Received: 19 June 2021 | Revised: 8 December 2021 | Accepted: 5 January 2022

DOI: 10.1111/1755-0998.13584

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications

Charles Pouchon[1] | Frédéric Boyer[1] | Cristina Roquet[1,2] | France Denoeud[3] | Jérome Chave[4] | Eric Coissac[1] | Inger Greve Alsos[5] | The PhyloAlps Consortium | The PhyloNorway Consortium | Sébastien Lavergne[1]

[1]LECA, Laboratoire d'Ecologie Alpine (LECA), Univ. Grenoble Alpes, CNRS, Univ. Savoie Mont Blanc, Grenoble, France

[2]Systematics and Evolution of Vascular Plants (UAB) – Associated Unit to CSIC, Departament de Biologia Animal, Biologia Vegetal i Ecologia, Universitat Autònoma de Barcelona, Bellaterra, Spain

[3]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

[4]Laboratoire Évolution et Diversité Biologique (EDB), UMR CNRS-IRD-UPS 5174, Toulouse, France

[5]The Arctic University Museum of Norway, UiT – The Arctic University of Norway, Tromsø, Norway

Correspondence
Charles Pouchon, Laboratoire d'Ecologie, UMR UGA-USMB-CNRS 5553, Université Grenoble Alpes, CS 40700, 38058 Grenoble cedex 9, France.
Email: charles.pouchon@orthoskim.org

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-10-INBS-09-08 and ANR-16-CE93-0004; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: SNF-310030L_170059; Norwegian Biodiversity Information Centre, Grant/Award Number: 14-14 and 70184209; Norges Forskningsråd, Grant/Award Number: 226134/F50 and 250963/F20

Abstract

Low-coverage whole genome shotgun sequencing (or genome skimming) has emerged as a cost-effective method for acquiring genomic data in nonmodel organisms. This method provides sequence information on chloroplast genome (cpDNA), mitochondrial genome (mtDNA) and nuclear ribosomal regions (rDNA), which are over-represented within cells. However, numerous bioinformatic challenges remain to accurately and rapidly obtain such data in organisms with complex genomic structures and rearrangements, in particular for mtDNA in plants or for cpDNA in some plant families. Here we introduce the pipeline ORTHOSKIM, which performs in silico capture of targeted sequences from genomic and transcriptomic libraries without assembling whole organelle genomes. ORTHOSKIM proceeds in three steps: (i) global sequence assembly, (ii) mapping against reference sequences and (iii) target sequence extraction; importantly it also includes a range of quality control tests. Different modes are implemented to capture both coding and noncoding regions of cpDNA, mtDNA and rDNA sequences, along with predefined nuclear sequences (e.g., ultraconserved elements) or collections of single-copy orthologue genes. Moreover, aligned DNA matrices are produced for phylogenetic reconstructions, by performing multiple alignments of the captured sequences. While ORTHOSKIM is suitable for any eukaryote, a case study is presented here, using 114 genome-skimming libraries and four RNA sequencing libraries obtained for two plant families, Primulaceae and Ericaceae, the latter being a well-known problematic family for cpDNA assemblies. ORTHOSKIM recovered with high success rates cpDNA, mtDNA and rDNA sequences, well suited to accurately infer evolutionary relationships within these families. ORTHOSKIM is released under a GPL-3 licence and is available at: https://github.com/cpouchon/ORTHOSKIM.

KEYWORDS
bioinformatics, Ericales, genome skimming, mitochondrion, nucleus, phylogenomic, plastome, transcriptome

Members of the PhyloAlps and PhyloNorway consortia are co-authors of this article and can be found under the heading Collaborators.

1 | INTRODUCTION

In the last decade, the advent of high-throughput sequencing techniques has offered unprecedented possibilities in evolutionary biology by greatly increasing both taxonomic coverage and the amount of sequence data at a relatively low cost. Alongside, the development of fast and efficient genome assemblers (Bankevich et al., 2012; Simpson et al., 2009; Zerbino & Birney, 2008) has also given access to large sets of genomic contigs for the genotyping of single nucleotide polymorphisms, among others (e.g., Li et al., 2019; Plomion et al., 2018; Twyford & Ness, 2017). The resulting profusion of genomic resources sheds new light on the phylogenomic inference of rapid diversification events (Pouchon et al., 2018; Vargas et al., 2017) or on DNA barcoding (Coissac et al., 2016), widely used for species identification (Braukmann, Kuzmina, et al., 2017; CBOL Plant Working Group et al., 2009; Tyagi et al., 2019), species monitoring and conservation (Fahner et al., 2016; Porter & Hajibabaei, 2018; Weigand et al., 2019), or biodiversity analyses (Clarke et al., 2019; Kennedy et al., 2020).

Low-coverage whole-genome shotgun sequencing, or genome skimming, is a powerful method for acquiring genomic data on nonmodel taxa. This approach consists of sequencing the total genomic DNA of a nonmodel organism at low coverage, typically less than 5× coverage of the nuclear genome, through random shearing with no targeted genomic enrichment (McKain et al., 2018; Straub et al., 2012). Genome skimming includes sequences from all genomic compartments, but especially those represented in higher copy numbers with a relatively high coverage (~>30×), such as the organelle genomes, plastid DNA (cpDNA) and mitochondrial DNA (mtDNA), and the nuclear ribosomal regions (rDNA). Genome skimming has been efficiently used for phylogenetic analyses (Barrett et al., 2016; Liu et al., 2018; Malé et al., 2014) and DNA barcoding questions (Bohmann et al., 2020; Coissac et al., 2016). It is widely applied to both fresh and degraded samples such as museum collections, and for a range of organisms such as plants (Alsos et al., 2020; Bakker et al., 2016; Nevill et al., 2020), animals (Grandjean et al., 2017; Linard et al., 2015; Trevisan et al., 2019) and fungi (Greshake et al., 2016; Meiser et al., 2017). Thus, genome skimming is scalable and cost-effective for large-scale projects of biodiversity genomics. For instance, we have recently produced genome skimming libraries for a total of 6655 samples and 5575 plant taxa (including species and subspecies) of the artic and alpine regions within the projects PhyloAlps and PhyloNorway (see phyloalps.org for more details, Alsos et al., 2020).

The assembly of large genome skimming projects leads to many bioinformatic challenges to rapidly and efficiently produce large data sets of reference barcodes or aligned sequences. Most existing computational workflows perform de novo assembly of whole mtDNA, cpDNA and rDNA regions through targeted organelle assembler and genome annotation pipelines (Freudenthal et al., 2020; McKain et al., 2018). Examples of such pipelines include: ORG.ASM (http://metabarcoding.org/org-asm), NOVOPLASTY (Dierckxsens et al., 2016), IOGA (Bakker et al., 2016), FAST-PLAST (https://github.com/mrmckain/Fast-Plast) and GETORGANELLE (Jin et al., 2020). One issue is that chloroplast genomes sometimes have complex structures due to gene translocation, inverted repeat expansion or contraction, or multiple sequence repeats. As a result, full cpDNA genome assembly is difficult (Bendich, 2004; Twyford & Ness, 2017), a particularly prominent issue in several Angiosperm families such as Campanulaceae, Cyperaceae, Ericaceae, Goodeniaceae, Orchidaceae and Poaceae (Alsos et al., 2020; Freudenthal et al., 2020; Nevill et al., 2020). Also, mtDNA genomes are currently underused in plant phylogenomic applications, because of their complex structure (Kozik et al., 2019), and the difficulty in assembling them (Van de Paer et al., 2018; Zhang et al., 2015; Zhang, Jin, et al., 2019). On the other hand, most phylogenetic studies based on organelle sequences mainly focus on genes and not on the whole genome structure (e.g., Givnish et al., 2018; Li et al., 2019), emphasizing the need for alternative approaches focusing on targeted sequences, such as ATRAM (Allen et al., 2015) or HYBPIPER (Johnson et al., 2016), but suitable for genome skimming data and not dependent on assembling the whole organelle. A further impediment is that genetic material has been frequently transferred from the plastid genome into the mitochondrial genome, called MTPTs (mitochondrial plastid DNAs), during the course of seed plant evolution (Gandini & Sanchez-Puerta, 2017; Sloan & Wu, 2014; Straub et al., 2013; Wang et al., 2018). Such DNA transfers may cause errors in genome assemblies and annotations (discussed in Jin et al., 2020), resulting in species misidentifications, and/or in deteriorating phylogenetic signal when using genes from within these genomic regions (Gandini & Sanchez-Puerta, 2017; Park et al., 2020).

Here, we present a user-friendly and generic pipeline called ORTHOSKIM, which addresses all of the aforementioned limitations. ORTHOSKIM performs in silico sequence capture from genomic and transcriptomic sequence data through mapping of global assemblies (i.e., on all sequencing reads) on a set of target reference sequences. It makes it possible to analyse large genome skimming data by capturing cpDNA, mtDNA and rDNA sequences (both coding and noncoding) in a single analysis. ORTHOSKIM extracts only target sequences in genomic assemblies without assembling the whole organelle genomes, so it works in organisms with complex cpDNA or mtDNA structures. Moreover, ORTHOSKIM can be used to capture nuclear markers (nuDNA, e.g., ultraconserved elements [UCEs]) and single-copy orthologues (i.e., from the BUSCO library of single-copy DNA sequences; see Simão et al., 2015; Waterhouse et al., 2018), which enhances its application to genome skimming with large enough sequencing depths (Berger et al., 2017; Liu et al., 2021; Vargas et al., 2019; Zhang, Ding, et al., 2019), RNA sequencing (RNAseq) libraries (Larson et al., 2020; Zhang et al., 2020) or hybrid capture libraries (Andermann et al., 2020; Koenen et al., 2020). Finally, ORTHOSKIM provides tools to rapidly perform multiple alignments of the captured sequences across a range of libraries. Although it is applicable to any eukaryotic taxa, here we demonstrate the utility of ORTHOSKIM to infer the phylogenetic

relationships for two flowering plant families, one for which previous organelle assemblers have regularly failed, Ericaceae, and another one for which they performed well, Primulaceae. These analyses illustrate the range of applications of this new pipeline.

2 | MATERIALS AND METHODS

2.1 | ORTHOSKIM overview

ORTHOSKIM is an open source pipeline, released under a GPL-3 licence and available at https://github.com/cpouchon/ORTHOSKIM. It is written in python and bash languages and runs as command line on UNIX environments. Its modular design takes advantage of multicore/processors architectures, for instance in Linux environments and HPC clusters. As it depends on different softwares and python libraries, ORTHOSKIM is contained within a conda package with all dependencies. Please refer to the online code documentation on the GitHub repository for installation and user instructions.

Different tasks, called "modes," are implemented in ORTHOSKIM. When calling ORTHOSKIM from the command line, the "mode" has to be specified as well as another parameter, the "target." Figure 1 provides an overview of ORTHOSKIM calling, with different "mode" and "target" options, to: (1) produce the sequence reference database for cpDNA, mtDNA and rDNA targets (pink arrows in Figure 1), (2) perform contig assembly and cleaning from whole sequencing reads (green arrows), (3) capture targeted sequences based on the most similar reference (blue arrows), and (4) obtain the alignments of these captured sequences among libraries (orange arrows). The sequence capture strategy can be aimed to cpDNA, mtDNA, rDNA, nuDNA (as UCE sequences) or BUSCO-type markers on both coding and noncoding sequences. A parameter file needs to be supplied to indicate the parameters and paths for data location, together with a sample description file (see documentation). Computation time is reported at the end of each call. A typical use of the ORTHOSKIM pipeline, as depicted in Figure 1, is described in the following section.

2.1.1 | Reference database collection

ORTHOSKIM first uses a multitaxon database of reference sequences for each of the targeted markers, from which the software will select the closest reference to perform sequence capture (see Section 2.1.3). The reference input files required for each of the target sequences are displayed by yellow boxes in Figure 1 and summarized in Appendix S1. We encourage users to customize their own reference database for the specific purposes and focus of their study. Please see the online documentation for recommendations on how to construct these input files.

For nuDNA sequences, the users have to provide their own database of reference sequences, consisting in a multifasta file of the target regions with a different data type depending on the capture mode: amino-acid sequences for the "*nucleus_aa*" target (suitable for coding sequences), or nucleotide sequences for the "*nucleus_nt*" target (for noncoding sequences). Sequence names need to be compliant with the ORTHOSKIM nomenclature (see documentation). ORTHOSKIM can also use the BUSCO single-copy nuDNA reference sequence database to capture such markers by using the "*busco*" target in the "*capture*" mode. In the BUSCO mode, ORTHOSKIM uses the amino acid consensus, or ancestral sequence variants of each BUSCO gene. Using amino acid sequences as references will maximize gene recovery during the capture even for the most divergent taxa. The different BUSCO data sets can be downloaded at: https://busco-data.ezlab.org/v4/data/lineages/. Note that no seeds are required for running ORTHOSKIM in the nuclear and BUSCO modes ("*nucleus_aa*," "*nucleus_nt*" or "*busco*" targets).

