

1 hybpiper-rbgv and yang-and-smith-rbgv: Containerization and additional options 2 for assembly and paralog detection in target enrichment data

3 Chris Jackson¹, Todd McLay^{1,2,3}, and Alexander N. Schmidt-Lebuhn^{2,4}

4 ¹ Royal Botanic Gardens Victoria, Birdwood Avenue, Melbourne Victoria 3004, Australia

5 ² Centre for Australian National Biodiversity Research, CSIRO, Clunies Ross Street, Canberra ACT 2601,
6 Australia

7 ³ School of Biosciences, The University of Melbourne, Parkville, Melbourne, 3010

8 ⁴ Author for correspondence: alexander.s-l@csiro.au

9 **ABSTRACT**

10 **PREMISE:** The HybPiper pipeline has become one of the most widely used tools for the assembly of
11 target enrichment (sequence capture) data for phylogenomic analysis. Between the production of locus
12 sequences and phylogenetic analysis, the identification of paralogs is a critical step ensuring accurate
13 inference of evolutionary relationships. Algorithmic approaches using gene tree topologies for the
14 inference of ortholog groups are computationally efficient and broadly applicable to non-model
15 organisms, especially in the absence of a known species tree. Unfortunately, software compatibility
16 issues, unfamiliarity with relevant programming languages, and the complexity involved in running
17 numerous subsequent analysis steps continue to limit the broad uptake of these approaches and constrain
18 their application in practice.

19 **METHODS AND RESULTS:** We updated the scripts constituting HybPiper and a pipeline for the
20 inference of ortholog groups (“Yang and Smith”) to provide novel options for the treatment of
21 supercontigs, remove bugs, and seamlessly use the outputs of the former as inputs for the latter. The
22 pipelines were containerised using Singularity and implemented via two Nextflow pipelines for easier
23 deployment and to vastly reduce the number of commands required for their use. We tested the pipelines
24 with several datasets, one of which is presented for demonstration.

25 **CONCLUSIONS:** hybpiper-rbgv and yang-and-smith-rbgv provide easy installation, user-friendly
26 experience, and robust results to the phylogenetic community. They are presently used as the analysis

27 pipeline of the Australian Angiosperm Tree of Life project. The pipelines are available at

28 <https://github.com/chrisjackson-pellicle>.

29 **KEY WORDS** containerised; HybPiper; orthologs; Nextflow; paralogs; polyploidy; phylogenomics;

30 sequence capture; Singularity; target enrichmenty.

31 Target enrichment (or sequence capture) is a widely used method for generating high-throughput, multi-
32 locus sequence data for phylogenomic analysis, and it is of greater utility at deeper phylogenetic levels
33 than most other marker systems (McCormack et al., 2013). The approach fragments genomic DNA and
34 then enriches the desired target loci, usually hundreds of genome/gene regions, with RNA baits while
35 removing fragments representing the non-target regions. Bait design consequently requires knowledge of
36 the sequence of the target regions in at least some species of a study group, or closely related species.
37 In recent years an increasing number of bait sets has been designed to enrich either protein coding genes
38 or highly conserved sites flanked by more variable regions (Bejerano et al., 2004; Lemmon et al., 2012)
39 for a variety of major taxonomic groups. In plants, bait sets have been published for flagellate plants
40 (GOFLAG) (Breinholt et al., 2020), flowering plants (PAFTOL / Angiosperms353) (Johnson et al.,
41 2019), Asteraceae (Mandel et al., 2014), mosses (Liu et al., 2019), and ferns (Wolf et al., 2018), among
42 other groups.

43 Since its publication, the bioinformatics software HybPiper (Johnson et al., 2016) has become one of the
44 most widely used tools for the assembly of target enrichment data (102 citations Web of Science, 166
45 Google Scholar, accessed 6 June 2021), partly because of its flexibility. It provides options for the
46 assembly of exon or intron sequences, to retrieve a single sequence per sample based on read coverage
47 and contig length, or to collect all potential paralogs for subsequent analysis with other tools using
48 different criteria. A recent adaption of HybPiper developed for the Plant And Fungal Tree Of Life project
49 (Baker et al., 2021), paftools (<https://github.com/RBGKew/KewTreeOfLife>), does not provide the latter
50 functionality.

51 The correct inference of ortholog groups is critical in groups showing frequent gene or genome
52 duplication such as many families of land plants, where polyploidy is prevalent. Phylogenetic analysis of
53 paralogous gene copies can produce incorrect topologies, as the evolutionary history of gene families
54 interferes with the evolutionary history of species lineages (Maddison, 1997). Some methods for the
55 inference of ortholog groups require the use of reference genomes (Dessimoz et al., 2012), which remain
56 unavailable in many groups of organisms. Others rely on *a priori* knowledge of ‘undisputed species trees’
57 (Altenhoff et al., 2016), which creates a conundrum for phylogeneticists, to whom the inference of the

58 species tree is the purpose of the entire exercise. Algorithmic approaches using gene tree topologies to
59 infer ortholog groups, on the other hand, are computationally efficient and have the advantage of broad
60 applicability even in the absence of this kind of data.

61 A collection of Python scripts published by Yang and Smith (2014) (subsequently Y&S) and recently
62 adapted by Morales-Briones et al. (2020) provides four such algorithms and has become a widely used
63 tool for ortholog inference (107 citations Web of Science, 165 Google Scholar, accessed 6 June 2021).

