# Cross-biome microbial networks reveal functional redundancy and suggest genome reduction through functional complementarity

**Fernando Puente-Sánchez[1,2,*], Alberto Pascual-García[3,4], Ugo Bastolla[3], Carlos Pedrós-Alió[1], Javier Tamames[1]**

[1]Systems Biology Program, Centro Nacional de Biotecnología (CSIC). C/ Darwin 3, Campus de Cantoblanco, 28049 Madrid, Spain.

[2]Department of Aquatic Sciences and Assessment, Swedish University for Agricultural Sciences (SLU), Lennart Hjelms väg 9, 756 51, Uppsala, Sweden

[3]Bioinformatics Unit, Centro de Biología Molecular Severo Ochoa (UAM-CSIC), C/ Nicolás Cabrera 1, Campus de Cantoblanco, 28049 Madrid, Spain.

[4] Institute of Integrative Biology, ETH-Zürich, Universitätstrasse 16, 8055, Zürich, Switzerland

*To whom correspondence may be addressed: fpusan@gmail.com

# Abstract

Microbial communities are complex and dynamic entities, and their structure arises from the interplay of a multitude of factors, including the interactions of microorganisms with each other and with the environment. Since each extant community has a unique eco-evolutionary history, it might appear that contingency rather than general rules govern their assembly. In spite of this, there is evidence that some general assembly principles exist, at least to a certain extent. In this work, we sought to identify those principles by performing a cross-study, cross-biome meta-analysis of microbial occurrence data in more than 5,000 samples from ten different environmental groups. We adopted a novel algorithm that allows the same taxa to aggregate with different partners in different habitats, capturing the complexity of interactions inherent to natural microbial communities. We tried to decouple function from phylogeny, the environment, and genome size, in order to provide an unbiased characterization of phylogenetic and functional redundancy in environmental microbial assemblages. We then examined the phylogenetic and functional composition of the resulting inferred communities, and searched for global patterns of assembly both at the community level and in individual metabolic pathways.

Our analysis of the resulting microbial assemblages highlighted that environmental communities are more functionally redundant than expected by chance. This effect is greater for communities appearing in more than one environment, suggesting a link between functional redundancy and environmental adaptation. In spite of this, certain pathways are observed in fewer taxa than expected by chance, suggesting the presence of auxotrophy, and presumably cooperation among community members, which is supported by our analysis of amino acid biosynthesis pathways. Furthermore, this hypothetical cooperation may play a role in genome reduction, since we observed a negative relationship between the size of bacterial genomes and the number of taxa of the community they belong to.

Overall, our results provide a global characterization of environmental microbial communities, and offer design principles for engineering robust bacterial communities.

# Introduction

Microorganisms are the second most abundant component of the global biomass on Earth (Bar-On *et al.*, 2018), and the first one in terms of biodiversity (Locey & Lennon, 2016). In addition, they are the only ones capable of performing key ecological functions, including nitrogen fixation, methanogenesis, and all kinds of anaerobic respirations. As such, they play a critical role in driving the essential biogeochemical cycles that sustain life on our planet (Falkowski *et al.*, 2008). Microorganisms interact among themselves and with the environment, giving rise to emergent community-level properties (Konopka *et* al., 2015; Louca *et al.*, 2018). These interactions are primary driving forces in microbial ecology, and determine the fate of microbial communities and, by extension, of their constituent microorganisms (Konopka *et* al., 2015). Therefore, the study of individual microorganisms is often not enough to predict how those very same microorganisms will behave in nature; instead, they have to be considered in the context of the community they live on.

Microbial communities are complex and dynamic entities, and their structure arises from the interplay of four key ecological processes: selection, diversification, dispersal and drift (Vellend, 2010; Nemergut *et al*., 2013). Among them, selection (i.e., the existence of fitness differences between individuals) is a primary force shaping microbial community assembly (Nemergut *et al*., 2013; Konopka *et* al., 2015; Louca *et al.*, 2017). Natural selection counteracts random fluctuations and acts over short timescales, which makes it experimentally tractable (Chuang *et al.*, 2009; Ribeck & Lenski 2015; *Yu et al.*, 2017). This has led to an increasing interest in synthetic microbial ecology as a tool to generate and test hypotheses regarding community assembly processes (reviewed in Dolinšek *et al*., 2016). However, the simplicity inherent to synthetic microbial communities, while facilitating their precise characterization, might also limit their usefulness as proxies of natural microbial communities (Yu *et al.*, 2016; Ehsani *et al.*, 2018).

A complementary approach is to study natural microbial communities and look for common assembly patterns, trying to unravel the bases of microbial association (Datta *et al.*, 2016; Rivett & Bell, 2018; Enke *et* al., 2019; Pascual-García & Bell, 2020; Ma *et al.*, 2020). It has been argued that each extant community has a unique evolutionary history, which makes the search for ‘laws’ in Ecology futile (O’Hara, 2005). Still, there is evidence that microbial dynamics can be generalized to a certain extent (Bashan *et al.*, 2016; Goldford *et al*., 2018), allowing to extract useful broad principles from the study

of multiple microbial communities. Such principles can be experimentally tested, improving the understanding of natural communities, and ultimately allowing to design robust synthetic communities (Konopka *et al.*, 2015; Gibson *et al.*, 2016).

In this work, we sought to identify general assembly principles by performing a cross-study, cross-biome meta-analysis of microbial occurrence data in more than 5,000 samples from ten different environments. We used a novel algorithm to create ecological assemblages from pairwise aggregations of microbial genera, which includes a statistical procedure to evaluate the significance of multi-genera assemblages. The significance is evaluated on the basis of a null model that is specific to each environmental class, attempting to separate the influence of the environment from the influence of biological interactions. This novel algorithm allowed the same taxa to aggregate with different partners in different habitats, thus capturing the complexity of interactions inherent to natural microbial communities. Finally, we analyzed the metabolic potential of the genera present in our ecological network in order to investigate the roles of redundancy and functional complementation in specific metabolic pathways for microbial community assembly.

# Results and discussion

## Generation of a modular ecological network

Our taxa-assembly algorithm generates ecological networks by following the steps summarized in **Figure 1**. Briefly, we collected environmental 16S rRNA gene sequences from the NCBI *env nt* database, assigned them a sample identifier and, when possible, classified them into a defined environmental hierarchy (Pignatelli *et al.*, 2009). We then clustered 16S sequences into OTUs at the 97% level, which we subsequently classified phylogenetically (Pignatelli *et al.*, 2009). For this study, we chose to classify our OTUs at the genus level. This provided a high taxonomic resolution while still allowing us to reliably combining results from different studies, which in many cases targeted different regions of the 16S rRNA gene.

Thus, we obtained a database that records the presence/absence patterns of microbial genera across thousands of samples from different environments (**Figure 1a**). Again, the use of presence/absence data was a necessary compromise in order to reliably aggregate data from studies that used very different methodologies. We demonstrated before the usefulness of this approach for generating cross-biome microbial association networks (Pascual-García *et al.*, 2014).

