Function and functional redundancy in microbial systems

Function and functional redundancy in microbial systems

Stilianos Louca[1,2,*], Martin F. Polz[3], Florent Mazel[1,4,5], Michaeline B. N. Albright[6], Julie A. Huber[7], Mary I. O'Connor[1,2], Martin Ackermann[8,9], Aria S. Hahn[10], Diane S. Srivastava[1,2], Sean A. Crowe[10,11,12], Michael Doebeli[1,2,13] & Laura Wegener Parfrey[1,2,4]

[1]*Biodiversity Research Centre, University of British Columbia, Canada*
[2]*Department of Zoology, University of British Columbia, Canada*
[3]*Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, USA*
[4]*Department of Botany, University of British Columbia, Canada*
[5]*Department of Biological Sciences, Simon Fraser University, Canada*
[6]*Department of Ecology and Evolutionary Biology, University of California, Irvine, USA*
[7]*Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA*
[8]*Department of Environmental Systems Science, Eidgenössische Technische Hochschule Zürich, Switzerland*
[9]*Department of Environmental Microbiology, Eawag, Switzerland*
[10]*Ecosystem Services, Commercialization Platforms, and Entrepreneurship (ECOSCOPE) Training Program, University of British Columbia, Canada*
[11]*Department of Microbiology and Immunology, University of British Columbia, Canada*
[12]*Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Canada*
[13]*Department of Mathematics, University of British Columbia, Canada*
[*]Corresponding author: louca.research@gmail.com

Abstract

Microbial communities often exhibit incredible taxonomic diversity, raising questions regarding the mechanisms enabling species coexistence and the role of this diversity in community functioning. On the one hand, many coexisting but taxonomically distinct microorganisms can encode the same energy-yielding metabolic functions, and this functional redundancy contrasts with the expectation that species should occupy distinct metabolic niches. On the other hand, the identity of taxa encoding each function can vary substantially across space or time with little effect on the function, and this taxonomic variability is frequently thought to result from ecological drift between equivalent organisms. Here we synthesize the powerful paradigm emerging from these two patterns, connecting the roles of function, functional redundancy and taxonomy in microbial systems. We conclude that both patterns are unlikely the result of ecological drift, but are inevitable emergent properties of open microbial systems resulting mainly from biotic interactions and environmental and spatial processes.

Keywords: *functional redundancy; metabolic niche; microbial community; biogeochemistry*

Introduction

Microorganisms are the most ancient, the most phylogenetically diverse and the most widespread form of life on Earth[1]. A single gram of soil can harbor thousands of microbial species[2]. The metabolic and biosynthetic versatility of microorganisms is equally impressive: the number of discovered prokaryotic protein-coding genes is orders of magnitude greater than those of all plants and animals combined[3,4]. Metabolic pathways encoded in microorganisms drive the bulk of elemental cycles in most ecosystems, shaping Earth's surface chemistry over billions of years[5]. Yet, our mechanistic understanding of microbial systems (microbial communities and coupled abiotic physicochemical processes) remains in its infancy. The enormous microbial diversity presents major challenges to modeling microbial systems and to explaining patterns of community variation across space and time. Moreover, many questions in ecosystem ecology and biogeochemistry require knowledge of the variation in microbial metabolic functions, rather than just taxonomic composition. Despite the high microbial diversity, most major biogeochemical reactions are driven by a limited set of energy-transducing metabolic pathways, each of which is found in a variety of microbial clades[5]. Functional community profiling — describing communities in terms of metabolic functions of interest — can simplify microbial systems to a level permissible to mathematical modeling and can reveal patterns of community structuring across environmental gradients[6–9]. A wave of recent studies in a multitude of environments, ranging from soil to the ocean and to the human gut[9–14], suggest that certain metabolic functions are strongly coupled to certain environmental factors and can, in many cases, appear decoupled from the species assemblages associated with them at a given place and time. Quantification of microbial diversity involved in various metabolic functions also revealed that communities typically exhibit high "functional redundancy" with respect to a multitude of functions, in the sense that each metabolic function can be performed by multiple coexisting, taxonomically distinct organisms[9,13–18]. Much confusion exists currently over the meaning of these patterns, however their proper interpretation is paramount to understanding the mechanisms controlling microbial community composition and function. In this synthesis we provide interpretations for these patterns and discuss the powerful paradigm emerging from them, uniting the roles that function, functional redundancy and taxonomy play in shaping microbial systems.

Disentangling function from taxonomy in microbial communities

In one of the first comparative metagenomic surveys of microbial communities, Tringe *et al.* [19] showed that functional profiles (in terms of the genes found in communities) were highly correlated with the type of sampled environment (seawater vs soil, etc), suggesting that the environment selected for specific functions. A subsequent comparison of gut microbiota between different human hosts revealed that the taxonomic composition of microbiomes varied strongly across hosts while their community gene content was strongly conserved[11]. Similarly, in a survey of bacterial communities on the macroalgae *Ulva australis*, communities appeared to be assembled on the basis of functional genes rather than species[12]. These findings suggest that alternative microbial assemblages can exhibit similar community gene profiles selected by their environment. In line with this perspective, a recent study of bacterial and archaeal communities inside the foliage "tanks"

of bromeliad plants[14] found that the functional composition of communities (in terms of genes involved in various energy-transducing functions; Fig. 1C,D) was highly conserved across bromeliads. In contrast, the taxa associated with each functional group (i.e., capable of performing a specific metabolic function) varied strongly between bromeliads[14], regardless of the taxonomic resolution used (up to class level; Fig. 1A,B). Hence, the taxonomic composition within functional groups must have been shaped by additional factors that are distinct from the factors shaping the functional structure of communities, that is, taxonomic composition and functional composition (genetic potential) appeared "decoupled". A similar decoupling between various metabolic functions and taxonomic community composition has been repeatedly observed in experiments with bioreactors, such as for nitrogen removal or methane production, where a high variation in community composition over time coincided with stable bioreactor performance[10,15,17,20–23]. In the following, we discuss conditions and mechanisms that could promote this frequently observed phenomenon.

The contrast between stable functional composition and variable taxonomic composition seen in the aforementioned studies[10–12,14,15,17,20–23] reflects a weak association between many functions and prokaryotic phylogeny. Indeed, a large fraction of metabolic functions are not monophyletic[24,25], that is, no single clade is the sole representative for any of those functions. Thus, while the phylogenetic placement of an organism in principle determines its metabolic potential (given sufficient resolution and/or trait conservatism), the reverse need not be true, i.e. metabolic potential is not necessarily indicative of a specific clade (a notable exception being oxygenic photosynthesis[25]). Adaptive loss of function or genome streamlining[26], convergent evolution, and horizontal gene transfer[27] all erode the phylogenetic signal of many traits[24]. Horizontal gene transfer also leads to low genetic linkage of traits within genomes and hence to reassortment of traits between genomes[28]. Some *Escherichia coli* strains, for example, overlap by less than 40% in their protein-coding genes[29]. The phylogenetic scale at which functions are conserved varies strongly between functions[25,30], and even for single functions phylogenetic conservatism can vary between clades (Figs. 2A,B). For example, the ability to respire sulfate is shared by all cultured members of the families *Desulfobacteraceae*, *Desulfohalobiaceae* and *Desulfomicrobiaceae*, but only by a subset of the genus *Archaeoglobus*[31]. Because a given metabolic function may be present and conserved within distinct clades of varying depths, there exists no taxonomic resolution at which taxa either always or never exhibit that function. Consequently, there exists no single taxonomic resolution at which taxonomic variation unambiguously reflects functional variation, and at which environmental selection of certain functions (e.g., the presence of oxygen selecting for aerobes) unambiguously translates to a selection of specific taxa.

