Received: 5 March 2020 | Revised: 24 July 2020 | Accepted: 25 August 2020

DOI: 10.1111/1755-0998.13255

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

SCRAPP: A tool to assess the diversity of microbial samples from phylogenetic placements

Pierre Barbera[1] | Lucas Czech[1] | Sarah Lutteropp[1] | Alexandros Stamatakis[1,2]

[1]Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

[2]Institute for Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

Correspondence
Pierre Barbera and Alexandros Stamatakis, Computational Molecular Evolution Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany.
Emails: pierre.barbera@h-its.org, alexandros.stamatakis@h-its.org

Funding information
Klaus Tschira Stiftung

Abstract

Microbial ecology research is currently driven by the continuously decreasing cost of DNA sequencing and the improving accuracy of data analysis methods. One such analysis method is phylogenetic placement, which establishes the phylogenetic identity of the anonymous environmental sequences in a sample by means of a given phylogenetic reference tree. However, assessing the diversity of a sample remains challenging, as traditional methods do not scale well with the increasing data volumes and/or do not leverage the phylogenetic placement information. Here, we present SCRAPP, a highly parallel and scalable tool that uses a molecular species delimitation algorithm to quantify the diversity distribution over the reference phylogeny for a given phylogenetic placement of the sample. SCRAPP employs a novel approach to cluster phylogenetic placements, called placement space clustering, to efficiently perform dimensionality reduction, so as to scale on large data volumes. Furthermore, it uses the phylogeny-aware molecular species delimitation method mPTP to quantify diversity. We evaluated SCRAPP using both, simulated and empirical data sets. We use simulated data to verify our approach. Tests on an empirical data set show that SCRAPP-derived metrics can classify samples by their diversity-correlated features equally well or better than existing, commonly used approaches. SCRAPP is available at https://github.com/pbdas/scrapp.

KEYWORDS

diversity, microbiome, phylogenetic placement, species delimitation

1 | INTRODUCTION

Environmental DNA sampling is increasingly becoming a standard practice, not least due to continuously decreasing sequencing costs. One, by now, established way to analyse such data is Phylogenetic Placement (Barbera et al., 2018b; Berger et al., 2011; Matsen et al., 2010). In phylogenetic placement, sequences from environmental samples (query sequences, QS) are placed on a phylogenetic tree comprising the biome under study (reference tree, RT), resulting in a set of QS placements on this reference tree. Commonly, this information is used to identify the taxonomic identity of individual QS in relation to the reference data, and through that the overall taxonomic composition of a given sample. Prominent examples of such compositional analyses include the study of protists in neotropical rainforest soils (Mahé et al., 2017) or the study of relationships between bacterial community composition and disease (Srinivasan et al., 2012).

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2020 The Authors. Molecular Ecology Resources published by John Wiley & Sons Ltd

However, a drawback of phylogenetic placement is its inability to resolve relationships between individual QS, even when they are placed in close proximity to each other on the RT. This is sensible as it substantially reduces the computational effort while still producing highly accurate results, especially for short read sequences with weak phylogenetic signal. Nonetheless, resolving relationships between QS constitutes a desired feature by many users. We expect this feature to become more important with the increasing adoption of fourth-generation sequencing technologies, which yield substantially longer reads. We have previously demonstrated the value of resolving between-QS relationships with longer read data (Jamy et al., 2020) and hope that the methods presented here constitute a step into this direction.

Another key goal of molecular studies is to assess microbial diversity. A plethora of distinct metrics already exist to quantify the diversity within a sample (*alpha* diversity) and between samples (*beta* diversity) (Tucker et al., 2017). For a subset of these metrics, phylogenetic information can be used to calculate both alpha (e.g. Phylogenetic Diversity [PD]; Faith, 1992), and Phylogenetic Species Variability (Helmus et al., 2007) and beta (e.g. the UniFrac distance [Lozupone & Knight, 2005]) diversities. A relatively recent approach to quantifying alpha diversity using sequence data is phylogeny-aware molecular species delimitation (Fujisawa & Barraclough, 2013; Kapli et al., 2017; Yang, 2015; Zhang et al., 2013). These methods rely on a given phylogenetic tree to identify species boundaries, essentially resulting in a clustering of the tips into distinct species.

Here, we combine previous work on phylogenetic placement (Barbera et al., 2018b) and species delimitation (Kapli et al., 2017; Zhang et al., 2013) to devise a measure of phylogeny-aware relative alpha diversity. Our SCRAPP (Species Counting on Reference trees viA Phylogenetic Placement) tool quantifies diversity by initially grouping QS by the branch on the reference tree (reference branch) to which they most likely belong with respect to their phylogenetic likelihood score. Subsequently, for each such group of QS placed onto the same reference branch, we infer a separate phylogenetic tree comprising the QS of that group, optionally including an outgroup sequence from the reference tree. We call such a tree a branch query phylogeny (BQP). Generating such BQP constitutes a major part of the analysis (in terms of run time) and is a feature that has, thus far, been missing for post analysing phylogenetic placements. Therefore, we include the set of inferred BQP in the SCRAPP output.

Finally, we apply mPTP (Kapli et al., 2017) to the BQP to obtain a species count for the corresponding reference branch. The output of SCRAPP is a branch-annotated reference tree that depicts how species diversity is distributed over the reference tree for a given sample.

SCRAPP is implemented in PYTHON and relies on mpi4py (Dalcin et al., 2011; Dalcín et al., 2005, 2008) for the respective parallel implementation targeting both, shared and distributed memory systems.

Some concepts are based on our admittedly difficult to use EPA-PTP tool, an early attempt to integrate phylogenetic placement with species delimitation (Zhang et al., 2013). The goal of SCRAPP is thus to quantify diversity for each branch of the reference tree individually and to improve usability. In contrast to SCRAPP, EPA-PTP used phylogenetic placement to calculate a single, overall species delimitation over the entire reference tree extended by all BQPs simultaneously.

