HybPhaser: a workflow for the detection and phasing of hybrids in target capture datasets

Lars Nauheimer[1,2,3,6], Nicholas Weigner[1], Elizabeth Joyce[1,3], Darren Crayn[1,2,3], Charles Clarke[1,4], Katharina Nargar[1,5]

[1] Australian Tropical Herbarium, James Cook University, McGregor Road, Smithfield, Qld 4878, Australia.
[2] Centre for Tropical Bioinformatics and Molecular Biology, James Cook University, McGregor Road, Smithfield, Qld 4878, Australia.
[3] Centre for Tropical Environmental Sustainability Science, James Cook University, McGregor Road, Smithfield, Qld 4878, Australia.
[4] Cairns Botanic Gardens, Collins Avenue., Edge Hill, Qld 4870, Australia.
[5] National Research Collections Australia, Commonwealth Industrial and Scientific Research Organisation (CSIRO), GPO Box 1700, Canberra, ACT 2601, Australia.

[6]Author for correspondence: larsn@mailbox.org

Abstract

- *Premise of the study:* Hybrids contain divergent alleles that can confound phylogenetic analyses but can provide insights into parental lineages when identified and phased. We developed HybPhaser to detect hybrids in target capture datasets and to phase reads according to haplotypes based on similarity and a phylogenetic framework.
- *Methods and Results:* HybPhaser is an extension to the HybPiper sequence assembly workflow. We used Angiosperms353 target capture data for *Nepenthes* including known hybrids to test the novel workflow. Reference mapping was used to record heterozygous sites and identify hybrid accessions that were phased by mapping sequence reads to multiple references. The parental lineages of known

hybrids were confirmed and conflicting phylogenetic signal reduced, improving the outcomes of phylogenetic analysis.

- *Conclusions:* HybPhaser is a novel pipeline for summarizing and optimizing target capture datasets, detecting hybrid accessions as well as paralogous genes, and generating phased accessions that can provide insights into reticulated evolution.

Key words: Angiosperm353; HybPiper; introgression; paralogs; polyploidy; reticulation

Introduction

Reticulation events caused by hybridization are common and important sources of novelty in angiosperm evolution (Wood et al. 2009, Palfalvi et al. 2020). However, detection, investigation and representation of hybridization remains a challenge in phylogenomics (Kellog 2016, Mallet et al. 2016, Spooner et al. 2020). The combination of divergent genomes in hybrids (herein used for any organism that contains divergent genomes due to a hybridization event) introduces conflicting phylogenetic signal and can lead to topologically incorrect or poorly resolved phylogenetic trees (McDade 1992, Soltis et al. 2008). However, the advancement of target-capture data and universal probe-kits such as Angiosperms353 (Johnson et al. 2018) provides an opportunity to gain insight into historical reticulations in angiosperm evolution and reduce phylogenetic conflict, if ortholog gene variants (or homeolog gene variants in polyploids) can be identified and separated (phased).

Previously, inclusion of phased gene variants in phylogenetic studies using single or low numbers of loci has been used to confirm hybrid status of organisms (Sang and Zhang 1999), reveal their parental lineages (Estep et al. 2014), enable reconstruction of past reticulations (Estep et al. 2014, Brassec and Blattner 2015), and date ancient hybridization events (Marcussen et al. 2015). However, this approach was limited by the resolution of single locus phylogenies and the lack of methods available to link phased alleles when multiple loci are used. The availability of universal target capture sequencing bait-kits such as Angiosperms353 has revolutionized the study of angiosperm evolution by enabling the study of evolution across a phylogenetically broad group with unprecedented resolution (Johnson et al. 2018). Moreover, it provides new opportunities for investigating reticulation events in angiosperm evolution, as target capture datasets contain sequence reads from all gene

variants in hybrids. However, no standard pipeline for the phasing of allele variants or paralogs currently exists for target capture data.

Common sequence assembly pipelines such as PHYLUCE (Faircloth 2016), HybPiper (Johnson et al. 2016) and SECAPR (Anderman et al. 2018) use *de novo* sequence assembly, which combines reads from divergent allele and gene variants into chimeric sequences leaving hybrid accessions undetected. Although alleles can be phased to some extent using read-backed phasing (Kates et al. 2018, Anderman et al. 2019), linking phased alleles has been performed only by manual inspection of gene trees, which is not feasible for target capture datasets (Kellogg 2016).

Here, we present HybPhaser, a bioinformatics workflow (Appendix 1) for the detection and phasing of hybrid accessions in target capture datasets. HybPhaser works as an extension to HybPiper by remapping reads onto HybPiper's *de novo* assembled sequences to generate consensus sequences with ambiguity codes where reads from alleles or gene variants differ. This enables the quantification of heterozygous sites across the dataset, and thus the detection of hybrids and paralogs. Customizable scripts can be used to optimize the data regarding missing data and putative paralogs, as well as collation of sequences for further analysis.

Finally, HybPhaser provides the option to phase sequence reads of hybrid samples into accessions that approximate haplotypes. This is achieved by simultaneously mapping sequence reads to multiple references from clades selected across a framework phylogeny and separating reads according to their similarity to those clades. The phased accessions can then be assembled and analyzed phylogenetically, revealing insights into reticulated evolution. Herein we detail the HybPhaser pipeline and demonstrate its utility for detecting hybrids and identifying parental lineages using Angiosperms353 data from the carnivorous plant genus *Nepenthes* L. (Nepenthaceae).

Methods and results

Preparation of the dataset and assembly

Study group

Nepenthes is a palaeotropical genus of c. 160 species known to readily hybridize in horticulture and nature (Clarke et al. 2018). Molecular phylogenetic studies of the genus based on a few Sanger-sequenced loci (e.g., Meimberg and Heubl 2006, Alamsyah and Ito 2013) have resulted in poorly resolved and supported trees due to limited resolution of the markers, incongruence between plastid and nuclear phylogenies, and the possible inclusion of ITS paralogs. Recently, phylogenomic approaches including genome skimming (Nauheimer et al. 2018) and Angiosperms353 target capture (Murphy et al. 2020) have provided great advances in resolving the *Nepenthes* phylogeny. In a study that sampled almost all recognized species in the genus, Murphy et al. (2020), included one known hybrid (*N.* x *hookeriana* H.Low) and inferred a new putative hybrid (*N. angustifolia* Mast.) from conflict between concatenated and gene tree analyses. Given the availability of suitable data and presence of known hybrids, *Nepenthes* is an ideal group for demonstrating hybrid detection and phasing using HybPhaser and the Angiosperms353 probe set.

