Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1060

*fusion*DB: assessing microbial diversity and environmental preferences via functional similarity networks

Chengsheng Zhu[1,†], Yannick Mahlich[1,2,3,4,*,†], Maximilian Miller[1,2,3,†] and Yana Bromberg[1,4,*]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA, [2]Computational Biology & Bioinformatics - i12 Informatics, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748 Garching/Munich, Germany, [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technical University of Munich (TUM), 85748 Garching/Munich, Germany and [4]Institute for Advanced Study, Technical University of Munich (TUM), Lichtenbergstrasse 2 a, 85748 Garching/Munich, Germany

Received August 11, 2017; Revised September 24, 2017; Editorial Decision October 12, 2017; Accepted October 22, 2017

ABSTRACT

Microbial functional diversification is driven by environmental factors, i.e. microorganisms inhabiting the same environmental niche tend to be more functionally similar than those from different environments. In some cases, even closely phylogenetically related microbes differ more across environments than across taxa. While microbial similarities are often reported in terms of taxonomic relationships, no existing databases directly link microbial functions to the environment. We previously developed a method for comparing microbial functional similarities on the basis of proteins translated from their sequenced genomes. Here, we describe *fusion*DB, a novel database that uses our functional data to represent 1374 taxonomically distinct bacteria annotated with available metadata: habitat/niche, preferred temperature, and oxygen use. Each microbe is encoded as a set of functions represented by its proteome and individual microbes are connected via common functions. Users can search *fusion*DB via combinations of organism names and metadata. Moreover, the web interface allows mapping new microbial genomes to the functional spectrum of reference bacteria, rendering interactive similarity networks that highlight shared functionality. *fusion*DB provides a fast means of comparing microbes, identifying potential horizontal gene transfer events, and highlighting key environment-specific functionality.

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 (1). 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 (2), COG (3), TIGRfam (4), RAST (5) 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 (6). While it is well known that environmental factors play an important role in microbial functionality (7), 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* (8), 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 vi-

*To whom correspondence should be addressed. Tel: +1 848 932 5638; Fax +1 848 932 8965; Email: ymahlich@bromberglab.org
Correspondence may also be addressed to Bromberg Yana. Tel: +1 646 220 3290; Email: yanab@rci.rutgers.edu
†These authors contributed equally to this work as first authors.

sualized, 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 to the fact that environmental factors drive 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 (8), 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, in the near future, *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 1374 bacterial genomes (December 2011 NCBI GenBank (9). For each bacterium, we store its (a) NCBI taxonomic information (10) and, where available, (b) environmental metadata (temperature, oxygen requirements, and habitat; GOLD (6). The environments are generalized, e.g. *thermophiles* include hyper-thermophiles. '*No data*' is used to indicate missing annotations (Supplementary Online Material, SOM Table S1, SOM Figure S1). The general *fusion* (functional repertoire similarity-based organism network) protocol is described in our previous work (8). Briefly, all proteins in our database are aligned against each other using three iterations of PSI-BLAST (11) and the alignment length and sequence identity are used to compute Homology-derived Secondary Structure of Proteins (HSSP) distances (12). A network of protein similarities is then clustered using the Markov Clustering Algorithm (MCL) (13). For *fusion*DB 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 1). Each bacterium was thus mapped to a set of functions, its functional repertoire (∼2400 functions on average, ranging from 118 to 6134 functions). Note that our functional repertoires include all the bacterial functions, regardless of annotation. We are thus able to make function predictions for proteins in new bacteria, even if these functions have not been annotated before.

Mapping new organisms to fusion

User submitted microbial proteomes and the associated functions are stored in a separate database (SOM Figure S2). For each query protein of the new organism, the mapping pipeline (SOM Figure S3, SOM Methods) (a) runs PSI-BLAST (reporting e-value 1e–10, inclusion e-value 1e–3, three iterations) against reference proteins in *fusion*DB and (b) maps the query to a *fusion* functional cluster, which contains the reference with the highest hit HSSP score. Note that novel proteins that cannot be assigned to existing functional groups (do not match any reference at HSSP distance $\geq$10) are reported as functional singletons even if they are similar among themselves. Additionally, protein alignments that exceed 12 CPU hours of run-time are currently eliminated from further consideration. In testing, we found that no $>$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 turnaround. Note that we also evaluated a number of other algorithms for mapping organism functional repertoires, of which the above-described algorithm performed best (SOM Methods).