64 Unfortunately, as originally published, it could not be used on the outputs of HybPiper without
65 reformatting of sequence names and changes to several scripts.

66 At a practical level, both HybPiper and the Y&S pipeline require the installation of a variety of
67 dependencies on the users' local system, and the user may be faced with software compatibility issues,
68 creating challenges for the wider adoption of these methods. Moreover, running HybPiper involves five to
69 eight individual terminal commands, and Y&S involves seven to ten (Table 1), depending on the desired
70 results and discounting additional scripts required to pipe HybPiper outputs into Y&S.

71 To address potential software installation and compatibility issues, we present a Singularity container
72 with all scripts and dependencies required by HybPiper and Y&S pre-installed in a portable software
73 'toolbox'. To simplify running HybPiper or Yang and Smith's (2014) scripts using this container, we
74 provide Nextflow scripts (hybpiper-rbgv and yang-and-smith-rbgv) that allow each improved pipeline to
75 be executed with a single command.

76 To run hybpiper-rbgv the only inputs required are a folder containing raw reads and a target file in fasta
77 format for the reads to be assembled against. It runs all steps comprising the original HybPiper pipeline,
78 including intronrate and paralog retrieval (<https://github.com/mossmatters/HybPiper/wiki/Introns>;
79 <https://github.com/mossmatters/HybPiper/wiki/Paralogs>). One of the outputs of HybPiper are sequence
80 files including all putative paralogs, and these are used as input to the yang-and-smith-rbgv script,
81 together with either a file of outgroup sequences or a list of designated outgroup samples that are already
82 in the HybPiper outputs. The latter outgroup information is required for two of the Y&S ortholog
83 inference algorithms. Additionally, bugs were fixed, and the modified HybPiper code produces more

84 accurate assemblies and flags final locus assemblies that may be built by concatenating SPAdes contigs
85 assembled from different paralogs.

86 **METHODS AND RESULTS**

87 **hybpiper-rbgv**

88 In the hybpiper-rbgv implementation (Fig. 1), several new features have been added to HybPiper as
89 follows. For each sample, multiple read files (e.g. from different Illumina sequencing machine lanes) can
90 be automatically combined prior to analyses. Input files can now be provided in compressed .gz format. If
91 read quality filtering has not yet been performed, hybpiper-rbgv can optionally run Trimmomatic before
92 assembly. If BLASTx is used for read mapping and the input target file provided contains nucleotide
93 sequences, it is automatically converted to amino acids before prior to BLASTX mapping. If desired, the
94 user can merge forwards and reverse reads prior to assembly using SPAdes.

95 By default, HybPiper attempts to unite several contigs that individually cover only part of a gene target
96 into a 'supercontig'. During development we observed that under some circumstances, this approach risks
97 the creation of chimeric supercontigs from different paralogs. Further, supercontig creation can lead to the
98 erroneous duplication of sequence areas at any sites of contig overlap. This latter issue has been fixed in
99 hybpiper-rbgv. To address the former issue, hybpiper-rbgv creates two output folders, one with all
100 supercontigs and one with suspected chimeras (assessed using read-mapping to supercontigs and
101 identification of discordant read-pairs) removed. Optionally, the creation of supercontigs can be
102 suppressed entirely.

103 In addition, minor bugs were fixed as documented in more detail on the project's Github site -
104 <https://github.com/chrisjackson-pellicle/HybPiper-RBGV>.

105 **yang-and-smith-rbgv**

106 Inference of ortholog groups with the Y&S scripts is based on examination of gene tree topologies. As a
107 first step, the yang-and-smith-rbgv pipeline (Fig. 2) aligns paralog sequences for each gene and infers
108 gene trees. Before the inference of ortholog groups, it conducts trimming of gene trees as implemented in
109 the original pipeline (Yang and Smith, 2014). First, the longer branch in very unbalanced sister terminals
110 is removed, under the assumption that it reveals an assembly or alignment error in the corresponding

111 sequence. Second, very closely related terminals (presumptive alleles) from the same sample are reduced
112 to one, as multiple closely related tips would interfere with the identification of paralogs. Third, very long
113 deep branches are pruned. Minimum parameters for pruning at all steps can be adjusted by the user.
114 The yang-and-smith-rbgv pipeline implements three of the four algorithms in the collection of Y&S
115 scripts. The Monophyletic Outgroups (MO) algorithm first removes all genes in which the outgroup is
116 non-monophyletic. In the remainder it then iteratively moves upwards from the root, checking at each
117 node if the two daughter clades share samples, and, if so, removes the smaller daughter clade, with the
118 rationale that these nodes represent the location of gene duplication events and that the more informative
119 ortholog group should be kept (Fig. 3a). This approach returns at most the same number of sequence
120 alignments as existed originally.

121 The other two algorithms make use of outgroups supplied as part of the paralog files or in a separate file.
122 Users who need to add outgroups to a dataset from custom baits for which little or no published data are
123 available can mine transcriptome data for sequences matching their HybPiper target file (McLay et al.,
124 2020).