Input data

We generated Angiosperms353 target capture data for 125 samples, including 17 of known hybrid origin. Further, we downloaded the Murphy et al. (2020) *Nepenthes* raw read files from the NCBI sequence read archive (Bioproject PRJEB35235, 185 taxa including one known and one suspected hybrid) and added them to the dataset. In total, 310 accessions representing 141 *Nepenthes* taxa and 18 known hybrids were analyzed (Appendix 2).

DNA extraction, library enrichment and sequencing

Total genomic DNA was isolated from leaf material stored in silica gel for 125 *Nepenthes* samples. DNA was extracted using the CTAB method and DNA extracts were further processed at the Australian Genomic Research Facility (AGRF; Melbourne, Australia) for library preparation and sequencing. Target sequence capture was carried out using the 'Angiosperms 353 v1' universal probe-set (ArborBioscience, Ann Arbor, USA) following Johnson et al. (2018). Sequencing was performed on an Illumina HiSeq 2500 sequencer producing 150bp paired-end reads.

Read trimming

The sequence reads were trimmed using Trimmomatic (v.0.39, Bolger et al. 2014) to remove sequencing adapters and poor-quality bases (illuminaclip 2:30:10, leading 20, trailing 20, sliding window 4:20). Final reads of less than 30 nucleotides were excluded.

HybPiper sequence assembly

Sequence assembly was performed with HybPiper using the Angiosperm353 target file in fasta format (https://github.com/mossmatters/Angiosperms353) and the -bwa flag. The script "intronerate.py" was run to retrieve supercontig sequences. Summary statistics were generated using the script "hybpiper_stats.py" (Appendix 3).

HybPhaser part I: Assessment of SNPs

Generation of consensus sequences

In the first part of the workflow (Table 1), heterozygous sites for each sample and locus were quantified using the bash script 'Bash1_generate_consensus_sequences.sh'. HybPiper's pre-selected reads were mapped to the *de novo* assembled contig for each locus using BWA (Li and Durbin 2009). To capture heterozygous sites across reads for each locus, bcftools (http://samtools.github.io/bcftools/bcftools.html) was then used to generate a consensus sequence with ambiguity characters at each site that contained a single nucleotide polymorphism (SNP). Ambiguity characters were only coded if the read depth was ≥10, the alternative nucleotide count was ≥4, and the proportion of alternative nucleotides was ≥0.15 to prevent cases of sequencing artefacts being coded as SNPs. The script "Rscript_1a_count_snps_in_consensus_seqs.R" was then used to collate information on the proportions of SNPs and length of sequences from all samples and loci (Appendix 4).

Reducing proportions of missing data

HybPhaser enables the removal of loci and samples that have poor sequence recovery using the R script 'R1b_optimize_dataset.R'. We removed all loci that had sequences recovered for less than 20% of samples and all samples that had less than 20% of loci or less than 50% of the target sequence length recovered. The threshold for target sequence length was reduced to 48% to include the hybrid accession *N. bokorensis* Mey x *ventricosa* Blanco. The output of this script reported that three samples and eight loci failed, without any sequences. Across the remaining loci, 32 were recovered for fewer than 62 of 302 samples, seven samples had fewer than 69 loci (of 351) recovered, and 13 more samples had less than 48% of the target sequence length recovered. In total, 35 loci and 23 low quality samples were removed from the dataset (Appendix 5).

Removal of paralogous genes

Genes with unusually high proportions of SNPs compared to other genes might have heterozygous sites due to the presence of paralogous genes or contamination. Recent paralogs might be restricted to a single species, while ancient paralogs can be shared across many taxa. Removing genes for all samples that have elevated levels of SNPs across several samples can remove paralogous genes where only a single copy is present. In the script 'R1b_optimize_dataset.R' putative paralogs can be removed across all samples and for each sample individually. Here, we removed outlier genes (defined as having values higher than 1.5x the interquartile range above the third quartile) of the mean proportion of SNPs across all samples, which affected 24 genes that had more than average of 0.007 SNP/bp across all samples. We further analyzed the distribution of proportion of SNPs per gene in each sample and removed genes being outliers for each sample individually, on average of 7.4 genes for each sample (Appendix 5).

Assessment of heterozygosity and allele divergence

Hybrid samples are expected to contain divergent alleles from their parental species and thus display high heterozygosity (here used as the proportion of genes with SNPs) and an allele divergence (proportion of SNPs per sequence) similar to the divergence between parental lineages. The script "Rscript_1c_summary_table.R" collects the information on SNPs across the dataset and generates tables and figures that enable the assessment of heterozygosity and allele divergence to identify putative hybrid samples.

Across all *Nepenthes* samples, mean heterozygosity was 53.1% (median 54.25%) and mean allele divergence was 0.21% (median 0.15), while known hybrids had a considerably higher mean heterozygosity of 78% (median: 92.9%) and mean allele divergence of 78% (median: 85%) (Appendix 6). Three samples identified as *Nepenthes angustifolia, N. kampotiana* Lecomte and *N. bellii* K.Kondo, had high heterozygosity (>90%) and high allele divergence (>1%) and thus represent previously unidentified hybrids. *Nepenthes angustifolia* was already suspected to be a hybrid by Murphy et al. (2020). Four known hybrids, *N. mirabilis* (Lour.) Druce x *tenax* C.Clarke & R.Kruger 1 & 2, *N. mirabilis* x *rowaniae* F.M.Bailey, and *N. boschiana* Kort x *glandulifera* Chi.C.Lee, had low proportions of heterozygous loci (20–40%) and very low allele divergence (0.09–0.22%), thus seemed not to contain divergent haplotypes. *Nepenthes boschiana* x *glandulifera* had poor sequence recovery (50.5% of target) and low sequence read coverage, which can result in underrepresentation of SNPs in the consensus sequences.

Finally, the script 'Rscript_1d_generate_sequence_lists.R' was used to collate the assembled and mapped sequences to generate multiple sequence lists containing HybPiper contigs and HybPhaser consensus sequences for each locus and sample, with (clean) and without dataset optimization (raw).

