# Quantifying microbial guilds

authors-labs

## ABSTRACT

The study of biological functions in microbiomes is prohibitively difficult, both in terms of information quantity and ambiguity. Using a quantifiable definition of guild would be useful for extracting general principles in the ecology of functions. The scope of this work is to provide a reimagined guild definition for the study of microbes by discriminating biostructures that are able to perpetrate a concrete function from those that are not. For functional characterization, we differentiate types of sequences both by their phylogenetic dissimilarity and by the physicochemical characteristics that condition the biocatalyst folding, allowing us to elucidate the contribution of taxonomic position and environmental convergence separatedly.

## INTRODUCTION

Microorganisms greatly modify their environment. A clear example of this is the dramatic change in atmospheric oxidation potential that occurred in the primitive Earth due to the appearance of oxygenic photosynthesis, probably during the neo-archaic period – around 2.8 Gyr ago (Cavalier-Smith 2006). This planetary shift was achieved solely through the multiplication of encoded biological functions. Unfortunately, if we rely on the assigned taxonomic position and on the automatic functional annotation, understanding how microbial functions contributed to this and other phenomena becomes challenging (Tikhonov 2017; Koskella et al. 2017).

Biological functions can be understood as the causal relationship between the physicochemical, structural information contained in a biosynthetic catalyst and the chemical interaction it facilitates on a specific substrate. This relationship may be curated by evolutionary forces or converge *de novo*, and it is environment-dependent (Huynen et al. 2000). If the environment changes, that causal relationship may be compromised or extinguished, resulting in loss-of-function (Forbes et al. 2019). In addition, the exploitation of the resource space is not due to a unique optimal biocatalyst, but rather there is a remarkable variability among them (Dourado et al. 2021), usually enzymes with prosthetic groups, which radiate through processes of drift (Masel 2011; Lynch et al. 2016), and duplication (Altenhoff and Dessimoz 2012).

Most functions are inherited vertically and, therefore, taxonomically related organisms will often share very similar set of functions (Baiser and Lockwood 2011). However, there are alternatives to vertical inheritance. On the one hand, horizontal transfer (van de Guchte 2017) has been observed even among bionts with markedly different taxonomic positions (Husnik and McCutcheon 2018). On the other hand, fundamentally dissimilar biostructures compromising the same function may emerge convergently populating the global biome with uneven, heterogenous patterns (Pagé et al. 2008; Storz 2016). Therefore, taxonomic position alone cannot predict the occurrence of some functions.

This is why we long for a non-taxonomic approach to explain the ecology of microbial functions. The ecological guild concept ~~would~~ could solve the latter problem. A guild is classically defined as: *"a*

*group of species that exploit the same class of environmental resources in a similar way (. . . ) without regard to taxonomic position, that overlap significantly in their niche requirements"* (Root 1967). Thus, guilds shall be interpreted as the functional modules into which the biocenosis can be subdivided.

The guild concept was designed for macroecology and became fashionable in the 70s. This viewpoint triggered a major inquiry into niche partitioning, so that different species would compete for the same resource. Consequently, and by way of example, all insect predators are now often regarded as members of the *insectivore* guild, regardless of the taxonomic group they currently belong to (Koran and Kropil 2014; Nebel et al. 2010).

The guild definition should help us to quantify and classify, taking into account the available techniques, the ecologically relevant information on microbes. Nevertheless, the classical definition does not completely fit the needs of microbial ecology. In macrofauna, guilds are defined by feeding behaviors (Hohberg 2003). These behaviors have to do with body plan, very complex genetic interactions leading to ethology, and nurturally transmitted information (Chiel and Beer 1997; Hillis and Mallory 1996). However, microbial feeding phenotypes are closer to their genotypes (Torsvik and Øvreås 2002). This is explained by the immediacy of their metabolism, which is focused on transforming the shared external environment locally (Paerl and Pinckney 1996; Shapiro 1998), as opposed to multicellulars, which instead try to regulate the internal environment (Wangemann and Schacht 1996) to withstand large extrinsic environment oscillations (Nemeskéri and Helyes 2019). In addition, we find that a single microorganism can belong to many different guilds. This is not so much the case in higher organisms, since functions affecting environment are typically implemented with fewer required elements in microbes (Gregory 2005; Gregory and DeSalle 2005). However, microbial communities also achieve functional complexity with environmental interaction networks. (Sanchez et al. 2022).

Despite the above considerations, many researchers have tried to use the classic guild concept because of its usefulness to explain the functional complexity of microbiomes (Veshareh and Nick 2021; Jones et al. 2014; Martinović et al. 2021). Because of this, there is a lack of consensus on

how to quantify microbial guilds. Below are some examples:

Wu et at. (Wu et al. 2021) use the term microbial guild to assign a functional value solely based on spatial co-occurrence among taxa. They elucidate guilds by correlating positive, neutral, and negative effects with CAGs (co-abundance groups). The problems with this approach are manifest, as spatiality does not necessarily correlate with function; especially in microorganisms.

A rather ingenious idea attempted to discriminate between different guilds of diatoms using a very similar standard to those of macroscopic guilds, based on the morphology and motility of a single-celled organism (Passy 2007). Passy's argument is that the nutritional trait of several diatoms seems to correlate with the motility profiles. In this case it seems a wise decision, since the ecological interpretation of the results leaves no room for doubt: "*The motile guild is comparatively free of both resource limitation and disturbance stress, because it has the physical capability of selecting the most suitable habitat*". However, it is not a standardizable solution, which is the main aim of the present study.

Nemergut et al. propose that the guild should be restricted to the co-occurrence in space and time of those taxa exploiting the same resource, and do not explore the question of how they exploit it (Nemergut et al. 2013). It is sensible to think that the guild concept should not be restricted spatially nor temporally, since what we want to quantify is how members are changing their contribution to the guild over time-space. We also consider it relevant to address how the function is being carried out in every case.

Pedrós-Alió defines more precisely what microbial guilds represent, as opposed to macro guilds: "*a group of microorganisms using the same energy and carbon sources and the same electron donors and acceptors*" (Pedrós-Alió 1989). However, microbes can share all energy and carbon sources and can still perform differently on the key substrate. For example, imagine two coexisting methanotrophs. They will share membership in *methane import* guild most of the time, but one of them removes methane only when it is abundant, and the other when it is scarce. The guild definition should consider the particularities of how the relevant function is carried out.

