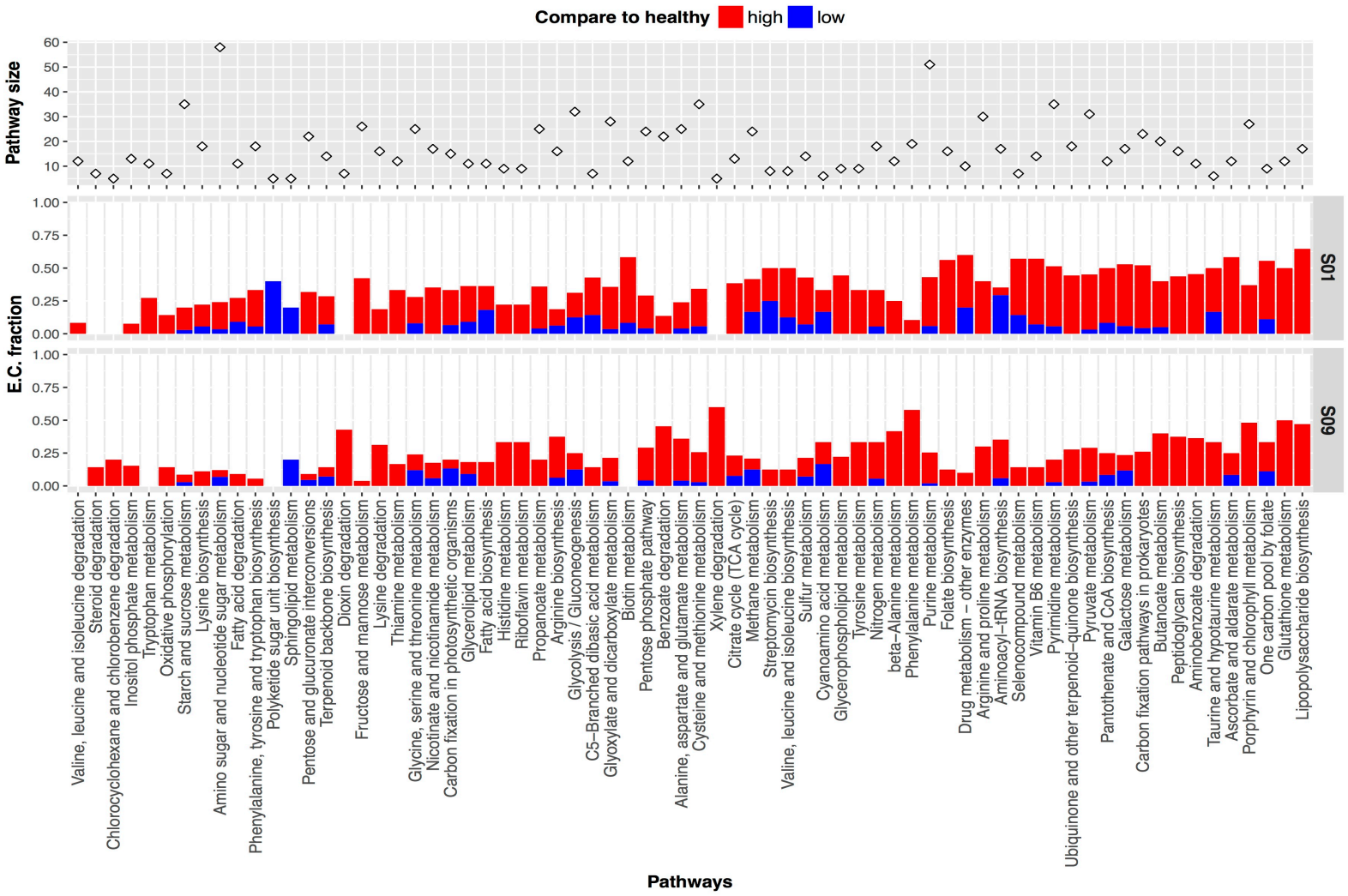


Figure 4-11 Enriched or depleted molecular pathways in microbiomes of CD-affected individuals. Changes in molecular pathways were obtained by counting the numbers of enriched or depleted E.C.s as compared to microbiome functional profiles of the healthy family members.

