Quantifying microbial guilds

Juan Rivas-Santisteban*[1], Pablo Yubero[2], Semidán Robaina-Estévez[3], José M. González[3], Javier Tamames[1], and Carlos Pedrós-Alió[1]

[1]Microbiome Analysis Laboratory, CNB-CSIC, Spain. E-mail*: jrisant@outlook.es

[2]Logic of Genomic Systems Laboratory, CNB-CSIC, Spain.

[3]Department of Microbiology, University of La Laguna, Spain.

ABSTRACT

The ecological role of microorganisms is of utmost importance due to their multiple interactions with the environment. However, assessing the contribution of individual taxonomic groups has proven difficult despite the availability of high throughput data, hindering our understanding of such complex systems. Here, we propose a quantitative definition of guild that is readily applicable to metagenomic data. Our framework focuses on the functional character of protein sequences, as well as their diversifying nature. First, we discriminate functional sequences from the whole sequence space corresponding to a gene annotation to then quantify their contribution to the guild composition across environments. In addition, we distinguish between sequence spaces that have different ways of carrying out the function. We demonstrate the validity of our approach by quantifying the guild of ammonia oxidation, and further reveal novel ecological dynamics of putrescine uptake guild in marine ecosystems. Thus, guilds help elucidate the functional role of different taxonomic groups with profound implications in the study of microbial communities.

INTRODUCTION

Organisms profoundly modify their environment. A clear example of this is the dramatic change in atmospheric oxidation potential that occurred in the primitive Earth, probably during the neo-archaic period – around 2.8 Gyr ago (Cavalier-Smith, 2006; Stüeken et al., 2012). The multiplication of biological functions related to oxygenic photosynthesis led to such planetary shift. Therefore, understanding ecology and evolution of functions is essential to predict such important changes. Specifically, functions carried out by microbes are, because of the simplicity, adaptability and ubiquity of these organisms, fundamental in the main nutrient cycles (Arrigo, 2005). Determination of microbial functions in ecosystems relies heavily on techniques based on the annotation of genomic information obtained from environmental DNA. This approach, however, presents a challenge to reach an insightful comprehension of such functions (Tikhonov, 2017; Koskella et al., 2017).

Biological functions can be understood as the causal relationship between the structural information contained in a biocatalyst (an effector) and the interaction it facilitates on a specific substrate. There are different types of biocatalysts, such as ribozymes or ribonucleoprotein complexes (Cech, 2009), but most consist of gene-encoded proteins with prosthetic groups. This causal relationship is affected by the environment (Zaks and Klibanov, 1986; Johansson et al., 2011) and is shaped by evolution, mainly through gene duplication, adaptation and drift (Masel, 2011; Altenhoff and Dessimoz, 2012; Lynch et al., 2016). If the environment changes, that causal relationship may be compromised or extinguished (Lanyi, 1974; Ladero et al., 2006). As a consequence of these processes, a variety of distinct proteins that fulfil the same function (Fig. 1) is generated (Liberles et al., 2012; Dourado et al., 2021).

Most functions are inherited vertically and therefore, taxonomically related organisms will often share a similar set of functions (Baiser and Lockwood, 2011). However, there are alternatives to vertical inheritance. A major example is horizontal transfer (van de Guchte, 2017), which has been observed even among organisms with markedly different taxonomic positions (Husnik and McCutcheon, 2018). Also, dissimilar sequences carrying out the same function may emerge by convergent evolution (Pagé et al., 2008; Storz, 2016). Therefore, the assigned taxonomic position alone may not be able to predict the occurrence of some functions.

Thus, there is a need for a non-taxonomic approach to explain the ecology of microbial functions. The ecological guild concept solves this problem. One classic definition of a guild is as follows: *"a group of species that exploit the same class of environmental resources in a similar way (. . .) without regard to taxonomic position, that overlaps significantly in their niche requirements"* (Root, 1967). Thus, guilds are broadly understood as the functional groups into which communities can be subdivided, unlike the concept of population, which consists of taxonomical groups.

The guild concept was designed for the ecology of macroorganisms and became popular in the 1970s. This viewpoint triggered research interest into niche partitioning. For example, all insect predators can be studied together as they are members of the *insectivorous* guild, without considering the taxonomic group they currently belong to (Koran and Kropil, 2014; Nebel et al., 2010).

However, the classical definition does not fit comfortably with the needs of microbial ecology. In macrofauna, guilds are defined by feeding behavior (Hohberg, 2003). Different behaviors have to do with very complex genetic interactions leading to ethology and information transmitted through nurture (Chiel and Beer, 1997; Hillis and Mallory, 1996). Thus, the nutrient acquisition carried out by an insectivorous guild: searching, capturing, ingesting and digesting an insect, is dependent on a vast series of genes with the corresponding molecular processes. In contrast, microbial feeding phenotypes are closer to their genotypes (Torsvik and Øvreås, 2002). In prokaryotes, the acquisition of a nutrient is almost exclusively dependent on a few proteins (Gregory, 2005; Gregory and DeSalle, 2005).

Despite the above considerations, several scientists have tried to use the classic guild concept to explain the functional complexity of microbiomes (Veshareh and Nick, 2021; Jones et al., 2014; Martinović et al., 2021). However, there is a lack of consensus on how to define and quantify microbial guilds. Below are some examples:

Wu and colleagues use the term microbial guild to assign a functional value solely based on spatial co-occurrence among taxa (Wu et al., 2021). The problems with this approach are clear, as co-occurrence in space does not necessarily imply sharing the same function, especially in microorganisms.

A rather ingenious idea attempted to discriminate between different guilds of diatoms based on the morphology and motility of these single-celled organisms (Passy, 2007). Passy's argument was that the nutritional traits of several diatoms seem to correlate with the presence of motility. However, it is not a suitable solution for all microbes, because it is a very specific case of limited application outside this group of organisms. This approach precludes the generalization of guilds, which is the main goal of the present study.

Other authors proposed that guilds should be restricted to taxa exploiting the same resource in a given space and time (Fauth et al., 1996; Nemergut et al., 2013). However, it is sensible to think that the guild concept should not limit itself either spatially nor temporally, since ecological dynamics are derived from spatiotemporal comparisons. In addition, we consider it relevant to understand how the function is performed in different scenarios in order to quantify its contribution to the guild.

Pedrós-Alió defined more precisely what microbial guilds represent, as opposed to guilds of macroorganisms: "*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, consider two coexisting methanotrophs: they will share membership in the *methane consumption* guild most of the time, but one of them may remove methane only when it is abundant, and the other when it is scarce. The guild definition must consider the particularities of how the relevant function is carried out.

All things considered, a guild is a diverse group of organisms benefiting from a key resource through evolving functional effectors, regardless of their taxonomic assignment, and where the success of the function in different circumstances is dependent on the diversification of its effectors. This definition not only fits the specific needs of molecular ecology, but is also applicable to all organisms.

Consequently, we postulate that each member of the guild can fulfil the definitory function, but different mechanisms to perform it inevitably emerge as a result of evolution. The classical definition was imprecise: not only the phylogenetic lineage, but the way in which the function is carried out are both irrelevant for an organism to be considered a member of a guild. However, to understand the ecology of the guild, it is still important to know its members and how they perform the function across different environments. In this work, we present a method for the quantification of guilds considering this redefinition, which reconciles the traditional view of guilds with its use to study microbial functions.

RESULTS

1. Quantification method for microbial guilds

To quantitatively introduce our definition of guild into the study of microbes, we considered the microbial guild as a 3D matrix that relates taxonomy, ways of performing the function (implementations) and environments. (Fig. 2A).

One dimension is the taxonomy assigned to the functional effector, in our case, the protein(s). This has been the only variable considered in the previous studies of guilds.

A second one is the different ways in which the function can be performed. That is, how the function is implemented in the different taxa. For example, high affinity vs. low affinity transport of dissolved ammonia, or thermophilic vs. psychrophilic oxidation of acetate. In the context of microbial guilds, an implementation can be associated with a specific gene or set of genes that encode the functional traits necessary for carrying out the desired function. Often, an implementation corresponds precisely to taxonomy. However, in many instances a single taxon may have more than one implementation and a given implementation may be shared by several taxa. Thus, the need to have this second dimension. For example, in the polyamine uptake guild analyzed later, the taxon UBA11654 sp001629325 (which represents an unclassified Gammaproteobacteria isolated from the Red Sea) contributes to the guild through two different implementations of the molecular function defined by the gene *potF*.

The third dimension, finally, is the environment in which the guild is conducting its activity. As will be seen later, the taxon just mentioned, for example, appears with two implementations in the epipelagic, but with only one in the mesopelagic, and it does not contribute to the function in the bathypelagic.

In practice we quantify the contribution of different taxa and implementations across environments to the guild of function f as the three-dimensional array $K_f \in \mathcal{M}_{T,I,E}$ with elements $k_{t,i,e}$, where the subindices t, i, and e represent taxa, implementations and environments, respectively. Its elements are calculated as:

$$k_{t,i,e} = \left(\sum_s a_s \right) \frac{d_{obs}}{d_{exp}} u, \tag{1}$$

The first term is the sum of sequence abundances a_s among the corresponding sequences (for which the subindices t, i and e are implied). Second, and because we postulate that the richness of unique sequences is a key factor for the resilience of the function over time, we include the ratio between observed and expected sequence richness, d_{obs} and d_{exp}, respectively (Fig. 2B). The term d_{obs} is calculated by the sum of unique sequences. The denominator d_{exp} is necessary to correct for the fact that greater richness is invariably observed when there are many sequences contributing to the function. Thus, d_{exp} empirically follows a power function of the summed abundances:

$$d_{exp} = c \left(\sum_s a_s \right)^{\gamma}, \tag{2}$$

where c and γ are gene and context specific constants. In this way we reward those instances with higher sequence richness than inferred from their abundance and we penalize those with lower than expected richness (Fig. 2C). The predictive ability of this empirical relationship allows the estimation of d_{exp} in several genes tested (Fig. 3).

Finally, the third term of (1) evaluates the univocity for the function $u \in [0, 1]$. Thus, $u = 1$ in the case that all the sequences fulfil the function perfectly, or $u = 0$ if the sequences are not capable of performing it. Although in our case we only consider a binary classification of function, this definition could take into account intermediate values, for example by assessing the average efficiency to accomplish the function.

In summary, higher k values reflect the availability in the environment of a given implementation, the occurrence of unexpected sequence diversity, and the likelihood of the sequences to perform the guild-definitory function. Therefore, and considering the dimensions of the guild hypervolume, we should have a lot of positions where $k = 0$ (there are no sequences contributing to the guild), and clouds where the values take $k > 0$. In this work, we relied on normalized abundance metagenomic outputs to quantify guilds.