A partial to complete decoupling of certain functions from particular taxonomic assemblages appears to be almost inevitable, given that the same functions can be performed by alternative taxa (Fig. 2C). Nutrient supply rates, irradiance, geochemical gradients, environmental transport processes and stoichiometric balances between pathways across organisms can strongly constrain reaction rates, and energy yields from metabolic pathways further affect the possible growth rates of functional groups[8,32,33]. While each function can of course only be performed by certain taxa, the aforementioned factors may exert little control over which of those taxa perform each function in a particular situation. Reciprocally, bulk biochemical flux rates may exhibit low sensitivity to taxonomic changes within functional groups over space or time. In support of this

interpretation, a global biogeographical study in soil found that abiotic soil characteristics largely explained the variation in the abundances of nitrogen cycling pathways, but only weakly explained the taxonomic composition within the corresponding functional groups[13]. Similar observations have also been made for a broad range of metabolic functions across the global ocean[6,9]. Reciprocally, a recent meta-analysis found that an inclusion of taxonomic community composition, in addition to environmental variables, as predictors of carbon and nitrogen process rates only improved predictive power in 29% of considered studies, with the adjusted R^2 only increasing from 0.56 to 0.65 on average[34]. Which functions are strongly controlled by the environment — thus being less sensitive to taxonomic variation — depends on the type of ecosystem, and in particular on the redox disequilibria available for energy gain and the physical-chemical boundary conditions. In experiments, broadly distributed functions such as respiration, overall carbon catabolism and biomass production often appear more resistant to changes in taxonomic community composition or diversity, than narrow functions such as the degradation of specific compounds[35–38]. A possible reason for this pattern is that broad functions may be more functionally redundant and thus better buffered against taxonomic shifts caused by biotic or abiotic disturbance[39]. Thermodynamically favored endpoints of linear catabolic pathways may be less sensitive to taxonomic variation than individual intermediate steps that can be performed in alternative ways. For example, models for methanogenic bioreactors fed continuously with glucose suggest that the relative flux rates through "alternative" catabolic pathways (e.g., the various alternative routes from glucose to volatile fatty acids and eventually to methane; Fig. 3A) may be less stable in the face of taxonomic shifts, than the overall methane production rate[40].

Some studies have observed strong correlations between functional and taxonomic community composition, for example across strong redox gradients[41]. We emphasize that when environmental conditions vary, selection for specific metabolic functions will generally cause changes in taxonomic community composition in addition to the taxonomic variation occurring within functional groups. Therefore, when comparing communities over space or time, the correlation between functional and taxonomic community composition will depend on the relative importance of mechanisms selecting for specific functions versus mechanisms causing variation within functional groups (discussed below), as well as on the phylogenetic distribution of those functions.

We point out that functional community structure can in principle be defined with respect to any arbitrary set of functions (and observed spatiotemporal patterns will depend on the choice of functions), although particular attention is typically devoted to energy-transducing metabolic functions involved in major elemental cycles[5] or of particular industrial importance[17]. We also mention that some authors define "functional response groups", i.e. organisms that respond similarly to specific environmental factors, and distinguish those from "functional effect groups", i.e. organisms with a similar effect on specific ecosystem functions[42]. Here we avoid this terminology, however, partly because (metabolic) functional groups (*sensu* this synthesis) can usually be seen both as effect groups and as response groups. Further, as discussed above, metabolic function and taxonomic variation within functional groups constitute complementary and disentangled facets of many microbial systems, and can yield insight into markedly different processes[9,14].

Functional redundancy is an omnipresent feature of open microbial systems

A large fraction of metabolic genes appeared early in Earth's history[27] and, as discussed above, over geological time propagated into multiple microbial clades[5,27]. Today, at global scales, most metabolic functions can be potentially performed by a wide range of extant taxa. More strikingly, even at local scales, the enumeration of taxa associated with each metabolic function, either by taxonomic binning of metagenomic sequences[13] or by functional classification of taxa[9], often reveals a coexistence of multiple distinct organisms capable of performing similar metabolic functions[9,13–18,39,43]. For example, hundreds of microorganisms capable of hydrogen oxidation can coexist in groundwater[18], and hundreds of oxygenic photoautotrophs can coexist in the ocean surface[9,44]. In a sub-seafloor aquifer, dozens of genomes had the potential to oxidize sulfide for energy and at least 15 genomes were capable of complete denitrification[43]. In methanogenic digesters cellulose hydrolysis can be concurrently performed by dozens of different organisms[17]. In nitrifying bioreactors, typically multiple ammonia oxidizing bacteria coexist and exhibit variable relative abundances over time[15,16]. Functional redundancy, it seems, is a common aspect of many microbial systems. That said, it is clear that the degree of functional redundancy in any given system depends on the function considered. In the sunlit and oxygen-rich ocean surface, for example, photoautotrophy and oxygen respiration are generally much more redundant than sulfate respiration and methanogenesis[9].

Functional community structure (and thus functional redundancy) could in principle be defined at various levels of detail, for example further differentiating functions based on reaction kinetics. Some authors consider organisms functionally redundant only if they can readily replace each other due to high ecological similarity[45], although the same authors acknowledge that this criterion is rarely met in practice. Other authors only define organisms as redundant if they are able to perform a function at the same rate, given the same environmental conditions[46]. The latter requirement can be hard to test in practice, and sequencing data rarely allow inference of enzyme kinetics beyond the type of reactions potentially catalyzed. The practicality of such a definition is also limited by the fact that the metabolic activity of a population depends on the overall community state, such as the presence of syntrophic partners, phages or bacteriocins. Moreover, bulk process rates could be largely constrained by physicochemical characteristics of the environment, such as spatial transport rates across sediment columns or substrate supply rates in bioreactors. Populations of distinct taxa with different reaction kinetics may thus induce different or similar biochemical flux rates, depending on the detailed environmental setup and the current state of the community. We thus argue that a definition of functional redundancy indicating the mere ability of multiple distinct organisms to perform a specific function, as used in this synthesis and as observed in many environments, is of greater practical relevance than the more stringent definitions by Fuhrman *et al.*[45] or by Allison *et al.*[46]. For example, functional redundancy (*sensu* this synthesis) is often linked to the stability of functions against environmental perturbations[39] and, as we discuss below, can yield insight into important community processes.