We wish to emphasize that, while our focus here is to demonstrate the utility of SCRAPP for microbial data, applying it to plant or animal data is also possible.

2 | DESCRIPTION

In the following, we initially provide a detailed description of the SCRAPP tool. An overview is provided in Figure 1. SCRAPP takes as input a jplace (Matsen et al., 2012) file containing the placements and the associated reference tree, as well as the corresponding multiple sequence alignment (MSA) of the QS. From this, we generate per-branch QS MSAs. These include all QS whose most likely placement was on the given branch. However, we remove those placements from this set, whose best likelihood weight ratio (LWR, von Mering et al., 2007) is below a given threshold (--min-weight, default 0.5).

If desired, an outgroup from a user-specified reference MSA is included in each branch QS MSA such that the corresponding BQP that is produced in the subsequent step can be rooted at this outgroup. We automatically choose the outgroup sequence for a given BQP as the leaf sequence in the reference tree that is most distant from the given branch. Note that, mPTP species delimitation operates on rooted phylogenies. Thus, specifying an outgroup can be beneficial if a more reliable root for the BQP is desired. If a root is not provided, mPTP will automatically root the BQP on its longest branch.

If the number of QS in a given branch QS MSA exceeds a user-specified maximum (500 by default), we reduce the number of QS to that maximum using the two-stage clustering method described in Section 2.1. This option is necessary to maintain BQP tree inference times within reasonable limits. On empirical data sets, specific reference branches can contain more than 100,000 QS; hence, yielding the inference of a BQP computationally challenging. We strongly recommend that the QS are dereplicated or even OTU-clustered prior to executing SCRAPP, or, for that matter, prior to performing placement.

Once the query MSAs have been generated for all branches of the reference tree, we infer a phylogeny for each of them separately using RAxML-NG (Kozlov et al., 2019). As there may be a large number of trees (potentially as many as there are branches in the reference tree) with highly variable sizes to infer, we use ParGenes (Morel et al., 2018) to orchestrate this tree inference process in a parallel, scalable and efficient way. The inferred BQPs are then processed using mPTP to obtain a species delimitation, and corresponding species count. The information produced by each mPTP run is tracked for each branch that contains QS in the reference tree.

We note that the species delimitation itself constitutes a clustering of the QS, which may represent a desirable output to the user. Particularly, if the original placement input data have *not* already been OTU clustered, the combination of placement with species

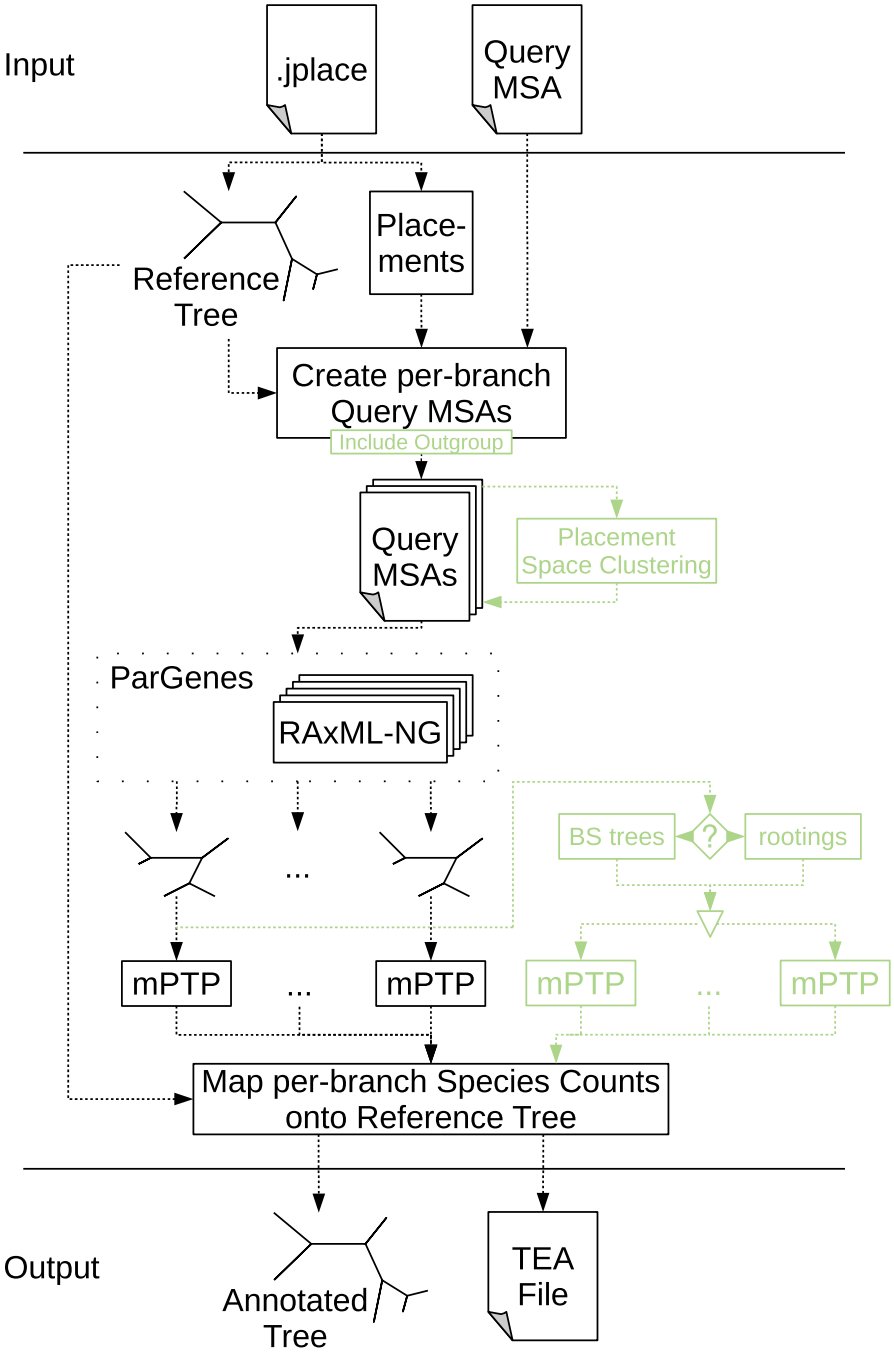

FIGURE 1 Overview of the major components of the SCRAPP pipeline. In green, we highlight optional components (inclusion of reference sequences for BQP outgroup rooting, *placement space clustering* for limiting computational effort, bootstrapping or re-rooting for delimitation variance assessment) [Colour figure can be viewed at wileyonlinelibrary.com]

