Title: The total microbiome functions in bacteria and fungi

Running title: The total functionality

Authors: Robert Starke[1], Petr Capek[2], Daniel Morais[1], Stephen J. Callister[3], Nico Jehmlich[4]

Affiliation:

[1]Laboratory of Environmental Microbiology, Institute of Microbiology of the CAS, Praha, Czech Republic, [2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, United States of America, [3]Biological Science Division, Pacific Northwest National Laboratory, Richland, Washington, United States of America, [4]Molecular Systems Biology, Helmholtz-Center for Environmental Research, UFZ Leipzig, Germany

Corresponding author:

Robert Starke, PhD

Phone: + 420 296 442 655

Email: robert.starke@biomed.cas.cz

Author Contributions

RS designed the study. DM performed the computational analysis. PC modelled the data. RS and NJ reviewed the analysis. The paper was written by RS, reviewed by all authors which approved the final version of the manuscript.

The authors declare no conflict of interest.

August 5, 2019

Significance

The functionality of and within a microbial community is generally inferred based on the taxonomic annotation of the organism. However, our understanding of functional diversity and how it relates to taxonomy is still limited. Here we predict the total microbiome functionality in bacteria and fungi on Earth using known and annotated protein-coding sequences in species accumulation curves. Our estimates reveal that the majority of functionality (>99%) could be assigned to yet unknown and rare functions, highlighting that our current knowledge is incomplete and functional inference is thus lackluster.

Robert Starke (on behalf of the authors)

The total microbiome functions in bacteria and fungi

Robert Starke[1], Petr Capek[2], Daniel Morais[1], Stephen J. Callister[3], Nico Jehmlich[4]

[1]Laboratory of Environmental Microbiology, Institute of Microbiology of the CAS, Praha, Czech Republic, [2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington, United States of America, [3]Biological Science Division, Pacific Northwest National Laboratory, Richland, Washington, United States of America, [4]Molecular Systems Biology, Helmholtz-Center for Environmental Research, UFZ Leipzig, Germany

Unveiling the relationship between phylogeny and function of the microbiome is crucial to determine its contribution to ecosystem functioning. However, while there is a considerable amount of information on microbial phylogenetic diversity, our understanding of its relationship to functional diversity is still scarce. Here we predicted the total microbiome functions of bacteria and fungi on Earth using the total known functions from level 3 of KEGG Orthology by modelling the increase of functions with increasing diversity of bacteria or fungi. For bacteria and fungi, the unsaturated model described the data significantly better (for both P < 2.2e-16), suggesting the presence of two types of functions. Widespread functions ubiquitous in every living organism that make up two thirds of our current knowledge of microbiome functions are separated from rare functions from specialized enzymes present in only a few species. Given previous estimates on species richness, we predicted a global total of 35.5 million functions in bacteria and 3.2 million in fungi; of which only 0.02% and 0.14% are known today. Our approach highlights the necessity of novel and more sophisticated methods to unveil the entirety of rare functions to fully understand the involvement of the microbiome in ecosystem functioning.

Ecosystem functioning is mediated by biochemical transformations performed by a community of microbes from every domain of life [1]. A wide range of ecosystem processes in the form of individual functions that contribute to the decomposition of organic carbon [2], deposition of recalcitrant carbon [3–6] and transformations of nitrogen and phosphorus [7,8] are performed by both bacteria and fungi. In every microbial community, multiple organisms from different taxonomic groups can play similar if not identical roles in ecosystem functioning and contribute with similar function, the so-called functional redundancy [9]. In fact, functional redundancy of certain functions was shown to be very high with several hundreds to thousands of different taxa expressing the same function within one habitat [10]. These functions can be statistically inferred based upon homology to experimentally characterized genes and proteins in specific organisms to find orthologs in other organisms present in a given