HybPhaser part II: Clade association

Framework phylogeny alignment generation

The HybPhaser output sequence lists containing consensus sequences for each locus were aligned using MAFFT (Kotah and Standley 2013) and columns with more than 50% gaps were removed using TrimAl (Capella-Gutierrez et al. 2009). Phylogenetic analyses were performed using maximum likelihood in the IQ-TREE2 package (version 2.0.3, Minh et al. 2019). A supermatrix phylogeny was created by concatenating all locus alignments and using the 'edge proportional' partition model with single loci as partitions. Phylogenetic clade support was estimated using ultra-fast bootstrap with 1000 replicates (Hoang et al. 2018) and gene and site concordance factors (gCF, sCF; Minh et al. 2020), which indicate the concordance of single gene trees and of 1000 randomly chosen parsimony informative sites with the phylogeny of the concatenated dataset. To determine gCF, single gene phylogenies were generated using IQ-TREE2 applying the best fit model for each locus selected by ModelFinder (Kalyaanamoorthy et al. 2017). The supermatrix phylogeny was rooted using both accessions of *Nepenthes pervillei* Blume, the sister to all other *Nepenthes* species (Murphy et al. 2020).

The supermatrix dataset consisted of 289 genes and 290 taxa, with an aligned sequence length of 423,105 base pairs (bp) and 30.4% missing data. Overall, clades in the recovered phylogeny were well supported, and largely corresponded to the clades retrieved by Murphy et al. (2020) (Fig. 2B, Appendix 7). Many major clades obtained maximum bootstrap support, although the gCF values were often low. Some of the backbone nodes received very low support. Most hybrid accessions fell within the clade containing one of the parental taxa where the hybrid was often found in a basally diverging position within the clade (e.g. *N. ampullaria* x *gracilis*; Appendix 7).

Selection of clade references

The haplotypes of putative hybrids can be separated when reads associate with divergent clades. In order to assess the clade association of each sample, HybPhaser maps sequence reads simultaneously to multiple references that represent major lineages of the study group. We chose samples with low heterozygosity and allele divergence and high sequence recovery (number of loci sequenced and proportion of target sequence) that were evenly distributed across the phylogeny as clade references. Based on the generated phylogeny, we selected 25 samples representing the major lineages in *Nepenthes* (Fig. 2A, Appendix 7, 8). The consensus sequences generated by HybPhaser that contained all loci without applying the dataset optimization step (consensus_samples_raw) were used as references.

Clade association

The proportion of reads that map to a certain reference is influenced by multiple factors, including the phylogenetic distance of the reads to the reference (a closer relationship leads to more reads matching), the phylogenetic divergence between the selected clade references (a closer relationship reduces the proportion of reads matching unambiguously to either reference), and the proportion of sequence reads in the original read files matching to the target capture baits. To mitigate the effects of variation in reads matching to the target sequences, we only used reads for the clade association that were mapped by HybPiper to the target sequences. We extracted those reads from the HybPiper generated bam file using the bash script "extract_mapped_reads.sh" generating a new read file containing only reads that mapped to the target sequences. These read files were mapped simultaneously to all 25 selected clade references recording the proportions of reads that mapped to either reference unambiguously using BBSplit (BBMap v38.47). The script 'Rscript_2a_prepare_BBSplit_script.R' was used to generate a BBSplit commands file with the selected clade reference names and abbreviations. The mapping results were summarized using the script 'Rscript_2b_collate_BBSplit_results.R'.

Interpretation of clade association must be done for each sample individually, taking into consideration the phylogenetic distance between that sample and the clade references, and the phylogenetic distance between clade references. In most *Nepenthes* samples reads matched to one or few clade references with high proportions (> 3%) relative to the remaining clade references (>1%; Table 2A, Appendix 8). The highest proportions of matching reads were found in phylogenetically distinct, basally diverging lineages (e.g. > 60% in *N. khasiana* Hook.f.), while multiple clade associations were found in taxa with low heterozygosity and

low allele divergence when they were not close to or in between two clade references (e.g. *N. suratensis* M.Catal). Of the 16 known hybrids, seven were clearly associated with two clades (i.e. had a high proportion of reads matching to references from those clades compared to other clade references), while *N. ampullaria* Jack x *tobaica* Danser were associated with three clades.

The known hybrids *N. burkei* Hort.Veitch x *ventricosa* 1 & 2 and *N. veichii* Hook.f. x *eymae* Sh.Kurata, had a single clade association with a moderate allele divergence (>55%) and high heterozygosity (>80%), suggesting they likely constitute hybrids of closely related parents. This was confirmed with a secondary, fine-scale clade association analysis, whereby each of these three hybrid accessions were mapped to nine references from the 'maxi' and 'merr' clades, and each successfully associated with two references closely-related to their parental taxa (Appendix 8).

The three "known" hybrids in the *N. mirabilis* complex (*N. mirabilis* x *rowaniae* and *N. mirabilis* x *tenax* 1, 2) were only associated with one clade, and exhibited very low heterozygosity and very low allele divergence in the HybPhaser heterozygosity assessment step. This is likely the result of these samples being from a species complex in need of taxonomic revision and suggests that these "known" hybrids are not hybrids but may be part of the same species. Of the three putative hybrid taxa identified in the heterozygosity assessment step (*N. angustifolia*, *N. kampotiana* 1 and *N. bellii*), only *N. bellii* was clearly associated with two clades. This supports the inference that *N. bellii* is of hybrid origin and that the parental lineages are distantly related enough for the accession to be phased. The high allele divergences of *N. angustifolia* and *N. kampotiana* 1 suggest that they are hybrids, but their parental clades were not sampled.

HybPhaser part III: Phasing

Phasing of sequence reads

The information provided by the clade association analysis was used to phase the read files into phased accessions by mapping reads only to the relevant clade references using BBSplit and generating new read files. Reads that matched unambiguously to a single reference were written into the respective read file, while reads that matched ambiguously to multiple references were written to all read files for that sample. This separated reads according to their similarity to divergent clades. Ideally, hybrids containing one haplotype from each parent will be phased into two accessions representing both haplotypes. Similarly, polyploids containing multiple haplotypes can be phased into multiple accessions. However, because

ambiguous reads were assigned to all phased accessions, the phased accessions only approximated haplotypes.

The previously selected twelve samples with multiple clade associations were phased using BBSplit run by a bash script prepared using the script "Rscript_3a_ prepare_phasing_script.R" and phasing stats were collated into a table using the script "Rscript_3b_collate_phasing_stats.R". The proportions of reads mapped to the selected references in the phasing step can give insights into the proportions of haplotypes and thus ploidy levels. Most phased accessions had approximate equal proportions indicating haplotypes were evenly distributed in the clades (Appendix 9). However, *Nepenthes bicalcarata* Hook.f. x *ampullaria* had a 3:1 proportion indicating a ploidy level that is a multiple of four. *Nepenthes burkei* x *ventricosa* and *N. veitchii* x *eymae* had proportions of roughly 3:2 indicating a more complex ploidy level.

