Deriving phylogenetic diversity metrics from barcode-based reference phylogenies in metabarcoding assays: modifying the QIIME2 workflow

Joppe Wesseling
S1079688

February 2025

Summary

This report presents a novel approach for analysing COI-5P metabarcoding data by integrating QIIME2 with external reference databases and phylogenetic trees. It addresses key challenges in biodiversity assessment, particularly the need for standardized phylogenetic diversity (PD) metrics that enable cross-study comparisons. The proposed workflow utilizes QIIME2 alongside Bio-Phylo-Forest-DBTree, HMMER and pplacer to derive PD metrics from raw sequencing data. By leveraging a curated reference tree, this approach ensures more consistent and biologically meaningful phylogenetic diversity estimates. The report outlines each step of the workflow, including sequence preprocessing, alignment, phylogenetic placement, and metric calculation, detailing both the methodological considerations and implementation strategies. The results from a proof-of-concept study demonstrate that integrating reference-based phylogenetic placement improves the reliability of PD metrics. Specifically, samples from geographically distinct regions clustered in expected biogeographical patterns, highlighting the potential of this approach for more robust biodiversity assessments. This work establishes a foundation for further refinement and broader application of reference-based phylogenetic placement in metabarcoding studies, with implications for ecological monitoring and conservation research.

Contents

1. Introduction

Accurate biodiversity assessment is fundamental to ecology, conservation, and environmental management. Traditional methods, such as field surveys, are labour-intensive and have limitations in scalability, particularly for microbial and large-scale ecosystem monitoring. As global biodiversity continues to decline at an alarming rate, the need for accurate, scalable, and efficient methods to monitor and assess biodiversity has become increasingly urgent.

Traditional field surveys have been the cornerstone of biodiversity assessments for decades. However, these methods are labour-intensive, time-consuming, not feasible for microorganisms and often limited in scope, making it difficult to scale them to the level required for biodiversity monitoring of larger communities [2]. As a result, there has been a growing interest in alternative approaches that can provide rapid, accurate and large-scale assessments. One such approach is metabarcoding, a technique that leverages high-throughput sequencing to identify species from environmental DNA (eDNA) samples based on specific genetic markers. While metabarcoding has the potential to transform biodiversity assessments by enabling the analysis of complex and diverse communities it is important to note that it will not solve challenges on smaller scale biodiversity analysis [3].

While species lists generated by metabarcoding provide valuable insights into community composition, they often fall short of capturing the full spectrum of biodiversity. Phylogenetic diversity (PD) metrics are a valuable tool, which, by incorporating evolutionary relationships among species, provide a more nuanced understanding of biodiversity. However, these metrics are commonly generated using samples itself and cannot be directly compared between different studies. For such comparisons to be valid, it would be imperative to calculate the metrics from not only the samples, but these samples placed on a common, well curated, and complete reference tree.

When considering a metabarcoding reference tree, particularly for the 5’ region of cytochrome c oxidase subunit I (COI-5P) barcode would be of great value as it that can be used across the animal kingdom [4]. The BACTRIA pipeline addresses this need by producing a robust, high-quality phylogenetic tree for the COI-5P marker [5]. This tree serves as a foundational reference, enabling consistent and comparable PD metrics across diverse environmental samples.

This study presents a novel workflow that integrates a standardized phylogenetic reference tree (BACTRIA-generated) into QIIME2, enabling consistent calculation of PD metrics across diverse metabarcoding datasets. This proof-of-concept assesses the feasibility of phylogenetic placement for improving biodiversity assessments. QIIME2 is already a widely used platform for microbiome multi-omics bioinformatics and data science. The integration of reference trees into QIIME2 requires specific modifications to the existing tools and workflows, which are detailed in this paper. This study specifically aims to produce a proof-of-concept to see if placement of sample sequences on a reference tree will provide reliable PD metrics. It does this by using geographical samples with expected grouping of neighbouring countries in PD metrics.

2. Materials and Methods

In Figure 2 a visualisation can be seen of the workflow incorporated in the program created for this project. It is based on the QIIME2 workflow with the alignment and phylogeny steps replaced by using external tools.