For all cpDNA, mtDNA and rDNA target sequences, the database is built from the "*database*" mode along with the corresponding target (e.g., "*chloroplast*" target for cpDNA, step 1 in Figure 1). For such purposes, ORTHOSKIM needs an annotated genome file with multitaxon accessions for each of the corresponding genome compartments, with "seeds" sequences for each target sequence. These seeds consist of multifasta files with a single reference of each of the targeted sequences (see online documentation for examples and formats). ORTHOSKIM extracts all gene sequences from the annotated genomes and maps them onto the given seed sequences to correctly identify targeted reference genes. It is important to note that each of the three annotation files has to be given for plant models, or both mtDNA and rDNA annotation files for other organisms, even if a single region is targeted (e.g., cpDNA sequences). This is because such files are also used to assign the genomic assemblies to the cpDNA, mtDNA or rDNA regions in order to take into account gene transfer between such regions (Section 2.1.3). For both cpDNA and mtDNA, seed files are given separately for the target coding (CDS) genes with amino acid sequences, and with nucleotide sequences for the noncoding RNA genes. For the cpDNA, a seed sequence file for the chloroplast *trnL-UAA* gene, a traditional plant barcode, must also be provided. Concerning the rDNA, the three rRNA genes sequences (i.e., *rrn18S*, *rrn5.8S* and *rrn26S*) have to be included in the corresponding seed sequence file. ORTHOSKIM then designs probes from these rRNA genes for both seeds and references, allowing the identification and capture of the two internal transcribed spacer regions (*ITS1* and *ITS2*).

The resulting reference sequence database consists of a multifasta file for each type of gene sequence (i.e., CDS, rRNA and tRNA), generated with amino acid sequences for CDS and nucleotide sequences for rRNA and tRNA genes. In addition, two free capture modes, working with any reference sequences, were also implemented for cpDNA and mtDNA ("chloroplast_nt" and "mitochondrion_nt" capture targets) that can be easily used to capture intergenic regions. For this purpose, a custom reference database has to be supplied for each of two modes, consisting of a multitaxon fasta file with nucleotide sequences of targeted regions and sequence names compliant with the ORTHOSKIM nomenclature (see documentation, Appendix S1). Users may also supply their own reference fasta files for each type of sequence (CDS, rRNA and tRNA),

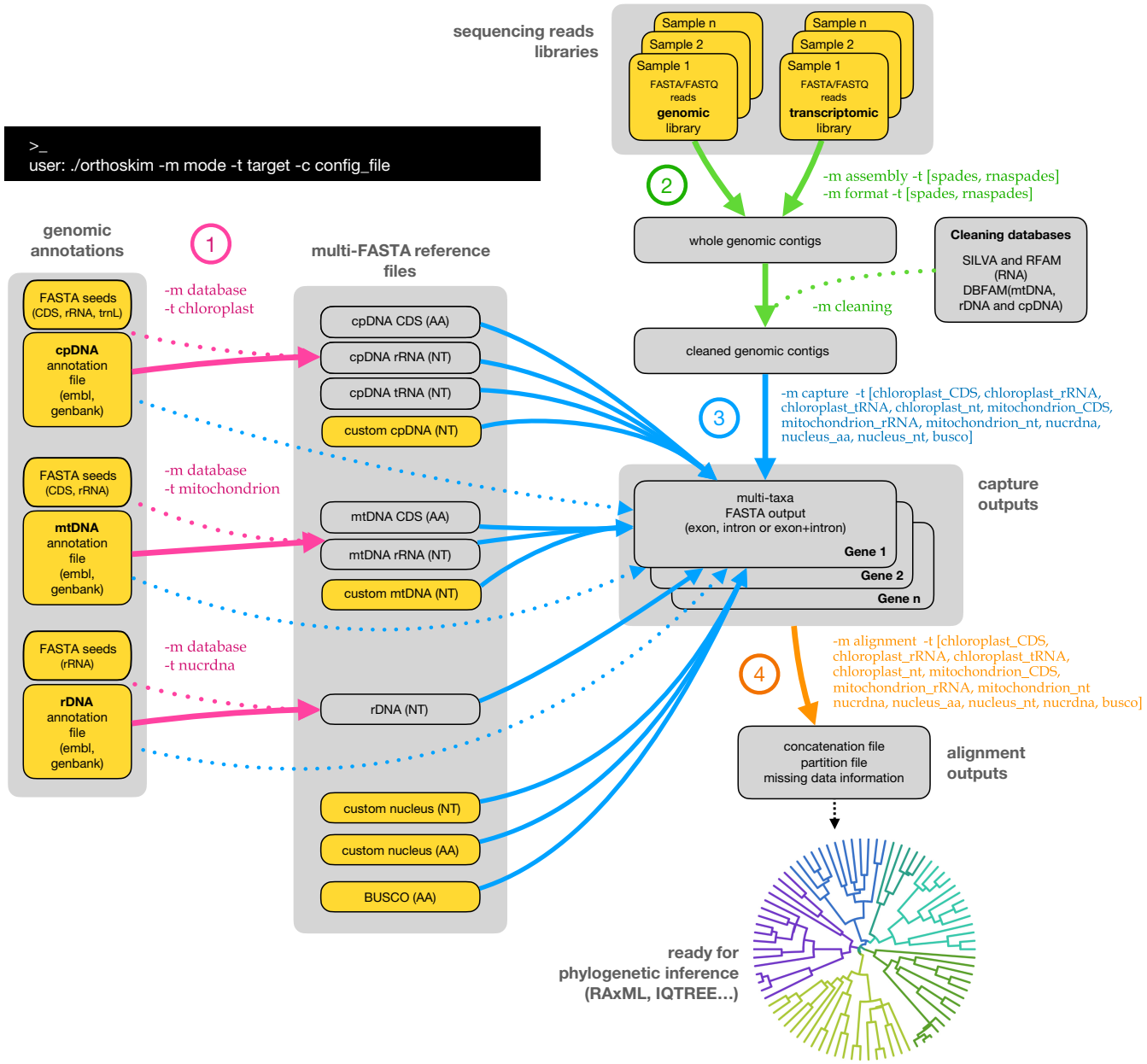

FIGURE 1 ORTHOSKIM workflow. The main steps, shown by coloured arrows are: (1) compute a multitaxa reference database for cpDNA, mtDNA and rDNA targets (in pink arrows); (2) assemble reads (including a cleaning step, green arrows); (3) capture targeted sequences from the database (blue arrows); and (4) align captured sequences across libraries and get alignment files for phylogenetic inference (orange arrows). The yellow boxes indicate user-supplied information. The command lines to call the software are printed next to the respective steps

but need to collect annotations for cpDNA, mtDNA and rDNA for the contig selection step (Section 2.1.3).

By default, ORTHOSKIM is supplied with a database of reference sequences designed to the study of green plants (i.e., Viridiplantaeae) genome skimming data sets: the BUSCO plant set (viridiplantaeae_odb10), 353 UCEs designed for angiosperms (Johnson et al., 2018), which can be used as "*nucleus_nt*" references, and a collection of annotations for plant cpDNA, mtDNA and rDNA genomic regions collected from the NCBI. For any other eukaryotic taxa, please refer to the online documentation for more information on how to develop and format input files.

2.1.2 | Global assembly and contaminant cleaning

The second step of ORTHOSKIM is to perform a global assembly of the sequencing reads into a set of contigs (without targeting a specific organelle genome), which are then used to capture target sequences (see step 2 in Figure 1). Assemblies are performed for each library using SPADES (Bankevich et al., 2012). SPADES is a nontargeted genome assembler, which combines good performance and low computational requirements (Bankevich et al., 2012), especially for organelle genomes (Freudenthal et al., 2020; Jin et al., 2020). This step is called in ORTHOSKIM using the "*assembly*" mode and

the "*spades*" target for genomic libraries or the "*rnaspades*" target for transcriptomic libraries. For genomic libraries, SPADES is run in read-correction mode and with the *–cov-cutoff auto* parameter to determine conservative read coverage cutoff values. The other parameters of the assembly (i.e., kmer sizes, threads, available memory) are read from the input parameter file. Once assemblies are completed, the "*format*" mode is used to extract and format the final scaffold of sequences from SPADES outputs.

Several sources of contamination could compromise the analysis of genomic data. The "*cleaning*" mode is next used to identify and remove potential cpDNA, mtDNA and rDNA contaminants in the final assemblies. For this purpose, all contigs of each library are blasted against the RFAM (Burge et al., 2013) and the SILVA (Ludwig et al., 2004) RNA databases, taken from the SORTMERNA software (Kopylova et al., 2012), and also against our own database, called DBFAM, including a subset of cpDNA, mtDNA and rDNA genomes covering a wide taxonomic range in Eukaryota with plant, animal and fungal sequences (see step 2 in Figure 1). For each contig, the taxonomic level of the best-hit BLAST output is next compared to an expected level supplied by the user in the input parameter file. Contigs outside the expected taxonomy are hence considered as contaminants and filtered out. This function was not designed to assign a contig to a specific taxonomic level (e.g., family, genus or species) but to quickly remove obvious contaminants. So, we recommend keeping a deep expected taxonomic level (e.g., kingdom, subkingdom or phylum), as the taxonomy of the SILVA, RFAM and DBFAM databases used is nonexhaustive. For example, for a plant model, the user can set the expected level to "*Embryophyta*" to remove all contaminants contigs in assemblies with a BLAST best-hit outside this level, which will filter out fungal, animal or bacterial contigs.

2.1.3 | Sequence capture

Step 3 of ORTHOSKIM consists in the capture of targeted sequences from cleaned assemblies (blue arrow in Figure 1). This step is called using the "*capture*" mode in accordance with the targeted genome compartment and sequence type, for example -t *chloroplast_CDS* for chloroplast CDS or -t *chloroplast_rRNA* for the chloroplast rRNA. Sequence capture follows the three following steps:

Selection of references and contigs

For each targeted sequence and study library, ORTHOSKIM first selects the closest reference from the reference sequence database. This selection follows the NCBI taxonomy: a valid taxonomy identifier (TaxId) is supplied for each library of the sample description file into the library name. If the TaxId is not provided, ORTHOSKIM will use all the provided seeds as reference, or the longest sequence for the nuclear targets ("*nucleus_aa*" and "*nucleus_nt*"). With the single-copy reference sequences ("*busco*" target), the procedure is slightly different: with a formatted set of references from BUSCO, the selection step is skipped because the ancestral sequence variants are provided.

Each contig is then assigned to a genomic compartment. This step is crucial to avoid cross-genome contamination, in particular MTPTs. A mitochondrial copy of a plastid gene should not be captured even if the homologous plastid gene is absent from the reconstructed contigs. To do so, each contig is blasted against the five closest organelle genomes and rDNA regions, and this procedure is repeated for each library and for all three genomic compartments: mtDNA, cpDNA and rDNA regions. As a result, annotations need to be supplied for all three compartments in plant models even if only one cpDNA, mtDNA or rDNA region is targeted, or only for mtDNA and rDNA for other study models. For modes without annotated genomes ("*busco*," "*nucleus_aa*" and "*nucleus_nt*" targets), it is also possible to identify which contigs align with targeted sequences. Mapping is performed using the DIAMOND software for amino acid sequences (Buchfink et al., 2015) or using BLAST for nucleotide sequences. Several thresholds should be specified by the user within the parameter file, to set kmer coverage, contig length and minimal BLAST *e*-value, so as to exclude all low-quality contigs following classic standards for these thresholds. Default values, which are good enough for most applications, are given in the supplied parameter file.

Exon–intron sequence prediction

The selected contigs are then aligned on the closest targeted references using the EXONERATE algorithm. This algorithm predicts the entire exonic structure of the targeted sequences by incorporating the appropriate gaps and frameshifts. In ORTHOSKIM, this step is achieved by using the *protein2genome* or the *genome2genome* mode, depending on the type of reference sequences (i.e., amino acid or nucleotide sequence). A gff3-formatted output is generated.

Extraction of targeted sequences

The homology of each targeted sequence is next assessed from sequence similarity thresholds. For each targeted sequence, the best alignment with the greatest sequence similarity is identified from the gff table and the sequence is outputted into a fasta file (e.g., *ycf1.fa* file, see documentation for output files). Options exist in the input parameter file to extract exonic regions, intronic regions or both. For coding targeted sequences, ORTHOSKIM quality-checks the output by verifying that the longest open reading frame (ORF) from the extracted exons covers at least a minimal fraction of the capture sequence. The fraction must be set in the input parameter file, with 80% the default value. The NCBI genetic code used for translation of the sequences must also be fixed to best accommodate the inference of ORF according to the studied model and the target sequences. For example, the vertebrate mitochondrial code must be specified if animal mtDNA sequences are targeted. This check flags variation or error in gene predictions, such as alternative start codons in the protein sequence of the reference. If such condition is not verified, due to pseudogenes or prediction errors, the sequence is tagged as a gene-like sequence (e.g., *ycf1-like*), and stored in a different file (e.g., *ycf1-like.fa* file). For plant models, a second control is performed to ensure the correct origin

of reconstructed organelle genes: the extracted sequences are aligned against the organelle seeds, to filter out chimeric organelle contigs. Such chimeras are assembled from conserved regions, and must be dealt with separately because they usually pass the previous quality-control steps.