Processing of haplotype accessions

The newly created read files of the phased accessions were passed through the HybPiper and HybPhaser pipelines using the same settings as for the non-phased samples. To optimize the dataset, we removed the same 24 loci flagged as putative paralogs for the non-phased accessions and performed the paralog removal step for each sample using the same threshold as for the non-phased samples (Appendix 10). Phased accessions had considerably lower heterozygosity and allele divergence compared to the same accessions before phasing (Appendix 10).

Combined dataset of phased and non-phased accessions

We combined the sequence lists of the phased accessions with the sequences of the non-phased samples using the script "Rscript_4a_merge_phased_with_normal_sequence_lists.R".

In order to reduce conflicting phylogenetic signal from this dataset, we excluded all accessions that were supposedly hybrids but were not phased due to unclear clade association (*N. angustifolia* and *N. kampotiana-1* as well as *N. clipeata* Danser x *rafflesiana* Jack*, N. bokorensis* x *ventricosa*, *N. rafflesiana* x *ampullaria, N. faizaliana* J.H.Adam & Wilcock x *campanulata* Sh.Kurata, *N. boschiana* x *glandulifera*). However, we kept *N. mirabilis* x *rowaniae* and *N. mirabilis* x *tenax* 1 & 2 in the dataset, as they appeared not to be hybrids but part of a species complex. Alignments and phylogenetic analyses were performed with the same methodology to select clade references described above.

Almost all phased accessions of known hybrids grouped in the clades they were phased to and sister to the known parent (Figure 2C, Appendix 11). The only exception was *N. ampullaria* x *tobaica*, which had one accession not grouping in the clade it was phased to ("raja") but sister to one parental clade indicating that the interpreted clade association was wrong. However, both other phased accessions grouped with the expected parents.
The previously unknown parental lineages of *N.* x *hookeriana* was revealed to be *N. ampullaria* and *N. rafflesiana* and the accession *N. bellii*-1 was identified as hybrid between *N. bellii* and *N. mirabilis* (Figure 2C, Appendix 11). Finally, the support for the backbone nodes and for the clades that contained hybrids increased compared to the phylogeny with non-phased accessions due to the reduction of conflicting phylogenetic signal. This was especially prominent regarding gene concordance factors showing that phasing hybrids reduced conflicting information and increased concordance in the dataset.

HybPhaser is compatible with Linux and consists as a collection of bash and R scripts that are protected under the terms of a free software license agreement GNU and are available online at the NCBI sequence read archive.

Conclusions

HybPhaser is a bioinformatics workflow that extends the functionality of the HybPiper assembly pipeline to detect and phase hybrids. Using *Nepenthes* as case study, we demonstrated that hybrids can be detected by assessing the occurrence of heterozygous sites in nuclear genes. The hybrid status of 14 known hybrids was confirmed and an additional three taxa were identified as putative hybrids. Twelve hybrids were phased into accessions that approximated divergent haplotypes. Their inclusion in phylogenetic analyses confirmed the parental lineages of known hybrids and revealed the origin of a previously unknown hybrid sample. Phasing of hybrids and the removal of non-phased hybrids reduced conflicting signal and improved the resolution and support on the resulting phylogenetic trees. Importantly, using HybPhaser for analysis of Angiosperms353 has great promise for investigating reticulation events throughout angiosperm evolution at broader scale.

Acknowledgements

This work was supported by the Australian Biological Resource Study (Department of the Environment, Australian Government, ABRS grant RFL214-62) as well as the Australian and

Pacific Science Foundation (APSF 16/4). We are thankful to Melissa Harrison for support in the laboratory and Todd McLay and Chris Jackson for feedback on the HybPhaser workflow.

Literature Cited

Alamsyah, F., and M. Ito. 2013. Phylogenetic analysis of Nepenthaceae, based on internal transcribed spacer nuclear ribosomal DNA sequences. *Acta Phytotaxonomica et Geobotanica* 64: 113–126.

Andermann, T., Á. Cano, A. Zizka, C. Bacon, and A. Antonelli. 2018. SECAPR—a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. *PeerJ* 6: e5175.

Andermann, T., A. M. Fernandes, U. Olsson, M. Töpel, B. Pfeil, B. Oxelman, A. Aleixo, et al. 2019. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Systematic Biology* 68: 32–46.

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

Brassac, J., and F. R. Blattner. 2015. Species-level phylogeny and polyploid relationships in *Hordeum* (Poaceae) inferred by next-generation sequencing and in silico cloning of multiple nuclear loci. *Systematic Biology* 64: 792–808.

Capella-Gutierrez, S., J. M. Silla-Martinez, and T. Gabaldon. 2009. TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.

Clarke, C., J. Schlauer, J. Moran, and A. Robinson. 2018. Systematics and evolution of *Nepenthes*. Oxford University Press.

Estep, M. C., M. R. McKain, D. V. Diaz, J. Zhong, J. G. Hodge, T. R. Hodkinson, D. J. Layton, et al. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences* 111: 15149–15154.

Faircloth, B. C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32: 786–788.

Hoang, D. T., O. Chernomor, A. von Haeseler, B. Q. Minh, and L. S. Vinh. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35: 518–522.

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. 2018. 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.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. von Haeseler, and L. S. Jermiin. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.

Kates, H. R., M. G. Johnson, E. M. Gardner, N. J. C. Zerega, and N. J. Wickett. 2018. Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *American Journal of Botany* 105: 404–416.

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.

Kellogg, E. A. 2016. Has the connection between polyploidy and diversification actually been tested? *Current Opinion in Plant Biology* 30: 25–32.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? *BioEssays* 38: 140–149.

Marcussen, T., L. Heier, A. K. Brysting, B. Oxelman, and K. S. Jakobsen. 2015. From gene trees to a dated allopolyploid network: insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology* 64: 84–101.

McDade, L. A. 1992. Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* 46: 1329–1346.

Meimberg H., and Heubl G. 2006. Introduction of a nuclear marker for phylogenetic analysis of Nepenthaceae. *Plant Biology* 8: 831–840.

Minh, B. Q., M. W. Hahn, and R. Lanfear. 2020. New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution* 37: 2727–2733.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37: 1530–1534.

Murphy, B., F. Forest, T. Barraclough, J. Rosindell, S. Bellot, R. Cowan, M. Golos, et al. 2020. A phylogenomic analysis of *Nepenthes* (Nepenthaceae). *Molecular Phylogenetics and Evolution* 144: 106668.

