BACTRIA: Barcode Tree Inference and Analysis

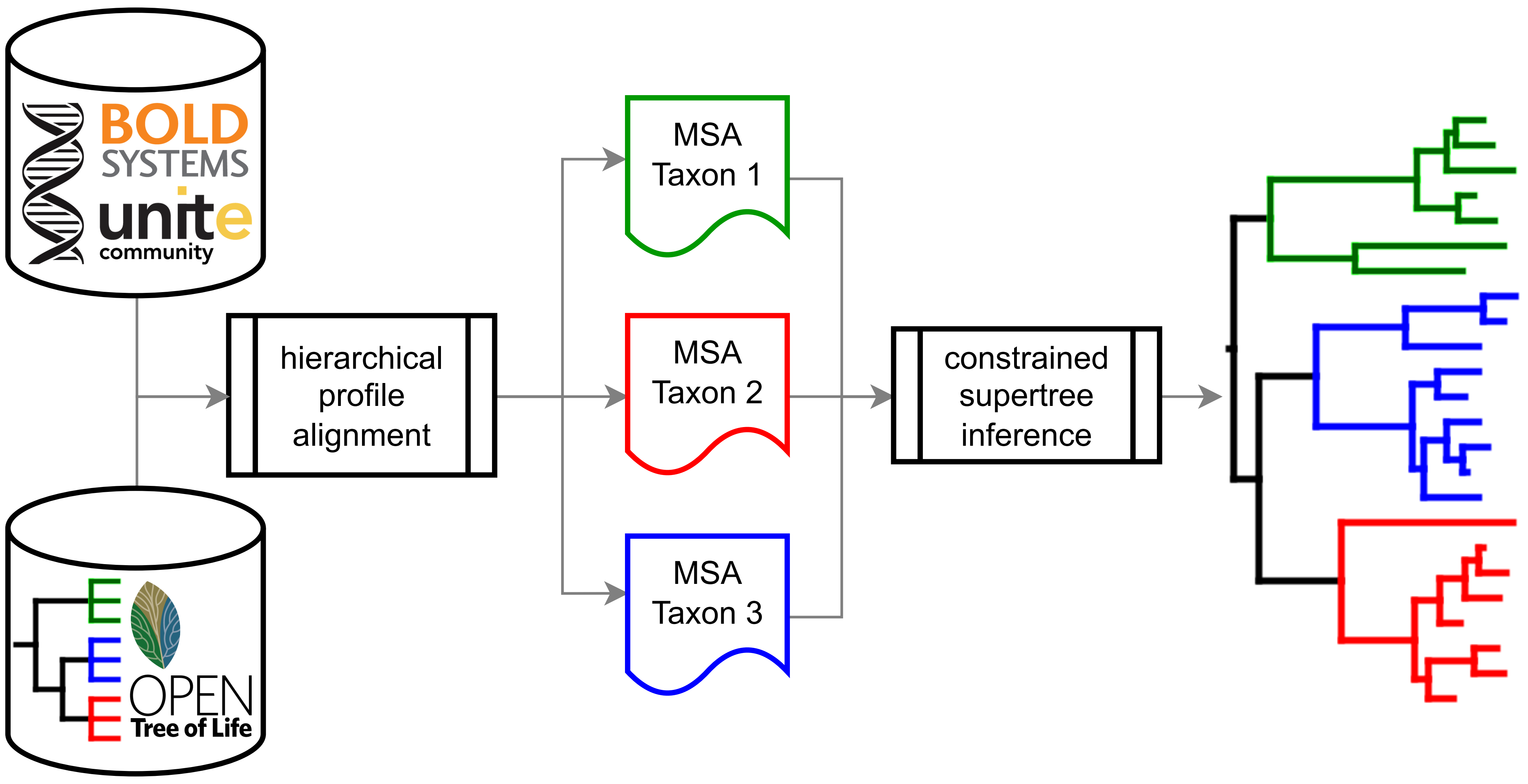
Reference phylogeny construction and phylogenetic placement for PD calculations from metabarcoding

Rutger Vos[1], Casper Carton[2], Naomi van Es[3], Lena ten Haaf[2], Luuk Romeijn[2], Noah Scheffer[4], Fons Verbeek[2]

The genetic diversity of environmental samples can be quantified using *DNA metabarcoding*, i.e. high-throughput sequencing of amplified barcode genes. Typical workflows to analyze the data result in species lists obtained by matching generated reads against reference libraries. Such results ignore the *phylogenetic diversity* (PD) context of the data. We develop workflows for 1) reference phylogeny construction as keystones to enable PD calculations, 2) perform scalable, multi-step phylogenetic placement

What is phylogenetic diversity (PD)?