2. Strict discrimination of functional paralogs improves guild assessment.

Prior to microbial guild quantification, a conservative criterion is needed in order to retain strictly functional sequences only (those with $u = 1$), since automatic classification relies on similarity with known sequences, and this produces many false positives, because similarity alone is often not accurate enough to discriminate functionality (Valencia, 2005). In this section we present a method that greatly improves automatic function annotation by using reference trees as functional sequence classifiers.

We will illustrate the procedure using ammonia oxidation as an example. The effector of this function is typically ammonia monooxygenase (AMO), which catalyzes the reaction of ammonia and oxygen to produce hydroxylamine. In particular, the gene encoding the A subunit (*amoA*), which conducts the catalytic activity of the enzyme complex (Ensign et al., 1993; Rotthauwe et al., 1997) has undergone extensive functional description. Therefore, the sequence spaces of *amoA* are well characterized in the literature (Martens-Habbena et al., 2009; Alves et al., 2018; Khadka et al., 2018; Wright et al., 2020). Thus, we discriminate between implementations carrying out mostly ammonia oxidation (mainly archaean AMO, AOA; and bacterial AMO, AOB) and those others that have higher affinities for methane or other simple aliphatic alkanes (Rochman et al., 2020), contributing to two or more guilds.

The particularity of this function is that its genes have evolved among taxonomical groups with different metabolic pathways. Thus, the enzyme has shifted from an ancestor with moderate affinity for a broad spectrum of substrates to a restricted substrate specificity, suboptimal for the substrate preference of each organism (Lau et al., 2016). Nonetheless, some groups of sequences remain with some promiscuity or ambivalence; for example, the pMMO effector is able to oxidize ammonium, while its main substrate is methane (Ward, 1987; Oudova-Rivera et al., 2023), or the particulate butane monooxygenase (pBMO) that oxydizes butane (Sayavedra-Soto et al., 2011). This means that specific *amoA*-like sequences behave as functional paralogs, so we must recognize them as false positives ($u = 0$) in order to evaluate the ammonia oxidation function present in the metagenomes.

Considering the above, a reference *amoA* tree was built to discriminate orthologs from paralogs in our metagenomic dataset using an in-house curated oceanic database (Methods; Sup. Fig. A). Then, we retrieved *amoA* sequences from Malaspina megatenomes and placed them onto the tree. We found that from a total of 129 unique automatically annotated sequences, 40 were discarded by the tree placement due to paralogy (31.0%), and only 82 were *bona fide* for ammonia oxidation (63.5%, Fig. 4). In this way, we don't rely on simple automatic annotations, as it is usually done. Instead, we used curated and annotated trees to infer the belonging of a particular sequence to one of the functional clusters of the tree defined by experimental evidence (full *amoA* clustering provided in Sup. Fig. B). This allows to accurately distinguish the functionality of particular genes.

3. High quality reference tree building without direct functional evidence.

Unfortunately, most genes have limited or ambiguous evidence of functional paralogies. This is the case of *potF*, which encodes a subunit of an ATP-binding cassette (ABC transporter) that binds and imports putrescine-like polyamines to be mainly used as source of N. These proteins must fulfil a specific function in the periplasm across a variety of external environments, leading to a vast diversity in transporter sequences sometimes not correlated with taxonomy (Offre et al., 2014), making automatic annotations challenging.

As before, to improve the automatic annotation we needed a reference tree (Fig. 5). Thus, we relied on the Hidden Markov Model corresponding to polyamine binding (KEGG K11073, Pistocchi et al. (1993)) to identify functional sequences across the same oceanic database used previously and to build the reference tree with them. We used the resulting tree to further discard long-branching metagenomic sequences (i.e. distances larger than the tree's original diameter), instead of discriminating functional paralogs (because we did not have this kind of information, as in the previous example with *amoA*).

We evaluated our metagenomes from Malaspina samples. Among the phylogenetic placements of the short environmental queries we discarded 71 queries (4.13%) as being false positives according to the reference tree. The rest of the queries populated all the tree (Fig. 6), suggesting that most known marine polyamine-like binding proteins are represented in our dataset. Moreover, most of the recovered sequences fit robustly in the reference tree (mean weighted likelihood ratio of 0.89), showing that the placement was robust.

4. Specific environmental features shape the protein sequence space of *potF*.

The reconstructed reference tree showed a collection of HMM-retrieved *potF*-like sequences grouped by their similarity where each sequence is represented by a leaf of the tree. On the one hand, the same organism may have more than one sequence, which may be either in distant or nearby positions in the tree, e.g., *Pseudomonas alcaligenes* appeared in three different clusters (Sup. Fig. C). On the other hand, sequences from distant taxa may unexpectedly converge in similarity. This is the case of *Oceanobacter kriegii* and *Thalassobius gelatinovorus*, an alpha- and gamma-proteobacterium, respectively, whose normalized phylogenetic distance on a 16S tree is large, while being remarkably close in the *potF*-like reference tree (0.67 vs 0.16; further details available in Sup. Fig. D). In summary, the phylogenetic signal poorly predicts divergence in *potF*-like protein sequences.

The ability of a protein to perform a function is influenced by the surrounding environment, thus requiring specific conditions to perform it effectively. In particular, transporters are exposed to changing environments. Therefore, we wished to address if the unexpected sequence divergence of the polyamine binding proteins depend on environmental conditions.

To that aim, we searched in the literature for the environmental preferences of the bacteria represented in the reference tree (Sup. Table 1), and searched for nodes grouping sequences that significantly shared similar environmental properties (one-tailed tests p-values < 0.003; Methods).

First, we observed significant nodes that grouped only a handful of sequences, highlighting properties that the taxonomy accurately predicts, e.g., hydrocarbon presence in growth conditions (Moreno-Ulloa et al., 2020). Representation of significant nodes is provided in Supplementary Figure E. Second, we found significant nodes that grouped all sequences in just a few clusters under common preferences to temperature, salinity and acidity, suggesting groups undergoing common environmental adaptation, or horizontal gene transfer (Fig. 5 and Sup. Table 2). These significant nodes were also represented in Supplementary Figure D, showing a clear correspondence between unexpected divergence of inner nodes and some environmental variables affecting protein folding. Therefore, we use the latter grouping to define different implementations and classify sequences accordingly, highlighting broader trends and properties across the reference tree of *potF*.

5. Functional clustering reveals ecological dynamics in the polyamine uptake guild.

Furnished with a functional *potF*-like sequence classifier, we finally proceeded to quantify the polyamine uptake guild in the Malaspina circumnavigation samples, showcasing the potential of our approach to reveal fundamental ecological dynamics among the oceanic layers. Since this specific guild is believed to be ubiquitous in the ocean, it was ideal to test whether there were differences within the guild between depths.

We decided to compare samples from three different marine environments: epipelagic (0 – 200 m), mesopelagic (200 – 1000 m) and bathypelagic (1000 – 4000 m). Using the classified sequences (Fig. 6) we calculated the k values for each taxon, cluster and environment. We represented the k values with radial plots (Fig. 7) to visualize the structure of the guild. In these graphs, each radial plot shows one environment, each direction represents an implementation of the function (or cluster in the tree), and the length of the spokes represents the impact coefficient k that, as explained, reflects abundance and sequence diversification. This representation of the data provides a visual and quantitative summary of the guild structure.

Overall, our results show that the polyamine uptake guild was important throughout the entire water column. First, the main forms of polyamine uptake were all saline implementations (*cIa*, *cIb*, and *cIII*); which is coherent with the fact that the samples were all marine. Implementation *incertae* included the placed sequences that were filtered out as false positives (Fig. 6) and is thus empty. In addition, the function exhibited considerable redundancy, since there were different implementations in every sample and several taxa with each implementation.

Specifically, the guild structure changed 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 changes slightly, with more top contributors above the fixed threshold in the bathypelagic. However, the polyamine uptake function persists throughout the water column despite the changes in taxonomic composition, evidencing a species turnover with depth. Therefore, our framework reveals non-trivial changes in guild structure in the ocean despite the ubiquity of the function.

In addition, the approach demonstrates its potential to track and estimate ecological dynamics. We can seamlessly measure changes between environments in the guild contribution by computing $\Delta k_e = k_e / k_{e-1}$, where k_e represents the contribution of all taxa and all implementations in environment e. In this example, Δk_{meso} polyamine uptake is equal to 1.66, while Δk_{bathy} is 0.84. Thus, even though several implementations are more important and taxonomically diverse in the bathypelagic, the main changes occur between the epipelagic and the mesopelagic.

In addition, the increment of k between environments can be calculated for the i-th implementation as $\Delta k_{i,e} = k_{i,e} / k_{i,e-1}$. Some examples of functional analysis with this ratio are shown in Figure 8. In our data, the most remarkable change correspond to $\Delta k_{cIIb,meso} = 7.92$ (dark pink in Fig. 8), suggesting an implementation-dependent bloom in the mesopelagic. This is an interesting finding, since the implementation $cIIb$ is the one that has the closest relationship with large pH variability; a variable that, coincidentally, reaches its minimum value in the mesopelagic. This confirms the ability of our guild approach to reveal environmental-dependencies of biological functions in general.

However, this is not the case for the most important contributors, such as implementation cIa, where the most important layer is barely the epipelagic, according to these metrics: $\Delta k_{cIa,meso} = 0.94$ and $\Delta k_{cIa,meso} = 0.72$. These analyses can be taken further to study the contribution of specific taxa to the k-value in a particular environment and implementation.

DISCUSSION

Diversification of functional protein sequences

As we have introduced, most functions are performed by evolving proteins. In addition, each function is often found in many different environments. Thus, diversification of the function-capable sequences is not only expected but frequently observed (Fay and Wu, 2003; Pascual-García et al., 2010; Soria et al., 2014). In order to quantify microbial guilds, the issue of how proteins diversify while maintaining function must be considered.

First, diversification of a protein can lead to promiscuity or pleiotropy (Hult and Berglund, 2007; Ruelens et al., 2023), especially when horizontal transfer events occur (Glasner et al., 2020). It is then likely that the protein may partial or totally lose its original function, undergoing a process of readaptation to its new genomic and environmental context (Deng et al., 2010; Manara et al., 2012; Husnik and McCutcheon, 2018). When there is sufficient functional evidence that these sequences do not play the definitory function, they can be filtered out (Methods). In the present work, all inferred guild marker sequences were carefully discriminated from those spurious sequences that were not functional. Where it could not be determined from the available evidence whether or not they fulfilled the function, they were categorized separately from true positives, as shown in Figure 1 and Figure 7 (see *incertae* implementation).