For this study, we focused on ten different environments: freshwater, marine water, marine sediments, hypersaline, oil, thermal, hypothermal/polar, soils, host-associated and water-treatment plants, which amounted to a total of 13,362 samples and 1,424 genera in our database. After filtering (see methods), we obtained a total of 5,369 samples and 966 genera for creating an agglomerative ecological network as follows. At the beginning of the process, each node represents one genus, and from the presence/absence profiles we compute all-against-all pairwise aggregation scores, which represent significant co-occurrences between pairs of genera (Pascual-García *et al.*, 2014; **Figure 1b**). The computation of the scores considers as a null model that co-occurrences occur by chance. To reduce the influence of the environment, we develop a different null model in any specific environment. We then iteratively cluster genera into larger environmental assemblages. At each step, we join the two nodes A and B with the highest aggregation score (**Figure 1c**). A novelty of our method is that the new node A+B only conserves the samples in which both nodes are present. We then assign the remaining samples from A and from B to two new nodes A* and B*. This strategy allows investigating the

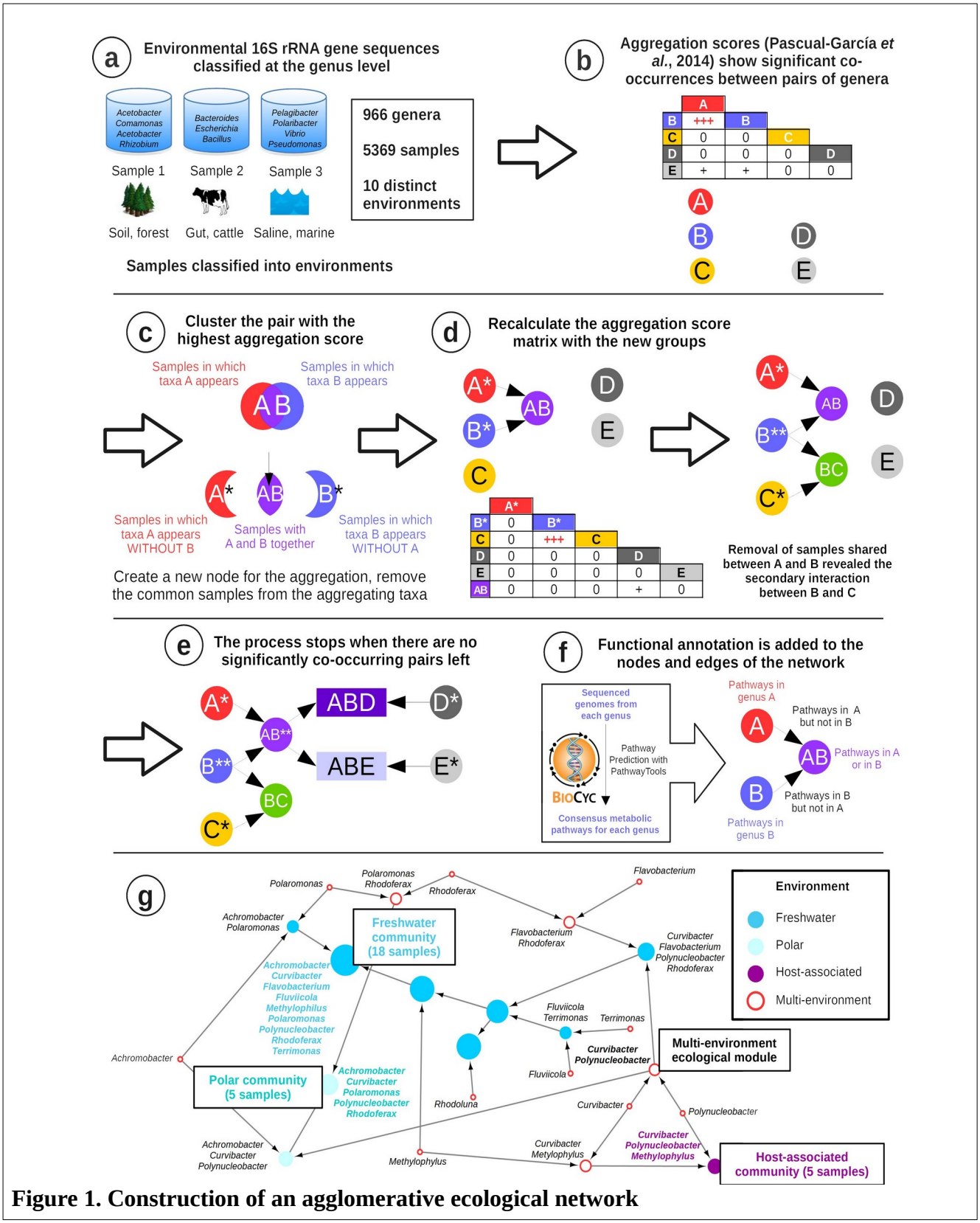


**Figure 1. Construction of an agglomerative ecological network**

aggregation of each genus with different partners in different environments. (**Figure 1c**). We then recalculate the aggregation score of the nodes A+B, A* and B* with respect to all the other nodes considering the samples in which each of them is observed (**Figure 1d**). We iterate this process until all pairwise scores fall below a significance threshold, obtaining a directed network that captures significant associations between increasingly large groups of genera (**Figure 1e**). Importantly, our procedure ensures that the whole assemblage is statistically significant. Finally, we use the Pathologic algorithm (Karp *et al.*, 2011) to predict the metabolic pathways present in the genera and assemblages included in our network (**Figure 1f**). In this way, we obtain a taxonomically and functionally annotated agglomerative ecological network that represents microbial associations at different levels of complexity (**Figure 1g**, **Supplementary Data S1**).

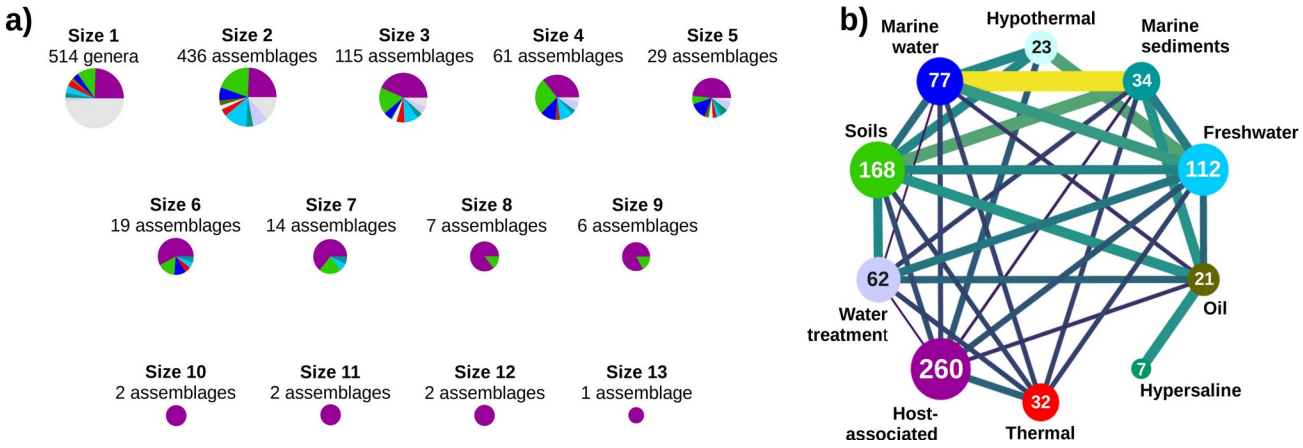


**Figure 2. Summary of the agglomerative ecological network. a)** Number of assemblages of different sizes, and their environmental distribution. Pie chart colors indicate environments as shown in **b)**, multi-environment assemblages are indicated in gray. **b)** Contribution of each environment to the network, and assemblages shared by different environments. Nodes are environments, the number of assemblages (size 2 or more) per environment is indicated inside the node. Link width and color show the assemblages that are shared between pairs of environments (as a percentage of the assemblages in the smallest environment of the pair, min 1.41%, max 29.41%). See **Supplementary Table S1** for details.

The final network included a total of 514 genera and 5,253 samples, resulting in 1,215 nodes and 1,428 edges. 701 nodes corresponded to assemblages of two or more genera, with the largest assemblage having 13 members (**Figure 2a**). The assemblages were distributed across the different environments, roughly following the number of input samples per environment (**Figure 2b**; **Supplementary Table S1**). Notably, some assemblages were reconstructed in more than one environment. For example, one third of the assemblages found in marine sediments were also found in marine water, highlighting the connectivity between both environments. Conversely, the host-associated environment, while having the highest number of assemblages, shared a small fraction of them with other environments (**Figure 2b**).