125 The Rooted subTrees (RT) algorithm first dismantles a gene tree into ingroup clades if the outgroups are
126 non-monophyletic. In each ingroup clade it then iteratively moves upwards from the root, checking at
127 each node if the two daughter clades share samples. If that is the case, it separates the smaller daughter
128 clade out as a new ortholog group under the assumption that a gene duplication occurred at this node (Fig.
129 3b). Consequently, this approach has the potential to output considerably more sequence files than in the
130 original input, and some ortholog groups may contain very few samples.

131 The Maximum Inclusion (MI) algorithm iteratively extracts the largest subtrees from an unrooted gene
132 tree that do not contain duplicated samples (Fig. 3c). In contrast to MO and RT, this approach does not
133 rely on a logic that locates putative gene duplication events and may consequently be considered less
134 theoretically defensible than the alternatives.

135 The final algorithm of Yang and Smith (2014), 1to1, simply removes all genes containing paralogs,
136 retaining only the paralog-free genes. While this not explicitly implemented in yang-and-smith-rbgv, the

137 user can select all files labeled ‘1to1ortho’ from the results of the Maximum Inclusion algorithm to
138 achieve the same outcome.

139 The yang-and-smith-rbgv pipeline produces gene alignments for each inferred ortholog group under each
140 of the three algorithms. These alignments are ready for phylogenetic analysis either separately or after
141 concatenation. The pipeline uses MAFFT v. 7.471 (Kato and Standley, 2013) or MUSCLE (Edgar,
142 2004) for alignment steps and IQ-TREE v. 2.0.3 (Nguyen et al., 2015) for gene tree inference.

143 **Example dataset**

144 We tested the two pipelines on several datasets predominantly of Asteraceae and Orchidaceae. Most
145 analyses used the Angiosperms353 bait set (Johnson et al., 2016), and one used the compositae1061 bait
146 set (Mandel et al., 2014). A small dataset of twelve ingroup and two outgroup Asteraceae is here used as
147 an example. It is drawn from tribe Gnaphalieae: subtribe Gnaphaliinae: Australasian clade (Schmidt-
148 Lebuhn and Bovill, 2021). The data were produced by the Australian Angiosperm Tree of Life project as
149 part of the Genomics for Australian Plants consortium (<https://www.genomicsforaustralianplants.com/>).
150 Raw reads were quality filtered and trimmed using Trimmomatic 0.38 (Bolger et al., 2014). Only paired
151 reads were used for subsequent assembly with hybpiper-rbgv (though the input can include single orphan
152 reads from a Trimmomatic run, as well as a new option to include merged reads). The target file for
153 assembly was produced by filtering the angiosperm megatarget file of McLay et al. (2020) for Asteraceae.
154 Ortholog groups were inferred for resulting sequence files including paralogs (‘11_paralogs’ directory)
155 using all algorithms implemented in yang-and-smith-rbgv under default settings. For the MO and RT
156 algorithms, *Acomis macra* F.Muell. and *Helichrysum calvertianum* (F.Muell.) F.Muell. were set as
157 outgroups. They were selected because they belong to the Waitzia clade of Australasian Gnaphalieae
158 (Schmidt-Lebuhn and Bovill, 2021). In each case, we removed genes or ortholog groups with data for less
159 than five samples.

160 Sequence alignments for each ortholog group were processed to ensure that they were all in the correct
161 frame and concatenated using custom Python scripts. We compared dataset characteristics and
162 phylogenetic results for five different approaches: the results from each algorithm for inference of
163 ortholog groups (MO, RT, MI); only the paralog-free genes; and the direct HybPiper output, which

164 selects a paralog to maximise contig length and read coverage. In each case, we reconstructed a species
165 tree using ASTRAL 5.7.7 (Zhang et al., 2018) after inferring individual gene trees with IQ-TREE 1.6.12
166 (Nguyen et al., 2015) under the HKY+I+G model, also partitioning by codon position.

167 **Comparison of ortholog inference approaches**

168 After filtering for read quality, the 14 samples in the example dataset retained 1,007,159 to 40,976,703
169 reads (median 5,895,305). Of these, between 5.1% and 56.1% were on-target (median 28.2%). hybpiper-
170 rbgv retrieved sequences for between 296 and 348 genes (median 342) per species, of which between 166
171 and 283 (median 251) were at least 75% of the length of the mean length of all target sequences for a
172 given gene. In total, hybpiper-rbgv produced gene files for 350 of the 353 targeted genes. Between 9 and
173 29 genes (median 20) generated paralog warnings; HybPiper statistics are available at DOI:

174 [10.25919/q42q-j056](https://doi.org/10.25919/q42q-j056).

175 Dataset sizes are summarised in [Table 2](#). Using the outputs of hybpiper-rbgv directly resulted in 296-345
176 genes per species (median 340.5), as five genes were excluded for having less than five terminals.

177 The MO algorithm of yang-and-smith-rbgv removed 51 genes for having non-monophyletic outgroups,
178 removed paralogs from 22 genes, and inferred no paralogs in 277 genes, for a total of 299 remaining
179 ortholog groups.

180 The RT algorithm inferred the existence of 642 ortholog groups but only resulted in 224-253 ortholog
181 groups per species carried over into phylogenetic analysis (median 235), because 335 ortholog groups
182 were excluded for having data for less than five species.

183 The MI algorithm inferred no paralogs for 277 and separated 139 ortholog groups out of the remaining
184 73, for a total of 416 resulting ortholog groups. It resulted in 306-352 ortholog groups per species (median
185 348), with 36 ortholog groups excluded for having less than five terminals.

