Novel Methods for Post-Processing Evolutionary Data Using Machine Learning Techniques

Dissertation
by

Lucas Czech

Karlsruhe Institute of Technology (KIT),
Karlsruhe, Germany

Heidelberg Institute for Theoretical Studies (HITS),
Heidelberg, Germany

First Reviewer:	Prof. Dr. Alexandros Stamatakis
Second Reviewer:	Prof. Dr. Emmanuel Müller

September 30, 2018

Zusammenfassung

Auf deutsch...

Abstract

In English...

Hiermit erkläre ich, dass ich diese Arbeit selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht habe. Ich habe die Satzung des Karlsruher Institutes für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis beachtet.

Karlsruhe, September 30, 2018

. . . .

(**Lucas Czech**)

Acknowledgment

There are many people who helped and supported me in different ways during the conduction of this dissertation.

First, I would like to thank my advisor Professor Dr. Alexandros Panavotis Stamatakis,

Many thanks also to my second advisor Professor Dr. Emmanuel Müller for his support and advice.

To the team of the Exelixis Lab— namely ... — also many thanks, you made ....

Finally, I want to thank my parents Peter and Maria and my sister Judith, who not only constantly supported me during this thesis, but through all of my years of study in Karlsruhe, Heidelberg and around the world.

This work was financially supported by the **Klaus Tschira Stiftung gGmbH** in Heidelberg, Germany.

We thank **S. Srinivasan** and **E. Matsen** for providing the Bacterial Vaginosis dataset [128] and for helping us understanding their methods and implementations. We also thank **M. Dunthorn**, **L. Rubinat-Ripoll**, **C. Berney**, **L. Guidi**, **G. Lentendu**, **A. Kozlov**, and **P. Barbera** for their feedback on our methods and help with this manuscript.

We thank **S. Srinivasan** and **E. Matsen** for providing the Bacterial Vaginosis dataset [128]; **M. Dunthorn**, **G. Lentendu**, and **A. Kozlov** for their feedback on our methods and this manuscript.

Contents

List of Figures

List of Tables

List of Acronyms

RT reference tree

RA reference alignment

QS query sequence

BT backbone tree

CT clade tree

LWR likelihood weight ratio

ART automatic reference tree

PCA Principal Components Analysis

MDS Multidimensional scaling

BV Bacterial Vaginosis

NTS Neotropical Soils

TO Tara Oceans

HMP Human Microbiome Project

1. Introduction

1.1 Motivation

1.2 Objective and Contribution

swarm code contrib [80, 81]

full list of publications is available in C

1.3 Structure

2. Foundations

This chapter is partially based on the peer-reviewed publications:

- **Lucas Czech**, Pierre Barbera, and Alexandros Stamatakis. ”Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement.” *Bioinformatics*, 2018.
- **Lucas Czech** and Alexandros Stamatakis. ”Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements.” *PLOS One*, 2018.

Contributions: Lucas Czech... Pierre Barbera... Alexandros Stamatakis... and...

sequences, sampling, alignments, trees, placement, distances

2.1 Evolution and Genetics

there is sequencing [120], also of the human genome [140], which gets cheaper all the time [147]

2.1.1 Metagenomics

The availability of high-throughput DNA sequencing technologies has revolutionized biology by transforming it into an ever more data-driven and compute-intense discipline [38]. In particular, Next Generation Sequencing (NGS) [72] has given rise to novel methods for studying microbial environments [37, 86, 101, 132]. NGS techniques are often used in metagenomic studies to sequence organisms in water [44, 46, 56] or soil [33, 82] samples, in the human microbiome [51, 95, 128], and a plethora of other environments. These studies yield a large set of short anonymous DNA sequences, so-called reads, for each sample. Reads that are obtained from specific parts of the genome are called meta-barcoding reads; most often, reads are amplified before sequencing and later de-replicated again, resulting in so-called amplicons.

High-throughput DNA sequencing technologies have revolutionized biology by transforming it into a data-driven computational discipline [38]. Next Generation Sequencing (NGS) methods now allow for studying microbial samples directly extracted from their environment [37, 86, 101, 107, 132] For each sample, these methods yield a set of short, anonymous DNA sequences, so-called reads.

challenges and bottlenecks in metag analysis [122] bottleneck is computational analysis. who is there, what are they doing [29] (aka taxonomic and functional diversity) metag review and caveats. it brings its problems [136]

2.1.2 Multiple Sequence Alignments

2.1.3 Consensus Sequences

ambiguity chars [52]

cons methods: see art

2.2 Phylogenetic Trees

phylogenetics is...

also, we can infer trees [42]

2.2.1 Tree Inference

2.3 Phylogenetic Placement

2.3.1 Motivation

A typical task in metagenomic studies is to identify and classify these sequences with respect to known reference sequences, either in a taxonomic or phylogenetic context.

Conventional methods like Blast [3] are based on sequence similarity. Such methods are fast, but only attain satisfying accuracy levels if the query sequences (e.g., the environmental reads or amplicons) are sufficiently similar to the reference sequences. Furthermore, the best Blast hit does often **not** represent the most closely related species [61].

Alternatively, so-called phylogenetic (or evolutionary) placement methods [7, 11, 90] identify query sequences based on a phylogenetic tree of reference sequences. Thereby, they incorporate information about the evolutionary history of the species under study and hence provide a more accurate means for read identification. The result of phylogenetic placement is a mapping of the query sequences to the branches of the reference tree. Such a mapping also elucidates the evolutionary distance between the query and the reference sequences.

This information represents useful biological knowledge **per se**. For example, the classification of query sequences can be summarized by means of sequence abundances [50, 108], or to obtain taxonomic annotations [62]. The data can also be utilized to derive further knowledge or hypotheses. Existing methods such as Edge PCA and Squash Clustering [87] are, for instance, able to visualize differences between sets of metagenomic samples, or to cluster samples based on placement similarity. Note that distinct samples from one study are typically mapped to the same underlying reference tree, thus facilitating such comparisons.

phylogenetic placement [7, 11, 90], cited more than 630 times (as of 2018-07-01) considered an established method

There also exist variants of phylogenetic placement that use maximum parsimony [11] and minimum evolution [43] instead of maximum likelihood, variants that calculate Bayesian posterior probabilities [90], and boosting methods to improve the accuracy of the placements [99]. Phylogenetic placement has been used for a variety of applications and derived pipelines, such as species delimitation [55, 152], genome and metagenome analysis [25], taxonomic identification and phylogenetic profiling [105], and identification and correction of taxonomically mislabeled sequences [62].

combine 1,010 more citations (as of 2018-07-01)

aligning

kommt unten noch mal!

, using programs such as PaPaRa [9, 10] or hmmalign, which is a subprogram of the HMMER suite [34, 35].

2.3.2 Introduction

In brief, phylogenetic placement calculates the most probable insertion branches for each given query sequence (QS) on a reference tree (RT). The QSs are reads or amplicons from environmental samples. The RT and the reference sequences it represents are typically assembled by the user so that they capture the expected species diversity in the samples. Nonetheless, we recently presented an automated approach for selecting and constructing appropriate reference sequences from large sequence databases [23]. As phylogenetic placement used a maximum likelihood criterion, the RT has to be strictly bifurcating. Prior to the placement, the QSs need to be aligned against the reference alignment of the RT by programs such as PaPaRa [9, 10] or hmmalign, which is part of the HMMER suite [34, 35]. The placement is then conducted by initially inserting one QS as a new tip into a branch of the tree, then re-optimizing the branch lengths that are most affected by the insertion, and thereafter evaluating the resulting likelihood score of the tree under a given model of nucleotide evolution, such as the Generalized Time-Reversible (GTR) model [134]. The QS is then removed from the current branch and subsequently placed into all other branches of the RT.

Thus, for each branch of the tree, the process yields a so-called **placement** of the QS, that is, an optimized position on the branch, along with a likelihood score for the whole tree. The likelihood scores for a QS are then transformed into probabilities, which quantify the uncertainty of placing the sequence on the respective branch [131, 145]. Those probabilities are called likelihood weight ratios (LWRs). The accumulated LWR sum over all branches for a single QS is 1.0. Figure 2.1 shows an example depicting the placements of one QS, including the respective LWRs.

Figure 2.1: Phylogenetic Placement of a Query Sequence. Each branch of the reference tree is tested as a potential insertion position, called a “placement” (blue dots). Note that placements have a specific position on their branch, due to the branch length optimization process. A probability of how likely it is that the sequence belongs to a specific branch is computed (numbers next to dots), which is called the **likelihood weight ratio** (LWR). The bold number (0.7) denotes the most probable placement of the sequence.

This process is repeated for every QS. Note that the placement process is conducted *independently* for each QS. That is, for each QS, the algorithm starts calculating placements from scratch on the original RT.

In summary, the result of a phylogenetic placement analysis is a mapping of the QSs in a sample to positions on the branches of the RT. Each such position, along with the corresponding LWR, is called a placement of the QS.

TODO: besser unterbringen The data is usually stored in so-called `jplace` files [91].

Each such sample represents a geographical location, a body site, a point in time, etc. In the following, we represent a sample by the placement locations of its metagenomic QSs, including the respective per-branch LWRs. Furthermore, for a specific analysis, we assume the standard use case, that is, all placements were computed on the same fixed reference tree (RT) and reference alignment.

2.4 Phylogenetic Placement Processing

When placing multiple samples, for instance, from different locations, typically, the same RT is used, in order to allow for comparisons of the phylogenetic composition of these samples. In this context, it sis important to consider how to properly normalize the samples. Normalization is required as the sample size (often also called library size), that is, the number of sequences per sample, can vary by several orders of magnitude, due to efficiency variations in the sequencing process or biases introduced by the amplification process. Selecting an appropriate normalization strategy constitutes a common problem in many metagenomic studies. The appropriateness depends on data characteristics [146], but also on the biological question asked. For example, estimating indices such as the species richness are often implemented via rarefaction and rarefaction curves [45], which however ignores a potentially large amount of the available valid data [94]. Furthermore, the specific type of input sequence data has to be taken into account for normalization: Biases induced by the amplification process can potentially be avoided if, instead of amplicons, data based on shotgun sequencing are used, such as $_{\mathrm{mi}}$tags [73]. Moreover, the sequences can be clustered prior to phylogenetic placement analysis, for instance, by constructing operational taxonomic units (OTUs) [36, 80, 81, 118]. Analyses using OTUs focus on species diversity instead of simple abundances. OTU clustering substantially reduces the number of sequences, and hence greatly decreases the computational cost for placement analyses. Lastly, one may completely ignore the abundances (which are called “multiplicities” of placements) of the placed sequences, reads, or OTUs, and only be interested in their presence/absence when comparing samples.

Which of the above analysis strategies is deployed, depends on the specific design of the study and the research question at hand. The common challenge is that the number of sequences per sample differs, which affects most post-analysis methods. Before introducing our methods, we therefore explain how the necessary normalizations of sample sizes can be performed in the following. We also describe general techniques for interpreting and working with phylogenetic placement data. Some of these techniques have been used before as building blocks for methods like Edge PCA and Squash Clustering [39, 87].

2.4.1 Edge Masses

Methods that compare samples directly based on their sequences, such as the UniFrac distance [75, 77], can benefit from rarefaction [146]. However, in the context of phylogenetic placement, rarefaction is not necessary. Thus, more valid data can be kept. To this end, it is convenient to think of the reference tree as a graph (when exploiting graph properties of the tree, we refer to the branches of the tree as edges). Then, the per-branch LWRs for a single QS can be interpreted as mass points distributed over the edges of the RT, including their respective placement positions on the branches, cf. Figure 2.1. This implies that each QS has a total accumulated mass of 1.0 on the RT. We call this the **mass interpretation** of the placed QSs, and henceforth use mass and LWR interchangeably. The **mass of an edge** refers to the sum of the LWRs on that edge for all QSs of a sample, as shown in Figure 2.2(a). The total mass of a sample is then the sum over all edge masses, which is identical to the number of QSs in the sample.

Figure 2.2: Edge Masses and Imbalances. (a) Reference tree where each edge is annotated with the normalized mass (first value, blue) and imbalance (second value, red) of the placements in a sample. The imbalance is the sum of masses on the root side of the edge minus the sum of the masses on the non-root side. The depicted tree is unrooted, hence, its top-level trifurcation (gray dot) is used as “root” node. An exemplary calculation of the imbalance is given at the top. Because terminal edges only have a root side, their imbalance is not informative. (b) The masses and imbalances for the edges of a sample constitute the rows of the first two matrices. The third matrix contains the available meta-data features for each sample. These matrices are used to calculate, for instance, the edge principal components or correlation coefficients.

The key idea is to use the distribution of placement mass points over the edges of the RT to characterize a sample. This allows for normalizing samples of different size by scaling the total sample mass to unit mass 1.0. In other words, absolute abundances are converted into relative abundances. This way, rare species, which might have been removed by rarefaction, can be kept, as they only contribute a negligible mass to the branches into which they have been placed. This approach is analogous to using proportional values for methods based on OTU count tables, that is, scaling each sample/column of the table by its sum of OTU counts [146]. Most of the methods presented here use normalized samples, that is, they use relative abundances. As relative abundances are compositional data, certain caveats occur [2, 74], which we discuss where appropriate.

When working with large numbers of QSs, the mass interpretation allows to further simplify and reduce the data: The masses on each edge of the tree can be quantized into b discrete bins, that is, each edge is divided into b intervals (or bins) of the corresponding branch length. All mass points on that edge are then accumulated into their respective nearest bin. The parameter b controls the resolution and accuracy of this approximation. In the extreme case $b := 1$, all masses on an edge are grouped into one single bin. This **branch binning** process drastically reduces the number of mass points that need to be stored and analyzed in several methods we present, while only inducing a negligible decrease in accuracy. As shown in Table 5.1. branch binning can yield a speedup of up to 75% for post-analysis run-times.

Furthermore, using masses allows to summarize a set of samples by annotating the RT with their average per-edge mass distribution. This procedure, also called **squashing** [87], sums over all sample masses per edge and then normalizes them once more to obtain unit mass for this resulting average tree. This normalized tree thereby summarizes the (sub-)set of samples it represents.

2.4.2 Edge Imbalances

So far, we have only considered the per-edge masses. Often, however, it is also of interest to “summarize” the mass of an entire clade. For example, sequences of the RT that represent species or strains might not provide sufficient phylogenetic signal for properly resolving the phylogenetic placement of short sequences [32]. In these cases, the placement mass of a sequence can be spread across different edges representing the same genus or species, thus blurring analyses based on per-edge masses. Instead, a clade-based summary can yield clearer analysis results. It can be computed by using the tree structure to appropriately transform the edge masses. Each edge splits the tree into two parts, of which only one contains the root (or top-level trifurcation) of the tree. For a given edge, its mass difference is then calculated by summing all masses in the root part of the tree and subtracting all masses in the other part, while ignoring the mass of the edge itself [87]. This difference is called the **imbalance** of the edge. It is usually normalized to represent unit total mass, as the absolute (not normalized) imbalance otherwise propagates the effects of differing sample sizes all across the tree. An example of the imbalance calculation is shown in Figure 2.2(a). The edge imbalance relates the masses on the two sides of an edge to each other. This implicitly captures the RT topology and reveals information about its clades. Furthermore, this transformation can also reveal differences in the placement mass distribution of nearby branches of the tree. This is in contrast to the KR distance, which yields low values for masses that are close to each other on the tree. Examples that illustrate the different use cases for edge mass and edge imbalance metrics are shown in the Results section.

The edge masses and edge imbalances per sample can be summarized by two matrices, which we use for all further downstream edge- and clade-related analyses, respectively. In these matrices, each row corresponds to a sample, and each column to an edge of the RT. Note that these matrices can either store absolute or relative abundances, depending on whether the placement mass was normalized.

Furthermore, many studies provide meta-data for their samples, for instance, the pH value or temperature of the samples' environment. Such meta-data features can also be summarized in a per-sample matrix, where each column corresponds to one feature. The three matrices are shown in Figure 2.2(b). Quantitative meta-data features are the most suitable for our purposes, as they can be used to detect correlations with the placement mass distributions of samples. For example, Edge principal components analysis (Edge PCA) [87] is a method that utilizes the imbalance matrix to detect and visualize edges with a high heterogeneity of mass difference between samples. Edge PCA further allows to annotate its plots with meta-data variables, for instance, by coloring, thus establishing a connection between differences in samples and differences in their meta-data [128]. In the following, we propose several new techniques to analyze placement data and their associated meta-data.

3. Automatic Reference Trees

This chapter is based on the peer-reviewed publication:

- **Lucas Czech**, Pierre Barbera, and Alexandros Stamatakis. ”Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement.” *Bioinformatics*, 2018.

Contributions: Lucas Czech... Pierre Barbera... Alexandros Stamatakis... and...

3.1 Motivation

Molecular environmental sequencing studies, particularly those that aim to conduct phylogenetic placement, often rely on a set of manually selected and aligned reference sequences to infer an RT [28, 82, 135, 137]. Creating and maintaining databases of such reference sequences constitutes a labor-intensive and potentially error-prone process. Moreover, this approach is impractical for samples that contain diverse sequences from many clades of the taxonomy, or samples obtained from unexplored environments. Lastly, even if a large RT is available, the visualization of placements on such an RT might be confusing and thus hard to interpret.

The reference tree (RT) used for phylogenetic placement should ideally (a) cover all major taxonomic groups that occur in the QSs, (b) use high-quality error-free reference sequences, and (c) not be too large to allow for unambiguous visualization and interpretation. These criteria can be met for small datasets by manually selecting curated sequences from databases. For large and taxonomically diverse samples one key challenge is that sequence databases such as Greengenes [30], Unite [1], PR2 [48], EzTaxon [58], Silva [115], and RDP [20] maintain reference collections of thousands to millions of taxonomically annotated sequences. Therefore, one needs to appropriately sub-sample sequences such that the RT can be inferred in reasonable time *and* and sufficiently covers the diversity of the sample.

Previous approaches mainly relied on phylogenetic diversity [41, 98, 109] and related methods [92]. The major drawback is that they require a comprehensive phylogeny as input. Inferring such large comprehensive phylogenies with hundreds of thousands of taxa, to subsequently reduce the taxon set again, is computationally inefficient and in certain cases infeasible.

To this end, we present a computationally efficient approach for obtaining sequences from large databases to infer an RT. This RT is then used for conducting phylogenetic placement analyses. The input of our method is a database of aligned sequences of known species including their taxonomic labels. Our approach then identifies sets of sequences that are similar to each other based on their entropy. It subsequently reduces the sequences in these sets to a predefined number of consensus sequences. This set of sequences is the output of the method, which represent clades of the taxonomy, and which is then used to infer the RT.

Here, we introduce methods to overcome the aforementioned limitations, that is, to (1) automatically obtain a high quality reference tree for phylogenetic placement, (2) split up the placement process into two steps using smaller phylogenies, and (3) accelerate the computation of placements via appropriate data pre-processing approaches. All methods are implemented as part of our GAPPA tool, which is freely available under GPLv3 at http://github.com/lczech/gappa.

3.2 Method

3.2.1 Sequence Entropy

Conventional methods for sequence similarity are often based on edit distance and other pairwise comparison methods [3, 102, 126]. This however necessitates to transform the pairwise distances to some form of ensemble measure that describes the similarity of all sequences to each other, for which there is no obvious approach [154]. There also exist methods that describe genetic variation and nucleotide diversity of sets of sequences [15, 103] which could be used for this purpose. We however decided to use entropy for measuring ensemble similarity of a set of sequences.

First, we define a measure to quantify the ensemble similarity of a set s of sequences, based on their entropy [123]. Variants of sequence entropy have been used before in numerous biological and phylogenetic contexts, for example, to asses the information content of sequences [21, 22, 68, 121, 142–144], or to measure substitution saturation [148]. Here, we use entropy for alignment sites, that is, we define the entropy (uncertainty) H at alignment site i as

$$H_i \;=\; -\sum_{c} f_{c,i} \times \log_2 f_{c,i}$$

where $c \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \mathtt{-}\}$ is the set of nucleotide states including gaps, and $f_{c,i}$ is the frequency of character c at site i of the alignment. Including gaps (-) in the summation reduces the contribution of sites that contain a large fraction of gaps. Their contribution is weighed down as all standard phylogenetic inference tools model gaps as undetermined states, that is, they do not contribute anything to the likelihood score. The entropy is 0 for sites that only contain a single character.

It increases the more different characters an alignment site contains, *and* the more similar their frequencies are. Its maximum occurs if all characters appear with the same frequency (each of them 20%). Note that we also treat ambiguous characters as gaps. As only 0.008% of the non-gap characters in our test database (Silva) are ambiguous, their influence is negligible. Ambiguous characters could however be incorporated by using fractional character counts.

Finally, the total entropy of a set s of aligned sequences is simply the sum over all per-site entropies: $H(s) = \sum_i H_i$. We use this entropy to quantify the ensemble similarity of a set of sequences. This can be regarded as an information content estimate of the sequences.

3.2.2 Sequence Grouping

The goal of this step is to group the sequences of a database into a given target number of groups/sets, such that the groups reflect the diversity of the sequences in the database. We use the taxonomy to identify potential candidate groups of sequences that could be represented by a consensus sequence. We interpret a taxonomy as a sequence labeling, where similar sequences have related labels. Thus, a taxonomy represents a pre-classification of similar sequences that can be exploited to group them.

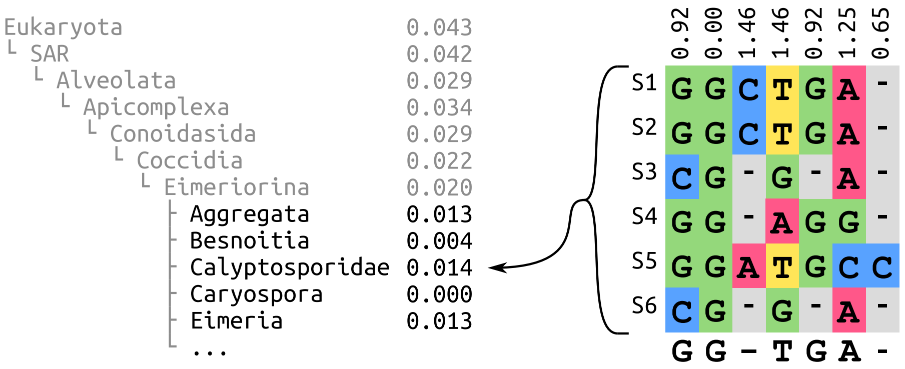

Figure 3.1: Entropy and consensus sequence of a taxonomic clade. The left hand side shows the exemplary clade *Eimeriorina* in its taxonomic context, listing its super- and sub-clades with the normalized entropy of their respective sequences. The right hand side is an excerpt from the alignment of six sequences that belong to the *Calyptosporidae* sub-clade. At its top, the per-site entropies for the alignment columns are shown. At the bottom, the majority rule consensus sequence is shown, which is used to represent the sub-clade.