microbiome. This so-called ortholog annotation can be performed by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [11,12] with a similar coverage of bacteria and fungi compared to other phylogenomic databases such as eggNOG or OMA (**Figure S1**). KEGG is a database resource that integrates genomic, chemical and systemic function information commonly used to describe functional traits in microbiomes as it covers a wide range of functional classes (level 1 of KEGG) comprising cellular processes, environmental information processing, genetic information processing, human diseases, metabolism, organismal system and *brite* hierarchies. However, the bottleneck of describing microbiome functions is the low number of fully annotated genomes of microbial species as they are mostly limited to those that have undergone isolation and extensive characterization while the vast majority of organisms were not yet studied [13,14] and the annotation is based on the similarity to the genomes of the very few studied model organisms. Hence, microbiome functions are normally inferred based on the composition of the microbiome and its relation to functional parameters [15] as indicated by the frequent use of amplicon sequencing of DNA or RNA to target the 16S rRNA gene in prokaryotes and the ITS2 rRNA gene in eukaryotes that only describes the bacterial microbiome composition but not its function directly (45,657 publications in PubMed with the search term "16S rRNA sequencing" and 537 with "ITS2 rRNA sequencing" as of December 2019) opposed to other possible techniques that describe both phylogeny and function such as metagenomics (9,322), metatranscriptomics (534) or metaproteomics (442). Logically, the relationship between phylogeny and function is still uncertain, and the inference of microbiome functions based on phylogeny derived from sequencing is prone to limitations such as quantitative accuracy [16,17] or methodological bias introduced by DNA extraction, PCR, sequencing and the bioinformatic pipeline [18]. To unveil the relationship between phylogeny and function, and to predict the total microbiome functions of bacteria and fungi on Earth, we used species accumulation curves (SAC) [19] with the publicly available data from KO to identify the total possible diversity of KO functions. Given the difference in lifestyle as bacteria are confined to micro-environments as niche specialists present in matrix pores or bound to surfaces in biofilms [20] compared to filamentous fungi that typically sense a much larger volume of the environment, i.e. through mycorrhizal networks in soil [21], we hypothesise that (i) functional redundancy is higher in fungi and that (ii) both species richness and functional richness in the KO database are higher in bacteria. For this, we extracted all protein-coding genes with taxonomic annotation on species level and functional annotation on level 3 of KO from Uniprot (https://www.uniprot.org/, as of November 5th, 2018 [22]) as it comprises all available information from different annotation tools, estimated the SAC with ten random permutations, and fitted the resulting data to a saturated and an unsaturated model.

In line with our first hypothesis of niche-specialised bacteria [20], fungi were indeed found to be functionally more redundant than bacteria as indicated by the median of the relative amount of bacterial (3±21%) and fungal organisms (46±22%) that share one randomly chosen function (**Figure 1a**). The number of KO functions shared between the two kingdoms (1,175) was smaller than the functions unique to either bacteria (8,597) or fungi (3,420) (**Figure 1b**), highlighting not only that the difference in lifestyle between the two kingdoms guides a difference in particular functions but also that two different KO functions could perform functionally similar processes. In fact and among others, the malate dehydrogenases (K00024-K00029) all perform the same function. Even though fungi were functionally more redundant than bacteria, one fungal species contributed 1.65±3.42 rare KO functions to the total functional richness as compared to only 0.18±0.88 in bacteria (**Figure 1c**), presumably due to the larger eukaryotic genomes derived from higher morphological complexity [23,24]. In addition to that, if each fungal genome adds more phylogenetic diversity, it could potentially also add more functional diversity as a higher taxonomic redundancy and closely relatedness exist in the bacterial genomes of the database (75.2% in bacteria compared to only 38.3% in fungi). In fact, the average fungi in the database contained more KO functions (1,918.8±1,709.2) than bacteria (1,404.5±974.7). The SAC were fitted to a saturated model (**Equation 1**) with the hypothesis of limited microbiome functions where the functional richness plateaus despite further increasing species richness. Otherwise, the unsaturated model (**Equation 2**) with the hypothesis of unlimited microbiome functions is the increase of functional richness with increasing species richness without ever reaching a plateau. The difference between the two is the addition of an additive term in the unsaturated model. The unsaturated model of both bacteria ($AIC_{bacteria}$ = 51,495.75) and fungi (AIC_{fungi} = 1,721.39) fitted the SAC significantly better (P < 2.2E-16) than the saturated model ($AIC_{bacteria}$ = 59,174.37; AIC_{fungi} = 2,327.99). The unsaturated model is described by the maximum functional richness f_{max}, the accretion rate of functions with an increasing number of species A_f and the constant of the additive term k. In line with our second hypothesis, we found more than double f_{max} in bacteria than in fungi and an A_f of only two fungal species but 22 in bacteria (**Figure 2**). Similarly, the unsaturated model described the SAC significantly better (for all P < 2.2E-16) when only 90%, 80%, 70%, 60% and 50% of the bacterial (**Table 1**) and fungal species were used (**Table 2**). The better fit of the unsaturated models inferred the presence of two types of microbiome functions within KO that are similarly present in bacteria and fungi based on the frequency of appearance in the database. On the one hand, widespread KO functions rapidly increase with the number of species and are ubiquitously abundant in every living organism. Among those, enzymes such as the glyceraldehyde 3-phosphate dehydrogenase *gapA* (EC 1.2.1.12, K00134) found in 79.9% of

