Molecular Phylogenetics and Evolution 186 (2023) 107826

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

REFMAKER: Make your own reference to target nuclear loci in low coverage genome skimming libraries. Phylogenomic application in Sapotaceae

Charles Pouchon [a,b,*], Carlos G. Boluda [a,b]

[a] Conservatoire et Jardin botaniques de la Ville de Genève, Chemin de l'Impératrice 1, 1292 Chambésy, Geneva, Switzerland
[b] PhyloLab, Department of Plant Sciences, Université de Genève, Chemin de l'Impératrice 1, 1292 Chambésy, Geneva, Switzerland

ARTICLE INFO

Keywords:
Bioinformatics
Nuclear phylogeny
Sapotaceae
Shotgun sequencing
Systematics

ABSTRACT

Genome skimming approach is widely used in plant systematics to infer phylogenies mostly from organelle genomes. However, organelles represent only 10 % of the produced libraries, and the low coverage associated with these libraries (<3X) prevents the capture of nuclear sequences, which are not always available in non-model organisms or limited to the ribosomal regions. We developed REFMAKER, a user-friendly pipeline, to create specific sets of nuclear loci that can be extracted directly from the genome skimming libraries. For this, a catalogue is built from the *meta*-assembly of each library contigs, and cleaned by selecting the nuclear regions and removing duplicates from clustering steps. Libraries are next mapped onto this catalogue and consensus sequences are generated to produce a ready-to-use phylogenetic matrix following different filtering parameters aiming at removing putative errors and paralogous sequences. REFMAKER allowed us to infer a well resolved phylogeny in *Capurodendron* (Sapotaceae) on 67 nuclear loci from low-coverage libraries (<1X). The resulting phylogeny is concomitant with one previously inferred on 638 nuclear genes from target enrichment libraries. While it remains preliminary because of this low sequencing depth, REFMAKER therefore opens perspectives in phylogenomics by allowing nuclear phylogeny reconstructions with genome skimming datasets.

1. Introduction

Genome skimming is a powerful approach to easily and rapidly collect phylogenetically informative markers in non-model organisms (McKain et al., 2018). This methods consists on the sequencing of the whole genomic DNA at low coverage (*i.e.* ~1−3x coverage of the nuclear genome) which is enough to provide sequences for genomic regions in high copy number in cells, such as the chloroplast (cpDNA), the mitochondria (mtDNA) or the nuclear ribosomal sequences (nrDNA), which are highly covered (~>30x; Malé et al., 2014; Straub et al., 2012). The total DNA is sequenced directly through random shearing without additional effort (McKain et al., 2018; Straub et al., 2012). This makes genome skimming an attractive, scalable and cost-effective approach in molecular systematics, applicable to both well preserved or degraded DNA (Alsos et al., 2020; Bakker et al., 2016; Grandjean et al., 2017; Trevisan et al., 2019), as it is expected for museum collections.

Over the last decade, genome skimming has been widely used in plants to infer phylogenies mostly from plastid markers (*e.g.* Givnish et al., 2018; Pouchon et al., 2022a; Thomson et al., 2018). However, this can be limiting to unravel the evolution of some plant taxa at the genus or the species level as both organelle genomes, *i.e.* cpDNA and mtDNA, are usually maternally inherited in Angiosperms, and, by essence, provide only a single evolutionary history (Gitzendanner et al., 2018). Additionally, the phylogenetic trees inferred on these markers can be conflicting with the nuclear species tree, mostly in presence of high levels of introgression and past hybridization where "foreign" organelle genome can be fixed through genome capture (Morales-Briones et al., 2018; Pouchon et al., 2018; Vargas et al., 2017) and organelle recombinations (Gandini and Sanchez-Puerta, 2017; Wang et al., 2018). On the other hand, the shallow coverage of the nuclear genome in genome skimming libraries limits the inference of nuclear trees to the nrDNA regions (Hollingsworth et al., 2016; McKain et al., 2018). The latter regions tend to be useful to infer relationships among closely related species but remain inefficient at deep phylogenetic levels as they evolve quickly, which could lead to high levels of homoplasy (Hughes

* Corresponding author.
E-mail addresses: charles.pouchon@unige.ch, charles.pouchon@ville-ge.ch (C. Pouchon).

https://doi.org/10.1016/j.ympev.2023.107826
Received 18 January 2023; Received in revised form 24 April 2023; Accepted 25 May 2023
Available online 29 May 2023

et al., 2006; Patwardhan et al., 2014).

Other genomic methods can be used to bring alternative evidence on such phylogenetic outcomes (Bohmann et al., 2020; McKain et al., 2018; Yu et al., 2018), as target enrichment (Boluda et al., 2022; Randriarisoa et al., 2022) or RAD-sequencing (Eaton and Ree, 2013; Vargas et al., 2017). Nevertheless, capturing nuclear loci within genome skimming libraries remains possible with higher sequencing depths (~>3–10x coverage; see Berger et al., 2017; Liu et al., 2021; Zhang et al., 2019), but often requires reference genomes or transcriptomes, which are not always available in non-model species, to target homologous regions (*e.g.* Vargas et al., 2019). In this context, assembly and alignment-free based methods have been developed to efficiently and quickly produce distance matrices from the kmer spectra in genome skimming libraries, which can be used to infer phylogenetic relationships (*e.g.* MASH, Ondov et al., 2016; SKMER, Sarmashghi et al., 2019). However, these methods are restricted to only few evolutionary models to correct distance estimates (*e.g.* Jukes-Cantor model), and can not be used for species tree inferences under coalescent model, molecular dating or species delimitation modeling as alignment-based approaches (*e.g.* Boluda et al. 2022). Besides, organelle sequences only constitute ~4–10 % of the genome skimming libraries regardless of the sequencing effort for the nuclear genome, therefore bringing lots of waste data (Steele et al., 2012; Straub et al., 2012). This reveals the need for alternative sequence-based methods to exploit the whole libraries even at low coverage by targeting any nuclear loci (*i.e.* within and outside genes) and without any specific references (*e.g.* Pouchon et al., 2018).