For a clade t of the taxonomic tree, we denote by $H(t)$ the entropy of all sequences that belong to that clade, including all sequences in its sub-clades, that is, its lower taxonomic ranks. Clades with low entropy imply that they contain highly similar sequences that can in turn be represented by a consensus sequence without sacrificing too much diversity. Inversely, clades with high entropy contain diverse sequences, implying that a consensus sequence is not likely to sufficiently capture the inherent sequence diversity. It is thus better to expand these clades and construct separate consensus sequences for their respective sub-clades. An example is shown in Figure 3.1. As the clade structure of a taxonomy forms a tree, this criterion can then be applied recursively, as shown in Algorithm 3.1.

Algorithm 3.1 Taxonomy Expansion

```
$Candidates \leftarrow$ list of highest ranking clades
$TaxaCount \leftarrow$ size of $Candidates$
while $TaxaCount < TargetCount$ do
    $MostDiverse \leftarrow \arg\max_{t \in Candidates} H(t)$
    remove $MostDiverse$ from $Candidates$
    add sub-clades of $MostDiverse$ to $Candidates$
    $TaxaCount \leftarrow TaxaCount - 1 +$ size of $MostDiverse$
return $Candidates$
```

The algorithm works as follows: We initialize a list of candidate clades with the highest ranking clades that we want to consider. In the most general case, these can be “Archaea”, “Bacteria”, and “Eukaryota”. We then select the most diverse candidate clade, that is, the clade t whose sequences exhibit the highest entropy $H(t)$. This clade is then expanded, and we do not consider it as a potential candidate for building a consensus sequence. The high entropy clade is then removed from our list and its immediate sub-clades are added as new candidates to the list. Finally, the current count of how many candidates we have already selected is updated accordingly. By expanding clades with high entropy, we descend into the lower ranks of the taxonomy. On average, this decreases the entropy, because low ranking clades generally tend to contain more similar sequences. This process is repeated until our list contains approximately as many candidate clades as the desired target count of reference sequences, which is provided as input. As the sizes of expanded clades can vary substantially, the target count cannot always be met exactly. In our tests, the average deviation was 0.2%, as shown in Table 3.1.

Given this list of clades from different taxonomic ranks, we can now compute the consensus sequences. For each clade, all sequences in that clade and its sub-clades are used to construct a consensus sequence, which represents the clade diversity, and serves as the reference sequence for that clade. A simple per-site majority rule consensus [27, 93] works well, but we also assessed alternative methods; see Figure 3.5 and Figure 3.6 for details.

Note that it would also be possible to directly use the relative character frequencies at each site to obtain more accurate representations. Maximum likelihood-based phylogenetic inference tools do, in principle, not require discrete input sequences. The likelihood model allows to account for uncertainty in the input data [42], although this is generally not implemented in the mainstream software packages. The above process yields a set of consensus reference sequences which capture the diversity of distinct taxonomic clades.

3.2.3 Inferring a Reference Tree

Once we have identified the consensus sequences, which are already aligned to each other, we can use them to infer a maximum likelihood tree, which we call an **Automatic Reference Tree** (ART). As each consensus sequence is associated with a taxonomic clade, the corresponding taxonomic path can be used to label the tips of the tree. Note that since clades with low entropy might not be expanded, the tip labels do not necessarily correspond to species or genera. Also, the ART will not necessarily be congruent to the taxonomy.

An ART obtained via our method satisfies all criteria we listed: (a) All taxonomic groups occurring in the QSs can be covered by using a suitable taxonomy as input. (b) By using consensus sequences, potential sequencing errors can be alleviated. (c) The size of the ART can be specified by the user. However, the resolution of the trees is ultimately limited by the used taxonomy, see Figure 3.8 for details. Thus, one needs to make sure that the resulting tree is suited for the dataset to be placed on it. We note however that this is also an issue when manually selecting reference sequences. Furthermore, using consensus sequences may obscure the degree of sequence diversity in sub-clades, which in turn can affect the accuracy of subsequent phylogenetic placements on that tree. Our algorithm as described here can not fully compensate for this. We present a method to address both issues (tree resolution and obscured diversity) in the next Section.

3.3 Multilevel Placement

When conducting phylogenetic placement, the computationally limiting factors are (i) the number of QSs to be placed (addressed in the next section) and (ii) the size of the RT (number of taxa) and corresponding alignment length (addressed below). Using RTs with more taxa increases the phylogenetic resolution of the placements, at the cost of increased computational effort for inferring the RT, aligning the QSs, and placing the QSs. Furthermore, longer reference alignments (if appropriate data is available) are required to accurately infer large trees under the maximum likelihood criterion [149], thus further increasing the computational costs. Lastly, placement on large trees that comprise reference sequences with high evolutionary distances can reduce placement accuracy [99]. Thus, using a large number of reference sequences is not always desirable in practice.

One solution is to divide the tree and its alignment into more conquerable subsets, for example as implemented in SATÉ [69, 70]. This approach has also been extended to phylogenetic placement in SEPP [99] and TIPP [105], which divide the tree into disjoint subsets of taxa and conduct placement on each of them. While yielding more accurate placements and taxonomic classifications in less computing time, this method might still result in large reference trees, which are hard to inspect and visualize.

To address this issue, we present an approach called **Multilevel** or **Russian Doll** Placement, which is summarized in Figure 3.2. Instead of working with one large RT comprising *all* taxa of interest, we use a smaller, but taxonomically broad backbone tree (BT) for pre-classifying the QSs (first level), and a set of refined clade trees (CTs) for the final, more accurate placements (second level). These CTs comprise the reference sequences that are of interest for a particular study. For example, if a study is concerned with *Apicomplexa* and *Cercozoa*, a broad *Eukaryotes* BT can be used for the first level, and two respective CTs for the second level, similar to [82]. Each CT is associated with the set of branches of a specific BT clade.

The method then works in three steps:

1. Align and place the QSs using the BT (first level).

2. For each CT, collect the QSs that are placed on the BT branches associated with the CT.

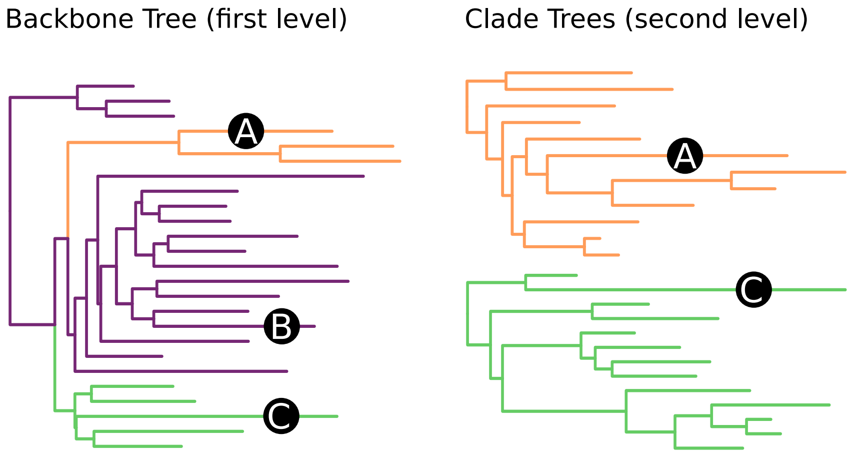

Figure 3.2: Multilevel Placement. The left shows a backbone tree (BT); the right shows two clade trees (CTs) in orange and green. Branches in the BT that are associated with a CT are marked in its color. The trees "overlap" each other, meaning that each CT is represented by multiple branches in the BT. Three sequences A , B and C are placed on the BT, which is the first level. A and C are placed on branches associated with a CT. Hence, their second level placement is conducted on the respective CT. B is placed on a branch that is not associated with any CT, and thus not used in the second level.

3. Align and place these QSs again, using their specific CTs (second level).

While this approach requires some additional bookkeeping, the total computational cost is reduced, because the QSs do not have to be placed on all branches of all CTs. The gain in speed depends on the sizes of the BT and CTs relative to the size of the (potentially imaginary) large comprehensive tree. For example, by splitting a tree with 10 000 taxa into a BT and 10 CTs with 1000 taxa each, computational cost is reduced to 20% of the original cost (two placement levels with 10% of the cost each). Furthermore, in each level, the amount of computer memory is limited to 10% of what is necessary for the large tree. Lastly, this method allows for fine-grained control over the clades of interest at both placement levels:

Firstly, the BT provides a means for phylogenetically informed sequence filtering – that is, to identify and remove "spurious" QSs. Sequences with low similarity to known references are often removed in environmental sequencing studies [130]. However, using sequence similarity as a filter criterion can remove too many QSs, particularly when studying new, unexplored environments [82]. By using phylogenetic placement as a filter instead, substantially more sequences can be retained for downstream analyses. Only the QSs that are placed onto the inner branches of the BT, that is, branches with no associated CT, are omitted at the second placement level.

Secondly, using specific clade trees for lower level taxonomic clades offers the phylogenetic resolution that is necessary for downstream analyses and for biological reasoning. It is, for example, possible to use manually curated "expert" trees for each clade of interest.

In this setup, the BT is only used for pre-classification, and can, for example, use our Automatic Reference Tree method. The aforementioned issue of obscured diversity in sub-clades can be circumvented by "overlapping" the CTs with the BT. That is, a CT can be associated with several branches of the BT, so that placements on each of these BT branches are collected and placed onto the same CT. See Figure 3.2 and Figure 3.7 for examples. We recommend to ensure that the branches of the BT that are associated with one CT are monophyletic, meaning that there is one split that separates these branches from the rest of the BT. This can be achieved by inferring the BT with a high-level constraint that maintains the monophyly of the CTs. It ensures phylogenetic consistency between the BT and the CTs, and improves the accuracy of the first placement level, as shown in Section 3.5.4. Lastly, it is also possible to use more than two levels, which might become necessary when working with RTs and datasets even larger than what is currently available.

3.4 Data Preprocessing for Phylogenetic Placement

Apart from the RT size, handling the sheer number of QSs also induces computational limitations for conducting phylogenetic placements. Most metagenomic studies publish their data in unprocessed formats, sometimes filtered to contain only reads from certain barcoding or marker regions. For instance, they store the raw sequencing output in `fasta` [111] or `fastq` [19] format. Those data often contain duplicates of exactly identical sequences, both *within* and *across* samples. Identical sequences are however treated the same in phylogenetic placement algorithms and therefore induce unnecessary computational overhead. Furthermore, sample sizes, that is, the number of sequences per sample, can vary by several orders of magnitude. If the placement algorithm is parallelized over samples, this leads to an uneven load balance across compute nodes.

In order to solve these issues, that is, reduce computational cost and achieve good load balancing, one can pre-process the sequences with our GAPPA tool. First, sequences are de-duplicated across all samples and fused into chunks of equal size. The chunk size should be chosen to allow aligning and placing a chunk within wall time on the intended hardware; we recommend chunk sizes of 50 000 or larger. Our tool assigns an identifier to each unique sequence, and computes a list of abundance counts for each sequence in a sample. Given an RT and its underlying alignment, the QS chunks are then aligned to the reference multiple sequence alignment, using programs such as PaPaRa [9, 10] or HMMALIGN [34, 35], and subsequently placed on the RT, for example by PPLACER, RAxML-EPA or EPA-NG [7, 11, 90]. The resulting per-chunk placement result files in combination with the per-sample abundance counts can then be parsed and analyzed by GAPPA to generate final per-sample placement files, containing a placement for each sequence in the original sample.

The speedup that can be gained via this preprocessing is proportional to the ratio of total versus unique sequences; the gain in parallel efficiency depends on the ratio of smallest to larges sample (in number of sequences). This approach allows to analyze datasets that are orders of magnitude larger than in previous published studies. For example, in 2012, an analysis of Bacterial Vaginosis (BV) data placed a total of 426 612 sequences, thereof 15 060 unique, on an RT with 796 tips [128]. Using a prototype of GAPPA, we were able to analyze a neotropical soils dataset with 50 118 536 total sequences, thereof 10 567 804 unique, with an RT comprising 512 taxa [82]. To demonstrate the scalability of our methods for this paper, we analyzed datasets with up to 116 520 289 total sequences, thereof 63 221 538 unique, from the HMP [51, 95], using RTs with up to 2059 tips. This corresponds to a computational effort that is four orders of magnitude greater than for the BV study.

3.5 Evaluation

To test the automatic reference tree (ART) method, we used the "SSU Ref NR 99" sequences of the Silva database [115] version 123.1 and the corresponding taxonomic framework [150]. The database contains 598 470 aligned sequences from all three domains of life, classified into 11 860 distinct taxonomic labels.

We constructed four sets of consensus sequences from the Silva database: a *General* set ("all of life"), as well as separate sets for the domains *Archaea*, *Bacteria*, and *Eukaryota*. For each set except the *Archaea*, the recursive expansion of taxonomic clades was applied to obtain approximately 2000 (*General*) and 1800 (*Bacteria*, *Eukaryota*) consensus reference sequences. This is large enough to cover the diversity well, while still being computationally feasible for the subsequent steps. The *Archaea* taxonomy in Silva is smaller, containing 248 taxa at *Genus* level, which is the lowest level in their taxonomy. Hence, the *Archaea* tree also comprises 248 taxa. Furthermore, in the three domain-specific trees, we included sequences at the *Phylum* level of the respective two other domains, to ensure that our methods also work with outgroups. The assembly of these four data sets required in total about 30 min and 10 GB of memory on a standard laptop computer. An overview of the tree sizes is shown in Supplementary Table 3.1. We then inferred constrained and unconstrained maximum likelihood trees for the consensus sequences. The constrained trees comply with the Silva taxonomy, and are used to assess how taxonomic constraints affect the phylogenetic placement and the subsequent analyses. Details are provided in Supplementary Section 3.5.3, which also discusses differences between the constrained and unconstrained trees. Details of the trees are shown in Supplementary Table 3.3; Figure 3.7 shows the unconstrained *Bacteria* tree as an example.

In total, our setup yields eight distinct RTs for evaluation: the *General* tree, the three domain trees, and the respective taxonomically constrained variants.

3.5.1 Accuracy

Here, we assess how using our ART affects phylogenetic placement accuracy. Each terminal branch of our RTs represents a consensus sequence, which is computed from species level sequences that share the same taxonomic label. We evaluate an RT by placing these species sequences onto the RT: Each species sequence is expected to be placed onto the branch leading to the consensus sequence that represents this particular species sequence. As the consensus sequences are derived from the taxonomy, all terminal branches of the tree have taxonomic labels. These labels thus identify the expected placement position for each species sequence. For example, sequences `S1-6` in Figure 3.1 are represented by the consensus sequence for the *Calyptosporidae* clade, which is shown below the 6 sequences in the Figure. They are thus expected to be placed onto the *Calyptosporidae* branch in the RT.

We placed the respective subset of the Silva database species sequences onto each of the eight RTs. We quantify placement accuracy for a sequence by the distance to its expected placement branch. More precisely, we measured (a) the (discrete) number of branches between the actual placement and the expected branch, and (b) the (continuous) distance in branch lengths units. As a sequence can have multiple placement locations, the distances, are, in fact, weighted averages incorporating the placement probabilities (likelihood weights). The results for the four unconstrained trees are shown in Figure 3.3; Figure 3.4 depicts the results for the constrained trees. Further details are provided in Supplementary Table 3.3.

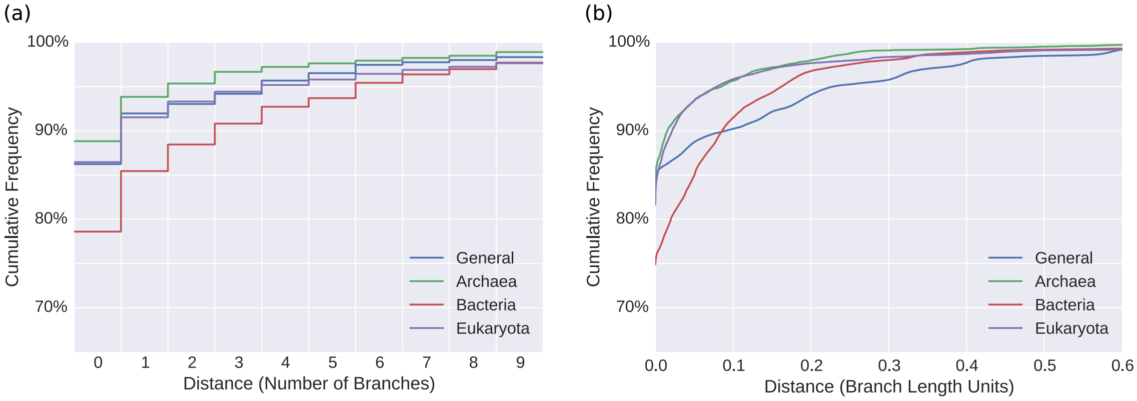

Figure 3.3: Weighted distances to expected edges for unconstrained trees. We evaluated the accuracy of our ARTs by placing sequences and measuring the weighted distances to their respective expected placement branches. The Figure shows the cumulative frequencies of number of sequences versus distances, measured (a) in number of branches and (b) in branch length units. In other words, it shows how many sequences are placed within a certain radius from their expected branches. For example, in (a), more than 85% of the sequences of the *Bacteria* (red) are placed within a radius of at most one branch from their expected branch, and in (b), more than 95% of the Eukaryota (purple) are within a radius of 0.1 branch length units from their expected branches.

Considering the size of the trees, most sequences are placed in close vicinity to their expected branches. This is corroborated by the short average distances reported in Supplementary Table 3.3. Furthermore, the average expected distance between placement locations [EDPL, 90] is low, indicating that the placements of a specific sequence mostly cluster in a small neighborhood of the tree. We observed that errors occur mostly in parts of the tree with short branches, which might be explained by the inability of 16S SSU sequences to properly resolve certain clades [53]. Also, the placement likelihood differences are small between neighboring, short branches, such that the placement signal is fuzzy.

With 77% of the sequences placed exactly on their expected branch, the accuracy is generally lowest for the *Bacteria* tree. This might be because the *Bacteria* have the most sequences in Silva, and exhibit a high diversity. In the other three trees, more than 90% of the sequences are placed at most one branch away from their respective expected branch. The constrained trees (Figure 3.4) exhibit similar placement accuracy. Particularly when using Multilevel Placement with overlapping RTs, placement differences of a few branches on the first level tree are acceptable, as they do not change the second level tree on which the sequence is placed. See Section TODO: sec:Results:sub:MultilevelPlacement for details.

As outlined in the method description, we represent clade diversity via majority rule consensus sequences. To assess the impact of the consensus method, we repeated the above evaluation, using two alternative consensus methods, but found little difference between the methods, see Figure 3.5. Furthermore, we also tested an automated approach that uses actual sequences from the database to represent the clades of the taxonomy, instead of using consensus sequences; see Figure 3.6. We however found that this approach yields less accurate trees.

3.5.2 Empirical Datasets

ARTs are intended for obtaining phylogenetic placements of environmental sequences. As the true evolutionary history of such sequences is unknown, we can not repeat the previous accuracy tests on empirical environmental datasets. Instead, we assess if the ARTs yield meaningful quantitative results for typical post-analysis methods. To this end, we placed two empirical metagenomic amplicon barcoding datasets on our unconstrained *Bacteria* tree. To asses the placement results obtained from the ART, we performed Squash Clustering and Edge PCA [87] post-analyses on the placement results (see A and Figure 3.8 and Figure 3.9 for details). The results reveal that the ART reproduces results of previous studies based on custom RTs with manually selected reference sequences. Furthermore, the ART is able to classify samples (e.g., healthy vs sick patients), at least to the extend that is expected from its phylogenetic resolution. That is, samples that only differ in placements at the species level cannot be classified using a broad, high-level tree such as our *Bacteria* tree. In order to obtain finer taxonomic resolution, it is thus necessary to either use an ART that contains more taxa, or to use our multilevel approach instead (see next Section).

3.5.3 Details about the Automatic Reference Tree Evaluation

To test the automatic reference tree (ART) method, we used the "SSU Ref NR 99" sequences of the Silva database [115] version 123.1 and the corresponding taxonomic framework [150], which are available at http://www.arb-silva.de. The database contains 598 470 aligned sequences from all three domains of life, classified into 11 860 distinct taxonomic labels, and mainly contains bacterial sequences. In detail, there are

- 22 913 sequences with 347 taxonomic labels for the *Archaea*,
- 62 436 sequences with 7441 taxonomic labels for the *Eukaryota*, and
- 513 121 sequences with 4072 taxonomic labels for the *Bacteria*.

The overall number of taxonomic labels is counted here, that is, it includes higher level labels.

As explained in the main text, we constructed four sets of consensus sequences from these data using our ART method: a *General* set ("all of life"), as well as separate sets for the domains *Archaea*, *Bacteria*, and *Eukaryota*. The assembly of the four data sets with our method required in total about 30 min and 10 GB of memory on a standard laptop computer. This includes counting alignment characters, calculating entropies and constructing consensus sequences. The resulting data set sizes and the fraction of sequences from each domain the ARTs contain are shown in Table 3.1.

Our implementation of the method furthermore contains some details that are worth mentioning for reproducibility: It is possible to constrain the maximal size of clades in order to not build a consensus sequence for an overly large clade, which might not be a good representative of that clade. For the same reason, it is possible to first expand the highest ranks of the taxonomy into separate candidates. We used conservative values for these two constraints (a maximal clade size of 2000 and an expansion of only the first two taxonomic ranks), in order to give more weight to the sequence entropy. Lastly, some clades contain only one sub-clade. Those were immediately expanded, as they do not change the length of the candidate list during the algorithm.

We then inferred unconstrained and constrained maximum likelihood trees on the sequences, running 50 independent tree searches for each tree and selecting the best-scoring tree. Unconstrained trees were inferred using RAxML 8.2.8 [129]. Constrained trees were inferred with Sativa 0.9-55 [62], which internally again relies on RAxML, and offers a convenient way to transform a taxonomy into a constraint tree.

The relative Robinson-Foulds distances [117] between the four pairs of trees (constrained versus unconstrained) are between 45.8% and 49.7%. The differences between the trees however mostly concern inner branches. As QSs generally tend to be placed more towards the terminal branches of the tree, the differences in the inner branches thus are acceptable for our evaluation purposes. Furthermore, we performed significance tests comparing the constrained trees to the unconstrained ones, as shown in Table 3.2. The tests show that in all cases, the unconstrained trees fit the data significantly better.

Given these eight ARTs, the evaluation was conducted as explained in the main text. As the sequences in Silva are already aligned to each other, no alignment step was necessary. As they contain no phylogenetic signal, we removed sites consisting entirely of gaps from the alignment, in order to reduce the memory footprint of downstream steps. Phylogenetic placement was conducted using EPA-ng [7], which is substantially faster and more scalable than RAxML-EPA [11] and pplacer [90].

We evaluated the accuracy of the placements using the taxonomic labels of the sequences in Silva as an indicator of the expected branch of each sequence. Thus, we have to assume the taxonomic label of each sequence to be correct. However, errors are expected due to incongruity between the taxonomy and the phylogeny [100], as well as due to taxonomically mislabeled sequences [62]. For example, Sativa [62], found 9934 mislabeled sequences in the Silva database. Furthermore, 17 452 sequences contain one of “incertae”, “unclassified” or “unknown” in their name, indicating that those sequences might not be reliable. In total, there are 25 910 (or 4.3%) such dubious sequences in version 123.1 of the Silva database. Not all sequences are hence expected to be placed on their expected branches. We also evaluated how these dubious sequences affect the accuracy of the trees. To this end, we used the same four trees as before (that is, they were constructed with all sequences, including the dubious sequences), but for the evaluation step excluded the dubious sequences.