bacteria and 59.2% of fungi in the KO database are counted. As comparison, the genome-based database comprises of in total 67,631 genomes with that protein found in archaea, bacteria and eukaryotes. The number of widespread KO functions is limited and the majority has been identified thus far amounting to, in total, 6,397 in bacteria and 3,173 in fungi. Logically, the difference in lifestyle between bacteria and fungi must influence widespread KO functions since the two kingdoms shared only 1,175 KO functions. On the other hand, rare KO functions, separated from widespread KO functions by the critical point, increase at a much slower rate with the number of species but require time and the evolution of "dead ends", i.e. species that were unable to evolve a particular function. Among those, specialised enzymes such as the D-nopaline dehydrogenase (EC 1.5.1.19, K00296) are only found in one bacterial taxonomic unit in the protein-based database. As comparison, the genome-based database comprises of in total eight organisms with that protein; all of which are bacteria. Given the estimated number of species of 100 million bacteria [25,26] and 1.5 million fungi [27] on Earth and assuming that the yet unknown microbiome functions are indeed rare KO functions, the propagation of the unsaturated model predicted the total microbiome functions on Earth to be 3,180,114±48,582 (with 3,084,709-3,275,474 as 95% confidence intervals) in fungi and 35,487,366±216,109 (with 35,063,316-35,910,407 as 95% confidence intervals) in bacteria. For bacteria, the estimate is likely imprecise as the confidence intervals of the prediction when only subsets of the SAC were used did not overlap (**Table 1**) whereas, otherwise for fungi, all confidence intervals for the prediction overlapped and the estimate of microbiome functions is therefore likely precise (**Table 2**). At large, based on our estimates, our understanding of microbiome functions is limited to 0.14% in fungi (4.5 thousand from 3.2 million) and 0.02% in bacteria (9.7 thousand from 35.5 million). Of those, the majority belong to ubiquitously present widespread KO functions. SAC with error bars as black area of the KO functions of functional classes (level 1 of KO) revealed that our current knowledge is differently divided among bacteria and fungi. Since the saturation means that most if not all of the functions are covered, cellular processes, environmental information and human disease in bacteria (**Figure S2**) together with cellular processes and genetic information in fungi are well-understood as of today (**Figure S3**). Otherwise, unsaturated relationships were found for genetic information, metabolism, organismal system, *brite* hierarchies and functions not included in the annotation of the two databases *pathway* or *brite* in bacteria together with environmental information, human disease, metabolism, organismal system, *brite* hierarchies and functions not included in the annotation of the two databases *pathway* or *brite*. The two latter represent the functional classes on which future research must focus to reach the saturation in total microbiome functions as our current knowledge is incomplete.

Taken together, we suggest based on known protein-coding sequences the presence of two types of microbiome functions in bacteria and fungi; widespread and rare KO functions. Our predictions revealed millions more yet unknown rare KO functions that, logically, can only be unveiled by novel and more sophisticated methods. However, due to the vast amount of yet unknown functions, it is questionable if the relationship between phylogeny and function is in fact explained by an unsaturated model, if only two types of KO functions (widespread and rare) exist and if it is similar when different phylogenomic tools for the functional annotation or genome-centric databases are used. In fact, as one isolated genome comprised of on average only 171 KO functions it is further questionable if rare and widespread KO functions is a feasible discrimination of functions or if the rather smooth transitions require more groups of functions such as less rare and super rare. Moving forward, estimates on total microbiome functions must include a higher coverage of species richness and functional richness through the discovery of both novel species and novel functions.