## Significant functional and phylogenetic redundancies in environmental microbial assemblages

Functional redundancy (i.e., the notion that multiple species can share similar roles in ecosystem functioning) has been previously reported in microbial communities, both for individual functions (Bell *et al.*, 2005; Wertz *et al.*, 2006; Jones *et al.*, 2008, Louca *et al.*, 2016,2017) and full metabolic reconstructions (Zelezniak *et al.*, 2015). On the other hand, its generality has also been challenged by several authors (Strickland *et al.*, 2009; Peter *et al.*, 2011; Fetzer *et al.*, 2015; Delgado-Baquerizo *et al.*, 2016; Galand *et al.*, 2016; Morrissey *et al.*, 2016). There are several issues that complicate the quantification of functional redundancy in microbial communities. In microorganisms, function is often associated to phylogeny (Martiny et al., 2012; Morrissey *et al.*, 2016; Tamames *et al.*, 2016). The presence of phylogenetically close taxa in a given community might thus increase the observed functional redundancy. Furthermore, taxa have themselves different environmental preferences (e.g., host-associated vs free living, saline vs non-saline, etc.; Tamames *et al.*, 2010; Nemergut *et al.*, 2011), which will aggravate this issue. Finally, some environments and lifestyles will favor organisms of certain genome sizes (Lauro *et al.*, 2009; Nikoh *et al.*, 2011; Bentkowski *et al.*, 2016; Cobo-Simón & Tamames, 2017). Since the prevalence of certain functional categories is also linked with genome size (Konstantinidis & Tiedje, 2004), selection based on genome size may indirectly enrich those functional categories, which would thus appear to be functionally redundant. In this work, we tried to decouple function from phylogeny, the environment, and genome size, in order to provide an unbiased characterization of phylogenetic and functional redundancy in environmental microbial assemblages.

We first compared the average pairwise phylogenetic and functional similarities of the microbial assemblages obtained by our approach (*environmental assemblages*) to that of random assemblages of genera (**Figure 3a,b** Random). The functional and phylogenetic distances in the environmental assemblages (blue and green boxplots in **Figure 3a,b**) were significantly lower than expected by chance (**Figure 3a,b**, Random vs Real, single environment), suggesting the existence of phylogenetic and functional redundancy. Furthermore, the assemblages that were detected in more than one environment (green boxplots) had a higher functional redundancy than single-environment ones (blue boxplots), pointing to a relationship between functional redundancy and the ability to cope with environmental change.

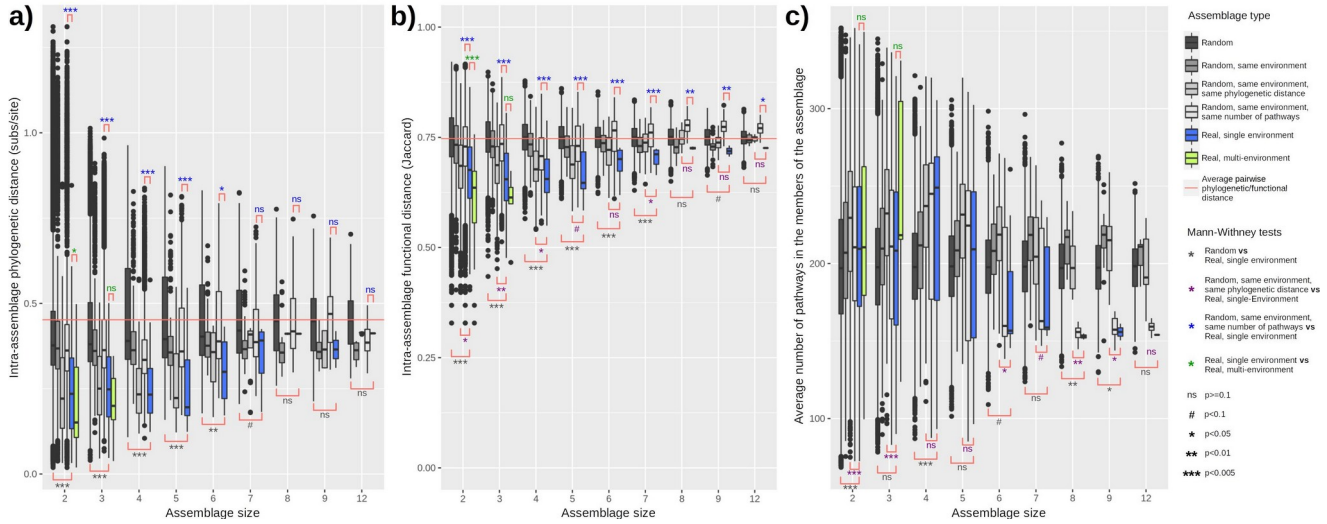


**Figure 3. Phylogeneticand functional redundancy in environmental versus random assemblages of microbial taxa.** Boxplots represent the distributions of average **a)** pairwise phylogenetic distances, **b)** pairwise functional distances and **c)** number of pathways in the members of increasingly large assemblages (x-axis). Boxplot colour shows assemblage type: **1)** fully random assemblages (dark grey), **2)** environmentally-equivalent random assemblages (medium gray), in which taxa come from the same environment, **3)** environmentally/phylogenetically equivalent assemblages, with taxa from the same environment and average phylogenetic distances similar to those found in environmental assemblages (grey), **4)** environmentally/genome size equivalent random assemblages, with taxa from the same environment and with the same average number of pathways as the environmental assemblages, **5)** environmental assemblages appearing in only one environment (blue), or **6)** environmental assemblages appearing in more than one environment (green). Significant differences between different types or assemblages were evaluated with the Mann-Whitney U test. Horizontal red lines represent the average pairwise phylogenetic or functional distance of all the genera included in our network.

We then aimed to control for possible confounding factors by creating random assemblages in which the genera came from the same environmental subtype, which is the most detailed environmental classification in the microDB database (differentiating for example between coastal, open and deep marine samples, see Pignatelli *et al.*, 2009 for details). After doing this, we further controlled the random assemblages so that their average phylogenetic similarities were the same as for the environmental assemblages (**Figure 3a,b**, Random, same environment, same phylogenetic distance). These phylogenetically-equivalent random assemblages had a higher functional redundancy (i.e., lower average distance) than completely random assemblages, which was expected since phylogenetically related organisms tend to be functionally similar (Tamames *et al*., 2016). However, the functional redundancy in the environmental assemblages was significantly higher than in these phylogenetically equivalent assemblages, showing that natural communities are constituted by organisms that are more functionally redundant than expected from their phylogenies.

Regarding the average number of pathways per genus (used here as a proxy for genome size), it was reduced for larger assemblages, in a behavior that deviated from that of the random assemblages (**Figure 3c**). In order to control for this factor, we created random assemblages in which the average number of pathways per genus was similar to that of the environmental assemblages (**Figure 3a,b**, Random, same environment, same number of pathways). Functional and phylogenetic redundancy was significantly higher in the environmental assemblages than in these genome-size-equivalent random assemblages, showcasing once again the apparent prevalence of phylogenetic and functional redundancy in environmental communities.

## Relationship between pathway redundancy, pathway specificity and community size in environmental microbial assemblages

The results presented in the previous section obeyed to assemblage-wide selection patterns, but we were also interested in the selection pressures affecting individual metabolic pathways. Selection may result in pathway specificity (i.e., a pathway appearing only in one member of an assemblage), due to competitive exclusion effects (only the best competitor for a contested resource involving that pathway is present in the community) or cooperative interactions (a complex route being divided among different organisms, or a common good being supplied by one member of the community; Morris *et al.*,

2012). Conversely, a metabolic pathway will have low specificity (i.e., it will be redundant) if it is required by most or all members of a microbial assemblage, as would happen for housekeeping pathways, or for pathways selected by a common abiotic constraint in a given environment (e.g., anoxia).

In order to investigate whether individual metabolic pathways are more redundant or specific in environmental assemblages than expected by chance, we first computed the number of times that each metabolic pathway appeared on each of the microbial assemblages obtained through our algorithm. We then compared these results to those obtained on 1,000 control assemblages with the same number of taxa, randomly assembled from taxa that belonged to the same environmental class as the real community and have similar pairwise phylogenetic distances (see Methods). Pathways whose prevalence in a real assemblage was more extreme (either higher or lower) than on 95% of the random control assemblages were subjected to further scrutiny.

