Recommendations for data analysis for the Genomics for Australian Plants (GAP) Phylogenomics projects

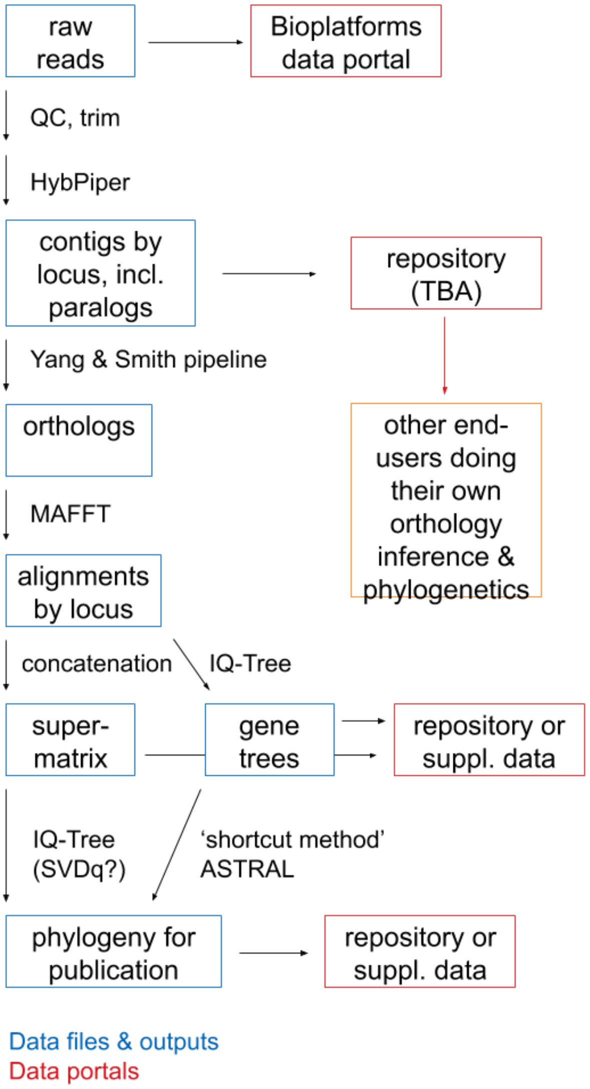
Last updated: 28 February 2022

GAP Phylogenomics Bioinformatics Working Group: Matthew Barrett, Edward Biffin, Jason Bragg, Richard Edwards, Johan Gustafsson, Chris Jackson, Todd McLay, Lars Nauheimer, Alexander Schmidt-Lebuhn, Anna Syme, Trevor Wilson

Purpose

The purpose of this document is to summarise the recommendations of the GAP Phylogenomics Bioinformatics Working Group regarding:

- What approaches and software to use to prepare and analyse sequence capture (hybrid enrichment) data for consortium-scale phylogenomics work;
- What suggestions, information and software training might be useful to individual GAP members who want to analyse subsets of the data;
- What data files to make available to the community after the end of the embargo period or to accompany publications;
- What questions still need to be clarified, and what additional genetic or technical resources would be helpful to achieve the aims of GAP phylogenomics work.



Preparation of reads

We recommend the use of **FastQC** (Andrews 2014) for quality control of raw reads. The software is widely available and includes a user-friendly graphical user interface (GUI) with a large number of potential warning flags and suggestions for improvement (e.g. sequence duplication, read quality, need to trim).

We recommend the use of **Trimmomatic** (Bolger *et al.* 2014) for cleaning, trimming, and pairing of raw reads.

Assembly

We recommend the use of **HybPiper** (Johnson *et al.* 2016) for the assembly of cleaned reads to the target loci. The tool consists of a set of Python scripts available at https://github.com/mossmatters/HybPiper and requires the unassembled reads and target sequences (usually the sequences used to design the bait set) as inputs.