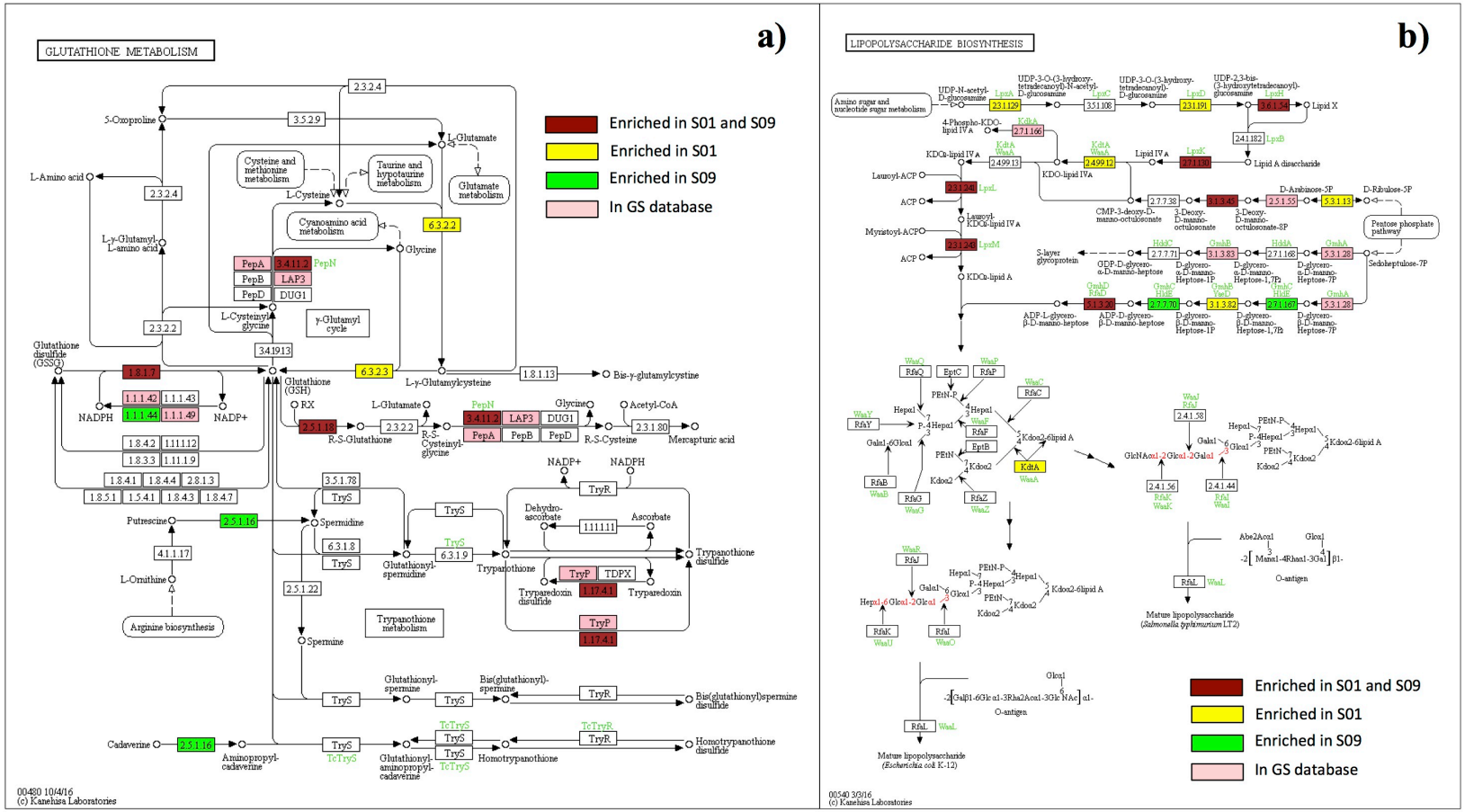


Figure 4-12 The pathways of glutathione metabolism and lipopolysaccharide biosynthesis contain E.C.s enriched in both S01 and S09. a) glutathione metabolism and **b)** lipopolysaccharide biosynthesis.The E.C.s enriched in both S01 and S09 (brown), E.C.s only enriched in S01 (yellow), and E.C.s only enriched in S09 (green). Pink colour indicates the rest E.C.s in the GS database.