ORTHOSKIM also includes a coverage cut-off option during the capture of cpDNA and mtDNA target sequences, to remove all possible contigs from organelle contaminants (e.g., alien sequenced DNA), which are characterized by lower sequencing depths. This step allows filtering out remaining contaminant contigs, which would have passed through the final "cleaning" step of global assembly. To achieve this, the software retrieves the coverage of each contig for which the best alignment of the targeted sequences is identified and then computes the mean coverage adjusted by the contig lengths. Each contig having a weighted coverage below this mean minus 3 standard deviations is then removed from the data set. We recommend using this option only for genome skimming libraries, for which homogeneous coverage is expected for cpDNA and mtDNA. For the rDNA target, *ITS1* and *ITS2* barcodes are extracted from the intronic regions of our rRNA probes. ORTHOSKIM stores the list of contigs for which sequences were extracted, in case the user prefers to use the contig sequences directly.

2.1.4 | Alignment of captured sequences

In a final step, ORTHOSKIM includes an "*alignment*" mode which generates multiple sequence alignments from the captured sequences for each of the analysed libraries (step 4 Figure 1). This facilitates downstream phylogenetic analyses. Multiple alignments are performed using MAFFT (Katoh & Standley, 2013) with the *--adjustdirectionaccurately* option, and poorly aligned regions are trimmed out using TRIMAL (Capella-Gutiérrez et al., 2009) optionally with the *automated1*, *gappyout* or *strictplus* method if indicated in the parameter file. At this step, the user chooses which libraries and which sequences will be aligned, from a list provided in the parameter file. We recommend checking for the presence of potential contaminants in the captured genes before aligning all extracted sequences. This can be done through the "*checking*" mode for a selection of captured sequences, including classic plant DNA barcodes (e.g., *matK*, *rbcL*, *trnL-UAA*). In these, the taxonomic assignation of the extracted sequences is verified by blasting the sequences to the NCBI database. Moreover, ORTHOSKIM is not designed to treat duplicated genes, so we recommend checking sequence alignments visually to ensure that homologous regions were well captured: plant mtDNA in particular is known to include divergent or chimeric gene copies (e.g., Kozik et al., 2019; Omelchenko et al., 2020; Palumbo et al., 2020). Softwares such as PREQUAL (Whelan et al., 2018) or SPRUCEUP (Borowiec, 2019), may be used to further check and correct the homology assignment of captured genes.

A concatenated alignment file is produced together with a partition file giving the location of the chosen sequences into the alignment in RAXML-style format. Libraries with missing data above a threshold specified in the parameter file are excluded. Output files generated by ORTHOSKIM can be used as inputs of phylogenetic tree reconstruction softwares (e.g., IQTREE; Minh et al., 2020) or pipelines (e.g., TREEASY; Mao et al., 2020).

2.2 | Pipeline illustration

Here we provide an example application of ORTHOSKIM, using a set of shotgun libraries sequenced for plant species from two families of Ericales: Primulaceae and Ericaceae. Inferring the phylogenetic relationships within this diverse and species-rich order, which found considerable interest among systematists, has been a recalcitrant problem in the past (Rose et al., 2018; Schönenberger et al., 2015). Moreover, Ericaceae are resistant to cpDNA assembly due to their complex plastid structure (Inger Greve Alsos et al., 2020). ORTHOSKIM was used to capture chloroplast and mitochondrial genes along with ribosomal regions, and to subsequently infer phylogenetic hypotheses from these markers.

2.2.1 | Data sets

We used a total of 114 genome skimming libraries generated within the framework of the PhyloAlps and PhyloNorway projects (Inger Greve Alsos et al., 2020). The libraries, with a mean sequencing depth estimated at 1.3× coverage (Appendix S2), were produced from 54 samples of Ericaceae, 52 of Primulaceae and eight of the Balsaminoid clade (one *Macrgravia* +seven *Impatiens* species). This sampling represented a total of 30 genera and 91 distinct species, including replicates and subspecies (see Appendix S2). Balsaminoid taxa were incorporated as outgroups for phylogenetic inference according to previous studies (Rose et al., 2018; Zhang et al., 2020).

This sampling included 82 libraries generated from leaf tissues freshly collected in the field and dried in silica gel (PhyloAlps project), but also 32 libraries obtained from the sequencing of herbarium leaf tissues (PhyloNorway project, see Appendix S2). This strategy was implemented to assess the efficiency of ORTHOSKIM to deal with both degraded and fresh DNA samples.

We also assessed the efficiency of ORTHOSKIM to extract phylogenetically informative sequences from both genomic and transcriptomic libraries. To that effect, we included four RNAseq data sets available from the European Nucleotide Archive (ENA) of Ericales generated from fresh leaf tissues, for the following species: *Lysimachia nummularia* L. (SRR6434984), *Primula vulgaris* Huds. (SRR1578145), *Pyrola americana* Sweet (SRR11994223) and *Vaccinium corymbosum* L. (SRR6472974).

2.3 | Pipeline run

ORTHOSKIM was executed on a single PC running Linux with 24 cores and 125 GB of RAM. The following versions of software dependences were used: BLAST version 2.9.0, DIAMOND version 0.9.13, EXONERATE version 2.2.0, MAFFT version 7.429, PYTHON version 3.7.5 and SPADES version 3.13.1; and we used the following python modules: BioPython version 1.74, numpy version 1.17.2, ete3 version 3.1.1 and joblib version 0.14.1.

A database was computed for the cpDNA and the rDNA sequence targets using the "*-m database*" mode of ORTHOSKIM, using 164 annotated cpDNA and 267 annotated rDNA regions produced for Ericales by the PhyloAlps and PhyloNorway projects (see Alsos et al., 2020), and sixx more annotated cpDNAs available from NCBI GenBank/Refseq (KX668174, MK550716, KU513437, MN418389, LC521967, MT533181). The database consisted of 79 chloroplast coding DNA sequences (CDS), four chloroplast noncoding rRNA (*rrn16S*, *rrn23S*, *rrn4.5S* and *rrn5S*) genes, the *trnL-UAA* gene, and the three nuclear ribosomic noncoding rRNA genes (*rrn18S*, *rrn26S* and *rrn5.8S*). For the mitochondrial genome, we computed the database from annotated mtDNA of 307 plant species available from NCBI GenBank/Refseq, including 39 coding and three noncoding rRNA genes (*rrn18S*, *rrn26S* and *rrn5S*). The seed sequences for the cpDNA and rDNA genes were extracted from the cpDNA genome of *Arabidopsis thaliana* (L.) Heynh. (AP000423), while seed sequences for the mtDNA genes were extracted from the mtDNA genomes of *Camellia sinensis* L. (NC_043914) and *Vaccinium macrocarpon* Aiton (NC_023338).

Global assemblies of sequence contigs were performed on each library with the "*-m assembly*" mode with a kmer size of 55, using the "*-t spades*" target for genome skimming libraries and the "*-t rnaspades*" target for the transcriptomic libraries. To capture targeted sequences, we included only contigs with kmer coverage ≥3 and size ≥500 bp. We set a minimal mapping *e*-value at 1*e*-5. We captured the exonic regions for each targeted sequence with the exception of the *trnL-UAA* for which the intronic region was included in the target. Sequence capture was considered successful if at least 60% of the reference was covered. Moreover, the minimal fraction of captured sequence covered by an ORF was set at 80%. ORTHOSKIM was run with the "-m capture" mode along with "*-t chloroplast_CDS*," "*-t chloroplast_rRNA*," "*-t chloroplast_tRNA*," "*-t mitochondrion_CDS*," "*-t mitochondrion_rRNA*" and "*-t nucrdna*" targets (see step 3 Figure 1). Contaminations were checked using the "*-m checking*" mode on *matK*, *rbcL*, *trnL-UAA*, *ITS1* and *ITS2*.

Next, the libraries were aligned for each captured sequence with the "*-m alignment*" mode by excluding gene-like tagged sequences and by trimming alignments using the heuristic *automated1* algorithm of TRIMAL. This produced a concatenated data set for each set of the cpDNA, mtDNA and rDNA captured regions. Libraries with <70% of the values in the concatenated alignments were filtered out. Alignments were checked for the possible capture of paralogues, in particular for the mtDNA, by using SPRUCEUP to identify and remove outlier sequences from the obtained alignments.

Phylogenetic inferences were conducted on the cpDNA, mtDNA and rDNA concatenated and partition files in IQTREE (Minh et al., 2020). We used MODELFINDER (Kalyaanamoorthy et al., 2017) to determine the best-fit model by partition (-m MFP), 1,000 ultrafast bootstrap (UFBoot) replicates (-bb 1000), the hill-climbing nearest neighbour interchange (NNI) search option (-bnni), and the SH-like approximate likelihood ratio test to assess branch support with 1000 replicates (-alrt 1000).

3 | RESULTS

3.1 | Sequence capture in Ericales

The median computational time was 01 h:04 min:00 s per genome skimming library and 03 h:02 min:24 s per RNAseq library using 24 cores and 125 GB of RAM per library (Appendix S2). This difference can be explained by the larger library sizes for the RNAseq libraries used here (Appendix S3). Moreover, the computational time within genome skimming libraries was highly correlated with the library size produced in terms of read number and the available number of assembled contigs, regardless of sample preservation (silicagel or herbarium) and independently of the sampled family (Appendix S3). Although some variability was shown in sequencing and assembly summary statistics (i.e., sequencing reads, contig number, N50, Appendix S1), the success rate of chloroplast gene capture from genome skimming libraries was 98% for Primulaceae, 93% for Ericaceae and 96% for Balsaminoids (including complete, partial and pseudogenes, Figure 2). This overall success of capture appeared to be correlated with the number of reconstructed contigs for cpDNA (Appendix S4), which are, as a whole, highly covered. Most of the genes were captured completely although some genes were only partially recovered, such as the genes with complex structure (e.g., *ycf1*, *ycf2* and *rps12*, particularly in Ericaceae), or were sometimes totally missing from the final concatenated alignment (e.g., *ycf2* in Ericaceae). We also found variant copies for some genes across families, such as *clpP* and *infA* in Primulaceae, or *ndh* genes in Ericaceae (tagged as gene-like in Figure 2). The capture rate of classical cpDNA plant barcodes (*matK*, *rbcL* and the intron of *trnL-UAA*) was 100% for Primulaceae, 94%, 98% and 100% for Ericaceae, respectively; and 75%, 100% and 75% for the Balsaminoids, respectively (Figure 2). In Balsaminoids, *matK* had a variant copy in 12.5% of samples (Figure 2).

Concerning mitochondrial genes, the capture rate from genome skimming libraries was 84% for Primulaceae, 69% for Ericaceae and 82% for Balsaminoids (Figure 2). Such rates correlated with the overall reconstructed size, and the amount and the coverage of mtDNA contigs (Appendix S4). More mtDNA genes were partially recovered in comparison to cpDNA, such as *nad2* or *nad5*, and *nad1*, *rps7*, *rps15* and *rpl16* were nearly missing from the captured genes (Figure 2). Concerning the rDNA targets, the capture rate was 97% in Primulaceae, 96% in Ericaceae and 100% in Balsaminoids (Figure 2) and highly correlated with the total reconstructed size of the rDNA contigs (Appendix S4). The *rrn18S* gene was captured with the lowest success rate. We recovered *ITS1* and *ITS2* barcodes in 96% and 100% of Primulaceae, in 98% and 100% of Ericaceae, and 100% for both ITS copies in Balsaminoids (Figure 2).

For the RNAseq libraries, we obtained a success rate for chloroplast, mitochondrion and ribosomal sequence recovery of respectively 94%/95%/80% for *Lysimachia nummularia*, 95%/74%/100% for *Primula vulgaris*, 62%/71%/80% for *Pyrola americana* and 93%/88%/60% for *Vaccinium corymbosum*.

FIGURE 2 Sequence capture rates in Ericales genome skimming data sets for cpDNA genes, mtDNA genes and rDNA regions. Grey shading indicates the length coverage of the reference, whereas gene-like captured sequences (pseudogenes or prediction errors) are shown in orange. Icons were retrieved from BioRender

3.2 | Phylogenetic inferences based on sequence capture

We inferred three different phylogenies, one for cpDNA genes, mtDNA genes and the rDNA sequences.