Materials and Methods

Metadata collection of the total known microbiome functions

The data used to quantify functional richness and species richness was downloaded from Uniprot (https://www.uniprot.org/, as of November 5th, 2018 [22]) using the search parameters “bacteria” and “fungi”. To standardise the overall classification system, we used only genes containing functional annotation from the level 3 of KEGG Orthology (KO) [11,12] with the search parameter “KEGG”. Level 3 of KO was used to provide sufficient depth to estimate the species accumulation curves as level 1 comprises of eight and level 2 of 54 categories, which would represent the maximum functional diversity. In total, our database comprised of 15,411,107 non-redundant protein-coding genes, which are related to 9,755 KO functions assigned to 4,092 bacteria, and 4,578 KO functions assigned to 196 fungi. A different organism was considered as taxonomic unit regardless of the depth of annotation. Of the 4,092 bacterial taxonomic units, 3,003 and 18 were annotated on the strain or the genus level, respectively. In total, the bacterial taxonomic units comprise of 1,017 bacterial genera. Of the 196 bacterial taxonomic units, 177 and 0 were annotated on the strain or the genus level, respectively. In total, the fungal taxonomic units comprise of 121 fungal genera. In every taxonomic unit, any one function can be encoded by multiple copies of a gene, the so-called gene redundancy [28]. Otherwise, one gene can have multiple functions, the so-called pleiotropy [29]. For our estimates, we only considered one gene for each function per taxonomic unit, disregarding both gene redundancy and pleiotropy. The functional redundancy in bacteria and fungi was determined as the median of the

functional redundancy of every individual function. For each individual functions, the number of bacterial or fungal species that share this specific function were estimated relative to the total number of species in each database. To determine the unique KO functions per kingdom, the number of KO functions unique to bacteria or fungi were extracted together with how many KO functions are shared between the two kingdoms. To determine the unique KO functions per species, the number of KO functions that are present in only one bacterial or fungal species were counted.

Species accumulation curves

For bacteria and fungi separately, species were randomly added in intervals of one up to the maximum species richness within each domain with ten permutations per step using the function *specaccum* from the R package *vegan* [30]. Similarly the species accumulation curves (SAC) were permuted for the functional classes (level 1 of KO) of KO functions. SAC of the database permutation from bacteria or fungi were then fitted to a saturated (**Equation 1**) and an unsaturated model (**Equation 2**) with the critical point estimated by the term $3A_f$ as previously described [31]. The fit of the models was compared by the analysis of variance (ANOVA) and the Akaike's An Information Criterion (AIC) [32] with a penalty per parameter set to *k* equals two. The total number of KO functions in bacteria and fungi on Earth was predicted using the global species richness estimates of 100 million bacteria [25,26] and 1.5 million fungi [27], and the function *predictNLS* in the R package *propagate* [33]. Lastly, the species richness of the bacterial and fungal SAC was subsampled to 90%, 80%, 70%, 60% and 50% of the maximum species richness and again fitted to a saturated and an unsaturated model as described above to validate the precision of the prediction.

Eq. 1: $$Functional\ richness = \frac{f_{max} * [Species\ richness]}{A_f + [Species\ richness]}$$

Eq. 2: $$Function\ richness = \frac{f_{max} * [Species\ richness]}{A_f + [Species\ richness]} + k * [Species\ richness]$$

Here, f_{max} is the maximum functional richness, A_f the accretion rate of functions with an increasing number of species and *k* the constant of the additive term.

References

[1] C.R. Woese, O. Kandler, M.L. Wheelis, Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya., Proc. Natl. Acad. Sci. (1990). doi:10.1073/pnas.87.12.4576.

[2] S.F. Chapin, P.A. Matson, P.M. Vitousek, Principles of terrestrial ecosystem ecology, 2012.

doi:10.1007/978-1-4419-9504-9.

[3] K.E. Clemmensen, A. Bahr, O. Ovaskainen, A. Dahlberg, A. Ekblad, H. Wallander, J. Stenlid, R.D. Finlay, D.A. Wardle, B.D. Lindahl, Roots and associated fungi drive long-term carbon sequestration in boreal forest, Science (80-.). (2013). doi:10.1126/science.1231923.

[4] I. Kögel-Knabner, The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter: Fourteen years on, Soil Biol. Biochem. (2017). doi:10.1016/j.soilbio.2016.08.011.

[5] M.W.I. Schmidt, M.S. Torn, S. Abiven, T. Dittmar, G. Guggenberger, I.A. Janssens, M. Kleber, I. Kögel-Knabner, J. Lehmann, D.A.C. Manning, P. Nannipieri, D.P. Rasse, S. Weiner, S.E. Trumbore, Persistence of soil organic matter as an ecosystem property, Nature. (2011). doi:10.1038/nature10386.

