PREDICTION OF ECOLOGICAL FUNCTION IN THE MICROBIOME USING MACHINE LEARNING ON THE GRAPH SPECTRA OF COEVOLVING SUBNETWORKS

Russell Y. Neches
Matthew D. McGee
Peter C. Wainwright
Jonathan A. Eisen

December 11, 2017

CONTENTS

1 ABSTRACT

Diversification is the process by which a single lineage evolves and branches into a group of distinct new lineages. It is rarely the case that diversification occurs in iso-

lation; evolving lineages interact with one another with varying intensity over time. This process is called codiversification or cospeciation when the terminal lineages are treated as separate species. When the the history of a diversifying lineage is reconstructed, it is usually represented as a tree; when the histories of interacting lineages are inferred, the interactions connect branches of the trees to one another. Such phylogenies resemble "tangled trees." Most existing methods for studying interacting phylogenies fall into two categories; either they test for codiversification by fitting data to an idealized model, or in cases where codiversification is known to have occurred, they endeavour to predict the most likely history. Here, we present a comparative approach to identifying and classifying codiversification events. Using the graph spectra of networks of interacting organisms, we construct a similarity metric on the topology of their interacting phylogenies that permits the use of established clustering and machine learning techniques to classify interactions that exhibit similar phylogenetic topology and patterns of ecological interaction. This is valuable for studying host-microbe interactions in microbiomes. For many groups of host-associated microorganisms, the pattern of interaction and phylogenies of the interacting organisms are known but the nature of the ecological relationship is not. Using this method, we are able to propose ecological relationships for host-associated bacterial clades based on their structural similarity to interactions with known ecological relationships.

2 INTRODUCTION

Ecology is the study of the relationships among organisms and between organisms and their environment. When making observations in a macroscopic system, a trained observer will find a vast amount of information in plain sight. Even if the scope of the observational protocol is defined narrowly, other potentially important information is nevertheless available. Behavior will be displayed. Aspects of life cycles may appear. Abiotic properties of the habitat, such as elevation and physical structure, will be known. Ecological connections with other organisms outside the scope of the protocol may present themselves. These pieces of data provide the observer with a great deal of context about the system under examination, and may help the identify weaknesses in the observational protocol, in the overall study design or in the hypothesis itself. Often, such contextual information only finds its way into rigorous experimental design by helping to frame and focus the hypothesis.

While it is well established that microorganisms exist in complex relationships with their environment and with one another, for the overwhelming majority of microorganisms known to science, our only direct knowledge are the sequences of marker genes, from which one may posit where the organisms fall in the tree of life. As with macroscopic communities, one can also make non-targeted observations of microbial communities using microscopy and environmental sequencing. However, one often cannot cross-reference or link these observations to one another in the intuitive way that is usually possible in macroscopic communities. In a macroscopic community, the link between molecular data and observations of morphology, behavior and life history is often an inevitable consequence of the need to physically collect the sample from individual specimens. In microbial communities, establishing such links is painstaking, deliberate work, and is often impractical.

Even so, there are other complications that tend to be much more pronounced in microscopic communities. Metabolically intimate relationships are often separated over vast distances and long intervals, and mediated by substances that cannot be observed optically and require sensitive, targeted chemical analysis to measure. Convergent evolution often drives distantly related organisms to present indistinguishable morphology, and closely related organism frequently present highly di-

vergent morphology. The great diversity of microorganisms routinely outstrips our ability to collect, organize and interpret meaningfully specific metadata.

The result is that knowledge of microbial communities tends to be much more fragmented than macroscopic communities, even when the total amount known is roughly the same. Key insights into microbial ecology are often obscured by methodological ambiguities that confound cross-referencing different categories of observational data.

Coevolution, the reciprocal evolutionary responses arising from interactions among (at least) two distinct populations, can serve as a point of leverage for building a more integrated understanding of ecological communities. Different types of ecological interactions impose distinct pressures on the evolution of the organisms they involve, which may leave patterns in the structures of the phylogenies of the interacting organisms that can be identified if the right conditions hold. [1, 2, 3] Of course, not all interactions can be inferred from the effect they have on the evolution of the organisms involved. As Janzen argued, coevolution is not synonymous with 'interaction' or 'symbiosis' or 'mutualism,' and applies only to the subset of interactions that drive reciprocal evolutionary responses. [4] In order for the patterns created by evolutionary responses to be observable, the interaction must persist for long enough to produce evolutionary responses, and those responses must be large enough to be picked up by the phylogenetic markers and methods applied.[1]

Although not all interactions that have a significant impact on the fitness of an organism lead to coevolution, coevolution is always the outcome of an ecological process that has a significant impact on the fitness of an organism. Organisms in predator-prey, host-parasite and other antagonistic interactions evolve with Red Queen dynamics [5, 1], where reciprocal adaptive responses drive cospeciation. Mutualistic interactions tend to obey different dynamics. [6] The Red King model, [2, 3] for example, predicts that in mutualistic interactions, advantages tend to accrue to slower-evolving organisms. This has the effect of reinforcing the development of complex, highly nested interactions. [7, 8] This is by no means an exhaustive list of types of interactions or the evolutionary dynamics they drive.

If ecology can drive evolutionary processes, then it ought to be possible in some cases to infer the nature of ecological interactions from the signature they imprint on the evolution of the organisms involved. Such inferences are potentially of great value for understanding microbial communities, where marker gene sequences and place of origin are usually the only pieces of data available for the bulk of observed biodiversity.

Most methods for predicting ecological function rely on correlations with a putative mechanism which is either observed directly (as in macroscopic systems) or inferred through proxies such as small molecules, [9, 10, 11] annotated gene functions, [12, 13] gene regulatory reponses, [14, 15, 16] or protein sequence profiles. [17] Among macroscopic organisms, identification of the mechanism of an interaction is usually carried out from visual observations. For example, a wasp using an ovipositor to lay eggs inside the body of a caterpillar is unlikely to be anything but a parasitic interaction. For interactions involving microbial organisms, one might infer that an interaction is parasitic through the observation of a molecular mechanism, such as a Type III secretion system. Type III secretion systems play a key role in the pathogenic behavior of several organisms, and one may reason that a homologous system has similar function in a new context, and so when the presence of a Type III secretion system is inferred from its DNA, RNA or protein sequence, or if its activity is inferred from transcription levels of its component genes, this may indicate that one is observing a parasitic interaction. Here, we describe the development and testing of a novel approach, which we call TangleSpace, that is agnostic to the mechanism, and instead relies on observing the evolutionary effect of occupying an ecological niche.

1 For a more thorough but less formal exploration of the uncertainties that come into play when linking phylogenies to other processes and to other phylogenies, please see the Main Introduction.

We conducted a survey of the literature to gather examples of interactions among groups of organisms, and labeled them according to the type of interaction (parasitism or mutualism). We were able to further subdivide mutualistic interactions into pollination and frugivory (more subdivisions are possible with more data). For each of these interactions, we obtained phylogenetic trees of the interacting organisms, or built them when they were not available. We also simulated interactions with no co-evolution and with perfect co-evolution to serve as null hypotheses. Using this database of interactions with known, labeled ecologies, we use machine learning to make predictions about the ecological roles of microbial clades in our system. This approach will work for any system where the microbial community is distributed among several host species, but it may be easily generalized to any system where dispersal through the system and the phylogenetic divergence exhibit significant, detectable variation on similar time scales.

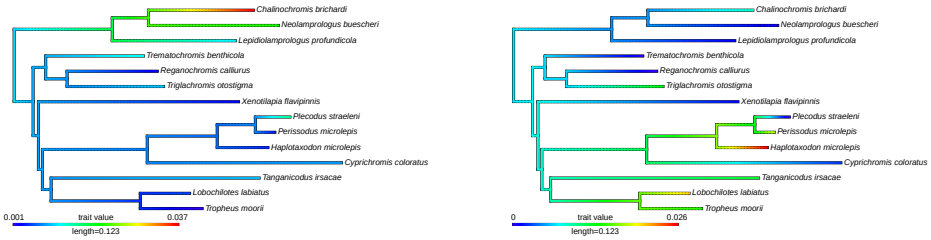
The habitat we have chosen to illustrate this method is a group of 14 species of Tanganyikan cichlid fish. With the exception of *Tylochromis polylepis* and *Oreochromis tanganicae*, the Tanganyikan cichlids are a species flock of more than two hundred described species sharing a common ancestor dating to the formation of the lake between ten and twenty million years ago. While some members of this species flock are older than hominids, other lineages have undergone much more recent speciation, including some in which the process of adaptive radiation appears to be ongoing. Within this species flock, a large variety of reproductive, defense and trophic strategies have evolved. Other than the anoxic abyssal region, cichlids have colonized every habitat in the lake and occupy every trophic level above primary producers. Taken together, this species flock represents roughly 0.5% of all described fish species and 0.25% of all vertebrate species. To interrogate the microbiomes of specimens of these 14 representative species, we performed 16S rRNA gene sequencing of stool samples and built a phylogenetic tree of all unique, non-chimeric sequences. By linking this tree with the host phylogeny through the presence/absence matrix of observed sequences, we created a "tanglegram" representing all putative coevolutionary events between the hosts and their associated organisms, as represented by their rRNA gene sequences. Because the host tree represents organisms that emerged much more recently than many of the deeper branches in the microbial rRNA tree, we searched for coevolutionary events by examining interactions with sub-clades of the rRNA tree.

3 BACKGROUND

Most of the literature on the topic addresses coevolution in one of two ways, depending on the initial assumptions. One may begin without the assumption coevolution has occurred, and then ascertain the likelihood that it did. Or, having established that coevolution has occurred either by testing for it or through some other piece of exogenous evidence, one may ascertain what history is most likely. The first approach frames the question as, *How likely is it that these two interacting groups have coevolved?* The second frames the question as, *Given that these two interacting groups have coevolved, what is the most likely sequence of events?* The first approach yields an overall likelihood, and the second yields a sequence of events (host switches, extinctions and bifurcations) that may have resulted in the observed associations. These two approaches are covered in sections 3.1 and 3.2, respectively.

We are interested in learning about the natural history of the system, and so we would like to ask a somewhat enlarged question. *In a system where the hosts are associated with a very large number of related organisms, which clades show evidence of coevolution with the hosts?* We will focus on this approach, which is summarized in section 3.3.

For each organism that appears among the microbiomes of a group of hosts, there exists a pattern of observed associations between the "guest" organisms and



(a) CHALBRI1_1198 (b) CHALBRI1_4931

OTU name	Tax. assignment	Tax. level	% id.	Rel. abund.	Bloomberg's K	Pagel's λ	Moran's Î	Abouheif's C_{mean}
CHALBRI1_1198	*Aeromonas*	genus	99%	0.14%	0.792551	0.640501	0.330350	0.394249
CHALBRI1_4931	Enterobacteriaceae	family	100%	0.12%	0.425427	0.000047	0.005716	0.046999

(c)

Figure 1: Two examples of bacterial associations examined as traits of their hosts, one with high phylogenetic signal **(a)** and one with low phylogenetic signal **(b)** (trait values in the tree legends are expressed in parts per hundred). However, the taxonomic assignments suggest that the second OTU may represent a more diverse group of organisms. If more than one of these organisms is present, then it is likely that each would have a different pattern of relationships with the host organisms. The sum of these patterns is likely to have a lower phylogenetic signal than any particular one of them. In principle, one could compensate for the different levels of diversity represented by different 16S sequences, but any such approach would require a relatively complete and unbiased database of microbial diversity. Unfortunately, such a database does not yet exist.

the host organisms. These associations can be treated as traits of the host, and the history of each association can be examined like any phenotypic trait of the host. Alternatively, the tendency to associate (or not) with a host could be treated as a trait distributed over a group of microbial organisms, although it would require dividing the microbiome into groups of related OTUs in order to be meaningful. Association with a particular host is more likely to correspond to different traits for more distantly related OTUs, and so it is probably not meaningful to naively apply stochastic trait mapping across all bacterial diversity.

3.1 Tests for phylogenetic signal

By treating OTUs, or groups of OTUs, as traits of the hosts, one can apply a variety of methods for estimating how likely it is that a trait is both conserved and vertically transmitted. These measures of "phylogenetic signal," attempt to characterize the statistical independence (or autocorrelation) among species traits due to phylogenetic relatedness. The less the trait distribution can be said to be statistically independent from the phylogeny, the more phylogenetic signal is present in that trait. [18] Münkemüller *et al.* [19] have prepared an excellent review covering the theory and application of Abouheif's C_{mean}, Pagel's λ, Moran's I, Blomberg's K and some of their variations. Unfortunately, trait-based models do not account for interactions *among* traits, or a way to account for traits that have their own evolutionary model (i.e., traits that are themselves organisms). For an interaction of a particular host and microbe, this might not present a problem. For examining interactions of microbiomes (or a significant subset of one), the assumption that traits are distributed independently is a significant weakness.