HybPiper produces large numbers of temporary files for each sample, potentially resulting in the user reaching the maximum number of files allowed on their supercomputer. It comes with a clean-up script that removes all unnecessary files, but in a high-throughput approach the user needs to script their loop so that this happens after each sample has been assembled. Instructions or practical assistance should therefore be provided to GAP members intending to run HybPiper themselves.

Identification of paralogs

Genome and gene duplication events lead to the existence of very similar gene families in the same organism. If a clade is derived from a polyploidisation event in a common ancestor, or if a single gene was duplicated in the ancestor, each descendant species may carry two copies of what was originally one gene, with both paralogs diversifying as independent gene trees or being lost differentially. A frequent problem in phylogenomics is to ensure that the sequences in any alignment used for analysis are orthologs of each other, i.e. members of the same gene tree since the duplication event.

HybPiper has a simplistic way of deciding which of several potential paralogs to retain for the final output, using first read coverage and then similarity to the target sequence as criteria. This means that it is easy for two different paralogs to be accepted in two samples, especially if one of the two paralogs was not assembled in one of the samples, or if orthologs had very different read coverage across samples.

We therefore recommend that HybPiper's **paralog investigator** script be run after assembly. It will produce a *.fasta file for each gene locus containing potentially multiple contigs for a given sample, allowing paralogy to be resolved with a dedicated tool. HybPiper also provides paralogy warnings, which together with the sequence data allow the user to decide whether gene alignments can be analysed directly or whether ortholog inference will be required

For datasets where paralogy is an issue, we recommend the use of the ortholog inference pipeline developed by **Yang & Smith** (2014). It is a set of Python scripts available at https://bitbucket.org/yangya/phylogenomic_dataset_construction/src/master/ and uses gene tree topology to infer paralogy. It has the advantages of not requiring a reference genome, being automated, and providing four different approaches of varying strictness in one package (1to1, MO, RT, and MI, from retrieving only those loci that have no paralogy at all to retrieving every part of the gene tree that shows a duplicated taxon as a separate ortholog). The outputs are *.fasta files with sequence alignments for each ortholog recognised by the specific approaches used (of four).

A limitation of the pipeline at the level of all angiosperms is that it uses outgroup information for its two intermediate-stringency approaches (MO, RT). Fortunately, sequences of *Austrobaileya*, which is a candidate for the sister group of all other extant angiosperms, are available for c. 95% of the genes in the 353 bait kit. It is likely that GAP or PAFTOL sequence data from other, related plant families can be used as outgroups for many, if not all, analyses at the family level.

A technical limitation of the pipeline, as originally published, is that it expects a different input than the output produced by HybPiper. The working group has developed and continues to improve scripts to connect the two into one workflow.

Phylogenetic analysis

Data matrices can be partitioned to allow different models of evolution to apply to different parts of the data. In datasets concatenated from different genes it is, for example, common to partition by gene, as they may evolve at different speeds. In the case of GAP Phylogenomics data, the targeted sequences are all of protein coding genes, in which every third nucleotide position (the third codon position) evolves considerably faster than the other two, because many mutations in that position do not affect the protein sequence. We therefore suggest that consortium members explore partitioning by codon position across their data.

Three main approaches are commonly used to infer phylogenies from multi-locus sequence data:

A full Bayesian multi-locus coalescent analysis as implemented in the BEAST add-on StarBEAST (Heled and Drummond 2010; Ogilvie *et al.* 2017) is widely considered the "gold standard", but it is only feasible for datasets of limited size with very little missing data. It would be desirable if a training workshop, or instructions, could be provided to GAP members who want to use BEAST to analyse subsets of the data. The consortium will not, however, be able to use this approach for an angiosperm phylogeny of hundreds of genera, as many are likely missing data for several of the gene loci.

The second approach is concatenation of the sequence data from all orthologs into a supermatrix. It allows the use of fast, robust and widely used likelihood phylogenetics, but discordance between gene trees can mislead estimation of tree topology and branch lengths.

The third is the collection of algorithms known as 'short-cut methods'. Their input is gene trees, usually themselves inferred with likelihood phylogenetics, which are then used to infer the species tree that best explains the patterns of gene tree discordance observed in the data. They can be misled by uncertain gene tree topologies, and only a few outlier genes can sometimes have a large effect on the results.