For this reasons, the guild concept, which was coined and mainly used for macroscopic studies,

has been extensively overhauled as a tool for the interests of modern molecular ecology. We propose the following definition: *a repertoire of bionts that, regardless of their current taxonomic position, benefit from a key resource through a set of biostructures converging to the very same function, which can be implemented differently to maintain an equivalent ecological role by adaptive radiation across all permitted environments.*

## Quantification of microbial guilds

As we have established, all biological functions are performed by biostructures, and these respond to an ideal range of action that is environment-dependent. Moreover, all biostructures are encoded in adaptive genetic polymers. Then, diversification of the functional secuences is not only expected but frequently observed (Pascual-García et al. 2010; Soria et al. 2014). Besides neutral drift, most likely the reason behind this phenomenon is that evolution rewards the pursuit of different optimal kinetics across various environments (Alam et al. 2009; Offre et al. 2014). So, if the entire biostructure repertoire for a function get clusterized by secuence dissimilarity, we will find groups that should work better under similar environments. Henceforth, we will call these sequence groups *implementations*, as they reveal how a function has been implemented in a particular set of biotic systems, which necessarily converge towards the unique action properties of the cluster.

These facts lead us to think that the guilds are structured differently, depending on the environmental circumstances (Figure 1). However, because there is no quantifiable guild definition, studying shifts in guild structure has been impossible until now. Consequently, we propose to determine the structure of a microbial guild as a vector, following:

$$Gi\ (k1, \ldots, kn)$$

where the structure of a guild $Gi$ is given by the ecological dominance **(k)** of its implementations per environment ***i***. Here we set one element of the vector for each known implementation of the function, and give each a *k-value* which can be defined as:

$$k = d \cdot u \cdot a$$

where, from left to right, are defined the terms: ***d*** as the total diversity measure inside a functional

repertoire or implementation; ***u*** as the functional univocity of each of the implementations that compose the guild, i.e., The closer the value is to 0, the more ambiguous the functional repertoire will be in occupying the same resource space; and ***a*** as the environmental abundance of the entire implementation. We construct the *k-value* in order to appraise the occurrence of diversity for an implementation, the likelihood of the implementation to perform the screened function, and the presence in the environment of this implementation.

We will obtain a vector of values for each screened environment and of as much length as implementations, so we can treat the quantified guilds as matrixes. We tested this approach all through with the *polyamine uptakers* guild along Malaspina (MP) watercolumns, to show the potential of the guild quantification concept to undestand complex ecological dynamics.

## RESULTS

### 1. Specific environmental features shape the functional biostructure

A reference tree has been constructed to classify the sequences found in the Malaspina metagenomes. For this purpose, the specific polyamine-uptake HMM has been searched against a curated oceanic database (details in methods).

The reference tree shows a collection of HMM-retrieved sequences segregated into dissimilarity groups. These sequences can belong to one or several taxa, which may be in degenerated positions around the tree. Sometimes these positions are distant, indicating that the same species may also have different sequence types. At this point, we consider it necessary to determine whether there is an environmental effect drifting the big groups of sequences (rather than taxonomic position) of these periplasmic proteins, which must maintain their function under physiological conditions.

Our results show that there are internal nodes significantly enriched for certain environmental variables (one-tailed p-values< 0.01). Consequently, we split the tree into two main clusters according to salinity, motility and acidity (Figure 2). The more external nodes would be also significant for other distinct environmental variables, indicating that taxonomic position summarizes the functional relationship only in relatively small nodes that do not contain sequences undergoing environmental convergence, nor re-adaptation, nor explained by horizontal transfer.

## 2. Determining functional sequences in metagenomic samples through placement filtering

The phylogenetic placement of the short environmental sequences shows, even after discarding false positives (71 placed sequences, 4.13% of total), a homogeneous distribution throughout the reference tree (Figure 3). Thus, the metagenomic sequences populate all the reference functional clusters that we have previously defined.

Moreover, most of the recovered sequences fit well in the reference tree (with a weighted likelihood ratio mean value of 0.89), indicating that the HMM would represent quite well the sequence diversity we found corresponding to polyamine uptake function within Malaspina metagenomes.

## 3. Functional clustering reveals ecological dynamics across environments

Overall, our results show that the polyamine uptake guild is present throughout the water column in the Malaspina samples, which is not only consistent with previous literature on the topic (Bergauer et al. 2018), but adds novel insights to how this function ~~is fluctuating~~ fluctuates with depth (Figure 4).

First, we found that the main forms of polyamine uptake were, regardless of depth, accounted for as saline implementations; this is coherent with the nature of the samples. In addition, the function appears to be redundant in the ocean layers we screened out, since we usually find all implementations in every sample. The latter effect seems to support the statement of functional redundancy being more prevalent than expected by chance in microbiomes (Puente-Sanchez et al. 2022).

Specifically, the guild pattern changes significantly between the epipelagic and mesopelagic, both in taxonomic composition and in the estimated strength of each of the implementations. Between the mesopelagic and bathypelagic the pattern is remarkably taxon-preserved, but the net contribution of each implementation to the overall function slightly changes, with major importance in the bathypelagic. Therefore, we conclude that there are unexpected non-linearities on how the polyamine uptake is carried out along the water column.

Another factor that denotes resilience in polyamine uptake is that, even in those environments where certain bacteria do not thrive (or that they cannot be detected while carrying a copy of this

function marker), other taxa replace them, evidencing a vertical ecological succession. We can easily quantify the taxonomic contribution to the total occurence of function in a concrete ocean layer.

In addition, the approach demonstrate its potential to track and estimate unexpected ecological traits. In our data, the occurrence of two different implementations in the same taxon is rare but possible. There is a clear example of this happening in the epipelagic between implementations IA and IIA.

We also noticed that there are non-strictly halophilic implementations contributing to the guild in the watercolumn, which would be adapted to alkaline environments. Of these alkaline implementations, the one with a more limited pH range would dominate in the epipelagic, while the one with more plasticity gains importance with depth.