Once the truly functional sequence spaces have been identified (i.e. the implementations of the function), the next question is to determine what is distinctive about these divergent groups. Neutral drift undoubtedly contributes to this diversification (Kimura, 1991). However, certain degree of functional flexibility in a population of sequences may be the product of selection, allowing proteins to have slightly different kinetics in diverse environments, as well as acting upon more than one substrate (Alam et al., 2009; Offre et al., 2014; Zhao, 2022). We can expect that sequence variants adapted to similar conditions will be closer to each other. Thus, if the entire sequence space performing exactly the same function is grouped into clusters of sequence similarity, groups of sequences that are expected to work alike in similar environments shall emerge (Figs. 5 and 6), instead of at-random groupings. For quantification of microbial guilds, we defined the function implementations as these groups of sequences that work in a similar way (i.e.: binding affinity, substrate spectrum, temperature, pH or salinity conditions, etc.).

How the environment constrains microbial protein diversification

Like all other organisms, microbes achieve proteostasis through expression regulatory feedbacks, tuning of non-covalent interactions between structural subunits, and sequence re-adaptation (Ullmann et al., 1968; Gidalevitz et al., 2011; Manara et al., 2012). All of these mechanisms act in multiple levels and can have an immediate impact on substrate accommodation (Thompson et al., 1999). A single amino acid change may be crucial for the specificity between the substrate and its binding site (Gierse et al., 1996; Price and Arkin, 2022). Moreover, modification of residues at sites other than the conserved regions of the protein can often be structurally important (Sadowski and Jones, 2009). Regarding the quaternary structure, the protein subunits evolve 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 protein is the fine-tuned product of a sequence to a very particular range of environmental conditions.

Most of the previous research has demonstrated that microbial proteins may have several adaptations to the environment (Bartlett, 1999; Rio et al., 2003; Spor et al., 2011). However, the process itself is poorly understood in a mechanistic way, despite continued efforts (Kreitman, 1996; Reed et al., 2013; Tamuri and Dos Reis, 2022). More recently, Panja et al. explored statistically how microbial proteins undergo selective changes to adapt to environments of different kinds, both in terms of amino acid composition and in their ordering (Panja et al., 2020), a result in line with the *weak selection* concept (Akashi et al., 2012). According to these and other previous results, salinity, pH and temperature would represent the major environmental drivers of how implementations evolve (Lanyi, 1974; Fisher et al., 1997; Kumar et al., 2009; Tamames et al., 2010), modifying the protein catalytic kinetics, substrate specificity or conditional stability while maintaining the same function (Huston et al., 2008; Zhao, 2022).

These facts lead us to think that the guilds are structured differently, depending on the environmental circumstances. However, in the absence of a quantitative definition of guild, the study of changes in guild structure under different environmental conditions has been difficult.

Determining microbial guild structure considering the nature of protein diversification

In order to test the usefulness of the guild quantification method, we chose a function that is difficult to explore and quantify, which is organic nitrogen acquisition through putrescine and other related polyamines. The difficulty of exploring this function is given by the following pitfalls: (i) substrate affinity is moderately unspecific and, although there may be a slight preferential binding to spermidine or putrescine depending on certain amino acids (Kashiwagi et al., 1996), our results indicate that it would be difficult to discriminate between tree regions with particular specificities (Sup. Fig. F); (ii) there are several gene names for very similar protein sequences; (iii) there is an extreme shortage of curated sequences with functional experimental evidence.

As stated above, the microbial guild quantification method aims to (1) discriminate sequence spaces that correspond to the same function, and then to (2) characterize groups of functional sequences that work in a similar way (implementations of the function). The first objective improved automatic gene annotation, while the second categorized it functionally. To do both, we built and used several reference phylogenetic trees for oceanic organisms as sequence space classifiers (Figs. 4 and 5). For more details on how the first objective was carried out, see Methods and the Supplementary Material.

Regarding the second, we wanted to classify the performances of ABC transporter-associated polyamine-binding proteins. Since we did not have sufficient information on the preferential binding to each polyamine or its kinetics, we decided to characterize groups that work alike in a different way. We manually obtained environmental preference information for as many of the cultured organisms present in our reference tree as possible. Then, we tested how good the tree topology was at discriminating groups of sequences putatively adapted to work in given ranges of the environmental variables.

When evaluating sequentially all nodes in the tree, we found that some internal nodes had a significant correspondence with particular environmental variables. These nodes divided the tree into highly paraphyletic clades containing sequences that correlate with salinity, pH, and temperature (Fig. 5). This result is consistent with the previous literature on the topic. In addition, motility was also very significant for the same group of sequences related to pH variability. One possibility would be that these environmentally consistent clades were phylogenetically close. That is, the adaptation of the functional protein to a certain environmental condition would have been vertically inherited. However, this was not the case. We found that the divergence of the bigger groups was not explained by taxonomy (Sup. Fig. D). We then defined the implementations of the function as the sequence spaces shown in Figure 5.

Decoupling taxonomy and function

We have argued above that taxonomic position is not, in many cases, synonymous with function. In addition to those arguments, there is some research actually focused on decoupling taxonomy from functional assets (Louca et al., 2016; Tamames et al., 2016). Moreover, machine learning approaches appear to outperform niche prediction with functions rather than phylogeny (Alneberg et al., 2020). This means that, at least in specific cases, it is possible to better predict the occurrence of function in an environment by its physicochemical features rather than by the taxonomic composition detected therein (Tamames et al., 2016).

Our guild definition can partially avoid the latter issue, because it can be used to discriminate these taxonomic effects from those caused by functional convergence in order to dissect how the function is implemented through a battery of environments. Even if the taxonomic assignment is biased or not very predictive, it is complemented by the information from the implementations that perform the function.

As shown in the results section, the acquisition of putrescine-like polyamines is a ubiquitous trait in the examined ocean layers, which is consistent with previous literature on the topic (Bergauer et al., 2018). However, we added novel insights about the guild changes with depth. The guild presents itself, however, in multiple forms; it changes both its taxonomic composition and the implementations mostly found, and seems to follow trends that correspond to the different physico-chemical characteristics intrinsically linked to the three analyzed zones of the ocean: bathypelagic (4000-1000m), mesopelagic (1000-200m) and epipelagic (200-0m).

There are characteristic guild patterns that seem to be better explained by depth than by sampling spot or latitude. Our results show that, in most cases, this function is carried out by a lot of different taxa and all types of polyamine uptake implementations. 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).

In Bergauer's study, different metabolic traits were studied to analyze microbial heterotrophy in different ocean layers. What our approach adds is, fundamentally, three things: (1) depuration of the truly functional space, (2) discrimination between purely taxonomic effects and those that are not related to taxonomy, and (3) determination whether function responds positively by unexpected diversification of its effectors. In addition, our approach allows rapid visual comparison of the guild pattern. Finally, it makes the comparison between different guilds easier, as the ecological values are standardized by the same theoretical framework, without assuming that the importance of a function in an environment depends solely on the abundance of automatic annotated genes.

Correction for expected richness of an implementation

A correction for expected sequence richness was introduced to estimate the importance of a molecular function because of the following reasons: (i) empirically, we observed that each gene grows in richness of unique sequences differently with relative abundance, as seen in Figure 3; (ii) abundance values for lower than expected richness can be explained by the strong dominance of an organism in a particular sample, but it does not imply that this function is responsible for the ecological success of the dominant organism, so low-richness abundances will be overestimating the importance of the function; (iii) higher than expected richness should result in a higher function robustness, since the loss of fitness for the global function regarding environmental changes should be reduced as the sequence space widens.

In other words, our model rewards versatile behaviors for the same function in an environment, inferred by the richness of its effectors, as long as d_{obs} is greater than d_{exp}. This is because we postulate that unexpected sequence diversification increases the odds that the function will persist in the environment when exposed to undefined changes. This phenomenon, although not formally described, has been proposed in a multitude of different biological systems (Wright et al., 2005; Hakes et al., 2007; Föhse et al., 2011; García-García et al., 2019).

For example, as can be seen in Figure 8, the fold change in k_{meso} has a similar behavior among implementations related to adaptation to plasticity in pH ($cIIa$ and $cIIb$). This very noticeable increase is shown to be exclusive for this type of implementations, and can be explained by the rapid depth-dependent acidification, a characteristic feature of the mesopelagic oxycline (Park, 1966; Dickson, 1993). In general, the lowest pH levels in the water column correspond with the presence of an oxygen minimum. The decrease in pH is mainly driven by the increased concentration of dissolved carbonic acid, which also relates to biological activity of upper layers (Sup. Fig. G). Values of pH are also dependent on more strictly abiotic factors such as temperature, salinity and pressure, acting as dissociation constant modifiers (Byrne et al., 1999; Ternon et al., 2001). Therefore, the pH minimum is strongly linked to mesopelagic depths and may exhibit some seasonality. So, as shown, we can relate or even anticipate complex functional dynamics in a particular ecosystem.

Importance of the guild concept to study microbial functions

There has been increasing interest in developing metagenomic studies based on guilds. A recent approach proposes a model that identifies potential functions through patterns of variation in species abundance and ecosystem properties across microbial communities (Shan and Cordero, 2023). Although we find this tool exciting and useful for identifying putative top contributors to a function in an environment, it has three shortcomings: (1) their model assumes a strict relationship between taxonomy and function (2) its usefulness strongly depends on the correlations with measurements of nutrients or substrates, which are costly and perhaps time-dependent, and

(3) it does not solve the existing problems in the guild definition as applied to microbes.

CONCLUSIONS

First, the original definition of guild suffers when applied to microbes, and has often been used in an intuitive way. Second, just as the genetic code is degenerate because the same amino acid can be the translation of different triplets, any protein function is also degenerate because an indefinite set of sequences can perform it. In fact, the set of sequences that can perform a specific function does not necessarily maintain a close evolutionary history. With these issues in mind, we propose a theoretical redefinition of the term guild to bring the ecology of microbial functions into a quantitative framework, considering its evolving nature. Furthermore, our definition of guild remains quantitative and easily applicable to all other organisms. We also developed methodological procedures and bioinformatics tools to facilitate its use by the community (https://github.com/pyubero/microguilds).

Regarding the technical issues, the potential for exploring functional ecology in microorganisms has been limited by the overwhelming amount of massive and imprecise omics data. Nevertheless, we have been able to partially avoid the dilemma of "automatic functional annotation black boxes" and describe some ecological trends within a complex function and ecosystem using reference phylogenetic reconstructions as functional sequence classifiers.