For chloroplast DNA, all 118 libraries were retained in the final concatenation of all captured gene alignments. Only *accD* and *clpP* were excluded from the concatenation because of the low alignment quality produced for both genes. The resulting alignment consisted in 68,010 bp with 18,587 informative sites and 16% missing data (Figure 3). Sequence homology was well estimated as only 0.22% of sites were identified as outliers in SPRUCEUP. The proportion of missing data tends to be slightly higher for herbarium and silica gel-dried materials (Mann–Whitney $p = .061$). Moreover, Ericaceae had significantly more missing data (Kruskal–Wallis $p = 2.2 \times 10^{-16}$). The resulting phylogeny was well resolved at both deep and shallow nodes with 86% of the nodes having a UFBoot ≥95 (Figure 3). In the analyses, all genera were monophyletic, except for *Primula* L. due to the placement of *Dodecatheon frigidum* Cham. & Schltdl within the *Primula* clade. All species replicates clustered together at the species level (*Impatiens noli-tangere* L., *Vaccinium myrtillus* L., *Trientalis europaea* L.) with the sole exception of *Androsace halleri* L. Samples from the transcriptomic libraries were also correctly placed: *Pyrola americana* was sister to the clade of *Pyrola media* Sw. and *Pyrola minor* L.; *Vaccinium corymbosum* sister to *Vaccinium myrtillus*; *Lysimachia nummularia* sister to its genomic replicate; and *Primula vulgaris* sister to *Primula veris veris* L.

For mtDNA, four genes (*nad1*, *rpl16*, *rps7*, *rps15*) had a consistently low capture rate and were removed from the concatenation (see Figure 2). Nine Ericaceae libraries had ≥70% of missing data and were removed. Concatenation of the remaining 38 gene alignments for 109 libraries resulted in 33,095 bp with 3717 informative sites, 12% missing data and 0.48% of outlier sites (Figure 4). In contrast to cpDNA, a higher proportion of missing data was found for herbarium specimen than for silica gel-dried plant material (Mann–Whitney $p = 7.1 \times 10^{-6}$). Primulaceae had a lower proportion of missing data than the other two families (Kruskal–Wallis $p = 3.80 \times 10^{-10}$).

The phylogenetic hypothesis inferred from mitochondrial genes was less resolved than that for cpDNA, notably for shallow nodes (e.g., within *Primula* or *Androsace* L. clades), with only 72% of the nodes having a UFBoot ≥95 (Figure 4). All families and genera were retrieved as monophyletic with the exception of *Primula* including *D. frigidum*, consistent with the cpDNA inference. Some sets of species replicates were not monophyletic (*Impatiens parviflora* DC. or *Vaccinium uliginosum* L.). Consistent with the cpDNA tree, *Androsace halleri* was not monophyletic. Taxa from transcriptomic-based libraries also placed as expected but had consistently longer branches than those inferred from genomic-based libraries. We also noted incongruences between cpDNA and mtDNA phylogenies for some low-support nodes (Appendix S5): *Vaccinium* L., some *Primula* subclades or *Soldanella calabrella* Kress.

Concerning the rDNA sequences, two accessions with ≥70% of missing data were removed. For the 116 remaining libraries, all five alignments of sequences were concatenated, resulting in 5,546 bp with 913 informative sites, 5.37% of missing data and 0.06% of outlier sites (Figure 5). The proportion of missing data did not differ significantly between herbarium and silica gel-dried materials (Mann–Whitney $p = .162$). The phylogeny was much less resolved than for both organelle data sets, for both deep and shallow nodes, with only 64% of the nodes displaying a UFBoot ≥95 (Figure 5). As for organelle phylogenies, all genera were monophyletic, except for *Primula* due to the inclusion of *D. frigidum* within the *Primula* clade. Species replicates were consistently placed with the exception of *Vaccinium uliginosum*, and RNAseq samples were also placed as expected. We also noted cytonuclear discordances (i.e., between cpDNA and nuDNA), shown in Appendix S6, either with low support values for the discordant nodes (e.g., relative position of *Daboecia cantabrica* (Huds.) K. Koch; and the *Erica* L. spp. Clade, within Ericaceae) or highly supported tree incongruences, such as the placement of *Soldanella alpina* L. spp. (Appendix S6).

4 | DISCUSSION

ORTHOSKIM offers an efficient method to capture both organelle and ribosomal sequences from genome skimming data sets through a single analysis. By also providing different modes to capture any coding or noncoding target regions, suitable for any eukaryotes and different types of genomic libraries, this pipeline opens new opportunities for phylogenomics and DNA barcoding studies.

4.1 | Sequence capture in genome skimming libraries

ORTHOSKIM streamlines the capturing of all cpDNA and mtDNA genes along with the rDNA regions, using both genomic and transcriptomic libraries. In this benchmark on Ericales samples, the approach displayed particularly high capture rates of the targeting sequences in genome skimming libraries despite very low genome coverage (median ~1.3×, *SD* ~0.8). This can be explained by high coverage obtained for these three regions in genomic assemblies (Appendix S7), as expected in genome skimming libraries. Such coverages depend highly on the sequencing effort of produced libraries (Appendix S8).

Concerning cpDNA targets, ORTHOSKIM performed notably well for shotgun data sets generated from herbarium and silica-dried material, and also from Ericaceae, a family known to display complex plastid structure, making these data difficult for *de novo* organelle assembly (Inger Greve Alsos et al., 2020). Here we captured almost all the cpDNA genes where previous attempts had failed, although some genes were only partially recovered in this family because of the complex nature of their cpDNA. Indeed, we found a higher number of cpDNA contigs assembled for Ericaceae data sets compared to other families (e.g., 13.26 contigs reconstructed on average in Ericaceae vs. 5.55 in Primulaceae, Appendix S4). Such fragmentation can thus explain why lower capture rates

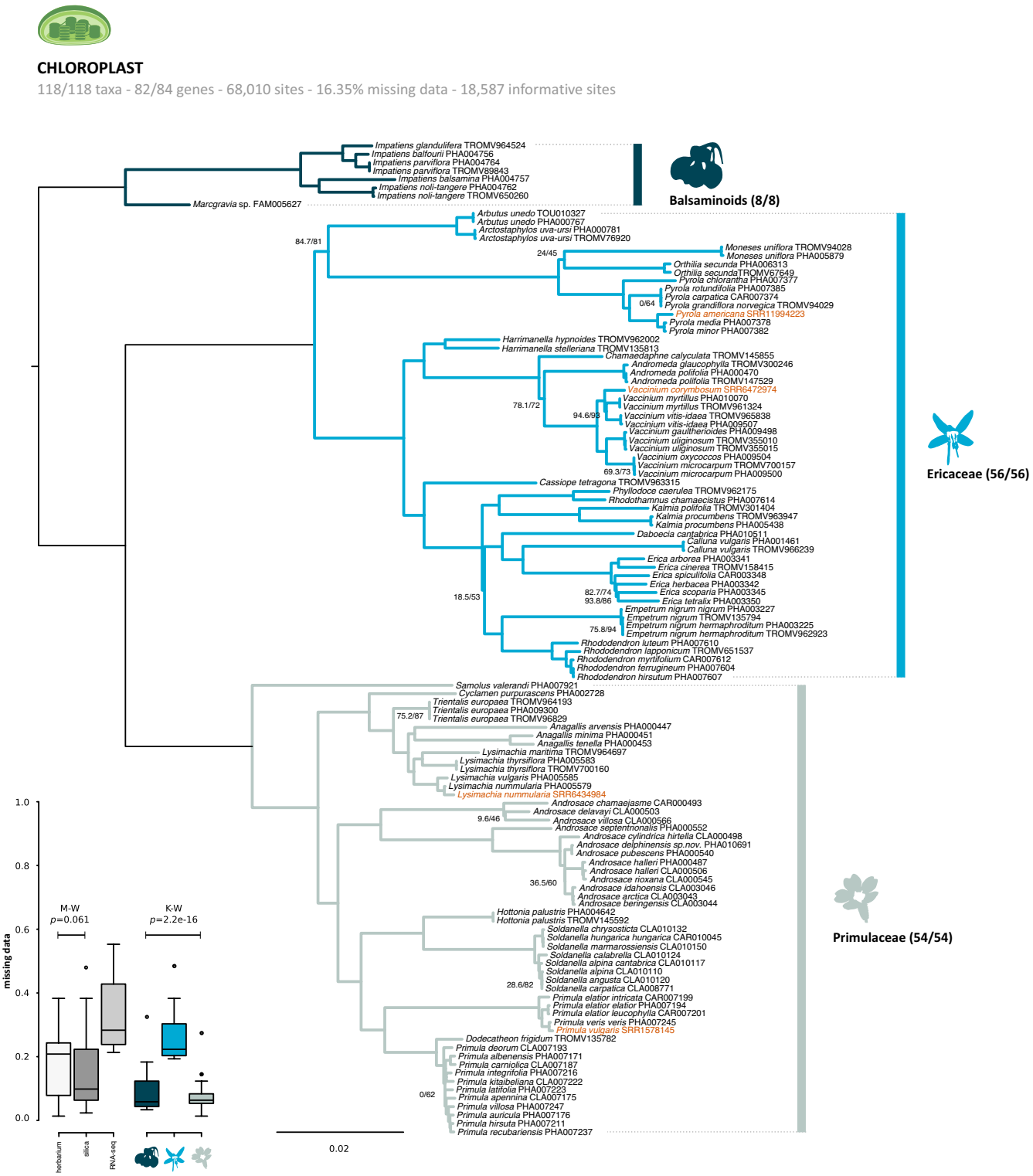

FIGURE 3 Phylogenetic relationships of Ericales families based on the concatenation of 82/84 chloroplast (cpDNA) genes and 118/118 taxa. Phylogenetic support (SH-aLRT/UFBoot) values are given for nodes for UFBoot values <95%. Bottom panels indicate the distribution of missing data according to the origin of the samples (silica-dried or herbarium-preserved materials for genome skimming libraries, or RNAseq libraries), and the clade (Balsaminoids, Ericaceae and Primulaceae). Taxa from transcriptomic-based libraries are highlighted in orange. The chloroplast icon was retrieved from BioRender

were obtained for Ericaceae, as the overall success of cpDNA capture seems to depend primarily on the amount of reconstructed contigs (Appendix S4).

Most of the missing cpDNA genes among the three studied families are known to be nonfunctional in Ericales (Braukmann, Broe, et al., 2017; Liu et al., 2016; Logacheva et al., 2016; Ren

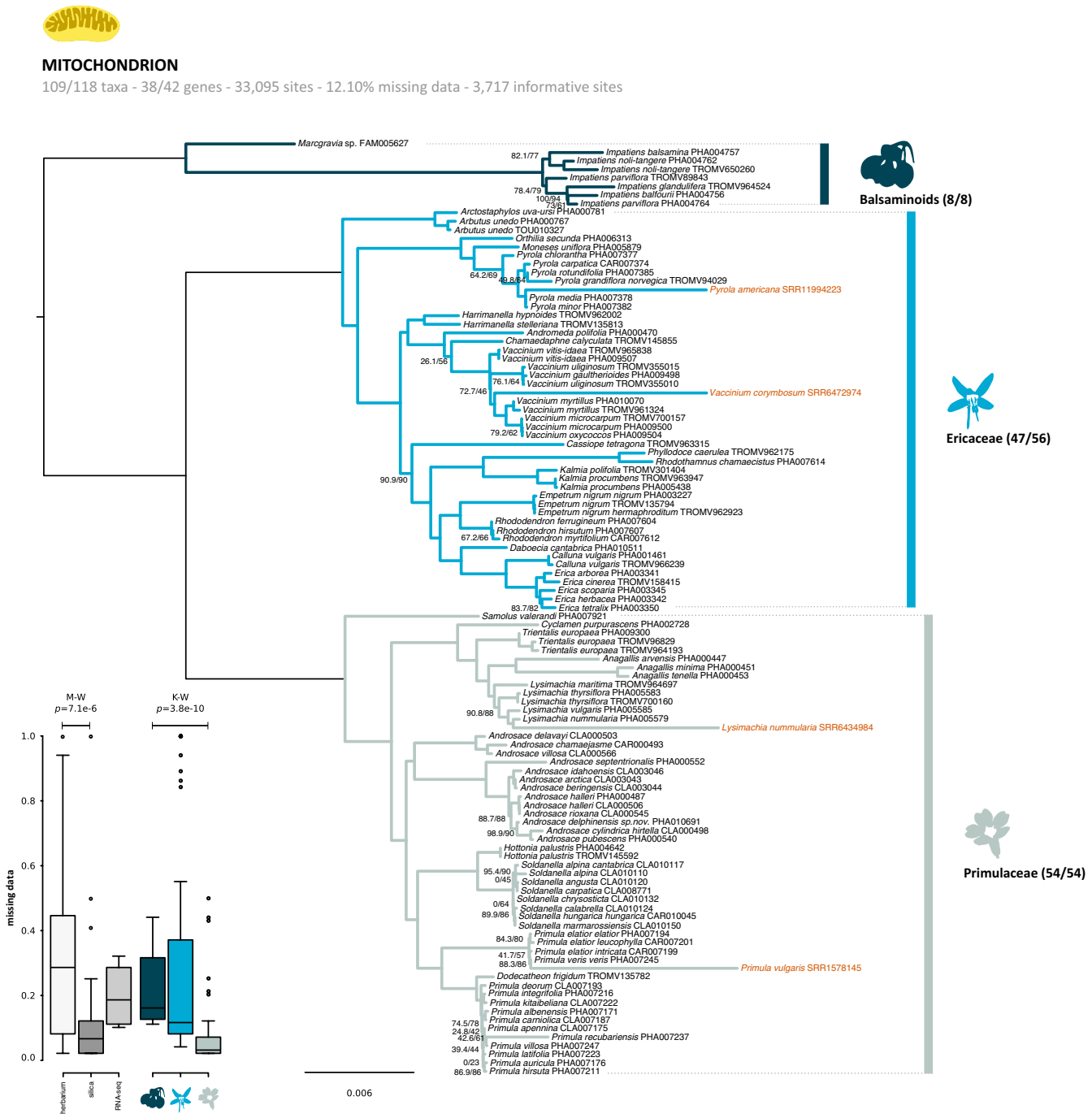

FIGURE 4 Phylogenetic relationships of Ericales families based on the concatenation of 38/42 mitochondrial (mtDNA) genes and 109/118 taxa. Phylogenetic support (SH-aLRT/UFBoot) values are given for UFBoot values <95%. Bottom panels indicate the distribution of missing data according to the origin of the samples (silica-dried or herbarium-preserved materials for genome skimming libraries, or RNAseq libraries), and the clade (Balsaminoids, Ericaceae and Primulaceae). Taxa from transcriptomic-based libraries are highlighted in orange. The mitochondrial icon was retrieved from BioRender