## DISCUSSION

Proteins constitute the bulk of biostructures that perform a function in living beings. Like all other organisms, microbes achieve proteostasis through expression regulatory feedbacks, tuning of non-covalent interactions between biostructural subunits, and sequence re-adaptation (Ullmann et al. 1968; Gidalevitz et al. 2011; Manara et al. 2012). All of these source-of-variation mechanisms act in multiple levels and can have an immediate impact on substrate accomodation (Thompson et al. 1999). A single amino acid change may be crucial for the stereospecificity between the substrate and its binding site (Gierse et al. 1996; Price and Arkin 2022). However, drift of residues at sites other than the conserved regions of the protein can often be important in elucidating folding (Sadowski and Jones 2009). As already mentioned above, the biochemical performance turns out to be mostly hold by proteic oligomers evolved to remain bound under physiological conditions, and to monomerize in out-of-range environments (Traut 1994). Sometimes, due in part to the non-covalent nature of these protein-protein bonds, it is possible to recover function when physiological conditions return (Traut 1994). For all these reasons, it can be stated that any functional biostructure is the fine-tuning product of a secuence to a very particular environmental configuration range.

Some have discussed the environmental effect on the functional footprint of microbiomes

before, although almost solely in the sense of extrinsic biological interactions. (Rio et al. 2003; Spor et al. 2011). However, Panja et al. explore quantitatively how functional biostructures would have selection pressure to adapt to certain types of extreme environments, both in amino acid composition and in their ordering (Panja et al. 2020). Halophilicity, pH and temperature would represent the major environmental drivers in how implementations evolve, while maintaining an equivalent function (Panja et al. 2020).

Therefore, in order to classify the performances of ABC transporter-associated polyamine-binding proteins, we decided to test which of the environmental variables from the cultured organisms present in our reference tree would have a significant effect per node; that is, to test how good the tree topology is at discriminating groups of sequences putatively adapted to certain environmental variables.

We have opted to study a function that is difficult to explore and quantify, which is organic nitrogen acquisition through putrescine and other related polyamines. The idea was to test the usefulness of the guild quantification method. The difficulty of exploring this function is given by the following pitfalls: i) substrate binding is highly degenerate and, although there may be a slight preferential binding to spermidine or putrescine depending on certain amino acids (Kashiwagi et al. 1996), the tests we have done indicate that it would be very difficult to discriminate between regions of the tree with preferential binding to one type of polyamine Supplementary Figure 1 ii) there are several gene names for very similar biostructures, all corresponding to the subunit of the ABC transporter that has activity with the ligand iii) there is an extreme shortage of curated sequences with functional evidence.

Our results show a clear significance on big internal nodes to some environmental variables, discriminating at least two dissimilarity groups behaving different in the reference tree. These are alkaline pH and halophilic organism-enriched clusters, which is consistent with the study by Panja et al. The ability to select a suitable local environment for the kinetics of physiological functions seems to be relevant, as motility remains wildly significant also at these deep levels of the tree. As one moves up to the leaves of the tree, some nodes are enriched in significance by other variables.

At the nodes closest to the leaves, the significant environmental variables were more related to taxonomic position and vertical inheritance. This is the case for some environmental variables. The main ones are PAHs presence and temperature. The first one is entirely to be expected, since hydrocarbon degrading bacteria that can thrive in PAHs media require sharing a fairly large metabolic machinery in order to inhabit the same environments, which would be explained by vertical inheritance. However, the case of temperature it is indeed surprising, since thermostability is reported to be one of the main drivers of bias in neutral drift of functional structures (Somero 1995; Panja et al. 2020). Our hypothesis is that temperature would bias all protein sequences of microorganisms equally (including "phylomarkers"), since temperature, unlike salinity and pH, cannot be regulated inside of unicellular organisms.

This opens a debate that needs to be explored further. The current paradigm uses certain types of preserved sequences to measure taxonomic position, such as 16s ribosomal subunit (Rajendhran and Gunasekaran 2011), under the assumption of the molecular clock (Bromham and Penny 2003). Phylogenetic reconstructions use more or less complex algorithms that can be summarized in Hamming distances and parsimony between sequences, assuming one of them being ancestral (Bruyn et al. 2014). However, this assumption does not contemplate that sequences can permute in an environmentally biased way. The problem is that two apparently phylogenetically distant bacteria could have a closer-than-expected vertical relationship in number of generations if there was a temperature readaptation process in one of them. The reverse, unfortunately, would also occur: two microorganisms very distant in number of vertical generations could have converged in environments of similar temperature, so that our phylogenetic estimates would be wrong; underestimating their actual divergence. In contrast, the molecular clock would have proven useful for estimating speciation in higher animals, since physicochemical variables such as temperature are internally regulated.

We previously introduced that taxonomic position is not, in many cases, synonymous with function. In addition to these arguments, there is some research actually focused on decoupling taxonomy from functional assets (Louca et al. 2016). Moreover, machine learning approaches

appear to outperform niche prediction with functions rather than phylogeny (Alneberg et al. 2020).

The guild concept can partially address the latter disquisition, because can be used to discriminate these taxonomic effects from those caused by functional convergence in order to dissect how the function behaves through a battery of environments.

It should be considered that, beside guilds, there are other ways to study microbial functions, such as metabolic pathways reconstruction from omics data (Gianoulis et al. 2009). However, this approach does not consider the ecological forces shaping the microbial community that makes up each sequenced sample, but instead correlate abundance of different process-related enzymes with metadata (Yang et al. 2021). A guild approach would be superior since (i) it simplifies and classifies ecologically valuable information (ii) allows standardized comparison between environments, taxa and functions among an unlimited number of samples (iii) allows us to discriminate between taxonomy-dependent and taxonomy-independent effects, whilst in metabolic pathways they remain mixed.

## CONCLUSIONS

The sequences that can perform a specific function are multiple and degenerate, not necessarily having a close evolutionary history. In addition, there would be abiotic extrinsic forces shaping the sequences capable of performing a function. The potential for exploring functional ecology in microorganisms has been limited by the overwhelming amount of irrelevant and imprecise omics data.

Despite this, we have been able to open the "automatic functional annotation black-box" and to mechanistically describe the ecological dynamics within a complex function and ecosystem. Therefore, we propose a theoretical redefinition of the term guild, as well as methodologic procedures and bioinformatics tools to facilitate the praxis.

There are four main arguments that justify the present work: (i) the original definition becomes inextricably ambiguous in the microscopic realm, as there is no consensus on what is a similar way to exploit the same kind of resources for microbes; (ii) the emergence of omics data that drags along technical biases and an insurmountable information quantity; (iii) the desire to establish a

universality of the term, which favors a referable use of the same by the scientific community; (iv) alternative concepts are neither quantifiable nor ecologically relevant.