There are four main arguments that justify the present work: (i) the original definition of guild 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 living beings; (ii) the emergence of omics data, involving technical biases and overwhelming 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 quantitative nor ecologically relevant.

MATERIALS AND METHODS.

Construction of the marine prokaryotic genomes database (1 in Sup. Figure A)

To facilitate the construction of the gene-specific reference databases, we compiled a database of peptide sequences obtained from a collection of prokaryotic, quality-filtered genomes (MAGs and SAGs) from marine environments. Specifically, we retrieved genomes from the following databases: a) the MAR database, all 1,270 complete genomes, and 5,521 partial genomes that had the "high quality" status as described in (Klemetsen et al., 2018); b) the OceanDNA database (Nishimura and Yoshizawa, 2022), all 52,325 genomes, since they had been quality-filtered based on their completeness and degree of contamination with the formula: percent completeness - 5 × percent-contamination $\geq$ 50); c) the collection compiled by (Paoli et al., 2022), which includes genomes from various origins such as TARA OCEANS (Sunagawa et al., 2015) and XORG, in this case, only genomes passing the same quality filter applied in OceanDNA were kept, amounting to a total of 26,942 additional genomes. All the genomes considered had assigned taxonomy obtained with the GTDB Toolkit (Chaumeil et al., 2020) version 2.0.0 available in their corresponding databases. Finally, all sequences were merged into a single database, reads were further quality filtered with fastp version 0.20.1, and sequence duplicates were removed with seqkit rmdud (Shen et al., 2016) version 2.0.0 using default parameters.

Functional marker selection (2a in Sup. Figure A)

The search for functional 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 maximize functional univocity. In order to choose an HMM as a guild marker, we followed the following conservative criteria: (i) the construction of the HMM must be congruent with the sequences that have reviewed functional evidence in literature, (ii) 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) or a consequence of the evolutive history of the gene (i.e.: similar sequences that have undergone functional drift). With this methodology, we selected the best minimal markers for the guilds analyzed in this work.

Construction of the gene-specific reference database (2b in Sup. Figure A)

We used the selected profile HMMs and HMMER3 (Johnson et al., 2010) to retrieve candidate sequences of the target gene from our collected marine peptide database. Gather score thresholds were used as a quality filter when available, otherwise, a minimum E-score threshold of 10^{-9} was employed.

To facilitate inference and later visual inspection of the phylogenetic trees, sequence hits were further filtered to set a maximum database size of N representative sequences. To this end, we applied a series of filters. First, we set minimal and maximal sequence length cutoff values of *l1* and *l2*, respectively. Second, we removed sequence duplicates through seqkit's rmdup sub-command with default parameters. Third, we applied CD-HIT (Fu et al., 2012) with default parameters to reduce redundancy in the peptide database. Finally, if the database size was larger than the allowed maximum after applying CD-HIT, we further reduced the number of representative peptides through RepSet (Libbrecht et al., 2018), an optimization-based algorithm that obtains a series of nested sets of representative peptides of decreasing size. Specifically, we selected the maximal set of representative peptides with a size lower than the established size threshold value.

Usage of synteny during gene-specific reference database construction (2c in Sup. Figure A)

In some cases, we used syntenic information to reduce uncertainty due to the potential presence of paralogs during the reference peptide database construction. To this end, we employed the Python package Pynteny (Robaina-Estévez, 2022), which facilitates synteny-aware profile, HMM-based searchers. After generating a list of synteny-complaint target-peptide matches, we followed the same protocol to reduce database size when required to meet the established reference database size threshold value.

Inference of the gene-specific reference trees (3a in Sup. Figure A)

Once the peptide reference database was constructed, we employed MUSCLE (Edgar, 2004) with default parameter values to perform a multiple sequence alignment of the reference database. Next, we used the previous alignment and IQ-TREE (Minh et al., 2020) with default parameter values to build a reference phylogenetic tree for each target gene. We determined the substitution model through ModelTest (Darriba et al., 2020) by selecting the model with the highest AIC score.

Classification of clusters within the reference phylogenetic tree (3b in Sup. Figure A)

Once we have constructed the reference tree, we can now propagate functional information that corresponds to the different regions of the reference tree. In the case of *amoA*, the functional information was obtained directly from the sequences we used to build the tree, and the clusters inferred from the similarity between sequences.

In the case of polyamine binding reference tree, we needed other criteria to classify clusters. To check whether different clusters are associated with different environmental conditions, we carried out an extensive literature search of the environmental preferences of 321 species that matched 478 leaves (41%). We assembled a curated collection of physicochemical preferences for these species that included tolerance ranges and optimal values of temperature, salinity and pH, as well as other variables such as motility (Sup. Table 1). For each internal node we calculated the average values of all its leaves. To determine whether the association with environmental variables of the cluster were significant, these node averages were compared to the distribution observed under $2 \cdot 10^4$ randomizations to obtain their z-scores. Nodes with an average value of the z-score larger than 3, i.e. p-value $\leq$ 0.003 were considered significant for the particular environmental variable (Sup. Table 2). To select the most general internal nodes, we focused on those that are significant but whose parent node is not. These are color coded in Figure 5 and Figure 6.

To determine whether these sequence clusters were expected by the taxonomy of the organisms, we constructed a null model of phylogenetic divergence with two ribosomal phylomarkers (16s and *rplB*). We then compared the divergence of *potF*-like sequences for the same organisms, finding that taxonomy does not explain, in most cases, the drift found in functional genes; however, environmental variables do, especially for nodes that separate more leaves on the tree, as shown in Supplementary Figure D.

This methodology can be applied to any type of functional evidence, not only environmental, but also kinetic, substrate preference, or any other type of evidence.

Preprocessing of query sequences (4a in Sup. Figure A)

Query sequences were retrieved from the metagenomes following SqueezeMeta's pipeline (Tamames and Puente-Sánchez, 2019).

Placement of query sequences (4b in Sup. Figure E)

To place query sequences (metagenomic output) in the reference tree, we first obtained an alignment between the query and the reference sequences with papara (Berger and Stamatakis, 2012) version 2.5. Then, we placed query sequences with the tool EPA-ng (Barbera et al., 2019) version 0.3.8. Additionally, we employed the Gappa toolkit version 0.8, specifically, the command gappa examine graft (Czech et al., 2020) to visualize the placed sequences on the reference tree using default parameters. The phylogenetic placement tree was visualized using the Interactive Tree of Life (Letunic and Bork, 2016).

Taxonomical and functional labeling of placed query sequences (4c in Sup. Figure A)

We employed the Gappa toolkit, specifically, the command *gappa examine assign* to assign taxonomy to placed sequences. Briefly, Gappa first assigns a consensus taxonomy to each internal node of the tree and then assigns to each query sequence the closest taxonomy in the reference tree weighted by the likelihood of each placement. We employed default parameters and the *best_hit* option to retrieve only the taxonomic assignments with the highest total placement likelihood for each query. To assign functional labels to placed queries, we selected the function of the tree cluster in which each query had been placed. To this end, we first added the cluster label to each taxonomic path of the reference sequences as an additional (artificial) taxon above the domain level. In this manner, we could employ *gappa examine assign* to assign both taxonomy and the cluster label (i.e., function) to each placed query.

Quantification of Polyamine-uptakers guild (5 in Sup. Figure A)

Once filtered sequences are classified by environment, taxon, and implementation, they are merged together with the corresponding normalized abundances into a single master table. This is the input for the first tool of our public repository (https://github.com/pyubero/microguilds), a python module called *guild_tensor_generate*. The module will extract all the required information for the calculation of each implementation, taxon and environment-dependent functional contribution, k. In the present case, we study three distinct environments, so the software will produce an array of dimension $3 \times m \times l$ (where m is the number of implementations established within the guild marker, and l the number of taxa).

The calculation contemplates three terms. The abundance, a_s, has been calculated as a summation of normalized metagenomic counts for all the sequences contained in the same implementation, taxon, and environment. The second term is d. Theoretical d is the unexpected sequence diversification according to the sum of a, the first term. Calculation of the theoretical d is complex and would require avoiding false negatives. Therefore, in our work it is limited by the technique of retrieving this kind of data, as shown in Figure 2. Finally, the term representing the univocity of the implementations, u, is equal to 1.0 since we discard the metagenomic sequences falling into non-functional sequence spaces of the reference tree, or false positives. In addition, we had a highly-conservative criteria to estimate the functional sequence space, as described also in methods. Ideally, environmental inhibition of the effector must be considered for the univocity calculation, but since we lack the data, we have decided that there is no inhibition for this example. An example of the k-tensor output is provided in the Supplementary Table 3.

The second tool, *guild_tensor_visualize*, helps to visualize this tensor, which can be of varying complexity. It does two things: (i) it filters by the taxonomic level to visualize the guild patterns and (ii) it takes the value of k by taxonomic contribution to each implementation and environment. To do the latter, it takes the contribution of each position in the tensor and plot them with different preferences (*top contributors* or *rare taxa*, *linear* or *log* representation, *polar* or *rectilinear* charting, etc.) as shown in Fig. 7, resulting in an easy way to visualize complex data.

16S and rplB sequences

To screen phylogenetic deviations between functions and phylomarkers (Sup. Fig. D), we obtained the nucleotide sequences of the 16S ribosomal subunit and the *rplB* 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.

AKCNOWLEDGEMENTS

We are grateful to Juan F. Poyatos (CNB-CSIC) for discussion of the methodologies and their implications, and to Alberto Pascual-García (CNB-CSIC) for the feedback and discussion on the epistemology of the microbial guild. We also thank the CSIC-LifeHUB forum (PIE-202120E047-Conexiones-Life) for generating space for discussion and allowing these ideas to mature.

AUTHOR CONTRIBUTIONS

J.R.S.: Conceptualization, Methodology, Validation, Writing - Original Draft, Investigation, Data curation, Visualization; P.Y.: Methodology, Software, Formal analysis, Writing - Review & Editing, Data curation, Visualization; S.R.E.: Methodology, Writing - Review & Editing; J.M.G.: Methodology; J.T. and C.P.A.: Writing - Review & Editing, Project administration, Supervision, Funding acquisition.

FUNDING

Project PID2019-110011RB-C31 funded by MCIN/AEI/10.13039/501100011033 and work supproted by Ph.D. fellowship PRE2020-096130 from the Spanish Ministerio de Ciencia e Innovación and the European Social Fund.

COMPETING INTERESTS

The authors declare no competing interests.

MATERIALS AND CORRESPONDENCE

Correspondence and request for materials should be addressed to J.R.S.

REFERENCES

Akashi, H., Osada, N., and Ohta, T. (2012). “Weak selection and protein evolution.” *Genetics*, 192(1), 15–31.