Individual OTUs may represent several distinct organisms, each with their own pattern of association. Figure 1) illustrates an example from the Cichlids dataset where this is likely the case. The taxonomic assignments suggest that the second OTU (CHALBRI1_4931) may represent a more diverse group of organisms than the first (CHALBRI1_1198). If more than one organism belonging to Enterobacteriaceae

is present, then it is likely that each would have a different pattern of relationships with the host organisms. The sum of these patterns is likely to have a lower phylogenetic signal than any particular one of them, and could account for the lower phylogenetic signal observed for this OTU. In principle, one could compensate for the different levels of diversity represented by different 16S sequences, but any such approach would require a relatively complete and unbiased database of microbial diversity. Unfortunately, such a database does not yet exist.

3.2 Permutation tests for cospeciation

Where phylogeneitic signal estimates the non-independence of a trait from the phylogeny of the organisms among which it is distributed, other methods exist for estimating the non-independence of interacting phylogneies from one another. For example, Hafner *et al.* [20] examine the relationship between pocket gophers and their chewing louse parasites using a tree reconciliation test implemented in `COMPONENT` by Roderic D. M. Page [21]. This is, in essence, a parsimony approach; the gene duplication and loss events necessary to achieve topological congruence between two trees are minimized. Hafner *et al.* first reject the null hypothesis (independent evolution) by applying the tree reconciliation test among gopher, louse and randomly drawn trees, and then examine rates of nucleotide divergence between the gopher and louse. This dataset has since been reanalyzed in many other publications, has become a sort of base-case for co-diversification methods. Notably, Huelsenbeck *et al.* use the gopher/louse data set to introduce the use of Bayesian inference to cospeciation. [22] In the context of a microbiome where many nested interactions must be examined, parsimony methods are inapplicable due to a lack of statistical consistency, [23] and Bayesian methods due to their large computational demands.

Hommola *et al.* [24] describe a method to extend the Mantel test. [25] The Mantel test is a statistic on the correlation of two matrixes, and requires that the matrixes be of equal rank. When applied to distance matrixes for trees containing differing numbers of taxa, one must delete or duplicate taxa until the matrix ranks are equal. This results in anomalous results which Hommola *et al.* explore in detail. To address this issue, they take every pair of linked tips (e.g., a parasite species that is observed to associate with a host species), and examine the correlation of distances through the two trees. This accommodates trees of arbitrary size and arbitrary patterns of association among their tips, and can be coupled with a straightforward permutation test to estimate the significance of the correlations. One can walk through every clade in the phylogeny of host-associated OTUs and, for every clade, compute the Hommola correlation with the host phylogeny (Figure 2).

However, there are some important difficulties when it comes to applying the Hommola cospeciation test to a large number of clades belonging to the same tree. First of all, this approach necessarily entails multiple comparisons, and so a correction to the significance values is required. However, a Bonferroni correction is not appropriate because the structure of each clade is not independent of the others. Rather, they are hierarchically nested within a tree, itself inferred from an probabilistic model (an approximate maximum likelihood model, in this case) based on nucleotide transitions inferred from an alignment. The autocorrelation within the tree would need to be accounted for, but calculating it would not be straightforward. At a minimum, one would need to take into account the fact that the multiple comparisons have hierarchical relationships.

The second difficulty arises from the need to perform unsupervised correlation tests. Hommola *et al.* use the Pearson product-moment correlation, which assumes that the processes are independent, identically distributed, and follow a bivariate normal distribution. Unfortunately, there is reason to suppose that the distances among tips of a phylogenetic tree would be normally distributed. Of course, other correlation statistics could be substituted at the expense of computational complexity (Spearman's rank correlation coefficient, for example, obeys quadratic scaling),

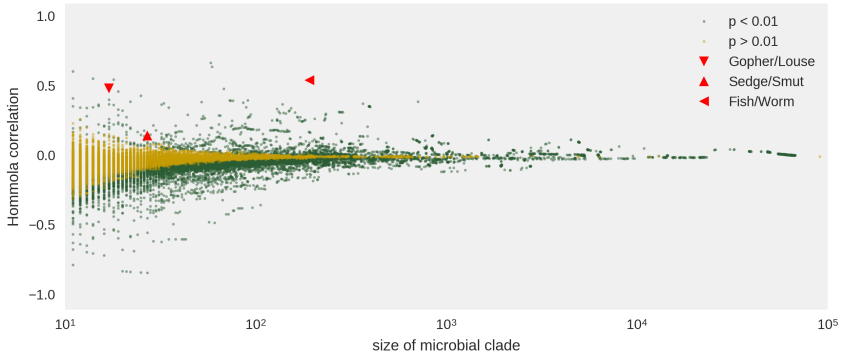


Figure 2: The Hommola correlation coefficient of the 70,343 host-associated microbial clades with more than five members. The average clade contains 228 host-associated OTUs, and the largest contains 90,858 OTUs. Statistically significant correlations (those with p values less than 0.01) are shown in blue, and non-significant correlations are shown in gray. For comparison, two reference case are included; the pocket gopher and chewing louse system from Hafner *et al.* ($r = 0.49$, $p = 1.38 \times 10^{-9}$, 17 host-associated taxa) and the sedge grasses and smut fungi system from Escudero ($r = 0.15$, $p = 1.27 \times 10^{-5}$, 27 host-associated taxa). [20, 26] Of thse, 343 clades show statistically significant Hommola correlations exceeding 0.1.

but the number of OTUs present in a typical microbiome would call for the use of a supercomputer.

The reliance on a summary statistic in Hommola *et al* makes detecting co-diversification in the microbiome difficlt. All summary statistics are a form of information reduction, usually a projection into a 1-dimensional space, and their interpretation depends on structural properties within the data. While it is often possible to construct complimentary tests to insure that assumptions hold, those tests are themselves likely to rely on summary statistics. Ultimately, someone has to actually inspect the data to make sure that the correlations are meaningful. [27] For example, compare the distribution of pairwise distances for the Gopher/Louse dataset [20] to Clade 72223, which appears just below it in Figure 3. Both have about the same correlation coefficient ($r = 0.490$ verses $r = 0.488$), very high significance (1.38×10^{-9} verses $p = 3.21 \times 10^{-223}$ and are within an order of magnitude in size (15×17 verses 14×68). Nevertheless, the two distributions are obviously different, and the vertical banding pattern in Clade 72223 is strongly reminiscent of Case 4 in

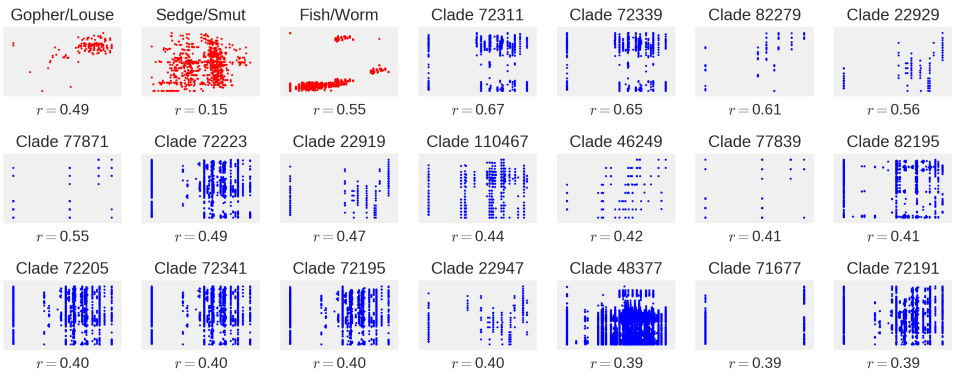
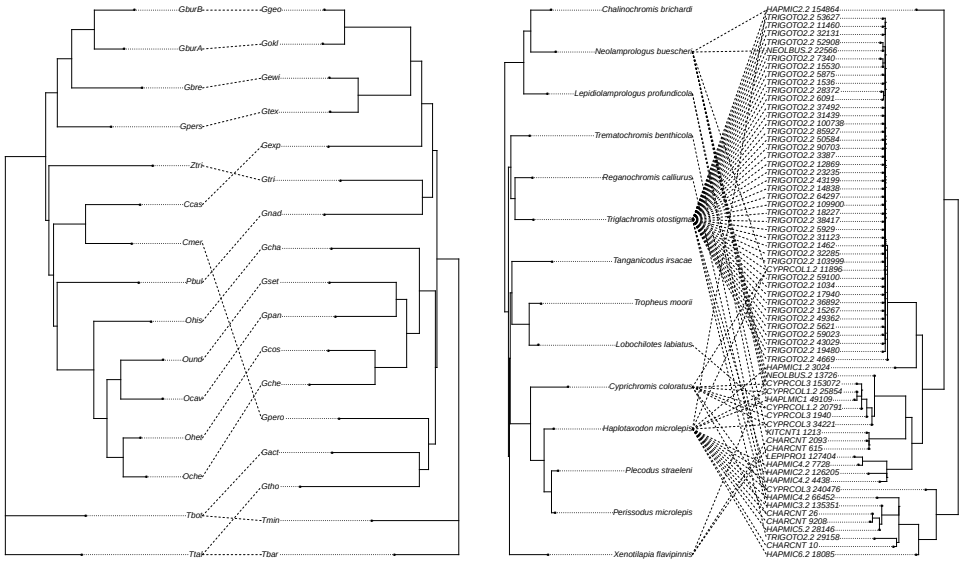


Figure 3: The distribution of linked distances for the top nineteen clades ranked by Hommola correlation coefficient (blue) and the Gopher/Louse case and Sedge/Smut case (red). Most distributions violate the assumptions of the Pearson correlation statistic, and are not likely to be meaningful.



(a) Gopher/Louse **(b)** Clade 72223

Figure 4: Tanglegrams illustrating the interactions among pocket gophers and their chewing lice parasites from Hafner *et al.* [20] **(a)**, and Clade 72223, a group of OTUs associated with Tanganyikan cichlid fishes **(b)**.

Anscombe's quartet. The structural differences between these interactions can also be seen clearly when represented as tanglegrams (Figure 4).

In principle, one could develop a hierarchical Bonferroni correction, design correlation tests that exclude artifactual features that crop up in distributions of patristic distances, and obtain the use of a powerful supercomputer. However, a third difficulty remains. All of the approaches mentioned so far are based upon an idealized model of coevolution. In the ideal case, the host and the guest organisms would diverge in lockstep and exhibit phylogenies of perfectly congruent structure. One might imagine this to be the case for vertically transmitted symbionts, but the reality is that it is not even true for different genes within the same organism (for example, due to incomplete lineage sorting). Coevolution can be expected to manifest alongside other effects. For example, a Red Queen interaction may "leak" lineages that escape from the reciprocal selective effects. A Red Queen interaction may emerge from a non-coevolutionary interaction, producing a embedded coevolution event. Organisms outside a coevolutionary interaction may impinge on the process, as may abiotic factors.

The model, and the statistical tests applied to it, must be sensitive enough to detect coevolution from a background of other process and selective enough to discriminate real cases of coevolution from patterns exhibiting spurious or artifactual resemblance to coevolution.

As mentioned before, the Gopher/Louse dataset from Hafner *et. al.* exhibits a Hommola correlation of $r = 0.49$, which is rather poor as correlated processes go. Nevertheless, the ecological case for cospeciation is solid and its physiological basis is sound. The Sedge/Smut interaction from Escudero [26] is also well established, but has a Hommola correlation of only $r = 0.15$. By itself, the Hommola test is neither sensitive nor selective enough to identify coevolution in microbiomes. However, it does measure an informative structural property of interacting phylogenies. Viewed in a broader context, it is a valuable feature to include in a more generalized framework for classifying these interactions.

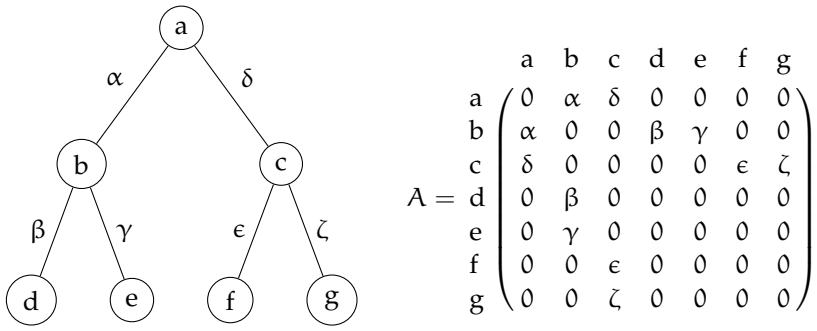


Figure 5: Construction of the graph adjacency matrix for a phylogenetic tree.