## MATERIALS AND METHODS.

### Oceanic reference DB construction

Semidan

### Guild marker selection

The search for guild markers was carried out by means of an extensive bibliographic comparison. This methodology is based on choosing public available Hidden Markov Models (HMMs) (Vasudevan et al. 2011) for one or several genes, trying to avoid functional paralogs. To integrate a HMM as a guild marker, we followed this conservative criteria: i) the construction of the HMM must be consistent, with an adequate number of seed sequences ii) it is well represented in our curated working database for reference tree reconstruction iii) there is sufficient functional evidence of the biostructure encoded by that gene iv) the metagemonic sequences retrieved with the tested HMM can be filtered out by a specific quality argument, derived from the inner workings of genomic architecture (i.e.: synteny). With this score system, we found that the best minimal marker for putrescine-like polyamines uptake is K11073, available from UniProt curated entry P31133 (https://www.uniprot.org/uniprotkb/P31133/external-links).

### Oceanic reference DB search

Semidan

### Phylogeny reconstruction

Semidan

### Functional clustering

Mr. Pabson

Before working with sequence clustering, we required functional features sequence-associated. Because we had no reliable evidence of how these polyamine transporter subunits perform, we

constructed a curated collection of physicochemical property vectors for each cultured organism in our functional tree (321 out of 1158 leaves, being XXX of them redundant). Supplementary Figure 2.

Supplementary table 1. Functional clustering was carried out by randomizing property vectors while maintaining the tree topology. Node enrichment.

Supplementary table 2.

**Metagenomic functional sequences search**

Squeezemeta (Tamames and Puente-Sánchez 2019)

**Placement and filtering of short query sequences**

Semidan

**Quantification of Polyamine-uptakers guild**

Once filtered sequences are classified by environment and functional cluster, is merged together with the corresponding normalized abundances into a single master table. This is the input for the first tool, microguilds.py. The module will extract all the required information for the calculation of each implementation contribution (k), with which we will establish the elements of a guild vector for an environment (Gi). In the present case, we study three distinct environments, so the software will produce an array of dimension n x 3 (where n is the number of implementations established within the guild marker).

The calculation contemplates three terms. The theoretical d is the total diversity of elements performing an equivalent ecological function. Calculation of the theoretical d is complex and would require avoiding false negatives. Therefore, in this example is calculated as the sum of unique sequences found within a functional cluster. The term representing the univocity of the implementations (u) is equalled to 1.0 since we discard the metagenomic sequences falling into non-functional clusters of the reference tree and we had a highly-conservative criteria to build the function's model. Finally, the abundance (a) has been calculated as a summation of normalized metagenomic counts for all sequence diversity falling in the same implementation. An example of

the k-matrix output is provided in the Supplementary Table 3.

The second tool (mgplots.py) helps to visualize this matrix, which can be of varying complexity. It does two things: (i) it filters by the taxonomic level to visualize the guild and (ii) it takes the value of k by taxonomic contribution to each functional cluster. To do this, it calculates the logarithm of each position in the matrix and plots it as shown in Fig. 4.

**16S and recA sequences**

To screen phylogenetic deviations between functions and phylomarkers, we obtained the nucleotide sequences of the 16S ribosomal subunit and the *recA* gene from the assembly genomic RNA and CDS provided by the NCBI for 319 out of the 321 species found in pure culture. When 16S sequences were < 1000bp, we used instead sequences from other strains as they should remain well conserved within the same species. All RefSeq assembly accession numbers and alternative GIs for 16S data were automatically retrieved from the NCBI, a detailed list is available in the Supplementary Table 4.

sup mat:

- Supplementary Figure 1 - potF/potD/spuD tree, mixed branches - refinement?
- Supplementary figure 2 - reference tree with distribution of evidence - maybe inside fig. 2
- Supplementary table 1 - environmental evidence table
- Supplementary table 2 - all significance nodes table
- Sup. table 3 - k-matrix of K11073 in MP
- Supplementary table 4 - 16s data

## REFERENCES

Alam, M. S., Garg, S. K., and Agrawal, P. (2009). "Studies on structural and functional divergence among seven whib proteins of mycobacterium tuberculosis h37rv." *The FEBS journal*, 276(1), 76–93.

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K., Riemann, L., Ekman, M., Jürgens, K., Labrenz, M., et al. (2020). "Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes." *Communications biology*, 3(1), 1–10.

Altenhoff, A. M. and Dessimoz, C. (2012). "Inferring orthology and paralogy." *Evolutionary genomics*, 259–279.

Baiser, B. and Lockwood, J. L. (2011). "The relationship between functional and taxonomic homogenization." *Global Ecology and Biogeography*, 20(1), 134–144.

Bergauer, K., Fernandez-Guerra, A., Garcia, J. A., Sprenger, R. R., Stepanauskas, R., Pachiadaki, M. G., Jensen, O. N., and Herndl, G. J. (2018). "Organic matter processing by microbial communities throughout the atlantic water column as revealed by metaproteomics." *Proceedings of the National Academy of Sciences*, 115(3), E400–E408.

Bromham, L. and Penny, D. (2003). "The modern molecular clock." *Nature Reviews Genetics*, 4(3), 216–224.

Bruyn, A. D., Martin, D. P., and Lefeuvre, P. (2014). "Phylogenetic reconstruction methods: an overview." *Molecular Plant Taxonomy*, 257–277.

Cavalier-Smith, T. (2006). "Cell evolution and earth history: stasis and revolution." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1470), 969–1006.

Chiel, H. J. and Beer, R. D. (1997). "The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment." *Trends in neurosciences*, 20(12), 553–557.

Dourado, H., Mori, M., Hwa, T., and Lercher, M. J. (2021). "On the optimality of the enzyme–substrate relationship in bacteria." *PLoS biology*, 19(10), e3001416.

Forbes, S., Morgan, N., Humphreys, G. J., Amézquita, A., Mistry, H., and McBain, A. J. (2019). "Loss of function in escherichia coli exposed to environmentally relevant concentrations of benzalkonium chloride." *Applied and environmental microbiology*, 85(4), e02417–18.

Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., Letunic, I., Yamada, T., Paccanaro, A., Jensen, L. J., Snyder, M., et al. (2009). "Quantifying environmental adaptation of metabolic pathways in metagenomics." *Proceedings of the National Academy of Sciences*, 106(5), 1374–1379.