Nauheimer, L., L. Cui, C. Clarke, D. M. Crayn, G. Bourke, and K. Nargar. 2019. Genome skimming provides well resolved plastid and nuclear phylogenies, showing patterns of deep reticulate evolution in the tropical carnivorous plant genus *Nepenthes* (Caryophyllales). *Australian Systematic Botany* 32: 243–254.

Palfalvi, G., T. Hackl, N. Terhoeven, T. F. Shibata, T. Nishiyama, M. Ankenbrand, D. Becker, et al. 2020. Genomes of the venus flytrap and close relatives unveil the roots of plant carnivory. *Current Biology* 30: 2312–2320.

Sang, T., and D. Zhang. 1999. Reconstructing hybrid speciation using sequences of low copy nuclear genes: hybrid origins of five *Paeonia* species based on adh gene phylogenies. *Systematic Botany* 24: 148–163.

Soltis, D. E., E. V. Mavrodiev, J. J. Doyle, J. Rauscher, and P. S. Soltis. 2008. ITS and ETS sequence data and phylogeny reconstruction in allopolyploids and hybrids. *Systematic Botany* 33: 7–20.

Spooner, D. M., H. Ruess, S. Ellison, D. Senalik, and P. Simon. 2020. What is truth: Consensus and discordance in next-generation phylogenetic analyses of *Daucus*. *Journal of Systematics and Evolution* early access.

Wood, T. E., N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg. 2009. The frequency of polyploid speciation in vascular plants. *Proceedings of the National Academy of Sciences* 106: 13875–13879.

Figures & Tables

Figure 1. Scatterplot displaying the heterozygosity (proportion of loci with SNPs) and the allele divergence of all included accessions. Known or suspected hybrid taxa are labeled. Red dots show previously known hybrids.

Figure 2. A: Heatmap showing the heterozygosity, allele divergence (sorted descending) and clade association of all samples. Extracts from several parts of the table are shown enhanced. Phased accessions are shown in various colors. B, C: Phylogenies of supermatrix without phased (B) and with phased (C) accessions. Clades chosen for clade association are collapsed

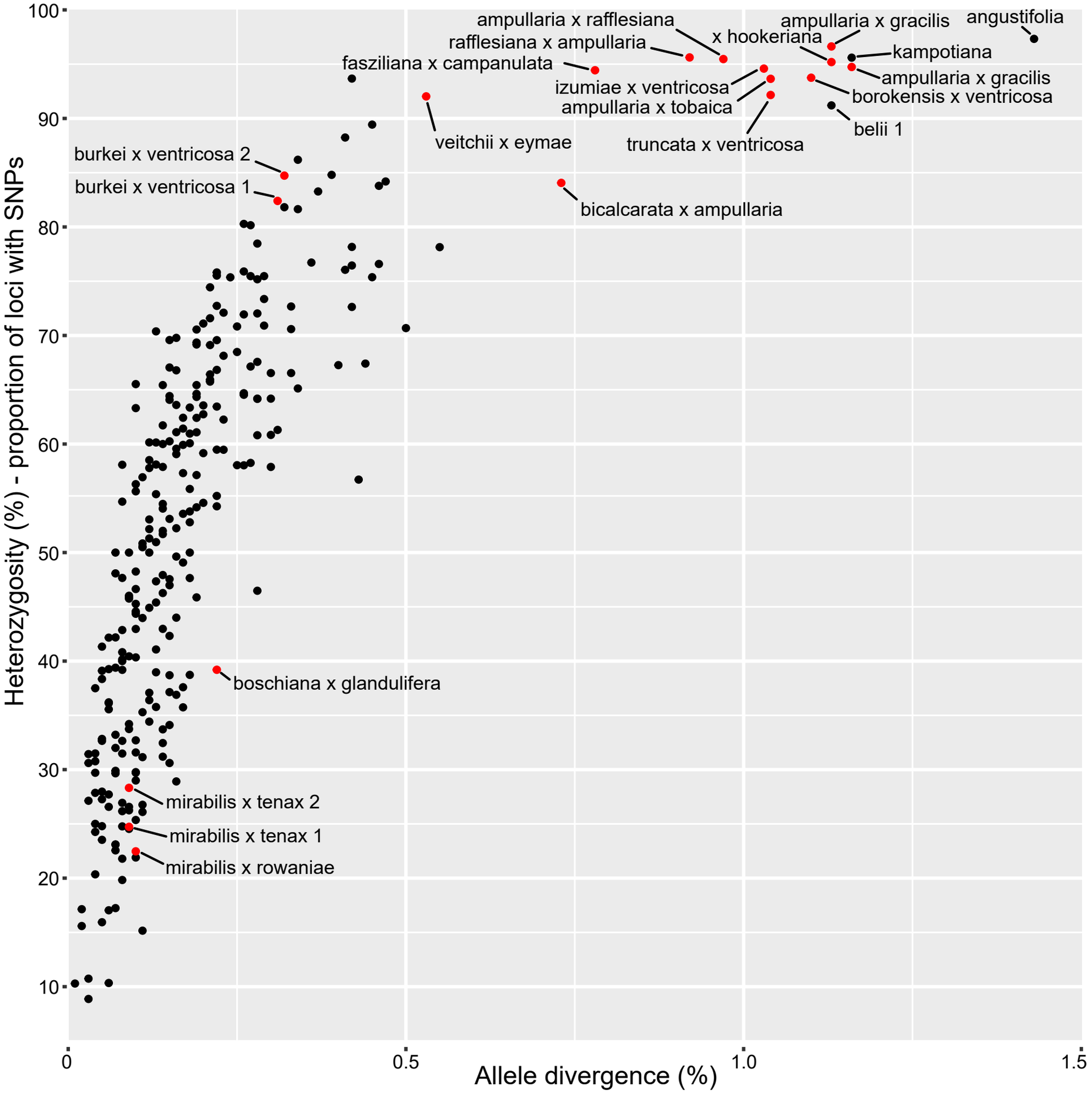
and phased accessions are shown in color. Circles on nodes show clades with moderate (blue: BS 95–99) and low (red: BS <95) bootstrap support. Nodes without circles have highest bootstrap support (BS=100). Node labels display gene and site concordance factors (gCF/sCF). Node labels in green show clades with improved CF support when phased accessions are included. Small triangle display clades for the fine scale clade association in the clades 'merr' and 'maxi'.

Table 1. Workflow overview containing all steps with short description, script/software used, required input and generated output for each step. Software marked with an asterisk are not part of the workflow.