PD encompasses metrics for species-specific **alpha diversity** (e.g., EDGE index), whole-tree metrics (e.g., sums of branch lengths), and **beta diversity** (e.g., UniFrac). These provide insights into **species richness, evolutionary relationships, and inter-community dissimilarity.** Alpha diversity metrics highlight PD within a sample. Beta diversity metrics quantify phylogenetic composition differences between samples.



hierarchical profile alignment

MSA Taxon 1

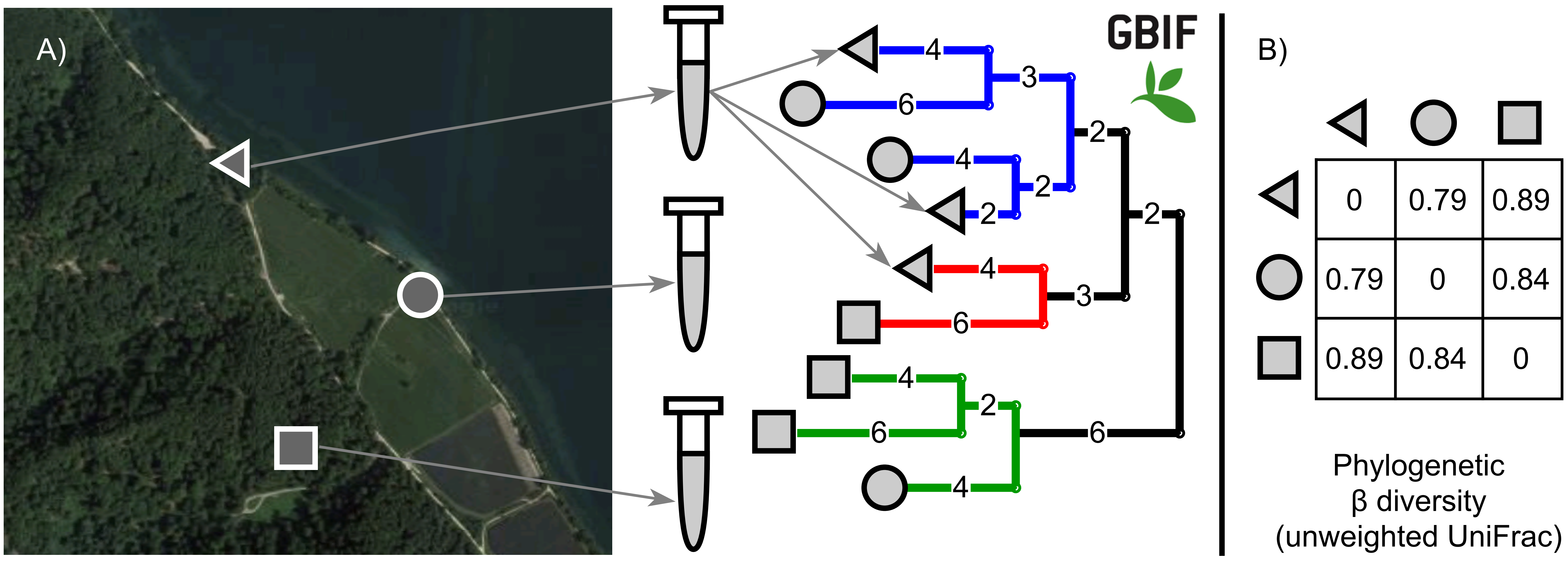
MSA Taxon 2

MSA Taxon 3

constrained supertree inference

Phylogeny construction

1. Using a backbone phylogeny (OpenTOL) or taxonomy, partition barcodes in chunks
2. Align chunks using protein-guide (e.g. COI) or raw nucleotides otherwise
3. Infer chunk subtrees with topologically constrained searches with RAxML-NG
4. Select 2 distal exemplars from each subtree, aggregate and build backbone tree
5. Graft subtrees on backbone



	◁	○	□
◁	0	0.79	0.89
○	0.79	0	0.84
□	0.89	0.84	0

Phylogenetic β diversity (unweighted UniFrac)

What is DNA metabarcoding?

Metabarcoding is a molecular technique used to **assess organism diversity** in environmental samples. By sequencing specific genetic markers, researchers can **identify multiple species** simultaneously. Metabarcoding provides insights into **community composition, species interactions, and ecosystem dynamics,** transforming biodiversity research.

What is phylogenetic placement?

Phylogenetic placement is a method used to assign environmental DNA sequences to their most likely position within a phylogenetic tree (panel A). By combining it with PD calculations, it **provides quantitative insights into the diversity and distribution of organisms** within and across environmental samples (panel B).

[1] Naturalis Biodiversity Center

[2] liacs Leiden Institute of Advanced Computer Science

[3] hogeschool Leiden

[4] HAN_UNIVERSITY OF APPLIED SCIENCES