3.3 Laplacian spectral density as a metric on phylogenetic topology

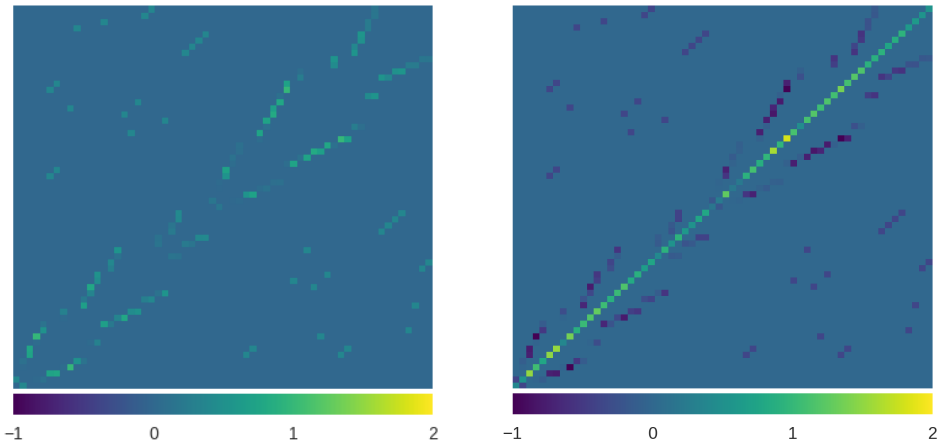
All four tests for phylogenetic signal are metrics for estimating the autocorrelation of a trait distribution with respect to the topology of a phylogenetic tree. The Hommola correlation test is a metric for estimating the autocorrelation of two trees. All five methods essentially sample the statistical *effect* of tree topology. There are mathematical approaches for interrogating the topology of trees, and graphs in general, that can make use of a greater portion of information present. One approach is examine moments of the spectral density distribution the graph Laplacian. The graph Laplacian is constructed by subtracting the graph adjacency matrix (Fig. 5) from the degree matrix (the total connectivity of each node). This representation of a general graph is often useful in graph analysis. For example, it can be used to calculate the number of spanning trees for the graph using Kirchhoff's theorem, and its second second smallest eigenvalue can be used to approximate to the sparsest cut through a given graph via Cheeger's inequality.

The eigenvalues of the graph Laplacian, sometimes called the Laplacian spectrum, correspond to the frequency with which a random walker would visit each node in steady state (i.e., in the limit of the number of random steps as the number of steps approaches infinity). It is important to note that the Laplacian spectrum of a graph is not unique. Graphs that have the same spectrum but are not isomorphic (i.e., do not have the same structure) are said to be isospectral or cospectral. For trees of intermediate size ($\approx 5 < n < n \to \infty$), the probability that two randomly selected trees will share the same spectrum is finite but negligible. [28] Most phylogenetic trees that express useful information fall within this range, and so the Laplacian spectrum can serve as a good description of a trees's general and local topology.

To compare the topology of two trees that have different numbers of nodes, it is necessary to do something about the fact that they will have different numbers of eigenvalues. Matsen and Evans [28] solve this problem by projecting the eigenvalues into a continuous 1-dimensional space using a kernel density estimator (KDE). The distributions produced by the KDE can then be compared directly. Lewitus and Morlon [29] use the Shannon-Jensen divergence between distributions, as well as several moments of the distributions, to estimate the topological dissimilarity among trees.

This suggests a spectral approach to coevolution problems. It is relatively straightforward to obtain the Shannon-Jensen divergence between a the phylogeny of a group of hosts and their parasites. However, interpretation of the result is not so straightforward. How much divergence is required to disqualify a pair of trees from being "similar" in topology? How sensitive is the divergence to tree size? How does one establish confidence intervals on this metric? Moreover, the divergence between the two trees does not take into account the way that the organisms in the trees interact with one another.

To address these questions, we propose four extensions to this approach. First, it is necessary to construct a generalized graph representing *both* trees. Second, the



(a) Graph adjacency matrix

(b) Graph adjacency matrix

Figure 6: The graph adjacency **(a)** and Laplacian **(b)** matrixes of the interactions among pocket gophers and their chewing lice parasites from Hafner *et al.* [20].

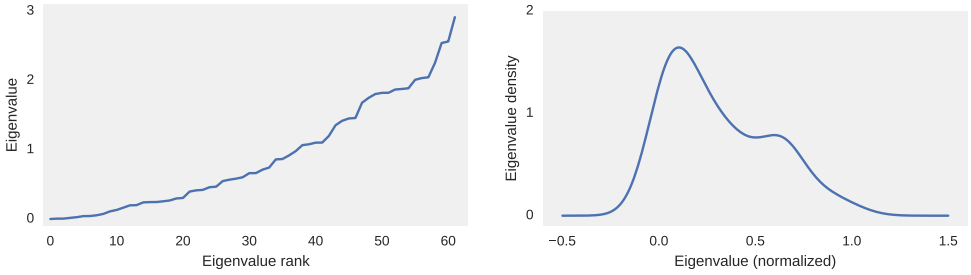
ecological interactions between the organisms represented in the trees are incorporated into the graph. Third, the phylogenetic distances in the trees are normalized with respect to one another and with respect to other interactions; we do this by setting the total distance in each tree from root to deepest tip to unity, and the links between the trees to the average of all branch lengths in each tree. Fourth, instead of looking for endogenous evidence of coevolution within an interaction, we compare interactions with unknown ecology to those with known ecology. By using a comparative approach, contextual information that would be difficult to obtain (and would likely result in overfitting) in a model-based approach can be integrated and tested as a supervised machine learning problem.

It is worth noting that almost all trees are cospectral, [30] whereas only a subset of generalized graphs are cospectral (uniform star graphs and complete graphs, for example). Spectral classification of graphs should thus suffer a lower rate of collisions in spectral space than trees alone.

4 MATERIALS & METHODS

Here, we propose an comparative method for co-diversifying systems that extends an approach developed by Lewitus and Morlon [29] for use with phylogenetic trees. The graph adjacency matrixes for each phylogenetic tree are constructed, as per Lewitus and Morlon. The total phylogenetic distance for each tree is then normalized to unity. Then, the adjacency matrixes for the two trees are joined along a common diagonal, creating two empty rectangular blocks symmetric about the diagonal. These rectangular blocks are where any links between the two joined graphs must exist, and so the interactions between the leaf nodes in the phylogeny are placed here, weighted at the mean branch length of the two trees. The graph Laplacian is constructed (Figure 6), the eigenvalues are computed and a spectral density distribution (Figure 7) is computed using a kernel density estimator with a Gaussian kernel, as in Lewitus and Morlon.

The eigenvalues of a network's Laplacian corresponds to the frequency with which a random walker would visit each node in steady state (i.e., in the limit of the number of random steps as the number of steps approaches infinity). In a network with edge weights (or a tree with branch lengths), the probability distribution of random steps made by the walker from each node is partitioned by the edge



(a) Eigenvalues of graph Laplacian

(b) Spectral density of eigenvalues

Figure 7: Eigenvalues in rank order of the graph Laplacian matrix of interactions among pocket gophers and their chewing lice parasites from Hafner *et al.* [20] **(a)** and the normalized density distribution of eigenvalues computed with a kernel density estimator using a Gaussian bandwidth of 0.4 **(b)**.

weights (or branch lengths) of each edge connected to that node. It is a measure of the relative connectivity of each node in the network.

The eigenvalues of a network comprise its spectrum. A network's spectrum is a not perfectly unique; very small networks with different topologies may share the same eigenvalues. The probability of one tree being a superset of another approaches unity in the limit of very large trees. For trees of intermediate size ($\approx 5 < n < n \to \infty$), the probability that two randomly selected trees will share the same spectrum is finite but negligible. [28] For networks constructed from two interconnected trees, individual internal nodes of the trees will have the same connectivity regardless of whether they are in a network or in a tree. Thus, the spectrum of such a network will be composed of eigenvalues that correspond to those of the internal nodes of each of the two trees, and a third group of eigenvalues corresponding to the leafs and their interactions. There are more possible configurations for networks of this general topology than for trees of a given number of nodes, and thus a larger number of spectra are possible for a network than for a tree.

Spectra are a discrete set of values equal to the number of nodes in the network, and so comparing the structure of networks with unequal sizes requires some additional transformation. As with Lewitus and Morlon, and Matsen and Evans [29, 28] work with trees, we map the Laplacian spectra into a continuous, unit space by applying a Gaussian kernel density estimator, yielding a continuous distribution function for each spectra. The dissimilarity between two distribution can be measured using the Kullback-Leibler divergence, D_{KL}.

$$D_{KL}(p, q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx \tag{1}$$

The Kullback-Leibler divergence measures the information lost when using distribution $q(x)$ to approximate distribution $p(x)$. This is almost what we want, but unfortunately D_{KL}, like subtraction and division, is not metric (in general, $D_{KL}(p, q) \neq D_{KL}(q, p)$ unless $p = q$, in which case $D_{KL} = 0$). However, the Jensen-Shannon divergence between the two distributions is metric.

$$D_{SJ}(p, q) = \sqrt{\frac{1}{2} D_{KL}(p, q) + \frac{1}{2} D_{KL}(q, p)} \tag{2}$$

To demonstrate how this metric performs, we compare permutations of the Gopher/Louse dataset from Hafner *et al.*. Keeping the tree structures intact, a collection of spectra were computed by randomly reassigning links between leaf nodes (Figure 8) and computing the Jensen-Shannon divergence between successively more permuted spectral distributions and the spectral distribution of the original, un-

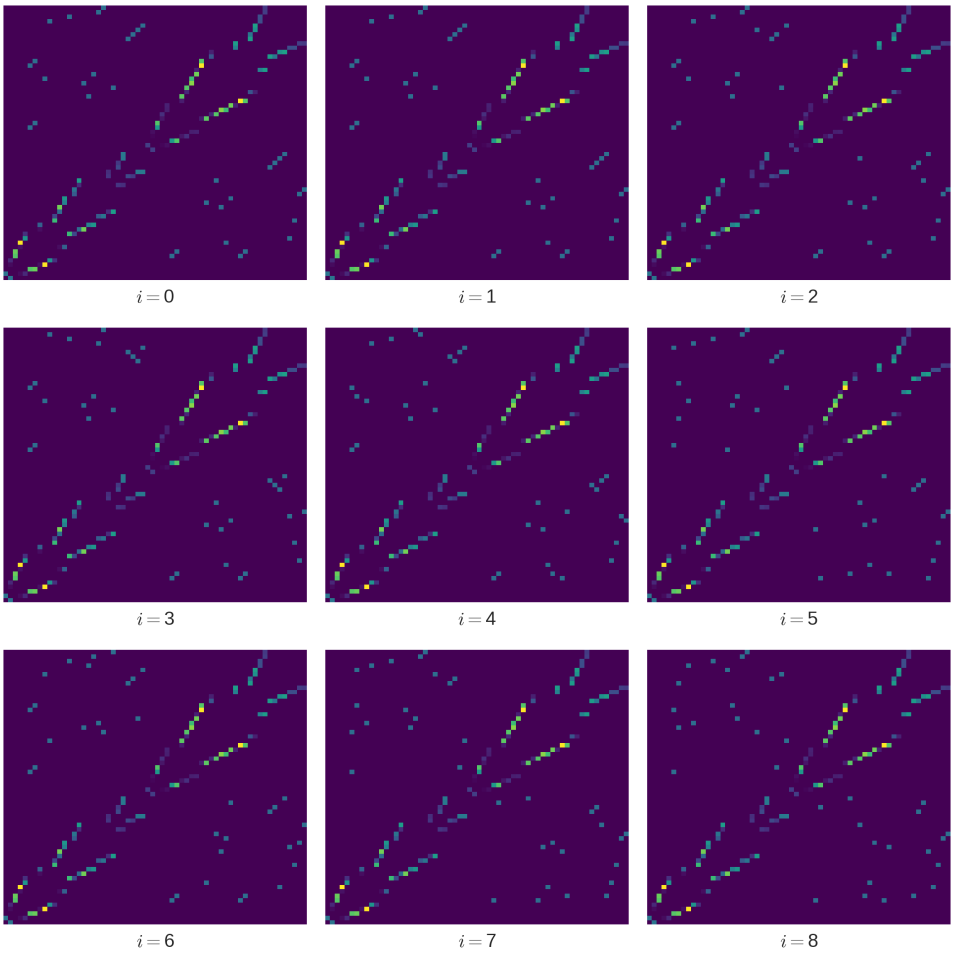


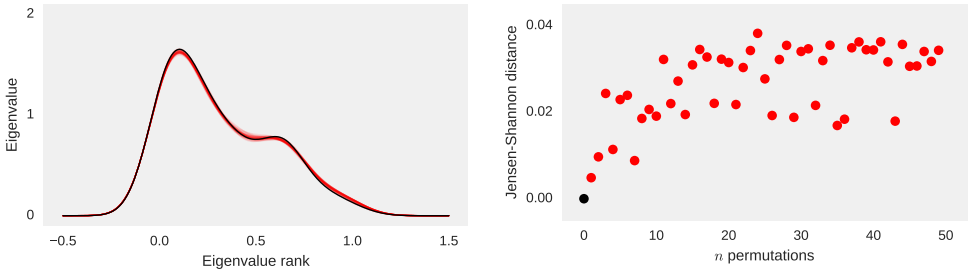
Figure 8: Successive permutations from none ($i = 0$) to eight ($i = 8$) of the graph adjacency matrix of interactions among pocket gophers and their chewing lice parasites from Hafner *et al.* [20]. Permutations are carried out by swapping the associations of randomly chosen links in between the host and parasite tree.

permuted graph (Figure 9). The metric is neither linear nor monotonic with these permutations, but it does diverge predictably.