Mechanisms promoting functional redundancy

A high functional redundancy with respect to energy-transducing metabolic pathways has long been observed in macrobial communities[47]. Almost all plants, for example, share a common metabolic niche — they are oxygenic photoautotrophs. In microbes and macrobes alike, functional redundancy indicates that additional factors beyond the mere availability of different energy sources must be controlling diversity. Indeed, Tilman's classical competition theory[48,49] asserts that at steady state and in a well-mixed system any given resource — such as an electron donor or acceptor — can only be limiting to at most a single persisting population. This population will be the one that can maintain a steady size at the lowest possible resource level, since all other populations are either outcompeted or limited by a different resource. While steady state and perfect mixing arguably represent an idealized situation, Tilman's competition theory provides a benchmark — a minimum expectation — to which observed diversity can be compared. The apparent disconnect between the theoretical expectation of one species persisting per limiting resource, and the observed diversity of life has been explained for macrobial communities in several ways[47]. First, spatial and temporal heterogeneity either in the identity of the limiting resource or in environmental conditions, combined with response differences between species, may effectively create multiple niches. Second, competitive exclusion can be disrupted by biotic interactions such as predation, or be offset by dispersal from a regional pool. Importantly, species may show tradeoffs between traits involved in resource competition, and traits involved in environmental tolerance, predator resistance or dispersal[47].

Similarly to macroorganisms, functional redundancy in microbial communities may be promoted by differentiation along other niche axes than just metabolic resources, including differences in their response to environmental perturbations, differences in attachment strategies to particles[17], differences in chemotactic strategies for exploring nutrient gradients and finding food particles[50,51], differences in the number and types of lyase genes for specific polysaccharides (e.g., alginate)[28], fluctuating nutrient concentrations combined with different growth kinetics[52], limitation by different trace nutrients[53] and predation by phages and protist grazers[54,55]. Trade-offs between nutrient acquisition and resistance to phage predation[56], for example, may enable coexistence of competitors[57], although the precise effects of phages on microbial communities remain uncertain[55,58]. Intransitive competitive dynamics, whereby multiple pairs of competing species collectively have no clear winner, may also play a role via antibiotic warfare[59,60]. It is likely that metabolically overlapping microorganisms differentiate ecologically in many more ways that we can currently identify, and hence community assembly takes place in a high-dimensional (multifactorial) space. Indeed, recent gene cataloging efforts across microbial genomes revealed hundreds of thousands of gene clusters with largely uncharacterized function[3]. In view of these observations, functional redundancy almost seems like an inevitable outcome in open microbial systems — systems where diversity is not limited by low immigration rates.

Care must be taken when assessing the metabolic niche utilized by an organism solely based on its metabolic potential, e.g., inferred from its genome. Populations with a similar metabolic repertoire ("fundamental" metabolic niche[61]) may specialize on distinct nutrients, thus exhibiting separate "realized" niches that may be expressed at the transcriptional level[51,62]. In particular, a functional group may appear as highly

redundant even if only a few members actively perform that function at a time, since some members can exhibit alternative modes to gain energy while others may simply be inactive. The metabolic functions performed by a given population generally depends on environmental conditions as well as on the presence and activity of other community members[58]. We emphasize that the predictions of classical competition theory, discussed above, still apply even if organisms in a community are metabolically multifunctional. That is, at steady state the number of coexisting organisms cannot exceed the number of resources (including metabolic byproducts) limiting the growth of at least one organism[49]. For example, while two hydrogenotrophic methanogens may coexist in the same environment, at steady state they cannot be limited by the same hydrogen pool. Fine-scale spatial segregation in a non-well-mixed environment is one possible mechanism enabling coexistence. For example, organisms with similar nutritional preferences can reside and obtain their nutrients within distinct biofilms and can thus co-exist at larger scales[51]. In these cases, however, it is important to realize that populations in distinct biofilms do not compete for the same nutrient pools and thus have distinct realized niches.

Functional redundancy does not imply neutrality

Loreau hypothesized that functional redundancy within a metabolic niche may reflect quasi-neutral coexistence of competitors[63]. However, as discussed above, coexisting microorganisms specializing on the same energy source not only typically differ in terms of their enzyme efficiencies and growth kinetics, but also in other traits influencing their growth rates under specific conditions. While differences between members of a functional group are generally acknowledged, controversy exists as to whether certain patterns of microbial community assembly may nevertheless be explained by neutral processes[64,65]. In analogy to neutral theories from macrobial ecology[66], Sloan *et al.*[67] developed a neutral model for local microbial community assembly based solely on stochastic immigration and ecological drift (fluctuations due to the stochasticity of birth/death events in finite populations), while omitting speciation — a common element of macrobial neutral theories. Sloan *et al.*[67] concluded that stochastic immigration and ecological drift are important factors in shaping prokaryotic communities, particularly within metabolic functional groups[67,68]. Following Sloan *et al.*[67], neutral models have been used to partly explain microbial biogeographical patterns in diverse environments, including animal guts[69], soil[70], bioreactors[71], tree holes[72] and biofilms[73]. It has also been suggested that ecological drift within functional groups may partly explain species turnover over time, for example in bioreactors[74,75], in subsurface waters[76] and in stream catchments[77].

We emphasize that complex or apparently stochastic changes in taxonomic composition within functional groups, even in closed systems, should not be confused for ecological drift. In fact, ecological drift is rarely a valid explanation for taxonomic turnover within functional groups, as observed for example in bioreactors over time[15,17,74,75]. This is because the importance of ecological drift, in contrast to selection processes, diminishes at large population sizes and/or large ecological differences between competitors[78,79]. In bioreactors and most natural environments, cell densities can be extremely high (up to 10^{13} cells $\cdot$ L^{-1} in bioreactors[80]) to the point that selection processes would clearly dominate over ecological drift. Indeed, neutral stochastic birth-death models predict that even at low population sizes (10^4 cells), it would take a relatively

rare organism (1% proportion) in a community consisting of equal competitors on average over 1,600 days to reach a proportion of 30% solely via ecological drift (based on a generation time of 1 day[40]). When even a weak competitive advantage is assumed for one of the organisms (5% higher expected growth rate), both populations closely follow the deterministic trajectory predicted from competitive exclusion (fraction of explained variance 0.98 ± 0.02 s.d.; Supplement S.1). Hence, the effect of drift on population trajectories becomes negligible even under weak competitive differences. We note that the above model parameters are quite conservative. Indeed, microbial populations typically comprise more than 10^4 cells and it is not uncommon to observe extremely rare taxa ($< 0.1\%$ proportion) replacing previously dominant and metabolically similar taxa within just a few weeks, even under constant environmental conditions[10,15,22,75]. Moreover, even strains of the same species can exhibit vastly different substrate affinities (e.g., up to 400% difference[81]) or distinct susceptibilities to specialist phages[55,58]. Consequently, the probability that competitors have sufficiently similar growth rates over a sufficient period of time for drift to be a noticeable driver of taxonomic turnover is extremely low. Hence, while functional redundancy — either at a local or regional scale — is a necessary condition for taxonomic turnover within functional groups, turnover itself is generally not explained by ecological drift. Consistent with this prediction, a recent large-scale analysis of human microbiomes[82] found that fewer than 1% of communities satisfied Hubbel's neutral theory of biodiversity[66]. Similarly, a survey of bromeliad microbiomes found that assembly within functional groups was far from neutral, despite their constant functional structure, high functional redundancies and highly variable taxonomic composition between bromeliads[14]. Even in plant and animal ecology, where population sizes are much lower than in typical microbial communities, clear evidence for a strong role of ecological drift (e.g., compared to selection) is rare[79].