We classified each metabolic pathway according to their presence in the members of the microbial community in one of the three following classes. (1) Missing, if the pathway is absent from all members but present in the random assemblages, suggesting that it is not needed in the habitat in which the community lives. (2) Specific, if it is present in at least one member of the real community, but in less members than in the random communities. The biochemical products of specific pathways are candidate for being shared in the community through cross-feeding interactions. (3) Redundant, if the pathway is possessed by more members of the real community than expected by chance, as expected for a capability that is useful in the given habitat and is seldom shared through cross-feeding.

The heatmap in **Figure 4a** shows the distribution of redundant, specific and missing pathways in the microbial assemblages detected by our approach. A hierarchical clustering of the assemblages based on the content of redundant, specific and missing pathways showed no clear relationship with their source environment (**Figure 4a**, color legend at the y axis). This suggests that we successfully controlled for biases coming from the source environment in our analysis, and that our results obey to other, more universal causes.

The number of pathways of the three types belonging to different MetaCyc categories is shown in **Figure 4b**. For most categories redundant pathways prevail, in particular pathways that belong to categories of energy metabolism, such as carbohydrate degradation, electron transfer, respiration, and

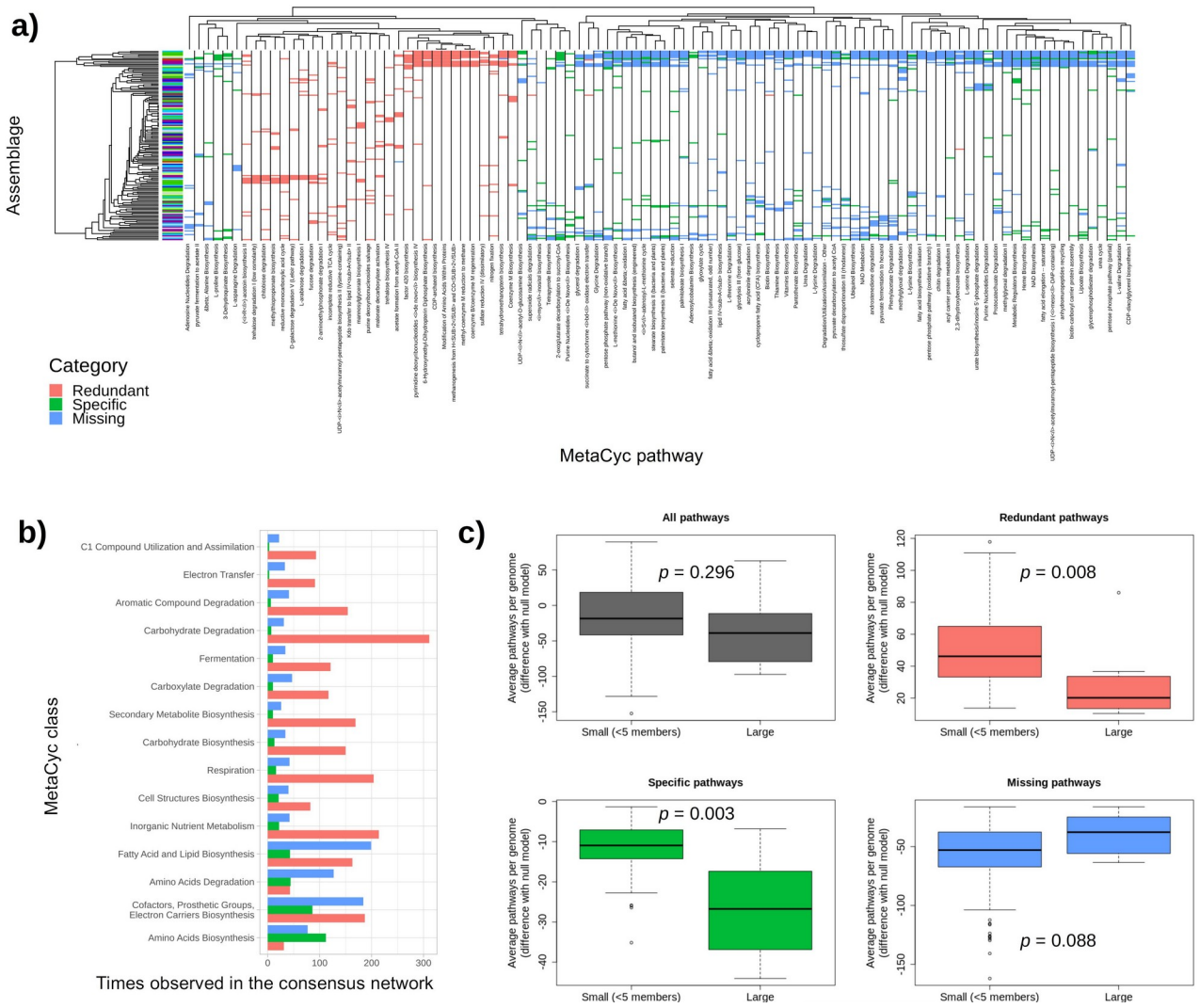


**Figure 4. Signs of selection in individual metabolic pathways a)** MetaCyc pathways redundant, specific and missing in the consensus network. Only pathways that were redundant, specific or missing in 10 or more communities are shown. The color legend in the y-axis dendrogram shows the source environment for each community, following the color code shown in **Figure 2.** Multi-environment communities are colored in light green. **b)** Number of times each MetaCyc class was redundant, specific and missing in the consensus network. Only the 15 MetaCyc classes with the highest deviation from the random communities are shown; **c)** Differences in average pathways per genome between environmental assemblages and control assemblages (null models). Environmental and control assemblages are separated in small (< 5 members) and large (5+ members). First panel (grey) shows the total differences between the real and the control assemblages, the other three (red, green, blue) show the contribution of redundant, specific, and missing pathways to the total differences.

carboxylate degradation. Pathways of inorganic nutrient metabolism, carbohydrate biosynthesis and secondary metabolite biosynthesis also tend to be redundant. We hypothesize that these pathways are redundant because they favor the use of the resources available in the given habitat, also consistent with the fact that the second most frequent type of these categories is "Missing". In contrast, in the category "Biosynthesis of amino acids" specific pathways prevail, and in the Biosynthesis of "cofactors, prosthetic groups, electron carriers", "fatty acid and lipids" and in "amino acid degradation" the pathways tend to be missing or specific. These results are consistent with our interpretation of specific and redundant pathways presented above.

An interesting result, presented in **Figure 3c**, is that environmental assemblages have on the average smaller genomes than expected by chance, particularly if they contain many members. To further explore this observation, we show in **Figure 4c** the difference in average pathways per genome (proxy of genome size) between the environmental and the control assemblages, for both small (< 5 members) and large (5+ members) assemblages. The figure reveals that large assemblages are indeed characterized by genomes with fewer pathways. In order to assess which types of pathways are responsible for this genome reduction, we separately considered redundant, specific, and missing pathways (**Figure 4c**).

Redundant pathways produced an increase of the number of pathways with respect to the control community, but this increase was not uniform: redundant pathways contribute 50 additional pathways per genome in small communities (with 4 or fewer members), but only 20 pathways per genome in large communities (**Figure 4c**, red). This difference is significant (Wilcoxon test, $p = 0.008$), and it might be attributed to interactions between species, suggesting that some of the members of large communities may benefit from the leakiness of some of the products of these otherwise redundant pathways. In contrast, missing pathways, which are also influenced by habitat filtering but cannot be shared because they are not present in the community, are not significantly different between small and large communities (the average reduction of the number of pathways is 50 and 40 respectively, **Figure 4c**, blue; Wilcoxon test, $p = 0.09$), supporting the idea that the comparison between small and large communities yield information about community interactions.