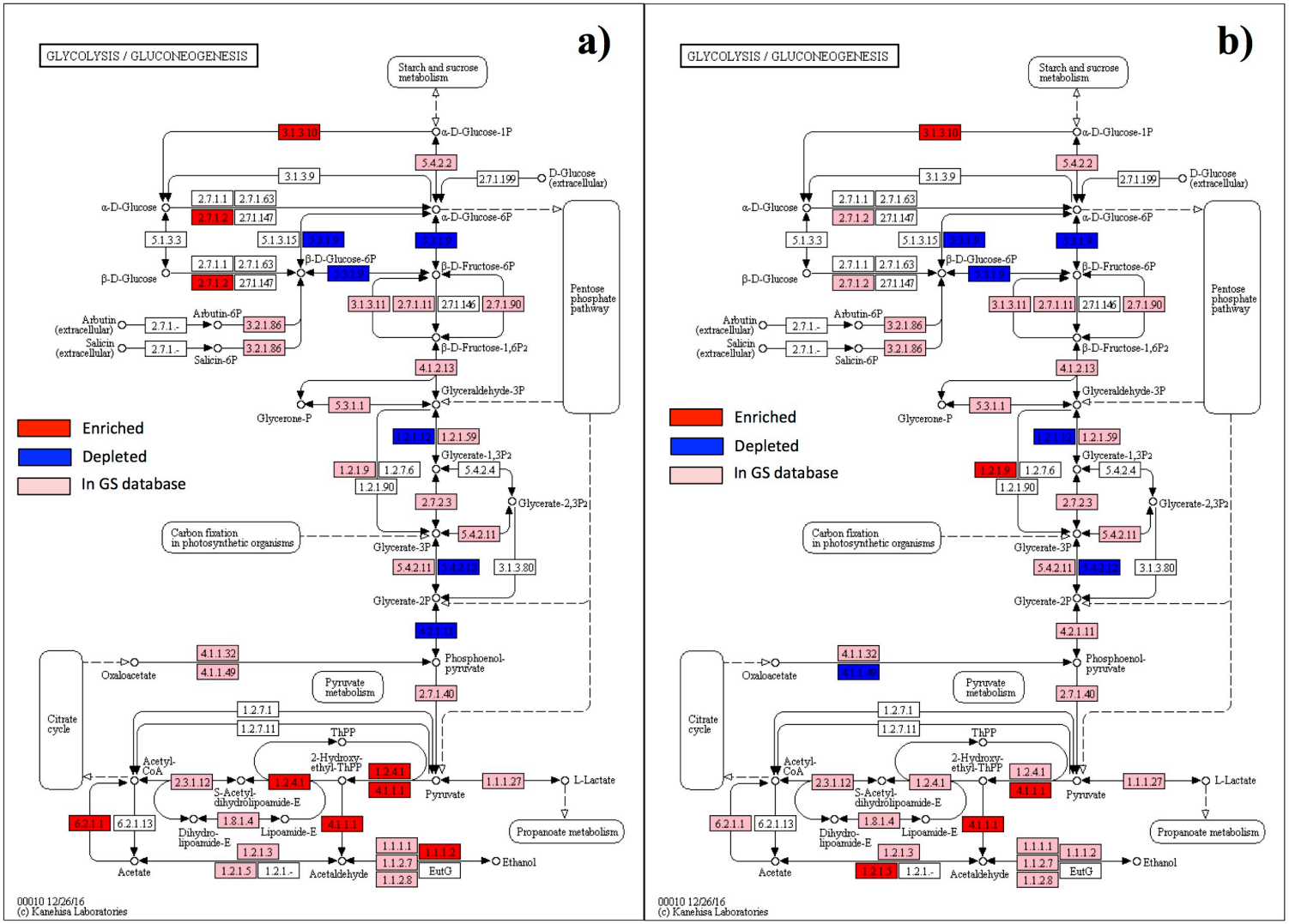


Figure 4-13 The E.C.s associated with acetaldehyde production in glycolysis/gluconeogenesis are enriched in both patients. a) S01 and **b)** S09**.** The red indicates enriched functions, while blue colour indicates depleted functions. Pink indicates the rest of the E.C.s in the GS database. Acetaldehyde is located at bottom center in both diagrams.

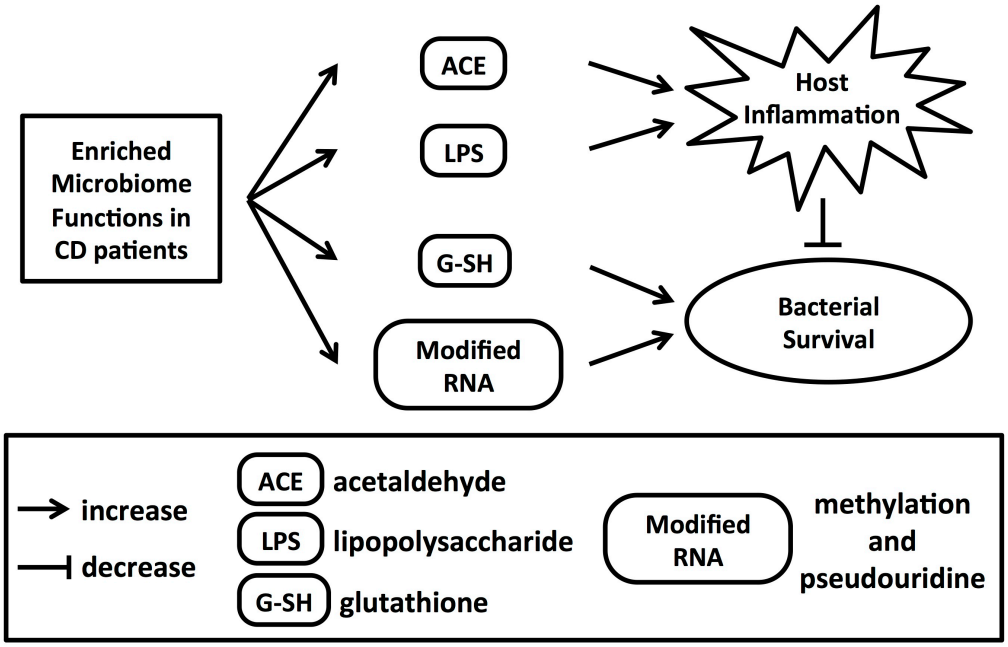


Figure 4-14 Microbial function shift in CD patients is involved in inflammation. Functions that are associated with inflammation inducers (acetaldehyde and lipopolysaccharide) are enriched in CD patient microbiomes, as are the functions that help bacteria survive inflammation conditions (glutathione metabolism, rRNA methytransferase and RNA pseudouridine synthase).

Conclusion

In this study, we compiled a “clean” protein dataset with experimentally confirmed E.C. annotations (gold standard, GS-set), and trained the *faser* algorithm to optimise transfer of function annotation from reference proteins to short peptides translated from sequencing reads. The *faser* algorithm significantly outperforms PSI-BLAST in differentiating functions at high-resolution levels. It also offers ~20% more annotations at comparable precision levels than the function annotation algorithm of MG-RAST. The *mi-faser* pipeline (*faser* in combination with GS*)* was able to identify, in BP oil spill data, unique candidate functions associated with oil-degradation, which were missed by the MG-RAST pipeline. Our pipeline also revealed that gastrointestinal microbiomes of related CD patients are functionally very different. We observed two types of functions enriched in CD patients: those that cause inflammation and those that help bacteria survive inflammatory stress; these may highlight the possible role of the microbiome in CD pathogenicity. Note that all *mi-faser* annotations, although highly informative, are based on the proteins making up the, currently limited, *GS-set*. We expect the growth in the number of proteins with experimentally verified functions to make our approach even more powerful in the near future.