This is why we developed REFMAKER, a user-friendly pipeline, which allows creating a set of nuclear reference loci directly from the assemblies of genome skimming libraries that can be next targeted from the sequencing reads of the same libraries. We tested its application for the systematics of *Capurodendron* (Sapotaceae), the second largest endemic genus of plants from Madagascar. The recent development of a bait kit for the target sequencing of 792 protein coding genes in Sapotaceae (Christe et al., 2021) has allowed species delimitation in *Capurodendron* with a well supported phylogeny (Boluda et al., 2022). We compared this phylogeny with the one obtained with REFMAKER from low coverage genome skimming libraries of some *Capurodendron* samples.

2. Material and methods

2.1. Pipeline overview

REFMAKER is an open source pipeline, written in bash and python languages, that needs to be run in command lines into UNIX environments. It is packaged within a conda environment with all required dependencies, released under a GPL-3 license, and available at https://github.com/cpouchon/REFMAKER.

REFMAKER is called by different modes, with an input parameter file (Fig. 1). These modes can be parametrized in order to: perform the assembly of each genome skimming library, perform the *meta*-assembly of these libraries to create a catalogue, clean the catalogue by selecting the nuclear regions and by removing duplicates, map the raw reads of each library into the cleaned catalogue, call the variants, get the consensus sequences for each library, and filter these sequences to create a phylogenetic matrix across the libraries.

The first step consists in a global assembly of each genome skimming library into a set of contigs. This is done with SPADES (Bankevich et al.,

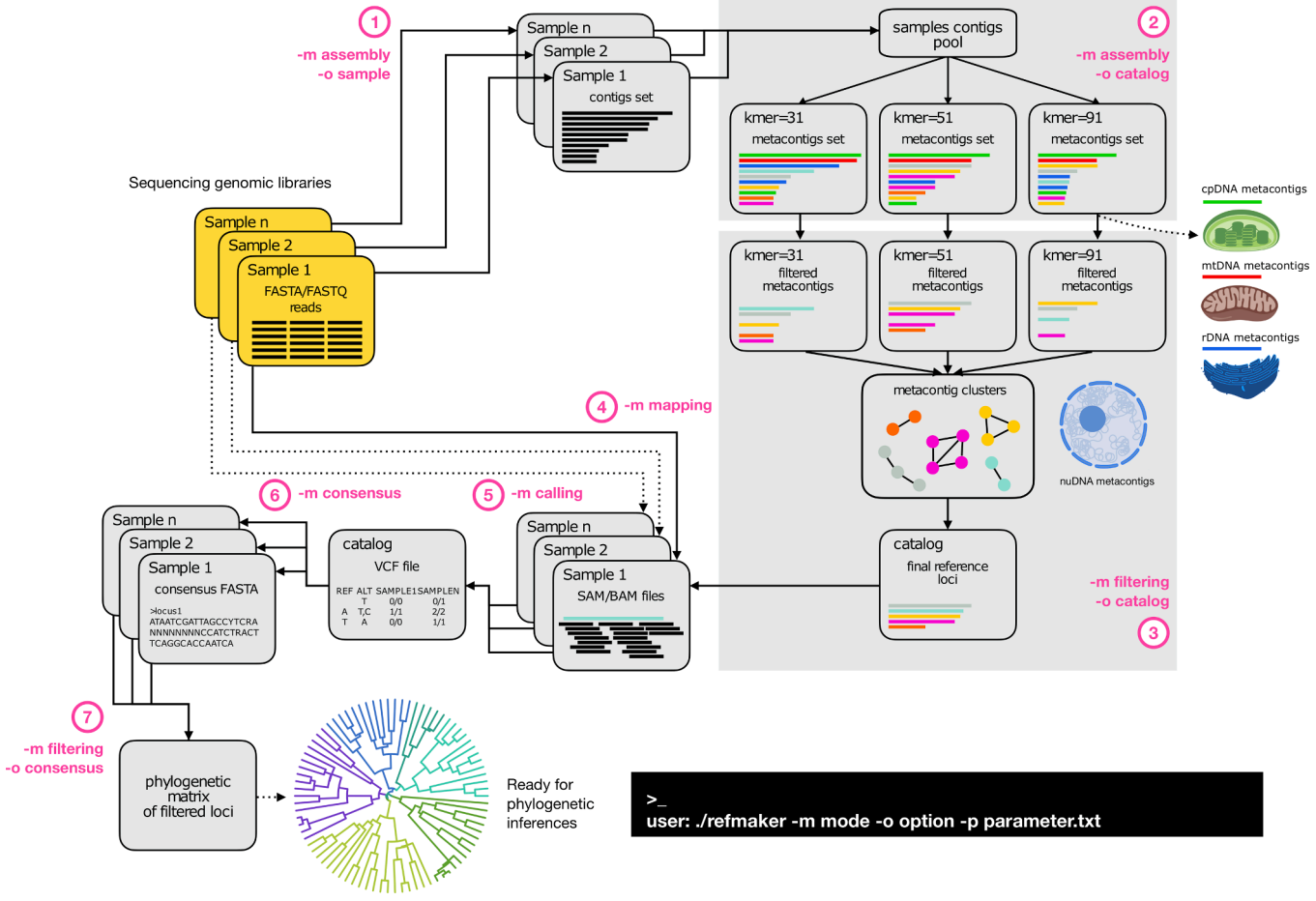

Fig. 1. REFMAKER workflow. The different steps are given as coloured numbers with their respective command lines.

2012), by using the *'assembly'* mode with the *'sample'* option (Fig. 1). The assembly is run with the read-correction mode and –cov-cutoff auto option of SPADES, along with predefined kmer size(s) and a maximal memory limit and thread numbers, given in the input parameter file.