In this way, a feature space is constructed over the Laplacian spectra of networks of interaction species with known ecologies (parasitic verses mutualistic interactions, for example). These networks are then projected into the space which they span, and a classifier is trained on their ecological labels. The spectra of interactions with unknown ecological labels are then projected into that space, and the trained classifier may then predict which ecological label is most likely for each unlabeled interaction.

For each interaction collected from the literature (which we'll denote $G_{L,i}$, there is a label for the type of ecological interaction taking place. Each spectral density distribution has a set of endogenous features (see also Table 3) :

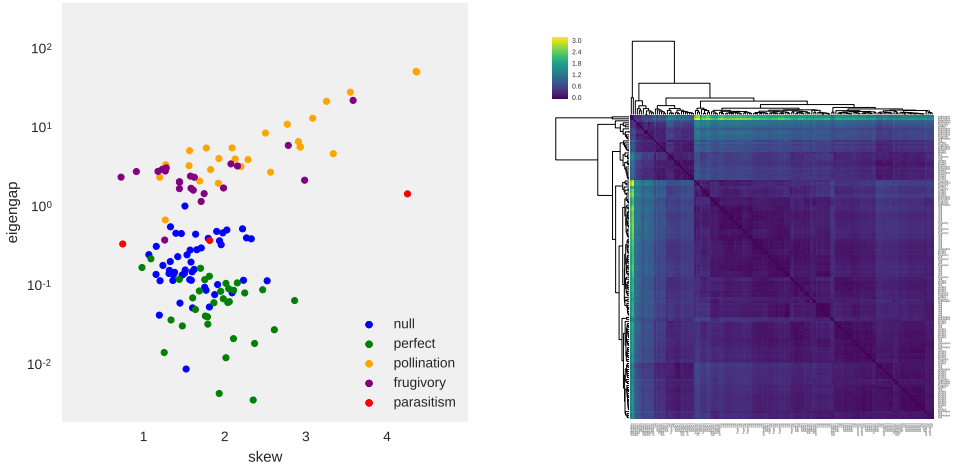
- **Links** (n_L) The number of links connecting the interacting trees.
- **Occupancy** (k) The ratio of the number of links to the number of leafs ($2n_L : n_{hosts} + n_{guests}$).
- **Squareness** (q) : The ratio of the number of leafs in each tree.
- **Eigengap** (λ_δ) The difference between the largest and second largest eigenvalues in the Laplacian spectrum.



(a) Permuted spectral densities

(b) Shannon-Jensen distances

Figure 9: Spectral density distributions of successively permuted interactions among pocket gophers and their chewing lice parasites from Hafner *et al.* [20] **(a)** and the Shannon-Jensen distances between each with respect to the spectral density distribution of the un-permuted interaction **(b)**.



(a) Skew verses eigengap

(b) Hierarchical clustering by distance

Figure 10: Spectral properties of examples of co-diversification gathered from the literature and simulated cases of independent and parallel diversification. The spectral distributions of these graphs can be separated into different groups either by extracting summary statistics, such as their skew and eigengap **(a)**, or by hierarchical clustering of their pairwise Shannon-Jensen distances **(b)**.

- **Kurtosis** (γ_2) The sharpness of the peak in the distribution (the fourth standardized moment).
- **Skew** (γ_1) The asymmetry of the distribution (the third standardized moment).
- **Hommola correlation** (r_H) The Hommola correlation of the interaction [24].
- **Hommola significance** (p_H) The significance of the Hommola correlation [24].
- **Tree distance** (D_t) The Jensen-Shannon divergence between the spectral density distributions of each of the two phylogenetic trees in the interactions.

These properties form an $n \times 9$ matrix of features, like so :

$$\psi = \begin{array}{c} \\ G_{L,0} \\ G_{L,1} \\ G_{L,2} \\ \vdots \\ G_{L,n} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \lambda_\delta & \gamma_1 & \gamma_2 & r_H & p_H & D_t & k & q & n_L \end{array} \\ \begin{pmatrix} \lambda_{\delta,0} & \gamma_{1,0} & \gamma_{2,0} & r_{H,0} & p_{H,0} & D_{t,0} & k_0 & q_0 & n_{L,0} \\ \lambda_{\delta,1} & \gamma_{1,1} & \gamma_{2,1} & r_{H,1} & p_{H,1} & D_{t,1} & k_1 & q_1 & n_{L,1} \\ \lambda_{\delta,2} & \gamma_{1,2} & \gamma_{2,2} & r_{H,2} & p_{H,2} & D_{t,2} & k_2 & q_2 & n_{L,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{\delta,n} & \gamma_{1,n} & \gamma_{2,n} & r_{H,n} & p_{H,n} & D_{t,n} & k_n & q_n & n_{L,n} \end{pmatrix} \end{array} \tag{3}$$

There are also the Shannon-Jensen distances between each pair of labeled interactions :

$$D_{L,L,i,j} = D_{SJ}(G_{L,i}, G_{L,j}) \tag{4}$$

These distances form an matrix of n^2 features, like so :

$$\xi_L = \begin{array}{c} \\ G_{L,0} \\ G_{L,1} \\ G_{L,2} \\ \vdots \\ G_{L,n} \end{array} \begin{array}{c} \begin{array}{ccccc} G_{L,0} & G_{L,1} & G_{L,2} & \cdots & G_{L,n} \end{array} \\ \begin{pmatrix} D_{L,L,0,0} & D_{L,L,0,1} & D_{L,L,0,2} & \cdots & D_{L,L,0,n} \\ D_{L,L,1,0} & D_{L,L,1,1} & D_{L,L,1,2} & \cdots & D_{L,L,1,n} \\ D_{L,L,2,0} & D_{L,L,2,1} & D_{L,L,2,2} & \cdots & D_{L,L,2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_{L,L,n,0} & D_{L,L,n,1} & D_{L,L,n,2} & \cdots & D_{L,L,n,n} \end{pmatrix} \end{array} \tag{5}$$

Together, these two matrixes form the feature space into which we will use to define the system.

$$\Psi_L = \left[\psi_{L,(n\times 9)} | \xi_{L,(n\times n)}\right] \tag{6}$$

Interactions with unknown ecology (unlabeled interactions) can then be placed into this space. Like the labeled interactions, the endogenous properties of the Laplacean spectra are tabulated :

$$\psi_U = \begin{array}{c} \\ G_{U,0} \\ G_{U,1} \\ G_{U,2} \\ \vdots \\ G_{U,n} \end{array} \begin{array}{c} \begin{array}{ccccccccc} \lambda_\delta & \gamma_1 & \gamma_2 & r_H & p_H & D_t & k & q & n_L \end{array} \\ \begin{pmatrix} \lambda_{\delta,0} & \gamma_{1,0} & \gamma_{2,0} & r_{H,0} & p_{H,0} & D_{t,0} & k_0 & q_0 & n_{L,0} \\ \lambda_{\delta,1} & \gamma_{1,1} & \gamma_{2,1} & r_{H,1} & p_{H,1} & D_{t,1} & k_1 & q_1 & n_{L,1} \\ \lambda_{\delta,2} & \gamma_{1,2} & \gamma_{2,2} & r_{H,2} & p_{H,2} & D_{t,2} & k_2 & q_2 & n_{L,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \lambda_{\delta,n} & \gamma_{1,n} & \gamma_{2,n} & r_{H,n} & p_{H,n} & D_{t,n} & k_n & q_n & n_{L,n} \end{pmatrix} \end{array} \tag{7}$$

Organism	NCBI Taxonomy ID	Individuals	Diet
Chalinochromis brichardi	34794	1	Algae
Xenotilapia flavipinnis	240481	1	Benthic crustaceans
Lepidiolamprologus profundicola	272719	1	Fish
Lobochilotes labiatus	28810	1	Benthic crustaceans/shrimp
Neolamprologus buescheri	329916	1	Benthic crustaceans/shrimp
Tropheus moorii	8150	1	Algae
Trematochromis benthicola	1171431	1	Shrimp
Tanganicodus irsacae	27763	1	Algae
Cyprichromis coloratus	559343	6	Zooplankton
Triglachromis otostigma	34813	6	Algae
Haplotaxodon microlepis	70786	6	Zooplankton/Fish
Plecodus straeleni	167846	1	Scales
Perissodus microlepis	32509	1	Scales
Reganochromis calliurus	70788	1	Shrimp

Table 1: NCBI Taxonomy identifier, number of individuals and diet of fishes used in this study.

For m unlabeled interactions and n labeled interactions, the Shannon-Jensen divergence between unlabeled and labeled interactions forms an $m \times n$ matrix of features.

$$\xi_U = \begin{array}{c} \\ G_{U,0} \\ G_{U,1} \\ G_{U,2} \\ \vdots \\ G_{U,m} \end{array} \begin{array}{c} \begin{array}{ccccc} G_{L,0} & G_{L,1} & G_{L,2} & \cdots & G_{L,n} \end{array} \\ \begin{pmatrix} D_{U,L,0,0} & D_{U,L,0,1} & D_{U,L,0,2} & \cdots & D_{U,L,0,n} \\ D_{U,L,1,0} & D_{U,L,1,1} & D_{U,L,1,2} & \cdots & D_{U,L,1,n} \\ D_{U,L,2,0} & D_{U,L,2,1} & D_{U,L,2,2} & \cdots & D_{U,L,2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_{U,L,m,0} & D_{U,L,m,1} & D_{U,L,m,2} & \cdots & D_{U,L,m,n} \end{pmatrix} \end{array} \tag{8}$$

Appending these matrixes yields a set of features that allows us to project each unlabeled interaction into the same feature space as the labeled interactions.

$$\Psi_U = \left[\psi_{U,(n\times 9)} | \xi_{U,(m\times n)}\right] \tag{9}$$

A neural network is then trained on Ψ_L and the corresponding ecological labels and used to predict the ecological labels for Ψ_U. [31]

4.1 Specimen collection and housing

Specimens were purchased as wild-caught individuals shipped from a single import facility (Old World Exotic Fish, Homestead, FL) via air cargo in accordance with guidelines and regulations set out by the US Fish and Wildlife Service. [32] The import facility packed the fish in individual 4-mil polyethylene bags with square-bottom bags, knotted with a large volume of air above the water. Bags were shipped in insulated foam boxes designed for air cargo shipment of live tropical fish. A direct flight was arranged from Miami International Airport (MIA) to Sacramento International Airport (SMF), and the shipment was received immediately at the Sacramento cargo terminal for transport to the laboratory by car. Specimens were housed in a vivarium at UC Davis under the supervision of M.D.M and P.C.W. Specimens were housed in single-species groups within 30-100L aquaria filtered by air-driven sponge filters.