Chapter 5

Microbial functional diversification is directly and tightly linked to the environment. The environmental records, the metadata, are thus very important to understand why evolutionarily closely-related bacteria are so different from each other functionally. In Chapter 2 and 3 we showed that relating bacterial functions to environments yields many interesting findings. However, the currently available metadata have many issues, as they are often 1) incomplete, *e.g.* only 7% of bacteria in our set are annotated with the tolerable pH value, 2) erroneous, *e.g.* some thermophiles are labelled as mesophiles, and 3) not standardized, *e.g.* different labs have different terminologies, blurring the boundary between, for example, aerobic and microaerobic, or thermophile and hyperthermophile. A manual curation (or a, less costly and less accurate, natural language processing) effort will help clarify and augment the current metadata, thus benefitting our future analyses.

With the curated metadata, we can further identify not only organisms, but also functions associated with specific environments. For slowly evolving functions, we can increase the clustering stringency to look for, if there are, environment-associated subgroups for the universal functions, *e.g.*, DNA polymerase in thermophiles should be more stable than DNA polymerase in mesophiles. On the other hand, for fast evolving, or horizontally transferred, functions that enable bacterial adaption to the environment, the signal should be clear. As the result, we will add one more layer to annotation of functions, which can be useful when

the function itself is unknown or function candidates under certain conditions, *e.g.*, low pH, are desired.

In chapter 2, we clustered bacteria based on the functions that they share. With further test for stability we can establish a new bacteria classification scheme. Comparing to the current 16S rRNA-based taxonomy, our advantages include: 1) function-based. We answer the "*what are these bacteria*" question with "*what do these bacteria do*", or to be more precise, "*what can the bacteria do*"; 2) easy assignment. New bacteria genomes can be mapped and assigned in no time to existing clusters via our fusionDB online service (chapter 3); 3) strain-level resolution. 16S rRNA can't differentiate the pathogenic strain from other strains of the same species, *e.g.*, *Escherichia coli* O157:H7. In our classification scheme, we are able to not only resolve strain level differences, but also pin the functions of interest. In addition, like how we clustered the bacteria, we can also cluster the functions based on the bacteria that they co-exist in. Functions that correlate among different bacteria are likely to be involved in the same pathway, which can potentially lead to discovery of novel pathways (previously unknown function clusters). In addition, we can further curate the main function clusters and assign them as signatures for the major bacteria clusters, which helps to make our bacterial classification scheme more informative and useful.

Another obvious application of this project is to combine *fusion*DB (chapter 3) and *mi-faser* (chapter 4). We can substitute the *GS-set* with the collection of proteins from all available bacterial genomes as reference database. In this way, we can offer: 1) wide range of function annotations, including hypothetical and

unknown functions with environment information; 2) pathway reconstructions, as we can now map to the function clusters that we define instead of KEGG pathway; 3) taxonomy annotations, derived from either the fraction of function-repertoire present, or the signature pathways detected. Therefore, for a given metagenomic sample, we hope to be able to answer "*what are these bacteria*", "*what do these bacteria do*", and even "*which bacteria do what*" at the same time. In the end, given limited time and with all the possible improvements in mind, I hope this work can bring in new views on bacteria classification, offer useful tools for (meta)genome annotation and eventually benefit the scientific community.

Reference

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Anderson, M. J. (2001). "A new method for non-parametric multivariate analysis of variance." Austral Ecology **26**(1): 32-46.

Atkinson, K. J. and R. K. Rao (2001). "Role of protein tyrosine phosphorylation in acetaldehyde-induced disruption of epithelial tight junctions." American Journal of Physiology - Gastrointestinal and Liver Physiology **280**(6): G1280.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke and O. Zagnitko (2008). "The RAST Server: rapid annotations using subsystems technology." BMC Genomics **9**: 75.

Bairoch, A., B. Boeckmann, S. Ferro and E. Gasteiger (2004). "Swiss-Prot: Juggling between evolution and stability." Briefings in Bioinformatics **5**(1): 39-55.

Baldridge, K. C. and L. M. Contreras (2014). "Functional implications of ribosomal RNA methylation in response to environmental stress." Critical Reviews in Biochemistry and Molecular Biology **49**(1): 69-89.

Bastian, M., S. Heymann and M. Jacomy (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and E. W. Sayers (2009). "GenBank." Nucleic Acids Res **37**(Database issue): D26-31.

Blondel, V. D. G., Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (2008). "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment(Issue 10): pp. 10008, 10012 pp.

Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: A flexible trimmer for Illumina Sequence Data." Bioinformatics.

Boutet, E., D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret and I. Xenarios (2016). "UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View." Methods Mol Biol **1374**: 23-54.

Brenner, D. J., G. R. Fanning, K. E. Johnson, R. V. Citarella and S. Falkow (1969). "Polynucleotide sequence relationships among members of Enterobacteriaceae." J Bacteriol **98**(2): 637-650.

Brenner, D. J., G. R. Fanning, A. V. Rake and K. E. Johnson (1969). "Batch procedure for thermal elution of DNA from hydroxyapatite." Anal Biochem **28**(1): 447-459.

Buchfink, B., C. Xie and D. H. Huson (2015). "Fast and sensitive protein alignment using DIAMOND." Nat Meth **12**(1): 59-60.