All functional cluster assignments of proteins in the query proteome are then combined into a functional repertoire where each functional cluster is unique; i.e. if two query proteins are assigned to the same functional cluster, this cluster is listed only once in the final repertoire.

Evaluating *fusion*DB performance

We evaluated the accuracy of functional mapping of new proteomes by iteratively mapping each of the *fusion*DB organisms back to the remaining 1373. We aligned each protein of the query organism to all proteins in other organisms and selected the alignment with highest HSSP score. We then assigned the query protein to the functional cluster of its match as described above for mapping new organisms.

The performance of this approach was evaluated on a per-function basis, i.e. for each function of each 'newly added' organism we retrieved counts of true positives (TP, proteins correctly assigned to this *fusion*DB function), false positives (FP, proteins falsely assigned to this *fusion*DB function), and false negatives (FN, proteins that are part of this *fusion*DB function in the reference database, but not correctly assigned). Note that reference singleton proteins that were not assigned to any *fusion*DB function were considered true positives. Averaged across all functions, the mean per-function precision and recall of correctly assigning proteins were 97.2% and 96.6%, respectively (3.1×10^{-8} mean per function false positive rate, FPR), while the overall precision of assigning any protein to a function was 98.2% (Eq. 1).

Individual organisms were assigned to their functional repertoires with 99.5% precision and 98.9% recall (Eq. 1,

Table 1. 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

SOM Figures S4 and S5). For this estimate we evaluated to overlap between reference and assigned repertoire; i.e. functional clusters that appear in both the reference and mapped functional repertoire are true positives. False positives are functional clusters in the mapped functional repertoire but not the reference repertoire, false negatives vice versa. The reported precision and recall are the mean precision and recall values averaged over all organism submissions.

$$precision = \frac{TP}{TP+FP},\ recall = \frac{TP}{TP+FN},\ FPR = \frac{FP}{FP+TN} \quad (1)$$

Web interface

*fusion*DB web interface has two functions: *explore* and *map new organisms*. The *explore* section contains access to all the 1374 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. A user-selected organism set can be 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 *toward* other organisms via common functions and *away* from others via unique functions (8). 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 primary function annotation of each functional cluster is the myRAST (5) description most commonly assigned to the cluster members. For each cluster we also include the corresponding Pfam (2) families. 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 (SOM Figure S3). The server sends out emails to users when mapping is finished. The *map* result page contains two tables containing (a) functional annotations, including the associated *fusion*DB reference sequences and proteins of the query organism that mapped to each functional cluster, as well as (b) similarity (Eq. 2) to the reference organisms in *fusion*DB, including functional repertoire size, functional overlap with the query, and metadata. 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 described above (Figure 1).

$$\text{similarity} = \frac{\text{shared functions}}{\text{the larger functional repertoire size}} \quad (2)$$

Analysis of environment-driven organism similarity

For each environmental condition in *fusionDB*, we sampled organism pairs where organisms were from (a) the same condition (SC, e.g. both mesophiles) and (b) different conditions (DC, e.g. thermophile versus 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 (SOM Table S1; 136 pairs—48 same phylum, 88 different phyla; due to high functional diversity of *Proteobacteria,* its classes were considered independent phyla). For all other environmental factors we sampled (bootstrap with 100 resamples) 136 organism pairs for both SC and DC sets, covering this same minimum taxonomic diversity. We calculated the pairwise functional similarity (Eq. 2) distributions and discarded organism pairs with <5% similarity.