186 Using only paralog-free genes resulted in 229-273 genes per species (median 268), with 3 genes excluded
187 for having less than five terminals.

188 The ASTRAL phylogeny inferred for direct HybPiper outputs without inference of ortholog groups
189 differs from that inferred for all ortholog inference approaches in the relationships of *Chthonocephalus*
190 *muellerianus* P.S.Short, *Epitriche demissus* (A.Gray) P.S.Short, *Gnephis tenuissima* Cass., and

191 *Trichanthodium skirrophorum* Sond. & F.Muell. ex Sond., suggesting that the analysis is misled by the
192 presence of unrecognized paralogy (Fig. 4). In addition, the placement of *Millotia tenuifolia* Cass. varies
193 across analyses, with data derived from the MI and RT algorithms favoring one placement, and those
194 from MO and only paralog-free genes another.

195 CONCLUSIONS

196 hybpiper-rbgv and yang-and-smith-rbgv are pipelines for the assembly of target enrichment data and the
197 inference of ortholog groups that facilitate installation and simplify use compared to the standalone
198 HybPiper and Yang-and-Smith softwares. They required little to no expertise in scripting and provide
199 several new options, increasing flexibility with regard to input data e.g. by allowing the use of read files
200 from multiple lanes.

201 By improving the method of joining contigs from the same gene together, hybpiper-rbgv does not
202 produce duplicated sequence regions during the generation of supercontig-derived loci sequences.

203 Additionally, it implements options for the removal of potentially chimeric supercontigs or of all
204 supercontigs, giving the user additional assembly options. yang-and-smith-rbgv implements the same
205 algorithms for ortholog inference as its original version but can use the outputs of hybpiper-rbgv directly
206 and provides greater flexibility for the use of outgroups.

207 Our testing of the algorithms implemented by Yang and Smith (2014) across different datasets, here
208 exemplified with a set of fourteen Australian Asteraceae, illustrated the benefit of the removal of
209 paralogs, the benefit of including genes exhibiting paralogy, and the relative performance of the topology-
210 based approaches. The phylogeny inferred without formal ortholog resolution deviated from all others,
211 suggesting that its topology is influenced by unrecognised paralogy (Fig. 4a). Removing all genes
212 showing paralogy, however, produced the smallest dataset, albeit with slightly more informative
213 characters than the results of RT (Table 2). This effect would be stronger in larger datasets, as the number
214 of gene files containing at least one paralog increases with the number of species in the analysis.

215 Similarly, the number of species with paralogs will increase with the number of genes, and vice-versa.

216 As expected, Maximum Inclusion (MI) produced the largest paralog-free dataset, and the resulting
217 phylogeny was not an outlier among those derived from the paralog-free datasets (Fig. 4b). Rooted

218 subTrees (RT) separated out the largest number of ortholog groups but resulted in the smallest dataset
219 after filtering for a minimum number of terminals per ortholog group, an artefact of the small size of the
220 example dataset. In larger test datasets, this approach frequently produced more informative datasets than
221 Monophyletic Outgroups (MO) (Schmidt-Lebuhn, unpubl. data).

222 Depending on the data, additional processing may be desirable before phylogenetic analysis, e.g. to
223 ensure that all genes are in the correct frame if protein-coding. Nevertheless, hybpiper-rbgv and yang-
224 and-smith-rbgv greatly streamline the assembly of target enrichment data and inference of ortholog
225 groups, making these methods more accessible and easier to use by those working with target capture
226 dataset.

227 **ACKNOWLEDGMENTS**

228 We are grateful to Matt G. Johnson, Stephen A. Smith, and Ya Yang for discussions on and clarifying the
229 use of their pipelines, and to Theodore Allnut, Jason Bragg, Johan Gustafsson, Cameron Jack, and Lars
230 Nauheimer for helpful conversations and feedback during development. Sequencing was funded by
231 Bioplatforms Australia through the Genomics for Australian Plants (GAP) initiative as part of the
232 phylogenomics project Australian Angiosperm Tree of Life (AAToL). We used the sequencing services
233 provided by the Australian Genome Research Facility (AGRF).

234 **DATA AVAILABILITY**

235 The hybpiper-rbgv and yang-and-smith-rbgv containers are available at [https://github.com/chrisjackson-](https://github.com/chrisjackson-pellicle)
236 [pellicle](https://github.com/chrisjackson-pellicle). The example dataset, HybPiper statistics, target file, and outgroup file are available at the CSIRO
237 Data Access Portal (DOI: [10.25919/q42q-j056](https://doi.org/10.25919/q42q-j056)). The raw reads of the example dataset are available in the
238 Bioplatforms Data Portal (<https://data.bioplatforms.com/>) under sample numbers 79649, 79652, 80014,
239 80042, 80066, 80070, 80071, 80082, 80088, 80089, 80105, 80109, 80123, and 80125.

240 **LITERATURE CITED**

Altenhoff, A. M., B. Boeckmann, S. Capella-Gutierrez, D. A. Dalquen, T. DeLuca, K. Forslund, J.

Huerta-Cepas, et al. 2016. Standardized benchmarking in the quest for orthologs. *Nature Methods*
13: 425–430.