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.

Alves, R. J. E., Minh, B. Q., Urich, T., von Haeseler, A., and Schleper, C. (2018). “Unifying the global phylogeny and environmental distribution of ammonia-oxidising archaea based on *amoA* genes.” *Nature communications*, 9(1), 1517.

Arrigo, K. R. (2005). “Marine microorganisms and global nutrient cycles.” *Nature*, 437(7057), 349–355.

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

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., and Stamatakis, A. (2019). “EPA-ng: massively parallel evolutionary placement of genetic sequences.” *Systematic biology*, 68(2), 365–369.

Bartlett, D. H. (1999). “Microbial adaptations to the psychrosphere/piezosphere.” *Journal of molecular microbiology and biotechnology*, 1(1), 93–100.

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.

Berger, S. A. and Stamatakis, A. (2012). "Papara 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension." *Heidelberg Institute for Theoretical Studies*, 12.

Byrne, R. H., McElligott, S., Feely, R., and Millero, F. (1999). "The role of pHT measurements in marine CO2-system characterizations." *Deep Sea Research Part I: Oceanographic Research Papers*, 46(11), 1985–1997.

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.

Cech, T. (2009). "Evolution of biological catalysis: ribozyme to RNP enzyme." *Cold Spring Harbor symposia on quantitative biology*, Vol. 74, Cold Spring Harbor Laboratory Press, 11–16.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). "GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database." *Bioinformatics*, 36(6), 1925–1927.

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.

Czech, L., Barbera, P., and Stamatakis, A. (2020). "Genesis and gappa: processing, analyzing and visualizing phylogenetic (placement) data." *Bioinformatics*, 36(10), 3263–3265.

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., and Flouri, T. (2020). "ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models." *Molecular biology and evolution*, 37(1), 291–294.

Deng, C., Cheng, C.-H. C., Ye, H., He, X., and Chen, L. (2010). "Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict." *Proceedings of the National Academy of Sciences*, 107(50), 21593–21598.

Dickson, A. G. (1993). “The measurement of sea water ph.” *Marine chemistry*, 44(2-4), 131–142.

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.

Edgar, R. C. (2004). “MUSCLE: a multiple sequence alignment method with reduced time and space complexity.” *BMC bioinformatics*, 5(1), 1–19.

Ensign, S. A., Hyman, M. R., and Arp, D. J. (1993). “In vitro activation of ammonia monooxygenase from *Nitrosomonas europaea* by copper.” *Journal of bacteriology*, 175(7), 1971–1980.

Fauth, J., Bernardo, J., Camara, M., Resetarits Jr, W., Van Buskirk, J., and McCollum, S. (1996). “Simplifying the jargon of community ecology: a conceptual approach.” *The American Naturalist*, 147(2), 282–286.

Fay, J. C. and Wu, C.-I. (2003). “Sequence divergence, functional constraint, and selection in protein evolution.” *Annual review of genomics and human genetics*, 4(1), 213–235.

Fisher, M., Gokhman, I., Pick, U., and Zamir, A. (1997). “A structurally novel transferrin-like protein accumulates in the plasma membrane of the unicellular green alga dunaliella salina grown in high salinities.” *Journal of biological chemistry*, 272(3), 1565–1570.

Föhse, L., Suffner, J., Suhre, K., Wahl, B., Lindner, C., Lee, C.-W., Schmitz, S., Haas, J. D., Lamprecht, S., Koenecke, C., et al. (2011). “High TCR diversity ensures optimal function andhomeostasis of Foxp3+ regulatory Tcells.” *European journal of immunology*, 41(11), 3101–3113.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). “CD-HIT: accelerated for clustering the next-generation sequencing data.” *Bioinformatics*, 28(23), 3150–3152.

García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C., and Puente-Sánchez, F. (2019). “Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions.” *The ISME journal*, 13(12), 2969–2983.

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.

Glasner, M. E., Truong, D. P., and Morse, B. C. (2020). "How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation." *The FEBS journal*, 287(7), 1323–1342.

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.

Hakes, L., Lovell, S. C., Oliver, S. G., and Robertson, D. L. (2007). "Specificity in protein interactions and its relationship with sequence diversity and coevolution." *Proceedings of the National Academy of Sciences*, 104(19), 7999–8004.

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.

Hult, K. and Berglund, P. (2007). "Enzyme promiscuity: mechanism and applications." *Trends in biotechnology*, 25(5), 231–238.

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

Huston, A. L., Haeggström, J. Z., and Feller, G. (2008). "Cold adaptation of enzymes: Structural, kinetic and microcalorimetric characterizations of an aminopeptidase from the Arctic psychrophile *Colwellia psychrerythraea* and of human leukotriene a4 hydrolase." *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1784(11), 1865–1872.

Johansson, H. E., Johansson, M. K., Wong, A. C., Armstrong, E. S., Peterson, E. J., Grant, R. E., Roy, M. A., Reddington, M. V., and Cook, R. M. (2011). "Bti1, an azoreductase with ph-dependent substrate specificity." *Applied and environmental microbiology*, 77(12), 4223–4225.

Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). "Hidden markov model speed heuristic and iterative HMM search procedure." *BMC bioinformatics*, 11(1), 1–8.

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.

Khadka, R., Clothier, L., Wang, L., Lim, C. K., Klotz, M. G., and Dunfield, P. F. (2018). "Evolutionary history of copper membrane monooxygenases." *Frontiers in microbiology*, 9, 2493.

Kimura, M. (1991). "The neutral theory of molecular evolution: a review of recent evidence." *The Japanese Journal of Genetics*, 66(4), 367–386.

Klemetsen, T., Raknes, I. A., Fu, J., Agafonov, A., Balasundaram, S. V., Tartari, G., Robertsen, E., and Willassen, N. P. (2018). "The MAR databases: development and implementation of databases specific for marine metagenomics." *Nucleic acids research*, 46(D1), D692–D699.

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.

Kreitman, M. (1996). “The neutral theory is dead. Long live the neutral theory.” *Bioessays*, 18(8), 678–683.

Kumar, S., Singh, S. K., and Gromiha, M. M. (2009). “Temperature-dependent molecular adaptations, microbial proteins.” *Encyclopedia of industrial biotechnology: bioprocess, bioseparation, and cell technology*, 1–22.

Ladero, M., Ruiz, G., Pessela, B., Vian, A., Santos, A., and Garcia-Ochoa, F. (2006). “Thermal and ph inactivation of an immobilized thermostable β-galactosidase from thermus sp. strain t2: Comparison to the free enzyme.” *Biochemical Engineering Journal*, 31(1), 14–24.

Lanyi, J. K. (1974). “Salt-dependent properties of proteins from extremely halophilic bacteria.” *Bacteriological reviews*, 38(3), 272–290.

Lau, E., Lukich, D., and Le, T. (2016). “The evolution of the *amoA*, *pmoA* and *bmoA* genes measured by ka/ks ratios.” *Proceedings of the West Virginia Academy of Science*, 88(1).

Letunic, I. and Bork, P. (2016). “Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees.” *Nucleic acids research*, 44(W1), W242–W245.

Libbrecht, M. W., Bilmes, J. A., and Noble, W. S. (2018). “Choosing non-redundant representative subsets of protein sequence data sets using submodular optimization.” *Proteins: Structure, Function, and Bioinformatics*, 86(4), 454–466.

Liberles, D. A., Teichmann, S. A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L. J., De Koning, A. J., Dokholyan, N. V., Echave, J., et al. (2012). “The interface of protein structure, protein biophysics, and molecular evolution.” *Protein Science*, 21(6), 769–785.

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.

Martens-Habbena, W., Berube, P. M., Urakawa, H., de La Torre, J. R., and Stahl, D. A. (2009). "Ammonia oxidation kinetics determine niche separation of nitrifying archaea and bacteria." *Nature*, 461(7266), 976–979.

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.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., and Lanfear, R. (2020). "IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era." *Molecular biology and evolution*, 37(5), 1530–1534.

Moreno-Ulloa, A., Sicairos Diaz, V., Tejeda-Mora, J. A., Macias Contreras, M. I., Castillo, F. D., Guerrero, A., Gonzalez Sanchez, R., Mendoza-Porras, O., Vazquez Duhalt, R., and Licea-Navarro, A. (2020). "Chemical profiling provides insights into the metabolic machinery of hydrocarbon-degrading deep-sea microbes." *Msystems*, 5(6), 10–1128.

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.

Nishimura, Y. and Yoshizawa, S. (2022). "The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments." *Scientific Data*, 9(1), 305.

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.

Oudova-Rivera, B., Wright, C. L., Crombie, A. T., Murrell, J. C., and Lehtovirta-Morley, L. E. (2023). "The effect of methane and methanol on the terrestrial ammonia oxidising archaeon '*Candidatus Nitrosocosmicus franklandus* c13'." *Environmental Microbiology*.

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.

Paoli, L., Ruscheweyh, H.-J., Forneris, C. C., Hubrich, F., Kautsar, S., Bhushan, A., Lotti, A., Clayssen, Q., Salazar, G., Milanese, A., et al. (2022). "Biosynthetic potential of the global ocean microbiome." *Nature*, 607(7917), 111–118.

Park, K. (1966). "Deep-sea pH." *Science*, 154(3756), 1540–1542.

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.

Pistocchi, R., Kashiwagi, K., Miyamoto, S., Nukui, E., Sadakata, Y., Kobayashi, H., and Igarashi, K. (1993). “Characteristics of the operon for a putrescine transport system that maps at 19 minutes on the *Escherichia coli* chromosome.” *Journal of Biological Chemistry*, 268(1), 146–152.

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*.

Reed, C. J., Lewis, H., Trejo, E., Winston, V., and Evilia, C. (2013). “Protein adaptations in archaeal extremophiles.” *Archaea*, 2013.

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.

Robaina-Estévez, S. (2022). “Pynteny: Synteny-aware hmm searches made easy.” *Zenodo*.

Rochman, F. F., Kwon, M., Khadka, R., Tamas, I., Lopez-Jauregui, A. A., Sheremet, A., V. Smirnova, A., Malmstrom, R. R., Yoon, S., Woyke, T., et al. (2020). “Novel copper-containing membrane monooxygenases (CuMMOs) encoded by alkane-utilizing Betaproteobacteria.” *The ISME journal*, 14(3), 714–726.

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

Rotthauwe, J.-H., Witzel, K.-P., and Liesack, W. (1997). “The ammonia monooxygenase structural gene *amoA* as a functional marker: molecular fine-scale analysis of natural ammonia-oxidizing populations." *Applied and environmental microbiology*, 63(12), 4704–4712.