Interestingly, specific pathways produce on the average a reduction of 10 pathways per genome in small communities and 25 pathways per genome in large assemblages (**Figure 4c**, green). This

difference is highly significant (Wilcoxon test, p = 0.003), despite the small number of large communities that we detected, and it is consistent with the results from redundant pathways. A possible interpretation is that large communities offer a larger variety of "public goods" that are shared by the members of the community, and that these conditions allow reduced metabolic cost and genomic streamlining, which act as a selective force favoring the formation and maintenance of these large communities. This interpretation is consistent with recent simulation studies (Thommes *et al.*, 2019; Wang *et al.*. 2021) and with our observations that amino acid biosynthesis is the biochemical class whose pathways are most frequently specific (see **Figure 4b**) and that the fraction of communities in which the biosynthesis of a given amino acid is specific is significantly correlated with its biochemical cost (**Figure 5**; next section). Overall, this suggests that the reduction of the biosynthetic cost is a relevant selective pressure behind the reduction of the number of pathways.

**Patterns of amino acid auxotrophy in environmental microbial assemblages**

As discussed above, environmental microbial communities are more functionally redundant than expected by chance (**Figure 2b**). In spite of this, it is also true that some pathways tend to appear in fewer members of the community, which we hypothesize is due to biotic effects. Microorganisms are well known to engage in complex interactions (Goldford *et al.*, 2018*)*, among which auxotrophy and cross-feeding are perhaps the most studied (Zengler & Zaramela, 2018). We therefore focused on the redundancy/specificity profiles of pathways related to amino acid biosynthesis, as they are the one of the metabolites most usually involved in such processes (Embree *et al*., 2015).

In order for auxotrophy to be a viable strategy, the potential benefits must be higher than the drawbacks derived from the resulting loss of autonomy (Oliveira *et al*., 2014). Accordingly, the environmental assemblages captured in our study contained more auxotrophs for expensive amino acids than for cheap ones, with the exception of tryptophan (**Figure 5,** $p = 0.049$ for all amino acids, $p < 0.001$ after removing tryptophan). The comparatively lower prevalence of tryptophan auxotrophs can be explained due to its tight regulation: not only is the use of this expensive amino acid minimized across the proteome (Akashi & Gojobori, 2002), but it is also seldom leaked into the environment (Mopper & Lindroth, 1982; Zomorrodi & Segrè, 2017). The difficulty of finding free tryptophan in nature might thus partly negate the potential benefits of auxotrophy. On the other hand, since tryptophan is only

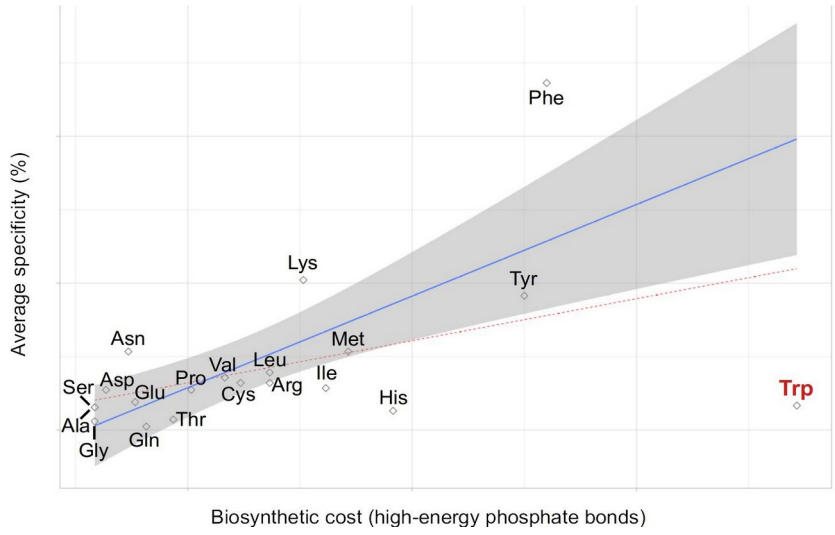


**Figure 5. Average specificity vs biosynthetic cost of amino acid biosynthesis pathways.** Blue line: linear regression model of specificity vs cost for all amino acids except for tryptophan ($R^2$=0.49, $p$ < 0.001). Grey area: 95% confidence interval for the linear models. Red dashed line: linear regression model of specificity vs cost for all amino acids including tryptophan ($R^2$=0.15, $p$=0.049). Amino acid biosynthesis costs were obtained from Akashi & Gojobori (2002).

required in small amounts, these benefits will be lower than otherwise suggested by its per-molecule biosynthetic cost.

These observations are consistent with those of Mee *et al.*, 2014, which similarly reported a larger prevalence of auxotrophy and cross-feeding for expensive amino acids. We note that this does not preclude the exchange of cheap amino acids as described in Wintermute & Silver, 2010. However, this exchange might not result in the emergence of auxotrophy, as the low cost of the exchanged metabolite might not be enough to offset the penalties associated with autonomy loss.

# Conclusions

We presented a cross-study, cross-biome meta-analysis of microbial occurrence data in more than 5,000 samples from ten different environments, using a novel network generation algorithm aimed to capture the conditional interactions that commonly appear in environmental microbial communities. This top-down approach complements the work already developed in synthetic communities (Wintermute & Silver, 2010; Mee *et al.*, 2014; Zomorrodi & Segrè, 2017), since it builds upon data from real environmental communities, and summarizes complex dynamics that may be difficult to replicate in experimental settings. For example, the establishment of cross-feeding interactions is expected to be subject to cost-to-benefit balance. However, the cost of the same metabolite is often context-dependent

and can vary widely across microbial species and environments (Pacheco *et al.*, 2019). Microbial communities can also have different degrees of spatial structuring, which affects the range of beneficial interactions that can be established (Germerodt *et al*., 2016). Microbial diversity is another key factor that influences community assembly, due to its effect on stability. A diverse community will have different species that perform the same function, and this functional redundancy will make such communities more resistant to perturbations (Shade *et al*., 2012). Additionally, the increased number of potential partners facilitates the establishment of weak interactions (Johnson *et al.*, 2020), which in turn allow for the development of mutualism without compromising community stability (McCann, *et* al., 1998; Butler & O'Dwyer, 2018).

In spite of the wide range of ecosystems analyzed in this study, we were able to detect consistent patterns of functional redundancy and auxotrophy, hinting at the existence of conserved, biome-agnostic principles governing the assembly of microbial communities. We found that functional redundancy is ubiquitous in environmental microbial communities, and it is, at least partly, decoupled from phylogeny. We hypothesize that it is driven by environmental selection for some biochemical processes. We also discovered that the number of biochemical pathways per genome (which is correlated with genome size) is negatively correlated with the size of the microbial community. This observation hints at interactions between members of the community, and in particular at "labor specialization", i.e. the possibility that some leaky biochemical functions possessed by some members of the community are exploited by other members, allowing them to reduce their biochemical work and their genome size. This labor specialization would generate a potential selective force behind the maintenance of large communities, as suggested by recent theoretical (Thommes *et al.*, 2019; Wang *et al.*, 2021) and observational studies (Anantharaman et al., 2016; Castelle *et al.*, 2018; Lannes *et al.*, 2019). In agreement with this interpretation, our results suggest that, in spite of the prevalence of functional redundancy, auxotrophy commonly occurrs in environmental microbial communities, particularly for costly compounds.

Overall, our results show that redundancy and auxotrophy are not mutually exclusive, but rather coexist in microbial communities from different origins. Combining a background of functional redundancy with cooperation in the biosynthesis of key nutrients might thus be a useful design principle for engineering more robust microbial communities in the future.