Case, R. J., Y. Boucher, I. Dahllof, C. Holmstrom, W. F. Doolittle and S. Kjelleberg (2007). "Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies." Applied and Environmental Microbiology **73**(1): 278-288.

Chastain, C. J., C. J. Failing, L. Manandhar, M. A. Zimmerman, M. M. Lakner and T. H. T. Nguyen (2011). "Functional evolution of C4 pyruvate,orthophosphate dikinase." Journal of Experimental Botany.

Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel and P. Bork (2006). "Toward automatic reconstruction of a highly resolved tree of life." Science **311**(5765): 1283-1287.

Cohan, F. M. (2001). "Bacterial Species and Speciation." Systematic Biology **50**(4): 513-524.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity and J. M. Tiedje (2009). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis." Nucleic Acids Research **37**: D141-D145.

Criscuolo, A. and S. Gribaldo (2011). "Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria." Mol Biol Evol **28**(11): 3019-3032.

Dagan, T., Y. Artzy-Randrup and W. Martin (2008). "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution." Proc Natl Acad Sci U S A **105**(29): 10039-10044.

Delmont, T. O., E. Prestat, K. P. Keegan, M. Faubladier, P. Robe, I. M. Clark, E. Pelletier, P. R. Hirsch, F. Meyer, J. A. Gilbert, D. Le Paslier, P. Simonet and T. M. Vogel (2012). "Structure, fluctuation and magnitude of a natural grassland soil metagenome." ISME J.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB." Applied and Environmental Microbiology **72**(7): 5069-5072.

Dicksved, J., J. Halfvarson, M. Rosenquist, G. Jarnerot, C. Tysk, J. Apajalahti, L. Engstrand and J. K. Jansson (2008). "Molecular analysis of the gut microbiota of identical twins with Crohn's disease." ISME J **2**(7): 716-727.

Dongen, S. v. (2000). "Graph Clustering by Flow Simulation." PhD thesis, University of Utrecht.

EC, W. (1992). Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes, Academic Press, San Diego, California.

Eddy, S. R. (2011). "Accelerated Profile HMM Searches." PLoS Comput Biol **7**(10): e1002195.

Eric Jones, T. O., Pearu Peterson and others (2001 -). "SciPy: Open Source Scientific Tools for Python."

Felsenstein, J. (2005). "PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.".

Fourment, M. and M. Gibbs (2006). "PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change." BMC Evolutionary Biology **6**(1): 1.

Fox, G. E., J. D. Wisotzkey and P. Jurtshuk, Jr. (1992). "How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity." Int J Syst Bacteriol **42**(1): 166-170.

Frank, D. N., C. E. Robertson, C. M. Hamm, Z. Kpadeh, T. Zhang, H. Chen, W. Zhu, R. B. Sartor, E. C. Boedeker, N. Harpaz, N. R. Pace and E. Li (2011). "Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases." Inflammatory bowel diseases **17**(1): 10.1002/ibd.21339.

Frank, D. N., A. L. St. Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz and N. R. Pace (2007). "Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases." Proceedings of the National Academy of Sciences of the United States of America **104**(34): 13780-13785.

Frank E Harrell Jr , w. c. f. C. D. a. m. o. (2016). "Hmisc: Harrell Miscellaneous. R package version 3.17-4." https://CRAN.R-project.org/package=Hmisc.

Frost, L. S., R. Leplae, A. O. Summers and A. Toussaint (2005). "Mobile genetic elements: the agents of open source evolution." Nat Rev Micro **3**(9): 722-732.

Fuerst, J. A. and E. Sagulenko (2011). "Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function." Nat Rev Microbiol **9**(6): 403-413.

Furnham, N., G. L. Holliday, T. A. P. de Beer, J. O. B. Jacobsen, W. R. Pearson and J. M. Thornton (2014). "The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes." Nucleic Acids Research **42**(Database issue): D485-D489.

Garrity GM, B. D., Castenholz RW, editors (2001). Bergey's Manual of Systematic Bacteriology, Volume 1. New York (NY), Springer.

Gil, R., F. J. Silva, J. Pereto and A. Moya (2004). "Determination of the core of a minimal bacterial gene set." Microbiol Mol Biol Rev **68**(3): 518-537, table of contents.

Glare, T., J. Caradus, W. Gelernter, T. Jackson, N. Keyhani, J. Kohl, P. Marrone, L. Morin and A. Stewart (2012). "Have biopesticides come of age?" Trends Biotechnol **30**(5): 250-258.

Goris, J., K. T. Konstantinidis, J. A. Klappenbach, T. Coenye, P. Vandamme and J. M. Tiedje (2007). "DNA-DNA hybridization values and their relationship to whole-genome sequence similarities." Int J Syst Evol Microbiol **57**(Pt 1): 81-91.

Guimaraes, A. M., A. P. Santos, N. C. do Nascimento, J. Timenetsky and J. B. Messick (2014). "Comparative genomics and phylogenomics of hemotrophic mycoplasmas." PLoS One **9**(3): e91445.

Guo, S., R. Al-Sadi, H. M. Said and T. Y. Ma (2013). "Lipopolysaccharide Causes an Increase in Intestinal Tight Junction Permeability in Vitro and in Vivo by Inducing Enterocyte Membrane Expression and Localization of TLR-4 and CD14." The American Journal of Pathology **182**(2): 375-387.