Since ecological drift generally can't explain taxonomic turnover within functional groups, this turnover must result from ecological differences between members of a functional group and, potentially, dispersal processes. Previous studies indeed suggested limited dispersal as an important source of taxonomic variation between sites, based on random phylogenetic structure of early colonists during succession[83], increasing taxonomic richness over time in semi-open incubations[84], or — more commonly — a decay of community similarity with increasing geographical distance[85,86]. The latter studies remain inconclusive, however, because a distance decay in community similarity can also be caused by spatially correlated environmental heterogeneity. For example, accounting for environmental heterogeneity was found to explain all or most of the correlation between distance and microbial community dissimilarity in salt marshes[87], in the global ocean[9] and between bromeliads[14]. Environmental heterogeneity is generally hard to rule out as a cause of spatial variation of taxonomic community composition without thorough environmental measurements.

In experiments with replicate bioreactors operated under constant conditions, microbial community composition followed complex but reproducible trajectories over periods ranging from weeks to months[20,22,88]. This suggests that taxonomic turnover within functional groups in the absence of obvious environmental variation can be driven by intrinsic and at least partly deterministic processes. Such intrinsic processes may include "killing-the-winner" type phage-host-interactions, where specialist phages repeatedly induce the collapse of dominant microbial populations, although experimental evidence for this mechanism remains rare[89].

Other proposed mechanisms include antibiotic warfare[59,60], rapid evolution of cross-feeding[90] and adaptive niche construction[91]. Every species may thus be affected by a distinct combination of biotic and abiotic factors that modulate its instantaneous growth rate, even if its metabolic potential overlaps with other members of the community[45]. These factors may be frequency-dependent and may include a stochastic component, for example due to mutations or horizontal gene transfer events. In practice, chaotic population dynamics[92] may obscure the distinction between deterministic and stochastic assembly processes. Further, at regional scales infrequent dispersal may add stochasticity to community assembly in a way that cannot be explained by intrinsic dynamics alone. Hence, even if all environmental factors were known at a specific moment in time, taxonomic community composition may not be perfectly predictable.

Conclusions

Frequently perceived as an indication of neutral assembly, functional redundancy is actually a manifestation of the ecological diversity of microorganisms capable of a particular metabolic function. Functional redundancy is an inevitable emergent property of open microbial systems that becomes visible when a high-dimensional trait space is projected to a lower-dimensional function space of interest. It may thus be seen as a partial measure of diversity, namely diversity within functional groups, that is mathematically complementary to functional richness of a community, just as the taxonomic composition within functional groups can be considered complementary to functional community structure[9,14]. We speculate that the degree of functional redundancy in open microbial systems may be a stabilized systemic property that is largely determined by the type of environment and the functions considered. This hypothesis may be particularly true for natural systems with continuous exposure to immigration, such as the open ocean, where a balance between immigration and local extinction could determine functional redundancy at ecological time scales.

Depending on the choice of functions, a distinction between functional community structure and composition within functional groups can yield important insight into biogeochemistry and community assembly mechanisms. Indeed, metabolic pathways involved in energy transduction can be strongly coupled to certain environmental factors and elemental cycles[5–7,33] and can appear decoupled from particular taxonomic assemblages[10,14,77]. Similar observations are known from macrobial ecology[93], which has had a long history of describing community structure in terms of guilds, lifeforms and strategies, all of which may be considered analogous to metabolic functional groups in microbes. More recently, there have been calls to entirely abandon modeling macroscopic communities in terms of species, but instead to focus on functional traits[94]. Reducing microbial communities to energy-transducing metabolic functions, and investigating functional redundancy with respect to these functions, may thus also be a fruitful approach for microbial ecology.

Beyond metabolic niche effects, several additional mechanisms, such as predation and antibiotic warfare, can modulate the taxonomic composition of microbial communities over space and time, even if the activity of certain metabolic functions is strongly conserved. It is clear that this apparent decoupling between function and taxonomy is not the simple result of stochastic ecological drift within functional groups. How and under which conditions various mechanisms lead to this decoupling, and what determines the extent of functional

redundancy in microbial systems, are becoming central questions in ecology.

Glossary

• **functional group**	The set of taxa potentially capable of performing a specific biochemical function, e.g., based on their genetic content.
• **functional richness**	(of a community) Number of focal biochemical functions or genes present.
• **functional redundancy**	(with respect to a given function) The coexistence of multiple distinct taxa or genomes capable of performing the same focal biochemical function.
• **functional structure**	(of a community) Relative abundances of various focal functional groups, or of genes associated with focal functions.
• **ecological drift**	Fluctuations in relative population sizes due to the stochastic nature of birth-death events in finite populations [79].
• **metabolic niche**	(in an ecosystem) The ability for organisms to gain energy for growth using a specific metabolic pathway (e.g., H_2/CO_2 methanogenesis) or half-reaction (e.g., use of a specific electron acceptor for respiration).
• **metabolic niche effects**	(on community assembly) Mechanisms selecting for organisms able to exploit specific metabolic niches. Such mechanisms may include the availability of light for photosynthesis, or of sulfate as an electron acceptor for respiration.
• **microbial system**	A microbial community, its metabolites in the extracellular environment and bidirectionally coupled abiotic physicochemical processes, including physical transport processes and abiotic chemical reactions. Analogous to "ecosystem", but focusing on microbial members instead of macrobial food webs.

Acknowledgements

We thank Matthew Pennell, Ford Doolittle, Adam C. Martiny and Ilan Rubin for discussions and for participation at a workshop from which this synthesis emerged. We thank the Canadian Institute for Ecology and Evolution (CIEE) for financial support of all authors, by means of a Thematic Working group grant on the "Evolution of microbial metabolic and genomic diversity at multiple scales". We thank the Biodiversity Research Centre and the Adapting Biosystems program, University of British Columbia, for financial support and Katie Beall for logistical support. S.L. was supported by an NSERC grant, and a postdoctoral fellowship from the Biodiversity Research Centre, UBC. J.A.H. was supported by the NSF Center for Dark Energy Biosphere Investigations (OCE-0939564).

Author contributions

S.L., L.W.P. and M.D. organized the workshop from which this synthesis emerged. S.L. performed the data analyses. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.L. Supporting text and tables, cited in the text, are provided as Supplementary Material.

References

[1] Hug, L.A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 EP (2016).

[2] Gans, J., Wolinsky, M. & Dunbar, J. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**, 1387–1390 (2005).

[3] Powell, S. *et al.* eggnog v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**, D231–D239 (2014).

[4] O'Leary, N.A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733 (2016).

[5] Falkowski, P.G., Fenchel, T. & Delong, E.F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034–1039 (2008).

[6] Raes, J., Letunic, I., Yamada, T., Jensen, L.J. & Bork, P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Syst. Biol.* **7**, 473 (2011).

[7] Reed, D.C., Algar, C.K., Huber, J.A. & Dick, G.J. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc. Natl. Acad. Sci. USA* **111**, 1879–1884 (2014).

[8] Louca, S. *et al.* Integrating biogeochemistry with multiomic sequence information in a model oxygen minimum zone. *Proc. Natl. Acad. Sci. USA* **113**, E5925–E5933 (2016).