The second step consists in the assembly of these contigs sets into *meta*-contigs by using the *'assembly'* mode with the *'catalog'* option (Fig. 1). We recommend to keep a set of different kmer sizes to build these *meta*-contigs in order to enhance the final number of loci into the catalogue, as these metacontigs are next merged and cleaned. The cleaning of this catalogue is achieved with the *'filtering'* mode and the *'catalog'* option (Fig. 1). This step allows selecting the putative nuclear loci and representative sequences within the catalogue by removing redundant and/or heterozygous loci. One representative sequence is selected by locus for each *meta*-contigs set performed on different kmer sizes by using CDHIT (Fu et al., 2012) with sequence similarity threshold at which two *meta*-contigs are identified as being homologous. Each set of reduced *meta*-contigs are next mapped with BLAST (Camacho et al., 2009) into the DBFAM database of ORTHOSKIM (Pouchon et al., 2022b), and the RFAM (Burge et al., 2013) and SILVA (Ludwig et al., 2004) RNA databases, to remove unwanted loci (*i.e.* cpDNA, mtDNA and rDNA loci). Users can also use their own database to remove additional sequences (*e.g.* repetitive elements). The reduced sets of putative nuclear loci of each *meta*-assembly are next merged and clustered together using BLAST. A connected graph is built between loci according to their sequence similarity and their overlapping positions using a MCL algorithm (Van Dongen, 2008) with optimised cluster inflation values. This clustering similarity threshold can be adjusted according to the level of phylogenetic divergence of the sampling. For each cluster, a representative sequence is then selected according to the length of these loci, weighted by their coverage, and added to the final catalogue of reference.

Raw reads are next mapped for each library onto the cleaned catalogue with BWA (Li and Durbin, 2009), by using the '*mapping*' mode (step 4). Duplicate reads are removed using the MarkDuplicates function of PICARD tools (https://broadinstitute.github.io/picard/), as well for low quality reads. A variant calling is next done with BCFTOOLS (Danecek et al., 2021) by using the *'calling'* mode (step 5). Variant positions are filtered for each library according to the read depth. Only Single Nucleotide Polymorphisms (SNPs) are kept.

Consensus sequences are next produced for each locus, using the IUPAC code with the*'consensus'* mode, and trimmed thanks to TRIMAL (Capella-Gutiérrez et al., 2009). In order to remove errors and/or paralogous sequences for the inference of the final phylogenetic matrix, these sequences are cleaned by using the *'filtering'* mode with the*'consensus'* option. We first identify outlier loci from the read coverage of each library and remove them according to the frequency of libraries sharing these outlier loci. Consensus sequences are next cleaned according to heterozygosity as proposed in different pipelines, such as in PYRAD (Eaton, 2014) or PPD (Zhou et al., 2021). A maximal number of heterozygous sites is set by locus and by library. This helps to remove low quality alignments with an excess of heterozygous sites. Loci are also filtered according to the frequency of libraries sharing an heterozygote site, which may represent a cluster of paralogous sequences and erroneous polymorphisms. We used a sliding window approach, as implemented in ORTHOSKIM (Pouchon et al., 2022b) or PPD (Zhou et al., 2021), to remove hypervariable regions, which can be produced by errors or paralogous loci evolving differently. For this, the sequences are removed by locus and by library according to a maximal number of polymorphic sites allowed within the window size. Finally, the sequences and the loci are also cleaned according to the missing data and the sharing between and within populations. This cleaning step results in the production of a concatenated alignment file with a partition file in RAxML-style format, which can be both used directly for phylogenetic inference.

2.2. Pipeline illustration

2.2.1. Dataset

Twenty four samples were selected, 11 from fresh material stored in silica gel for <5 years and 13 corresponding to 10 to 70 years-old herbarium samples. DNA was extracted using the CTAB method (Russell et al., 2010; Souza et al., 2012), and quantified with a Qubit® Fluorimeter version 3.0 (Invitrogen, Thermo Fisher Scientific, Waltham, MA, U.S.A.). Fragment sizes, estimated with a 2200 TapeStation (Agilent, Santa Clara, CA, U.S.A.), were around 20–100 bp for herbarium samples and above 700 bp for silica gel samples, which were therefore fragmented to 500 bp on average with a Bioruptor® sonicator (Diagenode, Seraing, Belgium). Library construction was performed with the KAPA HyperPrep Kit (Roche, Basel, Switzerland), following the protocol of Vanburen et al. (2018). The washing steps were done with Sera-Mag™ Speed Beads Carboxylate-Modified Magnetic Particles (GE Healthcare, Little Chalfont, Buckinghamshire, U.K.) in a PEG/NaCl buffer. For the herbarium samples washing PEG ratios were increased until 2.4X, to retain fragments as small as 75 bp. Libraries were quantified, pooled and sequenced on a HiSeq4000 Illumina machine (100 bp paired-end reads).

The genomic coverage of these samples were estimated from the Lander/Waterman equation with the two available Sapotoideae genome sizes available on NCBI (~670 Mb, GCA_003260245.2, GCA_019916065).

2.2.2. REFMAKER running

REFMAKER was executed on the High Performance Computing (HPC) nodes of the University of Geneva (https://www.unige.ch/eresearch/fr/services/hpc/), with 16 threads.

Contigs were assembled for each library using a kmer of 55 and a maximal memory of 48 Gb. We computed the catalogue assembly with four different kmer sizes (*i.e.* 31, 51, 71 and 91). This catalogue was next filtered with a clustering similarity of 0.80 for both CDHIT and BLAST steps, a minimal locus size of 250 bp and a minimal proportion of overlapping regions between *meta*-contigs of 0.25.

Raw reads were mapped onto this catalogue, by keeping reads with a minimal mapping quality of 60. For each library, a minimal depth of three reads by SNP was set during the variant calling. Consensus were next produced and filtered with a maximal frequency of heterozygous site (h) of 0.05 by sequence, a frequency of samples sharing heterozygous sites (H) of 1.0, a maximal frequency of missing data allowed by sequence (m) of 0.40, a maximal frequency of missing data allowed by sample of 0.85 across all the loci, a window size of 20 nucleotides, a maximal of 5 polymorphic site allowed within this window, a minimal locus length (l) of 200 bp and minimal frequency of samples within populations sharing a locus (r) of 0.25. All samples were assigned to the same population. The impact of some parameters on the produced matrix, *i.e.* h, H, r, l and m, was also assessed.