That is, those sequences were not placed on the trees, and their distance to the expected branch was not used for the evaluation. In most cases, this improved the results slightly (data not shown). However, we decided to only report the unfiltered results.

3.5.4 Sub-clades and Multilevel Placement

We selected five bacterial clades to evaluate ART accuracy on smaller clades, as well as to assess some properties of the Multilevel Placement approach. Figure 3.7 shows the *Bacteria* tree with the five test clades highlighted.

First, using the sequences and taxonomies of these five clades, we built unconstrained and constrained ARTs. We then conducted the same accuracy analysis as explained before on these ten trees. That is, we placed the Silva sequences of the five clades onto their respective ART and evaluated distances to expected branches. Thereby, we evaluated the accuracy of these ARTs when used as second level clade trees. The results are shown in Figure 3.10. The placement accuracy is slightly worse for the clade trees than for the eight comprehesive ARTs evaluated before. This is again likely due to 16S SSU sequences being unable to properly resolve lower taxonomic levels [53].

Next, using the five clades, we evaluated the accuracy of the first placement level when conducting Multilevel Placement. So far, our evaluation focused on the distance from a sequence placement to its expected placement branch. For the first placement level on a backbone tree (BT), it is however more important that a sequence is placed into the correct clade. Thus, we used the unconstrained *Bacteria* BT again, and assessed how many sequences were placed in the clades shown in Figure 3.7. Of the 450 313 sequences in Silva in these clades, 98.0% were placed (most likely placement) into a branch of their corresponding clade. Thus, for multilevel placement, they will be assigned to the correct second level clade tree (CT). More specifically, the *Firmicutes* perform worst, as only 94.7% of the *Firmicute* sequences are placed into the corresponding clade. This can be explained by the high amount of paraphyletic branches of this clade, cf. Figure 3.10, which is a known issue [110]. The sequences of the other four clades we tested achieve a clade identification accuracy exceeding 99%.

As mentioned before, a high-level taxonomic constraint can improve the accuracy of placing a sequence into the correct BT clade. To show this, we inferred the *Bacteria* RT again, but used a *Phylum* level constraint that separates the five clades from each other and from the rest of the tree. All branches within the clades were resolved using maximum likelihood. The tree (not shown) is similar to the tree in Figure 3.7, but all five clades are now monophyletic. Using this tree, 99.3% of the sequences were placed into the correct clade. Particularly the accuracy for *Firmicutes* improved, yielding an accuracy of 99.5%.

Overall, our experiments show that the first level placement is highly accurate, even if an extremely diverse “all bacteria” backbone tree is used. The accuracy on the second level is slightly worse when using ARTs as CTs.

3.6 Conclusion and Outlook

We presented algorithms and software tools to facilitate and accelerate phylogenetic placement of large environmental sequencing studies.

The automatic reference tree (ART) method provides a means for automatically obtaining suitable reference trees by using the taxonomy of large sequence databases. Using the Silva database as a test case, we showed that it can be applied for accurately (pre-)placing environmental sequences into taxonomic clades. The method can also be used for rapid data exploration in environmental sequencing studies: An ART might be useful to obtain an overview of the taxa that are necessary to capture the diversity of a sequence dataset, without the substantial human effort and potential bias of manually selecting reference sequences. To capture clade diversity with finer resolution, for example for a second placement level, clade-specific ARTs can be inferred. If species-level resolution is required, we recommend that the sequences are inspected by an expert, in order to make sure that the tree is suitable for the dataset to be placed on it. Furthermore, as our automated approach inevitably suffers from errors in the database it is based on, we recommend using Sativa [62] to identify potentially mislabeled sequences in the database. One should also keep in mind that phylogenetic placement does not necessarily provide resolution at the species level [32].

As we show, our multilevel placement method as well as the preprocessing pipeline accelerate the placement process without sacrificing accuracy. By first placing the query sequences on a broad backbone tree (BT), as described in the method, novel environments with sequences of unknown evolutionary origin can be classified without having to process a large tree comprising all taxa of interest. A second placement on a set of clade trees (CTs) provides sufficient resolution for biological interpretation. Placement accuracy can be further improved by inferring the BT with a high-level constraint that separates the clades of the CTs from each other and thus ensures monophyly of these clades.

The methods presented here are implemented as part of our gappa tool, which is freely available under GPLv3 at http://github.com/lczech/gappa (see B for an overview of the corresponding commands). All scripts and data used for this paper are available at http://github.com/lczech/placement-methods-paper.

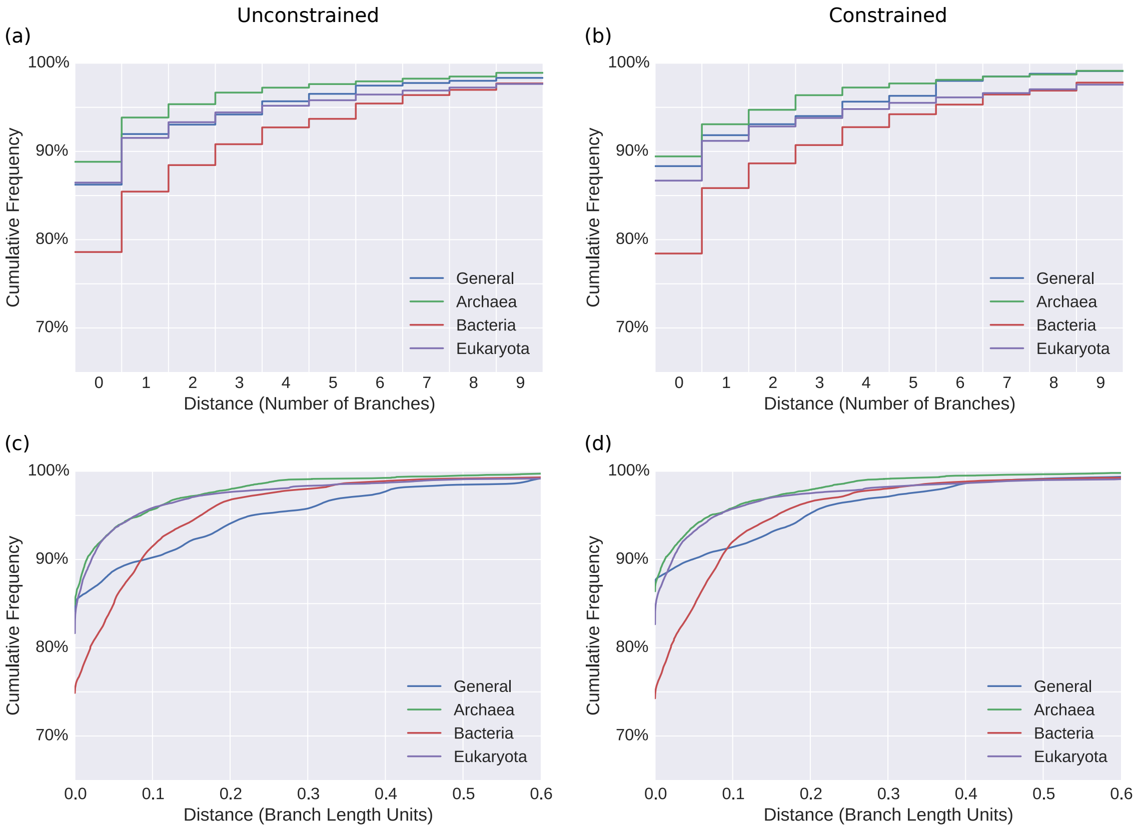

Figure 3.4: Accuracy comparison between unconstrained and constrained trees. Here, we compare how a taxonomic constraint changes the weighted distances to correct edges when placing our evaluation sequences on Automatic Reference Trees. The evaluation method is explained in the main text. Subfigures (a) and (c) are identical to Figure 3.3 of the main text and included here for ease of comparison. Subfigures (b) and (d) show the results when using the SILVA taxonomy as constraint for the tree inference. The relative Robinson-Foulds distances [117] between the four pairs of trees range between 45.8% and 49.7%. The differences probably occur because our trees span diverse clades, whose ancient branches are hard to resolve. Also, single gene data might not be sufficient to resolve these clades. See Supplementary Section 3.5.3 and Table 3.2 for a more detailed comparison of the constrained and unconstrained trees. However, the differences mostly concern inner branches. Most of the placements are however expected to be near the terminal branches, which are more stable across the trees.

This is confirmed by the fact that, overall, the results are similar between the unconstrained and constrained trees. A slight improvement can be observed for the constrained General tree (blue), which performs better according to both distance measures. However, when considering only the distance of the most likely placement (highest LWR) to its correct edge instead of using average distances weighted by the LWR per QS, the constrained trees consistently yield better results (data not shown). For example, the most significant change is observed for the Eukaryota tree, with 84% correct placements for the unconstrained tree, but 89% for the constrained one. We suspect that this is an artifact of our evaluation process, as we consider a sequence to be correctly placed if the placement branch belongs to the consensus sequence to which the sequence contributed. As the selection of sequences for each consensus sequence is guided by the taxonomy, using the same taxonomy as constraint for the tree thus might also improve the placement accuracy.

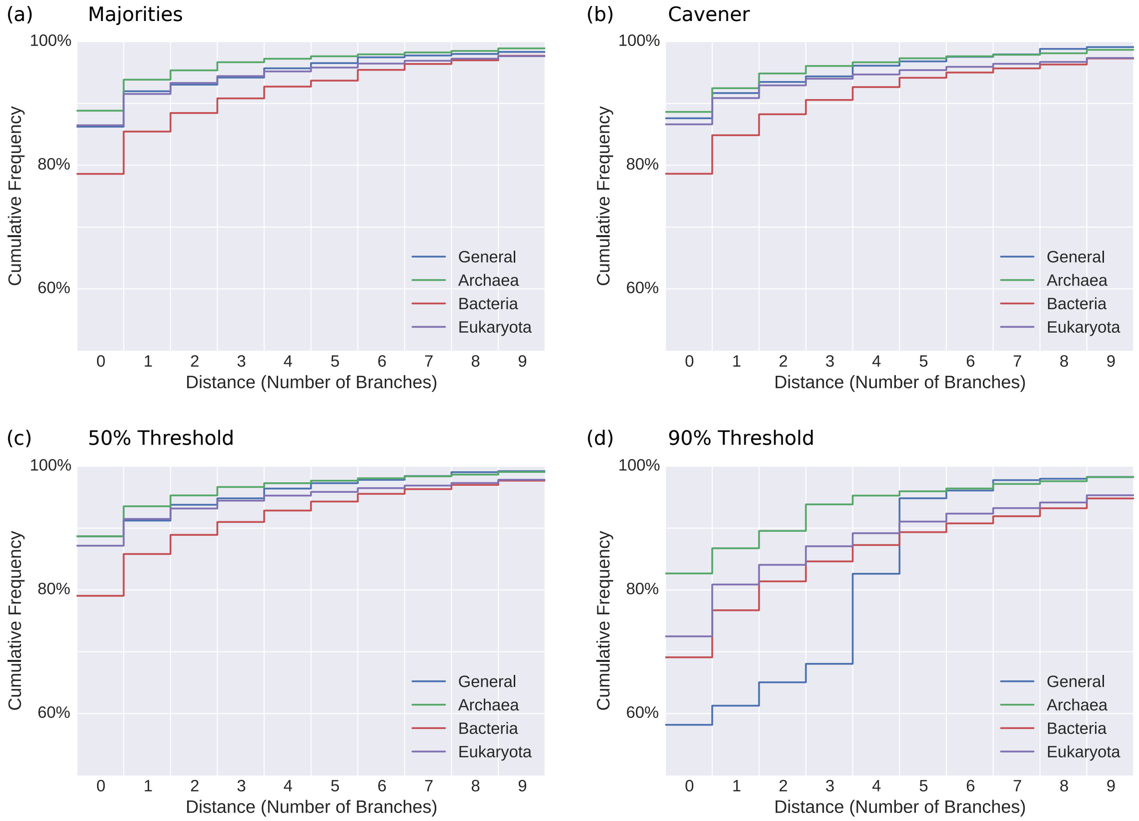

Figure 3.5: Effect of different consensus sequence methods on placement accuracy. In the main evaluation of our Automatic Reference Tree method, we used a reference tree and alignment based on majority rule consensus sequences [27, 93] of the Silva database sequences. Here, we evaluate the effect of using other consensus sequence methods on phylogenetic placement accuracy. The evaluation method is described in the main text. In addition to (a) majority rule consensus, we tested (b) Cavener's method [17, 18], as well as threshold consensus sequences [26, 27] using thresholds of 50%, 60%, 70%, 80%, and 90%, of which two are shown in (c) and (d). The three remaining threshold methods exhibit accuracies almost exactly in between the shown plots, that is, accuracy decreases with increasing thresholds. For comparison, we also included Figure 3.3(a) of the main text again, here as Subfigure (a), using the same y-axis scaling as the other plots. We only show distances measured in number of branches here, because this is more relevant in the context of our methods (e.g., Multilevel Placement).

By using alternative consensus methods, the consensus sequences and thus the sites in the alignments change. Hence, the obtained reference trees (not shown) differ substantially from each other. Across the corresponding trees of the consensus methods tested here, we observed an average relative RF distance of 49.5%. This is similar to our findings depicted in in Figure 3.4. Here too, the accuracy of the constrained variants of these trees (data not shown) does not change much compared to the accuracy obtained for the unconstrained trees shown here. Thus, the differences in accuracy seen here are most likely due to the interplay of alignment and placed sequences (which is what we are interested in), and not due to differences in the trees (which are not of interest here).

The first three plots (a)-(c) exhibit similar accuracies. On average, majority rule, Cavener's, and low threshold ($\leq$ 70%) consensus methods place 82-83% of the sequences on the expected branch. As a general trend, the *Archaea*, being the smallest tree, tend to have the highest accuracy. On the other hand, the *Bacteria*, having the most sequences in Silva, score worst. This changes for high consensus thresholds. At high thresholds, many sites contain ambiguity characters, thus blurring the phylogenetic signal. The *General* tree, representing the highest diversity, is most

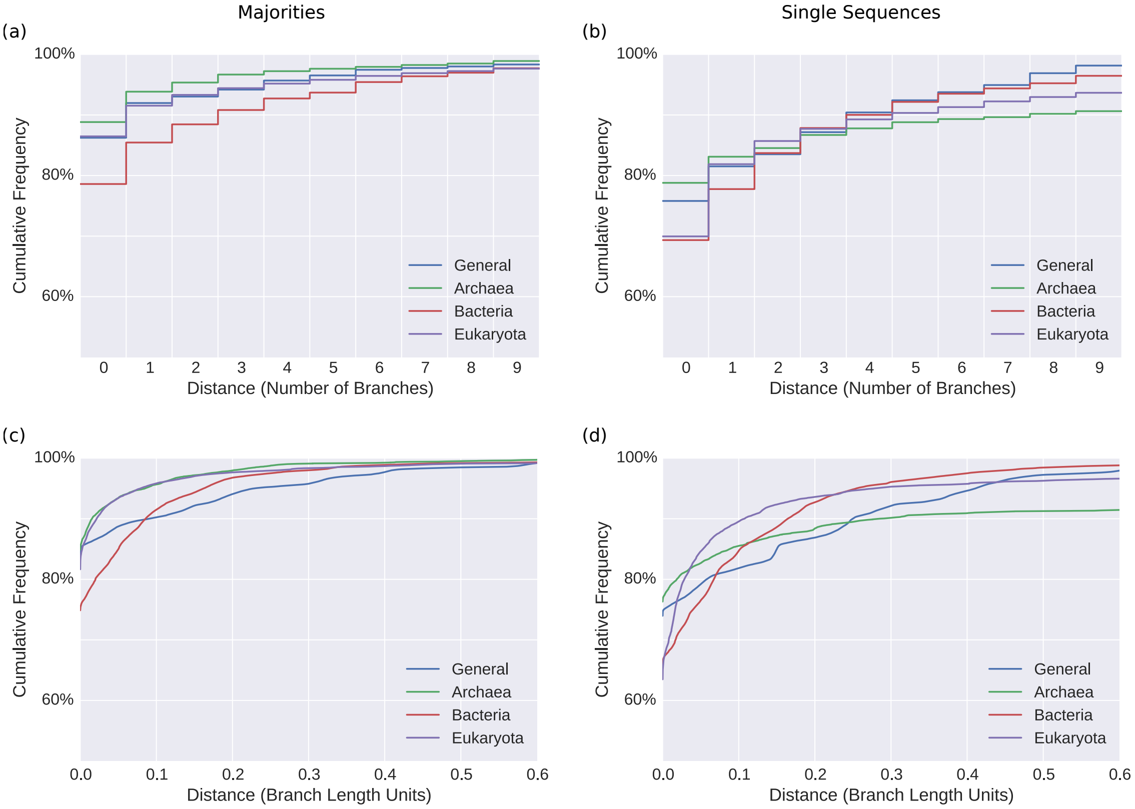

Figure 3.6: Effect of using actual sequences (instead of consensus sequences) on placement accuracy. For most of our evaluation of the Automatic Reference Tree method, we used some form of consensus sequence representation for the clades of the taxonomy, see e.g., Figures 3.4 and 3.5. However, we also tested how the method behaves when using actual sequences instead, thus avoiding to unnecessarily blur the phylogenetic signal, and other potential drawback of consensus sequences.

As manually selecting representative sequences from the database was not practical, we used the following automated approach. First, we took the 90% threshold consensus sequences of the ART method that were already evaluated in Figure 3.5(d). By using a high threshold, most of the diversity of each clade is included. Then, for each such consensus sequence, we calculated a score for all sequences from the database that were used to construct this consensus sequence. This score is the number of different nucleotides between the consensus sequence and the database sequence. The sequence with the lowest score (that is, with most matching nucleotides) was then used as representative of the clade. Thereby, the taxonomic clades are represented by actual sequences from the database. However, as these sequences are close to the respective consensus sequence, they are still good representatives of the diversity of the clade. Using this set of sequences, we then again inferred a tree and conducted the evaluation procedure by placing all sequences of the database on that tree, as described in the main text.

Subfigures (a) and (c) are identical to Figure 3.3 of the main text and included here for ease of comparison, however with the y-axis scaled to fit the remaining subfigures. That is, they show the evaluation of the majority rule consensus sequences. Subfigures (b) and (d) show the evaluation of the approach as described here. The resulting accuracy is worse in all cases. That is, on average, the sequences were placed further from their respective expected branch. We suspect that this is because single sequences do not capture the diversity of their clade as well as consensus sequences, and because they do not incorporate as much biological information (e.g., in form of ambiguity characters).

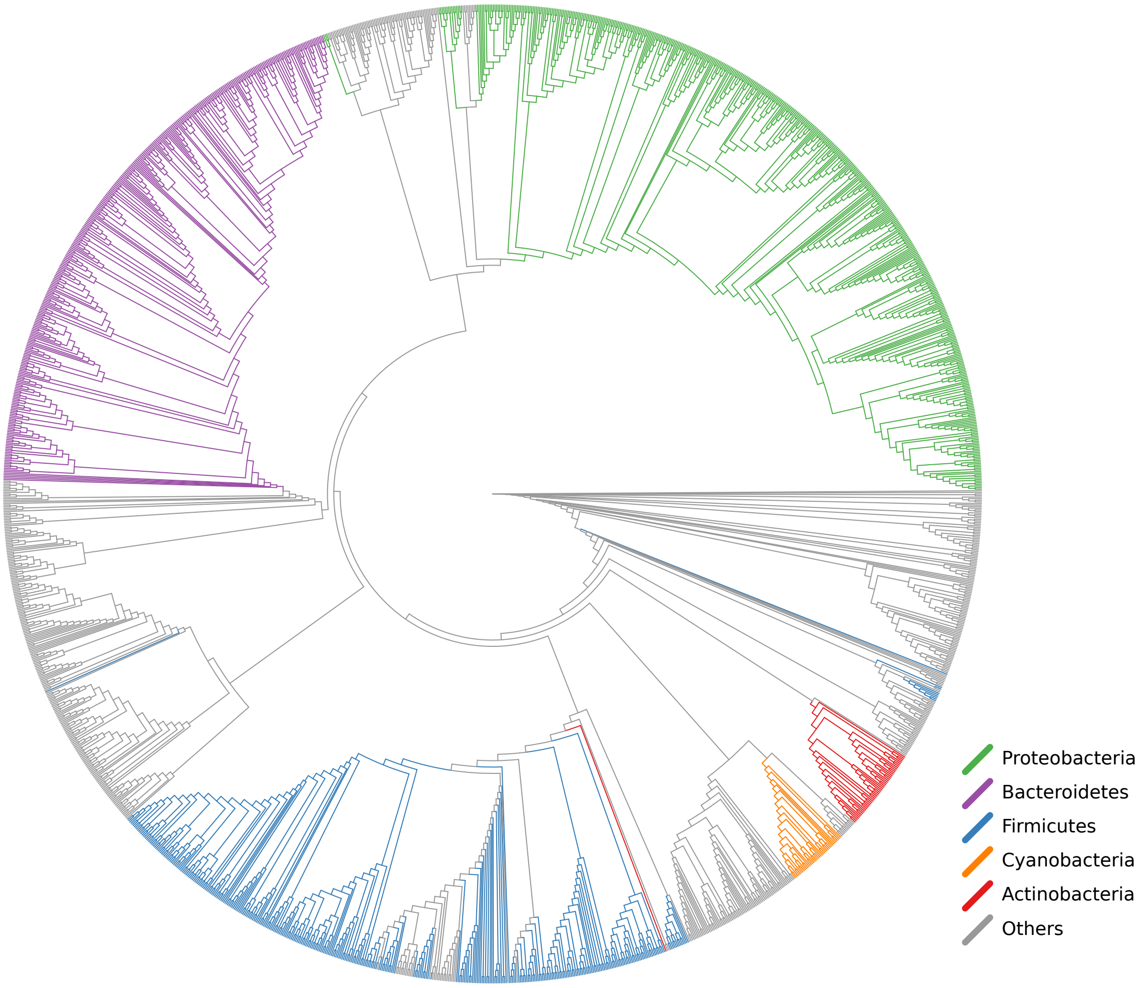

Figure 3.7: Unconstrained *Bacteria* tree with five bacterial sub-clades. This tree is the result of our Automatic Reference Tree method applied to the *Bacteria* sequences in Silva. The tree contains a total of 1914 taxa. Colorized are the five *Phylum* level sub-clades that we used for testing multilevel placement: *Proteobacteria* (505 taxa), *Bacteroidetes* (362 taxa), *Firmicutes* (360 taxa), *Cyanobacteria* (39 taxa) and *Actinobacteria* (53 taxa). The incongruence between taxonomy and phylogeny is visible here as non-monophyletic colored branches. We thus here define a clade to consist of all branches that are part of a monophyletic split of the tree with respect to the taxa in the clade. In other words, all branches on one side of a split are considered to belong to a clade, if that side of the split only contains taxa from that clade. These branches then receive the same color here. Then, for multilevel placement, a sequence is considered to be part of a clade if its most probable placement falls into that clade. For example, a sequence that is placed onto one of the orange branches on this tree is subsequently placed in the *Cyanobacteria* tree for the second level placement. Each of the five sub-clades is represented by multiple branches here, which we call the "overlap" with the *Bacteria* tree.