Gidalevitz, T., Prahlad, V., and Morimoto, R. I. (2011). "The stress of protein misfolding: from

single cells to multicellular organisms." *Cold Spring Harbor perspectives in biology*, 3(6), a009704.

Gierse, J. K., McDonald, J. J., Hauser, S. D., Rangwala, S. H., Koboldt, C. M., and Seibert, K. (1996). "A single amino acid difference between cyclooxygenase-1 (cox-1) and- 2 (cox-2) reverses the selectivity of cox-2 specific inhibitors." *Journal of Biological Chemistry*, 271(26), 15810–15814.

Gregory, T. R. (2005). "Genome size evolution in animals." *The evolution of the genome*, Elsevier, 3–87.

Gregory, T. R. and DeSalle, R. (2005). "Comparative genomics in prokaryotes." *The evolution of the genome*, Elsevier, 585–675.

Hillis, T. L. and Mallory, F. F. (1996). "Sexual dimorphism in wolves (canis lupus) of the keewatin district, northwest territories, canada." *Canadian Journal of Zoology*, 74(4), 721–725.

Hohberg, K. (2003). "Soil nematode fauna of afforested mine sites: genera distribution, trophic structure and functional guilds." *Applied Soil Ecology*, 22(2), 113–126.

Husnik, F. and McCutcheon, J. P. (2018). "Functional horizontal gene transfer from bacteria to eukaryotes." *Nature Reviews Microbiology*, 16(2), 67–79.

Huynen, M., Snel, B., Lathe, W., and Bork, P. (2000). "Predicting protein function by genomic context: quantitative evaluation and qualitative inferences." *Genome research*, 10(8), 1204–1210.

Jones, C. M., Spor, A., Brennan, F. P., Breuil, M.-C., Bru, D., Lemanceau, P., Griffiths, B., Hallin, S., and Philippot, L. (2014). "Recently identified microbial guild mediates soil n2o sink capacity." *Nature Climate Change*, 4(9), 801–805.

Kashiwagi, K., Pistocchi, R., Shibuya, S., Sugiyama, S., Morikawa, K., and Igarashi, K. (1996). "Spermidine-preferential uptake system in escherichia coli. identification of amino acids involved in polyamine binding in PotD protein." *Journal of Biological Chemistry*, 271(21), 12205–12208.

Koran, M. and Kropil, R. (2014). "What are ecological guilds? dilemma of guild concepts." *Russian Journal of Ecology*, 45(5), 445.

Koskella, B., Hall, L. J., and Metcalf, C. J. E. (2017). "The microbiome beyond the horizon of

ecological and evolutionary theory." *Nature ecology & evolution*, 1(11), 1606–1615.

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

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., and Foster, P. L. (2016). "Genetic drift, selection and the evolution of the mutation rate." *Nature Reviews Genetics*, 17(11), 704–714.

Manara, A., DalCorso, G., Baliardini, C., Farinati, S., Cecconi, D., and Furini, A. (2012). "Pseudomonas putida response to cadmium: changes in membrane and cytosolic proteomes." *Journal of proteome research*, 11(8), 4169–4179.

Martinović, T., Odriozola, I., Mašínová, T., Doreen Bahnmann, B., Kohout, P., Sedlák, P., Merunková, K., Větrovskỳ, T., Tomšovskỳ, M., Ovaskainen, O., et al. (2021). "Temporal turnover of the soil microbiome composition is guild-specific." *Ecology Letters*, 24(12), 2726–2738.

Masel, J. (2011). "Genetic drift." *Current Biology*, 21(20), R837–R838.

Nebel, S., Mills, A., McCracken, J., and Taylor, P. (2010). "Declines of aerial insectivores in north america follow a geographic gradient." *Avian Conservation and Ecology*, 5(2).

Nemergut, D. R., Schmidt, S. K., Fukami, T., O'Neill, S. P., Bilinski, T. M., Stanish, L. F., Knelman, J. E., Darcy, J. L., Lynch, R. C., Wickey, P., et al. (2013). "Patterns and processes of microbial community assembly." *Microbiology and Molecular Biology Reviews*, 77(3), 342–356.

Nemeskéri, E. and Helyes, L. (2019). "Physiological responses of selected vegetable crop species to water stress." *Agronomy*, 9(8), 447.

Offre, P., Kerou, M., Spang, A., and Schleper, C. (2014). "Variability of the transporter gene complement in ammonia-oxidizing archaea." *Trends in microbiology*, 22(12), 665–675.

Paerl, H. W. and Pinckney, J. (1996). "A mini-review of microbial consortia: their roles in aquatic production and biogeochemical cycling." *Microbial Ecology*, 31(3), 225–247.

Pagé, A., Tivey, M. K., Stakes, D. S., and Reysenbach, A.-L. (2008). "Temporal and spatial archaeal colonization of hydrothermal vent deposits." *Environmental Microbiology*, 10(4), 874–884.

Panja, A. S., Maiti, S., and Bandyopadhyay, B. (2020). "Protein stability governed by its structural

plasticity is inferred by physicochemical factors and salt bridges." *Scientific reports*, 10(1), 1–9.

Pascual-García, A., Abia, D., Méndez, R., Nido, G. S., and Bastolla, U. (2010). "Quantifying the evolutionary divergence of protein structures: the role of function change and function conservation." *Proteins: Structure, Function, and Bioinformatics*, 78(1), 181–196.

Passy, S. I. (2007). "Diatom ecological guilds display distinct and predictable behavior along nutrient and disturbance gradients in running waters." *Aquatic botany*, 86(2), 171–178.

Pedrós-Alió, C. (1989). "Toward an autecology of bacterioplankton." *Plankton Ecology*, Springer, 297–336.

Price, M. N. and Arkin, A. P. (2022). "Interactive analysis of functional residues in protein families." *Msystems*, e00705–22.

Puente-Sanchez, F., Pascual-Garcia, A., Bastolla, U., Pedros-Alio, C., and Tamames, J. (2022). "Cross-biome microbial networks reveal functional redundancy and suggest genome reduction through functional complementarity." *bioRxiv*.

Rajendhran, J. and Gunasekaran, P. (2011). "Microbial phylogeny and diversity: small subunit ribosomal rna sequence analysis and beyond." *Microbiological research*, 166(2), 99–110.