et al., 2018). For example, *ndh* genes (detected but classified in the "gene-like" category for *Orthilla* spp. and *Pyrola* spp.) have been shown to be pseudogenized in *Pyrola rotundifolia* (Logacheva et al., 2016). Probably due to their pseudogenization and size reduction (Braukmann, Broe, et al., 2017; Logacheva et al., 2016; Martínez-Alberola et al., 2013), some of the cpDNA genes were also discarded in Ericaceae, such as *accD* or *ycf1*, as they did not pass the reference coverage threshold. Moreover, both *accD* and *clpP* were shown under pseudogenization or unusual forms in both Ericaceae and Primulaceae (Braukmann, Broe, et al., 2017; Liu et al., 2016; Logacheva et al., 2016). This finding makes sense because these two genes have been shown to present accelerated rates of evolution in multiple independent lineages of Angiosperms, and at least in some cases it has been shown to be due to positive selection instead of

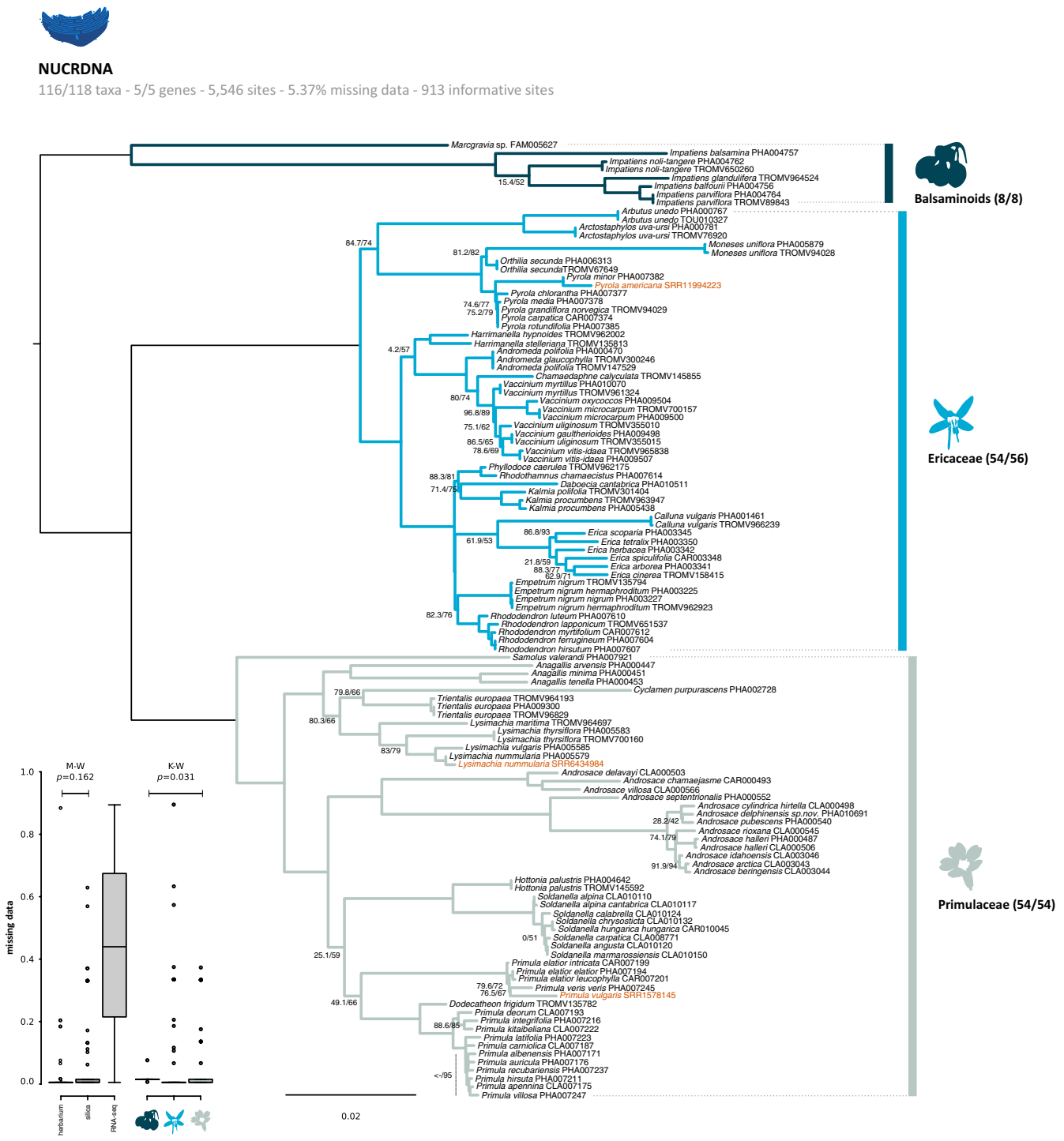

FIGURE 5 Phylogenetic relationships of Ericales families based on the concatenation of 5/5 ribosomal (rDNA) sequences and 116/118 taxa. Phylogenetic support (SH-aLRT/UFBoot) values are given for UFBoot values <95%. Bottom panels indicate the distribution of missing data according to the origin of the samples (silica-dried or herbarium-preserved materials for genome skimming libraries, or RNAseq libraries), and the family (Balsaminoids, Ericaceae and Primulaceae). Taxa from transcriptomic-based libraries are highlighted in orange. The ribosomal icon was retrieved from BioRender

pseudogenization (e.g., Rockenbach et al., 2016). This could explain the ambiguous alignment obtained for these genes, which resulted in their removal from the phylogenetic analysis.

Concerning mtDNA, their use in phylogenomic studies has long been limited in plants owing to their highly dynamic structure and variation in size, high recombination rates, and enrichment in repeated elements and intronic regions (Galtier, 2011; Gualberto et al., 2014). Here, ORTHOSKIM succeeded in capturing mtDNA genes, but with an overall lower success rate than for the other genomic compartments, in particular within Ericaceae. Lower capture rates were also found for herbarium specimens than for silica gel-dried materials, regardless of the age of the herbarium specimens

(Appendix S9). By investigating the assembly statistics, the overall mtDNA contigs were less covered than cpDNA and rDNA contigs (mean coverage of 14.3 vs. 190.2 and 434.5 respectively) for a higher global reconstructed size (mean size of 349,000 bp vs. 127,000 and 7200 bp, Appendix S7). Moreover, the success of mtDNA capture depends on the overall reconstructed size, the amount and the coverage of mtDNA contigs in global assemblies (Appendix S4). The lower capture rates in mtDNA may thus be explained by their underrepresentation in leaf cells and thus in genome skimming data sets in contrast to cpDNA or rDNA (Malé et al., 2014). This is particularly true for herbarium specimens, as overall cpDNA and mtDNA contigs were less covered in shotgun data sets produced from herbarium than from silica gel-dried materials (Appendix S10). The assembly of mtDNA contigs is consequently more sensitive to the coverage threshold set in ORTHOSKIM than for other genomic compartments. For rDNA, the overall success of capture depends on the total reconstructed size of the rDNA contigs (Appendix S4). This can be explained by the restrictions set in ORTHOSKIM to select only contigs including at least the *rrn5.8S* and one of the two other rRNA genes, as intergenic spacers are targeted. Finally, given this low success rate and the limited studies on mtDNA in Ericales and relatives, it remains unclear if the absence of some of mtDNA genes is due to pseudogenization or to the sampling coverage.

On the other hand, ORTHOSKIM appeared efficient for the retrieval of plant DNA barcodes, as traditional plant barcodes (i.e., *matK*, *rbcL*, *ITS1*, *ITS2*, *trnL-UAA*) were nearly always fully captured. ORTHOSKIM was recently used at larger scale to capture both *matK* and *rbcL* in 1815 genome skimming libraries for which these genes were missing following a targeted organelle assembly and annotation approach (Inger Greve Alsos et al., 2020). The capture rate of *matK* was 77% for silica-dried materials and 78% for herbarium materials, and 68%–70% for *rbcL*, when using the assembled annotated cpDNA genome obtained with ORG.ASM (http://metabarcoding.org/org-asm). ORTHOSKIM increased the capture rate to ≥95% for both *matK* and *rbcL* and was particularly efficient for Aspleniaceae, Campanulaceae, Cistaceae, Cyperaceae, Ericaceae, Gentianaceae, Geraniaceae and Juncaceae, which proved challenging to full cpDNA assembly. This thus opens new perspectives in DNA barcoding for other plant families identified as parasitic plants in the literature (Graham et al., 2017), Cupressaceae (Qu et al., 2017) or Geraniaceae (Weng et al., 2014), but also for animals with complex mitochondrial genome structures (Lavrov & Pett, 2016; Valach et al., 2017).

4.2 | Application to phylogenomics

As demonstrated here on two Ericales families, ORTHOSKIM produces aligned DNA sequence matrices suited for phylogenetic inference, using both genomic and transcriptomic libraries directly.

The phylogeny inferred from cpDNA genes was highly robust and concordant with the literature for both deep and shallow nodes among Ericaceae and Primulaceae taxa. In Ericaces, the relationships among genera were similar to those inferred by Rose et al. (2018), with: a basal clade of *Arbutus* L. – *Arctostaphylos* Adans., with a second divergence of *Moneses* Salisb. ex Gray – *Pyrola*. We also found the two main sister clades composed of: *Harrimanella* Coville – *Chamaedaphne* Moench – *Andromeda* L. – *Vaccinium* and one composed of other genera with similar relationships between *Daboecia* D. Don – *Calluna* Salisb. – *Erica* and between *Kalmia* L. – *Rhodothamnus* Rchb. – *Phyllodoce* Salisb. A similar concordance was found among the genera of Primulaceae (Boucher et al., 2016; Rose et al., 2018). Within genera, the relationships estimated among *Androsace* taxa were fully concordant with a previous study (Roquet et al., 2013). Indeed, we recovered a first divergence between a Central-Asian clade (*Androsace chamaejasme* Wulfen ex Host – *A. villosa* L.) and a European–North American clade with: a basal position of *A. septentrionalis* L. and subclades of taxa occurring in the Pyrenees (*A. rioxana* A. Segura – *A. halleri*), southeastern Europe (*A. delphinensis* – *A. pubsecens*) and North America (*A. idahoensis* – *A. beringensis*) mountains. The position of *Dodecatheon frigidum* within *Primula*, recovered in three phylogenies, was consistent with the literature (Mast et al., 2004).

In contrast to plastome data, phylogenies based on mitochondrial and ribosomal data sets yielded weaker and somewhat conflicting signals. Discordances between organellar DNA and rDNA can be explained by biological causes such as incomplete lineage sorting, lateral gene transfers or hybridization events with differential parental inheritance (Govindarajulu et al., 2015; Pouchon et al., 2018; Rice et al., 2013; Smith, 2014; Walker et al., 2017, 2019). In contrast to nuclear inheritance, plant organelles are usually maternally inherited (Corriveau & Coleman, 1988; Schneider et al., 2015; Van de Paer et al., 2016). However, cases of both paternal and biparental transmissions have also been documented in plants (Chybicki et al., 2016; McCauley, 2013; Shen et al., 2015), which could lead to heteroplasmy (i.e., different copies of organellar DNA in cells) and different phylogenetic signal (Sullivan et al., 2017). Such discordances between organellar DNA were detected among Rubiaceae genera (Rydin et al., 2017) or within octoploid *Fragaria* L. (Govindarajulu et al., 2015). However, they seem to at best play a minor role in our study as they concerned shallow nodes with low bootstrap supports on the mtDNA phylogeny (e.g., within *Vaccinium* or *Primula* clades, Appendix S5). Likewise, cytonuclear discordances, which have been shown in Ericales, due to genome duplications (Larson et al., 2020; Vargas et al., 2019), were not obviously present here as most of the conflicting nodes were not fully resolved in the rDNA phylogeny. However, even with low bootstrap support, some of these nodes were consistent with the literature. For example, the phylogenetic relationships inferred for *Soldanella* L., from rDNA, are the same previously reported based on the *ITS1*, *rrn5.8S* and *ITS2* markers (Bellino et al., 2015; Steffen & Kadereit, 2014) but were discordant from cpDNA and mtDNA inferences. The paraphyly retrieved here for *Androsace halleri* on organelle genomes in constrast to rDNA is also convergent with the previously demonstrated reticulate evolution highlighted in this species (Dixon et al., 2007).