delimitation can be regarded as phylogeny-aware OTU clustering. However, in this work we focus on the diversity metric aspects of SCRAPP and consider further potential applications as future work.

The set of inferred BQP can optionally be expanded to calculate species count variance. Two options are available to calculate this variance: *rootings*, generates a tree set on each BQP by enumerating all possible rootings for the unrooted BQP or *bootstraps*, generates a given number (20 by default) of bootstrapped branch QS MSAs and then re-optimizes the branch lengths on the original BQP for each of the bootstrapped branch QS MSAs. When using these expanded BQP sets, we calculate the final species count as median over all per-branch species delimitation results (i.e. over all rootings or all bootstrap replicates).

The *rootings* and *bootstraps* options constitute two of the three principal operating modes of SCRAPP. The third operating mode, the *outgroup* mode, offers the rooting of the BQP via inclusion of a reference outgroup (as described above).

Finally, SCRAPP generates two types of output files. Firstly, it outputs an annotated version of the reference tree in the extended NEWICK format, that can easily be visualized by a number of tree viewers (e.g. iTOL [Letunic & Bork, 2006] or Dendroscope [Huson et al., 2007]). This is useful for obtaining a high level overview of the diversity, as diversity is represented by just one species count value per reference tree branch.

To allow users to explore the results more thoroughly, for example, by inspecting the variance of the median species count, we also

produce a comprehensive output file in a json-based file format that is analogous to the jplace format (Matsen et al., 2012). This format, called Tree Edge Annotations (TEA), contains the reference tree with enumerated branches, as specified in jplace, followed by annotation information. The annotation comprises a list of per-branch values. In SCRAPP, this annotation includes the median species count, and the species count variance, among others. We provide a full specification and an example of the TEA format in the supplement, as well as online at https://github.com/pbdas/scrapp/wiki/TEA-format.

2.1 | Placement space clustering

In general, phylogenetic diversity metrics face a fundamental scalability issue, as they rely on a phylogeny inferred on the QS. With increasing sequencing volumes, inferring such phylogenies under maximum likelihood becomes prohibitively expensive. Moreover, as metabarcoding/metagenomic samples typically comprise short sequences, the available signal for reliable tree inference on thousands or tens of thousands of taxa is mostly insufficient (Bininda-Emonds et al., 2000). This was the key motivation for the development of phylogenetic placement methods as a scalable and more reliable alternative.

Nonetheless, SCRAPP faces this same computational issue again at a different level as a reference branch may contain tens of thousands of QS. To alleviate this, we have implemented a two-stage clustering method called placement space clustering (PSC) in SCRAPP. PSC leverages the fact that the insertion location of a maximum likelihood placement of QS along the reference branch (the so-called proximal length), and distance from that reference branch (the so-called pendant length) can be interpreted as an embedding into a two-dimensional euclidean space (hereafter called *placement space*). When using PSC, we map the set of placements on a branch into placement space and then perform a standard *k*-means clustering on the respective datapoints. Subsequently, we select a small number *x* of placements from each cluster as representatives of that cluster, such that k^*x equals the maximum desired number (as specified by the user) of sequences per branch QS MSA. More specifically, we select the top $x := 10$ sequences by number of informative (non-gap or non-undetermined) sites, thereby maximizing the potential phylogenetic signal for the subsequent tree inference.

3 | EVALUATION

We assessed the accuracy of SCRAPP using both, simulated and empirical data.

3.1 | Simulated data

We generated *true* species trees using the MSPRIME (Kelleher et al., 2016, version 0.7.3) coalescent simulator. We then used SEQ-GEN (Rambaut & Grass, 1997, version 1.3.4) to generate MSAs on those trees. We generated the trees and MSAs such as to evaluate SCRAPP under a broad range and combination of simulation parameters. The parameters include the following: the number of starting populations (which we call *species*) ([200,600]), the sequence length ([1,000,4,000]), the number of individuals per population (called *sample size* by MSPRIME) ([20,80]), the overall MSPRIME population size ($[10^5,10^7]$) and the mutation rate ($[10^{-7},10^{-8}]$). In particular, we investigated the influence on each parameter individually while keeping the remaining parameters fixed to a set of default values (see Supplement for details).

From each simulated *true* tree and MSA, we first pruned a set of QS by removing all but one individual from each starting population. To account for incomplete reference data with lower taxon sampling density, we subsequently further pruned a given fraction (denoted as prune_fract, [0.1,0.4]) of leaves uniformly at random from the trees. We then labelled the branches of the remaining reference tree by the number of query species (here assumed to be equal to the number of populations) whose *true* location is on that given branch.

We then used EPA-NG to place the query data back onto the tree. Next, we evaluated these phylogenetic placement results using SCRAPP, yielding an annotated NEWICK tree. Finally, we compare the reference tree with the inferred species count annotations (hereafter SCRAPP-tree) to the reference tree with the *true* species count annotations.