Baker, W. J., P. Bailey, V. Barber, A. Barker, S. Bellot, D. Bishop, L. R. Botigué, et al. 2021. A

Comprehensive Phylogenomic Platform for Exploring the Angiosperm Tree of Life. *Systematic Biology*.

Bejerano, G., M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler. 2004.

Ultraconserved elements in the human genome. *Science* 304: 1321–1325.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120.

Breinholt, J. W., S. B. Carey, G. P. Tiley, E. C. Davis, L. Endara, S. F. McDaniel, L. G. Neves, et al.

2020. A target enrichment probe set for resolving the flagellate plant tree of life. *bioRxiv*: 2020.05.29.124081.

Dessimoz, C., T. Gabaldón, D. S. Roos, E. L. L. Sonnhammer, J. Herrero, and the Q. for O. Consortium.

2012. Toward community standards in the quest for orthologs. *Bioinformatics* 28: 900–904.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.

Nucleic Acids Research 32: 1792–1797.

Johnson, M. G., E. M. Gardner, Y. Liu, R. Medina, B. Goffinet, A. J. Shaw, N. J. C. Zerega, and N. J.

Wickett. 2016. HybPiper: extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences* 4: 1600016.

Johnson, M. G., L. Pokorny, S. Dodsworth, L. R. Botigué, R. S. Cowan, A. Devault, W. L. Eiserhardt, et

al. 2019. A universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering. *Systematic Biology* 68: 594–606.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7:

improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Lemmon, A. R., S. A. Emme, and E. M. Lemmon. 2012. Anchored Hybrid Enrichment for massively

high-throughput phylogenomics. *Systematic Biology* 61: 727–744.

Liu, Y., M. G. Johnson, C. J. Cox, R. Medina, N. Devos, A. Vanderpoorten, L. Hedenäs, et al. 2019.

Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nature Communications* 10: 1485.

- Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.
- Mandel, J. R., R. B. Dikow, V. A. Funk, R. R. Masalia, S. E. Staton, A. Kozik, R. W. Michelmore, et al. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences* 2: 1300085.
- McCormack, J. E., S. M. Hird, A. J. Zellmer, B. C. Carstens, and R. T. Brumfield. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution* 66: 526–538.
- McLay, T. G. B., J. L. Birch, B. F. Gunn, W. Ning, J. A. Tate, L. Nauheimer, E. M. Joyce, et al. 2020. New targets acquired: improving locus recovery from the Angiosperms353 probe set. *bioRxiv*: 2020.10.04.325571.
- Morales-Briones, D. F., B. Gehrke, C.-H. Huang, A. Liston, H. Ma, H. E. Marx, D. C. Tank, and Y. Yang. 2020. Analysis of paralogs in target enrichment data pinpoints multiple ancient polyploidy events in *Alchemilla* s.l. (Rosaceae). *bioRxiv*: 2020.08.21.261925.
- Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274.
- Schmidt-Lebuhn, A. N., and J. Bovill. Phylogenomic data reveal four major clades of Australian Gnaphalieae (Asteraceae). *TAXON* doi: 10.1002/tax.12510.
- Wolf, P. G., T. A. Robison, M. G. Johnson, M. A. Sundue, W. L. Testo, and C. J. Rothfels. 2018. Target sequence capture of nuclear-encoded genes for phylogenetic analysis in ferns. *Applications in Plant Sciences* 6: e01148.
- Yang, Y., and S. A. Smith. 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.
- Zhang, C., M. Rabiee, E. Sayyari, and S. Mirarab. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19: 153.

242 **TABLES**

243 **TABLE 1.** Comparison of command line entries required to run containerized hybpiper-rbgv and yang-
 244 and-smith-rbgv against the original implementations of the pipelines, excluding command line arguments.
 245 Optional steps are bracketed. Note that additional steps were required to make HybPiper outputs directly
 246 usable in the Yang and Smith (2014) pipeline.

Commands to run containers	Commands to run original pipelines	Function
nextflow run hybpiper-rbgv-pipeline.nf	reads_first.py	Assemble reads to contigs, build exon sequences
	cleanup.py	Delete temporary files
	get_seq_lengths.py	Summarize gene reference lengths
	hybpiper_stats.py	Summarize gene recovery efficiency and paralog warnings for each sample
	retrieve_sequences.py	Generate sequence files for each gene, choosing one paralog each by length and read coverage
	(intronerate.py)	Retrieve intron sequences
	(paralog_investigator.py)	Report number of paralogs found for each gene
	(paralog_retriever.py)	Generate sequence files for each gene including all paralogs
nextflow run yang-and-smith-rbgv-pipeline.nf	fasta_to_tree.py	Align sequence files and infer gene trees
	trim_tips.py	Trim long terminals,

		suspected assembly errors
	mask_tips_by_taxonID_transcripts.py	Remove superfluous alleles from same species
	cut_long_internal_branches.py	Cut suspected deep paralogs
	write_fasta_files_from_trees.py	Create sequence files for samples left after trimming
	filter_1to1_orthologs.py prune_paralogs_MO.py prune_paralogs_RT.py prune_paralogs_MI.py	Infer ortholog groups using alternative algorithms
	write_alignments_from_orthologs.py	Create sequence files for each ortholog group