[6] J. Six, S.D. Frey, R.K. Thiet, K.M. Batten, Bacterial and fungal contributions to carbon sequestration in agroecosystems, Soil Sci. Soc. Am. J. (2006). doi:10.2136/sssaj2004.0347.

[7] R.L. Sinsabaugh, R.K. Antibus, A.E. Linkins, C.A. McClaugherty, L. Rayburn, D. Repert, T. Weiland, Wood Decomposition: Nitrogen and Phosphorus Dynamics in Relation to Extracellular Enzyme Activity, Ecology. (1993). doi:10.2307/1940086.

[8] R.L. Sinsabaugh, Enzymic analysis of microbial pattern and process, Biol. Fertil. Soils. (1994). doi:10.1007/BF00418675.

[9] S.P. Hubbell, Neutral theory in community ecology and the hypothesis of functional equivalence, Funct. Ecol. (2005). doi:10.1111/j.0269-8463.2005.00965.x.

[10] L. Žifčáková, T. Větrovský, V. Lombard, B. Henrissat, A. Howe, P. Baldrian, Feed in summer, rest in winter: microbial carbon utilization in forest topsoil, Microbiome. (2017). doi:10.1186/s40168-017-0340-0.

[11] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation, Nucleic Acids Res. (2016). doi:10.1093/nar/gkv1070.

[12] M. Kanehisa, Y. Sato, K. Morishima, BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences, J. Mol. Biol. (2016). doi:10.1016/j.jmb.2015.11.006.

[13] V.H.T. Pham, J. Kim, Cultivation of unculturable soil bacteria, Trends Biotechnol. (2012). doi:10.1016/j.tibtech.2012.05.007.

[14] A.C. Martiny, High proportions of bacteria are culturable across major biomes, ISME J. (2019).

[15] R. Starke, N. Jehmlich, F. Bastida, Using proteins to study how microbes contribute to soil ecosystem services: The current state and future perspectives of soil metaproteomics, J. Proteomics. (2018). doi:https://doi.org/10.1016/j.jprot.2018.11.011.

[16] B.A. Hungate, R.L. Mau, E. Schwartz, J. Gregory Caporaso, P. Dijkstra, N. van Gestel, B.J. Koch, C.M. Liu, T.A. McHugh, J.C. Marks, E.M. Morrissey, L.B. Price, Quantitative microbial ecology through stable isotope probing, Appl. Environ. Microbiol. (2015). doi:10.1128/AEM.02280-15.

[17] L.M. Feinstein, J.S. Woo, C.B. Blackwood, Assessment of bias associated with incomplete extraction of microbial DNA from soil, Appl. Environ. Microbiol. (2009). doi:10.1128/AEM.00120-09.

[18] M.R. McLaren, A.D. Willis, B.J. Callahan, Consistent and correctable bias in metagenomic sequencing experiments, BioRxiv. (2019). doi:http://dx.doi.org/10.1101/559831doi.

[19] N.J. Gotelli, R.K. Colwell, Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness, Ecol. Lett. (2001). doi:10.1046/j.1461-0248.2001.00230.x.

[20] M.E. Davey, G.A. O'toole, Microbial Biofilms: from Ecology to Molecular Genetics, Microbiol. Mol. Biol. Rev. (2000). doi:10.1128/MMBR.64.4.847-867.2000.

[21] S.W. Simard, K.J. Beiler, M.A. Bingham, J.R. Deslippe, L.J. Philip, F.P. Teste, Mycorrhizal networks: Mechanisms, ecology and modelling, Fungal Biol. Rev. (2012). doi:10.1016/j.fbr.2012.01.001.

[22] A. Bateman, M.J. Martin, C. O'Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-AJee, A. Cowley, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, L.G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. ArgoudPuy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas,

E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A.L. Veuthey, C.H. Wu, C.N. Arighi, L. Arminski, C. Chen, Y. Chen, J.S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D.A. Natale, K. Ross, C.R. Vinayaka, Q. Wang, Y. Wang, L.S. Yeh, J. Zhang, UniProt: The universal protein knowledgebase, Nucleic Acids Res. (2017). doi:10.1093/nar/gkw1099.

[23] J.I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M.R. Lewis, M. Maruf, C.A. Hutchison, H.O. Smith, J.C. Venter, Essential genes of a minimal bacterium, Proc. Natl. Acad. Sci. (2006). doi:10.1073/pnas.0510013103.