Part	**Step**	**Description**	**Script / software**	**Input**	**Output**
1) Assessment of SNPs	Consensus sequence generation	Reads are mapped to the *de novo* assembled contigs to generate consensus sequences with SNPs where reads differ.	Bash 1 (BWA, bcftools)	de novo contigs (HybPiper), reads mapped to each locus (HybPiper)	Consensus sequences
	Consensus sequence assessment	Proportion of SNPs and length of consensus sequences for each locus is collected.	R1a	Consensus sequences	Tables with SNPs/locus and sequence length
	Dataset optimization	Missing data can be reduced and putative paralogs removed.	R1b	Tables with SNPs/locus and sequence length	Lists with samples and loci to be removed
	Assessment of heterozygosity and allele divergence	Summary tables are generated.	R1c	Tables with SNPs/locus and sequence length (cleaned).	Summary table and graphs for the assessment of heterozygosity and the detection of hybrids
	Sequence lists generation	Sequences from HybPiper HybPhaser	R1d	Contigs and consensus sequences	Sequence lists for loci or samples with

		folder are collated into sequence lists.		and list with samples/loci to be removed.	contigs or consensus seq., raw or cleaned (optimized)
2) Clade association	Phylogenetic analysis	Alignments and phylogenetic analysis	e.g., MAFFT*, IQ-TREE*	Sequence lists	Phylogenetic tree
	Selection of clade references	Taxa that represent major clades are selected by the user.	user*	Info from phylogeny and summary table	Table (csv) with names of clade references
	Extraction of mapped reads	Generation of read files that contain only reads that mapped on the target sequences.	Bash 2 (samtools)	Bam file from HybPiper	Read files (mapped-only)
	BBSplit script preparation / execution	Generate and run BBSplit script to match reads (mapped-only) to clade references	R2a	Table (csv) with names of clade references	BBSplit stats files with proportions of reads mapped to each reference.
	Collation of BBSplit results	Generation of summary table for clade association	R2b	BBSplit stats files, summary table	Clade association summary table
3) Phasing	Selection of accessions for phasing		user*	Clade association summary table	Table (csv) with names of accessions to phase with respective references
	BBSplit phasing script preparation and execution	Generate and run BBSplit script to map and phase read files	R3a	Table (csv) with names of accessions to phase with respective references, sequence read files	Read files of phased accessions, BBSplit stats files
	Collation of BBSplit phasing results	Generation of summary table for phasing	R3b	BBSplit stats files	Summary table for phasing stats
4) Dataset merging	Assembly of phased accessions	Phased accessions are assembled using HybPiper and HybPhaser (part 1)	HybPiper, HybPhaser	Read files of phased accessions, target sequence list	Sequence lists of phased accessions
	Merging of datasets	Sequences of phased	R4	Sequence lists of phased	Merged sequence lists

		accessions are merged with sequences of non-phased accessions		and non-phased accessions	of phased and non-phased accessions
	Phylogenetic analysis	Alignments and phylogenetic analysis	e.g., MAFFT*, IQ-TREE*	Merged sequence lists	Phylogenetic tree including phased and non-phased accessions


Heterozygosity (%) - proportion of loci with SNPs
Allele divergence (%)
100
90
80
70
60
50
40
30
20
10
0
0.5
1.0
1.5
ampullaria x rafflesiana
rafflesiana x ampullaria
fasziliana x campanulata
ampullaria x gracilis
x hookeriana
angustifolia
kampotiana
ampullaria x gracilis
izumiae x ventricosa
ampullaria x tobaica
borokensis x ventricosa
belii 1
veitchii x eymae
truncata x ventricosa
burkei x ventricosa 2
burkei x ventricosa 1
bicalcarata x ampullaria
boschiana x glandulifera
mirabilis x tenax 2
mirabilis x tenax 1
mirabilis x rowaniae