[9] Louca, S., Parfrey, L.W. & Doebeli, M. Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**, 1272–1277 (2016).

[10] Fernández, A. *et al.* How stable is stable? Function versus community composition. *Appl. Environ. Microbiol.* **65**, 3697–3704 (1999).

[11] Turnbaugh, P.J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).

[12] Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S. & Thomas, T. Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. USA* **108**, 14288–14293 (2011).

[13] Nelson, M.B., Martiny, A.C. & Martiny, J.B.H. Global biogeography of microbial nitrogen-cycling traits in soil. *Proc. Natl. Acad. Sci. USA* **113**, 8033–8040 (2016).

[14] Louca, S. *et al.* High taxonomic variability despite stable functional structure across microbial communities. *Nat. Ecol. Evol.* **1**, 0015 (2016).

[15] Wittebolle, L., Vervaeren, H., Verstraete, W. & Boon, N. Quantifying community dynamics of nitrifiers in functionally stable reactors. *Appl. Environ. Microbiol.* **74**, 286–293 (2008).

[16] Wells, G.F. *et al.* Ammonia-oxidizing communities in a highly aerated full-scale activated sludge bioreactor: betaproteobacterial dynamics and low relative abundance of crenarchaea. *Environ. Microbiol.* **11**, 2310–2328 (2009).

[17] Vanwonterghem, I., Jensen, P.D., Rabaey, K. & Tyson, G.W. Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion. *Environ. Microbiol.* **18**, 3144–3158 (2016).

[18] Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 EP (2016).

[19] Tringe, S.G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).

[20] Kaewpipat, K. & Grady, C. Microbial population dynamics in laboratory-scale activated sludge reactors. *Water Sci. Technol.* **46**, 19–27 (2002).

[21] Wang, X. *et al.* Bacterial community dynamics in a functionally stable pilot-scale wastewater treatment plant. *Bioresour. Technol.* **102**, 2352–2357 (2011).

[22] Fernandez-Gonzalez, N., Huber, J.A. & Vallino, J.J. Microbial communities are well adapted to disturbances in energy input. *mSystems* **1**, e00117–16 (2016).

[23] Sheng, Y. *et al.* Geochemical and temporal influences on the enrichment of acidophilic iron-oxidizing bacterial communities. *Appl. Environ. Microbiol.* **82**, 3611–3621 (2016).

[24] Aguilar, D., Aviles, F.X., Querol, E. & Sternberg, M.J.E. Analysis of phenetic trees based on metabolic capabilites across the three domains of life. *J. Mol. Biol.* **340**, 491–512 (2004).

[25] Martiny, J.B.H., Jones, S.E., Lennon, J.T. & Martiny, A.C. Microbiomes in light of traits: A phylogenetic perspective. *Science* **350** (2015).

[26] Morris, J.J., Lenski, R.E. & Zinser, E.R. The black queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**, e00036–12 (2012).

[27] David, L.A. & Alm, E.J. Rapid evolutionary innovation during an archaean genetic expansion. *Nature* **469**, 93–96 (2011).

[28] Hehemann, J.H. *et al.* Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nat. Commun.* **7**, 12860 EP (2016).

[29] Welch, R.A. *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **99**, 17020–17024 (2002).

[30] Martiny, A.C., Treseder, K. & Pusch, G. Phylogenetic conservatism of functional traits in microorganisms. *ISME J.* **7**, 830–838 (2013).

[31] Whitman, W.B. (ed.) *Bergey's Manual of Systematics of Archaea and Bacteria* (John Wiley & Sons, Inc., Hoboken, New Jersey, 2015).

[32] Canfield, D.E. & Thamdrup, B. Towards a consistent classification scheme for geochemical environments, or, why we wish the term 'suboxic' would go away. *Geobiol.* **7**, 385–392 (2009).

[33] Reed, D.C. *et al.* Predicting the response of the deep-ocean microbiome to geochemical perturbations by hydrothermal vents. *ISME J.* **9**, 1857–1869 (2015).

[34] Graham, E.B. *et al.* Microbes as engines of ecosystem function: When does community structure enhance predictions of ecosystem processes? *Front. Microbiol.* **7**, 214 (2016).

[35] Girvan, M.S., Campbell, C.D., Killham, K., Prosser, J.I. & Glover, L.A. Bacterial diversity promotes community stability and functional resilience after perturbation. *Environ. Microbiol.* **7**, 301–313 (2005).

[36] Langenheder, S., Lindström, E.S. & Tranvik, L.J. Weak coupling between community composition and functioning of aquatic bacteria. *Limnol. Oceanogr.* **50**, 957–967 (2005).

[37] Langenheder, S., Lindström, E.S. & Tranvik, L.J. Structure and function of bacterial communities emerging from different sources under identical conditions. *Appl. Environ. Microbiol.* **72**, 212–220 (2006).

[38] Peter, H. *et al.* Function-specific response to depletion of microbial diversity. *ISME J.* **5**, 351–361 (2011).

[39] Jurburg, S.D. & Salles, J.F. *Functional Redundancy and Ecosystem Function – The Soil Microbiota as a Case Study*, chap. 2, 29–49 (INTECH, 2015).

[40] Louca, S. & Doebeli, M. Taxonomic variability and functional stability in microbial communities infected by phages. *Environ. Microbiol.* **19**, 3863–3878 (2017).

[41] Hawley, A.K., Brewer, H.M., Norbeck, A.D., Paša-Tolić, L. & Hallam, S.J. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proc. Natl. Acad. Sci. USA* **111**, 11395–11400 (2014).

[42] Lavorel, S. & Garnier, E. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the holy grail. *Funct. Ecol.* **16**, 545–556 (2002).

[43] Tully, B., Wheat, C.G., Glazer, B.T. & Huber, J. A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *ISME J.* **12**, 1–16 (2017).

[44] Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**, 416–420 (2014).

[45] Fuhrman, J.A., Cram, J.A. & Needham, D.M. Marine microbial community dynamics and their ecological interpretation. *Nat. Rev. Micro.* **13**, 133–146 (2015).

[46] Allison, S.D. & Martiny, J.B.H. Resistance, resilience, and redundancy in microbial communities. *Proc. Natl. Acad. Sci. USA* **105**, 11512–11519 (2008).

[47] Chesson, P. Mechanisms of maintenance of species diversity. *Annu. Rev. Ecol. Syst.* **31**, 343–366 (2000).

[48] Tilman, D. *Resource Competition and Community Structure* (Princeton University Press, Princeton, NJ, 1982).

[49] Freter, R., Brickner, H., Botney, M., Cleven, D. & Aranki, A. Mechanisms that control bacterial populations in continuous-flow culture models of mouse large intestinal flora. *Infect. Immun.* **39**, 676–685 (1983).

[50] Yawata, Y. *et al.* Competition–dispersal tradeoff ecologically differentiates recently speciated marine bacterioplankton populations. *Proc. Natl. Acad. Sci. USA* **111**, 5622–5627 (2014).

[51] Pereira, F.C. & Berry, D. Microbial nutrient niches in the gut. *Environ. Microbiol.* (2017).