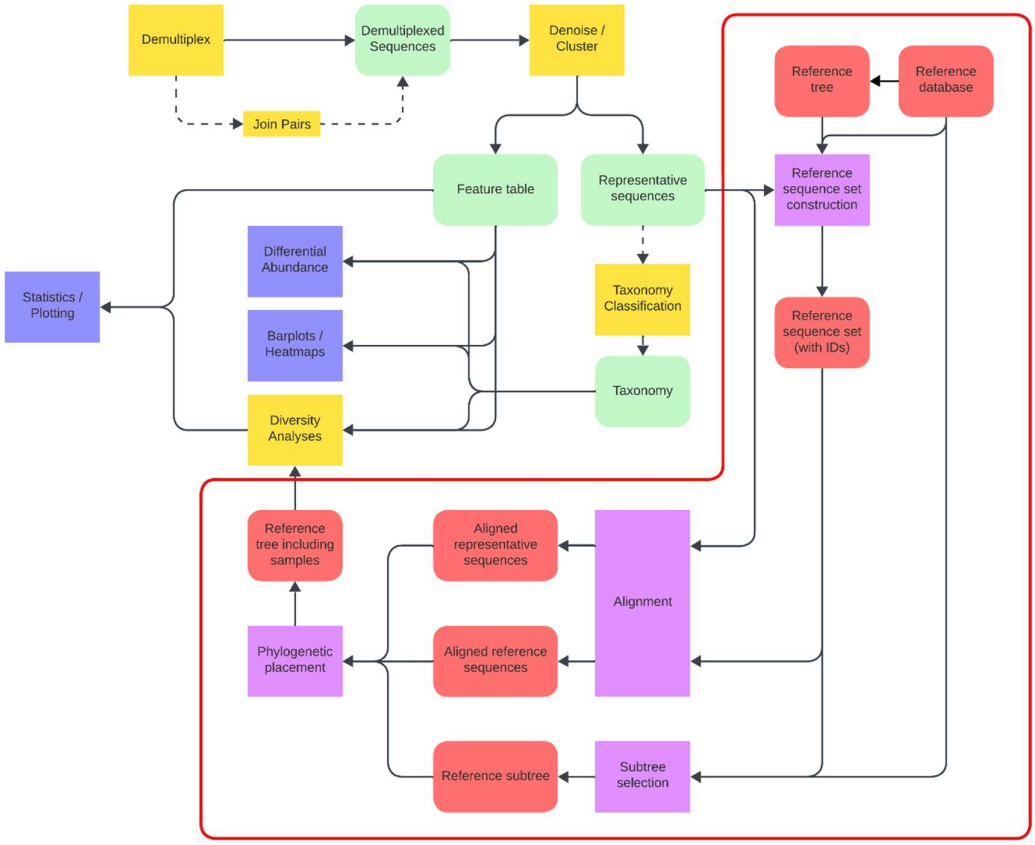

Figure 1: The workflow of QIIME2 supplemented with additional steps (in red) to visualise the workflow eventually used within this project (2025) [1]

2.1 QIIME2

QIIME2 was used for initial processing of raw sequencing data to representative sequences. After producing a reference tree with the representative sequences grafted on it was used again to produce the PD metrics.

Importing

The sequencing data used for this experiment was Illumina sequencing data. This data is demultiplexed and quality scored through the Illumina platform. This means the data is ready for denoising after importing into a QIIME2 artifact. To import the paired-end data a manifest file is constructed which specifies the location of the forward and the reverse reads so these will be paired in the QIIME2 artifact. This is the command use to import the sequencing data:

```
qiime tools import \
  --type SampleData[PairedEndSequencesWithQuality] \
  --input-path manifest.tsv \
  --output-path demux-paired-end.qza \
  --input-format PairedEndFastqManifestPhred33V2
```

Denoising

Denoising includes multiple steps, such as filtering out noisy and chimeric sequences, correcting erroneous sequences where possible, trimming, dereplication and truncating sequences to remove low-quality regions. In QIIME2 denoising can be executed using either DADA2 or Deblur. The main difference is that DADA2 creates an error model by processing each sample individually, while Deblur processes all samples at the same time using a predefined error model. Because of this, DADA2 does provide higher sensitivity and resolution, the downside is that this is at the cost of higher computing power and more erroneous amplicon sequence variants (ASVs)[6]. For this project, the script has been made so both deblur and dada2 can be used to denoise the samples. For the proof-of-concept dada2 was used through the following command:

```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux-paired-end.qza \
  --p-trunc-len-f 230 \
  --p-trunc-len-r 230 \
  --p-trim-left-f 13 \
  --p-trim-left-r 13 \
  --p-max-ee-f 2 \
  --p-max-ee-r 2 \
  --p-trunc-q 2 \
  --p-chimera-method consensus \
  --p-n-threads 4 \
  --o-table table.qza \
  --o-representative-sequences rep-seqs.qza \
  --o-denoising-stats denoising-stats.qza
```