Mitochondrial genes (CDS and rRNA genes) appeared useful in inferring deep relationships, as previously shown for Fabaceae (Choi et al., 2019), for Vitaceae (Zhang et al., 2015), for angiosperms (Xue et al., 2020) or here on Ericales. As an example, mtDNA fully supported

the sister clades of *Daboecia*/*Erica* and *Empetrum* L./*Rhododendron* L., unlike with cpDNA. However, the usefulness of mtDNA genes, which were less informative than plastid genes (i.e., 11.23% of informative sites in the mtDNA alignment vs. 27.32% in cpDNA), appeared limited for recent nodes (Van de Paer et al., 2018). As for mtDNA, rDNA genes yielded a lower proportion of informative sites than cpDNA alignments (16.46% of informative sites). Overall, mtDNA and rDNA had a relatively smaller contribution to phylogenetic inference than cpDNA within the families Ericaceae and Primulaceae. Such limitation over mtDNA and rDNA was also evidenced in Bignoniaceae (Fonseca & Lohmann, 2019). However, combining data from several genome compartments can be useful to resolve some problematic taxonomic groups. For example, our phylogenetic analyses placed *Vaccinium gaultherioides* Bigelow within *V. uliginosum* for both mtDNA and rDNA data sets, or as sister species for cpDNA. This supports past conclusions based on cpDNA, rDNA and AFLP markers over the systematic position of *V. gaultherioides* as a prostrate form of *V. uliginosum* (Alsos et al., 2005; Eidesen et al., 2007).

Finally, we obtained mixed results for the use of transcriptomic libraries along with genomic libraries for phylogenomics applications. Although RNAseq libraries seemed correctly positioned on phylogenies, the branch lengths recovered for these taxa differed markedly, in particular for mtDNA. Such a pattern could result from RNA-editing processes in transcriptomic libraries, which mostly concern organelle genomes and C–U changes (Ichinose & Sugita, 2016; Knie et al., 2016). Indeed, by comparing all CDS captured on genomic and transcriptomic libraries of *Lysimachia nummularia*, we found similar patterns of RNA editing as reported for *Arabidopsis thalania* (Chu & Wei, 2019), with ~74% of C–U variants detected and ~89% of them located in mtDNA genes (see Appendix S11). Incorporating such edited sequences in alignments could therefore result in long branch attraction with topological artefacts in the inferred phylogenetic trees.

4.3 | Advantages and perspectives of ORTHOSKIM

The ORTHOSKIM algorithm, presented here in detail, fills an important niche. It provides a generic, flexible and user-friendly tool suitable for both genomic and transcriptomic sequence data, and for a wide range of applications, including phylogeny, DNA barcoding and RNA-editing pattern. ORTHOSKIM is also a scalable tool as its computational burden is reasonable, and allows flexibility to the user by giving a wide range of options during sequence capture, with default parameters suitable for most applications.

Similar approaches have published, such as ATRAM (Allen et al., 2015), HYBPIPER (Johnson et al., 2016) and MITOFINDER (Allio et al., 2020). However, in contrast to these pipelines, ORTHOSKIM is especially designed to work with any type of genome skimming libraries. Indeed, organelle gene transfer is explicitly considered in ORTHOSKIM, such as mtDNA and cpDNA rearrangements, leading to genomic conflict and taxonomic mispositioning (Park et al., 2020; Rice et al., 2013; Wang et al., 2018). For example, *rbcL* copies have frequently been found in angiosperm mtDNA genomes (Park et al., 2020; Van de Paer et al., 2018). Moreover, ORTHOSKIM offers opportunities to capture and align lots of sequences from different genomic compartments in a single analysis, such as cpDNA, mtDNA and rDNA sequences from genome skimming libraries. It can also be applied more broadly to retrieve any targeted sequences including nuclear ones, which enhances its application beyond genome skimming data sets. This contrasts with MITOFINDER, focusing on mtDNA genes or UCE markers, with HYBPIPER for which ITS sequences cannot be directly targeted, or with traditional genome skimming approaches, mostly focusing on cpDNA for plants, and involving organelle assembler, annotators and alignment software (McKain et al., 2018). Furthermore, in contrast to other approaches, we also developed a specific mode to flag possible cpDNA, mtDNA and rDNA contaminants in libraries, which can be problematic when focusing on such regions, when microbiote communities (e.g., fungi or bacteria) are also sequenced and assembled (Chaudhry et al., 2021; Hsiang & Goodwin, 2003; Toju et al., 2019). In our benchmark on Ericales taxa, ORTHOSKIM cleaned genome-skimming libraries from algae cpDNA regions, and mtDNA and rDNA regions of fungi and oomycetes, avoiding misidentification of targeted sequences (Appendix S12). Another advantage of ORTHOSKIM is that it overcomes the issues of organelle assemblers dealing with complex organelle structures as it only focuses on key regions and does not attempt to fully assemble organelle genomes. Nevertheless, contrary to organelle assemblers, we cannot confirm whether a gene sequence is partial or missing, as shown here for *rps7*, *rpl16* or *rps12*, due to issues such as low coverage, fragmented assemblies or biological causes (e.g., pseudogenization, gene or exon loss events, see Xu et al., 2015).

Finally, we propose that two further improvements could be integrated in future releases of ORTHOSKIM. First, sequence homology is currently assessed in ORTHOSKIM by using sequence similarity thresholds, whose efficiency depends on the taxonomic coverage of reference sequences and the genomic coverage of genome skimming data sets. When the reference database is appropriate to the study scale, in particular in terms of taxonomic coverage, the bias of capturing paralogous sequence should be reduced. For example, in our benchmark tests, the homology was well estimated, with less than 0.5% of sites identified as outliers in SPRUCEUP for cpDNA, mtDNA and rDNA targets, which are highly conserved across taxa and well covered in genome skimming libraries. We currently recommend that users check and correct the homology assignment of captured genes, by using external software, such as PREQUAL (Whelan et al., 2018) or SPRUCEUP, as done here. Control steps for checking sequence paralogy in final alignments could be added in further versions of ORTHOSKIM, based on heterozygosity and hypervariable site detection, as recently performed on the PPD pipeline (Zhou et al., 2021). Second, the ability to assemble contigs and capture target sequences in ORTHOSKIM is improved when using higher coverage genomic libraries. For example, higher success rates of capture were obtained here for cpDNA than mtDNA due to a higher coverage of cpDNA contigs than mtDNA contigs in our shotgun data sets (Appendix S7). Moreover, while the very low overall genomic coverage of our benchmark tests (i.e., ~1.3×) was sufficient to retrieve

all cpDNA, mtDNA and rDNA targets, it did not allow us to capture any additional nuclear markers (e.g., UCEs or BUSCO-like). Capturing single copy nuclear sequences should be possible for greater sequencing depth (i.e., >3–10×; Berger et al., 2017; Liu et al., 2021; Zhang, Ding, et al., 2019; Zhang, Jin, et al., 2019). With lower genomic coverage, we recommend using alternative approaches based on read mapping when interested in nuDNA capture from genome skimming data sets (Vargas et al., 2019). Interestingly, Berger et al. (2017) showed that transcriptome assemblers implementing low k-mer values, such as SOAPDENOVO-TRANS (Xie et al., 2014) or TRINITY (Grabherr et al., 2011), performed better than SPADES to capture nuDNA regions, other than rDNA. Other assemblers, as implemented in MITOFINDER, could therefore be integrated in forthcoming versions of ORTHOSKIM to increase our capacity of capturing nuclear sequences in genome skimming libraries even at low coverage (Berger et al., 2017; Vargas et al., 2019; Zhang, Ding, et al., 2019).

5 | COLLABORATORS

Members of the PhyloAlps consortium: S. Lavergne, C. Pouchon, E. Coissac, C. Roquet, J. Smyčka, M. Boleda, W. Thuiller, L. Gielly, P. Taberlet, D. Rioux, F. Boyer, A. Hombiat, B. Bzeznick (Laboratoire d'Ecologie Alpine, CNRS, UGA, Grenoble, FR); A. Alberti, F. Denoeud, P. Wincker, C. Orvain (Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ. Evry, Univ. Paris-Saclay, FR); C. Perrier, R. Douzet, M. Rome, J.G. Valay, S. Aubert (Jardin Alpin du Lautaret, CNRS, UGA, Grenoble, FR); N. Zimmermann, R. O. Wüest, S. Latzin, S. Wipf (Swiss Federal Research Institute WSL, Birmensdorf, CH); J. Van Es, L. Garraud, J.C. Villaret, S. Abdulhak, V. Bonnet, S. Huc, N. Fort, T. Legland, T. Sanz, G. Pache, A. Mikolajczak (Conservatoire Botanique National Alpin, Gap, FR); V. Noble, H. Michaud, B. Offerhaus, M. Pires, Y. Morvant (Conservatoire Botanique National Méditerranéen, Hyères, FR); C. Dentant, P. Salomez, R. Bonet (Parc National des Ecrins, Gap, FR); T. Delahaye (Parc National de la Vanoise, Chambery, FR); M.F. Leccia, M. Perfus (Parc National du Mercantour, Nice, FR); S. Eggenberg, A. Möhl (Info-Flora, Bern, CH); B. Hurdu, M. Pușcaș (Babeș Bolyai University, Institute of Biological Research, Cluj Napoca, RO), M. Slovák (Institute of Botany, Bratislava, SK). Members of the PhyloNorway consortium: I.G. Alsos, M.K. Føreid Merkel, Y. Lammers (The Arctic University Museum of Norway, Tromsø, NO), E. Coissac, C. Pouchon (Laboratoire d'Ecologie Alpine, CNRS, UGA, Grenoble, FR); A. Alberti, F. Denoeud, P. Wincker (Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, FR).

ACKNOWLEDGMENTS

The research, including the salary of C.P., was funded by the joint ANR-SNF project Origin-Alps (ANR-16-CE93-0004, SNF-310030L_170059) and by the ECOGEN project funded by the Research Council of Norway grant 250963/F20. Sequencing was performed within the framework of the PhyloAlps project, funded by France Génomique (ANR-10-INBS-09-08). The PhyloNorway project was funded by the Research Council of Norway (226134/F50) and the Norwegian Biodiversity Information Centre (14-14, 70184209). Bioinformatics and statistical analyses were carried out with the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr). The sampling campaign and preliminary genomic analyses were partly funded by the European Research Council under the European Community's Seventh Framework Programme FP7/2007-2013 grant agreement 281422 (TEEMBIO). The LECA is part of Labex OSUG (ANR10 LABX56). The EDB is part of Labex CEBA (ANR-10-LABX-25-01) and Labex TULIP (ANR-10-LABX-0041). The PhyloAlps consortium also thanks a large network of field botanists who have contributed to sampling.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

CP, FB and SL designed the study and wrote the manuscript; CP developed ORTHOSKIM and analysed the data; sample collection, molecular laboratory work, and final editing of the manuscript were jointly performed by members of the PhyloAlps and the PhyloNorway projects.

DATA AVAILABILITY STATEMENT

ORTHOSKIM software is available at https://github.com/cpouchon/ORTHOSKIM. Genome skimming sequence data from the PhyloAlps and PhyloNorway projects have been deposited at the European Nucleotide Archive (PRJEB43865, PRJEB48693 and PRJEB48874).