# Materials and methods

## Description of the data set

We obtained the data from the microDB database (formerly envDB, http://botero.cnb.csic.es/envDB) (Pignatelli *et al.*, 2009), following the procedure in Tamames *et al.*, (2010). The database comprises more than 20,000 environmental samples and their associated 16S rRNA gene sequences, with each sample classified in a unique environment, thus informing of the presence or absence of taxa across a wide range of ecosystems. The genus level was chosen as the taxonomic working unit because it provided a good balance between the taxonomic resolution, the ability to accurately classify partial fragments of 16S coming from different regions, and the sparsity of the observations. In this study, we only considered samples coming from the following environments (as defined in the microDB classification): freshwater, marine water, marine sediments, hypersaline, oil, thermal, hypothermal/polar, soils, host-associated and water-treatment plants. In order to more reliably aggregate results from studies that used very different methodologies, data was binarized into a matrix that recorded the presence/absence of genera across samples. Samples with less than five genera and genera present in less than five samples were excluded from further analysis. This left a total of 966 genera distributed across 5369 samples.

## Detection of significant associations between pairs of taxa

For a given pair of taxa *i* and *j* that co-occurr in *N* out of *M* samples, we define its aggregation score $S_{ij}$, which represents their propensity to appear together in the same samples, as the negative logarithm of the conditional probability of i and j co-occurring in more than *N* out of *M* samples. The original implementation of the aggregation scores can be found in Pascual-García *et al.* (2014), the implementation used in this work is detailed in **Supplementary Note S1**. Briefly, we used the null model from Navarro-Alberto & Manly (2009) that estimates the probability that a given taxa is observed in a given sample under the assumption of no interaction between taxa. We developed a different null model in each of the ten studied environments. After inferring the parameters of the null models, we used them to generate 1000 random presence-absence matrices with the same row and column totals as the real matrix. These random matrices allow to assess the influence of cosmopolitanism (i.e. the number of samples in which taxa were present) into the aggregation scores.

To obtain the aggregation scores, we calculated he probability that two taxa co-occur in the number of samples observed following the algorithm in **Supplementary Note S1**. Aggregation scores were then transformed to Z-scores related to the mean and standard deviation of the null aggregation scores of pairs of taxa with similar cosmopolitanism. Finally, we derived a Z-score cutoff from the distribution of null Z-scores such that the False Positive Rate (i.e., the rate of significant aggregations in the null model) was not larger than 0.0001. Pairs of taxa with a Z-score higher than the cutoff were deemed significantly associated in our samples.

**Network generation**

We generated an ecological network representing significant associations between groups of taxa across multiple environments through the following steps:

1. For each of the ten environments included in this study:

    1.1. Compute aggregation Z-scores between pairs of taxa *i* in samples *a* from the binary presence-absence matrix $X_{ia}$ and the probability matrix $\pi_{ia}$ as described in **Supplementary Note S1**.

    1.2. Create 100 independent networks (in order to minimize path dependency during the clustering process) applying the following clustering procedure. We will refer generically to "nodes" for both individual taxa (e.g. elements at the beginning of the algorithm) and assemblages (taxa clustered together):

        1.2.1. While there are significantly associated pairs of nodes appearing together in more than 5 samples:

            1.2.1.1. Select one significantly associated pair *i*,*j* at random, weighted by its aggregation Z-score so that pairs with higher aggregation scores are more likely to be selected.

            1.2.1.2. Create a new node *k* that represents the aggregation of the selected pair of nodes *i*,*j* in the samples in which they appear together, with $X_{ka} = X_{ia} \cdot X_{ja}$ and $\pi_{ka} = \pi_{ia} \cdot \pi_{ja}$.

            1.2.1.3. Create the links $i \rightarrow k$ and $j \rightarrow k$.

1.2.1.4. Replace the values for *i* and *j* in the presence absence matrix and in the probability matrix, so that they represent the presence of *i* and *j* in the samples in which they do not appear together, with $X_{i'a} = X_{ia} \cdot (1 - X_{ja})$, $\pi_{i'a} = \pi_{ia} \cdot (1 - \pi_{ja})$, $X_{j'a} = X_{ja} \cdot (1 - X_{ia})$ and $\pi_{j'a} = \pi_{ja} \cdot (1 - \pi_{ia})$.

1.2.1.5. Recalculate the aggregation Z-scores from the new X and π matrices.

2. Combine the 1000 independent networks (100 networks from each of the 10 environments) into a single network as follows:.

2.1.1. The combined network contains all the nodes present in the individual networks. Nodes containing the same taxa in the individual networks are collapsed into a single node in the combined network.

2.1.2. All incoming and outgoing edges present in the individual networks are added to the collapsed nodes in the combined network.

2.1.3. For each node and edge, we define its *support* value as the number of individual networks in which that node or edge was observed. Nodes and edges with a support value smaller than 10 are discarded.

2.1.4. Nodes are annotated based on the source environment of the individual networks in which they were found.

**Environmental and bibliographic annotation of assemblages**

For each sample, the microDB database contains its isolation source, as originally found in the NCBI database, as well as the Pubmed ID (PMID) of any published work related to it. We annotated each assemblage representing a significant aggregation of two or more genera with the isolation sources and related PMIDs of the samples in which the genera appeared together.

**Functional annotation of assemblages and intra-assemblage functional redundancy**

We used the MetaCyc database version 19 (Caspi *et al.*, 2016) to download the predicted reactomes for all the sequenced genomes from the genera included in our network (**Supplementary Data S3**). For each genome, we predicted its metabolic pathways from its reactome using an in-house implementation

of the PathoLogic algorithm as described in Karp et al., 2011. As a deviation from the original algorithm, we did not add a more lenient prediction rule for energy metabolism pathways, as we found out that doing so would result in false positive predictions (e.g. sulfate respiration would be predicted for *Escherichia*). The fraction of genomes from each genus that contain each pathway is reported in **Supplementary Data S2**. We considered that a pathway is present in a genus if it is predicted in at least 25% of the complete genomes from that genus. We chose this threshold to reduce false positives due to pathways wrongly predicted in only few genomes within the genus. We then defined the pathways present in an assemblage $\{R\}_a$ as the set union of the pathways present in its constituent genera. We also defined the average pairwise functional distance of an assemblage as the average of the of the all-against-all Jaccard dissimilarities (1 – the Jaccard Index; Jaccard, 1912) between the pathway vectors of its constituent genera.

**Phylogenetic distance between genera and intra-assemblage phylogenetic distances**

We used 16S rRNA sequences from the GreenGenes database (DeSantis et al., 2006) to obtain estimates of the phylogenetic distances between genera. First, we selected a representative full-length 16S sequence for each prokaryotic species in the database, usually the type strain. Then, we calculated the distance between the aligned sequences as the number of substitutions per site using RaxML with a GTRGAMMA model (Stamatakis, 2014). We calculated distances between genera as the median of the distances between the species belonging to those genera. We then calculated the average pairwise phylogenetic distance between the constituent genera of each assemblage.

**Detection of significant functional and phylogenetic redundancies at different assemblage sizes**

For each assemblage size, ranging from 2 to 12 genera (the largest assemblage present in our graph for which all genera could be annotated) we compared the average functional and phylogenetic distance distributions of the assemblages present in our network to those of random assemblages of the same genera. Assemblages in which one or more genera could not be functionally annotated were ignored for this and subsequent computations. Multi-environment assemblages (i.e. assemblages of genera that were considered significant in more than one environment during our clustering process) were treated separately from single-environment ones. For each real assemblage, we generated four different kinds of random assemblages:

a) 1000 random assemblages with the same size.

b) 100 environmentally-equivalent random assemblages with the same size of the real assemblage, such that their genera came from the same environmental subtype (i.e. the finest environmental classification available in the microDB database, see Pignatelli *et al*., 2009).

c) 100 environmentally/phylogenetically - equivalent random assemblages with the same size, such that their genera came from the same environmental subtype and the average pairwise phylogenetic distances in the random assemblages differed by 0.05 substitutions per position or less from the average pairwise phylogenetic distance of the original assemblage. This was done in order to assess whether the functional redundancy was explained by phylogenetic similarity and source environment alone.

d) 100 environmentally/genome size – equivalent random assemblages with the same size,, such that their genera came from the same environmental subtype and the average number of pathways per genus differed by 20% of less from the average number of pathways in the original assemblage.