Haft, D. H., J. D. Selengut and O. White (2003). "The TIGRFAMs database of protein families." Nucleic Acids Res **31**(1): 371-373.

Halary, S., J. W. Leigh, B. Cheaib, P. Lopez and E. Bapteste (2010). "Network analyses structure genetic diversity in independent genetic worlds." Proc Natl Acad Sci U S A **107**(1): 127-132.

Hornef, M. W., M. J. Wick, M. Rhen and S. Normark (2002). "Bacterial strategies for overcoming host innate and adaptive immune responses." Nat Immunol **3**(11): 1033-1040.

Hug, L. A., B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas and J. F. Banfield (2016). "A new view of the tree of life." Nature Microbiology **1**: 16048.

Jacomy, M., T. Venturini, S. Heymann and M. Bastian (2014). "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." PLoS ONE **9**(6): e98679.

Jari Oksanen, F. G. B., Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, R. B. O'Hara, Gavin L. Simpson, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs and Helene Wagner (2016). "vegan: Community Ecology Package. R package version 2.4-0." https://CRAN.R-project.org/package=vegan.

Joshi, N. A. and J. N. Fass (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2016). "KEGG as a reference resource for gene and protein annotation." Nucleic Acids Research **44**(Database issue): D457-D462.

Katoh, K., K. Misawa, K. Kuma and T. Miyata (2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." Nucleic Acids Res **30**(14): 3059-3066.

Kim, S. E., J. S. Moon, W. S. Choi, S. H. Lee and S. U. Kim (2012). "Monitoring of horizontal gene transfer from agricultural microorganisms to soil bacteria and analysis of microbial community in soils." J Microbiol Biotechnol **22**(4): 563-566.

Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes." Proc Natl Acad Sci U S A **102**(7): 2567-2572.

Krieg NR, S. J., Brown DR, Hedlund BP, Paster BJ, Ward NL, Ludwig W and Whitman WB (2011). Bergey's Manual of Systematic Bacteriology, Volume 4. New York (NY), Springer.

Krishnan, M., C. Bharathiraja, J. Pandiarajan, V. A. Prasanna, J. Rajendhran and P. Gunasekaran (2014). "Insect gut microbiome - An unexploited reserve for biotechnological application." Asian Pac J Trop Biomed **4**(Suppl 1): S16-21.

Krupovic, M., M. Gonnet, W. B. Hania, P. Forterre and G. Erauso (2013). "Insights into Dynamics of Mobile Genetic Elements in Hyperthermophilic Environments from Five New Thermococcus Plasmids." PLoS ONE **8**(1): e49044.

Kruskal, J. B. (1964). "Nonmetric multidimensional scaling: A numerical method." Psychometrika **29**(2): 115-129.

Lambiotte, R. D., J.-C.; Barahona, M. (2008). "Laplacian Dynamics and Multiscale Modular Structure in Networks." New discussions on the selection of the most significant scales and the generalisation of stability to directed networks; IEEE Transactions on Network Science and Engineering **1**(2): pp 76-90.

Lawrence, J. G. (2002). "Gene transfer in bacteria: speciation without species?" Theor Popul Biol **61**(4): 449-460.

Lee, S. H. (2015). "Intestinal Permeability Regulation by Tight Junction: Implication on Inflammatory Bowel Diseases." Intestinal Research **13**(1): 11-18.

Leinonen, R., R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, G. Hoad, M. Jang, N. Pakseresht, S. Plaister, R. Radhakrishnan, K. Reddy, S. Sobhany, P. Ten Hoopen, R. Vaughan, V. Zalunin and G. Cochrane (2010). "The European Nucleotide Archive." Nucleic Acids Research.

Letunic, I. and P. Bork (2011). "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy." Nucleic Acids Res **39**(Web Server issue): W475-478.

Liu, L., X. Chen, G. Skogerbo, P. Zhang, R. Chen, S. He and D. W. Huang (2012). "The human microbiome: a hot spot of microbial horizontal gene transfer." Genomics **100**(5): 265-270.

Lloyd, S. (1982). "Least squares quantization in PCM." IEEE Transactions on Information Theory **28**(2): 129-137.

Manichanh, C., L. Rigottier−Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca and J. Dore (2006). "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach." Gut **55**(2): 205-211.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Markowitz, V. M., I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, A. Ratner, B. Jacob, J. Huang, P. Williams, M. Huntemann, I. Anderson, K. Mavromatis, N. N. Ivanova and N. C. Kyrpides (2012). "IMG: the integrated microbial genomes database and comparative analysis system." Nucleic Acids Research **40**(D1): D115-D122.

Markowitz, V. M., I.-M. A. Chen, K. Palaniappan, K. Chu, E. Szeto, M. Pillay, A. Ratner, J. Huang, T. Woyke, M. Huntemann, I. Anderson, K. Billis, N. Varghese, K. Mavromatis, A. Pati, N. N. Ivanova and N. C. Kyrpides (2014). "IMG 4 version of the integrated microbial genomes comparative analysis system." Nucleic Acids Research **42**(D1): D560-D567.

Marrero, G., K. L. Schneider, D. M. Jenkins and A. M. Alvarez (2013). "Phylogeny and classification of Dickeya based on multilocus sequence analysis." Int J Syst Evol Microbiol **63**(Pt 9): 3524-3539.