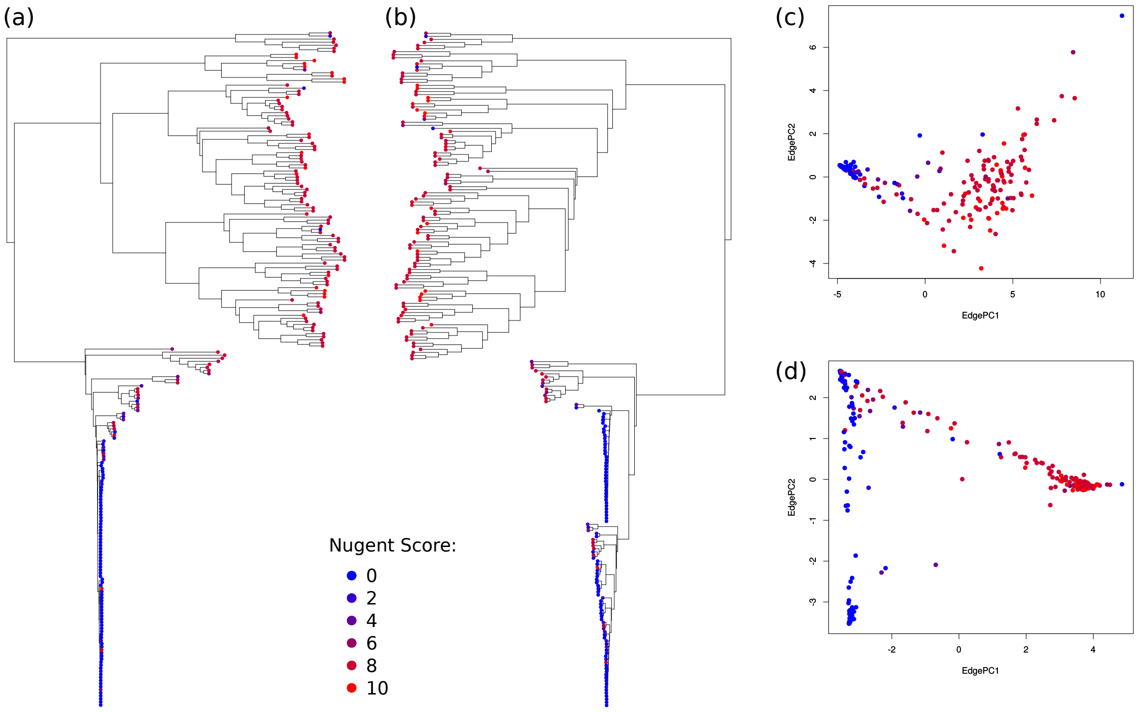

Figure 3.8: TODO: this is small. lets see if this needs to stay small! **Assessment of an Automatic Reference Tree for conducting Squash Clustering and Edge PCA.** Here, we test the behavior of the unconstrained *Bacteria* tree obtained via our ART method when used for standard post-analyses of phylogenetic placement results. For this, we used a sequence dataset of the vaginal microbiome of 220 women [128]. For details on the processing, see A. The original study showed associations between the presence of certain bacterial species and the diagnosis of Bacterial Vaginosis (BV), a condition caused by changes in the vaginal microbiome. In the study, the Nugent score [106] was used as a clinical diagnostic criterion for BV, which ranges from 0 (healthy) to 10 (severe illness). We placed the sequences of the dataset on their original tree and on our ART, and reproduce some of the results from the original study to assess differences induced by using distinct references trees.

Squash Clustering is a hierarchical clustering method for phylogenetic placement data [87] that uses the phylogenetic Kantorovich-Rubinstein (KR) distance [39] to measure cluster similarity. The result of Squash Clustering is a cluster tree of samples, where samples that are similar according to the KR distance are closer to each other, with branch lengths according to that distance. The left side of the figure compares the cluster trees resulting from using (a) our tree and (b) the original reference tree. Subfigure (b) is a recalculation of Figure 1(A) of [128]. The tips, which correspond to samples, are colored by the Nugent score of each sample, and thus indicate which women are affected by BV. The general features of the two cluster trees are comparable, indicating that our tree is able to distinguish between healthy and sick patients. However, there is a major difference in the lower half of the trees: While (b) shows some small branch lengths and even a separated sub-clade of samples with low Nugent score, these branches have a length of virtually zero in (a). As shown in [128], the healthy patients are divided into two classes, based on the presence of two species of *Lactobacillus*. The original reference tree contains sequences of those species, and can thus distinguish between them. Our broad *Bacteria* tree however does not have this degree of species-level resolution and thus treats them the same, yielding a negligible KR distance between the samples. Although this finding is expected, it serves as an example for the limits of our method. TODO: missing text!! dimensions too large!

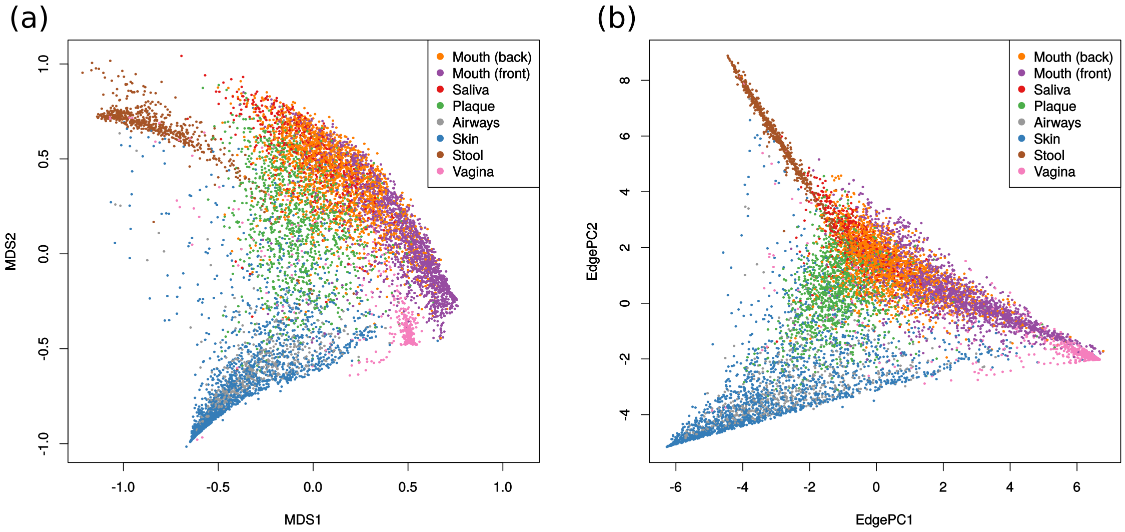

Figure 3.9: Assessment of an Automatic Reference Tree for large dataset analyses. Here, we test the unconstrained *Bacteria* tree generated by our Automatic Reference Tree method for placing and analyzing a large sequence dataset. For this, we use the Human Microbiome Project (HMP) [51, 95] data, and selected 9192 samples from different body sites with a total of 117 million sequences. For details on the processing, see A. We categorized the 19 original body site labels into 8 regions, in order to make the plot more readable. See Table 3.4 for the mapping between the original labels and the ones used here. The sequences were placed on the tree, and subsequently analyzed with two different methods. Both subfigures show that the tree, despite only representing higher taxonomic levels, suffices to separate different body site regions from each other.

In subfigure (a), we visualized the pairwise KR distances between all samples. The phylogenetic Kantorovich-Rubinstein (KR) distance [39, 87] between two samples, also known as the Earth Mover's distance, measures how much placement "mass" has to be moved by how much along the branches of the tree in order to transform one sample into the other. It is a generalization of the UniFrac distance [75, 77]. The high-dimensional pairwise KR distance matrix was then embedded into the plot by performing Multidimensional scaling (MDS). MDS [40, 64, 84] is a dimensionality reduction technique that finds an embedding of a distance matrix into lower dimensions (in this case, 2 dimensions) preserving higher dimensional distances as well as possible.

In subfigure (b), we also perform Edge PCA [87] on the samples. The grouping of body sites is again clearly visible with this method. Similar to Figure 3.8(d), the plot forms a triangle of samples, roughly separated by body site regions.

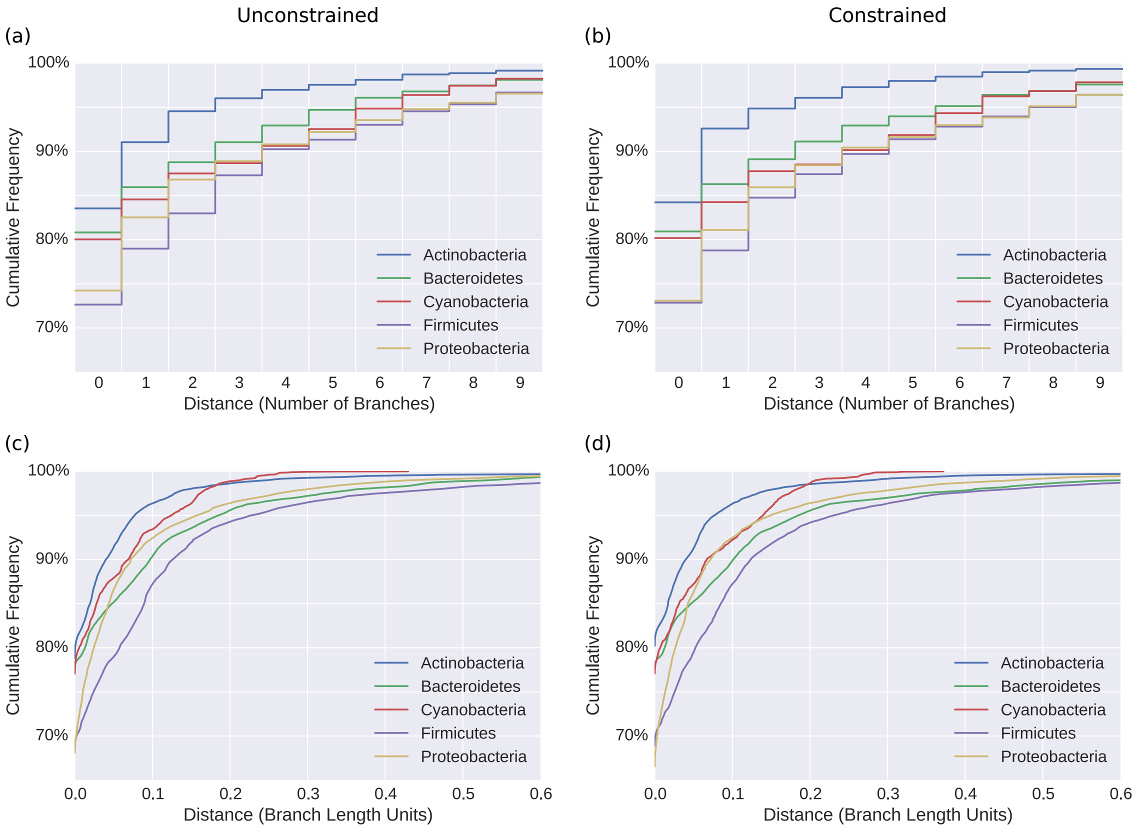

Figure 3.10: Accuracy of the Automatic Reference Trees of five bacterial sub-clades. We used five sub-clades of the *Bacteria* in Silva, which were already scrutinized in [62], to test how our Automatic Reference Tree method works for less diverse sets of sequences. These five clades are also highlighted in Figure 3.7; see there for a description of the clades. The evaluation was conducted as explained in the main text; in short, we placed the Silva sequences of the clades on their respective tree, and measured how far each of them is away from the branch of the consensus sequence it is represented by.

The placement accuracy on these sub-clades is slightly worse compared to the broad *Bacteria* tree, which can be seen by comparison to Figure 3.4. On average, 73.4% of the sequences were placed exactly on their expected branch, dominated by *Proteobacteria* and *Firmicutes*, which combined make up 75% of the sequences in the five clades, and have an accuracy of 71%. The *Actinobacteria* have the highest accuracy, with 82% of their sequences placed on the expected branch.

The two smallest clades, *Actinobacteria* and *Cyanobacteria*, exhibit the shortest distances in branch length units. In fact, the longest distance of any sequence from its expected branch in the *Cyanobacteria* clade is around 0.4, which is indicated by the end of the red line in the lower two plots. On the other hand, the *Firmicutes* generally have the lowest accuracy here. In Figure 3.7, which shows the unconstrained *Bacteria* tree, the *Firmicutes* clade exhibits many paraphyletic branches, which is a known issue [110]. This indicates that there is a high incongruence between the *Firmicutes* taxonomy and phylogeny in Silva, which might explain why the *Firmicutes* score worst here.

These results are likely due to the inability of 16S SSU sequences to properly resolve lower taxonomic levels [53, 97, 113]. For example, Table 2 of [53] lists 10 bacterial genera that are known to be hard to identify using 16S sequences. These genera account for 7.9% of the 2846 taxa that are represented by the five bacterial trees tested here. Furthermore, 95 553 of the 450 313 sequences that were placed on those trees (21.2%) belong to one of these genera. This might explain the worse scores of these clade trees. Lastly, the consensus sequences at the tips of the trees represent the *Genus* level. Thus, these have short branches, which increases the probability

Table 3.1: Taxonomic composition of the four Automatic Reference Trees. The table lists the four trees and their sizes (in number of tips), as well as how many of these tips originate from each of the three domains of life. The target size of the *General* tree was 2000 taxa, while the *Bacteria* and *Eukaryota* tree were targeting 1800 domain-specific taxa, which is approximately reached, but not exactly (underlined values). This is because the sizes of sub-clades in the taxonomy vary. Because each tip of the tree is a consensus sequence that represents the respective lowest taxonomic level, the number of available taxa is smaller than the total number of taxonomic labels in the Silva database. For example, the *Archaea* have a total of 347 taxonomic labels across all ranks, but only 248 labels at *Genus* level. Thus, the *Archaea* tree shown here represents the *Archaea* taxonomy resolved at the *Genus* level. In the three domain specific trees, we furthermore included consensus sequences at the *Phylum* level of the respective two remaining domains, in order to make sure that the evaluation also works well if such "outgroups" are included.

		Thereof number of		
Tree	Size	*Archaea*	*Bacteria*	*Eukaryota*
General	1998	210	508	1280
Archaea	511	248	205	58
Bacteria	1914	59	1797	58
Eukaryota	2059	59	205	1795

Table 3.2: Tree Topology Significance Tests. Here, we report significance tests comparing the four pairs of unconstrained (U) and constrained (C) trees used in our evaluation. The tests were performed with IQ-TREE v1.5.6 [104] under the GTR+G model and 10 000 resamplings using the RELL method [60]. The table shows that the unconstrained trees fit the data significantly better in all four cases and in all tests.

Columns are as follows. logL and deltaL: log likelihood and difference between constrained and unconstrained tree. bp: bootstrap proportion using RELL method [60]. p-(W)KH: p-value of the one sided and the weighted Kishino-Hasegawa test [59]. p-(W)SH: p-value of the (weighted) Shimodaira-Hasegawa test [125]. c-ELW: Expected Likelihood Weight [131]. p-AU: p-value of approximately unbiased (AU) test [124].

TODO: table size! set to small now, but maybe make the page landscape instead?!

Tree	logL	deltaL	bp	p-KH	p-WKH	p-SH	p-WSH	c-ELW	p-AU
General (U)	-725199.040		1.0	1.0	1.0	1.0	1.0	1.0	0.9987
General (C)	-731949.568	6750.528	0.0	0.0	0.0	0.0	0.0	0.0	0.0012
Archaea (U)	-131862.815		1.0	1.0	1.0	1.0	1.0	1.0	1.0000
Archaea (C)	-133110.463	1247.648	0.0	0.0	0.0	0.0	0.0	0.0	0.0000
Bacteria (U)	-405028.378		1.0	1.0	1.0	1.0	1.0	1.0	1.0000
Bacteria (C)	-412464.820	7436.442	0.0	0.0	0.0	0.0	0.0	0.0	0.0000
Eukaryota (U)	-745442.969		1.0	1.0	1.0	1.0	1.0	1.0	1.0000
Eukaryota (C)	-753944.998	8502.030	0.0	0.0	0.0	0.0	0.0	0.0	0.0000

Table 3.3: Overview of the Automatic Reference Trees and their evaluation statistics. Details of four unconstrained (U) and four constrained (C) trees are shown. "Size" is the number of leaves of the tree, that is, the number of consensus sequences that the tree was inferred from. "% Seqs." the percentage of sequences from Silva placed on it. The *General* tree does not cover all sequences, because there are some sequences labels in the database that could not be mapped to the taxonomy. "∅ Br. Len." is the average branch length in the tree. The evaluation results are reported in the remaining columns: Average distances of the sequences to their respective expected branch are listed in numbers of branches (Discrete) and in branch length units (Continuous), as explained in the text. Furthermore, "Exp. Br. Hits" shows how often the most probable placement was placed exactly on the expected branch. Lastly, the average expected distance between placement locations (EDPL) is shown. The EDPL is the sum of the distances between the placements of a sequences weighted by their probability [90].

				Average Distance			
Reference Tree	Size	% Seqs.	∅ Br. Len.	Discrete	Continuous	Exp. Br. Hits	∅ EDPL
General (U)	1998	98.7%	0.084	0.63	0.034	85.9%	0.00058
General (C)	1998	98.7%	0.086	0.57	0.027	88.2%	0.00046
Archaea (U)	511	3.4%	0.070	0.46	0.013	86.4%	0.00038
Archaea (C)	511	3.4%	0.071	0.45	0.013	88.2%	0.00041
Bacteria (U)	1914	84.6%	0.067	1.13	0.031	77.0%	0.00095
Bacteria (C)	1914	84.6%	0.071	1.11	0.031	76.6%	0.00091
Eukaryota (U)	2059	10.0%	0.080	0.79	0.022	84.9%	0.00032
Eukaryota (C)	2059	10.0%	0.083	0.81	0.024	85.7%	0.00031

Table 3.4: HMP Dataset Overview. The table lists the 19 body site labels used by the Human Microbiome Project (HMP) [51, 95]. We used this dataset to evaluate the applicability of typical analysis methods for phylogenetic placement using our Automatic Reference Trees, see A and Figure 3.9 for details. In order to simplify the visualization in Figure 3.9, we summarized some of the labels into eight location regions, as shown in the second column. The last column lists how many samples from each body site were used in our evaluation.

Body Site	Region	Samples
Tongue Dorsum	Mouth (back)	610
Palatine Tonsils	Mouth (back)	599
Throat	Mouth (back)	638
Attached Keratinized Gingiva	Mouth (front)	600
Hard Palate	Mouth (front)	566
Buccal Mucosa	Mouth (front)	597
Saliva	Saliva	529
Supragingival Plaque	Plaque	608
Subgingival Plaque	Plaque	595
Anterior Nares	Airways	541
Left Retroauricular Crease	Skin	596
Right Retroauricular Crease	Skin	604
Left Antecubital Fossa	Skin	290
Right Antecubital Fossa	Skin	328
Stool	Stool	600
Vaginal Introitus	Vagina	292
Mid Vagina	Vagina	298
Posterior Fornix	Vagina	301
Sum		9192

4. Visualization

This chapter is based on the peer-reviewed publication:

- **Lucas Czech** and Alexandros Stamatakis. ”Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements.” *PLOS One*, 2018.

Contributions: Lucas Czech... Alexandros Stamatakis...

4.1 Motivation

A first step in analyzing phylogenetic placement data is often to visualize them. For small samples, it is possible to mark individual placement locations on the RT, as offered for example by iTOL [66], or even to create a tree where the most probable placement per QS is attached as a new branch, as implemented in the guppy tool from the pplacer suite [90], RAxML-EPA [11, 129], and our tool gappa. For larger samples, one can alternatively display the per-edge placement mass, either by adjusting the line widths of the edges according to their mass, or by using a color scale, as offered in ggtree [151], guppy, and gappa. Using per-edge colors corresponds to binning all placement of an edge into one bin. For large datasets, the per-edge masses can vary by several orders of magnitude. In these cases, it is often preferable to use a logarithmic scaling, as shown in [82].

These simple visualizations directly depict the placement masses on the tree. When visualizing the accumulated masses of multiple samples at once, it is important to chose the appropriate normalization strategy for the task at hand. For example, if samples represent different locations, one might prefer to use normalized masses, as comparing relative abundances is common for this type of data. On the other hand, if samples from the same location are combined (e.g., from different points in time, or different size fractions), it might be preferable to use absolute abundances instead, so that the total number of sequences per sample can be visualized.

The visualizations provide an overview of the species abundances over the tree. They can be regarded as a more detailed version of classic abundance pie charts. When placing OTUs, or ignoring sequence abundances, the resulting visualizations depict species diversity. Moreover, these visualizations can be used to assess the quality of the RT. For example, placements into inner branches of the RT may indicate that appropriate reference sequences (i) have not been included or (ii) are simply not yet available.

Here, we introduce visualization methods that highlight (i) regions of the tree with a high variance in their placement distribution (called **Edge Dispersion**), and (ii) regions with a high correlation to meta-data features (called **Edge Correlation**).

4.2 Edge Dispersion

The Edge Dispersion is derived from the edge masses or edge imbalances matrix by calculating a measure of dispersion for each of the matrix columns, for example the standard deviation σ. Because each column corresponds to an edge, this information can be mapped back to the tree, and visualized, for instance, via color coding. This allows to examine which edges exhibit a high heterogeneity of placement masses across samples, and indicates which edges discriminate samples. As edge mass values can span many orders of magnitude, it might be necessary to scale the variance logarithmically. Often, one is more interested in the branches with high placement mass. In these cases, using the standard deviation or variance is appropriate, as they also indicate the mean mass per edge. On the other hand, by calculating the per-edge Index of Dispersion [40], that is, the variance-mean-ratio $^{\sigma^2}/_{\mu}$, differences on edges with little mass also become visible. As Edge Dispersion relates placement masses from different samples to each other, the choice of the normalization strategy *is* important. When using normalized masses, the magnitude of dispersion values needs to be cautiously interpreted [74]. The Edge Dispersion can also be calculated for edge imbalances. As edge imbalances are usually normalized to $[-1.0, 1.0]$, their dispersion can be visualized directly without any further normalization steps. An example for an Edge Dispersion visualization is shown in Figure 4.1(a), and discussed in Section Results.

Figure 4.1: Examples of Edge Dispersion and Edge Correlation. We applied our novel visualization methods to the BV dataset to compare them to the existing examinations of the data. (a) Edge Dispersion, measured as the standard deviation of the edge masses across samples, logarithmically scaled. (b) Edge Correlation, in form of Spearman's Rank Correlation Coefficient between the edge imbalances and the Nugent score. Tip edges are gray, because they do not have a meaningful imbalance. This example also shows the characteristics of edge masses and edge imbalances: The former highlights individual edges, the latter paths to clades.