4.2 Sample collection

After shipping, specimens were placed in tanks separated by species. Where only one member of a species was collected, they were placed in a solitary tank. Specimens were allowed to acclimatize to their tanks and observed for signs of injury or

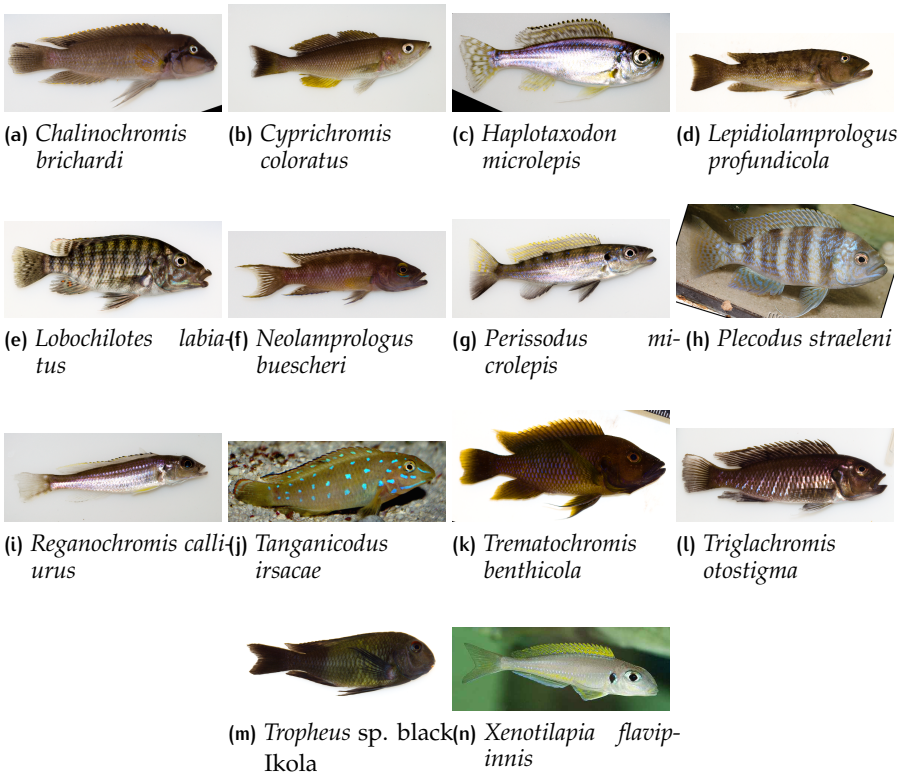


Figure 11: Live photographs of fishes used in this study (not to scale).

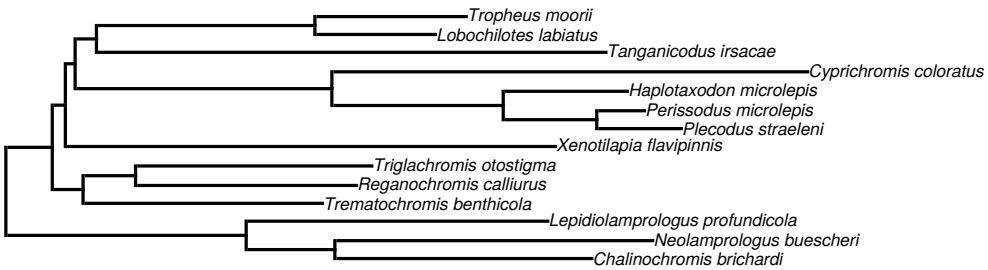


Figure 12: A maximum likelihood phylogeny of the host organisms.

Table 2: PCR cycling conditions

Temperature (°C)	Time (mm:ss)
95°C	2:00
95°C	0:15
52°C	0:30
72°C	1:30
Repeat	×10 cycles
72°C	3:00
4°C	Forever

stress. Because fish are usually starved prior to shipping to prevent stool decomposition from suffocating them in transit, feeding began as soon as was it deemed safe. All specimens were fed a common diet of kelp and *Arthrospira platensis* pellets (Omega One brand Veggie Mini Pellets) for six weeks.

For sample collection, specimens were fed until satiated and placed into 30L tanks. To minimize microbial contamination in the water and tank silicone, we lined each tank with a sterile autoclave bag prepared with 10 liters of molecular water augmented with a 5 grams each of sodium chloride and calcium chloride. To minimize potential contamination from biological filtration, we used chemical filtration via sterile charcoal pellets (Petco brand activated carbon pellets boiled in molecular water for 20 minutes) to sequester nitrogenous wastes produced by the fish, and a vinyl aquarium tube was submerged and connected to an air pump to aerate the water. For each collection, fresh tubing was flushed and wiped down with ethanol and dried before use. After feeding, specimens were transferred into the collection tanks and checked frequently until stool was observed. If no stool was observed within 24 hours, the specimen was returned to a standard vivarium tank and the experiment was repeated. After removing the specimen from the collection tank, stool was removed using a sterile serological pipette, placed in MoBio sample processing tubes and frozen.

4.3 Sample preparation, processing and sequencing

Stool samples were subjected to bead beating for 60 seconds and DNA was extracted using MoBio PowerSoil DNA Isolation kit in accordance with the manufacturer's protocol. DNA yields were measured using a Qubit Fluorometer (Invitrogen) and Quant-iT dsDNA Assay Kit, High Sensitivity (Invitrogen product no. Q33120). DNA was amplified by a two-step PCR enrichment of the V4 region of the 16S rRNA gene using "universal" primers 515F (TGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT), modified by addition of Illumina adapter and barcodes sequences (See Table 4).

PCR products were then purified using magnetic bead capture of high molecular weight DNA (Agricourt Ampure XP beads; product number A63880).

Libraries were sequenced using an Illumina MiSeq system, generating 250 base pair paired-end amplicon reads. The amplicon data was multiplexed using dual barcode combinations for each sample. A custom script (available in a GitHub repository `https://github.com/gjospin/scripts/blob/master/Demul_trim_prep.pl`) was used to assign each pair of reads to their respective samples when parsing the raw data. This script allows for one base pair difference per barcode. The paired reads were then aligned and a consensus was computed using `Trimmomatic` [33] with maximum overlap of 120 and a minimum overlap of 70 (other parameters were left as default). The custom script automatically demultiplexes the data into `fastq` files and parses the results to reformat the sequences in `fasta` format.

4.4 Building the observation table

Chimera were identified with `vsearch`, [34] and unique reads were identified using `hat-trie`. [35, 36] A table of observation counts were constructed as a `Pandas DataFrame` object, [37] and a count threshold was applied. Tables of raw counts and normalized counts were written as comma separated value files, and the corresponding sequences were written as a FASTA file. All analysis was carried out and documented in Jupyter Notebooks [38] and visualizations, including all those appearing in this manuscript, were constructed using `Matplotlib`. [39]

4.5 Building the OTU tree

The Illumina sequencing platform has a substitution error frequency of about 0.1%. [40] Trimming and filtering reads based on quality score removes about 69% of substitution errors, on average, [41] and overlap correcting paired-end reads improves this to about one substitution error in every 14 reads. [33] At the sacrifice of sensitivity, the error frequency can be reduced arbitrarily by excluding very rare sequences. For example, excluding sequences observed only once reduces the substitution error frequency to about one in every 200 sequences. Most substitution errors are likely to be corruptions of sequences already present in the data, and so phylogenetic reconstruction should place them as sister leafs of their uncorrupted sources. This confines the effect of substitution errors to the smallest topological scale, where it appears in the form of "extra" leafs that are randomly attached with the minimum branch length.

Alignment of the observed sequences is performed using `Clustal Omega`, [42, 43] and an approximate maximum likelihood phylogeny is constructed using `FastTree`. [44, 45]

4.6 Spectral analysis and machine learning

The host tree and guest tree are loaded as `SuchTree` objects, and linked together through the observation table as a `SuchLinkedTrees` object. The `SuchTree` class allows for extremely efficient traversals of large trees, enabling distance correlations to be efficiently computed. The `SuchLinkedTrees` class leverages this to compute graph adjacency and graph Laplacian matrixes of subtrees of host and guest phylogenies. Spectral decomposition and kernel densities of graph Laplacians are computed using `numpy`, and the Jensen-Shannon divergence is calculated between each pair of spectral densities using the `entropy` function in `scipy.stats`. [46] Feature tables were assembled using `Pandas` [37] and machine learning was carried out using `scikit-learn`. [31]

4.7 Correlation-based analysis

For each clade of guest organisms, the `SuchLinkedTrees.linked_distances` function is used to calculate the pairwise distances through the host and guest trees for every pair of non-null observations in the link matrix, as described by Hommola *et al.* [24] The Pierson's correlation for these distances is computed using the `pearsonr` function from `scipy.stats`. [46]

4.8 Literature search for comparative analysis

Trees and interaction matrixes were compiled from the supplementary material from Rezende *et al.*, [8] Hafner *et al.* [20] and Escudero [26] for a total of fifty data sets (Table 3). Taxonomic identifiers were reconciled between trees and interaction matrixes using Levenshtein distances [47] calculated using `fuzzywuzzy`, [48]

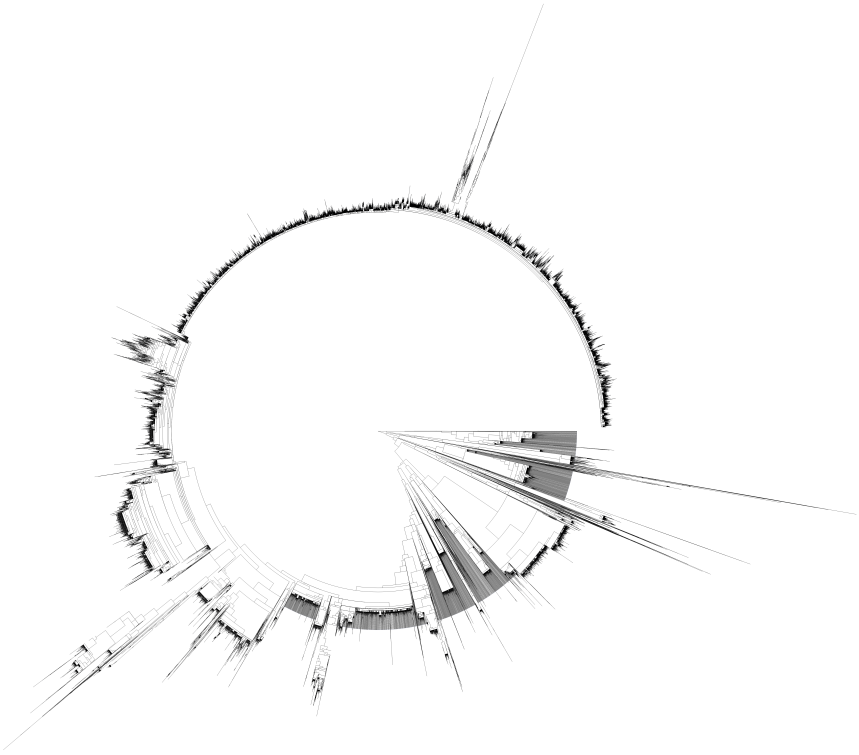


Figure 13: An approximate maximum likelihood phylogeny of the guest organisms.

and by hand where necessary. Corrected trees and matrixes are distributed with permission with `SuchTree` in Newick format and comma separated values, respectively.

4.9 Simulated datasets

Synthetic interactions with perfect phylogenetic congruence and with no phylogenetic congruence were simulated using `Dendropy`. [49] The birth and death rates of trees from the literature was estimated with the gamma statistic [50] using `Dendropy`, and random trees were generated so that the distribution of the number of taxa and their birth and death rates would match the trees from the literature. Two sets of random interactions were generated; a set of 'perfect' interactions, where the tips of two identical random trees were linked by a simple bijection, and a set of 'null' interactions, where the tips of independently generated random trees were linked with a random shuffle. The set of 'perfect' interactions simulates a coevolutionary process in which the evolution of the interacting groups proceeds in lockstep. The 'null' interactions simulate a small subset of possible ecological interactions in which no coevolution has taken place.

5 DISCUSSION AND RESULTS

Identification of ecological function of organisms or groups of organisms in microbial communities, particularly host-associated microbiomes, is a key goal for many avenues of theoretical and applied research. Broadly speaking, there are three ways to approach the question.

Ecological function can be predicted by correlating the relative or absolute abundance of microbial groups with a host response or environmental outcome. For many systems, this is the only approach available. In clinical applications, all hosts belong to a single species (*H. sapiens*) exhibiting very little genomic diversity. For this reason, correlation-based studies in humans, such as Genome Wide Association Studies (GWAS), usually require a very large sample size to achieve acceptable false discovery rates. [51] For marine communities, the cosmopolitan distribution of most organisms requires a planetary perspective on niche occupancy. [12, 52] While these approaches are powerful research tools, correlations alone cannot establish the ecological function of a organism.