2.2.3. Phylogenetic inferences

We used both coalescent and supermatrix approaches to reconstruct the phylogeny directly from the concatenated and the partition files. For the supermatrix method, IQTREE-2 v.2.1.3 (Minh et al., 2020) was used to estimate a ML concatenated tree. The best-fit model was determined for each partition (*i.e.* locus). We used 1,000 ultrafast bootstrap (UFBoot) replicates, along with the hill-climbing nearest neighbour interchange search option. For the coalescent approach, the locus trees were estimated in IQTREE-2, and used in ASTRAL-III v.5.7.2 (Zhang et al., 2018) in order to estimate a coalescent-based species tree.

In order to examine the sharing of loci across the trees, we generated a heat map from the number of shared loci per pair of samples with the presence/absence of each sample at each locus. The heat map was generated in R and ordered by the ML concatenated tree. The mean number of shared loci was computed for each node of the tree.

In comparison with REFMAKER, we also ran MASH v.2.3 and SKMER v.3.3.0 to produce distance matrices from the libraries using k = 31 (*i.e.*

the maximum kmer length allowed by MASH and the default size set in SKMER). Estimated distances from SKMER are transformed using the Jukes-Cantor model of substitution. Phylogenetic trees were inferred from both matrices with FASTME v.2.1.6.1 (Lefort et al., 2015). Support values were generated in SKMER using a subsampling procedure with a correction for the subsampled distance matrices obtained (Rachtman et al., 2022).

3. Results

An average of 3,112,706 reads were sequenced across samples, leading to a mean expected genome coverage of 0.93 for all the libraries according to the Lander/Waterman equation. On average, we generated 7,437 contigs by library and 2,124 *meta*-contigs by kmer size. After the first clustering step, 1,595 *meta*-contigs above 250 bp were merged, with 1,016 of them identified as putative nuclear loci. After the secondary clustering step, 309 representative loci were kept in the catalogue. Consensus sequences were produced for 303 loci after the alignment step. Among them, we removed three loci tagged as outliers according to the depth, 42 according to the minimal length, 118 according to the heterozygosity and 52 according to the population sharing.

A phylogenetic matrix of 89 loci was generated with 67 informative loci and 145,595 bp, including 1,390 informative sites and 36.93 % of missing sites on average (Fig. 2A). The median locus size was 787 bp. These loci contained a median of six informative sites (15.61 on average), and an average of 0.79 % heterozygous sites. The missing data by locus varied between 6 % and 82 % with an average of 42 %.

The resulting ML tree was well resolved with 86 % of the nodes having a UFBoot $\geq$ 90 (Fig. 2A). Both Sapoteae and Tseboneae were retrieved as monophyletic, as for all genera, with *Labourdonnaisia* sister to *Mimusops*, and *Bemangidia* sister to *Capurodendron*, respectively (UFBoot = 100). All species were monophyletic. Within *Capurodendron*, *C. madagascariense* was the first lineage to diverge (UFBoot = 100). We next recovered *C. ankaranense* as sister to the remaining *Capurondendron* lineages (UFBoot = 100), which formed two clades (UFboot = 100). The first one comprised *C. ludiifolium* + *C. apollonioides* (UFBoot = 95), and the second one *C. costatum* and *C. randrianaivoi* + *C. suarezense*, as sister taxa (UFboot = 90), but with a lower support (UFBoot = 83). The same topology was obtained for the ASTRAL tree (Fig. 2A). However, most of the nodes were not supported, in particular within the main *Capurodendron* clade (*C. ludiifolium* + *C. apollonioides* + *C. costatum* + *C. randrianaivoi* + *C. suarezense*) (Fig. 2A). Concerning alignment-free methods, SKMER produced a somewhat similar topology with the difference that *C. randrianaivoi* and *C. costatum* appeared as sister species (PP = 0.89; Fig S1). In contrast, MASH failed to recover *Capurodendron* as monophyletic, placing *C. madagascariense* sister to *Bemangidia* (Fig S1). In addition, *C. suarezense* was no longer monophyletic in the MASH-based tree (Fig S1).

The average proportion of shared loci was 43.5 % among all samples (Fig S2). Overall, closely related taxa shared slightly more loci than other taxa with 54.7 % of sharing between sister taxa and 43.1 % between non-sisters on average (Kruskal–Wallis p = 4.4 × 10−3, Fig S2). Moreover, more loci were shared within the *Capurodendron* lineages. Only the two *Bemangidia* samples shared less loci with others, which was consistent with their higher proportion of missing data within the phylogenetic matrix.