RESULTS AND DISCUSSION

Mapping a new Synechococcus genome 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 (9) and submitted it to our web interface. This 3,318 protein fresh water Cyanobacteria is isolated from a Sphagnum (peat moss) bog (6). 86% (2,853) of the bacterial proteins mapped to 2208 *fusion*DB functions, while 462 (14%) were functional singletons; three proteins exceeded runtime and were excluded (Methods). The whole process from submission to results notification e-mail took under three and a half hours. The mapping indicates that *Synechococcus* sp. PCC 7502 is most functionally similar (56%) to *Synechocystis* PCC 6803, a fresh water organism evolutionarily 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

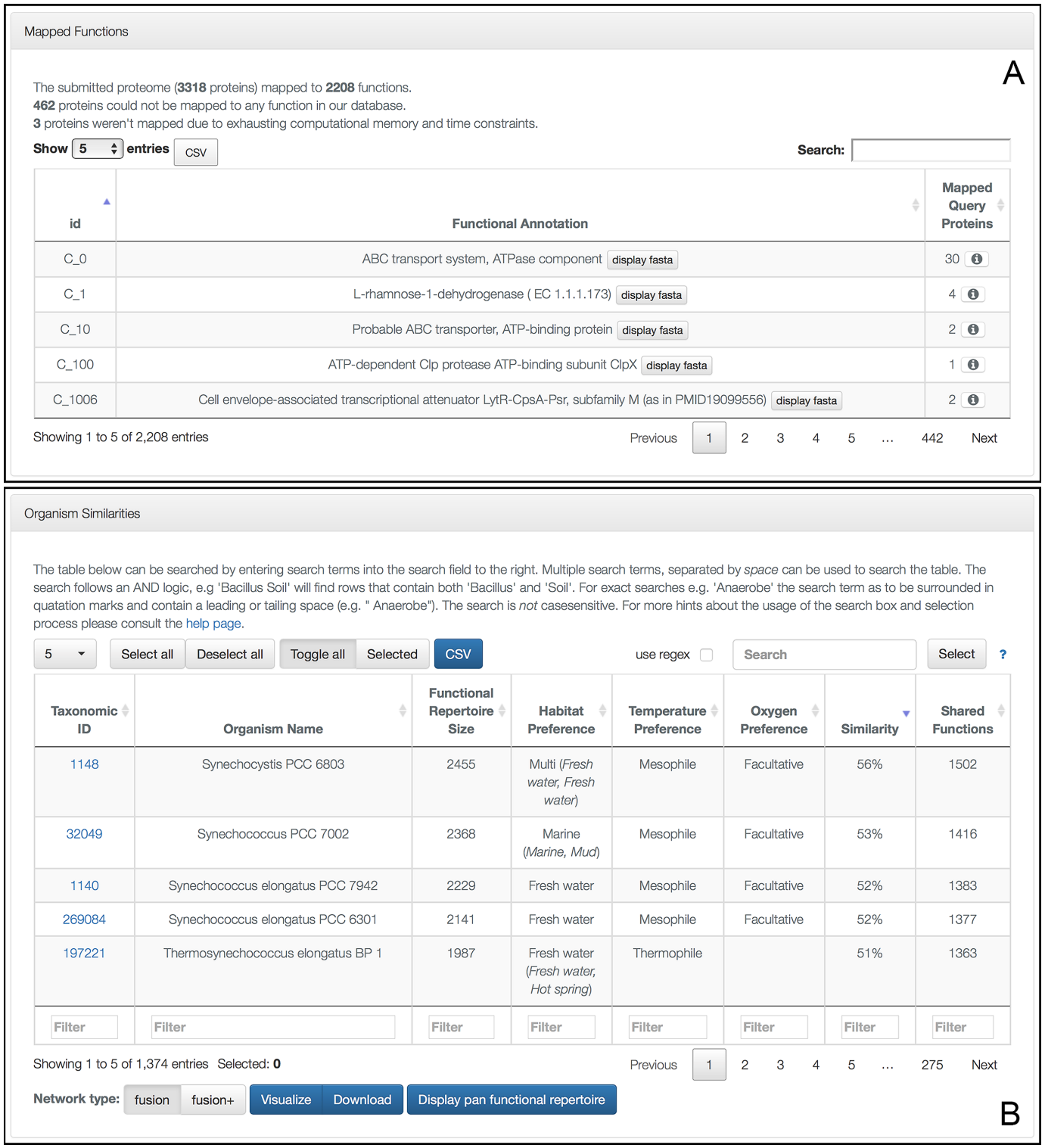

Figure 1. Screenshot of the organism mapping result page. (**A**) The 'Mapped Functions' table lists the functions that the submitted organism is mapped to. For