[24] T.R. Gregory, J.A. Nicol, H. Tamm, B. Kullman, K. Kullman, I.J. Leitch, B.G. Murray, D.F. Kapraun, J. Greilhuber, M.D. Bennett, Eukaryotic genome size databases, Nucleic Acids Res. (2007). doi:10.1093/nar/gkl828.

[25] P.D. Schloss, J. Handelsman, Toward a census of bacteria in soil, PLoS Comput. Biol. (2006). doi:10.1371/journal.pcbi.0020092.

[26] T.P. Curtis, W.T. Sloan, J.W. Scannell, Estimating prokaryotic diversity and its limits, Proc. Natl. Acad. Sci. (2002). doi:10.1073/pnas.142680199.

[27] D.L. Hawksworth, The magnitude of fungal diversity: The 1.5 million species estimate revisited, in: Mycol. Res., 2001. doi:10.1017/S0953756201004725.

[28] J.M. Pérez-Pérez, H. Candela, J.L. Micol, Understanding synergy in genetic interactions, Trends Genet. (2009). doi:10.1016/j.tig.2009.06.004.

[29] F.W. Stearns, One hundred years of pleiotropy: A retrospective, Genetics. (2010). doi:10.1534/genetics.110.122549.

[30] J. Oksanen, F.G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P.R. Minchin, R.B. O'Hara, G.L. Simpson, P. Solymos, M.H.H. Stevens, E. Szoecs, H. Wagner, vegan: Community Ecology Package. R package version 2.5-2, CRAN R. (2018). doi:ISBN 0-387-95457-0.

[31] P. Čapek, P. Kotas, S. Manzoni, H. Šantrůčková, Drivers of phosphorus limitation across soil microbial communities, Funct. Ecol. (2016). doi:10.1111/1365-2435.12650.

[32] P. V. Bertrand, Y. Sakamoto, M. Ishiguro, G. Kitagawa, Akaike Information Criterion Statistics., J. R. Stat. Soc. Ser. A (Statistics Soc. (2006). doi:10.2307/2983028.

[33] A.-N. Spiess, propagate: Propagation of Uncertainty, R Packag. Version 1.0-6. (2018).

Figures

Figure 1: Total share of KO functions within bacteria in blue and fungi in red as functional redundancy relative to the total number of species in each kingdom in the database (a), the number of KO functions unique to bacteria in blue or fungi in red and those common to both in grey (b), and the number of KO functions unique to one species within bacteria in blue and fungi in red (c).

Figure 2: The unsaturated model of the species accumulation curves as grey points with error bars for the total known microbiome functions derived from the KO database by ten random permutations for every one species richness of bacteria in blue and fungi in red with 95% confidence intervals. The maximum functional richness at infinite species richness is represented by f_{max}, A_f is the accretion rate of functions with increasing number of species, and k is the constant of the additive term. Significance of the parameter estimates are indicated by asterisks (*** equals $P < 0.001$).

Tables

Table 1: The fit of the saturated and the unsaturated model to the different amount of coverage of the species accumulation curve indicated by the Akaike's An Information Criterion (AIC), the P-value at which the unsaturated model gives a better fit than the saturated model, and the mean prediction with standard deviation (SD) and 95% confidence intervals (CI) at 100 million bacterial species.

Table 2: The fit of the saturated and the unsaturated model to the different amount of coverage of the species accumulation curve indicated by the Akaike's An Information Criterion (AIC), the P-value at which the unsaturated model gives a better fit than the saturated model, and the mean prediction with standard deviation (SD) and 95% confidence intervals (CI) at 1.5 million fungal species.

Acknowledgements

RS thanks the Czech Science Foundation for the project 18-25706S. The authors thank Iñaki Odriozola for the determination of the species accumulation curves and Petr Baldrian, Felipe Bastida and Richard Allen White III for critical revision.