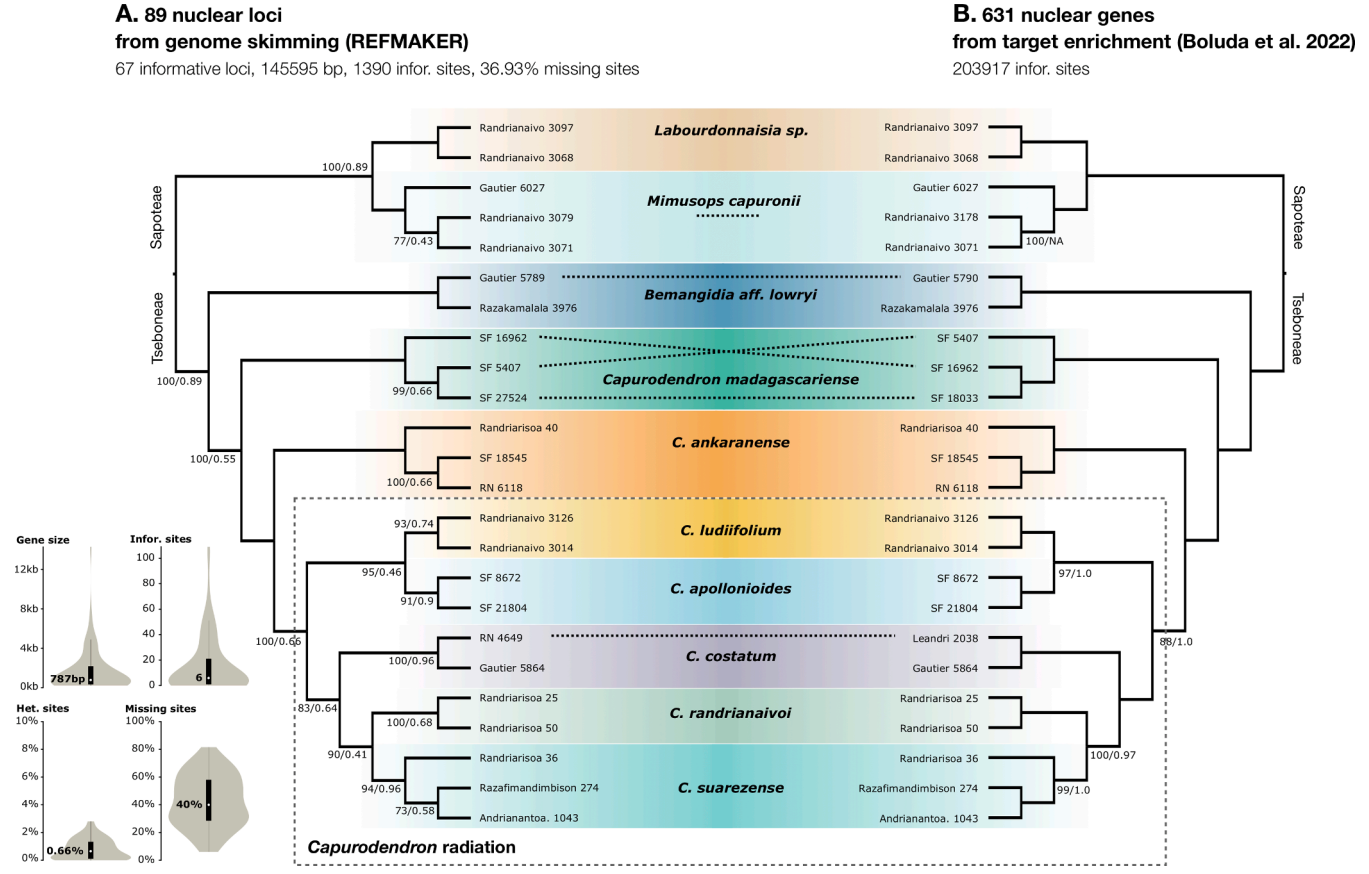

Fig. 2. Phylogenetic relationships of *Capurodendron* lineages inferred from the alignments of (A) 89 nuclear loci produced with REFMAKER on genome skimming libraries and (B) 638 nuclear genes captured from target enrichment libraries (Boluda et al. 2022). The use of different specimens or conflicting relationships between A and B are indicated with dotted lines. Node support (ML-UFBoot/ASTRAL-PP) values are given when not fully supported (100 %/1.0). Bottom panels show the distribution of the size, the informative and heterozygous, and the missing data proportion across the loci.

4. Discussion

4.1. Utility of REFMAKER in plant phylogenomic

Our study supports the ability of REFMAKER to find nuclear regions in genome skimming libraries and to use them to infer phylogenies.

The phylogenetic relationships estimated here on 67 nuclear loci are well resolved according to the concatenated tree, and interestingly, similar to those obtained by Boluda et al. (2022) on 631 loci from target enriched capture libraries, with 150 times more informative sites approximately (Fig. 2B). These relationships are also consistent with the habitats and geographic distribution of these species (Boluda et al., 2022). For instance, the clade comprising *Capurodendron ludiifolium* + *C. apollonioides* is composed of species distributed on the east coast of Madagascar in moist evergreen forests while the clade of *C. costatum* + *C. randrianaivoi* + *C. suarezense* is composed of species from the north and west dry deciduous forests. The only topological difference concerns the positioning of SF-16962 and SF-5407 in *C. madagascariense*. However, the specimen SF-27524 was not present in Boluda et al. (2022), leaving uncertainty about this positioning. Additionally, our phylogenetic estimates are also concordant to those obtained with SKMER from the kmer spectra distances, which strengthen our approach. The only discordance was about the *C. randrianaivoi* + *C. costatum* relationships on this distance-based tree, which could be explained by a very short branch supporting this sister relationship. The positioning of SF-16962 and SF-5407 in *C. madagascariense* was also concomitant to the one we obtained. In contrast, MASH failed to resolve the phylogeny of this closely related species.

REFMAKER also recovered both deep and shallow resolutions of the phylogeny of the studied species. According to Boluda et al. (2022), Tseboneae diverged from Sapoteae at ca. 52.5 Ma, *Bemangidia* from *Capurodendron* at ca. 44.5 Ma and *C. madagascariense* from the remaining *Capurodendron* lineages at ca. 40.2 Ma. Despite these divergence times, the proportion of shared loci is relatively high between the different species, the genera and the two tribes, although closely related taxa share slightly more loci. This can be expected since the catalogue of loci is built from the *meta*-assemblies of contigs from different taxa. The closer the taxa are, the more similar the contig sequences are and the more complete the *meta*-assemblies should be. The higher proportion of shared loci found between *Capurodendron* lineages could thus be explained by a sampling focused on this genus.

Some studies have previously reported the possibility of recovering nuclear regions in skimming genome libraries, but using a set of target loci as references (Vargas et al., 2019) and/or high covered libraries (Berger et al., 2017; Liu et al., 2021; Zhang et al., 2019). Here, we were able to identify nuclear loci on shotgun data with very low coverage (<1X), without the use of external nuclear references. This approach also works well on herbarium samples, regardless of the age of the samples (Fig. S3). As for other alignment-based approaches, REFMAKER allows using more complex substitution models for phylogenetic reconstructions, as well as species tree inference or species delimitation modeling in comparison to kmer-based approaches. This contrasts to SKMER, while it needs longer computational effort. Another advantage is that REFMAKER can be used in the analysis of genome skimming libraries complementarity to ORTHOSKIM (Pouchon et al., 2022b), designed for the capture of targeted sequences. Both approaches use the same set of contigs per sample at the beginning of their workflow. This step, which is the most time-consuming (Pouchon et al., 2022b), can thus be performed only once for the same libraries. ORTHOSKIM can be used to search for targeted sequences, such as chloroplastic genes or ribosomal regions, and REFMAKER to target nuclear genome regions when no reference is available or when the genomic coverage of the libraries is too low.