[52] Sommer, U. The paradox of the plankton: Fluctuations of phosphorus availability maintain diversity of phytoplankton in flow-through cultures. *Limnol. Oceanogr.* **29**, 633–636 (1984).

[53] Moore, C. *et al.* Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.* **6**, 701 (2013).

[54] Wildschutte, H., Wolfe, D.M., Tamewitz, A. & Lawrence, J.G. Protozoan predation, diversifying selection, and the evolution of antigenic diversity in salmonella. *Proc. Natl. Acad. Sci. USA* **101**, 10644–10649 (2004).

[55] Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).

[56] Bohannan, B.J.M., Kerr, B., Jessup, C.M., Hughes, J.B. & Sandvik, G. Trade-offs and coexistence in microbial microcosms. *Antonie Leeuwenhoek* **81**, 107–115 (2002).

[57] Chesson, P. & Kuang, J.J. The interaction between predation and competition. *Nature* **456**, 235–238 (2008).

[58] Cordero, O.X. & Polz, M.F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).

[59] Czárán, T.L., Hoekstra, R.F. & Pagie, L. Chemical warfare between microbes promotes biodiversity. *Proc. Natl. Acad. Sci. USA* **99**, 786–790 (2002).

[60] Hibbing, M.E., Fuqua, C., Parsek, M.R. & Peterson, S.B. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Micro.* **8**, 15–25 (2010).

[61] Hutchinson, G. Concluding remarks. In *Cold Spring Harbour Symposium in Quantitative Biology*, vol. 22, 415–427 (1957).

[62] Plichta, D.R. *et al.* Transcriptional interactions suggest niche segregation among microorganisms in the human gut. *Nat. Microbiol.* **1**, 16152 EP – (2016).

[63] Loreau, M. Does functional redundancy exist? *Oikos* **104**, 606–611 (2004).

[64] Curtis, T.P. & Sloan, W.T. Prokaryotic diversity and its limits: microbial community structure in nature and implications for microbial ecology. *Curr. Opin. Microbiol.* **7**, 221 – 226 (2004).

[65] Konopka, A., Lindemann, S. & Fredrickson, J. Dynamics in microbial communities: unraveling mechanisms to identify principles. *ISME J.* **9**, 1488–1495 (2015).

[66] Hubbell, S.P. *The Unified Neutral Theory of Biodiversity and Biogeography*, vol. 32 (Princeton University Press, 2001).

[67] Sloan, W.T. *et al.* Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8**, 732–740 (2006).

[68] Sloan, W.T., Woodcock, S., Lunn, M., Head, I.M. & Curtis, T.P. Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microb. Ecol.* **53**, 443–455 (2007).

[69] Burns, A.R. *et al.* Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* (2015).

[70] Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C. & Fitter, A.H. Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J.* **4**, 337–345 (2009).

[71] Curtis, T., Polchan, M., Baptista, J., Davenport, R. & Sloan, W. Microbial community assembly, theory and rare functions. *Front. Microbiol.* **4**, 68 (2013).

[72] Woodcock, S. *et al.* Neutral assembly of bacterial communities. *FEMS Microbiol. Ecol.* **62**, 171–180 (2007).

[73] Woodcock, S. & Sloan, W.T. Biofilm community succession: a neutral perspective. *Microbiology* **163**, 664–668 (2017).

[74] Ayarza, J.M. & Erijman, L. Balance of neutral and deterministic components in the dynamics of activated sludge floc assembly. *Microb. Ecol.* **61**, 486–495 (2010).

[75] Ofiţeru, I.D. *et al.* Combined niche and neutral effects in a microbial wastewater treatment community. *Proc. Natl. Acad. Sci. USA* **107**, 15345–15350 (2010).

[76] Stegen, J.C., Lin, X., Konopka, A.E. & Fredrickson, J.K. Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J.* **6**, 1653–1664 (2012).

[77] Frossard, A., Gerull, L., Mutz, M. & Gessner, M.O. Disconnect of microbial structure and function: enzyme activities and bacterial communities in nascent stream corridors. *ISME J.* **6**, 680–691 (2012).

[78] Lande, R., Engen, S. & Sæther, B. *Stochastic Population Dynamics in Ecology and Conservation*. Oxford series in ecology and evolution (Oxford University Press, 2003).

[79] Vellend, M. Conceptual synthesis in community ecology. *Q. Rev. Biol.* **85**, 183–206 (2010).

[80] Krakat, N., Westphal, A., Schmidt, S. & Scherer, P. Anaerobic digestion of renewable biomass: Thermophilic temperature governs methanogen population dynamics. *Appl. Environ. Microbiol.* **76**, 1842–1850 (2010).

[81] Ohtsubo, S. *et al.* Comparison of acetate utilization among strains of an aceticlastic methanogen, *Methanothrix soehngenii*. *Appl. Environ. Microbiol.* **58**, 703–705 (1992).

[82] Li, L. & Ma, Z.S. Testing the neutral theory of biodiversity with human microbiome datasets. *Scientific Reports* **6**, 31448 (2016).

[83] Dini-Andreote, F., Stegen, J.C., van Elsas, J.D. & Salles, J.F. Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc. Natl. Acad. Sci. USA* **112**, E1326–E1332 (2015).

[84] Albright, M.B.N. & Martiny, J.B.H. Dispersal alters bacterial diversity and composition in a natural community. *ISME J.* – (2017).

[85] Martiny, J.B.H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).

[86] Martiny, J.B.H., Eisen, J.A., Penn, K., Allison, S.D. & Horner-Devine, M.C. Drivers of bacterial β-diversity depend on spatial scale. *Proc. Natl. Acad. Sci. USA* **108**, 7850–7854 (2011).

[87] Horner-Devine, M.C., Lage, M., Hughes, J.B. & Bohannan, B.J. A taxa–area relationship for bacteria. *Nature* **432**, 750–753 (2004).

[88] Vanwonterghem, I. *et al.* Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *ISME J.* **8**, 2015–2028 (2014).

[89] Shapiro, O.H. & Kushmaro, A. Bacteriophage ecology in environmental biotechnology processes. *Curr. Opin. Biotechnol.* **22**, 449–455 (2011).

[90] Herron, M.D. & Doebeli, M. Parallel evolutionary dynamics of adaptive diversification in *Escherichia coli*. *PLoS Biol.* **11**, e1001490 (2013).

[91] Callahan, B.J., Fukami, T. & Fisher, D.S. Rapid evolution of adaptive niche construction in experimental microbial populations. *Evolution* **68**, 3307–3316 (2014).

[92] Graham, D.W. *et al.* Experimental demonstration of chaotic instability in biological nitrification. *ISME J.* **1**, 385–393 (2007).

[93] Fukami, T., Martijn Bezemer, T., Mortimer, S.R. & Putten, W.H. Species divergence and trait convergence in experimental plant community assembly. *Ecol. Lett.* **8**, 1283–1290 (2005).

[94] McGill, B.J., Enquist, B.J., Weiher, E. & Westoby, M. Rebuilding community ecology from functional traits. *Trends Ecol. Evol.* **21**, 178–185 (2006).

[95] Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2014).

[96] Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

[97] Haft, D.H. *et al.* TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41**, D387 (2013).

[98] Finn, R.D. *et al.* The pfam protein families database. *Nucleic Acids Res.* **36**, D281 (2008).