each function, associated proteins from *fusion*DB and mapped query proteins can be displayed. (**B**) The 'Organism Similarities' table displays, all 1374 *fusion*DB organisms and their functional similarities to the query organism, including additional information such as environmental metadata; the view can be toggled between all and user-selected organisms. *fusion*(+) networks of the query and user-selected organisms can be created for on-site visualization (see Figure 2) or download.

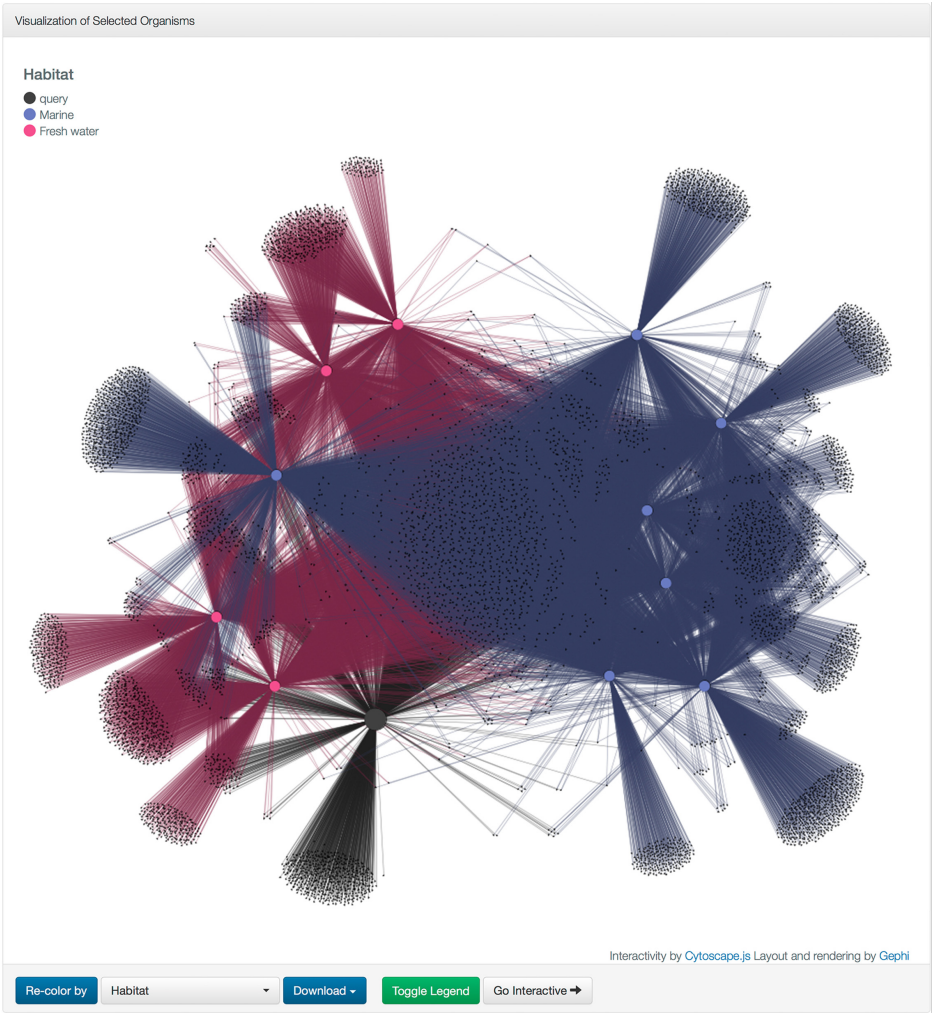

Figure 2. Screenshot of the fusion+ visualization of all Synechoccocus genomes. The submitted *Synechococcus* sp. PCC 7502 (query, black) clusters with the fresh water *Synechococcus* organisms (magenta). Note that *Synechococcus* sp. PCC 7002 – clustered among fresh water organisms; colored dark blue (marine) – is isolated from marine mud. It is salt tolerant but does not require salt for growth).