Ruelens, P., Wynands, T., and de Visser, J. A. G. (2023). "Interaction between mutation type and gene pleiotropy drives parallel evolution in the laboratory." *Philosophical Transactions of the Royal Society B*, 378(1877), 20220051.

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

Sayavedra-Soto, L. A., Hamamura, N., Liu, C.-W., Kimbrel, J. A., Chang, J. H., and Arp, D. J. (2011). "The membrane-associated monooxygenase in the butane-oxidizing gram-positive bacterium nocardioides sp. strain cf8 is a novel member of the amo/pmo family." *Environmental microbiology reports*, 3(3), 390–396.

Shan, X. and Cordero, O. X. (2023). "Identifying microbial guilds on the basis of ecological patterns." *Nature Ecology & Evolution*.

Shen, W., Le, S., Li, Y., and Hu, F. (2016). "SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation." *PloS one*, 11(10), e0163962.

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.

Stüeken, E. E., Catling, D. C., and Buick, R. (2012). "Contributions to late archaean sulphur cycling by life on land." *Nature Geoscience*, 5(10), 722–725.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., et al. (2015). “Structure and function of the global ocean microbiome.” *Science*, 348(6237), 1261359.

Tamames, J., Abellán, J. J., Pignatelli, M., Camacho, A., and Moya, A. (2010). “Environmental distribution of prokaryotic taxa.” *BMC microbiology*, 10(1), 1–14.

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

Tamames, J., Sánchez, P. D., Nikel, P. I., and 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.

Tamuri, A. U. and Dos Reis, M. (2022). “A mutation–selection model of protein evolution under persistent positive selection.” *Molecular Biology and Evolution*, 39(1), 309.

Ternon, J.-F., Oudot, C., Gourlaouen, V., and Diverres, D. (2001). “The determination of pHT in the equatorial atlantic ocean and its role in the sound absorption modeling in seawater.” *Journal of marine systems*, 30(1-2), 67–87.

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 β-galactosidase of *Escherichia coli*." *Journal of molecular biology*, 32(1), 1–13.

Valencia, A. (2005). "Automatic annotation of protein function." *Current opinion in structural biology*, 15(3), 267–274.

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." 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), 064.

Ward, B. (1987). "Kinetic studies on ammonia and methane oxidation by *Nitrosococcus oceanus*." *Archives of Microbiology*, 147, 126–133.

Wright, C. F., Teichmann, S. A., Clarke, J., and Dobson, C. M. (2005). "The importance of sequence diversity in the aggregation and evolution of proteins." *Nature*, 438(7069), 878–881.

Wright, C. L., Schatteman, A., Crombie, A. T., Murrell, J. C., and Lehtovirta-Morley, L. E. (2020). "Inhibition of ammonia monooxygenase from ammonia-oxidizing archaea by linear and aromatic alkynes." *Applied and environmental microbiology*, 86(9), e02388–19.

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.

Zaks, A. and Klibanov, A. M. (1986). "Substrate specificity of enzymes in organic solvents vs. water is reversed." *Journal of the American Chemical Society*, 108(10), 2767–2768.

Zhao, Q. (2022). "Molecular and thermodynamic mechanisms for protein adaptation." *European Biophysics Journal*, 51(7-8), 519–534.

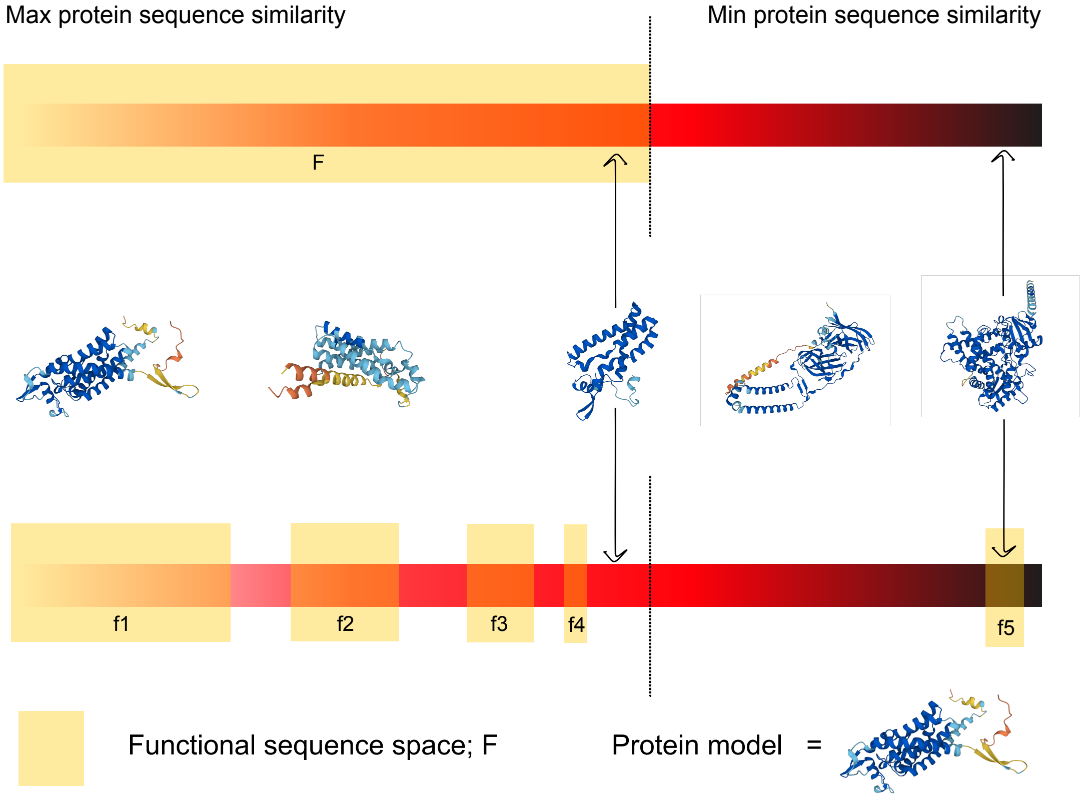
Max protein sequence similarity
Min protein sequence similarity
F
f1
f2
f3
f4
f5
Functional sequence space; F
Protein model =

Figure 1. Sequence spaces representing protein dissimilarity and function. A particular function can be performed by a plethora of different protein sequences, generated by adaptive evolution and drift. The traditional models for assigning function to the reconstructed genomic data are based solely on protein sequence similarity. The models automatically annotate as functional proteins those below a certain threshold and discards all those beyond the threshold (above). Although the threshold can be adjusted, nature seems to fit better the theoretical model below, which considers three casuistries: (i) the threshold value is prone to errors, (ii) the functional space may display discontinuties or gaps, and (iii) proteins beyond the dissimilarity threshold may be able to perform the function (f5).

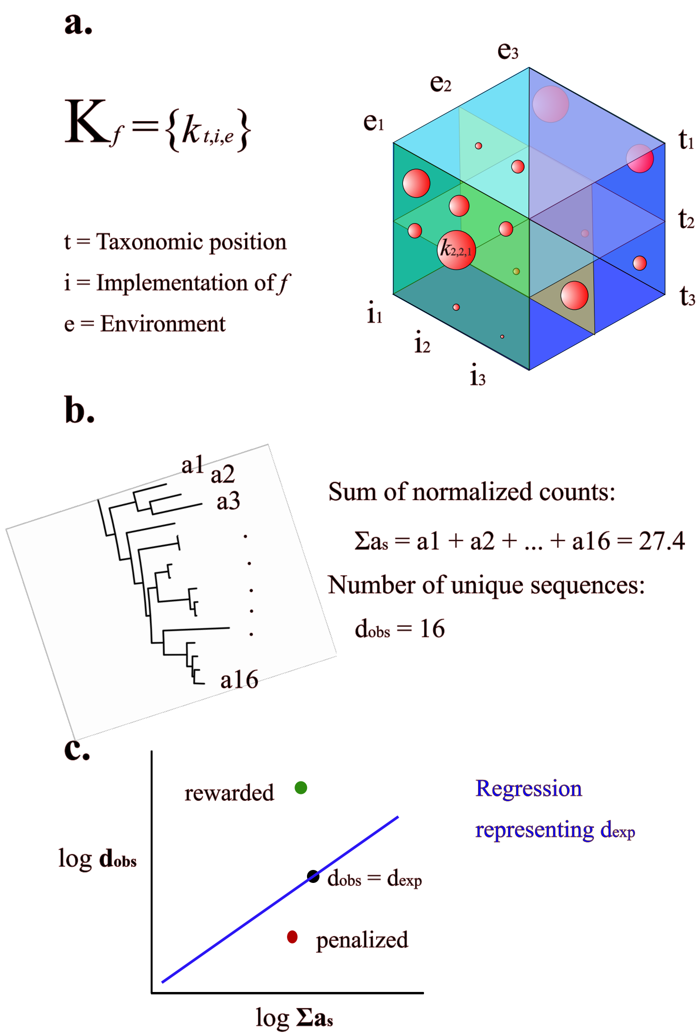

Figure 2. Quantification of guilds: mock example of *k* calculation. a. Visual concept of the three-dimensional object that quantifies the importance of the guild in different contexts. The guild structure can be defined as each of the impact coefficients k that the definitory function has on each triplet taxon, implementation, and environment ($k_{t,i,e}$). **b.** Mock example of sequence abundance and observed diversity calculation. To calculate k, we need to sum up all the corresponding sequence abundances at a particular position (taxon, implementation, environment). We also compute d_{obs} as the count of the unique sequences found in that position (a1 to a16 in this case). **c.** To calculate k, we need to correct the sequence abundances by the sequence diversification expected from abundance. This expected richness is calculated from log-log regression. Thus, values of k reward sequence diversities higher than expected for an abundance value (when $d_{obs} > d_{exp}$), and penalize them otherwise. The empirical model is based on all observations of a gene in our database.