Martinez-Medina, M., X. Aldeguer, M. Lopez-Siles, F. González-Huix, C. López-Oliu, G. Dahbi, J. E. Blanco, J. Blanco, J. L. Garcia-Gil and A. Darfeuille-Michaud (2009). "Molecular diversity ofEscherichia coliin the human gut: New ecological evidence supporting the role of adherent–invasiveE. coli(AIEC) in Crohn's disease." Inflammatory Bowel Diseases **15**(6): 872-882.

Martinezmurcia, A. J., S. Benlloch and M. D. Collins (1992). "Phylogenetic interrelationships of members of the genera Aeromonas and Plesiomonas as determined by 16S ribosomal DNA sequencing: lack of congruence with results of DNA-DNA hybridizations." International Journal of Systematic Bacteriology **42**(3): 412-421.

Masip, L., K. Veeravalli and G. Georgiou (2006). "The Many Faces of Glutathione in Bacteria." Antioxidants & Redox Signaling **8**(5-6): 753-762.

Morgan, X. C., T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier and C. Huttenhower (2012). "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment." Genome Biology **13**(9): R79.

Neefs, J. M., Y. Vandepeer, L. Hendriks and R. Dewachter (1990). "Compilation of small ribosdomal subunit RNA sequences." Nucleic Acids Research **18**: 2237-2317.

Neimark, H., K. E. Johansson, Y. Rikihisa and J. G. Tully (2001). "Proposal to transfer some members of the genera Haemobartonella and Eperythrozoon to the genus Mycoplasma with descriptions of 'Candidatus Mycoplasma haemofelis', 'Candidatus Mycoplasma haemomuris', 'Candidatus Mycoplasma haemosuis' and 'Candidatus Mycoplasma wenyonii'." Int J Syst Evol Microbiol **51**(Pt 3): 891-899.

Neimark, H., K. E. Johansson, Y. Rikihisa and J. G. Tully (2002). "Revision of haemotrophic Mycoplasma species names." Int J Syst Evol Microbiol **52**(Pt 2): 683.

Oliveros, J. C. (2007). "VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html..

Pagani, I., K. Liolios, J. Jansson, I. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz and N. C. Kyrpides (2012). "The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata." Nucleic Acids Res **40**(Database issue): D571-579.

Peters, G. (1991). "Azolla and other plant-cyanobacteria symbioses: Aspects of form and function." Plant and Soil **137**(1): 25-36.

Porter, J. R. (1976). "Antony van Leeuwenhoek: tercentenary of his discovery of bacteria." Bacteriol Rev **40**(2): 260-269.

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldon, T. Rattei, C. Creevey, M. Kuhn, L. J. Jensen, C. von Mering and P. Bork (2014). "eggNOG v4.0: nested orthology inference across 3686 organisms." Nucleic Acids Res **42**(Database issue): D231-239.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies and F. O. Gloeckner (2007). "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB." Nucleic Acids Research **35**(21): 7188-7196.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman and R. D. Finn (2012). "The Pfam protein families database." Nucleic Acids Research **40**(D1): D290-D301.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J. M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, M. Jian, Y. Zhou, Y. Li, X. Zhang, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork and S. D. Ehrlich (2010). "A human gut microbial gene catalogue established by metagenomic sequencing." Nature **464**(7285): 59-65.

Rippka, R., J. Deruelles, J. B. Waterbury, M. Herdman and R. Y. Stanier (1979). "Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria." Journal of General Microbiology **111**(1): 1-61.

Rodriguez-R, L. M., W. A. Overholt, C. Hagan, M. Huettel, J. E. Kostka and K. T. Konstantinidis (2015). "Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill." ISME J **9**(9): 1928-1940.

Rokach, L., and Oded Maimon (2005). "Clustering methods." Data mining and knowledge discovery handbook., Springer US.

Rossello-Mora, R. and R. Amann (2001). "The species concept for prokaryotes." Fems Microbiology Reviews **25**(1): 39-67.

Rost, B. (2002). "Enzyme function less conserved than anticipated." J Mol Biol **318**(2): 595-608.

Saye, D. J., O. Ogunseitan, G. S. Sayler and R. V. Miller (1987). "Potential for transduction of plasmids in a natural freshwater environment: effect of plasmid donor concentration and a natural microbial community on transduction in Pseudomonas aeruginosa." Applied and Environmental Microbiology **53**(5): 987-995.

Schirrmeister, B. E., M. Anisimova, A. Antonelli and H. C. Bagheri (2011). "Evolution of cyanobacterial morphotypes: Taxa required for improved phylogenomic approaches." Commun Integr Biol **4**(4): 424-427.

Schmieder, R. and R. Edwards (2011). "Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets." PLOS ONE **6**(3): e17288.

Schneider, R., A. de Daruvar and C. Sander (1997). "The HSSP database of protein structure-sequence alignments." Nucleic Acids Research **25**(1): 226-230.

Schnoes, A. M., S. D. Brown, I. Dodevski and P. C. Babbitt (2009). "Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies." PLoS Comput Biol **5**(12): e1000605.

Simon, C., A. Wiezer, A. W. Strittmatter and R. Daniel (2009). "Phylogenetic diversity and metabolic potential revealed in a glacier ice metagenome." Appl Environ Microbiol **75**(23): 7519-7526.