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* (SOM Data 1), the core-function repertoire for this genus, and 1128 functions shared among only the fresh water *Synechococcus* (SOM Data 2). These differential 254 functions (SOM Data 3) are likely important for living in fresh water, as opposed to marine, environment, e.g. low salinity and low osmotic pressure.

Environment significantly affects microbial function

In our evaluation of the effects of environmental pressures on microbial functionality we found that, in general, same environmental condition (SC) organisms across all environmental factors are more functionally similar than DC organisms (from different environments; Figure 3; with some exceptions mentioned below, Kolmogorov-Smirnov test (14) *P*-value $< 2.5e{-}6$). This finding is intuitive and many studies have demonstrated the presence of horizontal gene transfer (HGT) within environment-specific microbiomes (15–17). Our results, however, for the first time, quantify on a broad scale the environmental impact on microorganism function diversification.

SC-thermophile and SC-psychrophile pairs demonstrate significantly higher similarities when compared to DC pairs (Figure 3A). Notably, the higher functional similarity between thermophiles than between psychrophiles suggests that protein functional adaptation to low temperature may be less taxing than to high temperature – an interesting finding in itself. When contrasted with the extremophiles, mesophiles seem to have much larger functional diversity; in fact, SC-mesophile similarities are comparable to those of DC pairs (Figure 3A).

Different molecular pathways of aerobic-respiration and anaerobic-respiration/fermentation may explain the high level of dissimilarity between the aerobes and anaerobes (DC-anaerobe-aerobe; Figure 3B). Interestingly, the SC-anaerobe similarities are higher than the SC-aerobe similarities, likely because the more ancient anaerobic-respiration/fermentation machinery tends to be simpler (fewer reactions) (18) and more conserved.

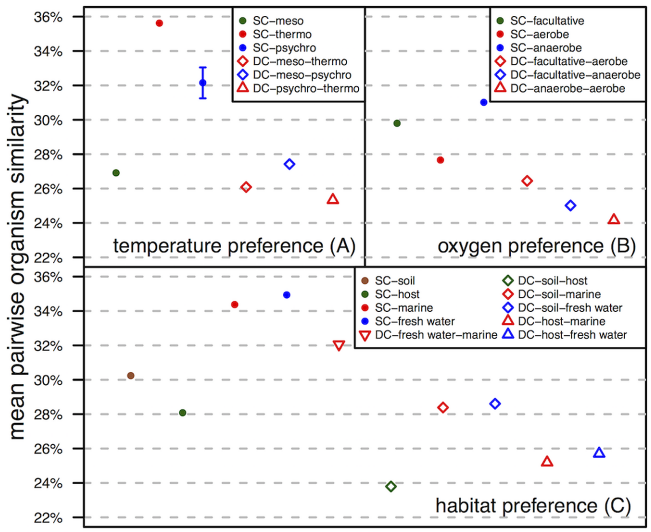

Figure 3. Organism pairwise similarity is higher among organisms living in the same environmental conditions. The mean pairwise similarity for same (SC) and different (DC) condition organisms according to (**A**) temperature, (**B**) oxygen and (**C**) habitat preferences. For all points without error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

Different habitat (DC) samples show lower pairwise organism similarity than SC samples as well (Figure 3C). Interestingly fresh water and marine organism similarity (DC-fresh water-marine) is fairly high, likely due to overlaps in requirements of the aquatic conditions. Note however, that the dissimilarity across fresh water and marine conditions is still high enough to differentiate organisms of the same taxa (e.g. strains of *Synechococcus* in Figure 2). SC-host has the lowest mean organism similarity of the habitat SC samples; we speculate this to be a result of differential adaptations necessary to deal with diverse host defense mechanisms (19). The soil organisms also share low functional similarity, which is likely due to soil heterogeneity at physical, chemical, and biological levels, from nano- to landscape scale (20).