QIIME2 is used once more at the end of the program to generate the PD metrics. For the PD metrics a metadata file will have to be created. This contains sample information and grouping from an experiment to which the PD metrics will be applied. Additionally, the sampling depth must be determined to perform rarefication. When the metadata file is created and the sampling depth is determined this command can be used to generate the PD metrics:

```
qiime diversity core-metrics-phylogenetic
  --i-phylogeny grafted_tree.qza
  --i-table feature-table.qza
  --p-sampling-depth 1000
  --m-metadata-file metadata.tsv
```

2.2 Reference sequence set construction

The goal for this project is to be able to use any reference tree and construct a reference database based on this tree and the database the tree was initially constructed with. For the proof-of-concept the reference tree used was constructed by BACTRIA using the barcode of life data (BOLD) database. As this program runs separately from BACTRIA the reference tree had to be used to obtain the process ids so an initial subset could be obtained from the BOLD database which corresponds with the tree and using the same process ids. This tree and database are still too big to be usable as a reference for the sample sequences. Using BLAST the ASVs generated through denoising are blasted against the BOLD database subset to obtain a smaller subset which consists of reference sequences with the closest match to the ASVs[7]. Below is the command used to execute the BLAST search.

```
# Create BLAST database from reference
makeblastdb -in reference_sequences.fasta -dbtype nucl -out
reference_db
```

```
# BLAST query sequences
blastn -query query_sequences.fasta -db reference_db \
  -out blast_results.txt \
  -outfmt "6 qseqid sseqid pident length mismatch gapopen
qstart qend sstart send evalue bitscore"
```

This command generates the 500 highest percent identity hits for every query sequence. For every query, the hits with an e-value of at least 1e-5 are selected to keep only similar species. Of that sub selection the five highest percent identity hits are chosen to limit the size of the reference database and reference tree. The hits from this are the reference database used to create the reference tree and execute pplacer.

2.3 Subtree selection

Using the reference database created by blasting we can prune the original reference tree into a more workable subtree. As we have been using the BOLD database and the original reference tree was created using this database we can keep using the BOLD process ids. Using the Bio-Phylo-Forest-DBTree toolkit we can use a list of process ids obtained after blasting to prune the tree. This is done by loading the reference Newick tree into an sqlite3 database to increase workability using megatree-loader. After this the text file containing the process ids will be used to prune the tree using megatree-pruner [8]. The commands used to do this are:

```
# Load tree into database
megatree-loader -i reference_tree.nwk -d reference_tree.db

# Prune tree based on IDs
megatree-pruner -d reference_tree.db -i reference_ids.txt >
pruned_tree.nwk
```

2.4 Alignment

Although multiple possibilities exist for alignment for this study HMMER is used. HMMER is used for creation of the BACTRIA tree and it is one of the recommended methods in pplacer documentation [9]. HMMER provides multiple output formats, Stockholm is chosen as this is the preferred format for pplacer. The scoring within the Stockholm allows for further refining of the results by choosing the sequence orientation with the highest value to ensure the correct alignment. Additionally, the scoring is used to filter out sequences with low alignment scores. Most of these steps happen without using HMMER for which the main command is:

```
hmmalign –trim -i reference_sequences.fasta -o
reference_alignment.sto -m COI-5P.hmm
```

After this has been done for the reference database, the alignment will be used to create a new HMM profile. This HMM profile is needed to align the queries to the reference which is needed for pplacer. The commands for profile creation are:

```
hmmbuild reference_alignment.hmm reference_sequences.sto
```

And the command for combining the reference alignment with the query sequence is:

```
hmmalign -o combined_alignment.sto --mapali refseqs.sto
refseqs.hmm query_sequences.fasta
```

2.5 Phylogenetic placement

For the main part of the project pplacer was used, pplacer needs three files as input, a combined alignment of the reference and query sequences, the reference tree, and a statistics file. The alignment and the reference tree have been made in the previous steps, but the statistics file will still need to be generated. This is done using RAxML as this is recommended by pplacer documentation using the following command [9]:

```
# Generate stats file
raxmlHPC-PTHREADS -T 4 -f e -t pruned_tree.nwk -s
combined_alignment.phy \
  -m GTRGAMMA -n stats -p 12345
```

With all the prerequisites present for pplacer, the generating of placement data is done with this command:

```
# Run pplacer
pplacer -t pruned_tree.nwk -s RAxML_info.stats -o
placements.jplace \
  -j 4 combined_alignment.sto
```

The placement data contains the reference tree with numbered leaves and per query the highest scoring leaves for placement. Within pplacer the guppy package can then be used to create the reference tree with the query sequences grafted onto it.

```
# Create grafted tree from placementdata
guppy tog -o grafted_tree.xml placements.jplace
```

3. Results

Presented here are results of a proof-of-concept to assess the reliability of placement of sample sequences on a reference tree. This is done in two runs with samples from European countries and samples from South American countries. For both cases, the ten biggest countries with at least 100 samples were chosen from the BOLD database. 100 samples per country were combined into a single FASTA file and registered in a feature table so it could eventually be used in QIIME2. Additionally, these samples were used to create a tree using MAFFT and FastTree, this tree is created de novo. These trees are used as a control.