Table 1

Click here to download Table: Table 1.xlsx

Coverage (%)	AIC_{unsat}	AIC_{sat}	P-value	Prediction	SD	Lower CI	Higher CI
100	51,495.75	59,174.37	2.20E-16	35,487,366	216,109	35,063,316	35,910,407
90	46,190.99	53,080.57	2.20E-16	38,841,560	249,651	38,352,496	39,330,613
80	40,893.62	47,028.51	2.20E-16	43,063,588	292,920	42,488,069	43,639,301
70	35,596.20	40,998.98	2.20E-16	48,487,877	350,996	47,801,109	49,177,219
60	30,379.72	34,969.69	2.20E-16	55,263,177	436,089	54,408,161	56,119,307
50	25,234.50	28,940.99	2.20E-16	64,176,709	574,565	63,049,975	65,302,445

Table 2

Click here to download Table: Table 2.xlsx

Coverage (%)	AIC_{unsat}	AIC_{sat}	P-value	Prediction	SD	Lower CI	Higher CI
100	1,721.39	2,327.99	2.20E-16	3,180,114	48,582	3,084,709	3,275,474
90	1,543.28	2,049.80	2.20E-16	3,304,410	59,462	3,187,645	3,421,034
80	1,386.50	1,799.95	2.20E-16	3,415,251	74,457	3,269,025	3,561,628
70	1,228.09	1,531.89	2.20E-16	3,411,586	100,613	3,214,075	3,609,491
60	1,074.71	1,282.71	2.20E-16	3,368,669	139,977	3,094,092	3,643,280
50	910.37	1,033.48	2.20E-16	3,345,225	212,630	2,927,000	3,762,746