All scripts used for generating the simulated data can be found in the SCRAPP repository: https://github.com/Pbdas/scrapp/tree/master/simtest

3.2 | Empirical data

In addition to the tests on simulated data, we replicated part of the evaluation of (McCoy and Matsen, 2013). McCoy and Matsen IV evaluated different diversity metrics by the quality of their fit with clinical metadata, which are known from literature to correlate with alpha diversity.

We chose to replicate and extend the evaluation of the Bacterial Vaginosis data set (Srinivasan et al., 2012; hereafter called BV), as we already had access to the data and the specific data set has been particularly well studied (Czech & Stamatakis, 2019). The clinical metadata included in the BV data set are based on two methodologies indicating the presence or absence of bacterial vaginosis for a patient: Amsel's criteria (Amsel et al., 1983), and the Nugent score (Nugent et al., 1991). Amsel's criteria comprise four distinct criteria, three of which need to be fulfilled to positively diagnose a patient with bacterial vaginosis. In the BV data set, 'Amsel' is provided as a binary value indicating whether a patient was diagnosed as positive or negative. The Nugent score is a composite score based on gram-stained vaginal swabs. The score ranges from negative (0–3), through intermediate (4–6) to positive (7–10).

Unfortunately, due to patient data protection issues, we cannot make the BV data set publicly available. Please refer to (Srinivasan

et al., 2012) and (Czech & Stamatakis, 2019) for an exhaustive exploration of the BV data set, and a detailed description of the phylogenetic placement of the per-sample data, respectively.

Firstly, to obtain the OTU-derived diversity measures used in the evaluation of (McCoy and Matsen, 2013), we performed OTU clustering using SWARM (Mahé et al., 2014, 2015, version 3.0, -d 1 -f) and utilizing VSEARCH (Rognes et al., 2016, version 2.6.2) for dereplication and filtering. We further analysed the resulting OTU table using the R package PHYLOSEQ (McMurdie & Holmes, 2013, version 1.22.3, function estimate_richness) to obtain the Shannon (Shannon, 1948), Simpson (Simpson, 1949), ACE (Chao & Lee, 1992) and Chao1 (Colwell & Coddington, 1994) indices.

Secondly, to assess the placement based methods, we computed a phylogenetic placement of the sample data. Note that, we did not use the reference tree given in the original publication (Srinivasan et al., 2012), as we found that the inclusion of multiple strains of the same bacterial species can produce a very flat likelihood distribution for potential placements of a single QS across individual branches of the tree (Czech & Stamatakis, 2019). Therefore, we used an appropriately modified version of the reference tree, as shown in Figure S1 in Czech and Stamatakis (2019). This modified reference tree only retains consensus sequences of all reference strains, such that only one taxon per species remains. The modified reference tree comprises 198 taxa.

Based on this placement data, we obtained the measures outlined in McCoy and Matsen (2013), on a per-sample basis, using the guppy command fpd (Matsen et al., 2010; McCoy & Matsen, 2013). Note that, we chose to omit the guppy fpd—include-pendant option to avoid overestimating diversity. The placement process does not resolve relationships between individual QS. Thus, the distance of each individual QS to the RT is denoted by a so-called pendant length. Consequently, if two or more QS are phylogenetically close to each other, but relatively distant to the RT, the common distance to the RT may be counted once per QS in the PD calculation. This can lead to potential overestimation.

Lastly, we applied SCRAPP to the placement data, running the analysis in the bootstrap operating mode and limiting the maximum number of taxa per BQP to 1,000. This again yields a SCRAPP-tree (see Section 3.1).

In the interest of comparability, we chose to re-implement the Balance Weighted Phylogenetic Diversity (BWPD) function using the genesis library (Czech et al., 2020), in a way such that it can be applied to SCRAPP-trees. The BWPD relies on a one-parameter function family interpolating between classical PD and an abundance weighted version of the PD. McCoy and Matsen IV chose to implement and evaluate the BWPD on placement results, which consist of precise locations and branch lengths of queries on the reference tree. In contrast to this, SCRAPP-trees comprise assignments of absolute numbers (species counts) to branches of the tree, without any more specific branch length information. To remedy this discrepancy, when calculating the BWPD on a SCRAPP-tree, we treat the species count of a branch as if it were a single placement, located at the middle of said branch, without a pendant length.

All data handling and analysis scripts used in the empirical data evaluation can be accessed online at https://github.com/Pbdas/diversity-compare.

3.3 | Clustering and showcase

Finally, we include a showcase test and analysis for two additional empirical data sets.

In one set of experiments, we use data from an study of eukaryotic community composition in neotropical soils (Mahé et al., 2017) to evaluate our PSC methodology (Section 2.1). These data are particularly challenging for phylogenetic placement, as the available reference data are too sparse to cover the diversity that was sampled. We will refer to this data set as the NEOTROP data set. For our purposes, we randomly selected small subsets of 1,000 QS from this data set and placed them on the reference tree described in Mahé et al. (2017; 512 reference taxa). We then executed SCRAPP for distinct settings of—cluster-above, thereby limiting the maximum number of sequences per branch used in the subsequent BQP tree searches. As the randomly selected 1,000 QS subsets produced a maximum of 298 QS placements per branch, a threshold value of 300 was selected as the benchmark against which all other runs are compared to, as this constitutes the 'no clustering' case. For each clustering threshold setting and each operating mode, we performed 5 independent runs of the same data in order to quantify variability introduced by the randomization component in the clustering algorithm. Scripts and data used in this experiment can be found in the repository at https://github.com/Pbdas/scrapp/tree/master/test.