A

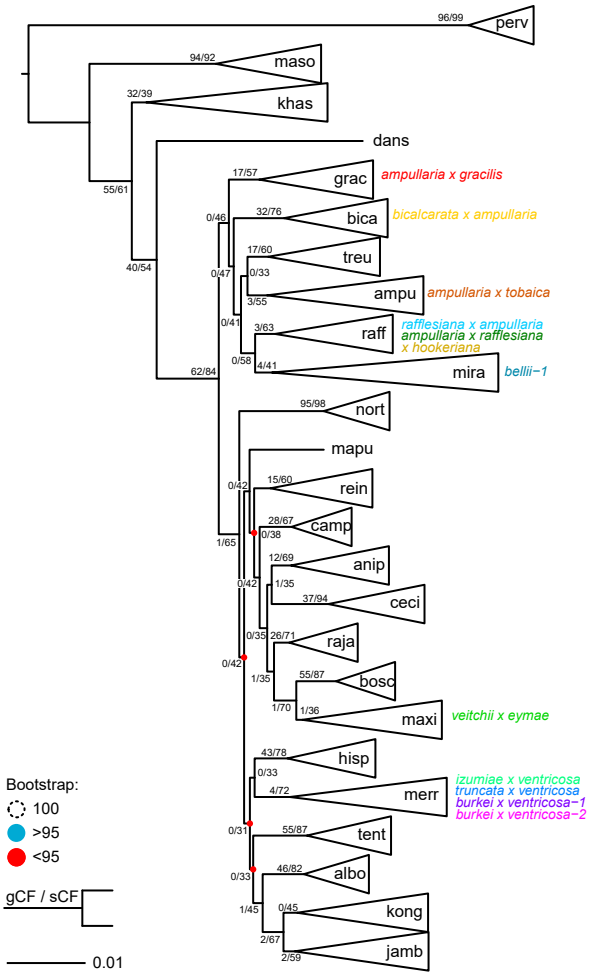
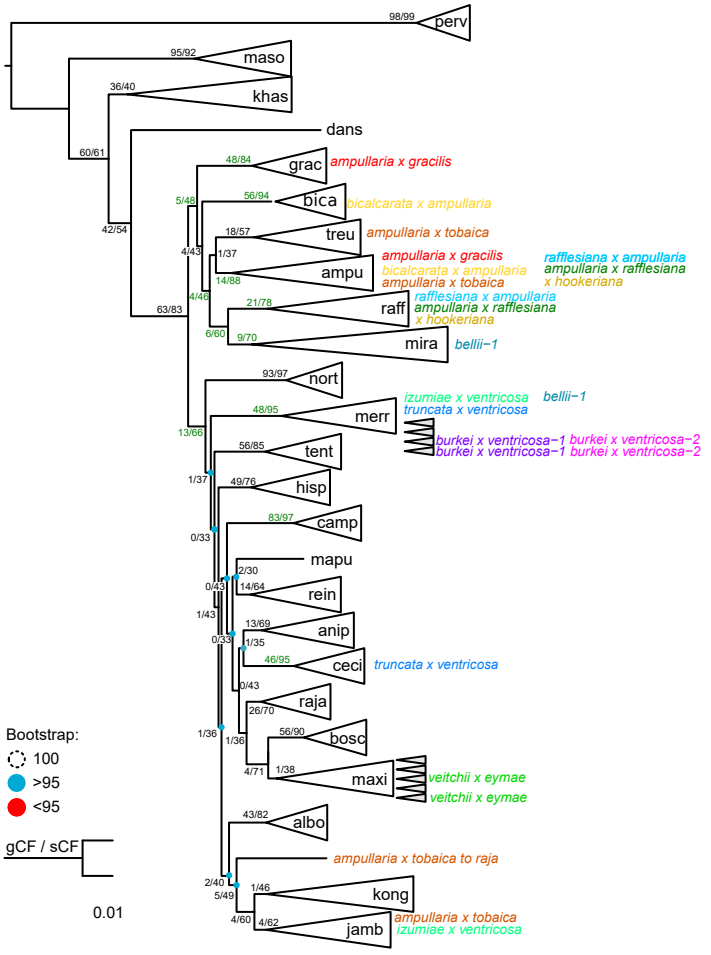
Accession	Heterozygosity	Allele divergence	perv	maso	khas	dans	grac	bica	ampu	treu	rafl	mira	nort	mapu	rein	camp	anip	ceci	raja	bosc	maxi	hisp	merr	tent	albo	kong	jamb
			Clade references for Nepenthes																								
angustifolia	97.3%	1.43%	0.1	0.3	0.4	0.3	11.9	0.4	0.1	0.4	0.1	0.1	0.5	0.3	0.2	0.3	0.3	0.1	0.6	0.1	0.1	1.2	0.3	0.3	0.1	0.3	0.2
x hookeriana	94.7%	1.16%	0.0	0.2	0.3	0.4	0.9	1.4	13.7	1.5	16.1	1.9	0.3	0.2	0.2	0.2	0.2	0.1	0.3	0.1	0.1	0.1	0.4	0.2	0.1	0.3	0.4
kampotiana 1	95.6%	1.16%	0.1	0.4	0.5	0.7	0.9	0.5	0.4	0.5	0.8	3.0	0.4	0.3	0.2	0.2	0.2	0.1	1.4	0.1	0.1	0.0	0.3	0.1	0.1	0.8	0.5
ampullaria x gracilis	96.6%	1.13%	0.2	0.4	0.5	0.7	7.8	0.6	4.2	0.5	0.3	0.2	0.3	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.1
bellii 1	91.2%	1.13%	0.1	0.3	0.3	0.4	0.3	0.3	0.2	0.4	0.8	5.6	0.2	0.2	0.1	0.1	0.1	0.1	0.5	0.1	0.1	0.0	9.6	0.2	0.1	0.1	0.2
clipeata x rafflesiana	95.2%	1.13%	0.1	0.2	0.3	0.4	0.3	0.1	0.1	0.2	0.1	0.0	0.3	0.3	0.3	0.2	0.4	0.1	1.1	0.9	1.6	0.2	8.3	0.2	0.3	0.1	0.2
bokorensis x ventricosa	93.8%	1.10%	0.1	0.4	0.5	0.8	0.9	0.4	0.3	0.5	0.2	0.1	0.6	0.6	0.3	0.3	0.2	0.1	1.3	0.2	0.1	0.1	3.7	0.1	0.1	0.8	0.5
ampullaria x tobaica	93.7%	1.04%	0.0	0.3	0.2	0.4	0.3	0.3	5.7	0.4	0.2	0.2	0.3	0.1	0.3	0.2	0.2	0.2	2.5	0.0	0.1	0.0	0.3	0.1	0.1	1.0	3.2
truncata x ventricosa	92.2%	1.04%	0.0	0.2	0.3	0.4	0.2	0.1	0.1	0.3	0.1	0.0	0.3	0.1	0.2	0.2	0.3	4.6	0.9	0.2	0.1	0.2	7.3	0.1	0.1	0.1	0.2
izumiae x ventricosa	94.6%	1.03%	0.1	0.2	0.3	0.5	0.3	0.1	0.1	0.3	0.1	0.1	0.3	0.2	0.2	0.1	0.2	0.1	0.5	0.2	0.1	0.1	7.7	0.2	0.1	0.8	4.8
ampullaria x rafflesiana	95.5%	0.97%	0.1	0.3	0.3	0.5	0.5	0.6	6.1	0.5	5.3	0.9	0.2	0.1	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.1	0.2	0.1
rafflesiana x ampullaria	95.6%	0.92%	0.2	0.4	0.5	0.5	1.1	0.7	3.0	0.6	2.6	0.5	0.3	0.2	0.1	0.1	0.1	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.1
faizaliana x campanulata	94.4%	0.78%	0.1	0.2	0.3	0.4	0.2	0.1	0.0	0.2	0.0	0.0	0.2	0.2	0.2	7.6	0.5	0.2	0.8	0.7	1.5	0.1	0.3	0.2	0.3	0.1	0.1
bicalcarata x ampullaria	84.1%	0.73%	0.1	0.4	0.3	0.4	0.4	9.2	5.8	0.4	0.2	0.1	0.2	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.1
ampullaria 1	78.1%	0.55%	0.1	0.4	0.4	0.5	0.5	0.6	12.5	0.6	0.4	0.3	0.3	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.2	0.1	0.0	0.1	0.2
veitchii x eymae	92.0%	0.53%	0.1	0.2	0.3	0.4	0.2	0.1	0.1	0.2	0.1	0.0	0.3	0.2	0.2	0.3	0.5	0.2	0.5	1.2	7.4	0.0	0.4	0.2	0.2	0.1	0.1
reinwardtiana 1	70.7%	0.50%	0.0	0.2	0.2	0.4	0.3	0.1	0.1	0.3	0.0	0.1	0.3	0.5	13.7	0.2	0.4	0.1	0.5	0.2	0.2	0.2	0.2	0.4	0.2	0.1	0.2
burkei x ventricosa 2	84.7%	0.32%	0.0	0.2	0.2	0.4	0.3	0.1	0.1	0.3	0.1	0.0	0.3	0.2	0.2	0.2	0.2	0.1	0.8	0.2	0.1	0.2	14.0	0.2	0.1	0.1	0.3
lingulata	81.8%	0.32%	0.0	0.2	0.2	0.3	0.3	0.3	0.1	0.4	0.1	0.1	0.4	0.5	0.5	0.4	0.5	0.2	1.8	0.4	0.5	0.2	0.7	0.4	0.4	2.9	22.0
ampullaria 7	61.3%	0.31%	0.0	0.1	0.3	0.3	0.5	1.0	18.3	0.9	0.6	0.4	0.3	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.0	0.0	0.3	0.1	0.0	0.3	0.1
burkei x ventricosa 1	82.4%	0.31%	0.1	0.2	0.2	0.4	0.2	0.1	0.1	0.3	0.1	0.0	0.3	0.2	0.1	0.2	0.2	0.1	0.8	0.3	0.1	0.1	14.0	0.2	0.1	0.1	0.3