We assessed significant differences between different types or assemblages with the Mann-Whitney U test.

**Detection of redundant and specific pathways in the assemblages of our network**

The procedure described above provided us with a per-assembly estimate of functional redundancy, but we were also interested in assessing functional redundancy on a per-pathway basis. For this, we first selected a subset of the network connected by highly supported (support > 70) edges. We then selected the terminal assemblages with no outgoing edges to larger assemblages, which represent the sink nodes of our clustering algorithm. For each of these assemblages, we then generated 1,000 phylogenetically and environmentally equivalent random assemblages (see previous section). In order to obtain a higher number of valid random assemblages, we increased the maximum difference in phylogenetic distances from 0.01 to 0.1 substitutions per position. Then, for each metabolic pathway, we compared its prevalence in the real assemblage with its prevalence in the random assemblages and classified it into one three categories:

1. Redundant, if its prevalence in the real assemblages was higher than its prevalence in 95% of

the random assemblages.

2. Specific, if its prevalence in the real assemblage was lower than its prevalence in 95% of the random assemblages.

3. Missing, if it was missing from the real assemblage, but present in 95% of the random assemblages.

Finally, for each metabolic pathway, we computed its *average specificity* as 1-(P/T), where P is the sum of its prevalence in the individual assemblages, and T is by the sum of the sizes of those assemblages. This value will become higher as more auxotrophs for the pathway exist in the environmental assemblages.

## Acknowledgements

AP-G was supported by the Simons Collaboration: Principles of Microbial Ecosystems (PriME), award number 542381. UB was supported through the grant PID2019-109041GB-C22/10.13039/501100011033 of the Spanish Agency of Research (AEI). Research at the CBMSO is facilitated by the Fundación Ramón Areces. FP-S was funded by grant CTM2016-80095-C2-1-R / NOVAMAR from the Spanish Ministerio de Economía y Competitividad and the the Marie Skłodowska-Curie grant agreement No 892961 from the European Union's Horizon 2020 research and innovation programme. The authors declare no conflict of interests.

## Data and code availability

The data used for this manuscript and the code used for analysis are publicly available in https://github.com/fpusan/cross-biome-microbial-networks.

# References


1. Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, 201711842.
2. Locey, K. J., & Lennon, J. T. (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences*, *113*(21), 5970-5975.
3. Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science*, *320*(5879), 1034-1039.
4. Konopka, A., Lindemann, S., & Fredrickson, J. (2015). qqqqqqqDynamics in microbial communities: unraveling mechanisms to identify principles. *The ISME journal*, *9*(7), 1488-1495.
5. Louca, S., Polz, M. F., Mazel, F., Albright, M. B., Huber, J. A., O'Connor, M. I., ... & Doebeli, M. (2018). Function and functional redundancy in microbial systems. *Nature ecology & evolution*, 1.
6. Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly review of biology*, *85*(2), 183-206.
7. Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., ... & Ferrenberg, S. (2013). Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews*, *77*(3), 342-356.
8. Louca, S., Jacques, S. M., Pires, A. P., Leal, J. S., Srivastava, D. S., Parfrey, L. W., ... & Doebeli, M. (2017). High taxonomic variability despite stable functional structure across microbial communities. *Nature ecology & evolution*, *1*(1), 0015.
9. Chuang, J. S., Rivoire, O., & Leibler, S. (2009). Simpson's paradox in a synthetic microbial system. *Science*, *323*(5911), 272-275.
10. Ribeck, N., & Lenski, R. E. (2015). Modeling and quantifying frequency-dependent fitness in microbial populations with cross-feeding interactions. *Evolution*, *69*(5), 1313-1320.
11. Yu, Z., Beck, D. A., & Chistoserdova, L. (2017). Natural selection in synthetic communities highlights the roles of Methylococcaceae and Methylophilaceae and suggests differential roles for alternative methanol dehydrogenases in methane consumption. *Frontiers in microbiology*, *8*, 2392.
12. Dolinšek, J., Goldschmidt, F., & Johnson, D. R. (2016). Synthetic microbial ecology and the dynamic interplay between microbial genotypes. *FEMS microbiology reviews*, *40*(6), 961-979.
13. Yu, Z., Krause, S., Beck, D. A., & Chistoserdova, L. (2016). A synthetic ecology perspective: How well does behavior of model organisms in the laboratory predict microbial activities in natural habitats?. *Frontiers in microbiology*, *7*, 946.
14. Ehsani, E., Hernandez-Sanabria, E., Kerckhof, F. M., Props, R., Vilchez-Vargas, R., Vital, M., ... & Boon, N. (2018). Initial evenness determines diversity and cell density dynamics in synthetic microbial ecosystems. *Scientific reports*, *8*(1), 340.
15. Datta, M. S., Sliwerska, E., Gore, J., Polz, M. F., & Cordero, O. X. (2016). Microbial interactions lead to

rapid micro-scale successions on model marine particles. *Nature communications*, *7*(1), 1-7.

16. Rivett, D. W., & Bell, T. (2018). Abundance determines the functional role of bacterial phylotypes in complex communities. *Nature microbiology*, *3*(7), 767-772.

17. Enke, T. N., Datta, M. S., Schwartzman, J., Cermak, N., Schmitz, D., Barrere, J., ... & Cordero, O. X. (2019). Modular assembly of polysaccharide-degrading marine microbial communities. *Current Biology*, *29*(9), 1528-1535.

18. Pascual-García, A., & Bell, T. (2020). Community-level signatures of ecological succession in natural bacterial communities. *Nature communications*, *11*(1), 1-11.

19. Ma, B., Wang, Y., Ye, S., Liu, S., Stirling, E., Gilbert, J. A., ... & Xu, J. (2020). Earth microbial co-occurrence network reveals interconnection pattern across microbiomes. *Microbiome*, *8*(1), 1-12.

20. O'hara, R. B. (2005). The anarchist's guide to ecological theory. Or, we don't need no stinkin'laws. *Oikos*, *110*(2), 390-393.

21. Bashan, A., Gibson, T. E., Friedman, J., Carey, V. J., Weiss, S. T., Hohmann, E. L., & Liu, Y. Y. (2016). Universality of human microbial dynamics. *Nature*, *534*(7606), 259.

22. Goldford, J. E., Lu, N., Bajić, D., Estrela, S., Tikhonov, M., Sanchez-Gorostiaga, A., ... & Sanchez, A. (2018). Emergent simplicity in microbial community assembly. *Science*, *361*(6401), 469-474.

23. Gibson, T. E., Bashan, A., Cao, H. T., Weiss, S. T., & Liu, Y. Y. (2016). On the origins and control of community types in the human microbiome. *PLoS computational biology*, *12*(2), e1004688.

24. Pignatelli, M., Moya, A., & Tamames, J. (2009). EnvDB, a database for describing the environmental distribution of prokaryotic taxa. *Environmental Microbiology Reports*, *1*(3), 191-197.

25. Pascual-García, A., Tamames, J., & Bastolla, U. (2014). Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions?. *BMC microbiology*, *14*(1), 284.

26. Navarro-Alberto, J. A., & Manly, B. F. (2009). Null model analyses of presence–absence matrices need a definition of independence. *Population ecology*, *51*(4), 505-512.

27. Karp, P. D., Latendresse, M., & Caspi, R. (2011). The pathway tools pathway prediction algorithm. *Standards in genomic sciences*, *5*(3), 424-429.

28. Tamames, J., Sánchez, P. D., Nikel, P. I., & Pedrós-Alió, C. (2016). Quantifying the relative importance of phylogeny and environmental preferences as drivers of gene content in prokaryotic microorganisms. *Frontiers in microbiology*, *7*, 433.

29. Bell, T., Newman, J. A., Silverman, B. W., Turner, S. L., & Lilley, A. K. (2005). The contribution of species richness and composition to bacterial services. *Nature*, *436*(7054), 1157.