In a second set of experiments, we used a large data set from the UniEuk project (Berney et al., 2017) as a showcase for deploying SCRAPP on a standard parallel compute cluster. For this test, we used a phylogenetic placement of 585,050 QS on a reference tree comprising 800 taxa, which resulted from an OTU clustering of roughly 300 million sequences (respective article in press). Unfortunately, the data set has not yet been published, so we are yet unable to make it available. From this, SCRAPP identified 254,103 QS as being placed with a LWR above the default 0.5 threshold (see Section 2). We limited the maximum number of sequences per branch to 800 and used the *bootstrap* operating mode, generating 100 bootstrap trees per BQP. This resulted in the inference of 1,070 trees, the largest tree containing 797 taxa. SCRAPP further evaluated each of them via 100 distinct bootstrap MSAs.

3.4 | Error metrics

For the simulated data, we calculate two distinct accuracy values. The first is the absolute difference between the inferred and the true species count on a branch in the reference tree. This absolute difference is then averaged over all branches of the reference tree that have non-zero values in either tree. We denote this accuracy metric as Mean Absolute per-branch Error (hereafter MAE).

More formally, let S and T be two trees with identical topologies and branch-associated values s_i and t_i for a given branch index i, respectively. T denotes the *true* tree, while S denotes the *SCRAPP-tree* (Section 3.1). Let B be the set of branch indices for which either S or T have non-zero values. We can now write the MAE as

$$\mathrm{MAE} = \frac{\sum_{i \in B} |t_i - s_i|}{|B|} \quad (1)$$

Our second accuracy metric is based on normalized per-branch species counts. For a given branch with index k, we calculate this normalized count based on a absolute species count x_k as

$$x_k^{\mathrm{norm}} = \frac{x_k}{\sum_{i \in B} x_i} \quad (2)$$

where k denotes the index of a given branch, and B is as defined above.

Further, instead of calculating the absolute difference, we calculate the relative difference:

$$\mathrm{rel}(t_k, s_k) = \frac{|t_k - s_k|}{(|t_k| + |s_k|)/2} \quad (3)$$

Again, s_k and t_k are the values for a given branch with index k, of two given trees S and T as defined above. Note that here we compute the relative difference by normalizing via the arithmetic mean of s_k and t_k. This ensures that the metric produces well-defined values in cases where $t_k = 0$. The term $\mathrm{rel}(t_k, s_k)$ is also known as the *Relative Percent Difference*. Note that $\mathrm{rel}(t_k, s_k)$ is bounded between 0 and 2.

Finally, we again calculate the average over all relative normalized species count differences across all branches that have non-zero value, resulting in the Normalized Mean Relative per-branch Error (NMRE).

$$\mathrm{NMRE} = \frac{\sum_{i \in B} \mathrm{rel}\left(t_i^{\mathrm{norm}}, d_i^{\mathrm{norm}}\right)}{|B|} \quad (4)$$

The MAE captures the deviation of the SCRAPP-based species count from the true species count. The NMRE quantifies the difference between the true and the inferred diversity distribution over the tree.

The accuracy of the methods used in the empirical evaluation is calculated as in McCoy and Matsen (2013). Here, the primary approach is to assess the correlation of the diversity measures with the clinical metadata (see Section 3.2). To quantify the correlation with the diagnosis based on Amsel's criteria, we first use the GLM function in R (R Core Team, 2017) to fit a generalized linear model to the data. We then calculate the *Amsel accuracy* as the proportion of correctly identified datapoints via a leave-one-out cross-validation. As McCoy and Matsen IV we perform independent 2-group t tests between the Amsel diagnosis and the investigated metrics, using the t test R function. The resulting p-value is presented here as the *Amsel p-value*. For comparing against the Nugent score, we fit the diversity measures using a linear regression model, via the lm function in R. The function also returns the R^2 of the fit, which is the proportion of the variation that is explained by the model.

TABLE 1 We report the mean NMRE and mean MAE, across all runs (last row) and across all runs of the specific operating modes (middle rows)

	NMRE	σ^2	CV	MAE	σ^2	CV
bootstrap	0.518	0.013	0.221	5.69	2.99	0.304
outgroup	0.535	0.017	0.241	7.71	5.48	0.304
rootings	0.471	0.019	0.289	8.15	5.76	0.294
across-all	0.508	0.016	0.254	7.18	5.86	0.337

Note: σ^2 denotes the variance of the given means, and CV denotes the coefficient of variation. As a reference, the mean variance among simulation replicates (identical parameter configurations but different random number seeds) was 1×10^{-3} and 3×10^{-2} for the NMRE and the MAE, respectively.

4 | RESULTS

4.1 | Simulated data

We performed a total of 270 independent simulation runs, covering all simulation space dimensions, all of their combinations with the SCRAPP operating modes and repeating runs for each individual configuration five times. We show high level results across all runs and stratified by operating mode, in Table 1. We observe a mean NMRE of 0.508 over all experiments. When stratified by the different operating modes, we observe the lowest overall NMRE for the rootings mode (0.471 mean NMRE).

To summarize our exploration of the impact of different simulation parameters, we find that result accuracy in terms of mean NMRE increases with increasing overall population size, sample size (number of individuals per population) and sequence length, as well as decreasing prune_fract (Section 3.1). While less pronounced, there is a trend for the NMRE to improve with increasing total tree size which may be attributed to improved taxon sampling density. This can be observed in Figure 2, which shows data for those simulation runs where we only varied the total number of starting populations (here called species).

Further, we observe a general trend for overestimating the species count across all simulation parameters, as indicated by the high MAE values (Table 1). Specifically, the *rootings* mode appears to overestimate the species count the most, while the *bootstrap* mode performs best in this regard. We therefore recommend that users deploy the *bootstrap* mode when the goal is to obtain as accurate as possible estimates of the absolute species counts. However if, one desires to obtain the most accurate *relative* distribution of species counts over the tree, we recommend the *rootings* mode, as it consistently showed the lowest NMRE.