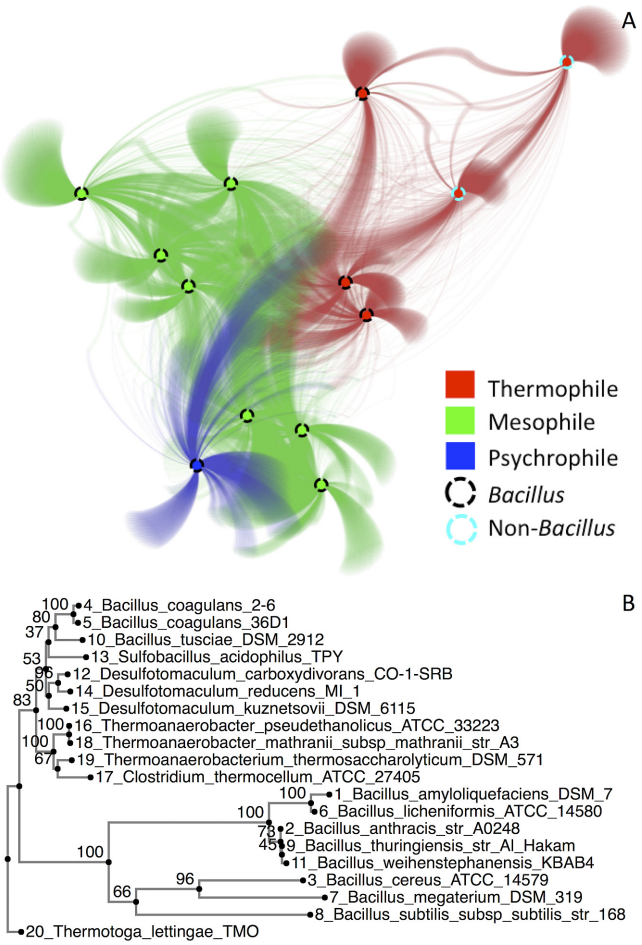

Figure 4. *fusion*DB reveals an HGT event between thermophilic *Bacilli* and thermophilic *Clostridia*. (**A**) *fusion*+ visualization of Bacillus and thermophilc Clostridia**.** Large organism nodes are connected via small function nodes. The two thermophilic *Clostridia* are connected to the thermophilic *Bacilli* via functions that are possibly horizontally transferred; (**B**) phylogenetic analysis of pyruvate, phosphate dikinase (PPDK) gene suggests HGT between thermophilic *Bacilli* and thermophilic *Clostridia.* The PPDK genes in thermophilic *Bacilli* are evolutionarily more related to those in thermophilic *Clostridia* than those in other *Bacilli*.

Case study of a temperature driven HGT event

Using the *fusion*DB *explore* functionality, we extracted thermophilic, mesophilic, and psychrophilic species representatives (one per species) of the *Bacillus* genus. We also added two other thermophilic *Clostridia*, *Desulfotomaculum carboxydivorans* CO-1-SRB and *Sulfobacillus acidophilus* TPY, to generate a *fusion*+ network (SOM Table S2; Figure S4A). As expected, note here that overall thermophilic bacteria are further removed from psychrophiles than from mesophiles. Moreover, the thermophilic *Bacilli* were more closely related to the non-*Bacillus* thermophiles than to other *Bacilli*. The three *Bacilli* thermophiles share 29 functions (SOM Data 4) that are not found in other *Bacilli* in this organism set, three of which also exist in the two thermophilc *Clostridia*. One is a likely pyruvate phosphate dikinase (PPDK) that, in extremophiles, works as a primary glycolysis enzyme (21). The thermophilic *Bacilli*'s PPDK proteins are more similar to those in thermophilic *Clostridia* (sequence identity = 0.65 ± 0.03), than to those in mesophilic/psychrophilic *Bacilli* (sequence identity = 0.17 ± 0.05). Phylogenetic analysis of the genes with additional thermophilic organisms (SOM Methods) suggests a likely HGT event between the thermophilic organisms (Figure 4B). 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 (22). Shared closely-related MGEs in distant organisms imply HGT (23). We thus suggest that *fusion*DB offers a fast and easy way to trace likely functionally necessary HGT events within niche-specific microbial communities.