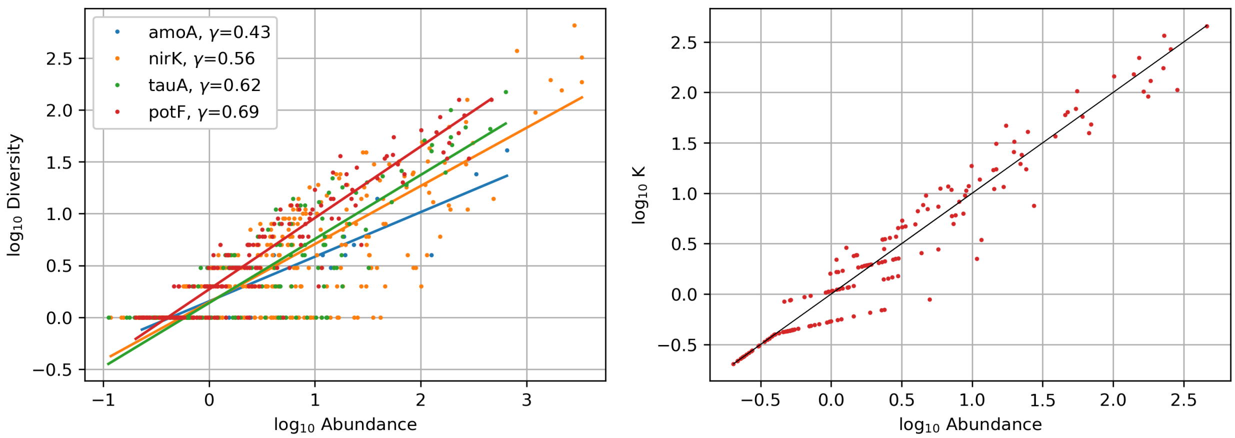

Figure 3. Practical *k* calculation with metagenomic data. a. Log-log regression of observed diversity and abundance in four different genes (function markers). In attempting to compare the sum of abundance for a taxon, environment, and implementation with its unique sequence richness, we expected an evident relationship. And certainly, most cases appear to follow a logarithmic trend. However, we found that (i) sometimes there are abundance values that do not predict the observed sequence richness (ii) although the r-squared values of the regression are very consistent (mean of $\simeq 0.8$), the slope seems characteristic of each gene. This translates as, apparently, each gene grows in sequence richness (γ) differently with abundance. **b.** Example of how the value of k changes with the value of the sum of abundances in *potF* gene. We use this empirical model to estimate the expected diversity for a given abundance, so we can positively weight which abundances are richer in unique sequences. Conversely, this model penalizes abundances that have lower than expected sequence diversity.

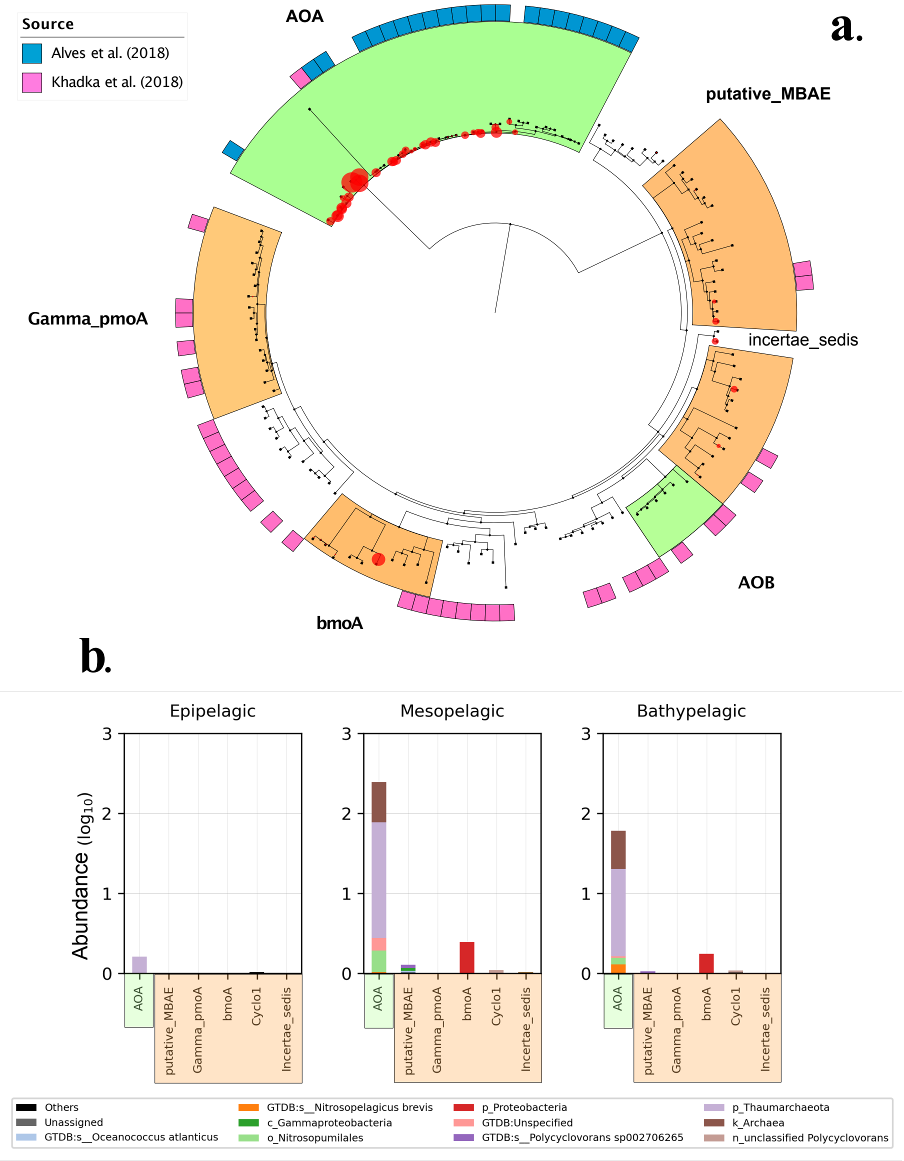

Figure 4. Distinguishing non-functional from functional sequences capable of oxidizing ammonia improves guild assessment. a. Reconstructed phylogeny of *amoA*, used as a reference tree to classify ammonia-oxidizing capable sequences. The tree contains 135 sequences with strong functional evidence based on either biochemical or physiological features, or inferred by homology to quality sequences (see Methods). For clarity, we only highlight clusters of sequences where metagenomic Malaspina samples have been successfully placed (full clustering in Sup. Fig. B). Among those, we distinguish sequence clusters with proven ammonium oxidation function (shaded greens) from sequences with probably a broader substrate spectrum and sequences with evidence of being non-functional for ammonia oxidation (shaded orange). Evidence of function was gathered from various sources, albeit the main ones are marked in pink and blue. **b.** Log representation of abundance values (TPM) of the *amoA* classified queries found in Malaspina metagenomes (red circles from a.). 31% of the unique sequences (1.01% of total TPM) obtained by automatic means are excluded with a conservative criterion (orange-shaded boxes corresponding to the non-univocal tree clusters).

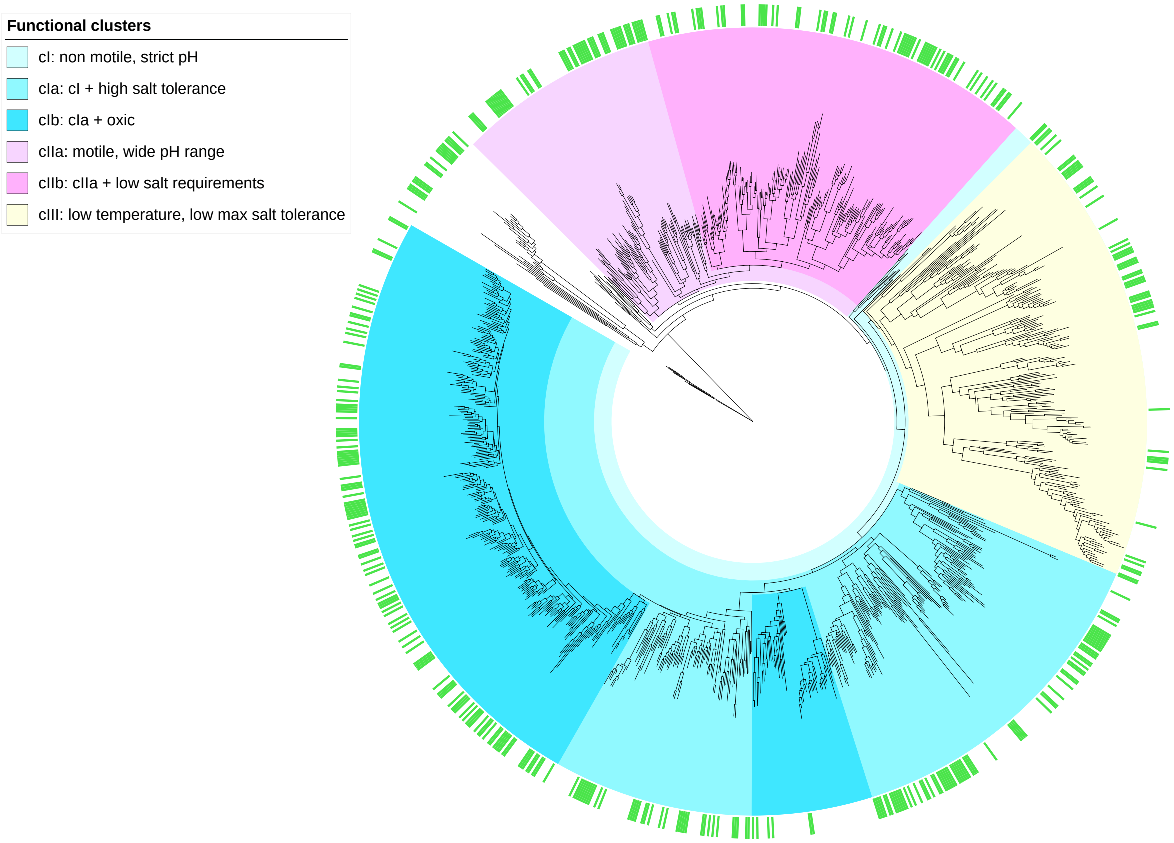
Functional clusters
cI: non motile, strict pH
cIa: cI + high salt tolerance
cIb: cIa + oxic
cIIa: motile, wide pH range
cIIb: cIIa + low salt requirements
cIII: low temperature, low max salt tolerance

Figure 5. Functional clustering of the putrescine-like binding protein reference tree. The phylogenetic reconstruction corresponds to a polyamine-binding subunit of an ABC transporter. This tree will act as a metagenomic query classifier. However, defining functional sequence spaces below the threshold becomes complicated when there is a lack of experimental evidence. To avoid this dilemma, we focused on determining sequence spaces that may be affected by, and therefore adapted to, environmental variables. To determine this, environmental evidence vectors have been established for each organism represented by one or more leaves in the tree. Fifteen environmental variables have been curated for 321 organisms in pure culture, representing 478 of 1158 possible tree locations (green tags). In addition, these evidence labels are well distributed throughout the tree. Then, we built a null model by randomizing the environmental evidence labels so that the topology holds, to see how enriched the nodes are for these screened variables. The result consist of the colored regions representing significant nodes (one-tailed p-values < 0.003).