In the case of South American countries, the ten biggest countries were Germany, Spain, France, Italy, Poland, Ukraine, The United Kingdom, Sweden, Norway, and Finland. The main PD metric used to visualise the analyse the results is Unweighted UniFrac as the composition including rare taxa will be better detected in comparison to weighted UniFrac. In Figure 2D, a dendrogram can be seen that was made using the unweighted UniFrac distance matrix to better visualize any grouping. In both the figure and the distance matrix it can be seen that some logical clustering exists, Poland and Ukraine, Italy, France and Spain, and Sweden, Finland and Norway. In Figure 3, the dendrogram for South America can be seen, some clusters like Columbia, Ecuador and Peru are logical, just as Argentina and Chile or Brazil and Paraguay. The branch containing Bolivia and Venezuela are farther apart and less expected to be in a similar cluster.

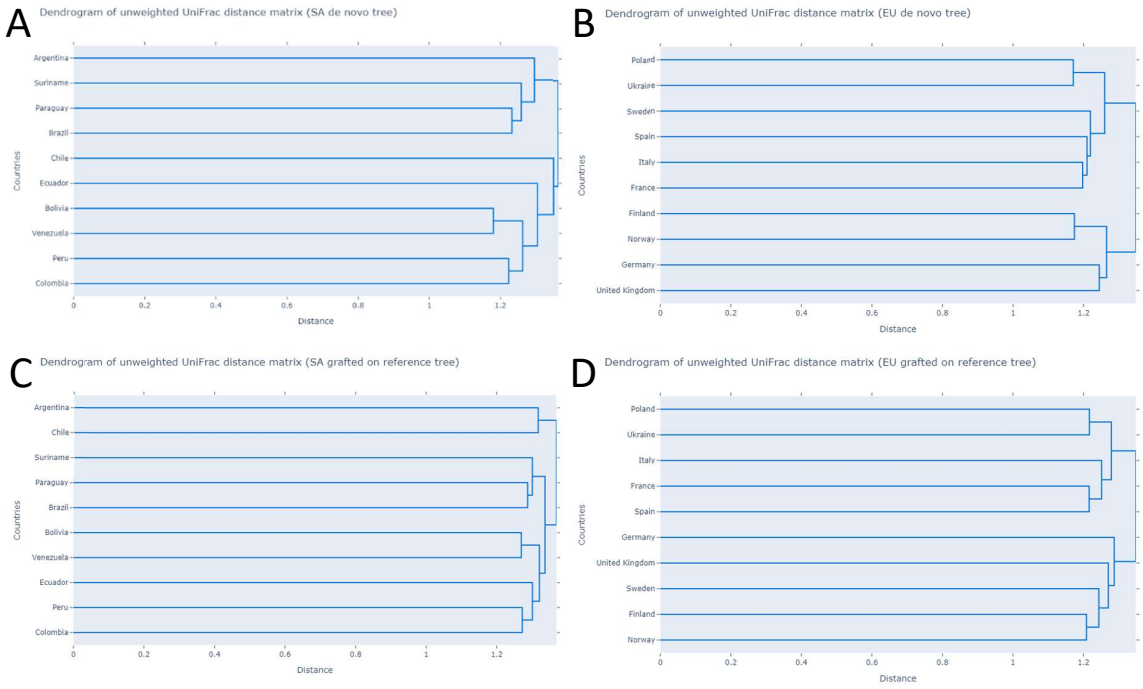

Figure 2: Figure 2: Dendrograms based on unweighted UniFrac distance matrices generated using QIIME2. A: Phylogenetic clustering of South American countries based on a tree built from query sequences alone, showing expected and unexpected groupings. B: Phylogenetic clustering of European countries using a tree built from query sequences alone, illustrating limitations in direct sequence-based tree construction. C: South American country clustering using a reference tree with grafted query sequences, demonstrating improved geographical coherence. D: European country clustering using a reference tree with grafted query sequences, highlighting better separation of expected biogeographical patterns.