Rio, R. V., Lefevre, C., Heddi, A., and Aksoy, S. (2003). "Comparative genomics of insect-symbiotic bacteria: influence of host environment on microbial genome composition." *Applied and Environmental Microbiology*, 69(11), 6825–6832.

Root, R. B. (1967). "The niche exploitation pattern of the blue-gray gnatcatcher." *Ecological monographs*, 37(4), 317–350.

Sadowski, M. and Jones, D. (2009). "The sequence–structure relationship and protein function prediction." *Current opinion in structural biology*, 19(3), 357–362.

Sanchez, A., Bajic, D., Diaz-Colunga, J., Skwara, A., Vila, J., and Kuehn, S. (2022). "The community-function landscape of microbial consortia.

Shapiro, J. A. (1998). "Thinking about bacterial populations as multicellular organisms." *Annual review of microbiology*, 52(1), 81–104.

Somero, G. N. (1995). "Proteins and temperature." *Annual review of physiology*, 57(1), 43–68.

Soria, P. S., McGary, K. L., and Rokas, A. (2014). "Functional divergence for every paralog." *Molecular biology and evolution*, 31(4), 984–992.

Spor, A., Koren, O., and Ley, R. (2011). "Unravelling the effects of the environment and host genotype on the gut microbiome." *Nature Reviews Microbiology*, 9(4), 279–290.

Storz, J. F. (2016). "Causes of molecular convergence and parallelism in protein evolution." *Nature Reviews Genetics*, 17(4), 239–250.

Tamames, J. and Puente-Sánchez, F. (2019). "Squeezemeta, a highly portable, fully automatic metagenomic analysis pipeline." *Frontiers in microbiology*, 9, 3349.

Thompson, J., Reese-Wagoner, A., and Banaszak, L. (1999). "Liver fatty acid binding protein: species variation and the accommodation of different ligands." *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1441(2-3), 117–130.

Tikhonov, M. (2017). "Theoretical microbial ecology without species." *Physical Review E*, 96(3), 032410.

Torsvik, V. and Øvreås, L. (2002). "Microbial diversity and function in soil: from genes to ecosystems." *Current opinion in microbiology*, 5(3), 240–245.

Traut, T. W. (1994). "Dissociation of enzyme oligomers: a mechanism for allosteric regulation." *Critical reviews in biochemistry and molecular biology*, 29(2), 125–163.

Ullmann, A., Jacob, F., and Monod, J. (1968). "On the subunit structure of wild-type versus complemented $\beta$-galactosidase of escherichia coli." *Journal of molecular biology*, 32(1), 1–13.

van de Guchte, M. (2017). "Horizontal gene transfer and ecosystem function dynamics." *Trends in microbiology*, 25(9), 699–700.

Vasudevan, S., Vinayaka, C., Natale, D. A., Huang, H., Kahsay, R. Y., and Wu, C. H. (2011). "Structure-guided rule-based annotation of protein functional sites in uniprot knowledgebase." *Bioinformatics for Comparative Proteomics*, Springer, 91–105.

Veshareh, M. J. and Nick, H. M. (2021). "A novel relationship for the maximum specific growth rate of a microbial guild." *FEMS Microbiology Letters*, 368(12), fnab064.

Wangemann, P. and Schacht, J. (1996). "Homeostatic mechanisms in the cochlea." *The cochlea*,

Springer, 130–185.

Wu, G., Zhao, N., Zhang, C., Lam, Y. Y., and Zhao, L. (2021). "Guild-based analysis for understanding gut microbiome in human health and diseases." *Genome medicine*, 13(1), 1–12.

Yang, Z., Zhou, Q., Sun, H., Jia, L., Zhao, L., and Wu, W. (2021). "Metagenomic analyses of microbial structure and metabolic pathway in solid-phase denitrification systems for advanced nitrogen removal of wastewater treatment plant effluent: A pilot-scale study." *Water Research*, 196, 117067.