4.2. Limitations and perspectives

The main issue of our approach concerns the final number of loci found in the catalogue, which depends on the sequencing effort, the filtering parameters and the taxonomic diversity of the sampling. Many filters are performed to remove alignment errors and paralogs before/after the catalogue building, which can strongly reduce the final number of loci. This was shown for most of the filtering parameters we tested (Fig S4). In our main dataset, 39 % of the loci were removed according to heterozygosity for example. This reduced number of loci in the catalogue may impact phylogenetic estimates. This is illustrated with a weak node support in the ASTRAL tree, which is inferred from locus trees requiring a sufficient number of informative sites to be resolved. We recovered few loci for a relatively low median number of informative sites, which could explain this low support. This is particularly important within the *Capurodendron* radiation where many clades emerged in a time lapse estimated to be around 5 million years (Boluda et al., 2022), leading to short speciation times and internodes. The phylogenetic signal displayed by each locus is probably not sufficient to resolve these internodes, reflecting shared ancestry and high incomplete lineage sorting (Kong et al., 2021). In the meantime, one can expect a high level of incomplete lineage sorting on such a part of the tree, where ancestral allele copies can be maintained in diverging species through the coalescence with short speciation times (Pinho and Hey, 2010; Townsend et al., 2012), and thus supporting alternative locus history (Degnan and Rosenberg, 2009; Rosenberg and Tao, 2008). This was highlighted in Boluda et al. (2022), with frequently unsupported or incongruent gene trees for these lineages. The same pattern was shown when varying some filtering parameters as reducing the maximal frequency of samples allowed sharing heterozygous sites or increasing the minimal frequency of samples allowed within populations sharing a locus, for example (Fig S4). In these cases, fewer loci and consequently fewer informative sites were kept. Topological discordances started emerging below 80 remaining loci on the catalogue (Fig S4). However, we recommend keeping stringent filters to avoid incorporating noise in the recovered phylogenetic signal.

Another issue concerns the phylogenetic range of our approach. Although we found a correct proportion of shared loci between Tseboneae and Sapoteae, one can expect a drop in the phylogenetic signal and the number of shared loci at a deeper phylogenetic scale as a consequence of the catalogue construction method and the sequence divergence between taxa. For this, the use of universal capture kits, such as the Angiosperms353 kit (Johnson et al., 2018), is still recommendable. However, an increase in the sequencing effort of the libraries would allow adding more loci into the catalogue, and/or a better coverage between taxa.

Several improvement points can also be raised. First, other assemblers could be added to increase the probability of finding nuclear loci such as MEGAHIT (Li et al., 2015) or SOAPDENOVO (Xie et al., 2014), which produce more fragmented contigs (Berger et al., 2017). Secondly, a filtering step could be added before the assemblies to remove bad quality reads and contaminants directly on the libraries, *e.g.* using FASTP (Chen et al., 2018) and KRAKEN-2 (Wood et al., 2019), respectively. Both approaches can be used before running REFMAKER until further development. Moreover, some loci in the final catalogue are uninformative and could be filtered according to a minimum number of informative sites. Finally, the same filtering parameters used for the consensus sequences could be applied in a forthcoming version directly on the variant catalogue (*i.e.* VFC file) for population genetics applications, as proposed in other programs (*e.g.* PYRAD; Eaton, 2014).

5. Conclusions

This study opens interesting perspectives for phylogenomics and systematics by showing the possibility of collecting nuclear loci in very low coverage shotgun libraries, on herbarium samples, and without any

external reference. This pipeline can be easily used with ORTHOSKIM in order to fully exploit the produced libraries. However, the phylogenies inferred on these loci should remain preliminary because of the low number of loci and/or phylogenetic signals which can be produced due to the low sequencing effort. We thus recommend using SKMER together with REFMAKER to have a better understanding about the phylogenetic relationships among the libraries. Besides, nuclear gene enrichment techniques seem to be more appropriate for inferring robust and larger scale phylogenies.

CRediT authorship contribution statement

Charles Pouchon: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Carlos G. Boluda:** Investigation, Resources, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code is available on Github at https://github.com/cpouchon/REFMAKER

Acknowledgements

We thank the Malagasy Government and the Malagasy people on site, for their help in collecting Sapotaceae specimens, Yamama Naciri (YN) and Laurent Gautier (LG) for discussions and advice, and the University of Geneva High Performance Computing Cluster for computer resources. We are grateful to the curators of the herbaria MO, P, TAN and TEF for allowing us to study their specimens and to perform limited destructive sampling. This work was supported by a grant No. 31003A-166349, 2016–2019 from the Swiss National Foundation to YN and LG, a grant from the Schmidheiny Foundation attributed in 2018 to YN, and the grant No. 2019-20 from the Franklinia Foundation to LG which allowed to hire Carlos Boluda (2017-2023).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2023.107826.

References

Alsos, I.G., Lavergne, S., Merkel, M.K.F., Boleda, M., Lammers, Y., Alberti, A., Pouchon, C., Denoeud, F., Pitelkova, I., Puşcaş, M., Roquet, C., Hurdu, B.-I., Thuiller, W., Zimmermann, N.E., Hollingsworth, P.M., Coissac, E., 2020. The treasure vault can be opened: large-scale genome skimming works well using herbarium and silica gel dried material. Plants 9, 432.