4. Discussion

This study presents a proof-of-concept for integrating phylogenetic placement into metabarcoding workflows using a standardized reference tree. By leveraging external tools such as HMMER, pplacer and RAxML in combination with QIIME2, the approach successfully generates phylogenetic diversity (PD) metrics. The results demonstrate that placing sample sequences onto a common phylogenetic framework can yield meaningful biodiversity insights, as evidenced by the clustering of geographical samples in expected biogeographical patterns. Showing that this could become a valuable tool in biodiversity research in standardising the PD metrics calculation.

For the current proof-of-concept pplacer is used. There were two other possibilities, EPA and EPA-NG, both developed by the RAxML team. Literature suggests that pplacer performs comparably to EPA under most conditions, while EPA-NG has been reported to outperform both EPA and pplacer in cases with larger number of query sequences [9, 10]. However, our use case differs in that the computational bottleneck is the size of the reference tree rather than the number of query sequences. Previous studies from our department have also found that EPA-NG did not consistently outperform pplacer in our specific context. Additionally, none of these tools have seen significant updates in recent years, which may impact their long-term viability for this type of analysis [11, 12].

While this study did not perform an exhaustive computational benchmarking, the results from the proof-of-concept shows that pplacer did work with the current setup. EPA-NG, which is designed for high-throughput placements, might be more beneficial in cases where thousands of query sequences need to be processed simultaneously. Future work should assess runtime and memory usage more systematically, particularly for scaling the workflow to larger datasets.

While challenges remain, particularly regarding computational efficiency, this proof-of-concept lays the groundwork for further refinement and expansion of phylogenetic placement in metabarcoding workflows. With continued optimization and broader application, this approach has the potential to significantly enhance biodiversity monitoring and ecological research.

Literature

[1] E. Bolyen *et al.*, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nature Biotechnology,* vol. 37, no. 8, pp. 852-857, 2019/08/01 2019, doi: 10.1038/s41587-019-0209-9.

[2] P. de Witt and C. A. de Witt, "RESEARCH ARTICLE: How Long Does It Take to Prepare an Environmental Impact Statement?," (in en), *Environmental Practice,* vol. 10, no. 4, pp. 164-174 %U https://www.cambridge.org/core/journals/environmental-practice/article/abs/research-article-how-long-does-it-take-to-prepare-an-environmental-impact-statement/C1B14ECB03EBB159A2CE6B3A43CB5FAB, 2008.

[3] Z. G. Compson, B. McClenaghan, G. A. C. Singer, N. A. Fahner, and M. Hajibabaei, "Metabarcoding From Microbes to Mammals: Comprehensive Bioassessment on a Global Scale," (in English), *Frontiers in Ecology and Evolution,* vol. 8 %U https://www.frontiersin.org/articles/10.3389/fevo.2020.581835, 2020.

[4] P. D. Hebert, S. Ratnasingham, and J. R. de Waard, "Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species," (in eng), *Proc Biol Sci,* vol. 270 Suppl 1, no. Suppl 1, pp. S96-9, Aug 7 2003, doi: 10.1098/rsbl.2003.0025.

[5] *Bactria: BarCode TRee Inference and Analysis*. (2023). Github. [Online]. Available: https://github.com/naturalis/barcode-constrained-phylogeny

[6] J. T. Nearing, G. M. Douglas, A. M. Comeau, and M. G. I. Langille, "Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches," (in en), *PeerJ,* vol. 6, p. e5364 %U https://peerj.com/articles/5364, 2018.

[7] C. Camacho *et al.*, "BLAST+: architecture and applications," *BMC Bioinformatics,* vol. 10, no. 1, pp. 421 %U https://doi.org/10.1186/1471-2105-10-421, 2009.

[8] "DBTree: Very large phylogenies in portable databases - Vos - 2020 - Methods in Ecology and Evolution - Wiley Online Library %U https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13337," ed.

[9] 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,* vol. 11, no. 1, pp. 538 %U https://doi.org/10.1186/1471-2105-11-538, 2010.

[10] P. Barbera *et al.*, "EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences," *Systematic Biology,* vol. 68, no. 2, pp. 365-369, 2019, doi: 10.1093/sysbio/syy054.

[11] *pierrebarbera/epa-ng %* AGPL-3.0 %U https://github.com/pierrebarbera/epa-ng*. (2025).

[12] "Releases · matsen/pplacer %U https://github.com/matsen/pplacer/releases," in *GitHub*, ed.