[99] Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).

[100] Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* (2017).

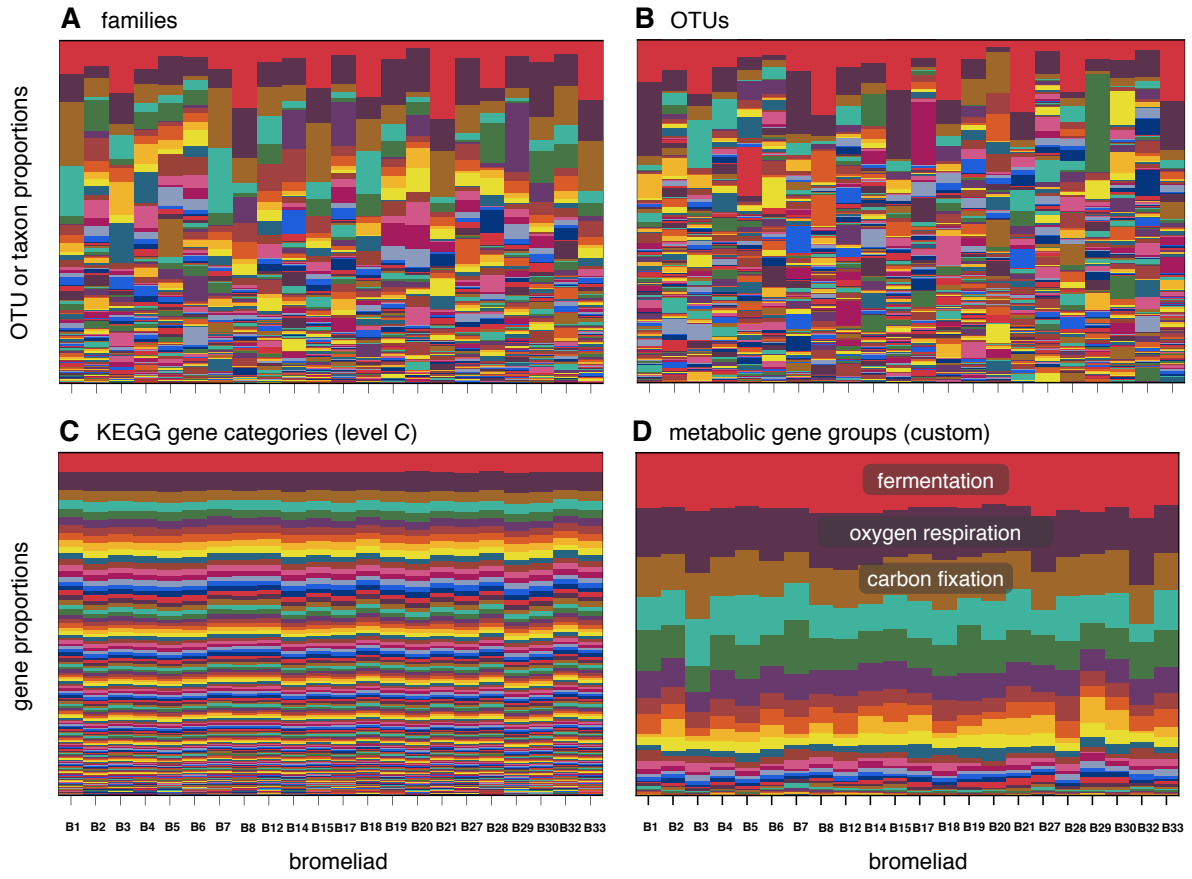

Figure 1: **Gene-centric structure of microbial communities can decouple from taxonomic composition.** (A,B) Relative abundances of bacterial and archaeal families (A) and OTUs (B; at 99% 16S gene similarity), found in the foliage of 22 similar and concurrently sampled *Aechmea nudicaulis* bromeliads in Juruba Tiba National Park, Brazil[14] (one column per bromeliad, one color per taxon). (C,D) Corresponding metagenomic community composition in terms of KEGG standard categories (C) and custom metabolic gene groups (D), as defined in[14] (one column per sample, one color per gene group). Note the more variable taxonomic composition across bromeliads (A, B), compared to the relatively conserved metagenomic composition (C, D).

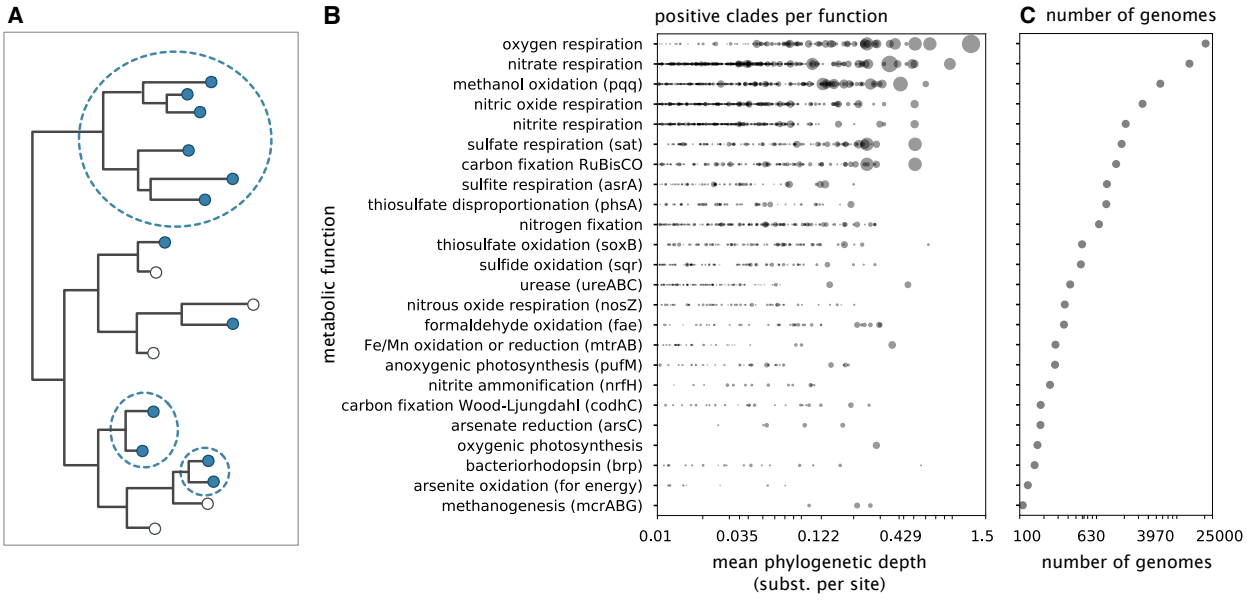

Figure 2: **Phylogenetic conservatism varies between functions and between clades.** (A) Schematic illustration of a phylogenetic tree, where filled and empty tips indicate the presence and absence, respectively, of a specific function. Depending on the location in the tree, a function may be conserved in deep or shallow clades (dashed circles). (B) Prokaryotic clades positive in various metabolic functions (i.e. with the function present in $\geq$95% of tips), represented as circles (one circle per positive clade per function). Circles are positioned on the horizontal axis according to the clade's mean phylogenetic depth. Larger circles correspond to clades containing more tips (logarithmic scale). The majority of functions are conserved in a multitude of clades of variable depths and sizes, with oxygenic photosynthesis being a notable exception. Thus, for most functions there exists no taxonomic resolution at which taxa either always or never exhibit that function. (C) Number of non-redundant prokaryotic genomes (i.e., with unique NCBI taxon IDs), downloaded from NCBI RefSeq[4] and found to exhibit each function. B and C are based on genes detected in ~59,000 nearly-complete sequenced genomes. See Methods for details.