247

248

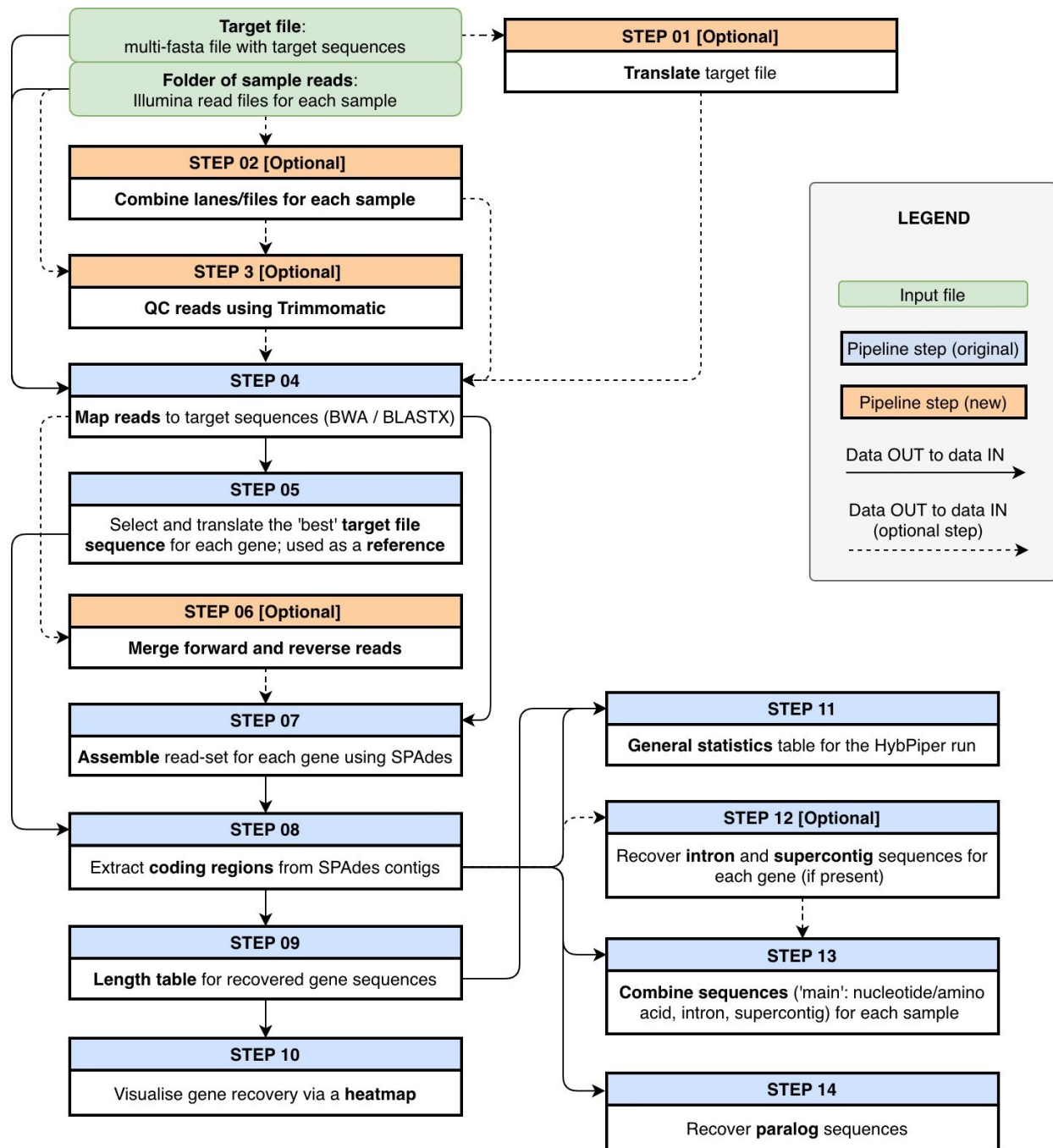
249 **TABLE 2.** Dataset sizes resulting from different algorithms for the inference of ortholog groups in a test
250 dataset of fourteen Australian Asteraceae. In larger datasets, the use of paralog-free genes only is likely to
251 result in relatively smaller datasets, and that of the Rooted subTrees algorithm in relatively larger ones.

Algorithm	Ortholog groups per species, min-max (median), after filtering for ≥ 5 terminals	Characters			
		Total	Parsimony informative	Variable but uninformative	Constant
No ortholog inference	296-345 (340.5)	273,042	34,485	49,600	188,957
Monophyletic Outgroups	245-293 (286)	210,090	27,836	36,530	145,724
Rooted subTrees	224-253 (235)	209,613	19,933	34,319	155,361
Maximum Inclusion	306-352 (348)	251,499	32,095	42,815	176,589
Only paralog- free genes	229-273 (268)	195,822	25,883	34,239	135,700