4.3 Edge Correlation

In addition to the per-edge masses, the Edge Correlation further takes a specific meta-data feature into account, that is, a column of the meta-data matrix. The

Edge Correlation is calculated as the correlation between each edge column and the feature column, for example by using the Pearson Correlation Coefficient or Spearman's Rank Correlation Coefficient [40]. This yields a per-edge correlation of the placement masses or imbalances with the meta-data feature, and can again be visualized via color coding of the edges. It is inexpensive to calculate and hence scales well to large datasets. As typical correlation coefficients are within $[-1.0, 1.0]$, there is again no need for further normalization. This yields a tree where edges or clades with either a high linear or monotonic correlation with the selected meta-data feature are highlighted. Figure 4.1(b) shows an example of this method. In contrast to Edge PCA [87] that can use meta-data features to annotate samples in its scatter plots, our Edge Correlation method directly represents the influence of a feature on the branches or clades of the tree. It can thus, for example, help to identify and visualize dependencies between species abundances and environmental factors such as temperature or nutrient levels. Again, the choice of normalization strategy is important to draw meaningful conclusions. However, the correlation is **not** calculated between samples or sequence abundances. Hence, even when using normalized samples, the pitfalls regarding correlations of compositional data [74] do not apply here.

4.4 Results

4.4.1 BV Dataset

We re-analyzed the BV dataset by inferring a tree from the original reference sequence set and conducting phylogenetic placement of the 220 samples. The characteristics of this dataset were already explored in [128] and [87]. We use it here to give exemplary interpretations of our Edge Dispersion and Edge Correlation methods, and to evaluate them in comparison to existing methods.

Figure 4.1 shows our novel visualizations of the BV dataset. Edge Dispersion is shown in Figure 4.1(a), while Figure 4.1(b) shows Edge Correlation with the so-called Nugent score. The Nugent score [106] is a clinical standard for the diagnosis of Bacterial Vaginosis, ranging from 0 (healthy) to 10 (severe illness). The connection between the Nugent score and the abundance of placements on particular edges was already explored in [87], but only visualized indirectly (i.e., not on the RT itself). For example, Figure 6 of the original study plots the first two Edge PCA components colorized by the Nugent score. We recalculated this figure for comparison in Figure 5.3(i). In contrast, our Edge Correlation measure directly reveals the connection between Nugent score and placements on the reference tree: The clade on the left hand side of the tree, to which the red and orange branches lead to, are *Lactobacillus iners* and *Lactobacillus crispatus*, respectively, which were identified in [128] to be associated with a healthy vaginal microbiome. Thus, their presence in a sample is anti-correlated with the Nugent score, which is lower for healthy subjects. The branches leading to this clade are hence colored in red. On the other hand, there are several other clades that exhibit a positive correlation with the Nugent score, that is, were green and blue paths lead to in the figure, again a finding already reported in [128].

Both trees in Figure 4.1 highlight the same parts of the tree: The dark branches with high deviation in Figure 4.1(a) represent clades attached to either highly correlated (blue) or anti-correlated (red) paths Figure 4.1(b). This indicates that edges that have a high dispersion also vary between samples of different Nugent score.

We further compared our methods to the visualization of Edge PCA components on the reference tree. To this end, we recalculated Figures 4 and 5 of [87], and visualized them with our color scheme in Figure 4.5 for ease of comparison. They show the first two components of Edge PCA, mapped back to the RT. The first component reveals that the *Lactobacillus* clade represents the axis with the highest heterogeneity across samples, while the second componentfurther distinguishes between the two aforementioned clades within *Lactobacillus*. Edge Correlation also highlights the *Lactobacillus* clade as shown in Figure 4.1(b), but does not distinguish further between its sub-clades. This is because a high Nugent score is associated with a high abundance of placements in either of the two relevant *Lactobacillus* clades.

Further examples of variants of Edge Dispersion and Edge Correlation on this dataset are shown in Figure 4.2 and Figure 4.3. We also conducted Edge Correlation using Amsel's criteria [4] and the vaginal pH value as shown in Figure 4.4, both of which were used in [128] as additional indicators of Bacterial Vaginosis. We again found similar correlations compared to the Nugent score.

4.4.2 Tara Oceans Dataset

We analyzed the Tara Oceans (TO) dataset to provide further exemplary use cases for our visualization methods. To this end, we used the unconstrained *Eukaryota* RT with 2059 taxa as provided by our Automatic Reference Tree method [23]. The meta-data features of this dataset that best lend themselves to our methods are the sensor values for chlorophyll, nitrate, and oxygen concentration, as well as the salinity and temperature of the water samples. Other available meta-data features such as longitude and latitude are available, but would require more involved methods. This is because geographical coordinates yield pairwise distances between samples, whose integration into our correlation analysis methods is challenging. The Edge Correlation of the 370 samples with the nitrate concentration, the salinity, the chlorophyll concentration, and the water temperature are shown in Figure 4.6.

We selected the diatoms and the animals as two exemplary clades for closer examination of the results. In particular, the diatoms show a high correlation with the nitrate concentration, as well as an anti-correlation with salinity, which represent well-known relationships [76, 114]. See Figure 4.6 for details. These findings indicate that the method is able to identify known relationships. It will therefore also be useful to investigate or discover insights of novel relationships between sequence abundances and environmental parameters.

4.4.3 Performance

Both methods (Edge Dispersion and Edge Correlation) are computationally inexpensive, and thus applicable to large datasets. The calculation of the above visualizations took about 30 s each, which were mainly required for reading in the data. Furthermore, in order to scale to large datasets, we reimplemented Edge PCA, which was originally implemented as a command in the GUPPY program [90]. For the BV dataset with 220 samples, GUPPY required 9 min and used 2.2 GB of memory, while our implementation only required 33 s on a single core, using less than 600 MB of main memory. For the HMP dataset, as it is only single-threaded, GUPPY took 11 days and 75.1 GB memory, while our implementation needed 7.5 min on 16 cores and used 43.5 GB memory.

4.5 Conclusion and Outlook

Edge Dispersion highlights branches of the phylogenetic tree that exhibit variations in the number of placements, and thus allows to identify regions of the tree with a high placement heterogeneity. Edge Correlation additionally takes meta-data features into account, and identifies branches of the tree that correlate with quantitative features, such as the temperature or the pH value of the environmental samples. These methods complement existing methods such as Edge PCA, and are data exploration tools that can help unravel new patterns in phylogenetic placement data. The variants of the methods presented here are hence best used in combination with each other.

Figure 4.2: Examples of variants of Edge Dispersion. We re-analyzed the BV dataset to show variants of our Edge Dispersion method. All subfigures highlight the same branches and clades as found by other methods such as Edge PCA. The method is useful as a first exploratory tool to detect placement heterogeneity across samples. In contrast to Edge Correlation, it can however not explain the reasons of heterogeneity. Subfigure (a) shows the standard deviation of the absolute edge masses, without any further processing. It is striking that one outlier, marked with an arrow, is dominating, thus hiding the values on less variable edges. This outlier occurs at the species *Prevotella bivia* in one of the 220 samples, where 2781 out of 2782 sequences in the sample have placement mass on that branch. Upon close examination, this outlier can also be seen in Figure 1D of [128], but is less apparent there. Subfigure (b) is identical to Figure 4.1(a) of the main text and shows the standard deviation again, but this time using logarithmic scaling, thus revealing more details on the edges with lower placement mass variance. Furthermore, when comparing it to Figure 4.3(c), we see that the same clades that exhibit a high correlation or anti-correlation with meta-data there are also highlighted here. There are only few medium values, which indicates that there are two classes of edges: Those which distinguish patients and those who have almost no placement on them at all. Subfigure (c) shows the Index of Dispersion of the edge masses, that is, the variance normalized by the mean. Hence, edges with a higher number of placements are also allowed to have a higher variance. We again use a logarithmic scale because of the outlier. The figure reveals more details on the edges with lower variance, highlighted in medium green colors. Subfigure (d) shows the standard deviation of edge imbalances. Because we used imbalances of unit mass samples, the values are already normalized. The path to the *Lactobacillus* clade is again clearly visible, indicating that the placement mass in this clade has a high variance across samples. Note that imbalances can be negative; thus, the Index of Dispersion is not applicable to them.

Figure 4.3: Examples of variants of Edge Correlation. We again use the BV dataset, and show the correlation of edge masses and imbalances with the Nugent score. The Nugent score measures the severeness of Bacterial Vaginosis, and ranges from 0 for healthy subjects to 10 for heavily affected patients. Subfigures (a) and (b) use the Pearson Correlation Coefficient, that is, they show the linear correlation with the meta-data feature, while subfigures (c) and (d) use Spearman's Rank Correlation Coefficient and thus show monotonic correlations. Subfigure (d) is identical to Figure 4.1(b) of the main text. All subfigures show red edges or red paths at the *Lactobacillus* clade. This indicates that presence of placements in this clade is anti-correlated with the Nugent score, which is consistent with the findings of [128] and [87]. In other words, presence of *Lactobacillus* correlates with a healthy vaginal microbiome. On the other hand, blue and green edges, which indicate positive correlation, are indicative of edges that correlate to Bacterial Vaginosis. The extent of correlation is larger for Spearman's Coefficient, indicating that the correlation is monotonic, but not strictly linear.

Figure 4.4: Edge Correlation with more meta-data features. Here, we use additional meta-data features of the BV dataset to show that Edge Correlation yields consistent results with existing methods. In particular, we caltucated Spearman's Coefficient with Amsel's criteria [4] in subfigures (a) and (b), as well as with the vaginal pH value in subfigures (c) and (d). Both features were also used in [128] as indicators of Bacterial Vaginosis. The figures are almost identical to the ones shown in Figure 4.3; that is, they yield results that are consistent with the previously used Nugent score, as well as consistent with existing methods.

Figure 4.5: Recalculation of the Edge PCA tree visualization. Subfigures (a) and (b) are recalculations of Figures 4 and 5 of [87], respectively. However, we show them here in our coloring scheme in order to facilitate comparison with other figures. The original publication instead uses two colors for a positive and a negative sign of the principal components, and branch width to show their magnitude. Note that the actual sign is arbitrary, as it is derived from principal components.
The figure shows the first two Edge PCA components, visualized on the reference tree. This form of visualization is useful to interpret results such as the Edge PCA projection plot as shown in Figure 5.1(e) of the main text. It reveals which edges are mainly responsible for separating the samples into the PCA dimensions. Here, the first principal component in (a) indicates that the main PCA axis separates samples based on the presence of placements in the *Lactobacillus* clade, which is what the blue and green path leads to. The second component in (b) then further distinguishes between two species in this clade, namely *Lactobacillus iners* and *Lactobacillus crispatus*.

Figure 4.6: Examples of Edge Correlation using Tara Oceans samples. The figure shows the correlation of Tara Oceans sequence placements with (a) the nitrate, (b) the salinity, (c) the chlorophyll, and (d) the temperature sensor data of each sample. The sensor values range from −2.2 to 33.1 µmol/l (nitrate), from 33.2 to 40.2 psu (salt), from −0.02 to 1.55 mg/m^3 (chlorophyll), and from −0.8 to 30.5 °C (temperature), respectively. The negative nitrate and chlorophyll concentrations are values below the detection limit of the measurement method (pers. comm. with L. Guidi), and hence simply denote low concentrations. We used Spearman's Rank Correlation Coefficient, and examine two exemplary clades, namely the *Animals* and the *Diatoms*.
Diatoms are mainly photosynthetic, and thus depend on nitrates as key nutrients, which is clearly visible by the high correlation of the clade in (a). Furthermore, the diatoms exhibit positive correlation with the chlorophyll concentration (c), which again is indicative of their photosynthetic behavior. On the other hand, they show a high anti-correlation with the salt content (b). Salinity is a strong environmental factor which heavily affects community structures and species abundances [76], particularly diatoms [114].
The correlations of the animal clade are less pronounced. They exhibit a negative correlation with nitrate (a), as well as an increase in absolute abundance with higher temperatures (d). While these findings are not surprising, they show that the method is able to find meaningful relationships in the data.

5. Clustering

This chapter is based on the peer-reviewed publication:

- **Lucas Czech** and Alexandros Stamatakis. ”Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements.” *PLOS One*, 2018.

Contributions: Lucas Czech... Alexandros Stamatakis...

5.1 Motivation

Given a set of metagenomic samples, one key question is how much they differ from each other. A common distance metric between microbial communities is the (weighted) UniFrac distance [75, 77]. It uses the fraction of unique and shared branch lengths between phylogenetic trees to determine their difference. UniFrac has been generalized and adapted to phylogenetic placements in form of the phylogenetic Kantorovich-Rubinstein (KR) distance [39, 87]. In other contexts, the KR distance is also called Wasserstein distance, Mallows distance, or Earth Mover’s distance [67, 83, 116, 141]. The KR distance between two metagenomic samples is a metric that describes by at least how much the normalized mass distribution of one sample has to be moved across the RT to obtain the distribution of the other sample. The distance is symmetrical, and becomes larger the more mass needs to be moved, and the larger the respective displacement (moving distance) is. As the two samples being compared need to have equal masses, the KR distance operates on normalized samples; that is, it compares relative abundances.

Given such a distance measure between samples, a fundamental task consists in clustering samples that are similar to each other. Standard linkage-based clustering methods like UPGMA [65, 96, 127] are solely based on the distances between samples, that is, they calculate the distances of clusters as a function of pairwise sample distances. Squash Clustering [87, 128] is a method that also takes into account the intrinsic structure of phylogenetic placement data. It uses the KR distance to perform agglomerative hierarchical clustering of samples. Instead of using pairwise sample distances, however, it merges (**squashes**) clusters of samples by calculating their average per-edge placement mass. Thus, in each step, it operates on the same type of data, that is, on mass distributions on the RT. This results in a hierarchical clustering tree that has meaningful, and hence interpretable, branch lengths.

Remark: Squash clustering uses the KR distance, which is based on mass distributions on the edges of the RT. It thus "suffers" from the same shortcomings that Edge PCA is solving by using mass imbalance instead. For instance, figure 7 of [87] shows the result of a normal principal component analysis on a dataset of Bacterial Vaginosis, a disease of the vagina. On the left hand side of the figure, the blue data points, representing healthy women, cluster close together, while the red data points, which belong to sick patients, spread over the rest of the graph. The result of Squash Clustering on this dataset is presented in figure 1 of [128]. The bottom half of the clustering tree, containing mainly healthy samples, has short branches, while the top half, with mostly samples from sick patients, has many long branches. Thus, Squash Clustering and normal PCA represent almost equivalent information in this case.

5.2 Phylogenetic k-means

The number of tips in the resulting clustering tree obtained through Squash Clustering is equal to the number n of samples that are being clustered. Thus, for datasets with more than a few hundred samples, the clustering result becomes hard to inspect and interpret visually. We propose a variant of k-means clustering [78] to address this problem, which we call **Phylogenetic k-means**. It uses a similar approach as Squash Clustering, but yields a predefined number of k clusters. It is hence able to work with arbitrarily large datasets. Note that we are clustering samples here, instead of sequences [57]. We discuss choosing a reasonable value for k later.

The underlying idea is to assign each of the n samples to one of k cluster centroids, where each centroid represents the average mass distribution of all samples assigned to it. Note that all samples and centroids are of the same data type, namely, they are mass distributions on a fixed RT. It is thus possible to calculate distances between samples and centroids, and to calculate their average mass distributions, as described earlier. Our implementation follows Lloyd's algorithm [71], as shown in Algorithm 5.1.

Algorithm 5.1 Phylogenetic k-means

```
initialize $k$ Centroids
while not converged do
    assign each Sample to nearest Centroid
    update Centroids as mass averages of their Samples
return Assignments and Centroids
```

By default, we use the k-means++ initialization algorithm [6] to obtain a set of k initial centroids. It works by subsequent random selection of samples to be used as initial centroids, until k centroids have been selected. In each step, the probability of selecting a sample is proportional to its squared distance to the nearest already selected sample. An alternative initialization is to select samples as initial clusters entirely at random. This is however more likely to yield sub-optimal clusterings [54].

Then, each sample is assigned to its nearest centroid, using the KR distance. Lastly, the centroids are updated to represent the average mass distribution of all samples that are currently assigned to them. This iterative process alternates between improving the assignments and the centroids. Thus, the main difference to normal k-means is the use of phylogenetic information: Instead of euclidean distances on vectors, we use the KR distance, and instead of averaging vectors to obtain centroids, we use the average mass distribution.

The process is repeated until it converges, that is, the cluster assignments do not change any more, or until a maximum number of iterations have been executed. The second stopping criterion is added to avoid the super-polynomial worst case running time of k-means, which however almost never occurs in practice [5, 16].

The result of the algorithm is an assignment of each sample to one of the k clusters. As the algorithm relies on the KR distance, it clusters samples with similar relative abundances. The cluster centroids can be visualized as trees with a mass distribution, analogous to how Squash Clustering visualizes inner nodes of the clustering tree. That is, each centroid can be represented as the average mass distribution of the samples that were assigned to it. This allows to inspect the centroids and thus to interpret how the samples were clustered. Examples of this are shown in Figure 5.4.

The key question is how to select an appropriate k that reflects the number of “natural” clusters in the data. There exist various suggestions in the literature [14, 49, 112, 119, 138, 139]; we assessed the Elbow method [138] as explained in Figure 5.6, which is a straight forward method that yielded reasonable results for our test datasets. Additionally, for a quantitative evaluation of the clusterings, we used the k that arose from the number of distinct labels based on the available meta-data for the data. For example, the HMP samples are labeled with 18 distinct body sites, describing where each sample was taken from, c.f. Figure 5.2.

5.2.1 Algorithmic Improvements

In each assignment step of the algorithm, distances from all samples to all centroids are calculated, which has a time complexity of $\mathcal{O}(n \cdot k)$. In order to accelerate this step, we can apply branch binning as introduced in Section 2.4.1. For the BV dataset, we found that even using just 2 bins per edge does not alter the cluster assignments. Branch binning reduces the number of mass points that have to be accessed in memory during KR distance calculations; however, the costs for tree traversals remain. Thus, we observed a maximal speedup of 75% when using one bin per branch, see S5.1 Table for details.

Furthermore, during the execution of the algorithm, empty clusters can occur, for example, if k is greater than the number of natural clusters in the data. Although this problem did not occur in our tests, we implemented the following solution: First, find the cluster with the highest variance. Then, choose the sample of that cluster that is furthest from its centroid, and assign it to the empty cluster instead. This process is repeated if multiple empty clusters occur at once.

5.3 Imbalance k-means

We further propose **Imbalance k-means**, which is a variant of k-means that makes use of the edge imbalance transformation, and thus also takes the clades of the tree into account. In order to quantify the difference in imbalances between two samples, we use the euclidean distance between their imbalance vectors (that is, rows of the imbalance matrix). This is a suitable distance measure, as the imbalances implicitly capture the tree topology as well as the placement mass distributions. As a consequence, the expensive tree traversals required for Phylogenetic k-means are not necessary here. The algorithm takes the edge imbalance matrix of normalized samples as input, as shown in Figure 2.2(b), and performs a standard euclidean k-means clustering following Lloyd's algorithm.

This variant of k-means tends to find clusters that are consistent with the results of Edge PCA, as both use the same input data as well as the same distance measure. Furthermore, as the method does not need to calculate KR distances, and thus does not involve tree traversals, it is several orders of magnitude faster than the Phylogenetic k-means. For example, on the HMP dataset, it runs in mere seconds, instead of several hours needed for Phylogenetic k-means; see Section Performance for details.

5.4 Results

We now evaluate our Phylogenetic k-means clustering (which uses edge masses and KR distances) and Imbalance k-means clustering (which uses edge imbalances and euclidean distances) methods in terms of their clustering accuracy. We used the BV as an example of a small dataset to which methods such as Squash Clustering [87] are still applicable, and the HMP dataset to showcase that our methods scale to datasets that are too large for existing methods.

5.4.1 BV Dataset

We again use the re-analyzed BV dataset to test whether our methods work as expected, by comparing them to the existing analysis of the data in [128] and [87]. To this end, we ran both Phylogenetic k-means and Imbalance k-means on the BV dataset. We chose $k := 3$, inspired by the findings of [128]. They distinguish between subjects affected by Bacterial Vaginosis and healthy subjects, and further separate the healthy ones into two categories depending on the dominating clade in the vaginal microbiome, which is either *Lactobacillus iners* or *Lactobacillus crispatus.* Any choice of $k > 3$ would simply result in smaller, more fine-grained clusters, but not change the general findings of these experiments. An evaluation of the number of clusters using the Elbow method is shown in Figure 5.6. We furthermore conducted Squash Clustering and Edge PCA on the dataset, thereby reproducing previous results, in order to allow for a direct comparison between the methods, see Figure 5.1. The figure shows the results of Squash Clustering, Edge PCA, and two alternative dimensionality reduction methods, colorized by the cluster assignments **PKM** of Phylogenetic k-means (in red, green, and blue) and **IKM** of the Imbalance k-means (in purple, orange, and gray), respectively. We use two different color sets for the two methods, in order to make them distinguishable at first glance. Note that the

Figure 5.1: Comparison of k-means clustering to Squash Clustering and Edge PCA. We applied our variants of the k-means clustering method to the BV dataset in order to compare them to existing methods. See [128] for details of the dataset and its interpretation. We chose $k := 3$, as this best fits the features of the dataset. For each sample, we obtained two cluster assignments: First, by using Phylogenetic k-means, we obtained the cluster assignment **PKM**. Second, by using Imbalance k-means, we obtained assignment **IKM**. In each subfigure, the 220 samples are represented by colored circles: red, green, and blue show the cluster assignments **PKM**, while purple, orange, and gray show the cluster assignments **IKM**. (a) Hierarchical cluster tree of the samples, using Squash Clustering. The tree is a recalculation of Figure 1(A) of [128]. Each leaf represents a sample; branch lengths are KR distances. We added color coding for the samples, using **PKM**. The lower half of red samples are mostly healthy subjects, while the green and blue upper half are patients affected by Bacterial Vaginosis. (b) The same tree, but annotated by **IKM**. The tree is flipped horizontally for ease of comparison. The healthy subjects are split into two sub-classes, discriminated by the dominating species in their vaginal microbiome: orange and purple represent samples were *Lactobacillus iners* and *Lactobacillus crispatus* dominate the microbiome, respectively. The patients mostly affected by BV are clustered in gray. (c) Multidimensional scaling using the pairwise KR distance matrix of the samples, and colored by **PKM**. (d) Principal component analysis applied to the distance matrix by interpreting it as a data matrix. This is a recalculation of Figure 4 of [88], but colored by **PKM**. (e) Edge PCA applied to the samples, which is a recalculation of Figure 3 of [88], but colored by **IKM**.

mapping of colors to clusters is arbitrary and depends on the random initialization of the algorithm.

As can be seen in Figure 5.1(a), Squash Clustering as well as Phylogenetic k-means can distinguish healthy subjects from those affected by Bacterial Vaginosis. Healthy subjects constitute the lower part of the cluster tree. They have shorter branches between each other, indicating the smaller KR distance between them, which is a result of the dominance of *Lactobacillus* in healthy subjects. The same clusters are found by Phylogenetic k-means: As it uses the KR distance, it assigns all healthy subjects with short cluster tree branches to one cluster (shown in red). The green and blue clusters are mostly the subjects affected by the disease.