Bakker, F.T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., van de Kerke, S., Gravendeel, B., Nieuwenhuis, M., Staats, M., Alquezar-Planas, D.E., Holmer, R., 2016. Herbarium genomics: plastome sequence assembly from a range of herbarium specimens using an iterative organelle genome assembly pipeline. Biol. J. Linn. Soc. 117, 33–43.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V. M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Berger, B.A., Han, J., Sessa, E.B., Gardner, A.G., Shepherd, K.A., Ricigliano, V.A., Jabaily, R.S., Howarth, D.G., 2017. The unexpected depths of genome-skimming data: A case study examining Goodeniaceae floral symmetry genes1. Appl. Plant Sci. 20:5 (10), 1700042.

Bohmann, K., Mirarab, S., Bafna, V., Gilbert, M.T.P., 2020. Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. Mol. Ecol. 29, 2521–2534.

Boluda, C.G., Christe, C., Naciri, Y., Gautier, L., 2022. A 638-gene phylogeny supports the recognition of twice as many species in the Malagasy endemic genus *Capurodendron* (Sapotaceae). Taxon 71, 360–395.

Burge, S.W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E.P., Eddy, S.R., Gardner, P.P., Bateman, A., 2013. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res. 41, D226–D232.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinf. 10, 421.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973.

Chen, S., Zhou, Y., Chen, Y., Gu, J., 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.

Christe, C., Boluda, C.G., Koubínová, D., Gautier, L., Naciri, Y., 2021. New genetic markers for Sapotaceae phylogenomics: More than 600 nuclear genes applicable from family to population levels. Mol. Phylogenet. Evol. 160, 107123.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecol. Evol. (Amst.) 24, 332–340.

Eaton, D.A.R., 2014. PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30, 1844–1849.

Eaton, D.A.R., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62, 689–706.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Gandini, C.L., Sanchez-Puerta, M.V., 2017. Foreign plastid sequences in plant mitochondria are frequently acquired via mitochondrion-to-mitochondrion horizontal transfer. Sci. Rep. 7, 43402.

Gitzendanner, M.A., Soltis, P.S., Wong, G.-K.-S., Ruhfel, B.R., Soltis, D.E., 2018. Plastid phylogenomic analysis of green plants: A billion years of evolutionary history. Am. J. Bot. 105, 291–301.

Givnish, T.J., Zuluaga, A., Spalink, D., Soto Gomez, M., Lam, V.K.Y., Saarela, J.M., Sass, C., Iles, W.J.D., de Sousa, D.J.L., Leebens-Mack, J., Chris Pires, J., Zomlefer, W. B., Gandolfo, M.A., Davis, J.I., Stevenson, D.W., dePamphilis, C., Specht, C.D., Graham, S.W., Barrett, C.F., Ané, C., 2018. Monocot plastid phylogenomics, timeline, net rates of species diversification, the power of multi-gene analyses, and a functional model for the origin of monocots. Am. J. Bot. 105, 1888–1910.

Grandjean, F., Tan, M.H., Gan, H.M., Lee, Y.P., Kawai, T., Distefano, R.J., Blaha, M., Roles, A.J., Austin, C.M., 2017. Rapid recovery of nuclear and mitochondrial genes by genome skimming from Northern Hemisphere freshwater crayfish. Zool. Scr. 46, 718–728.

Hollingsworth, P.M., Li, D.-Z., van der Bank, M., Twyford, A.D., 2016. Telling plant species apart with DNA: from barcodes to genomes. Philos. Trans. R. Soc. Lond. B Biol. Sci. 371, 20150338.

Hughes, C.E., Eastwood, R.J., Donovan Bailey, C., 2006. From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. Philos. Trans. R. Soc. Lond.. Series B, Biol. Sci. 361, 211–225.

Johnson, M.G., Pokorny, L., Dodsworth, S., Botigué, L.R., Cowan, R.S., Devault, A., Eiserhardt, W.L., Epitawalage, N., Forest, F., Kim, J.T., Leebens-Mack, J.H., Leitch, I. J., Maurin, O., Soltis, D.E., Soltis, P.S., Wong, G.K., Baker, W.J., Wickett, N.J., 2018. A universal probe set for targeted sequencing of 353 Nuclear genes from any flowering plant designed using k-medoids clustering. Syst. Biol. 68, 594–606.

Kong, H., Condamine, F.L., Yang, L., Harris, A.J., Feng, C., Wen, F., Kang, M., 2021. Phylogenomic and macroevolutionary evidence for an explosive radiation of a plant genus in the miocene. Syst. Biol. syab068.

Lefort, V., Desper, R., Gascuel, O., 2015. FastME 2.0: A comprehensive, accurate, and fast distance-based phylogeny inference program. Mol. Biol. Evol. 32, 2798–2800.

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

Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676.

Liu, B.-B., Ma, Z.-Y., Ren, C., Hodel, R.G.J., Sun, M., Liu, X.-Q., Liu, G.-N., Hong, D.-Y., Zimmer, E.A., Wen, J., 2021. Capturing single-copy nuclear genes, organellar genomes, and nuclear ribosomal DNA from deep genome skimming data for plant phylogenetics: A case study in Vitaceae. J. Syst. Evol. 59, 1124–1138.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., Brettske, I., Gerber, S., Ginhart, A.W., Gross, O., Grumann, S., Hermann, S., Jost, R., König, A., Liss, T., Lüssmann, R., May, M., Nonhoff, B., Reichel, B., Strehlow, R., Stamatakis, A., Stuckmann, N., Vilbig, A., Lenke, M., Ludwig, T., Bode, A., Schleifer, K.H., 2004. ARB: a software environment for sequence data. Nucleic Acids Res. 32, 1363–1371.

Malé, P.-J.-G., Bardon, L., Besnard, G., Coissac, E., Delsuc, F., Engel, J., Lhuillier, E., Scotti-Saintagne, C., Tinaut, A., Chave, J., 2014. Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol. Ecol. Resour. 14, 966–975.

McKain, M.R., Johnson, M.G., Uribe-Convers, S., Eaton, D., Yang, Y., 2018. Practical considerations for plant phylogenomics. Appl. Plant Sci. 6, e1038.