gracilis 3	57.9%	0.30%	0.0	0.1	0.2	0.4	35.2	0.4	0.3	0.5	0.4	0.4	0.2	0.2	0.1	0.1	0.2	0.1	0.2	0.1	0.1	0.0	0.3	0.1	0.0	0.3	0.1
albomarginata 2	59.5%	0.22%	0.2	0.2	0.4	0.4	0.3	0.1	0.1	0.2	0.1	0.1	0.4	0.2	0.2	0.2	0.2	0.1	0.6	0.3	0.5	0.2	0.3	0.2	5.3	0.3	0.2
boschiana x glandulifera	39.2%	0.22%	0.2	0.1	0.3	0.1	0.2	0.0	0.1	0.2	0.1	0.0	0.2	0.1	0.3	0.4	0.3	0.1	1.6	0.9	1.6	0.1	1.8	0.3	0.2	0.2	0.5
burkei 2	69.6%	0.22%	0.0	0.1	0.2	0.4	0.5	0.3	0.1	0.4	0.2	0.1	0.6	0.4	0.3	0.4	0.4	0.3	1.5	0.4	0.3	0.3	29.8	0.4	0.2	0.4	0.7
kongkandana 2	55.2%	0.22%	0.1	0.1	0.4	1.1	0.2	0.3	0.2	0.9	0.2	0.1	0.3	0.3	0.5	0.3	0.3	0.1	1.1	0.1	0.2	0.2	0.7	0.3	0.2	30.9	1.1
mikei	72.7%	0.22%	0.0	0.1	0.2	0.3	0.5	0.4	0.2	0.4	0.3	0.2	0.7	0.6	1.0	0.7	0.6	0.3	2.1	0.3	0.3	0.2	1.0	0.5	0.7	4.3	9.4
mirabilis 19	30.6%	0.15%	0.0	0.2	0.2	0.4	0.7	0.7	2.5	1.1	2.5	16.9	0.3	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.0	0.4	0.2	0.1	0.4	0.1
mirabilis 21	34.1%	0.15%	0.1	0.2	0.3	0.4	0.8	1.0	2.2	1.3	2.7	20.7	0.3	0.1	0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.1	0.4	0.2	0.1	0.4	0.2
sibuyanensis 2	69.6%	0.15%	0.0	0.2	0.1	0.4	0.5	0.4	0.2	0.3	0.2	0.1	0.5	0.4	0.3	0.5	0.6	0.3	1.5	0.4	0.3	0.4	31.4	0.4	0.2	0.5	0.7
stenophylla 1	47.0%	0.15%	0.1	0.2	0.3	0.2	0.2	0.1	0.1	0.2	0.1	0.0	0.3	0.1	0.2	0.2	0.4	0.3	0.4	0.8	2.5	0.1	0.2	0.2	0.3	0.1	0.1
suratensis	42.3%	0.15%	0.0	0.1	0.2	0.6	0.7	0.3	0.2	0.4	0.3	1.0	0.5	0.7	1.1	0.7	0.5	0.3	1.6	0.2	0.2	0.1	0.6	0.2	0.7	5.3	3.2
maxima 7	50.0%	0.09%	0.0	0.1	0.3	0.2	0.1	0.2	0.1	0.3	0.0	0.1	0.3	0.2	0.4	0.4	0.5	0.3	0.4	1.2	20.1	0.0	0.6	0.2	0.1	0.2	0.1
mirabilis x tenax 1	24.7%	0.09%	0.2	0.4	0.6	0.8	1.4	0.8	0.9	0.9	1.4	8.4	0.2	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.1
mirabilis x tenax 2	28.3%	0.09%	0.1	0.3	0.4	0.4	0.6	0.3	0.6	0.7	1.2	8.8	0.2	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.2	0.0	0.1	0.1
mirabilis 5	24.5%	0.09%	0.2	0.4	0.5	0.8	1.2	0.8	1.1	0.8	1.2	7.3	0.2	0.2	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.1	0.1	0.1
northiana 2	40.4%	0.09%	0.0	0.2	0.2	0.3	0.1	0.2	0.2	0.2	0.1	0.1	47.4	0.3	0.1	0.1	0.1	0.1	0.3	0.1	0.1	0.0	0.7	0.2	0.1	0.2	0.2
khasiana 2	30.6%	0.03%	0.0	0.7	67.7	0.7	0.2	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.0	0.1	0.2	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.0	0.2	0.1
maxima 8	31.4%	0.03%	0.0	0.1	0.2	0.4	0.2	0.2	0.1	0.2	0.0	0.1	0.1	0.4	0.4	0.8	0.6	0.4	0.6	2.7	19.0	0.1	0.5	0.3	0.4	0.3	0.1
monticola	17.1%	0.02%	0.1	0.3	0.4	0.7	1.1	1.0	1.9	11.0	1.4	1.3	0.4	0.4	0.2	0.2	0.3	0.2	0.4	0.1	0.2	0.1	0.3	0.4	0.2	0.4	0.4
xiphioides 2	15.6%	0.02%	0.1	0.1	0.2	0.4	0.3	0.3	0.2	0.5	0.1	0.3	0.4	0.4	0.6	0.8	0.4	0.2	2.3	0.1	0.4	0.3	0.4	0.6	0.5	4.4	13.5
oblanceolata	10.3%	0.01%	0.0	0.1	0.3	0.4	0.1	0.2	0.1	0.2	0.0	0.1	0.1	0.2	0.3	0.5	0.4	0.4	0.9	2.1	17.5	0.2	0.5	0.3	0.2	0.2	0.1

Heterozygosity										
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Allele divergence										
0%	0.15%	0.3%	0.45%	0.6%	0.75%	0.9%	1.05%	1.2%	1.35%	1.5%

Percentage of reads mapped to reference									
0%	0.5%	1%	1.5%	2%	2.5%	3%	3.5%	4%	72%

B



C