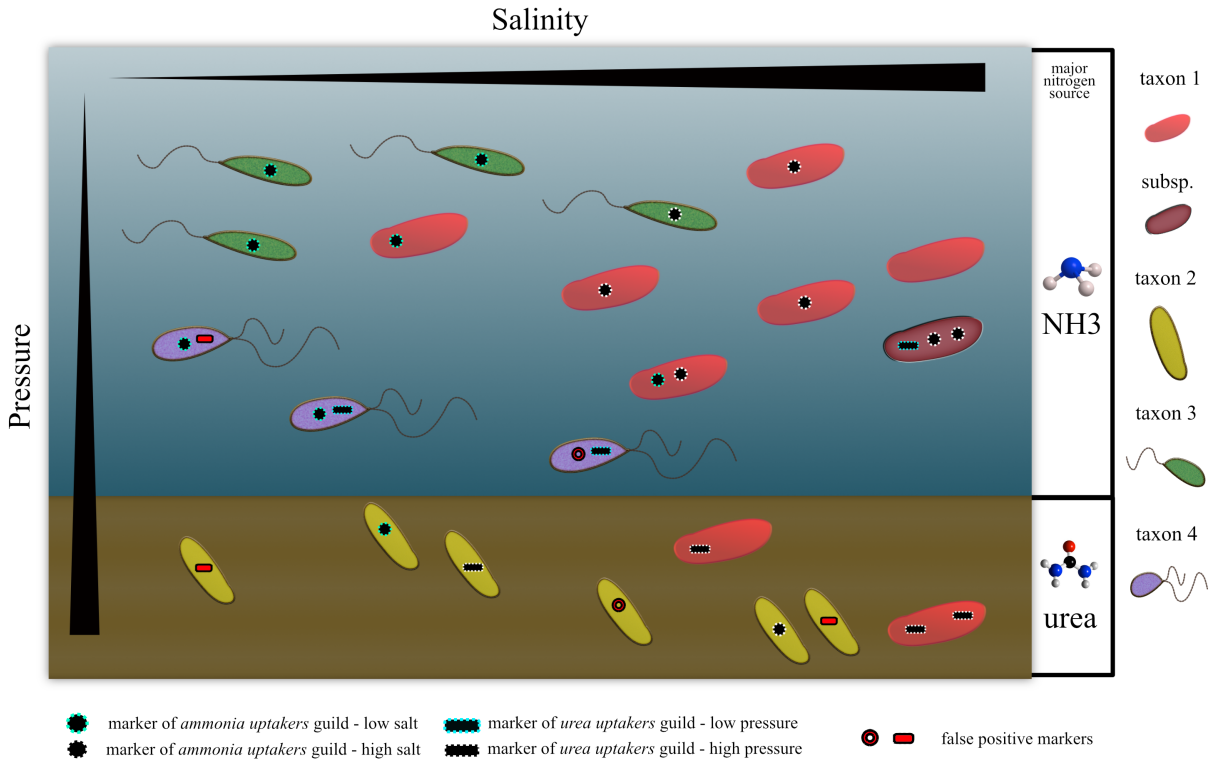


**Figure 1. Example of two cooccurring oceanic microbial guilds.** Environment selects for the ecological interactions of various taxa, whose functions are implemented in different ways. The functional set that each species maintains over time depend on the relative fitness in the actual environment. Therefore, it is expected to observe genotypic variations in populations of the same species across environments. Using massive sequencing techniques we cannot detect false negatives, because we do not know if there are relevant unknown proteins that have an equivalent ecological function. However, we can rule out false positives (functional paralogs) by comparison with certain types of sequences that exhibit evidence of catalysis. Then, we can cluster the diversity of sequences that do perform the function. For clustering, we will take into account the environmental variables that constitute a major agent of sequence drift for each guild marker. In the above example, taxonomic position alone rarely predict the guild membership, nor the final ecologic contribution to urea or ammonia uptake. On the other hand, environmental conditions (salinity, pressure and major nitrogen source) are better at predicting the transporter kind. Screening circumstances where the guild pattern drastically changes would allow us to draw major ecologic knowledge.

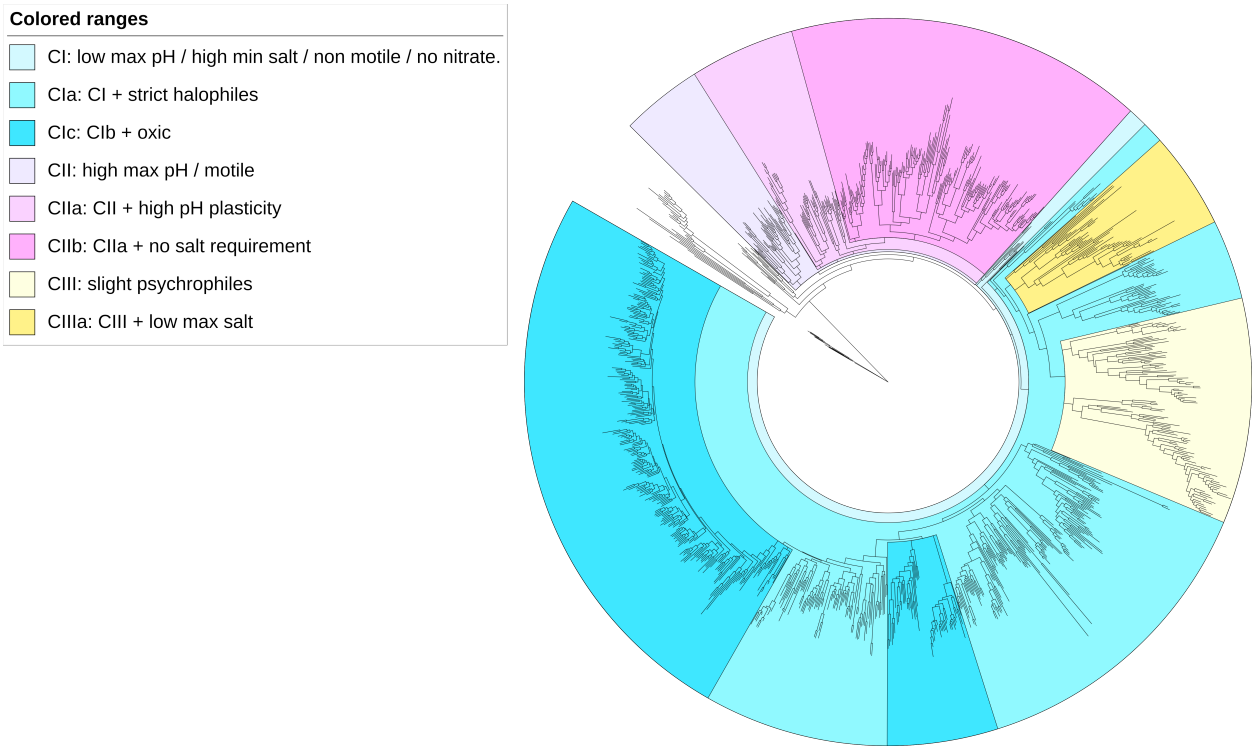


**Figure 2. Functional Clustering of K11073 Reference Tree.** The tree's colored regions represent the most distant parental nodes that remain significant for a feature that is not explained by taxonomic position. Therefore, we can establish that there are two markedly different groups of dissimilarity: the sequences that would be adapted to salinity conditions and those that would be adapted to a wide range of pH. Within these groups we find significant nodes for other environmental variables, which allow us to differentiate subgroups of dissimilarity for specific conditions. Studying the prevalence of each of these functional clusters in different environments will provide us with a more detailed picture of the ecological dynamics concerning this guild, which would be the amount of putrescine-like polyamines.