Ecological function can be predicted by examining the impact of an organism on the fitness of its host. Of course, this requires both a meaningful definition of host fitness and a reliable way to measure it. [53, 54]

Ecological function can be predicted by conceptually treating a host-associated organism as a trait of the host, and examining how the "trait" assorts with the evolution of other host phenotypes. [55, 56] This phylogenetic signal approach [18, 57, 19] is very powerful, but in practice is can be confounded by the extraordinary nuance of microbial diversity. Very closely related microbial organisms can have radically divergent ecological functions – for example, *Escherichia coli* Nissle 1917 is evidently beneficial to the host, [58] while *Escherichia coli* O157:H7 is a deadly pathogen. [59] It is not always apparent where to draw the boundary between two such "traits," nor is it always straightforward to determine which side of the boundary an observation falls. This is not to say that the microbial-relationship-as-host-trait model is not useful, but rather that it is useful for exploring relationships that have a well-characterized scope with respect to the diversity of the microbial component. Its usefulness for discovering new relationships is limited.

Ecological function can be predicted by the likelihood of coevolution between a group of hosts and a group of microbes. [20, 24] This approach is limited by the

fact that coevolution can result from different and perhaps antithetical ecological scenarios. [5, 4, 2]

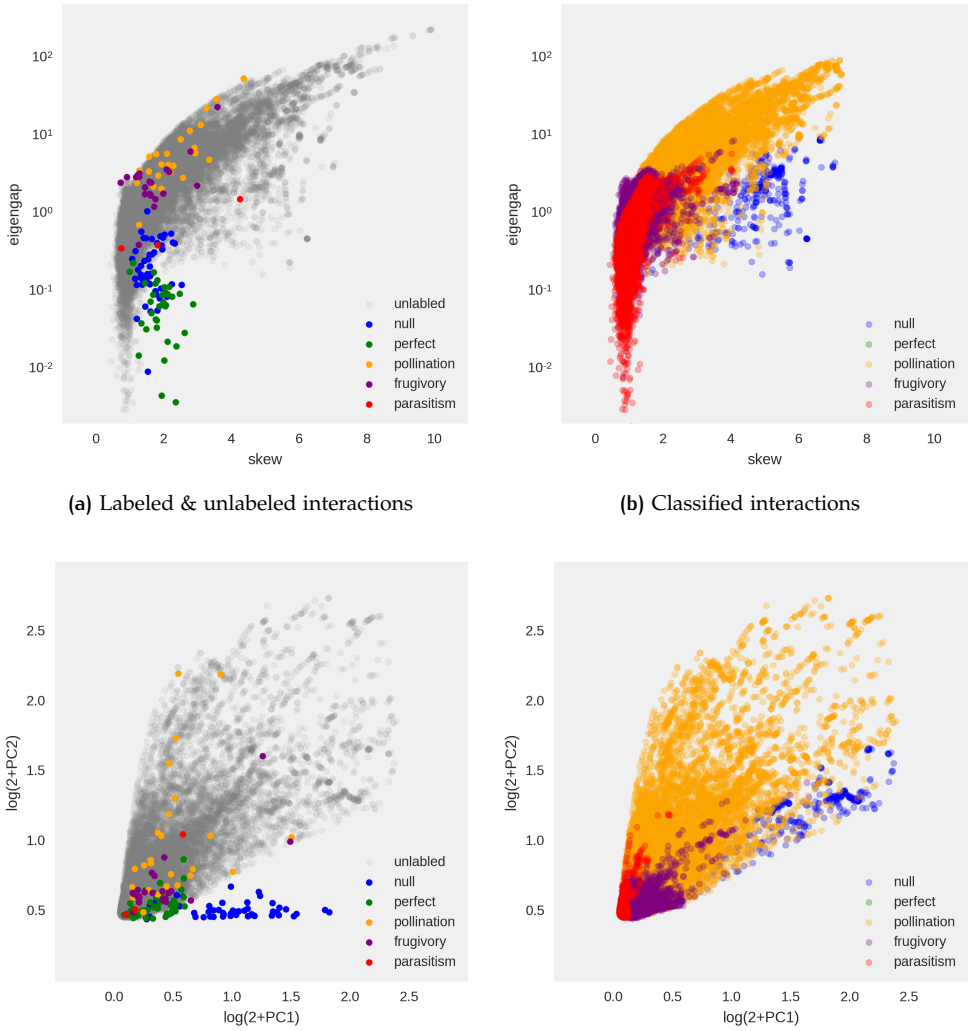
By placing unknown interactions in the microbiome into context with the evolution and ecology of characterized interactions, we overcome most of these difficulties. Because the effect of the relationship is implicit in the characterization of labeled interactions, we do not draw *post hoc* conclusions from correlations with outcomes. The effect of the relationship on host fitness is implicit in the phylogeny of both organisms. The scope of microbial diversity is explicitly addressed by the use of the microbial phylogeny. If the phylogeny fails to capture the necessary scope, the method fails gracefully, yielding an inconclusive result. Finally, by directly addressing the underlying driver of coevolution, we can make positive predictions about these mechanisms.

All model-based methods must convincingly demonstrate that they provide the appropriate selectivity, sensitivity and parameter selection for their application. This is more challenging for some applications than others, and searching for coevolution in microbiomes appears to be particularly challenging. Comparative methods exchange the statistical vulnerabilities of model-based methods for the epistemological vulnerabilities of database bias. Fortunately, the number of multi-species interactions is combinatorial with the number of species (though constrained by propinquity), and the ecological literature is rich with examples. A database drawn from the existing literature would suffer from bias, but it would not be small.

Rather than comparing candidate cases to a model, a comparative approach calls for a metric that scales with dissimilarity among cases. Measures of dissimilarity are not summary statistics (avoiding the problem of supervision), and it is possible to construct them with fewer assumptions. The construction of a feature space of topological properties of interactions and dissimilarities with respect to interactions of known ecologies makes it possible to cast the question of the ecological function as a machine learning problem. This sketches out a powerful and flexible framework for extracting inferences on the nature of many kinds of ecological interactions without direct observation of their mechanism. The cost is that one must assemble a collection of relevant training data, and it is limited to cases where the interaction has persisted long enough to leave an significant imprint in the evolutionary history of the interacting groups. It should work particularly well in cases where two or more adaptive radiations have interacted.

The training problem can be visualized by selecting a subspace spanned by two axes of the feature space and projecting the labeled training data and the unlabeled experimental data into it. Alternatively, one can project into a subspace spanned by principle components. Similarly, the predictions can be visualized by projecting the experimental data into one of these subspaces with labels corresponding to the predictions (Figure 14).

This study is limited by the small number of labeled interactions we were able to extract from the literature (51 interactions from the literature and 100 simulated interactions). With a training set of suitable size, the machine learning process calls for the refinement of the classifier by splitting the training data into training and testing sets. The classifier should be trained using the training set and its performance scored using the testing set. The tuning parameters of the classifier would be adjusted, re-trained and re-scored using multidimensional gradient descent to minimize the classification error. Here, we show only a single iteration of this process using a trained but unoptimized neural network as our classifier. Our trained neural network predicts the correct labels for the interactions with which it was trained about 97-98% of the time (there is some stochastic variation), but this does not represent a rigorous examination of its accuracy on unlabeled data. Fortunately, our training set represents a negligible fraction of the interactions found in the ecology literature. The effort of extracting, reformatting and standardizing it is the only thing that limited us to 50 examples. An automated approach is under development, but is beyond the scope of this study.



(a) Labeled & unlabeled interactions

(b) Classified interactions

(c) Labeled & unlabeled interactions

(d) Classified interactions

Figure 14: Projections of labeled and predicted interactions. Projections of unlabeled microbiome interactions and ecological interactions from the literature into subspaces spanned by two important axes of the feature space **(a)** and the first and second principle components of the feature space of the labeled interactions **(c)**. A neural network was trained using the feature space of labeled interactions and labels were predicted for the microbiome interactions. These predicted interactions were projected into subspaces spanned by the same two axes of the feature space of labeled interactions **(b)** and into a space spanned by their first and second principle components **(d)**.

6 CONCLUSION

Ascertaining the ecological function of organisms in the microbiome is a critical step along the path to understanding how microbiomes work. For relationships that have persisted long enough to shape the evolution of the host and the microbe, patterns in the evolution of the host and the microbe may provide insight into the type of mutual adaptation operating in the relationship. This information, if it is present, will be embodied in the topology of the phylogenetic trees of the two groups of interacting organisms. Graph theory provides a sophisticated suite of tools for interrogating the topology of trees, but not a framework for assaying the significance of the features it illuminates.

Rather than constructing an implicit or explicit model of how we expect coevolution to influence the topology of the host and microbial phylogenies, we propose a comparative approach. The graph theoretic view of tree topology will admit generalized graphs without any difficulty, and so one may construct graph objects that represent complete ecological relationships in their evolutionary context. Graph theory furnishes a dissimilarity metric, as well as a way of extracting moments on topological features, that serve as the basis of a feature space in which the topology of graphs are projected. In this feature space, one may project graphs for which the nature of the ecological relationship shaping the interaction is understood. These labeled interactions can then be used to train and evaluate classifiers, which can then be used to make predictions about the nature ecological relationships of interactions that are unknown.

6.1 Future directions

We demonstrate this process using a small dataset of microbiome data from a group of 14 host organisms and a neural network trained on small collection of 51 labeled interactions. While the results are promising, we stress that this is a proof-of-concept. A much larger collection of labeled interactions would be preferable for training the neural network, and would make it possible to conduct a rigorous evaluation of the error rate. A more comprehensive collection of host organisms could be subdivided into different adaptive innovations, opening up the possibility of identifying key microbial players in, for example, the evolution of trophic strategies. In this study, we use a region of the 16S rRNA gene to reconstruct the microbial phylogeny. While this choice maximizes the microbial diversity we are able to observe, the slow evolution of this gene limits the sensitivity of our proof-of-concept study to interactions that take place on long time scales. A metagenomic approach would make it possible to target more quickly evolving genes and specific functions. Lastly, and perhaps most interestingly, we note that the generalization to a graph structure is capable of representing much more than binary ecological interactions. It would be trivial to incorporate three trees representing the evolution and ranges of a host, a parasite and a vector. With an arbitrary number of trees, one could use this framework as a phylogenomic tool for classifying coalescent events by the patterns left in the way orthologs and paralogs assort within and among genomes.

7 FUNDING

RYN was funded by a grant from the Alfred P. Sloan Foundation to Jonathan A. Eisen.

References

[1] Amanda Kyle Gibson and Jesualdo Arturo Fuentes. "A phylogenetic test of the Red Queen Hypothesis: outcrossing and parasitism in the nematode phylum". In: *Evolution* 69.2 (2015), pp. 530–540.

[2] Carl T Bergstrom and Michael Lachmann. "The Red King effect: when the slowest runner wins the coevolutionary race". In: *Proceedings of the National Academy of Sciences* 100.2 (2003), pp. 593–598.

[3] Chaitanya S Gokhale and Arne Traulsen. "Mutualism and evolutionary multiplayer games: revisiting the Red King". In: *Proceedings of the Royal Society of London B: Biological Sciences* 279.1747 (2012), pp. 4611–4616.

[4] Daniel H Janzen. "When is it coevolution". In: *Evolution* 34.3 (1980), pp. 611–612.

[5] Leigh Van Valen. "A new evolutionary law". In: *Evol Theory* 1 (1973), pp. 1–30.

[6] Kristina Linnea Hillesland. "Evolution on the bright side of life: microorganisms and the evolution of mutualism". In: *Annals of the New York Academy of Sciences* (Nov. 2017). DOI: 10.1111/nyas.13515. URL: https://doi.org/10.1111/nyas.13515.

[7] Ugo Bastolla et al. "The architecture of mutualistic networks minimizes competition and increases biodiversity". In: *Nature* 458.7241 (2009), pp. 1018–1020.

[8] Enrico L Rezende et al. "Non-random coextinctions in phylogenetically structured mutualistic networks". In: *Nature* 448.7156 (2007), p. 925.

[9] Peter E Larsen, Frank R Collart, and Yang Dai. "Predicting ecological roles in the rhizosphere using metabolome and transportome modeling". In: *PloS one* 10.9 (2015), e0132837.

[10] Jordi Sardans, Josep Penuelas, and Albert Rivas-Ubach. "Ecological metabolomics: overview of current developments and future challenges". In: *Chemoecology* 21.4 (2011), pp. 191–225.