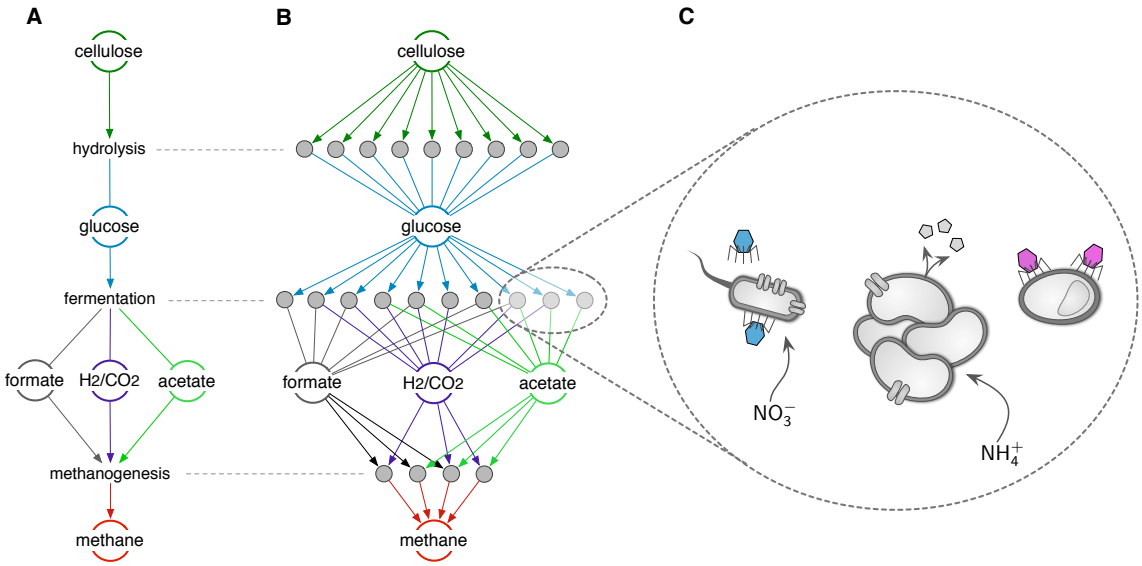

Figure 3: **Functional redundancy in methanogenic communities (schematic illustration).** (A) Illustration of a typical metabolic network spanned by microbial communities in methanogenic cellulose-fed bioreactors, driving the catabolism of cellulose to methane. Circles represent substrates or end-products, and edge color indicates the associated substrate. (B) Expansion of each catabolic step, showing multiple distinct organisms capable of performing the same reaction. Filled dots represent distinct population genomes. Schematic illustration of roughly analogous findings by Vanwonterghem *et al.* [17]. (C) Focus on 3 seemingly redundant organisms, catabolizing glucose to acetate. Realized niche differentiation and coexistence can be enabled by trait differences beyond the type of substrates used, potentially including susceptibility to different phages (blue vs purple), different strategies for foraging, attachment to particles and biofilm formation, different nitrogen pools used (nitrate NO_3^- vs ammonium NH_4^+), as well as production and resistance to different antibiotics (small pentagons).

Methods

Phylogenetic distribution of metabolic functions

To examine the phylogenetic distribution of various metabolic functions (Figs. 2B,C), we proceeded as described below. Unless otherwise mentioned, all online files were downloaded on September 12, 2017. A total of 92,315 sequenced prokaryotic genomes with a completion status "Complete Genome", "Contig" or "Scaffold", and a gap fraction not greater than 1%, were downloaded from NCBI RefSeq[4]. Downloaded genomes were further checked for completeness and contamination with checkM 1.0.6[95], using the option "`reduced_tree`". Genomes estimated to be less than 98% complete, exhibiting a contamination level above 1%, exhibiting a strain heterogeneity above 1% or lacking a protein prediction file (files `protein.faa`, provided by NCBI), were discarded, leaving us with 59,092 nearly-complete genomes for downstream analysis. In each genome, we used Hidden Markov Models (HMMs) and hmmsearch v3.1b2[96] to search for proteins associated with various metabolic functions, such as photosynthesis or methanogenesis. HMMs were obtained from the Jillian Banfield lab GitHub page (`https://github.com/banfieldlab`)[18], the TIGRFAM database v15.0[97] and the Pfam protein database v30.0[98]. Pre-calibrated noise cutoff values (included in each HMM) and a maximum E-value of 10^{-50} were used as hit criterion for each HMM. For some functions, multiple proteins were used as alternative proxies for the function, while for other functions multiple proteins had to be all present for the function. Proteins used as proxies for each function and corresponding HMM accession numbers are listed in Table S1.

Each prokaryotic OTU in the SILVA NR99 small subunit ribosomal RNA database (release 128[99]) was mapped to one of the genomes whenever possible, based on the ID of the NCBI taxonomy project ("taxid") provided by SILVA (file `https://www.arb-silva.de/no_cache/download/archive/release_128/Exports/taxmap_embl_ssu_ref_128.txt.gz`) and the taxid provided by NCBI for each genome (table `ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/*/assembly_summary.txt`, where "`*`" is either "`bacteria` or "`archaea`"). A total of 54,043 OTUs could be mapped to a genome. A phylogenetic tree was constructed for the mapped OTUs by pruning the official SILVA NR99 tree. For each mapped OTU, we assumed a metabolic function to be present if it was found to be present in the mapped genome. To determine the clades within which a particular function was conserved (Fig. 2B), we proceeded as follows: We traversed the tree from the root to the tips in breadth-first search mode until reaching a node whose descending clade was positive in the function (i.e. where the function was present in at least 95% of descending tips), recording the mean phylogenetic depth of the clade (average distance of the node to its tips) and the total size of the clade (total number of descending tips). All descending tips and nodes were subsequently excluded from the remainder of the traversal, and traversal continued with the next node in the traversal queue. Thus, every positive clade recorded (and plotted in Fig. 2B) was maximal, in the sense that it was not part of any bigger positive clade. Single positive tips with no positive sister tip were not counted, because in that case no information was available on the phylogenetic depth at which the function is locally conserved. Occasional positive clades with a mean phylogenetic depth below 1% are not shown in Fig. 2B, since OUTs in SILVA are officially resolved at 1% dissimilarity. The above analysis has been implemented in the R package `castor`,

function `get_trait_depth`[100].

To calculate the number of non-redundant genomes exhibiting a particular function (Fig. 2C), i.e. accounting for the fact that some RefSeq genomes are genomes of the same strains or very closely related strains, we only counted genomes with a unique NCBI taxid. Among the 59,092 genomes, there were 22,660 unique taxids. We emphasize that the number of genomes per function shown in Fig. 2C should not be compared between functions, due to biases in the types of prokaryotes represented in RefSeq.