Figure 1

Click here to download high resolution image

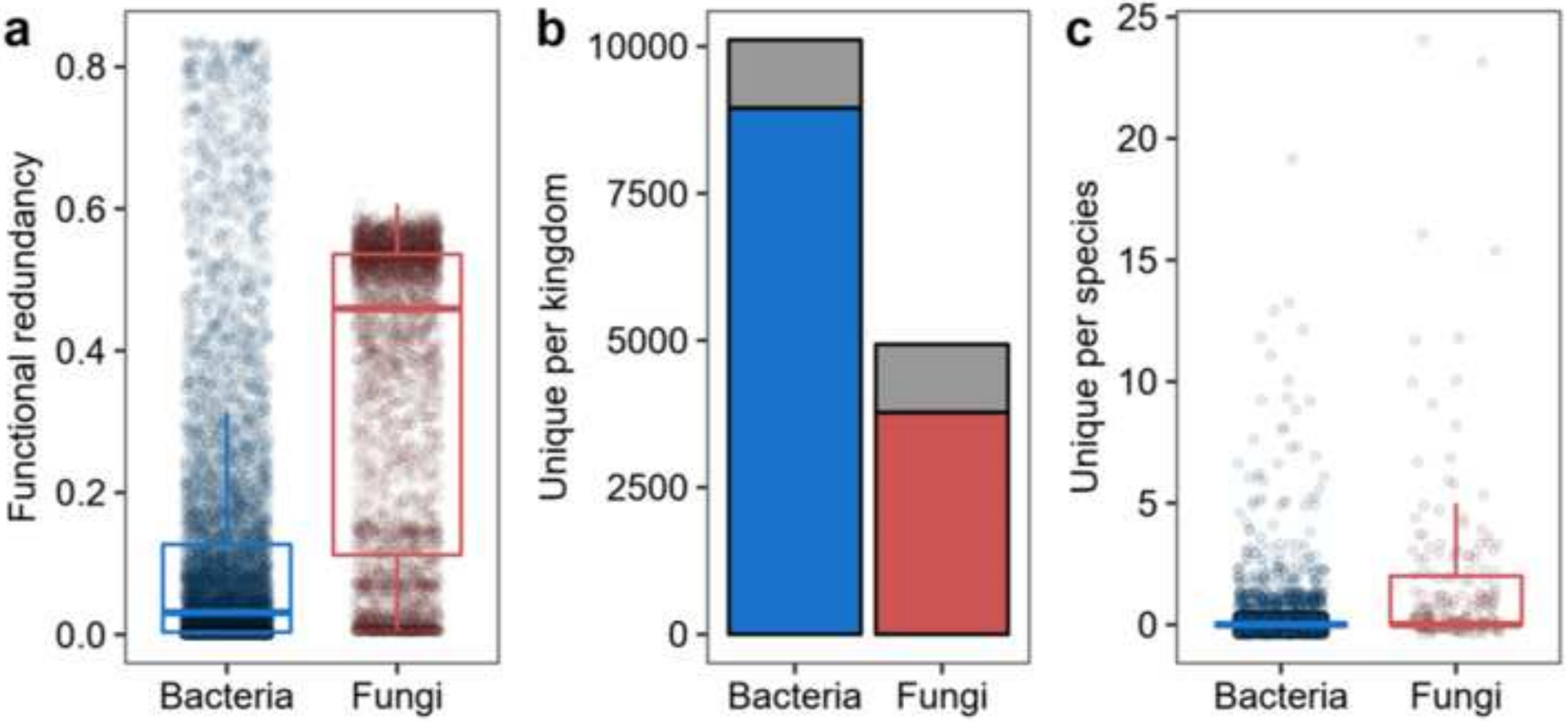

Figure 2

Click here to download high resolution image

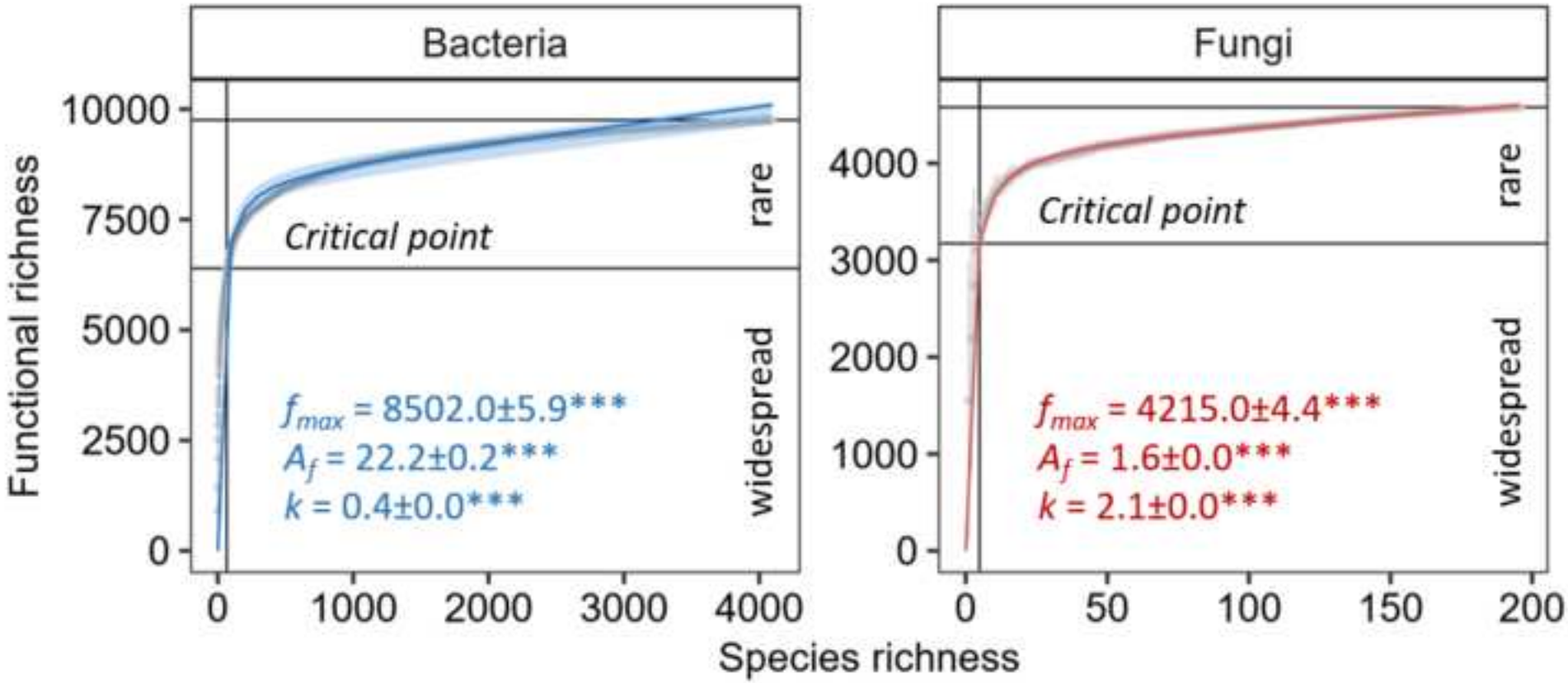

Figure S1

Click here to download Supplementary material: Figure S1 .png

Figure S2

Click here to download Supplementary material: Figure S2 - Bacteria.png

Figure S3

Click here to download Supplementary material: Figure S3 - Fungi.png