252

253

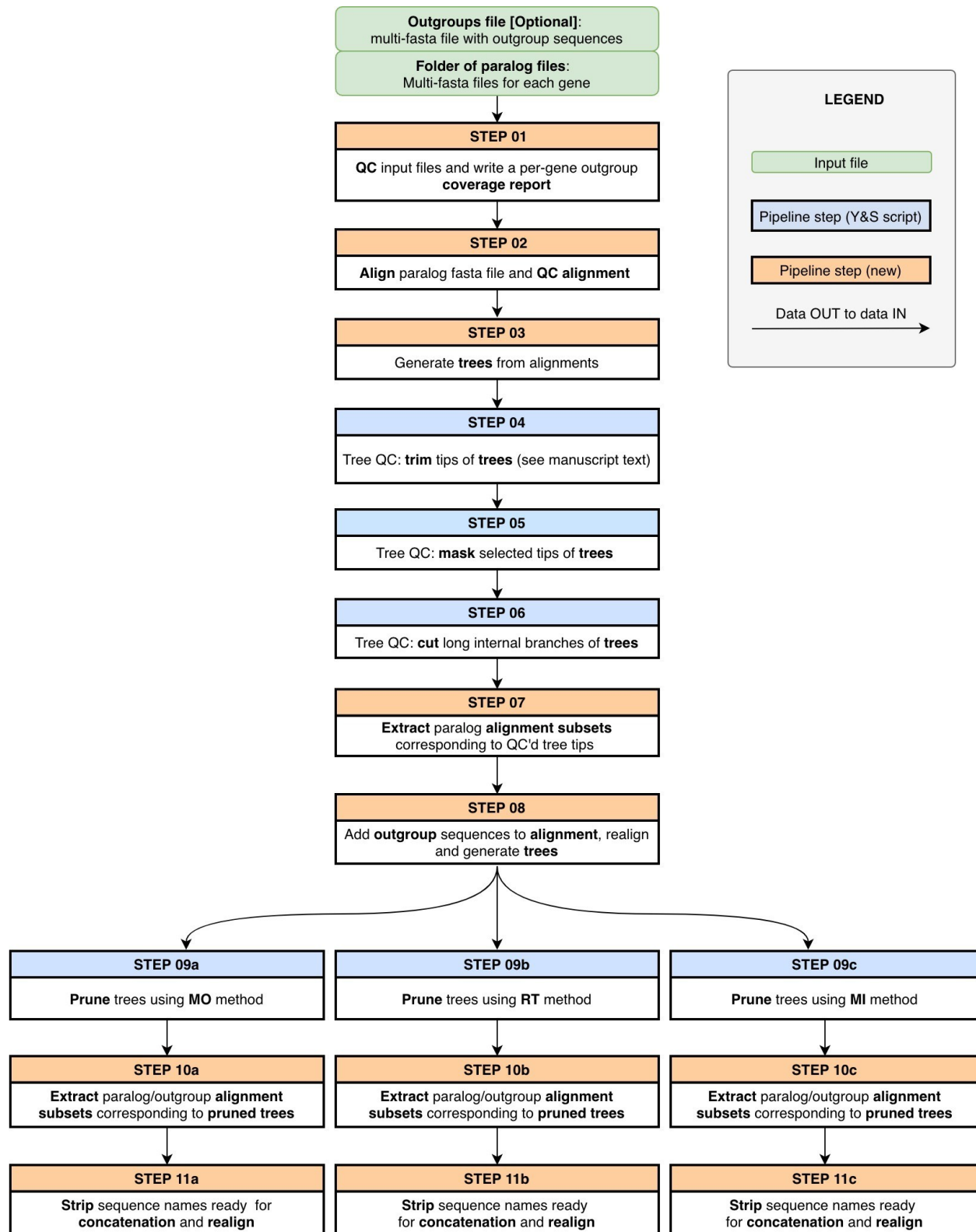
254 **FIGURE LEGENDS**



255

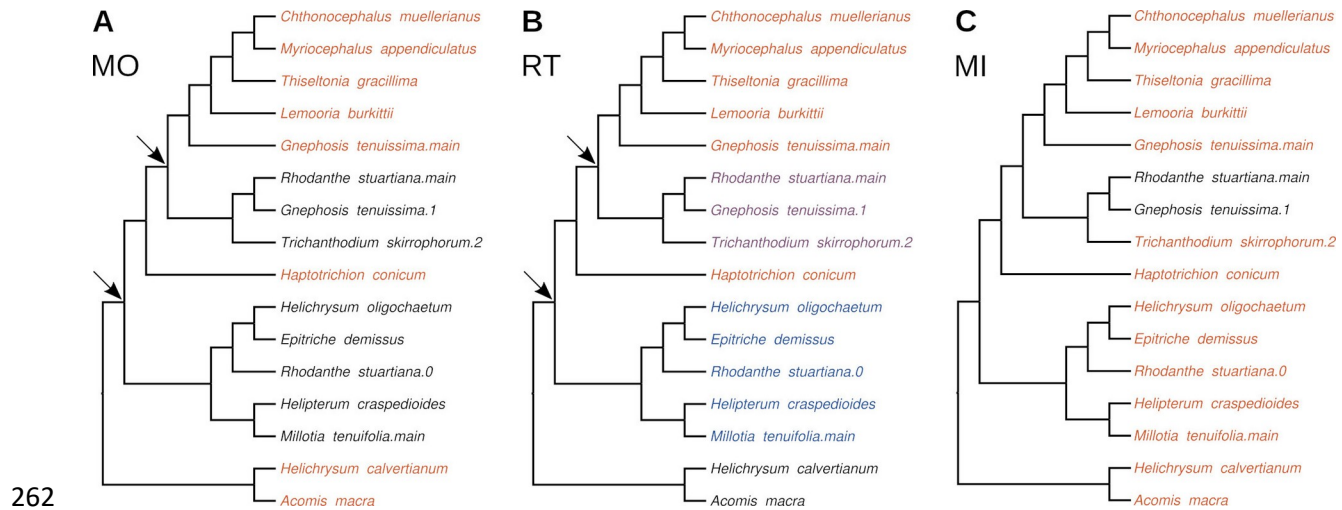
256 **FIGURE 1.** Flowchart summarizing the hybpiiper-rbgv pipeline for assembly of sequence capture or

257 target enrichment data.