30. Wertz, S., Degrange, V., Prosser, J. I., Poly, F., Commeaux, C., Freitag, T., ... & Roux, X. L. (2006). Maintenance of soil functioning following erosion of microbial diversity. *Environmental microbiology*, *8*(12), 2162-2169.

31. Jones, B. V., Begley, M., Hill, C., Gahan, C. G., & Marchesi, J. R. (2008). Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proceedings of the National Academy of Sciences*, *105*(36), 13580-13585.

32. Louca, S., Parfrey, L. W., & Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science*, *353*(6305), 1272-1277.

33. Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., & Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences*, *112*(20), 6449-6454.

34. Strickland, M. S., Lauber, C., Fierer, N., & Bradford, M. A. (2009). Testing the functional significance of microbial community composition. *Ecology*, *90*(2), 441-451.

35. Peter, H., Beier, S., Bertilsson, S., Lindström, E. S., Langenheder, S., & Tranvik, L. J. (2011). Function-specific response to depletion of microbial diversity. *The ISME journal*, *5*(2), 351.

36. Fetzer, I., Johst, K., Schäwe, R., Banitz, T., Harms, H., & Chatzinotas, A. (2015). The extent of functional redundancy changes as species' roles shift in different environments. *Proceedings of the National Academy of Sciences*, *112*(48), 14888-14893.

37. Delgado-Baquerizo, M., Giaramida, L., Reich, P. B., Khachane, A. N., Hamonts, K., Edwards, C., ... & Singh, B. K. (2016). Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. *Journal of ecology*, *104*(4), 936-946.

38. Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C., & Debroas, D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *The ISME journal*, *12*(10), 2470-2478.

39. Martiny, A. C., Treseder, K., & Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *The ISME journal*, *7*(4), 830-838.

40. Morrissey, E. M., Mau, R. L., Schwartz, E., Caporaso, J. G., Dijkstra, P., van Gestel, N., ... & Hungate, B. A. (2016). Phylogenetic organization of bacterial activity. *The ISME journal*, *10*(9), 2336-2340.

41. Tamames, J., Abellán, J. J., Pignatelli, M., Camacho, A., & Moya, A. (2010). Environmental distribution of prokaryotic taxa. *BMC microbiology*, *10*(1), 1-14.

42. Nemergut, D. R., Costello, E. K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S. K., ... & Knight, R. (2011). Global patterns in the biogeography of bacterial taxa. *Environmental microbiology*, *13*(1), 135-144.

43. Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., ... & Cavicchioli, R. (2009). The genomic basis of trophic strategy in marine bacteria. *Proceedings of the National Academy of Sciences*, *106*(37), 15527-15533.

44. Nikoh, N., Hosokawa, T., Oshima, K., Hattori, M., & Fukatsu, T. (2011). Reductive evolution of bacterial genome in insect gut environment. *Genome biology and evolution*, *3*, 702-714.

45. Bentkowski, P., Van Oosterhout, C., & Mock, T. (2015). A model of genome size evolution for

prokaryotes in stable and fluctuating environments. *Genome biology and evolution*, *7*(8), 2344-2351.

46. Cobo-Simón, M., & Tamames, J. (2017). Relating genomic characteristics to environmental preferences and ubiquity in different microbial taxa. *BMC genomics*, *18*(1), 1-11.

47. Konstantinidis, K. T., & Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proceedings of the National Academy of Sciences*, *101*(9), 3160-3165.

48. Morris, J. J., Lenski, R. E., & Zinser, E. R. (2012). The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, *3*(2), e00036-12.

49. Thommes M, Wang T, Zhao Q, Paschalidis IC, Segrè D (2019). Designing Metabolic Division of Labor in Microbial Communities. mSystems. 4:e00263-18. doi: 10.1128/mSystems.00263-18.

50. Wang M, Liu X, Nie Y, 2, Xiao-Lei Wu XL (2021). Selfishness driving reductive evolution shapes interdependent patterns in spatially structured microbial communities. ISME J. 15:1387-1401. doi: 10.1038/s41396-020-00858-x.

51. Zengler, K., & Zaramela, L. S. (2018). The social network of microorganisms—how auxotrophies shape complex communities. *Nature Reviews Microbiology*, *16*(6), 383-390.

52. Embree, M., Liu, J. K., Al-Bassam, M. M., & Zengler, K. (2015). Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences*, *112*(50), 15450-15455.

53. Oliveira, N. M., Niehus, R., & Foster, K. R. (2014). Evolutionary limits to cooperation in microbial communities. *Proceedings of the National Academy of Sciences*, *111*(50), 17941-17946.

54. Mopper, K., & Lindroth, P. (1982). Diel and depth variations in dissolved free amino acids and ammonium in the Baltic Sea determined by shipboard HPLC analysis 1. Limnology and Oceanography, 27(2), 336-347.

55. Zomorrodi, A. R., & Segrè, D. (2017). Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. *Nature Communications*, *8*(1), 1563.

56. Mee, M. T., Collins, J. J., Church, G. M., & Wang, H. H. (2014). Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences*, *111*(20), E2149-E2156.

57. Wintermute, E. H., & Silver, P. A. (2010). Emergent cooperation in microbial metabolism. *Molecular systems biology*, *6*(1), 407.

58. Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proceedings of the National Academy of Sciences*, *99*(6), 3695-3700.

59. Pacheco, A. R., Moel, M., & Segrè, D. (2019). Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature communications*, *10*(1), 1-12.

60. Germerodt, S., Bohl, K., Lück, A., Pande, S., Schröter, A., Kaleta, C., ... & Kost, C. (2016). Pervasive selection for cooperative cross-feeding in bacterial communities. *PLoS computational biology*, *12*(6),

e1004986.

61. Shade, A., Peter, H., Allison, S. D., Baho, D., Berga, M., Bürgmann, H., ... & Handelsman, J. (2012). Fundamentals of microbial community resistance and resilience. *Frontiers in microbiology*, *3*, 417.

62. Johnson, W. M., Alexander, H., Bier, R. L., Miller, D. R., Muscarella, M. E., Pitz, K. J., & Smith, H. (2020). Auxotrophic interactions: A stabilizing attribute of aquatic microbial communities?. *FEMS Microbiology Ecology*.

63. McCann, K., Hastings, A., & Huxel, G. R. (1998). Weak trophic interactions and the balance of nature. Nature, 395(6704), 794-798.

64. Butler, S., & O'Dwyer, J. P. (2018). Stability criteria for complex microbial communities. *Nature communications*, *9*(1), 1-10.

65. Caspi, R., Billington, R., Ferrer, L., Foerster, H., Fulcher, C. A., Keseler, I. M., ... & Ong, Q. (2016). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, *44*(D1), D471-D480.

66. Jaccard, P. (1912). The distribution of the flora in the alpine zone. 1. *New phytologist*, *11*(2), 37-50.

67. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., ... & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology*, *72*(7), 5069-5072.

68. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

69. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

70. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 24;7:13219. doi: 10.1038/ncomms13219.

71. Castelle CJ, Brown CT, Anantharaman K, Probst AJ, Huang RH, Banfield JF (2018) Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. Nat Rev Microbiol. 16:629-645. doi: 10.1038/s41579-018-0076-2.

72. Lannes R, Olsson-Francis K, Lopez P, Bapteste E (2019) Carbon Fixation by Marine Ultrasmall Prokaryotes. Gen Biol Evol. 11:1166-1177. doi: 10.1093/gbe/evz050.

# Supplementary Material

**Supplementary Note S1: Calculation of aggregation scores assessing the propensity of pairs of taxa to appear together in the same samples**

**Supplementary Table S1. General statistics on the network**

**Supplementary Figure S1. Average number of pathways per genus in environmental versus random assemblages of microbial taxa.** Boxplot color denotes assemblage type as described for **Figure 1**.

**Supplementary Data S1. Annotated network in the Cytoscape format**

**Supplementary Data S2. Fraction of genomes containing each pathway in each generality**

**Supplementary Data S3. Number of genomes per genus in the MetaCyc19 database**