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

Morales-Briones, D.F., Romoleroux, K., Kolář, F., Tank, D.C., 2018. Phylogeny and evolution of the neotropical radiation of *Lachemilla* (Rosaceae): Uncovering a history

of reticulate evolution and implications for infrageneric classification. Syst. Bot. 43, 17–34.
Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132.
Patwardhan, A., Ray, S., Roy, A., 2014. Molecular markers in phylogenetic studies-A review. Journal of Phylogenetics & Evolutionary Biology 02.
Pinho, C., Hey, J., 2010. Divergence with gene flow: Models and data. Annu. Rev. Ecol. Evol. Syst. 41, 215–230.
Pouchon, C., Fernández, A., Nassar, J.M., Boyer, F., Aubert, S., Lavergne, S., Mavárez, J., 2018. Phylogenomic analysis of the explosive adaptive radiation of the *Espeletia* Complex (Asteraceae) in the tropical andes. Syst. Biol. 67, 1041–1060.
Pouchon, C., Gauthier, J., Pitteloud, C., Claudel, C., Alvarez, N., 2022a. Phylogenomic study of *Amorphophallus* (Alismatales; Araceae): When plastid DNA gene sequences help to resolve the backbone subgeneric delineation. J. Syst. Evol. 61, 64–79.
Pouchon, C., Boyer, F., Roquet, C., Denoeud, F., Chave, J., Coissac, E., Alsos, I.G., Lavergne, S., The PhyloAlps Consortium, The PhyloNorway Consortium, 2022b. ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications. Mol. Ecol. Resour. 22, 2018–2037.
Rachtman, E., Sarmashghi, S., Bafna, V., Mirarab, S., 2022. Quantifying the uncertainty of assembly-free genome-wide distance estimates and phylogenetic relationships using subsampling. Cell Syst. 13, 817–829.e3.
Randriarisoa, A., Naciri, Y., Armstrong, K., Boluda, C.G., Dafreville, S., Pouchon, C., Gautier, L., 2022. One in, one out: Generic circumscription within subtribe Manilkarinae (Sapotaceae). Taxon 72, 98–125.
Rosenberg, N.A., Tao, R., 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57, 131–140.
Russell, A., Samuel, R., Rupp, B., Barfuss, M.H.J., Šafran, M., Besendorfer, V., Chase, M. W., 2010. Phylogenetics and cytology of a pantropical orchid genus *Polystachya* (Polystachyinae, Vandeae, Orchidaceae): Evidence from plastid DNA sequence data. Taxon 59, 389–404.
Sarmashghi, S., Bohmann, K.P., Gilbert, M.T., Bafna, V., Mirarab, S., 2019. Skmer: assembly-free and alignment-free sample identification using genome skims. Genome Biol. 20, 34.
Souza, H.A.V., Muller, L.A.C., Brandão, R.L., Lovato, M.B., 2012. Isolation of high quality and polysaccharide-free DNA from leaves of *Dimorphandra mollis* (Leguminosae), a tree from the Brazilian Cerrado. Genet. Mol. Res. 11, 756–764.
Steele, P.R., Hertweck, K.L., Mayfield, D., McKain, M.R., Leebens-Mack, J., Pires, J.C., 2012. Quality and quantity of data recovered from massively parallel sequencing: Examples in Asparagales and Poaceae. Am. J. Bot. 99, 330–348.
Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., Liston, A., 2012. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99, 349–364.
Thomson, A.M., Vargas, O.M., Dick, C.W., 2018. Complete plastome sequences from *Bertholletia excelsa* and 23 related species yield informative markers for Lecythidaceae. Appl. Plant Sci. 6, e01151.
Townsend, J.P., Su, Z., Tekle, Y.I., 2012. Phylogenetic signal and noise: Predicting the power of a data set to resolve phylogeny. Syst. Biol. 61, 835.
Trevisan, B., Alcantara, D.M.C., Machado, D.J., Marques, F.P.L., Lahr, D.J.G., 2019. Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. PeerJ 7, e7543.
Van Dongen, S., 2008. Graph clustering via a discrete uncoupling process. SIAM J. Matrix Anal. Appl. 30, 121–141.
Vanburen, R., Paris, M., Wai, J., Zhang, J., Huang, L., Zhou, H., Wang, H., Hwa, T.-Y., Kao, S.-M., Choi, J., Liao, Z., Lin, Z., Wang, L., Zhang, X., Yue, J., Sharma, A., Singh, R., Song, J., Yim, W.C., Ming, R., 2018. Sexual Recombination and selection during domestication of clonally propagated pineapple. SSRN Electron. J.
Vargas, O.M., Ortiz, E.M., Simpson, B.B., 2017. Conflicting phylogenomic signals reveal a pattern of reticulate evolution in a recent high-Andean diversification (Asteraceae: Astereae: *Diplostephium*). New Phytol. 214, 1736–1750.
Vargas, O.M., Heuertz, M., Smith, S.A., Dick, C.W., 2019. Target sequence capture in the Brazil nut family (Lecythidaceae): Marker selection and in silico capture from genome skimming data. Mol. Phylogenet. Evol. 135, 98–104.
Wang, X.-C., Chen, H., Yang, D., Liu, C., 2018. Diversity of mitochondrial plastid DNAs (MTPTs) in seed plants. Mitochondrial DNA Part A 29, 635–642.
Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20, 257.
Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T.-W., Li, Y., Xu, X., Wong, G.-K.-S., Wang, J., 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666.
Yu, X., Yang, D., Guo, C., Gao, L., 2018. Plant phylogenomics based on genome-partitioning strategies: Progress and prospects. Plant Diversity 40, 158–164.
Zhang, F., Ding, Y., Zhu, C.-D., Zhou, X., Orr, M.C., Scheu, S., Luan, Y.-X., 2019. Phylogenomics from low-coverage whole-genome sequencing. Methods Ecol. Evol. 10, 507–517.
Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinf. 19, 153.
Zhou, W., Soghigian, J., Xiang (Jenny), Q.-Y., 2021. A new pipeline for removing paralogs in target enrichment data. Syst. Biol. 71, 410–425.