ORCID

Charles Pouchon https://orcid.org/0000-0001-7766-3732
Cristina Roquet https://orcid.org/0000-0001-8748-3743
Eric Coissac https://orcid.org/0000-0001-7507-6729
Sébastien Lavergne https://orcid.org/0000-0001-8842-7495

REFERENCES

Allen, J. M., Huang, D. I., Cronk, Q. C., & Johnson, K. P. (2015). aTRAM – Automated target restricted assembly method: A fast method for assembling loci across divergent taxa from next-generation sequencing data. *BMC Bioinformatics*, *16*(1), 98. https://doi.org/10.1186/s12859-015-0515-2

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2020). MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Molecular Ecology Resources*, *20*(4), 892–905. https://doi.org/10.1111/1755-0998.13160

Alsos, I. G., Engelskjøn, T., Gielly, L., Taberlet, P., & Brochmann, C. (2005). Impact of ice ages on circumpolar molecular diversity: Insights from an ecological key species. *Molecular Ecology*, *14*(9), 2739–2753. https://doi.org/10.1111/j.1365-294X.2005.02621.x

Alsos, I. G., Lavergne, S., Merkel, M. K. F., Boleda, M., Lammers, Y., Alberti, A., Pouchon, C., Denoeud, F., Pitelkova, I., Pușcaș, 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*(4), 432. https://doi.org/10.3390/plants9040432

Andermann, T., Torres Jiménez, M. F., Matos-Maraví, P., Batista, R., Blanco-Pastor, J. L., Gustafsson, A. L. S., Kistler, L., Liberal, I. M., Oxelman, B., Bacon, C. D., & Antonelli, A. (2020). A guide to carrying out a phylogenomic target sequence capture project. *Frontiers in Genetics*, *10*, 1407. https://doi.org/10.3389/fgene.2019.01407

Bakker, F. T., Lei, D. I., 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. *Biological Journal of the Linnean Society*, *117*(1), 33–43. https://doi.org/10.1111/bij.12642

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. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barrett, C. F., Baker, W. J., Comer, J. R., Conran, J. G., Lahmeyer, S. C., Leebens-Mack, J. H., Li, J., Lim, G. S., Mayfield-Jones, D. R., Perez, L., Medina, J., Pires, J. C., Santos, C., Wm. Stevenson, D., Zomlefer, W. B., & Davis, J. I. (2016). Plastid genomes reveal support for deep phylogenetic relationships and extensive rate variation among palms and other commelinid monocots. *New Phytologist*, *209*(2), 855–870. https://doi.org/10.1111/nph.13617

Bellino, A., Bellino, L., Baldantoni, D., & Saracino, A. (2015). Evolution, ecology and systematics of *Soldanella* (Primulaceae) in the southern Apennines (Italy). *BMC Evolutionary Biology*, *15*, 158. https://doi.org/10.1186/s12862-015-0433-y

Bendich, A. J. (2004). Circular chloroplast chromosomes: The grand illusion. *The Plant Cell*, *16*(7), 1661–1666. https://doi.org/10.1105/tpc.160771

Berger, B. A., Han, J., Sessa, E. B., Gardner, A. G., Shepherd, K. A., Ricigliano, V. A., & Howarth, D. G. (2017). The unexpected depths of genome-skimming data: A case study examining Goodeniaceae floral symmetry genes. *Applications in Plant Sciences*, *5*(10), 1700042. https://doi.org/10.3732/apps.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. *Molecular Ecology*, *29*(14), 2521–2534. https://doi.org/10.1111/mec.15507

Borowiec, M. L. (2019). Spruceup: Fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. *Journal of Open Source Software*, *4*(42), 1635. https://doi.org/10.21105/joss.01635

Boucher, F. C., Zimmermann, N. E., & Conti, E. (2016). Allopatric speciation with little niche divergence is common among alpine Primulaceae. *Journal of Biogeography*, *43*(3), 591–602. https://doi.org/10.1111/jbi.12652

Braukmann, T. W. A., Broe, M. B., Stefanović, S., & Freudenstein, J. V. (2017). On the brink: The highly reduced plastomes of nonphotosynthetic Ericaceae. *New Phytologist*, *216*(1), 254–266. https://doi.org/10.1111/nph.14681

Braukmann, T. W. A., Kuzmina, M. L., Sills, J., Zakharov, E. V., & Hebert, P. D. N. (2017). Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. *PLoS One*, *12*(1), e0169515. https://doi.org/10.1371/journal.pone.0169515

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

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 Research*, *41*(Database issue), D226–D232. https://doi.org/10.1093/nar/gks1005

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

CBOL Plant Working Group, Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K.-J., Kress, W. J., Schneider, H., van AlphenStahl, J., Barrett, S. C., van den Berg, C., Bogarin, D., … Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106

Chaudhry, V., Runge, P., Sengupta, P., Doehlemann, G., Parker, J. E., & Kemen, E. (2021). Shaping the leaf microbiota: Plant–microbe–microbe interactions. *Journal of Experimental Botany*, *72*(1), 36–56. https://doi.org/10.1093/jxb/eraa417

Choi, I.-S., Schwarz, E. N., Ruhlman, T. A., Khiyami, M. A., Sabir, J. S. M., Hajarah, N. H., Sabir, M. J., Rabah, S. O., & Jansen, R. K. (2019). Fluctuations in Fabaceae mitochondrial genome size and content are both ancient and recent. *BMC Plant Biology*, *19*(1), 448. https://doi.org/10.1186/s12870-019-2064-8

Chu, D., & Wei, L. (2019). The chloroplast and mitochondrial C-to-U RNA editing in *Arabidopsis thaliana* shows signals of adaptation. *Plant Direct*, *3*(9), e00169. https://doi.org/10.1002/pld3.169

Chybicki, I. J., Dering, M., Iszkuło, G., Meyza, K., & Suszka, J. (2016). Relative strength of fine-scale spatial genetic structure in paternally vs. biparentally inherited DNA in a dioecious plant depends on both sex proportions and pollen-to-seed dispersal ratio. *Heredity*, *117*(6), 449–459. https://doi.org/10.1038/hdy.2016.65

Clarke, C. L., Edwards, M. E., Gielly, L., Ehrich, D., Hughes, P. D. M., Morozova, L. M., Haflidason, H., Mangerud, J., Svendsen, J. I., & Alsos, I. G. (2019). Persistence of arctic-alpine flora during 24,000 years of environmental change in the Polar Urals. *Scientific Reports*, *9*(1), 1–11. https://doi.org/10.1038/s41598-019-55989-9

Coissac, E., Hollingsworth, P. M., Lavergne, S., & Taberlet, P. (2016). From barcodes to genomes: Extending the concept of DNA barcoding. *Molecular Ecology*, *25*(7), 1423–1428. https://doi.org/10.1111/mec.13549

Corriveau, J. L., & Coleman, A. W. (1988). Rapid screening method to detect potential biparental inheritance of plastid DNA and results for over 200 angiosperm species. *American Journal of Botany*, *75*(10), 1443–1458. https://doi.org/10.1002/j.1537-2197.1988.tb11219.x

Dierckxsens, N., Mardulyn, P., & Smits, G. (2016). NOVOPlasty: *De novo* assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), e18. https://doi.org/10.1093/nar/gkw955

Dixon, C. J., Schönswetter, P., & Schneeweiss, G. M. (2007). Traces of ancient range shifts in a mountain plant group (*Androsace halleri* complex, Primulaceae). *Molecular Ecology*, *16*(18), 3890–3901. https://doi.org/10.1111/j.1365-294X.2007.03342.x

Eidesen, P. B., Alsos, I. G., Popp, M., Stensrud, Ø., Suda, J., & Brochmann, C. (2007). Nuclear vs. plastid data: Complex Pleistocene history of a circumpolar key species. *Molecular Ecology*, *16*(18), 3902–3925. https://doi.org/10.1111/j.1365-294X.2007.03425.x

Fahner, N. A., Shokralla, S., Baird, D. J., & Hajibabaei, M. (2016). Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. *PLoS One*, *11*(6), e0157505. https://doi.org/10.1371/journal.pone.0157505

Fonseca, L. H. M., & Lohmann, L. G. (2019). Exploring the potential of nuclear and mitochondrial sequencing data generated through genome-skimming for plant phylogenetics: A case study from a clade of neotropical lianas. *Journal of Systematics and Evolution*, *58*(1), 18–32. https://doi.org/10.1111/jse.12533

Freudenthal, J. A., Pfaff, S., Terhoeven, N., Korte, A., Ankenbrand, M. J., & Förster, F. (2020). A systematic comparison of chloroplast genome assembly tools. *Genome Biology*, *21*(1), 254. https://doi.org/10.1186/s13059-020-02153-6

Galtier, N. (2011). The intriguing evolutionary dynamics of plant mitochondrial DNA. *BMC Biology*, *9*(1), 61. https://doi.org/10.1186/1741-7007-9-61

Gandini, C. L., & Sanchez-Puerta, M. V. (2017). Foreign plastid sequences in plant mitochondria are frequently acquired via mitochondrion-to-mitochondrion horizontal transfer. *Scientific Reports*, *7*, 43402. https://doi.org/10.1038/srep43402

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. *American Journal of Botany*, *105*(11), 1888–1910. https://doi.org/10.1002/ajb2.1178

Govindarajulu, R., Parks, M., Tennessen, J. A., Liston, A., & Ashman, T.-L. (2015). Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. *American Journal of Botany*, *102*(4), 544–554. https://doi.org/10.3732/ajb.1500026

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., & Regev, A. (2011). Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Graham, S. W., Lam, V. K. Y., & Merckx, V. S. F. T. (2017). Plastomes on the edge: The evolutionary breakdown of mycoheterotroph plastid genomes. *New Phytologist*, *214*(1), 48–55. https://doi.org/10.1111/nph.14398

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. *Zoologica Scripta*, *46*(6), 718–728. https://doi.org/10.1111/zsc.12247

Greshake, B., Zehr, S., Dal Grande, F., Meiser, A., Schmitt, I., & Ebersberger, I. (2016). Potential and pitfalls of eukaryotic metagenome skimming: A test case for lichens. *Molecular Ecology Resources*, *16*(2), 511–523. https://doi.org/10.1111/1755-0998.12463

Gualberto, J. M., Mileshina, D., Wallet, C., Niazi, A. K., Weber-Lotfi, F., & Dietrich, A. (2014). The plant mitochondrial genome: Dynamics and maintenance. *Biochimie*, *100*, 107–120. https://doi.org/10.1016/j.biochi.2013.09.016

Hsiang, T., & Goodwin, P. H. (2003). Distinguishing plant and fungal sequences in ESTs from infected plant tissues. *Journal of Microbiological Methods*, *54*(3), 339–351. https://doi.org/10.1016/S0167-7012(03)00067-8

Ichinose, M., & Sugita, M. (2016). RNA editing and its molecular mechanism in plant organelles. *Genes*, *8*(1), 5. https://doi.org/10.3390/genes8010005

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., & Li, D.-Z. (2020). GetOrganelle: A fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biology*, *21*(1), 241. https://doi.org/10.1186/s13059-020-02154-5

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. C., & Wickett, N. J. (2016). HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, *4*(7), 1600016. https://doi.org/10.3732/apps.1600016

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.-S., 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. *Systematic Biology*, *68*(4), 594–606. https://doi.org/10.1093/sysbio/syy086

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kennedy, S. R., Prost, S., Overcast, I., Rominger, A. J., Gillespie, R. G., & Krehenwinkel, H. (2020). High-throughput sequencing for community analysis: The promise of DNA barcoding to uncover diversity, relatedness, abundances and interactions in spider communities. *Development Genes and Evolution*, *230*(2), 185–201. https://doi.org/10.1007/s00427-020-00652-x

Knie, N., Grewe, F., Fischer, S., & Knoop, V. (2016). Reverse U-to-C editing exceeds C-to-U RNA editing in some ferns – A monilophyte-wide comparison of chloroplast and mitochondrial RNA editing suggests independent evolution of the two processes in both organelles. *BMC Evolutionary Biology*, *16*, 134. https://doi.org/10.1186/s12862-016-0707-z

Koenen, E. J. M., Kidner, C., Souza, É. R., Simon, M. F., Iganci, J. R., Nicholls, J. A., Brown, G. K., Queiroz, L. P., Luckow, M., Lewis, G. P., Pennington, R. T., & Hughes, C. E. (2020). Hybrid capture of 964 nuclear genes resolves evolutionary relationships in the mimosoid legumes and reveals the polytomous origins of a large pantropical radiation. *American Journal of Botany*, *107*(12), 1710–1735. https://doi.org/10.1002/ajb2.1568

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics (Oxford, England)*, *28*(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Kozik, A., Rowan, B. A., Lavelle, D., Berke, L., Schranz, M. E., Michelmore, R. W., & Christensen, A. C. (2019). The alternative reality of plant mitochondrial DNA: One ring does not rule them all. *PLoS Genetics*, *15*(8), e1008373. https://doi.org/10.1371/journal.pgen.1008373

Larson, D. A., Walker, J. F., Vargas, O. M., & Smith, S. A. (2020). A consensus phylogenomic approach highlights paleopolyploid and rapid radiation in the history of Ericales. *American Journal of Botany*, *107*(5), 773–789. https://doi.org/10.1002/ajb2.1469

Lavrov, D. V., & Pett, W. (2016). Animal Mitochondrial DNA as we do not know it: Mt-genome organization and evolution in nonbilaterian lineages. *Genome Biology and Evolution*, *8*(9), 2896–2913. https://doi.org/10.1093/gbe/evw195

Li, H.-T., Yi, T.-S., Gao, L.-M., Ma, P.-F., Zhang, T., Yang, J.-B., Gitzendanner, M. A., Fritsch, P. W., Cai, J., Luo, Y., Wang, H., van der Bank, M., Zhang, S.-D., Wang, Q.-F., Wang, J., Zhang, Z.-R., Fu, C.-N., Yang, J., Hollingsworth, P. M., … Li, D.-Z. (2019). Origin of angiosperms and the puzzle of the Jurassic gap. *Nature Plants*, *5*(5), 461–470. https://doi.org/10.1038/s41477-019-0421-0

Linard, B., Crampton-Platt, A., Gillett, C. P. D. T., Timmermans, M. J. T. N., & Vogler, A. P. (2015). Metagenome skimming of insect specimen pools: Potential for comparative genomics. *Genome Biology and Evolution*, *7*(6), 1474–1489. https://doi.org/10.1093/gbe/evv086