258

259 **FIGURE 2.** Flowchart summarizing the yang-and-smith-rbgv pipeline, which uses gene tree topology to
 260 resolve paralogy, assuming that gene or genome duplication events caused samples to be duplicated in
 261 different gene tree clades.



262

263 **FIGURE 3.** Illustration of algorithms for inference of orthologs using one gene tree as example. (A)

264 Monophyletic Outgroups (MO) moves iteratively through the tree from the root, checks at each node if

265 samples are duplicated between the descendent sister clades, and, if so, removes the smaller descendent

266 sister clade, here retrieving the terminals marked in red. (B) Rooted subTrees (RT) proceeds as MO but

267 separates the smaller descendent sister clades into distinct ortholog groups. In this case, this approach

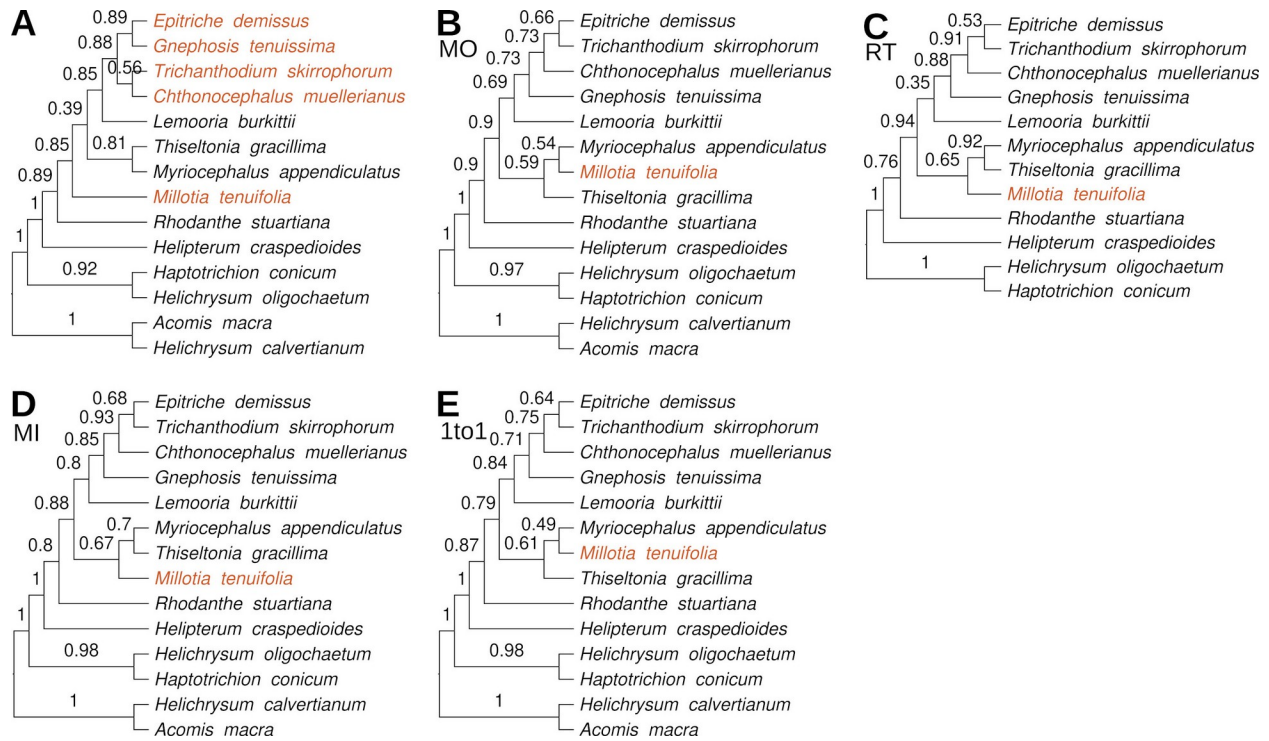
268 results in the retrieval of three ortholog groups marked in red, blue, and purple. (C) Maximum Inclusion

269 (MI) iteratively retrieves the largest unrooted subtrees without duplicated samples, in this case resulting in

270 a single ortholog group marked in red. The gene tree is presented in cladogram view. Arrows indicate

271 instances of ancestral gene duplication inferred by MO and RT. Name elements after stops are paralog

272 identifiers assigned by HybPiper.



273

274 **FIGURE 4.** Results of phylogenetic analysis of the example dataset with ASTRAL, using data from
 275 orthology inference by (A) hybpiper-rbgv directly, based on length and read coverage, (B) Monophyletic
 276 Outgroups, (C) Rooted subTrees, (D) Maximum Inclusion, and (E) exclusion of all genes with paralogs.
 277 Outgroup is missing in (C) because the RT algorithm removes it. Numbers above branches indicate clade
 278 support from local posterior probability. Red font color marks a species placed in differing positions and a
 279 clade whose internal structure differs in (A), whereas the remainder of the topology is constant across
 280 analyses.

281