Snel, B., P. Bork and M. A. Huynen (1999). "Genome phylogeny based on gene content." Nature Genetics **21**(1): 108-110.

Sokal, S. a. (1973). Numerical taxonomy — The principles and practice of numerical classification. San Francisco, W H Freeman & Co (Sd) (June 1973).

Sokol, H., P. Seksik, J. P. Furet, O. Firmesse, I. Nion-Larmurier, L. Beaugerie, J. Cosnes, G. Corthier, P. Marteau and J. Doré (2009). "Low counts ofFaecalibacterium prausnitziiin colitis microbiota." Inflammatory Bowel Diseases **15**(8): 1183-1189.

Stackebrandt, E. and B. M. Goebel (1994). "A PLACE FOR DNA-DNA REASSOCIATION AND 16S RIBOSOMAL-RNA SEQUENCE-ANALYSIS IN THE PRESENT SPECIES DEFINITION IN BACTERIOLOGY." International Journal of Systematic Bacteriology **44**(4): 846-849.

Staley, J. T. (2006). "The bacterial species dilemma and the genomic-phylogenetic species concept." Philos Trans R Soc Lond B Biol Sci **361**(1475): 1899-1909.

Sun, W., G. Yu, T. Louie, T. Liu, C. Zhu, G. Xue and P. Gao (2015). "From mesophilic to thermophilic digestion: the transitions of anaerobic bacterial, archaeal, and fungal community structures in sludge and manure samples." Appl Microbiol Biotechnol **99**(23): 10271-10282.

Swingley, W. D., M. Chen, P. C. Cheung, A. L. Conrad, L. C. Dejesa, J. Hao, B. M. Honchak, L. E. Karbach, A. Kurdoglu, S. Lahiri, S. D. Mastrian, H. Miyashita, L. Page, P. Ramakrishna, S. Satoh, W. M. Sattley, Y. Shimada, H. L. Taylor, T. Tomo, T. Tsuchiya, Z. T. Wang, J. Raymond, M. Mimuro, R. E. Blankenship and J. W. Touchman (2008). "Niche adaptation and genome expansion in the

chlorophyll d-producing cyanobacterium Acaryochloris marina." Proc Natl Acad Sci U S A **105**(6): 2005-2010.

Sykora, P. (1992). "Macroevolution of plasmids: a model for plasmid speciation." J Theor Biol **159**(1): 53-65.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale (2003). "The COG database: an updated version includes eukaryotes." BMC Bioinformatics **4**: 41.

Thompson, A. W., R. A. Foster, A. Krupke, B. J. Carter, N. Musat, D. Vaulot, M. M. M. Kuypers and J. P. Zehr (2012). "Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga." Science **337**(6101): 1546-1550.

Titsworth, E., E. Grunberg, G. Beskid, R. Cleeland, Jr. and W. F. Delorenzo (1969). "Efficiency of a multitest system (Enterotube) for rapid identification of Enterobacteriaceae." Appl Microbiol **18**(2): 207-213.

Uilenberg, G., F. ß. Thiaucourt and F. Jongejan (2004). "On molecular taxonomy: what is in a name?" Experimental & Applied Acarology **32**(4): 301-312.

Vandamme, P., B. Pot, M. Gillis, P. DeVos, K. Kersters and J. Swings (1996). "Polyphasic taxonomy, a consensus approach to bacterial systematics." Microbiological Reviews **60**(2): 407-+.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers and H. O. Smith (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

Weissgerber, T., R. Zigann, D. Bruce, Y. J. Chang, J. C. Detter, C. Han, L. Hauser, C. D. Jeffries, M. Land, A. C. Munk, R. Tapia and C. Dahl (2011). "Complete genome sequence of Allochromatium vinosum DSM 180(T)." Stand Genomic Sci **5**(3): 311-330.

Wilke, A., J. Bischof, T. Harrison, T. Brettin, M. D'Souza, W. Gerlach, H. Matthews, T. Paczian, J. Wilkening, E. M. Glass, N. Desai and F. Meyer (2015). "A RESTful API for Accessing Microbial Community Data for MG-RAST." PLOS Computational Biology **11**(1): e1004008.

Wilke, A., T. Harrison, J. Wilkening, D. Field, E. M. Glass, N. Kyrpides, K. Mavrommatis and F. Meyer (2012). "The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools." BMC Bioinformatics **13**(1): 1-5.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: the primary kingdoms." Proceedings of the National Academy of Sciences of the United States of America **74**(11): 5088-5090.

Wommack, K. E., J. Bhavsar and J. Ravel (2008). "Metagenomics: read length matters." Appl Environ Microbiol **74**(5): 1453-1463.

Wu, D., M. Wu, A. Halpern, D. B. Rusch, S. Yooseph, M. Frazier, J. C. Venter and J. A. Eisen (2011). "Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees." Plos One **6**(3).

Wu, M. and J. A. Eisen (2008). "A simple, fast, and accurate method of phylogenomic inference." Genome Biol **9**(10): R151.

Zengler, K., G. Toledo, M. Rappe, J. Elkins, E. J. Mathur, J. M. Short and M. Keller (2002). "Cultivating the uncultured." Proceedings of the National Academy of Sciences of the United States of America **99**(24): 15681-15686.

Zhu, C., T. O. Delmont, T. M. Vogel and Y. Bromberg (2015). "Functional Basis of Microorganism Classification." PLoS Comput Biol **11**(8): e1004472.