[11] Jacob G Bundy, Matthew P Davey, and Mark R Viant. "Environmental metabolomics: a critical review and future perspectives". In: *Metabolomics* 5.1 (2009), p. 3.

[12] Xingpeng Jiang et al. "Functional biogeography of ocean microbes revealed through non-negative matrix factorization". In: *PLoS One* 7.9 (2012), e43866.

[13] Holly M Bik. "Deciphering diversity and ecological function from marine metagenomes". In: *The Biological Bulletin* 227.2 (2014), pp. 107–116.

[14] Olivia U Mason et al. "Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill". In: *The ISME journal* 6.9 (2012), p. 1715.

[15] Tim Urich et al. "Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome". In: *PloS one* 3.6 (2008), e2527.

[16] Scott M Gifford et al. "Expression patterns reveal niche diversification in a marine microbial assemblage". In: *The ISME journal* 7.2 (2013), p. 281.

[17] Da-Zhi Wang et al. "Environmental Microbial Community Proteomics: Status, Challenges and Perspectives". In: *International journal of molecular sciences* 17.8 (2016), p. 1275.

[18] Joseph Felsenstein. "Phylogenies and the comparative method". In: *The American Naturalist* 125.1 (1985), pp. 1–15.

[19] Tamara Münkemüller et al. "How to measure and test phylogenetic signal". In: *Methods in Ecology and Evolution* 3.4 (2012), pp. 743–756.

[20] Mark S Hafner et al. “Disparate rates of molecular evolution in cospeciating hosts and parasites”. In: *Science* 265.5175 (1994), p. 1087.

[21] Roderic DM Page. “Genes, organisms, and areas: the problem of multiple lineages”. In: *Systematic Biology* 42.1 (1993), pp. 77–84.

[22] John P Huelsenbeck, Bruce Rannala, and Bret Larget. “A Bayesian framework for the analysis of cospeciation”. In: *Evolution* 54.2 (2000), pp. 352–364.

[23] Joseph Felsenstein. “Cases in which parsimony or compatibility methods will be positively misleading”. In: *Systematic zoology* 27.4 (1978), pp. 401–410.

[24] Kerstin Hommola et al. “A permutation test of host–parasite cospeciation”. In: *Molecular biology and evolution* 26.7 (2009), pp. 1457–1468.

[25] Nathan Mantel. “The detection of disease clustering and a generalized regression approach”. In: *Cancer research* 27.2 Part 1 (1967), pp. 209–220.

[26] Marcial Escudero. “Phylogenetic congruence of parasitic smut fungi (Anthracoidea, Anthracoideaceae) and their host plants (Carex, Cyperaceae): Cospeciation or host-shift speciation?” In: *American journal of botany* 102.7 (2015), pp. 1108–1114.

[27] Francis J Anscombe. “Graphs in statistical analysis”. In: *The American Statistician* 27.1 (1973), pp. 17–21.

[28] Frederick A Matsen and Steven N Evans. “Ubiquity of synonymity: almost all large binary trees are not uniquely identified by their spectra or their immanantal polynomials”. In: *Algorithms for Molecular Biology* 7.1 (2012), p. 14.

[29] Eric Lewitus and Helene Morlon. “Characterizing and comparing phylogenies from their Laplacian spectrum”. In: *Systematic biology* (2015), syv116.

[30] Allen J Schwenk. “Almost all trees are cospectral”. In: *New directions in the theory of graphs* (1973), pp. 275–307.

[31] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *Journal of Machine Learning Research* 12.Oct (2011), pp. 2825–2830.

[32] *Guidelines for the Humane Transportation of Research Animals*. National Academies Press, July 2006. DOI: 10.17226/11557. URL: https://doi.org/10.17226/11557.

[33] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* (2014), btu170.

[34] Torbjørn Rognes et al. “VSEARCH: a versatile open source tool for metagenomics”. In: *PeerJ* 4 (2016), e2584.

[35] Nikolas Askitis and Justin Zobel. “Cache-conscious collision resolution in string hash tables”. In: *International Symposium on String Processing and Information Retrieval*. Springer. 2005, pp. 91–102.

[36] Nikolas Askitis and Ranjan Sinha. “HAT-trie: a cache-conscious trie-based data structure for strings”. In: *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*. Australian Computer Society, Inc. 2007, pp. 97–105.

[37] Wes McKinney et al. “Data structures for statistical computing in python”. In: *Proceedings of the 9th Python in Science Conference*. Vol. 445. van der Voort S, Millman J. 2010, pp. 51–56.

[38] Fernando Pérez and Brian E Granger. “IPython: a system for interactive scientific computing”. In: *Computing in Science & Engineering* 9.3 (2007).

[39] John D Hunter. “Matplotlib: A 2D graphics environment”. In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95.

[40] Michael G Ross et al. “Characterizing and measuring bias in sequence data”. In: *Genome biology* 14.5 (2013), R51.

[41] Melanie Schirmer et al. "Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data". In: *BMC bioinformatics* 17.1 (2016), p. 125.

[42] Mickael Goujon et al. "A new bioinformatics analysis tools framework at EMBL–EBI". In: *Nucleic acids research* 38.suppl 2 (2010), W695–W699.

[43] Fabian Sievers et al. "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega". In: *Molecular systems biology* 7.1 (2011), p. 539.

[44] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. "FastTree: computing large minimum evolution trees with profiles instead of a distance matrix". In: *Molecular biology and evolution* 26.7 (2009), pp. 1641–1650.

[45] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. "FastTree 2–approximately maximum-likelihood trees for large alignments". In: *PloS one* 5.3 (2010), e9490.

[46] Stéfan van der Walt, S Chris Colbert, and Gael Varoquaux. "The NumPy array: a structure for efficient numerical computation". In: *Computing in Science & Engineering* 13.2 (2011), pp. 22–30.

[47] Vladimir I Levenshtein. "Binary codes capable of correcting deletions, insertions, and reversals". In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[48] Jose Diaz-Gonzalez. *fuzzywuzzy*. https://github.com/seatgeek/fuzzywuzzy. 2017.

[49] Jeet Sukumaran and Mark T Holder. "DendroPy: a Python library for phylogenetic computing". In: *Bioinformatics* 26.12 (2010), pp. 1569–1571.

[50] Oliver G Pybus and Paul H Harvey. "Testing macro–evolutionary models using incomplete molecular phylogenies". In: *Proceedings of the Royal Society of London B: Biological Sciences* 267.1459 (2000), pp. 2267–2272.

[51] Thomas A Pearson and Teri A Manolio. "How to interpret a genome-wide association study". In: *Jama* 299.11 (2008), pp. 1335–1344.

[52] Cyrille Violle et al. "The emergence and promise of functional biogeography". In: *Proceedings of the National Academy of Sciences* 111.38 (2014), pp. 13690–13696.

[53] Megan AM Kutzer and Sophie AO Armitage. "Maximising fitness in the face of parasites: a review of host tolerance". In: *Zoology* 119.4 (2016), pp. 281–289.

[54] Evelyn C Rynkiewicz, Amy B Pedersen, and Andy Fenton. "An ecosystem approach to understanding and managing within-host parasite community dynamics". In: *Trends in parasitology* 31.5 (2015), pp. 212–221.

[55] Cole G Easson and Robert W Thacker. "Phylogenetic signal in the community structure of host-specific microbiomes of tropical marine sponges". In: *Frontiers in microbiology* 5 (2014).

[56] Sandra Schöttner et al. "Relationships between host phylogeny, host type and bacterial community diversity in cold-water coral reef sponges". In: *PloS one* 8.2 (2013), e55505.

[57] Liam J Revell, Luke J Harmon, and David C Collar. "Phylogenetic signal, evolutionary process, and rate". In: *Systematic Biology* 57.4 (2008), pp. 591–601.

[58] Kerstin Gronbach et al. "Safety of probiotic Escherichia coli strain Nissle 1917 depends on intestinal microbiota and adaptive immunity of the host". In: *Infection and immunity* 78.7 (2010), pp. 3036–3046.

[59] David L Gally and Mark P Stevens. "Microbe Profile: Escherichia coli O157: H7–notorious relative of the microbiologist's workhorse". In: *Microbiology* 163.1 (2017), pp. 1–3.

[60] Maarten PM Vanhove et al. "Hidden biodiversity in an ancient lake: phylogenetic congruence between Lake Tanganyika tropheine cichlids and their monogenean flatworm parasites". In: *Scientific reports* 5 (2015), p. 13669.

[61] Bruce Beehler. "Frugivory and polygamy in birds of paradise". In: *The Auk* (1983), pp. 1–12.

[62] Mary T Kalin Arroyo, Richard Primack, and Juan Armesto. "Community studies in pollination ecology in the high temperate Andes of central Chile. I. Pollination mechanisms and altitudinal variation". In: *American journal of botany* (1982), pp. 82–97.

[63] John W Baird. "The selection and use of fruit by birds in an eastern forest". In: *The Wilson Bulletin* (1980), pp. 63–73.

[64] Tomás A Carlo, Jaime A Collazo, and Martha J Groom. "Avian fruit preferences across a Puerto Rican forested landscape: pattern consistency and implications for seed removal". In: *Oecologia* 134.1 (2003), pp. 119–131.

[65] Frederic E Clements. "Experimental Pollination An Outline Of The Ecology Of Flower And Insects". In: (1923).

[66] FHJ Crome. "The ecology of fruit pigeons in tropical Northern Queensland." In: *Wildlife Research* 2.2 (1975), pp. 155–185.

[67] LV Dicks, SA Corbet, and RF Pywell. "Compartmentalization in plant–insect flower visitor webs". In: *Journal of Animal Ecology* 71.1 (2002), pp. 32–43.

[68] Yoko L Dupont, Dennis M Hansen, and Jens M Olesen. "Structure of a plant–flower-visitor network in the high-altitude sub-alpine desert of Tenerife, Canary Islands". In: *Ecography* 26.3 (2003), pp. 301–310.

[69] Heidi Elberling and Jens M Olesen. "The structure of a high latitude plant-flower visitor system: the dominance of flies". In: *Ecography* 22.3 (1999), pp. 314–323.

[70] Heidi Elberling and Jens M Olesen. Data from Zackenberg Research Station.

[71] Eskildsen et al. Data from Mauritius Island.

[72] PGH Frost. "Fruit-frugivore interactions in a South African coastal dune forest". In: *Acta XVII Congr Int Orn Deutschen Orn Ges, Berlin, West Germany* (1980), pp. 1179–1184.

[73] Mauro Galetti and Marco Aurelio Pizo. "Fruit eating by birds in a forest fragment in southeastern Brazil." In: *Revista Brasileira de Ornitologia-Brazilian Journal of Ornithology* 4.5 (2013), p. 9.

[74] Andreas Hamann and Eberhard Curio. "Interactions among frugivores and fleshy fruit trees in a Philippine submontane rainforest". In: *Conservation Biology* 13.4 (1999), pp. 766–773.

[75] Javier Herrera. "Pollination relationships in southern Spanish Mediterranean shrublands". In: *The Journal of Ecology* (1988), pp. 274–287.

[76] Brian Hocking. "Insect-flower associations in the high Arctic with special reference to nectar". In: *Oikos* (1968), pp. 359–387.

[77] Pedro Jordano. "El ciclo anual de los paseriformes frugívoros en el matorral mediterráneo del sur de España: importancia de su invernada y variaciones interanuales". In: *Ardeola* 32.1 (1985), pp. 69–94.

[78] David W Inouye and Graham H Pyke. "Pollination biology in the Snowy Mountains of Australia: comparisons with montane Colorado, USA". In: *Austral Ecology* 13.2 (1988), pp. 191–205.

[79] Gail E Kantak. "Observations on some fruit-eating birds in Mexico". In: *The Auk* 96.1 (1979), pp. 183–186.

[80] Peter G Kevan. "High arctic insect-flower relations: the inter-relationships of arthropods and flowers at Lake Hazen, Ellesmere Island, N. W. T., Canada". PhD thesis. University of Alberta Edmonton,, Canada, 1970.

[81] Caroline EG Tutin et al. "The primate community of the Lopé Reserve, Gabon: diets, responses to fruit scarcity, and effects on biomass". In: *American Journal of Primatology* 42.1 (1997), pp. 1–24.

[82] Andrew L Mack and Debra D Wright. "Notes on occurrence and feeding of birds at Crater Mountain biological research station, Papua New Guinea". In: *Emu* 96.2 (1996), pp. 89–101.

[83] Diego Medan et al. "Plant-pollinator relationships at two altitudes in the Andes of Mendoza, Argentina". In: *Arctic, Antarctic, and Alpine Research* (2002), pp. 233–241.