The distinguishing features between the green and the blue cluster are not apparent in the Squash cluster tree. This can however be seen in Figure 5.1(c), which shows a Multidimensional scaling (MDS) plot of the pairwise KR distances between the samples. MDS [40, 64, 84] is a dimensionality reduction method that can be used for visualizing levels of similarity between data points. Given a pairwise distance matrix, it finds an embedding into lower dimensions (in this case, 2 dimensions) that preserves higher dimensional distances as well as possible. Here, the red cluster forms a dense region, which is in agreement with its short branch lengths in the cluster tree. At the same time, the green and blue cluster are separated in the MDS plot, but form a coherent region of low density, indicating that $k := 3$ might be too large with Phylogenetic k-means on this dataset. That is, the actual clustering just distinguishes healthy from sick patients (Figure 5.6).

A similar visualization of the pairwise KR distances is shown in Figure 5.1(d). It is a recalculation of Figure 4 in the preprint [88], which did not appear in the final published version [87]. It is a recalculation of Figure 4 of [88]. but can also be found at [89]. The figure shows a standard Principal Components Analysis (PCA) [40, 64] applied to the distance matrix by interpreting it as a data matrix, and was previously used to motivate Edge PCA. However, although it is mathematically sound, the direct application of PCA to a distance matrix lacks a simple interpretation. Again, the red cluster is clearly separated from the rest, but this time, the distinction between the green and the blue cluster is not as apparent.

In Figure 5.1(b), we compare Squash Clustering to Imbalance k-means. Here, the distinction between the two *Lactobacillus* clades can be seen by the purple and orange cluster assignments. The cluster tree also separates those clusters into clades. The separate small group of orange samples above the purple clade is an artifact of the tree ladderization. The diseased subjects are all assigned to the gray cluster, represented by the upper half of the cluster tree. It is apparent that both methods separate the same samples from each other.

Lastly, Figure 5.1(e) compares Imbalance k-means to Edge PCA. The plot is a recalculation of Figure 3 of [88], which also appeared in Figure 6 in [87] and Figure 3 in [128], but colored using our cluster assignments. Because both methods work on edge imbalances, they group the data in the same way, that is, they clearly separate the two healthy groups and the diseased one from each other. Edge PCA forms a plot with three corners, which are colored by the three Imbalance k-means cluster assignments.

In Figure 5.3, we report more details of the comparison of our k-means variants to the dimensionality reduction methods used here. Furthermore, examples visualizations of the cluster centroids are shown in Figure 5.4, which further supports that our methods yield results that are in agreement with existing methods.

5.4.2 HMP Dataset

The HMP dataset is used here as an example to show that our method scales to large datasets. To this end, we used the unconstrained *Bacteria* RT with 1914 taxa as provided by our Automatic Reference Tree method [23]. The tree represents a broad taxonomic range of *Bacteria*, that is, the sequences were not explicitly selected for the HMP dataset, in order to test the robustness of our clustering methods. We then placed the 9192 samples of the HMP dataset with a total of 118 701 818 sequences on that tree, and calculated Phylogenetic and Imbalance k-means on the samples. The freely available meta-data for the HMP dataset labels each sample by the body site were it was taken from. As there are 18 different body site labels, we used $k := 18$. The result is shown in Figure 5.2. Furthermore, in Figure 5.5, we show a clustering of this dataset into $k := 8$ broader body site regions to exemplify the effects of using different values of k. This is further explored by using the Elbow method as shown in Figure 5.6.

Ideally, all samples from one body site would be assigned to the same cluster, hence forming a diagonal on the plot. However, as there are several nearby body sites, which share a large fraction of their microbiome [51], we do not expect a perfect clustering. Furthermore, we used a broad reference tree that might not be able

Figure 5.2: k-means cluster assignments of the HMP dataset with $k := 18$. Here, we show the cluster assignments as yielded by Phylogenetic k-means (a) and Imbalance k-means (b) of the HMP dataset. We used $k := 18$, which is the number of body site labels in the dataset, in order to compare the clusterings to this "ground truth". Each row represents a body site; each column one of the 18 clusters found by the algorithm. The color values indicate how many samples of a body site were assigned to each cluster. Similar body sites are clearly grouped together in coherent blocks, indicated by darker colors. For example, the stool samples were split into two clusters (topmost row), while the three vaginal sites were all put into one cluster (rightmost column). However, the algorithm cannot always distinguish between nearby sites, as can be seen from the fuzziness of the clusters of oral samples. This might be caused by our broad reference tree, and could potentially be resolved by using a tree more specialized for the data/region (not tested). Lastly, the figure also lists how the body site labels were aggregated into regions as used in Figure 5.5. Although the plots of the two k-means variants generally exhibit similar characteristics, there are some differences. For example, the samples from the body surface (ear, nose, arm) form two relatively dense clusters (columns) in (a), whereas those sites are spread across four of five clusters in (b). On the other hand, the mouth samples are more densely clustered in (b).

to resolve details in some clades. Nonetheless, the clustering is reasonable, which indicates a robustness against the exact choice of reference taxa, and can thus by used for distinguishing among samples. For example, stool and vaginal samples are clearly clustered. Furthermore, the sites that are on the surface of the body (ear, nose, and arm) also mostly form two blocks of cluster columns.

5.4.3 Performance

The complexity of Phylogenetic k-means is in $\mathcal{O}(k \cdot i \cdot n \cdot e)$, with k clusters, i iterations, and n samples, and e being the number of tree edges, which corresponds to the number of dimensions in standard euclidean k-means. As the centroids are randomly initialized, the number of iterations can vary; in our tests, it was mostly below 100. For the BV dataset with 220 samples and a reference tree with 1590 edges, using $k := 3$, our implementation ran 9 iterations, needing 35 s and 730 MB of main memory on a single core. For the HMP dataset with 9192 samples and 3824 edges, we used $k := 18$, which took 46 iterations and ran in 2.7 h on 16 cores, using 48 GB memory.

In contrast to this, Imbalance k-means does not need to conduct any expensive tree traversals, and instead operates on compact vectors, using euclidean distances. It is hence several orders of magnitude faster than Phylogenetic k-means. For example, using again $k := 18$ for the HMP dataset, the algorithm executed 75 iterations in 2 s. It is thus also applicable to extremely large datasets.

Furthermore, as the KR distance is used in Phylogenetic k-means as well as other methods such as Squash Clustering, our implementation is highly optimized and outperforms the existing implementation in GUPPY [90] by orders of magnitude (see below for details). The KR distance between two samples has a linear computational complexity in both the number of QSs and the tree size. As a test case, we computed the pairwise distance matrix between sets of samples. Calculating this matrix is quadratic in the number of samples, and is thus expensive for large datasets. For example, in order to calculate the matrix for the BV dataset with 220 samples, GUPPY can only use a single core and required 86 min. Our KR distance implementation in GENESIS is faster and also supports multiple cores. It only needed 90 s on a single core; almost half of this time is used for reading input files. When using 32 cores, the matrix calculation itself only took 8 s. This allows to process larger datasets: The distance matrix of the HMP dataset with 9192 samples placed on a tree with 3824 branches for instance took less than 10 h to calculate using 16 cores in GENESIS. In contrast, GUPPY needed 43 days for this dataset. Lastly, branch binning can be used to achieve additional speedups, as shown in S5.1 Table.

5.5 Conclusion and Outlook

Furthermore, we presented adapted variants of the k-means method, which exploit the structure of phylogenetic placement data to identify clusters of environmental samples. The method builds upon ideas such as Squash Clustering and can be applied to substantially larger datasets, as it uses a pre-defined number of clusters. For future exploration, other forms of cluster analyses could be extended to work on phylogenetic placement data, for example, soft k-means clustering [13, 31] or density-based methods [63]. The main challenge when adopting such methods consists in making them phylogeny-aware, that is, to use mass distributions on trees instead of the typical $\mathbb{R}^n$ vectors.

Figure 5.3: Comparison of k-means clustering to MDS, PCA, and Edge PCA. Here, we show and compare the dimensionality reduction methods MDS, PCA, and Edge PCA (one per row). MDS and PCA were calculated on the pairwise KR distance matrix of the BV dataset, Edge PCA was calculated using the placements on the re-inferred RT of the original publication [128]. The plots are colored by the cluster assignments as found by our k-means variants (first two columns), and by the Nugent score of the samples (last column). The Nugent score is included to allow comparison of the health status of patients with the clustering results. (a), (d) and (h) are identical to Figs 5.1(c), (d) and (e) of the main text, respectively. (f) and (i) are recalculations of Figures 4 and 3 of [88], respectively. This figure reveals additional details about how the k-means method works, that is, which samples are assigned to the same cluster. For example, the purple cluster found by Imbalance k-means forms a dense cluster of close-by samples on the left in (b) and (e), which is in accordance with the short branch lengths of this cluster as shown in Figure 5.1(b) of the main text.

Figure 5.4: Example of k-means cluster centroids visualization. Here we show the cluster centroids as found by our k-means variants using the BV dataset, visualized on the reference tree via color coding. The cluster assignments are the same as in Figure 5.1 of the main text; the first row show the three clusters found by Phylogenetic k-means, the second row the clusters found by Imbalance k-means. Each tree represents one centroid around which the samples were clustered, that is, it shows the combined masses of the samples that were assigned to that cluster. The edges are colored relative to each other, using a linear scaling of light blue (no mass), purple (half of the maximal mass) and black (maximal mass).
As explained in the main text, the samples can be split into three groups: The diseased subjects, which have placement mass in various parts of the tree, as well as two groups of healthy subjects, with placement mass in one of two *Lactobacillus* clades (marked with black arcs on the left of the trees). This grouping is also clearly visible in these trees. The red cluster for example represents all healthy subjects, and thus most of its mass is located in the two *Lactobacillus* clades. The purple and orange clusters on the other hand show a difference in placement mass between those clades. Furthermore, the placement mass of the gray cluster is mostly a combination of the masses of the green and blue cluster, all of which represent diseased subjects. These observations are in accordance with previous findings as explained in the main text.

Figure 5.5: Clustering using Phylogenetic k-means on the HMP dataset. k is set to 8, instead of $k := 18$ as in the main text, based on a coarse aggregation of the original body site labels. See Figure 5.1 for the cluster assignment where k is set to the original number of labels; there, we also list how the labels were aggregated. Each row represents a body site; each column one of the 8 clusters. The color values indicate how many samples of a body site were assigned to each cluster. Some of the body sites can be clearly separated, while particularly the samples from the oral region are distributed over different clusters. This might be due to homogeneity of the oral samples.

Figure 5.6: Variances of k-means clusters in our test datasets. The figures show the cluster variance, that is, the average squared distance of the samples to their assigned cluster centroids, for different values of k. The first row are clusterings of the BV dataset, the second row of the HMP dataset. They were clustered using Phylogenetic k-means (first column), and Imbalance k-means (second column), respectively. Accordingly, (a) and (c) use the KR distance, while (b) and (d) use the euclidean distance to measure the variance. These plots can be used for the Elbow method in order to find the appropriate number of clusters in a dataset [138]. Low values of k induce a high variance, because many samples exhibit a large distance from their assigned centroid. On the other hand, at a given point, higher values of k only yield a marginal gain by further splitting clusters. Thus, if the data has a natural number of clusters, the corresponding k produces an angle in the plot, called the “elbow”.

For example, (a) and (b) exhibit the elbow at $k := 2$ and 3, respectively, which are marked with orange circles. These values are consistent with previous findings, for instance, Figure 5.1: There, Phylogenetic k-means splits the samples into a distinct red cluster and the nearby green and blue clusters, while Imbalance k-means yields three separate clusters in purple, orange, and gray.

For the HMP dataset, the elbow is less pronounced. We suspect that this is due to the broad reference tree not being able to adequately resolve fine-grained differences between samples, see S1 Text for details. Likely candidates for k are $4-6$ for (c) and around 7 for (d). These values are consistent with the number of coherent “blocks” of clusters, which can be observed in Figure 5.2. Clearer results for this dataset might be obtained with other methods for finding “good” values for k, although we did not test them here.

Table 5.1: Effect of Branch Binning on the KR Distance of the HMP Dataset. Here we show the effect of per-branch placement binning on the run-time and on the resulting relative error when calculating the pairwise KR distance matrix between samples, by example of the Human Microbiome Project (HMP) [51, 95] dataset. Because of the size of the dataset (9192 samples) and reference tree (1914 taxa), we executed this evaluation in parallel on 16 cores. The first row shows the baseline performance, that is, without binning. When using fewer bins per branch, the run-time decreases, at the cost of slightly increasing the average relative error. Still, even when compressing the placement masses into only one bin per branch (that is, just using per-branch masses), the average relative error of the KR distances is around 1%, which is acceptable for most applications. However, considering that the run-time savings are not substantially better for a low number of bins, we recommend using a relatively large number of bins, e.g., 32 or more. This is because run-times of KR distance calculations also depend on other effects such as the necessary repeated tree traversals. We also conducted these tests on the BV dataset, were the relative error is even smaller.

Bins	Time (h:mm)	Speedup	Relative Δ
-	9:46	1.00	0.000000
256	6:58	1.40	0.000008
128	6:39	1.47	0.000015
64	6:30	1.50	0.000035
32	6:25	1.52	0.000124
16	6:13	1.57	0.000272
8	6:08	1.59	0.000669
4	6:07	1.60	0.002747
2	6:04	1.61	0.004284
1	5:35	1.75	0.011585

6. Conclusion and Outlook

we showed...

in the future...

A. Empirical Datasets

This chapter is based on the peer-reviewed publications:

- **Lucas Czech**, Pierre Barbera, and Alexandros Stamatakis. ”Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement.” *Bioinformatics*, 2018.
- **Lucas Czech** and Alexandros Stamatakis. ”Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements.” *PLOS One*, 2018.

Contributions: Lucas Czech... Alexandros Stamatakis...

We used three real world datasets to evaluate our methods:

- Bacterial Vaginosis (BV) [128]. This small dataset was already analyzed with phylogenetic placement in the original publication. We used it as an example of an established study to compare our results to. It has 220 samples with a total of 15 060 unique sequences.
- Tara Oceans (TO) [47, 56, 133]. This world-wide sequencing effort of the open oceans provides a rich set of meta-data, such as geographic location, temperature, and salinity. Unfortunately, the sample analysis for creating the official data repository is still ongoing. We thus were only able to use 370 samples with 27 697 007 unique sequences.
- Human Microbiome Project (HMP) [51, 95]. This large data repository intends to characterize the human microbiota. It contains 9192 samples from different body sites with a total of 63 221 538 unique sequences. There is additional meta-data such as age and medical history, which is available upon special request. We only used the publicly available meta-data.

Details of the datasets (download links, data statistics, data preprocessing, etc.) are provided in S1 Text. At the time of writing, about one year after we initially downloaded the data, the TO repository has grown to 1170 samples, while the HMP even published a second phase and now comprises 23 666 samples of the 16S region. This further emphasizes the need for scalable methods to analyze such data.

These datasets represent a wide range of environments, number of samples, and sequence lengths. We use them to evaluate our methods and to exemplify which method is applicable to what kind of data. To this end, the sequences of the datasets were placed on appropriate phylogenetic RTs as explained in S1 Text, in order to obtain phylogenetic placements that our methods can be applied to. In the following, we present the respective results, and also compare our methods to other methods where applicable. As the amount and type of available meta-data differs for each dataset, we could not apply all methods to all datasets. Lastly, we also report the run-time performance of our methods on these data.

The analyses and figures presented here were conducted on distinct reference alignments and trees. Firstly, for the BV dataset, we used the original set of reference sequences, and re-inferred a tree on them. Secondly, for the TO and HMP datasets, we used our Automatic Reference Tree (ART) method [23] to construct sets of suitable reference sequences from the SILVA database [115, 150]. We used the 90% threshold consensus sequences; see [23] for details.

For all analyses, we used the following software setup: Unconstrained maximum likelihood trees were inferred using RAxML v8.2.8 [129]. For aligning reads against reference alignments and reference trees, we used a custom MPI wrapper for PaPaRa 2.0 [9, 10], which is available at [8]. We then applied the `chunkify` procedure as explained in [23] to split the sequences into chunks of unique sequences prior to conducting the phylogenetic placement, in order to minimize processing time. Phylogenetic placement was conducted using EPA-ng [7], which is a faster and more scalable phylogenetic placement implementation than RAxML-EPA [11] and pplacer [90]. Lastly, given the per-chunk placement files produced by EPA-ng, we executed the `unchunkify` procedure of [23] to obtain per-sample placement files. These subsequently served as the input data for the methods presented here.

A.1 Bacterial Vaginosis

We used the Bacterial Vaginosis dataset [128] in order to compare our novel methods to existing ones such as Edge PCA and Squash Clustering [39, 87]. The dataset contains metabarcoding sequences of the vaginal microbiome of 220 women, and was kindly provided by Sujatha Srinivasan. This small dataset with a total of 426 612 query sequences, thereof 15 060 unique, was already analyzed with phylogenetic placement methods in the original publication. We re-inferred the reference tree of the original publication using the original alignment, which contains 797 reference sequences specifically selected to represent the vaginal microbiome. As the query sequences were already prepared, no further preprocessing was applied prior to phylogenetic placement. The available per-sample quantitative meta-data for this dataset comprises the Nugent score [106], the value of Amsel's criteria [4], and the vaginal pH value. We used all three meta-data types in our analyses.

TODO: from art:

For testing the accuracy of our unconstrained *Bacteria* tree on real data, we used a vaginal microbiome dataset of 220 women [128], which was provided by Sujatha Srinivasan. See Figure 3.8 for the respective results. This small dataset with a total of 426 612 query sequences, thereof 15 060 unique, was already analyzed with phylogenetic placement methods in the original publication. We used it as an example of a well-designed study to asses our results using an ART as reference tree. In addition to the *Bacteria* ART, we re-inferred the reference tree of the original publication using their alignment, again using RAxML 8.2.8 [129]. The query sequences of the dataset were then aligned to our reference tree and alignment, as well as to the reference alignment of the original publication and our re-inferred tree. For aligning, we used a custom MPI wrapper of PaPaRa 2.0 [9, 10], which is available at [8]. Finally, the query sequences were placed on these trees using EPA-ng [7], and the analyses were subsequently performed as explained in Figure 3.8.

A.2 Tara Oceans

The Tara Oceans (TO) dataset [47, 56, 133] that we used here contains amplicon sequences of protists, and is available at https://www.ebi.ac.uk/ena/data/view/PRJEB6610. At the time of download, there were 370 samples available with a total of 49 023 231 sequences. As the available data are raw `fastq` files, we followed [79] to generate cleaned per-sample `fasta` files. For this, we used our tool PEAR [153] to merge the paired-end reads; cutadapt [85] for trimming tags as well as forward and reverse primers; and vsearch [118] for filtering erroneous sequences and generating per-sample `fasta` files. We filtered out sequences below 95 bps and above 150 bps, to remove potentially erroneous sequences. No further preprocessing (such as chimera detection) was applied. This resulted in a total of 48 036 019 sequences, thereof 27 697 007 unique. The sequences were then used for phylogenetic placement as explained above. We placed the sequences on the unconstrained *Eukaryota* reference tree obtained via our ART method [23], which comprises 2059 taxa, thereof 1795 eukaryotic sequences. The remaining 264 taxa are *Archaea* and *Bacteria*, and were included as a broad outgroup. The TO dataset has a rich variety of per-sample meta-data features; in the context of this paper, we mainly focus on quantitative features such as temperature, salinity, as well as oxygen, nitrate and chlorophyll content of the water. Furthermore, each sample has meta-data features indicating the date, longitude and latitude, depth, etc. of the sampling location. This data might be interesting for further correlation analyses based on geographical information. We did not use them here, as for example longitude and latitude would require a more involved method that also accounts for, e.g., ocean currents. Furthermore, geographical coordinates yield pairwise distances between samples, which are not readily usable with our correlation analysis. Lastly, in order to use features such as the date, that is, in order to analyze samples over time, the same sampling locations would need to be visited at different times during the year, which is not the case for the Tara Oceans expedition.

A.3 Human Microbiome Project

We used the Human Microbiome Project (HMP) dataset [51, 95] for testing the scalability of our methods. In particular, we used the “HM16STR” data of the initial phase “HMP1”, which are available from http://www.hmpdacc.org/hmp/. The dataset consists of trimmed 16S rRNA sequences of the `V1V3`, `V3V5`, and `V6V9` regions. The data are further divided into a “by_sample” set and a “healthy” set, which we merged in order to obtain one large dataset, with a total of 9811 samples. We then removed 10 samples that were larger than 70 MB as well as 605 samples that had fewer than 1500 sequences, because we considered them as defective or unreliable outliers. Finally, we also removed 2 samples that did not have a valid body site label assigned to them. This resulted in a set of 9192 samples containing a total of 118 702 967 sequences with an average length of 413 bps. From these samples, sequences with a length of less than 150 bps as well as sequences longer than 540 bps were removed, as we considered them potentially erroneous. No further preprocessing (such as chimera detection) was applied. This resulted in a total of 116 520 289 sequences, of which 63 221 538 were unique. We then used the unconstrained *Bacteria* tree of our ART method [23] for phylogenetic placement. The tree comprises 1914 taxa, thereof 1797 bacterial sequences. The remaining 117 taxa are *Archaea* and *Eukaryota*, and were included as a broad outgroup. Each sample is labeled with one of 18 human body site locations where it was sampled. This is the only publicly available meta-data feature.

TODO: from art:

We used the Human Microbiome Project (HMP) dataset [51, 95] for further testing our methods (see Figure 3.9). In particular, we used the “HM16STR” data of their initial phase “HMP1”, which are available from http://www.hmpdacc.org/hmp/. The dataset consists of trimmed 16S rRNA sequences of the `V1V3`, `V3V5`, and `V6V9` regions. Each sample of the dataset is labeled with the human body site where it was obtained. See Table 3.4 for an overview of those labels. The data are further divided into a “by_sample” set and a “healthy” set, which we merged in order to obtain one large dataset, with a total of 9811 samples. We then removed 10 samples that were larger than 70 MB as well as 605 samples that had fewer than 1500 sequences, because we considered them as outliers, and finally 2 samples that did not have a valid body site label assigned to them. This resulted in a set of 9192 samples containing a total of 118 702 967 sequences with an average length of 413 bps. From these samples, sequences with a length less than 150 bps as well as sequences longer than 540 bps were removed. No further preprocessing (e.g., chimera detection) was applied. This resulted in a total of 116 520 289 sequences, of which 63 221 538 were unique. These were split into 1265 chunks of size 50 000 each, which were subsequently aligned to and placed on the unconstrained *Bacteria* tree with 2059 tips using the steps explained above. The chunk placements were then transformed again into per-sample placement files, before finally running the steps explained in Figure 3.9.

B. Pipeline and Implementation

This chapter is based on the peer-reviewed publications:

- **Lucas Czech**, Pierre Barbera, and Alexandros Stamatakis. ”Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement.” *Bioinformatics*, 2018.
- **Lucas Czech** and Alexandros Stamatakis. ”Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements.” *PLOS One*, 2018.

