dartR support for Phylogenetic Analysis

The objective of phylogenetic analysis is to extract relationships between taxonomic entities, be they species, evolutionarily significant units (ESUs) or other diagnosable units (Felsenstein, 2004; Swofford & Berlocher, 1987). The goal is thus to extract the pattern of ancestry and descent among such taxonomic entities (call them operational taxonomic units or OTUs). The OTUs need to be diagnosable because one assumes that differences among them reflect divergence through time, unobscured by contemporary or recent tokogenic exchange. That is, they are considered to be on evolutionary trajectories that are independent by virtue of reproductive or long-standing geographic isolation. If the analysis is applied to individuals or populations subject to tokogenic exchange (horizontal transfer), the resultant tree is a summary of genetic similarity, not a phylogeny. or a loose combination of the two.

The true evolutionary history of the OTUs is in the form of a bifurcating tree, that is, the true divergences among OTUs satisfy both the conditions of a metric and the four-point condition (Buneman, 1973). Metric is used here in the sense that, given the positions of two OTUs to represent the distance between them, the distances to a third OTU uniquely defines its position. The four- point condition is used here in the sense that for any four OTUs, there exists a simple tree accurately depicting the distances between them (that is, there exists a non-negative internal branch). A set of OTUs and pairwise distances among them that satisfy the metric and four-point conditions will define a unique bifurcating tree.

Unfortunately, in nature, phylogenies do not follow a process that maintains tree distances or strict tree-like structure in the data. Homoplasy and differential histories among the genes (independent lineage sorting) means that the process is not one of mining the true tree, but one of estimating the “best tree” given data that contains conflicting signals on the true trajectory of ancestry and descent. The best tree might be the most parsimonious tree, or the maximum likelihood tree, or the best bet optimised using all available information. There are various approaches to gaining consensus across multiple gene trees.

This is a controversial area, and so we have simply provided streamlined pathways to some more common analyses, and leave it to you to decide which you think is most appropriate to your case. The options are

gl2phylip.r — produces an input file for Phylip (Felsenstein, 1989), including an option for bootstrapping. This is a distance approach to recovering phylogenies.

```
gl2phylip(gl,outfile="infile",outpath=getwd(),
bstrap=1000)
```

gl2svdquartets.r — produces an input nexus file for PAUP* (Swofford, 2003) to undertake a phylogeny using SVD quartets. This approach is one way to deal with the challenges of independent lineage sorting. Method 1 has one line per population (OTU) and uses ambiguity codes (Swofford pers. comm.); Method 2 has two lines per population (OTU) to allow representation of both SNP states (Chifman & Kubatko, 2014).

```
gl2svdquartets(gl,outfile="svd.nex",outpath=getwd(),
method=1)
```

gl2treemix.r — produces an input file for treemix (Pickrell & Pritchard, 2012), which looks for stress between the data matrix and the ML tree, and reduces that stress by hypothesising migration events (gene flow).

```
gl2treemix(gl,
outfile="treemix.pnet",outpath=getwd())
```

gl2phylonet.r — produces an input file for Phylonet (Zhu, Wen, Yu, Meudt, & Nakhleh, 2018), which has options for analysing data not as a bifurcating tree, but as a network. It thus accommodates tokogeneic exchange. This script has been tested only so far as confirming the format of the input file for Phylonet. Let us know how you go.

```
gl2phylonet(gl,outfile="phynet.nex",outpath=getwd())
```

Note that the output file needs to be gzipped before it is passed to Phylonet.

gl2fasta.r — produces a fastA file in a number of formats suitable for phylogenetic analysis. Methods 1 and 2 concatenate the sequence tags for analysis by Garli which uses a mutational model that has been generated from the base frequencies and ts/tv ratios. Method 1 replaces heterozygous sites with the relevant ambiguity code; method 2 selects one SNP state at random. Methods 3 and 4 combines only the SNP base information, which is suitable for use by programs like RAXML. Again, Method 1 replaces heterozygous sites with the relevant ambiguity code; method 2 selects one SNP state at random.

```
gl2fasta(gl, method=1,outfile="output.fasta",
outpath = tempdir())
```

References

Buneman, P. (1973). A note on the metric properies of trees. *Journal of Combinatorial Theory*, *17B*, 48–50.