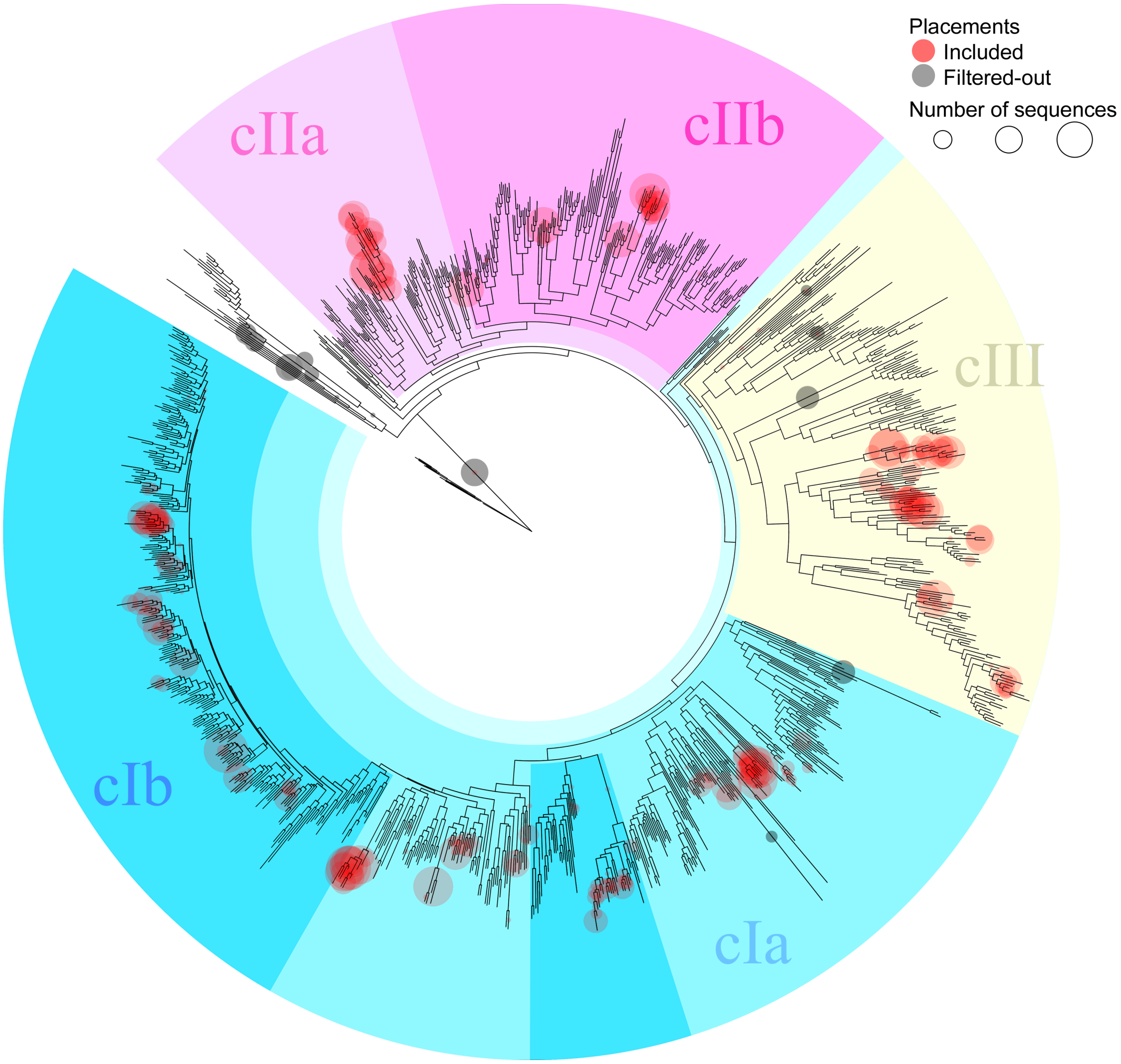

Figure 6. Placement of short environmental sequences from Malaspina samples in the putrescine-like polyamines uptake reference tree. The tree acts as a classifier of placed sequences (circles) that are close to being functionally synonymous by sharing environmental features. However, placed metagenomic sequences that do not fit well in the tree will be subject to further filtering (Methods). Most metagenomic queries are considered functional (red circles) while a small fraction are filtered out (4.13%, grey circles). All five implementations are represented in the 75 metagenomic samples used in this study distributed in the three main oceanic environments: epipelagic, mesopelagic, and bathypelagic.

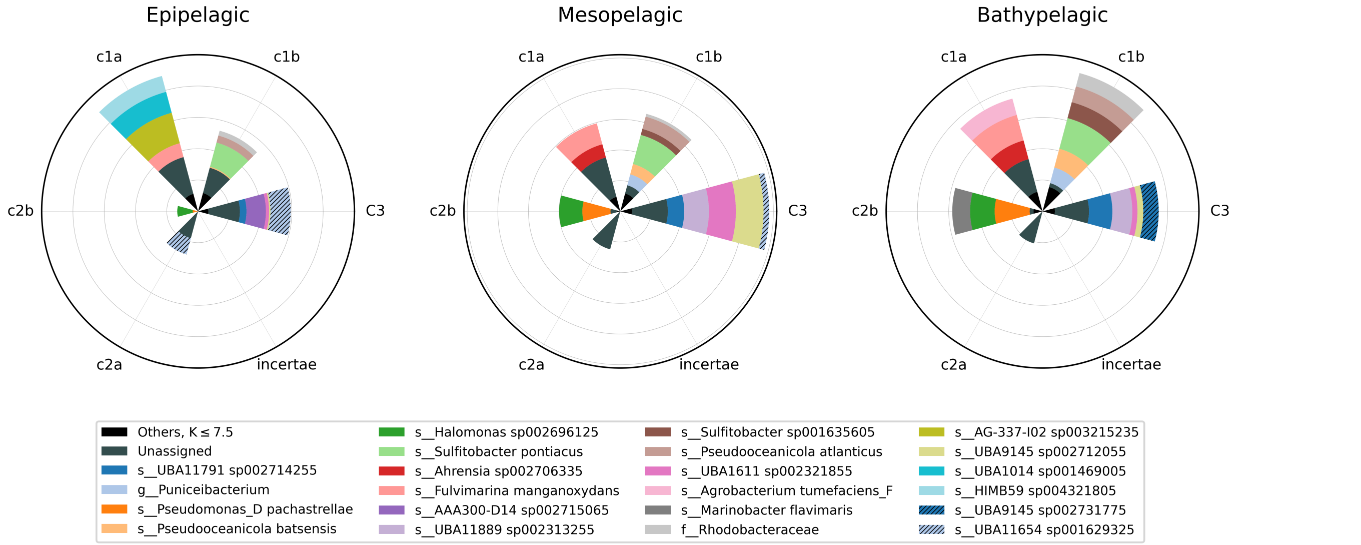

Figure 7. Patterns of putrescine-like polyamines uptake guild. This is a logarithmic representation of the top contributions to the k value for the taxonomic level of *Species*, assigned by GTDB. Here we observe only the largest contributors, whose k exceeds an adjustable threshold value, which in this case is $k = 7.5$. It is found that the contribution to the function fluctuates in both taxonomic identity and implementation preference, and that it is not an obvious relationship with depth. For example, the taxon UBA11654 sp001629325 (striped blue) contributes in the epipelagic with two different implementations, cIII and cIIa, it only contributes through cIII in the mesopelagic, while disappearing in the bathypelagic. Note how easy it is to observe distinct functional trends for each taxon, even in this particular case where the size of the input is unmanageable with traditional approaches. The *incertae* implementation is representing the absence of k values in the undefined sequence spaces of the reference tree.

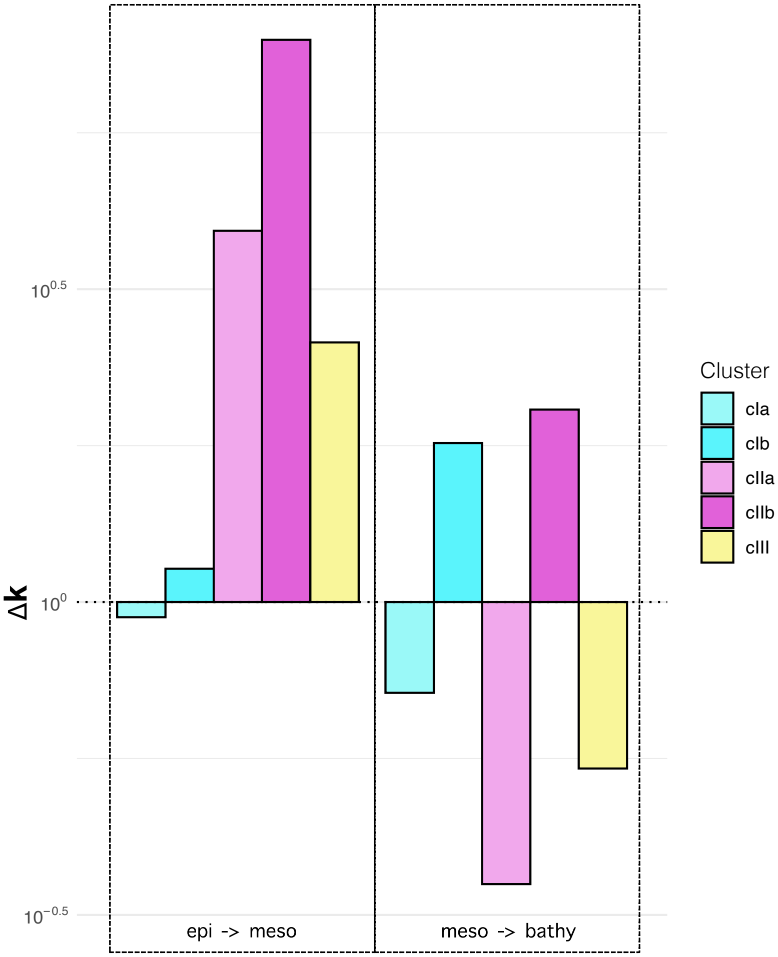

Figure 8. Changes with depth in the importance of the polyamine uptake guild. An advantage of using the K_f tensor to determine the structure of a guild is that we can visualize the functional contribution in a variety of ways. For example, here we look at the fold changes in the contribution between different ocean layers at the implementation level. It is easily observed which implementations depend the most on depth, which in this case are $cIIa$ and $cIIb$, sequence spaces putatively adapted to a wide pH range. Coincidentally, the acute changes in these two implementations correspond with the area of the water column with the largest shift in pH toward acidity in the oxycline (Sup. Fig. G).