In this work, 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

annotations. By mapping metagenome assemblies to *fusion*DB, both the functional and taxonomical annotations can be obtained. Moreover, our recent work (Zhu *et al.* 2017, Functional sequencing read annotation for high precision microbiome analysis, *submitted*) suggests that accurate functional annotations can also be obtained without assembly. We thus also expect to make *fusion*DB useful in this type of analyses in the near future.

CONCLUSIONS

*fusion*DB links microbial functional similarities and environmental preferences. Our analysis reveals environmental factors driving microbial functional diversification. By mapping new organisms to the reference functional space, our database 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.

AVAILABILITY

*fusion*DB is publicly available at http://services.bromberglab.org/fusiondb/

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Drs Burkhard Rost (TU Munich), Max Haggblom, Tamar Barkay (both Rutgers), and Tom O. Delmont (U Chicago) for all help with interpreting our data and understanding the community needs. Big thanks to Yanran Wang and Dr Anton Molyboha (both Rutgers) for all discussions. We want to thank the anonymous reviewers for their thorough review and suggestions to improve this manuscript. We are also grateful to all those who deposit their data in public databases – fusionDB wouldn't be possible without them.

FUNDING

NSF CAREER Award [1553289 to Y.B. and C.Z.]; USDA-NIFA [1015:0228906]; TU München – Institute for Advanced Study Hans Fischer Fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme [291763 to Y.B. and Y.M.]; German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.
Conflict of interest statement. None declared.

REFERENCES

1. Garrity,G.M., Boone,D.R. and Castenholz,R.W. (eds). (2001) *Bergey's Manual of Systematic Bacteriology*. 2nd edn. Springer, NY, Vol. **1**.
2. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
3. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
4. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
5. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
6. Pagani,I., Liolios,K., Jansson,J., Chen,I.M., Smirnova,T., Nosrat,B., Markowitz,V.M. and Kyrpides,N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **40**, D571–D579.
7. Sun,W., Yu,G., Louie,T., Liu,T., Zhu,C., Xue,G. and Gao,P. (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**, 10271–10282.
8. Zhu,C., Delmont,T.O., Vogel,T.M. and Bromberg,Y. (2015) Functional basis of microorganism classification. *PLoS Comput. Biol.*, **11**, e1004472.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
10. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
13. Dongen,S.V. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht.
14. Massey,F.J. (1951) The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Statist. Assoc.*, **46**, 68–78.
15. Kim,S.E., Moon,J.S., Choi,W.S., Lee,S.H. and Kim,S.U. (2012) Monitoring of horizontal gene transfer from agricultural microorganisms to soil bacteria and analysis of microbial community in soils. *J. Microbiol. Biotechnol.*, **22**, 563–566.
16. Liu,L., Chen,X., Skogerbo,G., Zhang,P., Chen,R., He,S. and Huang,D.W. (2012) The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics*, **100**, 265–270.
17. Saye,D.J., Ogunseitan,O., Sayler,G.S. and Miller,R.V. (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. *Appl. Environ. Microbiol.*, **53**, 987–995.
18. Raymond,J. and Segre,D. (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science (New York, N.Y.)*, **311**, 1764–1767.
19. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
20. Lehmann,J., Solomon,D., Kinyangi,J., Dathe,L., Wirick,S. and Jacobsen,C. (2008) Spatial complexity of soil organic matter forms at nanometre scales. *Nat. Geosci.*, **1**, 238–242.
21. Chastain,C.J., Failing,C.J., Manandhar,L., Zimmerman,M.A., Lakner,M.M. and Nguyen,T.H.T. (2011) Functional evolution of C4 pyruvate,orthophosphate dikinase. *J. Exp. Bot.*, **62**, 3083–3091.
22. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Micro*, **3**, 722–732.
23. Krupovic,M., Gonnet,M., Hania,W.B., Forterre,P. and Erauso,G. (2013) Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. *PLoS ONE*, **8**, e49044.