Contributions: Lucas Czech... Alexandros Stamatakis...

The methods described here are implemented in our tool GAPPA, which is freely available under GPLv3 at http://github.com/lczech/gappa. GAPPA internally uses our C++ library GENESIS, which offers functionality for working with phylogenies and phylogenetic placement data, and also contains methods to work with taxonomies, sequences and many other data types. GENESIS is also freely available under GPLv3 at http://github.com/lczech/genesis.

GAPPA offers a command line interface for conducting typical tasks when working with phylogenetic placements. The methods that we described here are implemented via the following sub-commands:

- `dispersion`: The command takes a set of jplace files (called samples), and calculates and visualizes the Edge Dispersion per edge of the reference tree.
- `correlation`: The command takes a set of jplace samples, as well as a table containing metadata features for each sample. It then calculates and visualizes the Edge Correlation with the metadata features per edge of the reference tree.
- `phylogenetic-kmeans` and `imbalance-kmeans`: Performs k-means clustering of a set of jplace files according to our methods.

- `squash` and `edgepca`: Reimplementations of the two existing methods [39, 87].

These are the GAPPA commands that are relevant for this paper. The tool also offers additional commands that are useful for phylogenetic placement data, such as visualization or filtering. At the time of writing this manuscript, GAPPA is under active development, with more functions planned in the near future. Lastly, we provide prototype implementations, scripts, data, and other tools used for the tests and figures in this paper at http://github.com/lczech/placement-methods-paper.

TODO: ART:

An implementation of the methods described here is freely available in our tool GAPPA, which is published under GPLv3 at http://github.com/lczech/gappa. GAPPA is based on our C++ library GENESIS, which offers functionality concerning phylogenies and phylogenetic placement data, but also has functions to work with sequences, taxonomies and many other data types. GENESIS is also published under GPLv3 and is available at http://github.com/lczech/genesis.

GAPPA offers a command line interface for typical tasks of working with phylogenetic placements. The methods described in this paper are implemented via four subcommands:

- `art`: Automatic Reference Tree method. The command expects a taxonomy file and a sequence file of a sequence database, e.g., SILVA [115, 150], as well as the target number of consensus sequences to be generated for the intended phylogeny. The result is a `fasta` file with consensus sequences representing taxonomic clades. The command can be further customized, e.g., by changing the consensus sequence method, using only a specified subclade of the taxonomy for running the algorithm, as well as several detail settings for the method. It can also output additional info files that allow to inspect details of the calculations, like the number of sequences and their entropy per clade.

- `extract`: Extract/collect placements in specific sub-clades of the tree. The command performs the main step of the multilevel placement approach. Its input is a set of `jplace` files containing placements on the backbone tree, as well as a file listing the clade name that each taxon of the backbone tree belongs to. For each clade, it then writes a new `jplace` file, containing all queries that were placed in that clade with more than a customizable threshold of their placement mass.
  Furthermore, if provided with the sequence files that were used to make the input `jplace` files, the corresponding sequence of each query are also written to `fasta` files per clade. Thus, a per-clade collection of sequences is created, where each result file contains the sequences that were placed in this clade of the backbone tree. These can then be used for the second level placement on separate clade-specific trees.

- `chunkify`: Split a set of `fasta` files into chunks of equal size, and write abundance maps. The command re-names the sequences using a configurable hash function (MD5, SHA1 or SHA256), and de-duplicates across all input sequences. Its output are chunk files of sequences, as well as an abundance map file for each input sequences file. The sequence chunk files can then be used to perform phylogenetic placement to obtain per-chunk `jplace` files.

- `unchunkify`: Take the per-chunk `jplace` files as well as the abundance map files, and generate a `jplace` for each original sequence file, including the correct abundances. This command is the second step of the `chunkify` command, and reverts its effect, so that the resulting `jplace` files are as if they were created using the original sequence files.

These are the commands of GAPPA relevant for this paper, but it also offers more commands that are useful when working with phylogenetic placements. At the time of writing, it is under active development, and more functions are planned for the near future. Furthermore, we provide prototype implementations, scripts, data and other tools used for the tests and figures in this paper at http://github.com/lczech/placement-methods-paper.

C. List of Publications

unieuk: [12] art: [23] pppp: [24]

1. **L. Czech**, S. Berger, D. Krompaß, J. Zhang, P. Kapli, P. Pavlidis and A. Stamatakis. Evolutionary Placement of Short Reads - Methods, Applications, and Visualization. Poster at EMBO/EMBL Symposium: A New Age of Discovery for Aquatic Microeukaryotes, Heidelberg, Germany, January 2016, and at Hellenic Bioinformatics Conference (HBio) 2016, Thessaloniki, Greece, November 2016.

2. F. Mahé, C. de Vargas, D. Bass, **L. Czech**, A. Stamatakis, E. Lara, D. Singer, J. Mayor, J. Bunge, S. Sernaker, T. Siemensmeyer, I. Trautmann, S. Romac, C. Berney, A. Kozlov, E. A. D. Mitchell, C. V. W. Seppey, E. Egge, G. Lentendu, R. Wirth, G. Trueba, and M. Dunthorn. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution*, vol. 1, no. 4, p. 0091, 2017.

3. **L. Czech**, J. Huerta-Cepas, and A. Stamatakis. A Critical Review on the Use of Support Values in Tree Viewers and Bioinformatics Toolkits. *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 383–384, 2017.

4. T. Flouri, J. Zhang, **L. Czech**, K. Kobert, A. Stamatakis. An efficient approach to merging paired-end reads and the incorporation of uncertainties. Chapter in Algorithms for Next-Generation Sequencing Data: Techniques, Approaches and Applications. 1st ed., M. Elloumi, Ed. Springer International Publishing AG, 2017, pp. 299–326.

5. P. Barbera, A. Kozlov, T. Flouri, D. Darriba, **L. Czech** and A. Stamatakis. Massively Parallel Evolutionary Placement of Genetic Sequences. Poster at ISC 2017 PhD Symposium, Frankfurt am Main, Germany, June 2017.

6. C. Berney, A. Ciuprina, S. Bender, J. Brodie, V. Edgcomb, E. Kim, J. Rajan, L. W. Parfrey, S. Adl, S. Audic, D. Bass, D. A. Caron, G. Cochrane, **L. Czech**, M. Dunthorn, S. Geisen, F. O. Glöckner, F. Mahé, C. Quast, J. Z. Kaye, A.

G. B. Simpson, A. Stamatakis, J. del Campo, P. Yilmaz, and C. de Vargas. UniEuk : Time to Speak a Common Language in Protistology! *Journal of Eukaryotic Microbiology*, vol. 38, no. 1, pp. 42–49, 2017.

7. X. Zhou, S. Lutteropp, **L. Czech**, A. Stamatakis, M. von Looz, A. Rokas. Quartet-based computations of internode certainty provide accurate and robust measures of phylogenetic incongruence. *bioRxiv*, 168526, 2017.

8. P. Barbera, A. Kozlov, **L. Czech**, B. Morel, A. Stamatakis. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *bioRxiv*, 291658, 2018.

9. D. Bass, **L.Czech**, B. Williams, C. Berney, M. Dunthorn, F. Mahe, G. Torruella, G. Stentiford and T. Williams. Clarifying the Relationships between Microsporidia and Cryptomycota. *Journal of Eukaryotic Microbiology*, 2018.

10. **L. Czech**, A. Stamatakis. Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement. *bioRxiv*, 299792, 2018.

11. **L. Czech**, A. Stamatakis. Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements. *bioRxiv*, 346353, 2018.

12. TODO: 1KITE

13. TODO: long reads

14. TODO: swarm 3?

Bibliography

[1] K. Abarenkov, R. Henrik Nilsson, K.-H. Larsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjøller, E. Larsson, T. Pennanen, and Others. The UNITE database for molecular identification of fungi–recent updates and future perspectives. **New Phytologist**, 186(2):281–285, 2010.

[2] J. Aitchison. **The statistical analysis of compositional data**. Chapman and Hall London, 1986.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. **Journal of Molecular Biology**, 215(3):403–410, oct 1990. ISSN 00222836. doi: 10.1016/S0022-2836(05)80360-2. URL http://linkinghub.elsevier.com/retrieve/pii/S0022283605803602.

[4] R. Amsel, P. A. Totten, C. A. Spiegel, K. C. S. Chen, D. Eschenbach, and K. K. Holmes. Nonspecific vaginitis: Diagnostic Criteria and Microbial and Epidemiologic Associations. **The American Journal of Medicine**, 74(1): 14–22, 1983. ISSN 0002-9343. doi: 10.1016/0002-9343(83)91112-9. URL http://www.sciencedirect.com/science/article/pii/0002934383911129.

[5] D. Arthur and S. Vassilvitskii. How Slow is the K-means Method? In **Proceedings of the Twenty-second Annual Symposium on Computational Geometry**, SCG '06, pages 144–153, New York, NY, USA, 2006. ACM. ISBN 1-59593-340-9. doi: 10.1145/1137856.1137880.

[6] D. Arthur and S. Vassilvitskii. k-means++ : The Advantages of Careful Seeding. In **Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.**, pages 1027–1035. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.

[7] P. Barbera, A. M. Kozlov, L. Czech, B. Morel, and A. Stamatakis. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. **bioRxiv**, 2018. doi: 10.1101/291658. URL https://www.biorxiv.org/content/early/2018/03/29/291658.

[8] S. Berger and L. Czech. PaPaRa 2.0 with MPI, September 2016. URL https://github.com/lczech/papara_nt. Accessed: 2017-11-04.

[9] S. Berger and A. Stamatakis. Aligning short reads to reference alignments and trees. **Bioinformatics**, 27(15):2068–2075, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btr320.

[10] S. Berger and A. Stamatakis. PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension. Technical report, Heidelberg Institute for Theoretical Studies, Heidelberg, 2012.

[11] S. Berger, D. Krompass, and A. Stamatakis. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. **Systematic Biology**, 60(3):291–302, 2011. ISSN 10635157. doi: 10.1093/sysbio/syr010.

[12] C. Berney, A. Ciuprina, S. Bender, J. Brodie, V. Edgcomb, E. Kim, J. Rajan, L. W. Parfrey, S. Adl, S. Audic, D. Bass, D. A. Caron, G. Cochrane, L. Czech, M. Dunthorn, S. Geisen, F. O. Glöckner, F. Mahé, C. Quast, J. Z. Kaye, A. G. B. Simpson, A. Stamatakis, J. del Campo, P. Yilmaz, and C. de Vargas. UniEuk : Time to Speak a Common Language in Protistology! **Journal of Eukaryotic Microbiology**, 38(1):42–49, 2017. ISSN 10665234. doi: 10.1111/jeu.12414. URL http://doi.wiley.com/10.1111/jeu.12414.

[13] J. C. Bezdek. **Pattern Recognition with Fuzzy Objective Function Algorithms**. Advanced applications in pattern recognition. Plenum Press, 1981. doi: 10.1007/978-1-4757-0450-1.

[14] H. Bischof, A. Leonardis, and A. Selb. MDL Principle for Robust Vector Quantisation. **Pattern Analysis & Applications**, 2(1):59–72, apr 1999. ISSN 1433-7541. doi: 10.1007/s100440050015. URL https://doi.org/10.1007/s100440050015.

[15] B. E. Blaisdell. A measure of the similarity of sets of sequences not requiring sequence alignment. **Proceedings of the National Academy of Sciences**, 83(14):5155–5159, 1986. URL http://www.pnas.org/content/83/14/5155.abstract.

[16] L. Bottou and Y. Bengio. Convergence properties of the k-means algorithms. In **Advances in neural information processing systems**, pages 585–592, 1995.

[17] D. R. Cavener. Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates. **Nucleic Acids Research**, 15(4), 1987.

[18] D. R. Cavener and S. C. Ray. Eukaryotic start and stop translation sites. **Nucleic Acids Research**, 19(12):3185–3192, jun 1991. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC328309/.

[19] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. **Nucleic Acids Research**, 38(6):1767–1771, 2009. ISSN 03051048. doi: 10.1093/nar/gkp1137.

[20] J. R. Cole, Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske, and J. M. Tiedje. Ribosomal database project: data and tools for high throughput rRNA analysis. **Nucleic Acids Res**, 42, 2014. doi: 10.1093/nar/gkt1244. URL https://doi.org/10.1093/nar/gkt1244.

[21] M. Comin and D. Verzotto. Alignment-free phylogeny of whole genomes using underlying subwords. **Algorithms for molecular biology : AMB**, 7(1):34, 2012. ISSN 1748-7188. doi: 10.1186/1748-7188-7-34. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549825{&}tool=pmcentrez{&}rendertype=abstract.

[22] A. Criscuolo and S. Gribaldo. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. **BMC Evolutionary Biology**, 10(1):210, 2010. ISSN 1471-2148. doi: 10.1186/1471-2148-10-210. URL http://www.biomedcentral.com/1471-2148/10/210{%}5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/20626897{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3017758{%}5Cnhttp://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-10-210.

[23] L. Czech and A. Stamatakis. Methods for Automatic Reference Trees and Multilevel Phylogenetic Placement. **bioRxiv**, page 299792, 2018. doi: 10.1101/299792. URL https://www.biorxiv.org/content/early/2018/04/11/299792.full.pdf+html.

[24] L. Czech and A. Stamatakis. Scalable Methods for Post-Processing, Visualizing, and Analyzing Phylogenetic Placements. **bioRxiv**, 2018. doi: 10.1101/346353. URL https://www.biorxiv.org/content/early/2018/06/14/346353.

[25] A. E. Darling, G. Jospin, E. Lowe, F. A. Matsen, H. M. Bik, and J. a. Eisen. PhyloSift: phylogenetic analysis of genomes and metagenomes. **PeerJ**, 2:e243, 2014. ISSN 2167-8359. doi: 10.7717/peerj.243. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3897386{&}tool=pmcentrez{&}rendertype=abstract.

[26] W. H. E. Day and F. R. McMorris. Threshold consensus methods for molecular sequences. **Journal of Theoretical Biology**, 159(4):481–489, 1992. ISSN 10958541. doi: 10.1016/S0022-5193(05)80692-7.

[27] W. H. E. Day and F. R. McMorris. Critical comparison of consensus methods for molecular sequences. **Nucleic acids research**, 20(5):1093–1099, 1992.

[28] C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahe, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horak, O. Jaillon, G. Lima-Mendez, J. Luke, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, E. Karsenti, E. Boss, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. B. Sullivan, and D. Velayoudon. Eukaryotic plankton diversity in the sunlit ocean. **Science**, 348(6237): 1261605–1261605, 2015. ISSN 0036-8075. doi: 10.1126/science.1261605. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1261605.

[29] N. Desai, D. Antonopoulos, J. A. Gilbert, E. M. Glass, and F. Meyer. From genomics to metagenomics. **Current Opinion in Biotechnology**, 23(1): 72–76, 2012. ISSN 09581669. doi: 10.1016/j.copbio.2011.12.017.

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

[31] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. **Journal of Cybernetics**, 3(3): 32–57, 1973. doi: 10.1080/01969727308546046. URL https://doi.org/10.1080/01969727308546046.

[32] M. Dunthorn, J. Otto, S. A. Berger, A. Stamatakis, F. Mahé, S. Romac, C. De Vargas, S. Audic, A. Stock, F. Kauff, T. Stoeck, B. Edvardsen, R. Massana, F. Not, N. Simon, and A. Zingone. Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. **Molecular Biology and Evolution**, 31(4):993–1009, 2014. ISSN 07374038. doi: 10.1093/molbev/msu055.

[33] A. Ö. C. Dupont, R. I. Griffiths, T. Bell, and D. Bass. Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. **Environmental Microbiology**, 18(6):2010–2024, 2016. ISSN 14622920. doi: 10.1111/1462-2920.13220.

[34] S. R. Eddy. Profile hidden Markov models. **Bioinformatics**, 14(9):755—-763, 1998. URL http://bioinformatics.oxfordjournals.org/content/14/9/755.short.

[35] S. R. Eddy. A new generation of homology search tools based on probabilistic inference. In **Genome Informatics**, volume 23, pages 205–211. World Scientific, 2009.

[36] R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. **Bioinformatics**, 26(19):2460–2461, 2010. doi: 10.1093/bioinformatics/btq461. URL +http://dx.doi.org/10.1093/bioinformatics/btq461.

[37] D. J. Edwards and K. E. Holt. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. **Microbial informatics and experimentation**, 3(1):2, 2013. ISSN 2042-5783. doi: 10.1186/2042-5783-3-2. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3630013{&}tool=pmcentrez{&}rendertype=abstract.

[38] A. Escobar-Zepeda, A. Vera-Ponce De León, and A. Sanchez-Flores. The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. **Frontiers in Genetics**, 6(348):1–15, 2015. ISSN 16648021. doi: 10.3389/fgene.2015.00348.

[39] S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, 74:569–592, 2012. ISSN 13697412. doi: 10.1111/j.1467-9868.2011.01018.x.

[40] B. S. Everitt and A. Skrondal. **The Cambridge Dictionary of Statistics**. Cambridge University Press, 4th edition, 2010. ISBN 978-0521766999.

[41] D. P. Faith. Conservation evaluation and phylogenetic diversity. **Biological Conservation**, 61:1–10, 1992. ISSN 00063207. doi: 10.1016/0006-3207(92) 91201-3.

[42] J. Felsenstein. **Inferring Phylogenies**. Sinauer Associates Sunderland, MA, 2 edition, 2004. ISBN 978-0878931774.

[43] A. Filipski, K. Tamura, P. Billing-Ross, O. Murillo, and S. Kumar. Phylogenetic placement of metagenomic reads using the minimum evolution principle. **BMC genomics**, 16(1):6947, dec 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S1-S13. URL http://www.biomedcentral.com/1471-2164/16/S1/S13.

[44] C. R. Giner, I. Forn, S. Romac, R. Logares, and C. D. Vargas. Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes. **Applied and Environmental Microbiology**, 82 (15):4757–4766, 2016. doi: 10.1128/AEM.00560-16.Address.

[45] N. J. Gotelli and R. K. Colwell. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. **Ecology Letters**, 4(4):379–391, jul 2001. ISSN 1461023X. doi: 10.1046/j.1461-0248.2001.00230.x. URL http://doi.wiley.com/10.1046/j.1461-0248.2001.00230.x.

[46] S. Gran-Stadniczeñko, L. Šupraha, E. D. Egge, and B. Edvardsen. Haptophyte Diversity and Vertical Distribution Explored by 18S and 28S Ribosomal RNA Gene Metabarcoding and Scanning Electron Microscopy. **Journal of Eukaryotic Microbiology**, pages 1–19, 2017. ISSN 10665234. doi: 10.1111/jeu.12388. URL http://doi.wiley.com/10.1111/jeu.12388.

[47] L. Guidi, S. Chaffron, L. Bittner, D. Eveillard, A. Larhlimi, S. Roux, Y. Darzi, S. Audic, L. Berline, J. Brum, L. P. Coelho, J. C. I. Espinoza, S. Malviya, S. Sunagawa, C. Dimier, S. Kandels-Lewis, M. Picheral, J. Poulain, S. Searson, Tara Oceans coordinators, L. Stemmann, F. Not, P. Hingamp, S. Speich, M. Follows, L. Karp-Boss, E. Boss, H. Ogata, S. Pesant, J. Weissenbach, P. Wincker, S. G. Acinas, P. Bork, C. de Vargas, D. Iudicone, M. B. Sullivan, J. Raes, E. Karsenti, C. Bowler, and G. Gorsky. Plankton networks driving carbon export in the oligotrophic ocean. **Nature**, 532 (7600):465–470, apr 2016. ISSN 0028-0836. doi: 10.1038/nature16942. URL http://europepmc.org/articles/PMC4851848.

[48] L. Guillou, D. Bachar, S. Audic, D. Bass, C. Berney, L. Bittner, C. Boutte, G. Burgaud, C. De Vargas, J. Decelle, and Others. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. **Nucleic acids research**, 41(D1): D597—-D604, 2012.

[49] G. Hamerly and C. Elkan. Learning the k in k-means. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, **Advances in Neural Information Processing Systems 16**, pages 281–288. MIT Press, 2004.

[50] P. Hugenholtz, B. M. Goebel, and N. R. Pace. Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. **Journal of Bacteriology**, 180(18):4765–4774, 1998.

[51] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, M. G. Giglio, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-Vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. Bonazzi, J. Paul Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. G. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. DeSantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. Michael Dunne, A. Scott Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. J. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, S. Kinder Haake, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H. Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, N. B. King, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.-C. Lo, C. A. Lozupone, R. Dwayne Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavromatis, J. M. McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya, J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. Pop, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren, R. A. Gibbs, S. K. Highlander, B. A. Methé, K. E. Nelson, J. F. Petrosino, G. M. Weinstock, R. K. Wilson, and O. White. Structure, function and diversity

of the healthy human microbiome. **Nature**, 486(7402):207–214, jun 2012. ISSN 0028-0836. doi: 10.1038/nature11234. URL http://www.ncbi.nlm.nih.gov/pubmed/22699609http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3564958http://www.nature.com/doifinder/10.1038/nature11234.

[52] IUPAC-IUB Commission on Biochemical Nomenclature (CBN). Abbreviations and Symbols for nucleic acids, polynucleotides and their constituents. **Journal of Molecular Biology**, 55(3):299–310, 1971. ISSN 00222836. doi: 10.1016/0022-2836(71)90319-6.

[53] J. M. Janda and S. L. Abbott. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. **Journal of Clinical Microbiology**, 45(9):2761–2764, 2007. doi: 10.1128/JCM.01228-07. URL http://jcm.asm.org/content/45/9/2761.short.

[54] T. Kanungo, D. M. Mount, N. S. Netanyahu, A. Y. Wu, C. D. Piatko, R. Silverman, and A. Y. Wu. A Local Search Approximation Algorithm for k-Means Clustering. **Computational Geometry**, 28(2-3):89–112, 2003.

[55] P. Kapli, S. Lutteropp, J. Zhang, K. Kobert, P. Pavlidis, A. Stamatakis, and T. Flouri. Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. **Bioinformatics**, 33(11):1630–1638, 2017.

[56] E. Karsenti, S. G. Acinas, P. Bork, C. Bowler, C. de Vargas, J. Raes, M. Sullivan, D. Arendt, F. Benzoni, J. M. Claverie, M. Follows, G. Gorsky, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, U. Krzic, F. Not, H. Ogata, S. Pesant, E. G. Reynaud, C. Sardet, M. E. Sieracki, S. Speich, D. Velayoudon, J. Weissenbach, and P. Wincker. A holistic approach to marine Eco-systems biology. **PLoS Biology**, 9(10):7–11, 2011. ISSN 15449173. doi: 10.1371/journal.pbio.1001177.

[57] D. R. Kelley and S. L. Salzberg. Clustering metagenomic sequences with interpolated Markov models. **BMC Bioinformatics**, 11(1):544, nov 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-544. URL https://doi.org/10.1186/1471-2105-11-544.

[58] O.-S. Kim, Y.-J. Cho, K. Lee, S.-H. Yoon, M. Kim, H. Na, S.-C. Park, Y. S. Jeon, J.-H. Lee, H. Yi, and Others. Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. **International journal of systematic and evolutionary microbiology**, 62(3):716–721, 2012.