Further, we observe a divergent relationship between the MAE and NMRE scores for the population size, sample size,

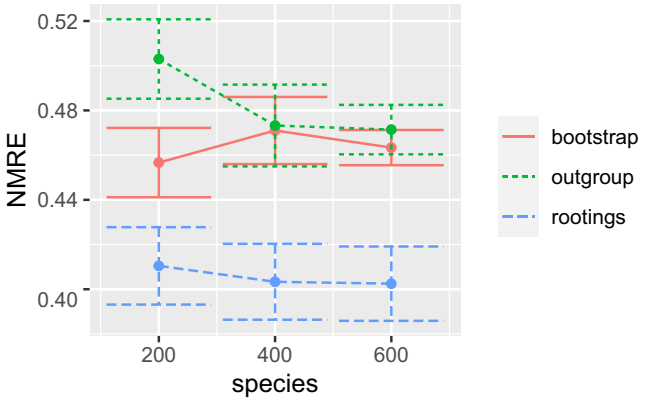

FIGURE 2 NMRE (Equation 4) for several runs on simulated data sets where we only varied the total species count of the 'true' tree (the number of individual populations). Error bars denote the first standard deviation from the mean. Data were stratified by the three different operating modes of SCRAPP (see Section 2) [Colour figure can be viewed at wileyonlinelibrary.com]

sequence length and species parameters. For the first three parameters, this is due to a decrease in the fraction of mPTP delimitation results that yield the null model. Note that, when mPTP yields the null model, it cannot distinguish between a delimitation into one or *n* species, where *n* is the number of leaves in the tree (the BQP in our case). As the fraction of null model results decreases, the *relative* mPTP accuracy increases, yielding more accurate results with respect to the NMRE metric. At the same time, this also increases the average *absolute* species count and thereby generates higher MAE values.

For the species parameter (the number of populations in the coalescent simulation), the negative relationship between MAE and NMRE is less pronounced. It can be explained by the fact that an increase in the species parameter yields a larger simulated tree, but, at the same time, unlike the other three parameters does not increase the phylogenetic signal for reconstructing the BQPs. As a consequence, the fraction of mPTP null model results remains constant over the species parameter. Further, as phylogenetic placement is not exact, a larger reference tree with an increased number of branches also implies a larger potential for misplacing QS. This increases the chance of reference tree branches for which the true number of placed QS should be 0, to contain misplaced QS, and thereby yield a minimum species count of 1. As the delimited species to which a misplaced QS belongs may already be accounted for on another branch, the total species count increases. As a result, the MAE will increase as well.

For detailed figures exploring the effect of varying individual simulation parameters on the MAE and NMRE metrics, as well as the fraction of null model results, please refer to the supplement.

4.2 | Empirical data

The most important results of our evaluation based on the BV data set are shown in Table 2. We were able to closely replicate the results of (McCoy and Matsen, 2013; their Table 2), although we observe generally higher values for the Amsel accuracy and Nugent R^2. The exception to this is the R^2 values obtained from the ACE and Chao1 measures, that substantially underperform compared with the results of (McCoy and Matsen, 2013). As ACE and Chao1 are the only tested OTU-based metrics that specifically assign a higher weight to rare observations (i.e. OTUs observed only once or twice), we speculate that our data handling approach has reduced the number of rare OTUs. However, our results confirm the general trend that phylogenetic methods outperform OTU methods with respect to the aforementioned metrics.

Further, we observe a high level of agreement between metrics directly calculated from phylogenetic placement results, and metrics derived from SCRAPP results.

TABLE 2 Correlation and predictive power of SCRAPP in comparison with analogous approaches on the Bacterial Vaginosis data

Measure	Amsel accuracy	Nugent R^2	Amsel *p*-value	Mean rank
bwpd_0.25.guppy	0.877	0.777	2.01e−35	2.33
bwpd_0.25.scrapp	0.874	0.785	4.02e−34	2.33
phylo_entropy.scrapp	0.873	0.782	4.70e−34	4.00
bwpd_0.5.guppy	0.873	0.757	1.03e−34	4.33
bwpd_0.5.scrapp	0.872	0.786	1.37e−33	4.67
bwpd_0.scrapp	0.873	0.767	1.49e−33	5.67
quadratic.scrapp	0.869	0.779	1.60e−32	8.33
bwpd_0.75.guppy	0.870	0.725	2.46e−33	9.00
bwpd_0.75.scrapp	0.868	0.772	5.10e−32	10.33
quadratic.guppy	0.869	0.718	7.97e−33	10.33
bwpd_0.guppy	0.872	0.713	2.00e−31	11.17
unrooted_pd.guppy	0.872	0.713	2.00e−31	11.17
phylo_entropy.guppy	0.869	0.716	1.43e−32	11.33
rooted_pd.guppy	0.871	0.701	5.73e−31	13.00
bwpd_1.scrapp	0.861	0.741	1.30e−29	13.67
bwpd_1.guppy	0.867	0.691	8.36e−32	14.33
Shannon	0.826	0.387	5.03e−18	17.00
ACE	0.822	0.242	1.41e−10	18.00
Chao1	0.810	0.213	6.35e−09	19.00
Simpson	0.788	0.168	3.61e−08	20.00

Note: Amsel accuracy, Nugent R^2, Amsel *p*-value and mean rank are calculated exactly as in McCoy and Matsen (2013). Rows are sorted by mean rank. Measures suffixed by ".guppy" are calculated using guppy fpd (Matsen et al., 2010), whereas measures suffixed by ".scrapp" were calculated based on results produced by SCRAPP. Shannon, ACE, Chao1 and Simpson values were calculated based on an OTU clustering of the same data (see Section 3.2).