[84] Jane Memmott. "The structure of a plant-pollinator food web". In: *Ecology letters* 2.5 (1999), pp. 276–280.

[85] Theodore Mosquin and Jeh Martin. "Observations on the pollination biology of plants on Melville Island, NWT, Canada". In: *Canadian Field Naturalist* 81 (1967), pp. 201–205.

[86] Nathaniel T Wheelwright et al. "Tropical fruit-eating birds and their food plants: a survey of a Costa Rican lower montane forest". In: *Biotropica* (1984), pp. 173–192.

[87] Alexander F Motten. "Pollination ecology of the spring wildflower community of a temperate deciduous forest". In: *Ecological Monographs* 56.1 (1986), pp. 21–42.

[88] CK McMullen. "Flower-visiting insects of the Galapagos Islands". In: *Pan-Pacific Entomologist* 69.1 (1993), pp. 95–106.

[89] Pedro Jordano. Nava Correhuelas. S. Cazorla, SE Spain.

[90] Pedro Jordano. Nava Noguera, Sierra de Cazorla, SE Spain.

[91] Heidi Elberling and Jens M Olesen. Data from Garajonay, Gomera, Spain.

[92] Mary Percival. "Floral ecology of coastal scrub in southeast Jamaica". In: *Biotropica* (1974), pp. 104–129.

[93] Nelson Ramirez and Ysaleny Brito. "Pollination biology in a palm swamp community in the Venezuelan Central Plains". In: *Botanical Journal of the Linnean Society* 110.4 (1992), pp. 277–302.

[94] Nelson Ramirez. "Biologia de polinizacion en una comunidad arbustiva tropical de la Alta Guayana Venezolana". In: *Biotropica* (1989), pp. 319–330.

[95] N Noma. "Annual fluctuations of sapfruits production and synchronization within and inter species in a warm temperate forest on Yakushima Island". In: *Tropics* 6 (1997), pp. 441–449.

[96] Douglas W Schemske et al. "Flowering ecology of some spring woodland herbs". In: *Ecology* 59.2 (1978), pp. 351–366.

[97] Ernest Small. "Insect pollinators of the Mer Bleue peat bog of Ottawa". In: *Canadian field-naturalist* (1976).

[98] Barbara K Snow and DW Snow. "The feeding ecology of tanagers and honeycreepers in Trinidad". In: *The Auk* 88.2 (1971), pp. 291–322.

[99] Wesley R Silva et al. "28 Patterns of Fruit–Frugivore Interactions in Two Atlantic Forest Bird Communities of South-eastern Brazil: Implications for Conservation". In: *Seed Dispersal and Frugivory: Ecology, Evolution, and Conservation* (2002), p. 423.

[100] Barbara K Snow and David W Snow. *Birds and Berries*. T & AD Poyser. Academic Press, 1990. ISBN: 0856610496.

Name	Label	Links (n_L)	Occupancy (k)	Squareness (q)	Eigengap (λ_δ)	Kurtosis (γ_2)	Skew (γ_1)	Hommola cor. (r_H)	Hommola sig. (p_H)	treedist (D_t)
Gopher, Lice [20]	parasitism	17	1.06	0.88	0.36	-0.70	0.74	0.490	1.4e-09	0.14
Sedge, Smut [26]	parasitism	41	1.24	1.44	0.39	1.99	1.81	0.152	1.3e-05	0.18
Fish, Worm [60]	parasitism	191	1.80	0.11	1.54	17.55	4.25	0.548	$< 10^{-9}$	2.60
beeh [61]	frugivory	119	5.95	3.44	1.81	1.07	1.58	-0.036	2.4e-03	0.11
arr1 [62]	pollination	307	3.41	0.88	7.06	7.52	2.91	0.045	4.2e-22	0.09
arr2 [62]	pollination	170	3.33	0.73	3.12	1.88	1.83	0.054	8.3e-11	0.06
arr3 [62]	pollination	62	2.14	1.64	0.72	0.02	1.27	0.091	7.6e-05	0.09
bair [63]	frugivory	50	3.57	0.33	1.23	1.67	1.71	-0.032	2.6e-01	0.47
cacg [64]	frugivory	65	3.42	1.53	2.49	1.09	1.63	-0.019	3.8e-01	0.11
caco [64]	frugivory	47	2.61	1.77	2.19	0.52	1.44	-0.009	7.6e-01	0.09
caci [64]	frugivory	94	3.55	1.65	3.68	2.90	2.08	-0.047	1.9e-03	0.03
cafr [64]	frugivory	51	2.83	1.40	2.16	0.51	1.44	-0.036	2.0e-01	0.06
cllo [65]	pollination	871	5.09	0.39	4.96	10.25	3.34	0.018	4.1e-29	0.17
crom [66]	frugivory	125	3.25	11.83	2.29	7.72	2.98	-0.020	8.5e-02	0.24
dihi [67]	pollination	141	3.76	0.29	22.86	9.61	3.25	0.078	1.2e-14	0.08
dish [67]	pollination	82	3.28	0.47	5.88	2.94	2.09	0.001	9.6e-01	0.17
dupo [68]	pollination	104	4.43	0.31	2.24	1.38	1.69	0.025	7.0e-02	0.15
eol [69]	pollination	230	3.38	0.21	2.89	5.34	2.56	-0.011	8.3e-02	0.26
eolz [70]	pollination	428	8.15	0.42	4.26	3.40	2.13	-0.017	5.3e-07	0.12
eski [71]	pollination	33	2.54	1.17	3.58	0.04	1.27	-0.071	1.1e-01	0.04
fros [72]	frugivory	88	7.33	2.00	2.94	-0.76	0.91	-0.040	1.4e-02	0.15
gen1 [73]	frugivory	36	3.00	0.41	2.58	1.27	1.58	-0.030	4.5e-01	0.22
gen2 [73]	frugivory	129	4.16	1.21	1.79	0.75	1.44	0.035	1.6e-03	0.56
hamm [74]	frugivory	139	4.71	2.69	2.99	0.26	1.27	0.033	1.2e-03	0.45
herr [75]	pollination	359	3.78	0.16	54.63	18.54	4.36	-0.004	3.6e-01	0.15
hock [76]	pollination	161	3.19	0.40	13.93	8.40	3.09	0.007	4.5e-01	0.08
hrat [77]	frugivory	118	7.38	1.00	3.15	0.18	1.23	-0.028	2.2e-02	0.17
inpk [78]	pollination	220	3.61	0.53	3.42	3.50	2.20	-0.002	8.1e-01	0.03
kant [79]	frugivory	86	5.38	0.19	3.46	3.28	2.16	-0.016	3.4e-01	0.70
kevn [80]	pollination	165	3.47	0.27	29.79	11.72	3.55	-0.016	6.0e-02	0.21
lope [81]	frugivory	50	4.00	2.12	2.50	-1.02	0.72	-0.029	3.0e-01	0.16
mack [82]	frugivory	35	1.09	1.00	0.40	0.47	1.26	0.042	3.1e-01	0.63
med1 [83]	pollination	39	1.22	0.49	5.41	0.91	1.57	0.023	5.3e-01	0.05
med2 [83]	pollination	114	2.51	0.34	6.03	7.39	2.93	0.017	1.7e-01	0.09
memm [84]	pollination	114	2.51	0.34	6.03	7.39	2.93	0.017	1.7e-01	0.09
moma [85]	pollination	35	2.50	0.65	2.49	-0.10	1.20	0.077	6.2e-02	0.17
mont [86]	frugivory	626	6.05	4.31	23.47	11.98	3.58	0.005	2.1e-02	0.64
mott [87]	pollination	104	3.78	0.31	9.05	4.95	2.50	-0.023	9.5e-02	0.29
mull [88]	pollination	173	2.40	2.69	53.75	18.57	4.37	-0.005	5.5e-01	0.12
ncor [89]	frugivory	148	5.10	0.76	1.82	2.82	1.99	-0.004	6.7e-01	0.66
nnog [90]	frugivory	129	5.61	0.64	1.55	1.87	1.75	-0.022	5.0e-02	0.60
olau [91]	pollination	76	1.83	0.54	2.09	2.21	1.92	-0.020	3.0e-01	0.13
perc [92]	pollination	140	3.04	1.97	11.72	6.43	2.77	0.026	9.2e-03	0.03
rabr [93]	pollination	83	2.10	0.72	5.85	1.60	1.77	0.001	9.7e-01	0.17
rmrz [94]	pollination	147	3.13	1.04	4.31	2.28	1.93	-0.052	6.9e-08	0.06
sapf [95]	frugivory	34	1.94	3.38	2.96	0.05	1.18	-0.060	1.6e-01	0.26
schm [96]	pollination	64	3.28	0.22	4.17	3.85	2.29	0.004	8.4e-01	0.34
smal [97]	pollination	132	5.87	0.41	3.49	1.03	1.56	0.081	3.2e-14	0.17
snow [98]	frugivory	222	7.16	3.43	1.71	1.14	1.61	0.035	3.2e-08	0.45
wes [99]	frugivory	859	6.01	2.58	6.31	6.67	2.78	0.061	3.0e-296	0.93
wyth [100]	frugivory	47	3.76	0.79	3.28	0.36	1.28	-0.007	8.3e-01	0.38

Table 3: Spectral features of ecological interactions gathered from the literature.

Sample ID	Sample source	Barcode 1	Barcode 2	Reads after QC
CYPRCOL1	*Chalinochromis brichardi*	AACGCTAA	GACTGGTT	174,725
CYPRCOL2	*Chalinochromis brichardi*	AACGCTAA	TTAGCGTT	271,491
CYPRCOL3	*Chalinochromis brichardi*	AACGCTAA	GTAGTCTT	263,386
CYPRCOL4	*Chalinochromis brichardi*	AACGCTAA	TATCGATT	284,670
CYPRCOL5	*Chalinochromis brichardi*	AACGCTAA	CAATTGGT	242,392
CYPRCOL6	*Chalinochromis brichardi*	AACGCTAA	ACTTCAGT	241,783
HAPLMIC1	*Haplotaxodon microlepis*	AACGCTAA	GCGGCAAT	100,413
HAPLMIC2	*Haplotaxodon microlepis*	AACGCTAA	TAAGGTTG	72,009
HAPLMIC3	*Haplotaxodon microlepis*	AACGCTAA	TTATTAGG	55,773
HAPLMIC4	*Haplotaxodon microlepis*	AACGCTAA	ATCAGAGG	115,762
HAPLMIC5	*Haplotaxodon microlepis*	AACGCTAA	CTCGACCG	145,978
HAPLMIC6	*Haplotaxodon microlepis*	AACGCTAA	GCCATTAG	162,213
TRIGOTO1	*Triglachromis otostigma*	AACGCTAA	CGCATGAG	301,107
TRIGOTO2	*Triglachromis otostigma*	AACGCTAA	CTCCGTTC	326,061
TRIGOTO3	*Triglachromis otostigma*	AACGCTAA	GCGTAGGC	207,092
TRIGOTO4	*Triglachromis otostigma*	AACGCTAA	GGTAACGC	146,460
TRIGOTO5	*Triglachromis otostigma*	AACGCTAA	CAGCCTCC	177,713
TRIGOTO6	*Triglachromis otostigma*	AACGCTAA	TGGCATCC	45,174
LEPIPRO1	*Lepidiolamprologus profundicola*	AACGCTAA	CCTAATCC	179,020
TROPMOO1	*Tropheus moorii*	AACGCTAA	CGGCCAAC	246,050
CHALBRI1	*Cyprichromis coloratus*	AACGCTAA	AGTTAATA	201,117
TANGIRA1	*Tanganicodus irsacae*	AACGCTAA	AGCAGTCA	289,405
LOBOLAB1	*Lobochilotes labiatus*	AACGCTAA	AATCGCCA	121,979
NEOLBUS1	*Neolamprologus buescheri*	AACGCTAA	TCGCTGAA	106,657
TREMBEN1	*Trematochromis benthicola*	CAACCTTA	GACTGGTT	107,462
XENOFLA1	*Xenotilapia flavipinnis*	CAACCTTA	TTAGCGTT	152,933
KITCNT1	Kit control	CAACCTTA	GTAGTCTT	14,329
CHARCNT	Charcoal control	CAACCTTA	TATCGATT	21,176
WATERCNT	Water control	CAACCTTA	CAATTGGT	10,326
FOOD	Food control	CAACCTTA	ACTTCAGT	76,932
NEGCNT	PCR control	CAACCTTA	TAAGGTTG	174

Table 4: **Samples, barcodes and yields.** Except for controls, all samples are from stool of the organism indicated.