Chifman, J., & Kubatko, L. (2014). Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, *30*, 3317–3324.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, *5*, 164–166.

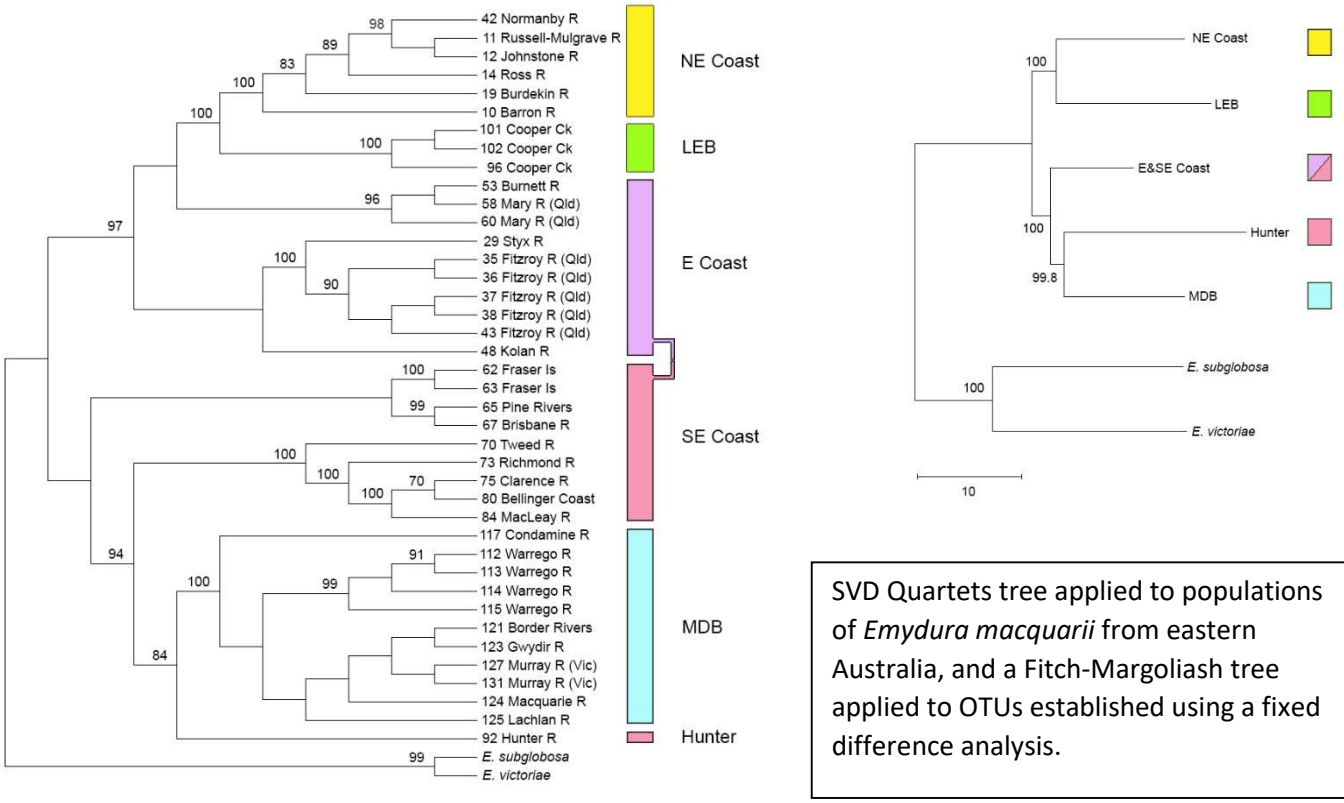
Felsenstein, J. (2004). *Inferring Phylogenies*. Sunderdland, MA USA: Sinauer Associates.

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genetics*, *8*, e1002967.

Swofford, D. L. (2003). *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sunderland, Massachusetts, USA: Sinauer Associates.

Swofford, D. L., & Berlocher, S. H. (1987). Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Systematic Zoology*, *36*, 293–325.

Zhu, J., Wen, D., Yu, Y., Meudt, H. M., & Nakhleh, L. (2018). Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLOS Computational Biology*, *14*, e1005932.



SVD Quartets tree applied to populations of *Emydura macquarii* from eastern Australia, and a Fitch-Margoliash tree applied to OTUs established using a fixed difference analysis.