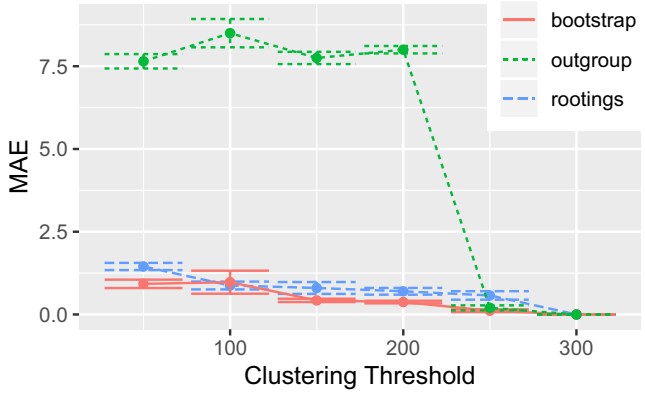

FIGURE 3 MAE (Equation 1) of multiple runs of SCRAPP, using different thresholds down to which placement space clustering (PSC) reduces the maximum per-branch data volume. The MAE is calculated with reference to the case of the threshold being 300, as 300 was the maximum number of QS that were placed per-branch. The underlying query and reference data are from the neotrop data set (Section 3.3, (Mahé et al., 2017)) [Colour figure can be viewed at wileyonlinelibrary.com]

4.3 | Clustering and Showcase

The results of evaluating PSC with varying clustering thresholds are shown in Figure 3. Both, the *bootstrap* and *rootings* operating modes produced stable results, which are qualitatively similar to the tests on simulated data. However, the *outgroup* operating mode proved to be highly sensitive to the PSC, yielding high species count deviations starting at a clustering threshold of 200 (a data reduction of approx. 33%). Due to the known issues with the eukaryotic soil reference data set at hand, we hypothesize that the cause for this behaviour is the sparse taxon sampling in the reference MSA. This incomplete taxon sampling induces a high branch length distance between the ingroup QS and the outgroup, as SCRAPP selects the phylogenetically most distant taxon in the reference tree as outgroup.

As a final showcase for the scalability of SCRAPP on distributed computing clusters, we performed an analysis of a large data set of 585,050 QS placed on a 800 taxon reference tree, utilizing 50 compute nodes comprising a total of 800 cores. Running this analysis involved handling about 1 million files, of which approximately 8,500 had to be retained as intermediate results for further downstream analysis. The total runtime under this setting was 26.5 hr. We regard this as being fast, since the overall computational task includes hundreds of tree inferences with up to 797 taxa and handling approximately 1 million intermediate files.

5 | CONCLUSION

We presented SCRAPP, a highly scalable and fully automated pipeline for diversity quantification of phylogenetic placement data. The primary goal of SCRAPP is to quantify the diversity distribution of a given sample over the reference tree. We show that, on simulated data sets, SCRAPP yields phylogenetic diversity distributions with a comparatively low per-branch error rate. On empirical data, we show that alpha diversity metrics calculated on the results obtained from SCRAPP rank among the top of those tested in terms of predictive power, and correlation with clinical metadata.

By using MPI (Message Passing Interface), SCRAPP achieves a high level of parallelism, enabling the user to use an arbitrary number of cores in a cluster computing environment. In a selected showcase, we were able to run SCRAPP on a data set with 585,050 QS on 50 cluster nodes, using a total of 800 cores, in 26.5 hr. This run involved hundreds of tree inferences with up to 797 taxa, and the handling of approximately 1 million intermediate files.

Using placement space clustering, a novel clustering method for placements, SCRAPP is able to to efficiently perform dimensionality reduction of the branch QS MSA input data. This enables SCRAPP to tackle the scalability challenge induced by the metagenomic and metabarcoding data flood. Finally, it should be noted that issues with the underlying reference data regarding taxon sampling density may negatively affect the results when clustering is used.

ACKNOWLEDGEMENTS

We thank S. Srinivasan and E. Matsen for providing the Bacterial Vaginosis dataset (Srinivasan et al., 2012). We extend our thanks to Paschalia Kapli and Aggelos Koropoulis for discussions and advice regarding the simulations. We thank Cédric Berney and Laura Rubinat-Ripoll for the insight that led to the development of placement space clustering. This work was financially supported by the Klaus Tschira Foundation.

AUTHOR CONTRIBUTIONS

The software was designed by P.B., L.C. and A.S.; software was implemented by P.B. and L.C.; a feasibility prestudy was performed by S.L.; manuscript was written by P.B. and A.S.

DATA AVAILABILITY STATEMENT

The *neotrop* data set is included in the repository at https://github.com/Pbdas/scrapp/tree/master/test, as used in the evaluation of the placement space clustering. Data from the *neotrop* dataset can also be accessed via the supplementary repository to our previous publication, at https://doi.org/10.5061/dryad.kb505nc (Barbera et al., 2018a). Available in the same repository is also a anonymized version of the *BV* dataset; however, no disambiguation into individual samples can be obtained from it, as this was one of the goals of patient data protection. For the same reason, we are unable to share the data set, as used in this work, publicly. The *UniEuk*-associated data set used in the showcase is not yet published, and we are unable to share it. Simulated data may be recreated from the provided scripts and instructions.

ORCID

Pierre Barbera https://orcid.org/0000-0002-3437-150X
Lucas Czech https://orcid.org/0000-0002-1340-9644
Alexandros Stamatakis https://orcid.org/0000-0003-0353-0691

REFERENCES

Amsel, R., Totten, P. A., Spiegel, C. A., Chen, K. C., Eschenbach, D., & Holmes, K. K. (1983). Nonspecific vaginitis: Diagnostic criteria and microbial and epidemiologic associations. *The American Journal of Medicine*, *74*(1), 14–22. https://doi.org/10.1016/0002-9343(83)91112-9

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2018a). Data from: EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Dryad*. https://doi.org/10.5061/dryad.kb505nc

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2018b). EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, *68*(2), 365–369. https://doi.org/10.1093/sysbio/syy054

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

Berney, C., Ciuprina, A., Bender, S., Brodie, J., Edgcomb, V., Kim, E., & de Vargas, C. (2017). UniEuk: Time to speak a common language in protistology! *Journal of Eukaryotic Microbiology*, *64*(3), 407–411. https://doi.org/10.1111/jeu.12414

Bininda-Emonds, O. R., Brady, S., Kim, J., & Sanderson, M. J. (2000). Scaling of accuracy in extremely large phylogenetic trees. In R. B. Altman A. K. Dunker L. Hunter & T. E. Klein (Eds.), *Biocomputing 2001* (pp. 547–558). World Scientific.

Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, *87*(417), 210–217. https://doi.org/10.1080/01621459.1992.10475194

Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *345*(1311), 101–118.

Czech, L., Barbera, P., & Stamatakis, A. (2020). Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics*, *36*, 3263–3265. https://doi.org/10.1093/bioinformatics/btaa070

Czech, L., & Stamatakis, A. (2019). Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *PLoS One*, *14*(5), 1–50. https://doi.org/10.1371/journal.pone.0217050

Dalcin, L. D., Paz, R. R., Kler, P. A., & Cosimo, A. (2011). Parallel distributed computing using Python. *Advances in Water Resources*, *34*(9), 1124–1139. https://doi.org/10.1016/j.advwatres.2011.04.013

Dalcín, L., Paz, R., & Storti, M. (2005). MPI for Python. *Journal of Parallel and Distributed Computing*, *65*(9), 1108–1115. https://doi.org/10.1016/j.jpdc.2005.03.010

Dalcín, L., Paz, R., Storti, M., & D'Elía, J. (2008). MPI for Python: Performance improvements and MPI-2 extensions. *Journal of Parallel and Distributed Computing*, *68*(5), 655–662. https://doi.org/10.1016/j.jpdc.2007.09.005

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

Fujisawa, T., & Barraclough, T. G. (2013). Delimiting species using single-locus data and the generalized mixed yule coalescent approach: A revised method and evaluation on simulated data sets. *Systematic Biology*, *62*(5), 707–724. https://doi.org/10.1093/sysbio/syt033

Helmus, M., Bland, T., Williams, C., & Ives, A. (2007). Phylogenetic measures of biodiversity. *The American Naturalist*, *169*(3), E68–E83. https://doi.org/10.1086/511334

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., & Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, *8*(1), 460. https://doi.org/10.1186/1471-2105-8-460

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., & Burki, F. (2020). Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, *20*(2), 429–443. https://doi.org/10.1111/1755-0998.13117

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., & Flouri, T. (2017). Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics*, *33*(11), 1630–1638. https://doi.org/10.1093/bioinformatics/btx025

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, *12*(5), 1–22. https://doi.org/10.1371/journal.pcbi.1004842

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*, 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Letunic, I., & Bork, P. (2006). Interactive tree of life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128.

Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., Singer, D., Mayor, J., Bunge, J., Sernaker, S., Siemensmeyer, T., Trautmann, I., Romac, S., Berney, C., Kozlov, A., Mitchell, E. A. D., Seppey, C. V. W., Egge, E., Lentendu, G., … Dunthorn, M. (2017). Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution*, *1*(4), 0091. https://doi.org/10.1038/s41559-017-0091

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. https://doi.org/10.7717/peerj.593

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420. https://doi.org/10.7717/peerj.1420

Matsen, F. A., Hoffman, N. G., Gallagher, A., & Stamatakis, A. (2012). A format for phylogenetic placements. *PLoS One*, *7*(2), 1–4. https://doi.org/10.1371/journal.pone.0031009

Matsen, F. A., Kodner, R. B., & Armbrust, V. E. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BioMed Central Bioinformatics*, *11*(1), 1–16. https://doi.org/10.1186/1471-2105-11-538

McCoy, C. O., & Matsen, F. A. IV (2013). Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*, *1*, e157. https://doi.org/10.7717/peerj.157

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, *8*(4), 1–11. https://doi.org/10.1371/journal.pone.0061217

Morel, B., Kozlov, A. M., & Stamatakis, A. (2018). ParGenes: A tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics*, *35*(10), 1771–1773. https://doi.org/10.1093/bioinformatics/bty839

Nugent, R. P., Krohn, M. A., & Hillier, S. L. (1991). Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. *Journal of Clinical Microbiology*, *29*(2), 297–301. https://doi.org/10.1128/JCM.29.2.297-301.1991

R Core Team. (2017). *R: A language and environment for statistical computing [Computer software manual]*. Austria.

Rambaut, A., & Grass, N. C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(3), 235–238. https://doi.org/10.1093/bioinformatics/13.3.235

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688. https://doi.org/10.1038/163688a0

Srinivasan, S., Hoffman, N. G., Morgan, M. T., Matsen, F. A., Fiedler, T. L., Hall, R. W., Ross, F. J., McCoy, C. O., Bumgarner, R., Marrazzo, J. M., & Fredricks, D. N. (2012). Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLoS One*, *7*(6), 1–15. https://doi.org/10.1371/journal.pone.0037818

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., Grenyer, R., Helmus, M. R., Jin, L. S., Mooers, A. O., Pavoine, S., Purschke, O., Redding, D. W., Rosauer, D. F., Winter, M., & Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715. https://doi.org/10.1111/brv.12252

von Mering, C., Hugenholtz, P., Raes, J., Tringe, S. G., Doerks, T., Jensen, L. J., Ward, N., & Bork, P. (2007). Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, *315*(5815), 1126–1130. https://doi.org/10.1126/science.1133420

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, *61*(5), 854–865. https://doi.org/10.1093/czoolo/61.5.854

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, *29*(22), 2869–2876. https://doi.org/10.1093/bioinformatics/btt499

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Barbera P, Czech L, Lutteropp S, Stamatakis A. SCRAPP: A tool to assess the diversity of microbial samples from phylogenetic placements. *Mol Ecol Resour* 2021;21:340–349. https://doi.org/10.1111/1755-0998.13255