[59] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. **Journal of Molecular Evolution**, 29(2):170–179, aug 1989. ISSN 1432-1432. doi: 10.1007/BF02100115. URL https://doi.org/10.1007/BF02100115.

[60] H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. **Journal of Molecular Evolution**, 31(2):151–160, aug 1990. ISSN 1432-1432. doi: 10.1007/BF02109483. URL https://doi.org/10.1007/BF02109483.

[61] L. B. Koski and G. B. Golding. The closest BLAST hit is often not the nearest neighbor. **Journal of molecular evolution**, 52(6):540–2, jun 2001. ISSN 0022-2844. doi: 10.1007/s002390010184. URL http://www.ncbi.nlm.nih.gov/pubmed/11443357.

[62] A. M. Kozlov, J. Zhang, P. Yilmaz, F. O. Glöckner, and A. Stamatakis. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. **Nucleic Acids Research**, 44(11):5022–5033, jun 2016. ISSN 13624962. doi: 10.1093/nar/gkw396. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw396.

[63] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, 1(3):231–240, 2011.

[64] W. J. Krzanowski and F. Marriott. **Multivariate Analysis**. Wiley, 1994. ISBN 9780340593257.

[65] P. Legendre and L. F. J. Legendre. **Numerical Ecology**. Developments in Environmental Modelling. Elsevier Science, 1998. ISBN 9780080537870.

[66] I. Letunic and P. Bork. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. **Nucleic acids research**, 44(W1):W242–5, jul 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw290. URL http://www.ncbi.nlm.nih.gov/pubmed/27095192http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4987883.

[67] E. Levina and P. Bickel. The earth mover's distance is the Mallows distance: some insights from statistics. **Eighth IEEE International Conference on Computer Vision**, pages 251–256, 2001. doi: 10.1109/ICCV.2001.937632. URL http://ieeexplore.ieee.org/xpls/abs{_}all.jsp?arnumber=937632.

[68] C. Li and J. Wang. Relative entropy of DNA and its application. **Physica A: Statistical Mechanics and its Applications**, 347:465–471, 2005. ISSN 03784371. doi: 10.1016/j.physa.2004.08.041.

[69] K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. **Science**, 324(5934):1561—-1564, jun 2009. ISSN 1095-9203. doi: 10.1126/science.1171243. URL http://www.ncbi.nlm.nih.gov/pubmed/19541996.

[70] K. Liu, T. J. Warnow, M. T. Holder, S. M. Nelesen, J. Yu, A. P. Stamatakis, and C. R. Linder. SATé-II: Very Fast and Accurate Simultaneous Estimation of Multiple Sequence Alignments and Phylogenetic Trees. **Systematic Biology**, 61(1):90, jan 2012. ISSN 1076-836X. doi: 10.1093/sysbio/syr095. URL https://academic.oup.com/sysbio/article-lookup/doi/10.1093/sysbio/syr095.

[71] S. P. Lloyd. Least Squares Quantization in PCM. **IEEE Transactions on Information Theory**, 28(2):129–137, 1982. ISSN 15579654. doi: 10.1109/TIT.1982.1056489.

[72] R. Logares, T. H. Haverkamp, S. Kumar, A. Lanzén, A. J. Nederbragt, C. Quince, and H. Kauserud. Environmental microbiology through the lens of high-throughput DNA sequencing: Synopsis of current platforms and bioinformatics approaches. **Journal of Microbiological Methods**, 91(1):106–113, oct 2012. ISSN 0167-7012. doi: 10.1016/J.MIMET.2012.07.017. URL https://www.sciencedirect.com/science/article/pii/S0167701212002527.

[73] R. Logares, S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, J. Raes, J. Poulain, O. Jaillon, P. Wincker, S. Kandels-Lewis, E. Karsenti, P. Bork, and S. G. Acinas. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. **Environmental Microbiology**, 16(9):2659–2671, sep 2014. ISSN 14622920. doi: 10.1111/1462-2920.12250. URL http://doi.wiley.com/10.1111/1462-2920.12250.

[74] D. Lovell, V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bähler. Proportionality: A Valid Alternative to Correlation for Relative Data. **PLOS Computational Biology**, 11(3):e1004075, mar 2015. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004075. URL http://dx.plos.org/10.1371/journal.pcbi.1004075.

[75] C. Lozupone and R. Knight. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. **Applied and Environmental Microbiology**, 71(12):8228–8235, 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.12.8228.

[76] C. A. Lozupone and R. Knight. Global patterns in bacterial diversity. **Proceedings of the National Academy of Sciences**, 104(27):11436–11440, 2007. ISSN 0027-8424. doi: 10.1073/pnas.0611525104. URL http://www.pnas.org/content/104/27/11436.

[77] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities. **Applied and Environmental Microbiology**, 73(5):1576–1585, mar 2007. ISSN 00992240. doi: 10.1128/AEM.01996-06. URL http://aem.asm.org/content/73/5/1576.abstract.

[78] J. MacQueen. Some methods for classification and analysis of multivariate observations. **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**, 1(233):281–297, 1967. ISSN 00970433. doi: citeulike-article-id:6083430.

[79] F. Mahé. Fred's metabarcoding pipeline, November 2016. URL https://github.com/frederic-mahe/swarm/wiki/Fred's-metabarcoding-pipeline. Accessed: 2018-01-15.

[80] F. Mahé, T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. Swarm: Robust and fast clustering method for amplicon-based studies. **PeerJ**, 2:1–12, 2014. ISSN 2167-9843. doi: http://dx.doi.org/10.7287/peerj.preprints.386v1. URL http://dx.doi.org/10.7717/peerj.593.

[81] F. Mahé, T. Rognes, C. Quince, C. De Vargas, and M. Dunthorn. Swarm v2: Highly-scalable and high-resolution amplicon clustering. **PeerJ**, 2015. URL https://peerj.com/articles/1420/.

[82] F. Mahé, C. de Vargas, D. Bass, L. Czech, A. Stamatakis, E. Lara, D. Singer, J. Mayor, J. Bunge, S. Sernaker, T. Siemensmeyer, I. Trautmann, S. Romac, C. Berney, A. Kozlov, E. A. D. Mitchell, C. V. W. Seppey, E. Egge, G. Lentendu, R. Wirth, G. Trueba, and M. Dunthorn. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. **Nature Ecology & Evolution**, 1(4):0091, mar 2017. ISSN 2397-334X. doi: 10.1038/s41559-017-0091. URL http://www.nature.com/articles/s41559-017-0091.

[83] C. L. Mallows. A Note on Asymptotic Joint Normality. **Ann. Math. Statist.**, 43(2):508–515, 1972. doi: 10.1214/aoms/1177692631. URL http://dx.doi.org/10.1214/aoms/1177692631.

[84] K. V. Mardia. Some Properties of Classical Multi-Dimesional Scaling. **Communications in Statistics-Theory and Methods**, 7(13):1233–1241, 1978.

[85] M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. **EMBnet.journal**, 17(1):10, may 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200. URL http://journal.embnet.org/index.php/embnetjournal/article/view/200.

[86] F. A. Matsen. Phylogenetics and the Human Microbiome. **Systematic Biology**, 64(1):e26–e41, jul 2015. doi: 10.1093/sysbio/syu053. URL +http://dx.doi.org/10.1093/sysbio/syu053http://arxiv.org/abs/1407.1794.

[87] F. A. Matsen and S. N. Evans. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. **PLOS ONE**, 8(3):1–17, jan 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0056859. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0056859.

[88] F. A. Matsen and S. N. Evans. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. **arXiv**, 2011.

[89] F. A. Matsen and S. N. Evans. Edge principal components analysis example. Online: `http://matsen.fredhutch.org/pplacer/demo/pca.html`, July 2011. URL http://matsen.fredhutch.org/pplacer/demo/pca.html. Accessed: 2018-01-15.

[90] F. A. Matsen, R. B. Kodner, and E. V. Armbrust. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. **BMC Bioinformatics**, 11(1):538, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-538. URL http://www.biomedcentral.com/1471-2105/11/538.

[91] F. A. Matsen, N. G. Hoffman, A. Gallagher, and A. Stamatakis. A format for phylogenetic placements. **PLoS ONE**, 7(2):1–4, jan 2012. ISSN 19326203. doi: 10.1371/journal.pone.0031009. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031009.

[92] F. A. Matsen, A. Gallagher, and C. O. McCoy. Minimizing the average distance to a closest leaf in a phylogenetic tree. **Systematic Biology**, 62(6):824–836, 2013. ISSN 10635157. doi: 10.1093/sysbio/syt044.

[93] K. O. May. A set of independent necessary and sufficient conditions for simple majority decision. **Econometrica: Journal of the Econometric Society**, pages 680–684, 1952.

[94] P. J. McMurdie and S. Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. **PLoS Computational Biology**, 10(4): e1003531, apr 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003531. URL http://dx.plos.org/10.1371/journal.pcbi.1003531.

[95] B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, H. Jonathan, A. T. Chinwalla, A. M. Earl, M. G. Fitzgerald, R. S. Fulton, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-vercoe, J. Eric, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, J. V. Bonazzi, P. Brooks, G. A. Buck, J. Christian, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, M. Dawn, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. Desantis, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, J. H. Badger, A. T. Chinwalla, A. M. Earl, M. G. Fitzgerald, R. S. Fulton, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. R. Bonazzi, P. Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. Desantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. Michael Dunne, A. Scott Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H. Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, S. Kinder-Haake, N. B. King, R. Knight, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu,

C.-C. Lo, C. A. Lozupone, R. Dwayne Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavrommatis, J. M. McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya, J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Qing Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren, R. A. Gibbs, S. K. Highlander, G. M. Weinstock, R. K. Wilson, and O. White. A framework for human microbiome research. **Nature**, 486(7402):215–221, jun 2012. ISSN 0028-0836. doi: 10.1038/nature11209.A. URL http://www.ncbi.nlm.nih.gov/pubmed/22699610http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3377744http://www.nature.com/doifinder/10.1038/nature11209.

[96] C. D. Michener and R. R. Sokal. A quantitative approach to a problem in classification. **Evolution**, 11(2):130–162, 1957.

[97] S. Mignard and J. P. Flandrois. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. **Journal of microbiological methods**, 67(3):574–581, 2006.

[98] B. Minh, S. Klaere, and A. Haeseler. Phylogenetic Diversity within Seconds. **Systematic Biology**, 55(5):769–773, 2006. ISSN 1063-5157. doi: 10.1080/10635150600981604. URL https://academic.oup.com/sysbio/article-lookup/doi/10.1080/10635150600981604.

[99] S. Mirarab, N. Nguyen, and T. Warnow. SEPP: SATé-Enabled Phylogenetic Placement. **Biocomputing**, pages 247–258, 2012.

[100] D. Moreira and H. Philippe. Molecular phylogeny: pitfalls and progress. **International Microbiology**, 3(1):9–16, 2000.

[101] J. L. Morgan, A. E. Darling, and J. A. Eisen. Metagenomic sequencing of an in vitro-simulated microbial community. **PLoS ONE**, 5(4):1–10, 2010. ISSN 19326203. doi: 10.1371/journal.pone.0010209.

[102] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **Journal of molecular biology**, 48(3):443–453, 1970.

[103] M. Nei and W. H. Li. Mathematical model for studying genetic variation in terms of restriction endonucleases. **Proceedings of the National Academy**

of Sciences of the United States of America, 76(10):5269–5273, oct 1979. ISSN 0027-8424. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC413122/.

[104] L.-T. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. **Molecular Biology and Evolution**, 32(1):268–74, jan 2015. ISSN 1537-1719. doi: 10.1093/molbev/msu300. URL http://www.ncbi.nlm.nih.gov/pubmed/25371430http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4271533.

[105] N.-p. Nguyen, S. Mirarab, B. Liu, M. Pop, and T. Warnow. TIPP: taxonomic identification and phylogenetic profiling. **Bioinformatics**, 30(24):3548–3555, 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu721. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu721.

[106] R. P. Nugent, M. A. Krohn, and S. L. Hillier. Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. **Journal of clinical microbiology**, 29(2):297–301, 1991.

[107] A. Oulas, C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis, and I. Iliopoulos. Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies, 2015. ISSN 1177-9322 (Electronic).

[108] N. R. Pace. A molecular view of microbial diversity and the biosphere. **Science**, 276(5313):734–740, 1997.

[109] F. Pardi and N. Goldman. Species Choice for Comparative Genomics: Being Greedy Works. **PLOS Genetics**, 1(6):1, 2005. doi: 10.1371/journal.pgen.0010071. URL https://doi.org/10.1371/journal.pgen.0010071.

[110] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz. A proposal for a standardized bacterial taxonomy based on genome phylogeny. **bioRxiv**, 2018. doi: 10.1101/256800. URL https://www.biorxiv.org/content/early/2018/01/31/256800.

[111] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. **Proceedings of the National Academy of Sciences**, 85(8): 2444–2448, 1988. URL http://www.pnas.org/content/85/8/2444.abstract.

[112] D. Pelleg, A. W. Moore, and Others. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In **ICML**, volume 1, pages 727–734, 2000.

[113] C. A. Petti. Detection and identification of microorganisms by gene amplification and sequencing. **Clinical infectious diseases**, 44(8):1108–1114, 2007.

[114] M. Potapova. **Patterns of Diatom Distribution In Relation to Salinity**. Springer, 2011. ISBN 978-94-007-1326-0. doi: 10.1007/978-94-007-1327-7.

[115] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glockner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. **Nucleic Acids Research**, 41 (D1):D590–D596, jan 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1219. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1219.

[116] S. T. Rachev. The Monge-Kantorovich Mass Transference Problem and its Stochastic Applications. **Theory of Probability and its Applications**, 29 (4):647–676, 1985.

[117] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. **Mathematical Biosciences**, 53(1):131–147, 1981. ISSN 00255564. doi: 10.1016/0025-5564(81)90043-2.

[118] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé. VSEARCH: a versatile open source tool for metagenomics. **PeerJ**, 4:e2584, 2016.

[119] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, 20:53–65, 1987.

[120] F. Sanger and A. Coulson. A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase. **Journal of Molecular Biology**, 94(3):441–448, may 1975. ISSN 0022-2836. doi: 10.1016/0022-2836(75)90213-2. URL https://www.sciencedirect.com/science/article/pii/0022283675902132?via{%}3Dihub.

[121] a. O. Schmitt and H. Herzel. Estimating the entropy of DNA sequences. **Journal of theoretical biology**, 188(3):369–377, 1997. ISSN 00225193. doi: 10.1006/jtbi.1997.0493.

[122] M. B. Scholz, C. C. Lo, and P. S. G. Chain. Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. **Current Opinion in Biotechnology**, 23(1):9–15, 2012. ISSN 09581669. doi: 10.1016/j.copbio.2011.11.013.

[123] C. E. Shannon and W. Weaver. **The Mathematical Theory of Communication**. University of Illinois Press, 1951. ISBN 0-252-72546-8.

[124] H. Shimodaira. An Approximately Unbiased Test of Phylogenetic Tree Selection. **Systematic Biology**, 51(3):492–508, 2002. doi: 10.1080/10635150290069913. URL +http://dx.doi.org/10.1080/10635150290069913.

[125] H. Shimodaira and M. Hasegawa. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. **Molecular Biology and Evolution**, 16(8):1114, 1999. doi: 10.1093/oxfordjournals.molbev.a026201. URL +http://dx.doi.org/10.1093/oxfordjournals.molbev.a026201.

[126] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. **Journal of molecular biology**, 147(1):195–197, 1981.

[127] R. R. Sokal. A statistical method for evaluating systematic relationship. **University of Kansas science bulletin**, 28:1409–1438, 1958.

[128] S. Srinivasan, N. G. Hoffman, M. T. Morgan, F. A. Matsen, T. L. Fiedler, R. W. Hall, F. J. Ross, C. O. McCoy, R. Bumgarner, J. M. Marrazzo, and D. N. Fredricks. Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. **PLOS ONE**, 7(6):e37818, jan 2012. ISSN 19326203. doi: 10.1371/journal.pone.0037818. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0037818.

[129] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. **Bioinformatics**, 30(9):1312–1313, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu033.

[130] T. Stoeck, D. Bass, M. Nebel, R. Christen, M. D. M. Jones, H.-W. BREINER, and T. A. Richards. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. **Molecular Ecology**, 19(s1):21–31, 2010. doi: 10.1111/j.1365-294X.2009.04480.x. URL http://dx.doi.org/10.1111/j.1365-294X.2009.04480.x.

[131] K. Strimmer and A. Rambaut. Inferring confidence sets of possibly misspecified gene trees. **Proceedings of the Royal Society of London B: Biological Sciences**, 269(1487):137–142, 2002.

[132] S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. a. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Doré, S. D. Ehrlich, A. Stamatakis, P. Bork, J. Dore, S. D. Ehrlich, A. Stamatakis, and P. Bork. Metagenomic species profiling using universal phylogenetic marker genes. **Nature Methods**, 10(12):1196, oct 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2693. URL http://www.ncbi.nlm.nih.gov/pubmed/24141494http://www.nature.com/doifinder/10.1038/nmeth.2693http://dx.doi.org/10.1038/nmeth.2693http://10.0.4.14/nmeth.2693https://www.nature.com/articles/nmeth.2693{#}supplementary-information.

[133] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-castillo, P. I. Costea, C. Cruaud, F. Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, F. D\textquoterightOvidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, and P. Bork. Structure and function of the global ocean microbiome. **Science**, 348(6237):1–10, 2015. ISSN 0036-8075. doi: 10.1126/science.1261359. URL http://science.sciencemag.org/content/348/6237/1261359.

[134] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. **American Mathematical Society: Lectures on Mathematics in the Life Sciences**, 17:57–86, 1986. doi: citeulike-article-id:4801403.

URL citeulike-article-id:4801403{%}5Cnhttp://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20{&}amp{%}5Cnpath=ASIN/0821811673.

[135] L. Tedersoo, M. Bahram, S. Põlme, U. Kõljalg, N. S. Yorou, R. Wijesundera, L. V. Ruiz, A. M. Vasco-Palacios, P. Q. Thu, A. Suija, and Others. Global diversity and geography of soil fungi. **Science**, 346(6213):1256688, 2014.

[136] B. Temperton and S. J. Giovannoni. Metagenomics: Microbial diversity through a scratched lens. **Current Opinion in Microbiology**, 15(5): 605–612, 2012. ISSN 13695274. doi: 10.1016/j.mib.2012.07.001. URL http://dx.doi.org/10.1016/j.mib.2012.07.001.

[137] L. R. Thompson, J. G. Sanders, D. McDonald, A. Amir, J. Ladau, K. J. Locey, R. J. Prill, A. Tripathi, S. M. Gibbons, G. Ackermann, J. A. Navas-Molina, S. Janssen, E. Kopylova, Y. Vázquez-Baeza, A. González, J. T. Morton, S. Mirarab, Z. Zech Xu, L. Jiang, M. F. Haroon, J. Kanbar, Q. Zhu, S. Jin Song, T. Kosciolek, N. A. Bokulich, J. Lefler, C. J. Brislawn, G. Humphrey, S. M. Owens, J. Hampton-Marcell, D. Berg-Lyons, V. McKenzie, N. Fierer, J. A. Fuhrman, A. Clauset, R. L. Stevens, A. Shade, K. S. Pollard, K. D. Goodwin, J. K. Jansson, J. A. Gilbert, R. Knight, and T. E. M. P. Consortium. A communal catalogue reveals Earth's multiscale microbial diversity. **Nature**, nov 2017. URL http://dx.doi.org/10.1038/nature24621http://10.0.4.14/nature24621https://www.nature.com/articles/nature24621{#}supplementary-information.

[138] R. L. Thorndike. Who belongs in the family? **Psychometrika**, 18(4):267–276, 1953.

[139] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, 63(2):411–423, 2001.

[140] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter,

G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The Sequence of the Human Genome. **Science**, 291 (5507):1304–51, feb 2001. ISSN 0036-8075. doi: 10.1126/science.1058040. URL http://www.ncbi.nlm.nih.gov/pubmed/11181995.

[141] C. Villani. **Optimal transport: old and new**. Springer Science & Business Media, 2008. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9.

[142] S. Vinga. Information theory applications for biological sequence analysis. **Briefings in Bioinformatics**, 15(3):376–389, 2014. doi: 10.1093/bib/bbt068. URL +http://dx.doi.org/10.1093/bib/bbt068.

[143] S. Vinga and J. Almeida. Alignment-free sequence comparison - A review. **Bioinformatics**, 19(4):513–523, 2003. ISSN 13674803. doi: 10.1093/bioinformatics/btg005.

[144] S. Vinga and J. S. Almeida. Rényi continuous entropy of DNA sequences. **Journal of Theoretical Biology**, 231(3):377–388, 2004. ISSN 00225193. doi: 10.1016/j.jtbi.2004.06.030.

[145] C. von Mering, P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork. Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. **Science**, 315(5815):1126–1130, 2007. ISSN 0036-8075. doi: 10.1126/science.1133420. URL http://science.sciencemag.org/content/315/5815/1126.

[146] S. Weiss, Z. Z. Xu, S. Peddada, A. Amir, K. Bittinger, A. Gonzalez, C. Lozupone, J. R. Zaneveld, Y. Vázquez-Baeza, A. Birmingham, E. R. Hyde,

and R. Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. **Microbiome**, 5(1):27, dec 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0237-y. URL http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0237-y.

[147] K. A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), 2018. URL https://www.genome.gov/sequencingcostsdata.

[148] X. Xia, Z. Xie, M. Salemi, L. Chen, and Y. Wang. An index of substitution saturation and its application. **Molecular Phylogenetics and Evolution**, 26(1):1–7, 2003. ISSN 1055-7903. doi: https://doi.org/10.1016/S1055-7903(02)00326-3. URL http://www.sciencedirect.com/science/article/pii/S1055790302003263.

[149] Z. Yang. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. **Syst. Biol.**, 43(3):329–342, 1994.

[150] P. Yilmaz, L. W. Parfrey, P. Yarza, J. Gerken, E. Pruesse, C. Quast, T. Schweer, J. Peplies, W. Ludwig, and F. O. Gl?ckner. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. **Nucleic Acids Research**, 42(D1):D643–D648, jan 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1209. URL https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkt1209.

[151] G. Yu, D. K. Smith, H. Zhu, Y. Guan, and T. T.-Y. Lam. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. **Methods in Ecology and Evolution**, 8(1):28–36, 2017. ISSN 2041-210X. doi: 10.1111/2041-210X.12628. URL http://dx.doi.org/10.1111/2041-210X.12628.

[152] J. Zhang, P. Kapli, P. Pavlidis, and A. Stamatakis. A general species delimitation method with applications to phylogenetic placements. **Bioinformatics**, 29(22):2869–2876, 2013. ISSN 13674803. doi: 10.1093/bioinformatics/btt499.

[153] J. Zhang, K. Kobert, T. Flouri, and A. Stamatakis. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. **Bioinformatics**, 30(5):614–620, 2014. ISSN 13674803. doi: 10.1093/bioinformatics/btt593.

[154] S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. **IEEE transactions on pattern analysis and machine intelligence**, 28(6):917–929, 2006.