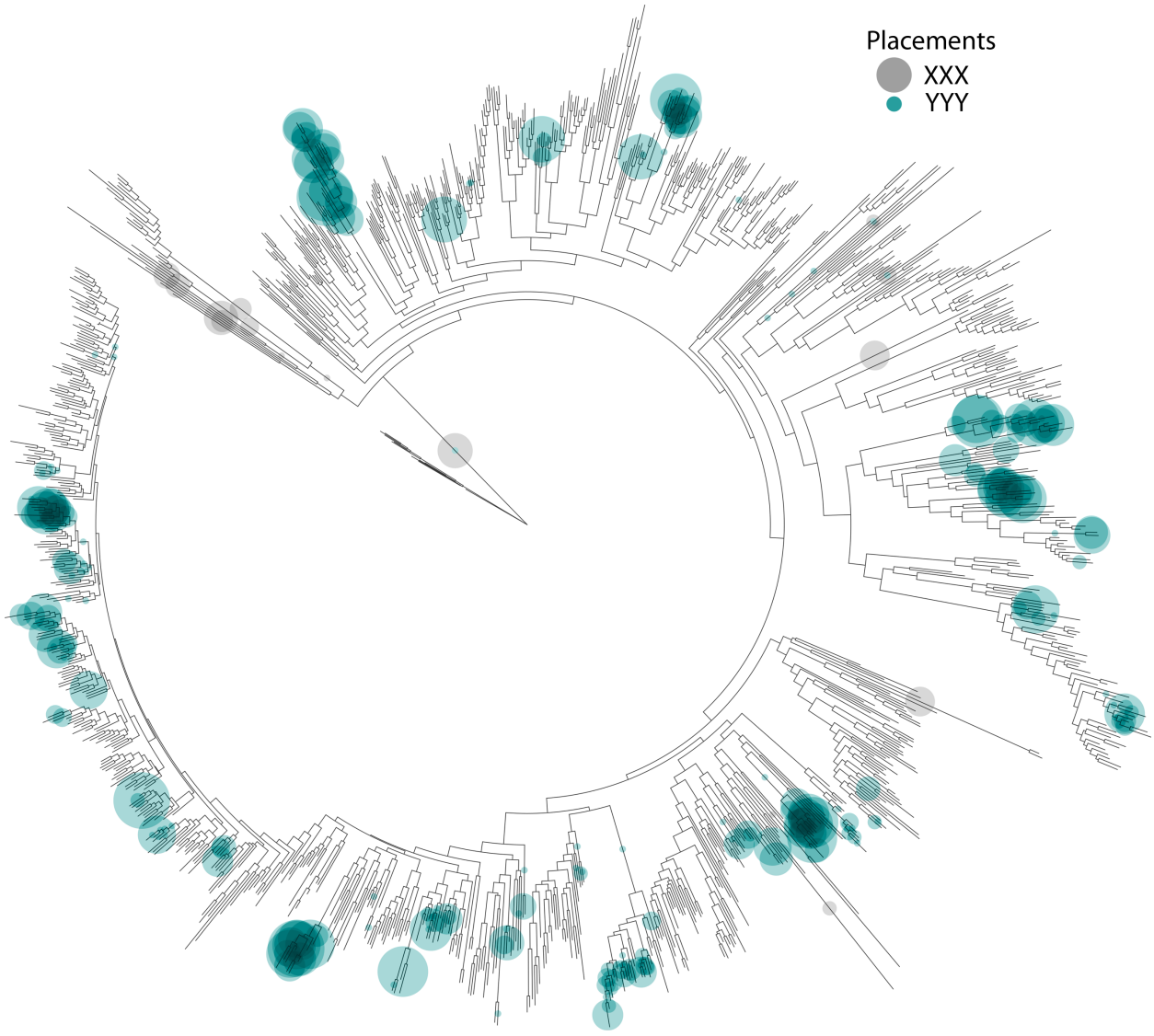


**Figure 3. Short environmental sequences from MP watercolumns placed in K11073 Reference Tree.** Part of the placed sequences will not pass the first filter, which consists of eliminating sequences that do not align well with the reference tree (gray). Others have fallen into regions where there is functional evidence for not binding polyamines, so they are not considered for guild quantification. Note that the emplaced leaves represent 2/3 of the tree's total, and come from more than 70 different samples, divided into epipelagic, mesopelagic and bathypelagic regions of the ocean. The complexity of the functional data in metagenomes requires a suitable quantitative method for their subsequent analysis.

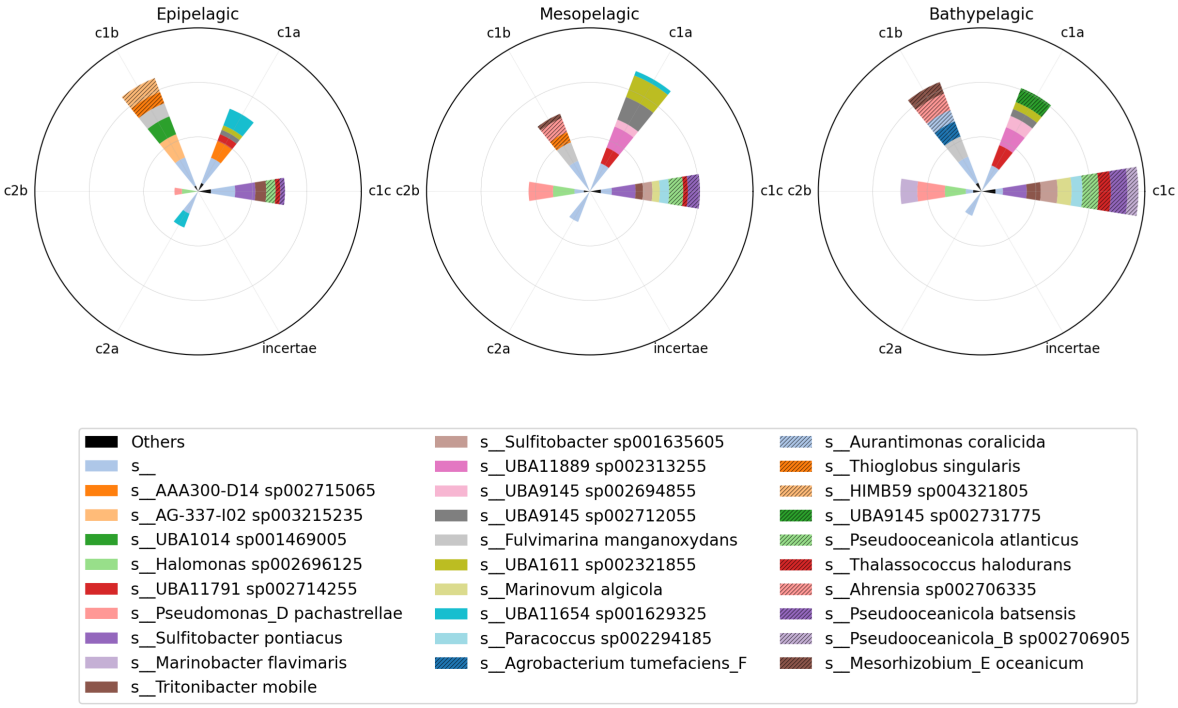


**Figure 4. Patterns of "Polyamine uptakers" guild.** K-matrix is calculated only with We need to introduce the abundance. Implementation c1b == cIII