It is assumed that gene tree incongruence is a greater problem at shallow phylogenetic scales, suggesting that short-cut methods are more suitable when studying young groups of closely related species, and that concatenation is more suitable for deep relationships (Bryant and Hahn 2020). For large datasets in particular we recommend the exploration of **both concatenation and short-cut methods**.

We recommend the use of **IQ-TREE** (http://www.iqtree.org/) (Nguyen *et al.* 2015) for likelihood phylogenetics of concatenated data and to produce individual gene trees. The software is extremely fast and user-friendly due to the implementation of many sensible default settings. It allows the combination of model testing, partition finding, and phylogenetic analysis in one command, and provides a variety of branch support measures (traditional bootstrap, ultrafast bootstrap, concordance factors, likelihood ratio tests, etc.).

We recommend the use of **ASTRAL** (https://github.com/smirarab/ASTRAL) (Mirarab and Warnow 2015) for short-cut phylogenetics. We recommend that gene trees be tested for formation of separate clusters with conflicting topologies, for example using the approach of Foster *et al.* (2018). Consortium members may also be interested in exploring ASTRAL-PRO (https://github.com/chaoszhang/A-pro), which uses paralog sequences directly, bypassing bioinformatic ortholog inference (Zhang et al. 2020).

Making data available to the community

We expect that data should be made publicly available at three steps of the analysis pipeline.

Raw reads will be made available through the Bioplatforms Data Portal. However, it would be inconvenient if researchers who want to build on GAP resources only had raw data at their disposal, because they would have to repeat the entire bioinformatics workflow, and raw read files are extremely large compared to processed target enrichment data.

Two intermediate data outputs are suggested for upload: contigs for each gene and specimen, either after assembly or after ortholog inference. We recommend that the former, i.e. **contigs including all paralogs**, be made available. When future researchers attempt to integrate their own enrichment/capture data with existing GAP data they will need to assign their contigs to the correct orthologs, and the most reliable way of doing so is by repeating ortholog inference with all available information.

Finally, we expect that **data matrices used for phylogenetic analysis, gene trees and species trees** be made available either in suitable repositories (e.g. TreeBASE, Dryad, CSIRO Data Access Portal) or as supplementary data on a journal website.

References

Andrews S (2014) ‘Fast QC, a quality control tool for high throughput sequence data.’ http://www.bioin-formatics.babraham.ac.uk/projects/fastqc/.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bryant D, Hahn MW (2020) The concatenation question. In: Scornavacca C, Delsuc F, Galtier N (eds), *Phylogenetics in the Genomic Era*, pp 3.4:1-3.4:23. https://hal.archives-ouvertes.fr/hal-02535651/document

Foster CSP, Henwood MJ, Ho SYW (2018) Plastome sequences and exploration of tree-space help to resolve the phylogeny of riceflowers (Thymelaeaceae: *Pimelea*). *Molecular Phylogenetics and Evolution* **127**, 156–167. doi:10.1016/j.ympev.2018.05.018.

Heled J, Drummond AJ (2010) Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution* **27**, 570–580. doi:10.1093/molbev/msp274.

Johnson MG, Gardner EM, Liu Y, Medina R, Goffinet B, Shaw AJ, Zerega NJC, Wickett NJ (2016) HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* **4**, 1600016. doi:10.3732/apps.1600016.

Mirarab S, Warnow T (2015) ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52. doi:10.1093/bioinformatics/btv234.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268–274. doi:10.1093/molbev/msu300.

Ogilvie HA, Bouckaert RR, Drummond AJ (2017) StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution* **34**, 2101–2114. doi:10.1093/molbev/msx126

Yang Y, Smith SA (2014) Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Molecular Biology and Evolution* **31**, 3081–3092. doi:10.1093/molbev/msu245.

Zhang C, Scornavacca C, Molloy EK, Mirarab S (2020) ASTRAL-Pro: Quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution* **37**, 3292–3307.