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. *Journal of Systematics and Evolution*, *59*(5), 1124–1138. https://doi.org/10.1111/jse.12806

Liu, S.-H., Edwards, C. E., Hoch, P. C., Raven, P. H., & Barber, J. C. (2018). Genome skimming provides new insight into the relationships in Ludwigia section Macrocarpon, a polyploid complex. *American*

Journal of Botany, *105*(5), 875–887. https://doi.org/10.1002/ajb2.1086

Liu, T.-J., Zhang, C.-Y., Yan, H.-F., Zhang, L., Ge, X.-J., & Hao, G. (2016). Complete plastid genome sequence of *Primula sinensis* (Primulaceae): Structure comparison, sequence variation and evidence for *accD* transfer to nucleus. *PeerJ*, *4*, e2101. https://doi.org/10.7717/peerj.2101

Logacheva, M. D., Schelkunov, M. I., Shtratnikova, V. Y., Matveeva, M. V., & Penin, A. A. (2016). Comparative analysis of plastid genomes of nonphotosynthetic Ericaceae and their photosynthetic relatives. *Scientific Reports*, *6*, 30042. https://doi.org/10.1038/srep30042

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., 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., … Schleifer, K.-H. (2004). ARB: A software environment for sequence data. *Nucleic Acids Research*, *32*(4), 1363–1371. https://doi.org/10.1093/nar/gkh293

Malé, P.-J., 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. *Molecular Ecology Resources*, *14*(5), 966–975. https://doi.org/10.1111/1755-0998.12246

Mao, Y., Hou, S., Shi, J., & Economo, E. P. (2020). TREEasy: An automated workflow to infer gene trees, species trees, and phylogenetic networks from multilocus data. *Molecular Ecology Resources*, *20*(3), 832–840. https://doi.org/10.1111/1755-0998.13149

Martínez-Alberola, F., del Campo, E. M., Lázaro-Gimeno, D., Mezquita-Claramonte, S., Molins, A., Mateu-Andrés, I., Pedrola-Monfort, J., Casano, L. M., & Barreno, E. (2013). Balanced gene losses, duplications and intensive rearrangements led to an unusual regularly sized genome in *Arbutus unedo* chloroplasts. *PLoS One*, *8*(11), e79685. https://doi.org/10.1371/journal.pone.0079685

Mast, A. R., Feller, D. M. S., Kelso, S., & Conti, E. (2004). Buzz-pollinated *Dodecatheon* originated from within the heterostylous *Primula* subgenus *Auriculastrum* (Primulaceae): A seven-region cpDNA phylogeny and its implications for floral evolution. *American Journal of Botany*, *91*(6), 926–942. https://doi.org/10.3732/ajb.91.6.926

McCauley, D. E. (2013). Paternal leakage, heteroplasmy, and the evolution of plant mitochondrial genomes. *The New Phytologist*, *200*(4), 966–977. https://doi.org/10.1111/nph.12431

McKain, M. R., Johnson, M. G., Uribe-Convers, S., Eaton, D., & Yang, Y. (2018). Practical considerations for plant phylogenomics. *Applications in Plant Sciences*, *6*(3), e1038. https://doi.org/10.1002/aps3.1038

Meiser, A., Otte, J., Schmitt, I., & Grande, F. D. (2017). Sequencing genomes from mixed DNA samples—Evaluating the metagenome skimming approach in lichenized fungi. *Scientific Reports*, *7*(1), 14881. https://doi.org/10.1038/s41598-017-14576-6

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. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Nevill, P. G., Zhong, X., Tonti-Filippini, J., Byrne, M., Hislop, M., Thiele, K., van Leeuwen, S., Boykin, L. M., & Small, I. (2020). Large scale genome skimming from herbarium material for accurate plant identification and phylogenomics. *Plant Methods*, *16*(1), 1. https://doi.org/10.1186/s13007-019-0534-5

Omelchenko, D. O., Makarenko, M. S., Kasianov, A. S., Schelkunov, M. I., Logacheva, M. D., & Penin, A. A. (2020). Assembly and analysis of the complete mitochondrial genome of *Capsella bursa-pastoris*. *Plants*, *9*(4), 469. https://doi.org/10.3390/plants9040469

Palumbo, F., Vitulo, N., Vannozzi, A., Magon, G., & Barcaccia, G. (2020). The mitochondrial genome assembly of fennel (*Foeniculum vulgare*) reveals two different atp6 gene sequences in cytoplasmic male sterile accessions. *International Journal of Molecular Sciences*, *21*(13), https://doi.org/10.3390/ijms21134664

Park, H.-S., Jayakodi, M., Lee, S. H., Jeon, J.-H., Lee, H.-O., Park, J. Y., Moon, B. C., Kim, C.-K., Wing, R. A., Newmaster, S. G., Kim, J. Y., & Yang, T.-J. (2020). Mitochondrial plastid DNA can cause DNA barcoding paradox in plants. *Scientific Reports*, *10*(1), 6112. https://doi.org/10.1038/s41598-020-63233-y

Plomion, C., Aury, J. M., Amselem, J., Leroy, T., Murat, F., Duplessis, S., Salse, J. (2018). Oak genome reveals facets of long lifespan. *Molecular Ecology*, *4*(7), 440–452. https://doi.org/10.1038/s41477-018-0172-3

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338. https://doi.org/10.1111/mec.14478

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. *Systematic Biology*, *67*(6), 1041–1060. https://doi.org/10.1093/sysbio/syy022

Qu, X.-J., Wu, C.-S., Chaw, S.-M., & Yi, T.-S. (2017). Insights into the existence of isomeric plastomes in Cupressoideae (Cupressaceae). *Genome Biology and Evolution*, *9*(4), 1110–1119. https://doi.org/10.1093/gbe/evx071

Ren, T., Yang, Y., Zhou, T., & Liu, Z.-L. (2018). Comparative plastid genomes of primula species: Sequence divergence and phylogenetic relationships. *International Journal of Molecular Sciences*, *19*(4), 1050. https://doi.org/10.3390/ijms19041050

Rice, D. W., Alverson, A. J., Richardson, A. O., Young, G. J., Sanchez-Puerta, M. V., Munzinger, J., & Palmer, J. D. (2013). Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. *Science (New York, N.Y.)*, *342*(6165), 1468–1473. https://doi.org/10.1126/science.1246275

Rockenbach, K., Havird, J. C., Monroe, J. G., Triant, D. A., Taylor, D. R., & Sloan, D. B. (2016). Positive selection in rapidly evolving plastid-nuclear enzyme complexes. *Genetics*, *204*(4), 1507–1522. https://doi.org/10.1534/genetics.116.188268

Roquet, C., Boucher, F. C., Thuiller, W., & Lavergne, S. (2013). Replicated radiations of the alpine genus *Androsace* (Primulaceae) driven by range expansion and convergent key innovations. *Journal of Biogeography*, *40*(10), 1874–1886. https://doi.org/10.1111/jbi.12135

Rose, J. P., Kleist, T. J., Löfstrand, S. D., Drew, B. T., Schönenberger, J., & Sytsma, K. J. (2018). Phylogeny, historical biogeography, and diversification of angiosperm order Ericales suggest ancient Neotropical and East Asian connections. *Molecular Phylogenetics and Evolution*, *122*, 59–79. https://doi.org/10.1016/j.ympev.2018.01.014

Rydin, C., Wikström, N., & Bremer, B. (2017). Conflicting results from mitochondrial genomic data challenge current views of Rubiaceae phylogeny. *American Journal of Botany*, *104*(10), 1522–1532. https://doi.org/10.3732/ajb.1700255

Schneider, A., Stelljes, C., Adams, C., Kirchner, S., Burkhard, G., Jarzombski, S., Broer, I., Horn, P., Elsayed, A., Hagl, P., Leister, D., & Koop, H.-U. (2015). Low frequency paternal transmission of plastid genes in Brassicaceae. *Transgenic Research*, *24*(2), 267–277. https://doi.org/10.1007/s11248-014-9842-8

Schönenberger, J., Anderberg, A. A., & Sytsma, K. J. (2015). Molecular phylogenetics and patterns of floral evolution in the Ericales. *International Journal of Plant Sciences*, *166*(2), 265–288. https://doi.org/10.1086/427198

Shen, J., Zhao, J., Bartoszewski, G., Malepszy, S., Havey, M., & Chen, J. (2015). Persistence and protection of mitochondrial DNA in the generative cell of cucumber is consistent with its paternal transmission. *Plant & Cell Physiology*, *56*(11), 2271–2282. https://doi.org/10.1093/pcp/pcv140

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation

completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

Sloan, D. B., & Wu, Z. (2014). History of plastid DNA insertions reveals weak deletion and AT mutation biases in angiosperm mitochondrial genomes. *Genome Biology and Evolution*, *6*(12), 3210–3221. https://doi.org/10.1093/gbe/evu253

Smith, D. R. (2014). Mitochondrion-to-plastid DNA transfer: It happens. *New Phytologist*, *202*(3), 736–738. https://doi.org/10.1111/nph.12704

Steffen, S., & Kadereit, J. W. (2014). Parallel evolution of flower reduction in two alpine *Soldanella* species (Primulaceae). *Botanical Journal of the Linnean Society*, *175*(3), 409–422. https://doi.org/10.1111/boj.12174

Straub, S. C. K., Cronn, R. C., Edwards, C., Fishbein, M., & Liston, A. (2013). Horizontal transfer of DNA from the mitochondrial to the plastid genome and its subsequent evolution in milkweeds (Apocynaceae). *Genome Biology and Evolution*, *5*(10), 1872–1885. https://doi.org/10.1093/gbe/evt140

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. *American Journal of Botany*, *99*(2), 349–364. https://doi.org/10.3732/ajb.1100335

Sullivan, A. R., Schiffthaler, B., Thompson, S. L., Street, N. R., & Wang, X.-R. (2017). Interspecific plastome recombination reflects ancient reticulate evolution in *Picea* (Pinaceae). *Molecular Biology and Evolution*, *34*(7), 1689–1701. https://doi.org/10.1093/molbev/msx111

Toju, H., Okayasu, K., & Notaguchi, M. (2019). Leaf-associated microbiomes of grafted tomato plants. *Scientific Reports*, *9*(1), 1787. https://doi.org/10.1038/s41598-018-38344-2

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. https://doi.org/10.7717/peerj.7543

Twyford, A. D., & Ness, R. W. (2017). Strategies for complete plastid genome sequencing. *Molecular Ecology Resources*, *17*(5), 858–868. https://doi.org/10.1111/1755-0998.12626

Tyagi, K., Kumar, V., Kundu, S., Pakrashi, A., Prasad, P., Caleb, J. T. D., & Chandra, K. (2019). Identification of Indian Spiders through DNA barcoding: Cryptic species and species complex. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-50510-8

Valach, M., Moreira, S., Hoffmann, S., Stadler, P. F., & Burger, G. (2017). Keeping it complicated: Mitochondrial genome plasticity across diplonemids. *Scientific Reports*, *7*(1), 14166. https://doi.org/10.1038/s41598-017-14286-z

Van de Paer, C., Bouchez, O., & Besnard, G. (2018). Prospects on the evolutionary mitogenomics of plants: A case study on the olive family (Oleaceae). *Molecular Ecology Resources*, *18*(3), 407–423. https://doi.org/10.1111/1755-0998.12742

Van de Paer, C., Hong-Wa, C., Jeziorski, C., & Besnard, G. (2016). Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. *Gene*, *594*(2), 197–202. https://doi.org/10.1016/j.gene.2016.09.007

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. *Molecular Phylogenetics and Evolution*, *135*, 98–104. https://doi.org/10.1016/j.ympev.2019.02.020

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*). *The New Phytologist*, *214*(4), 1736–1750. https://doi.org/10.1111/nph.14530

Walker, J. F., Walker-Hale, N., Vargas, O. M., Larson, D. A., & Stull, G. W. (2019). Characterizing gene tree conflict in plastome-inferred phylogenies. *PeerJ*, *7*, e7747. https://doi.org/10.7717/peerj.7747

Walker, J. F., Yang, Y., Moore, M. J., Mikenas, J., Timoneda, A., Brockington, S. F., & Smith, S. A. (2017). Widespread paleopolyploidy, gene tree conflict, and recalcitrant relationships among the carnivorous Caryophyllales. *American Journal of Botany*, *104*(6), 858–867. https://doi.org/10.3732/ajb.1700083

Wang, X.-C., Chen, H., Yang, D., & Liu, C. (2018). Diversity of mitochondrial plastid DNAs (MTPTs) in seed plants. *Mitochondrial DNA. Part A, DNA Mapping, Sequencing, and Analysis*, *29*(4), 635–642. https://doi.org/10.1080/24701394.2017.1334772

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, S. A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., … Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, *678*, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Weng, M.-L., Blazier, J. C., Govindu, M., & Jansen, R. K. (2014). Reconstruction of the ancestral plastid genome in Geraniaceae reveals a correlation between genome rearrangements, repeats, and nucleotide substitution rates. *Molecular Biology and Evolution*, *31*(3), 645–659. https://doi.org/10.1093/molbev/mst257

Whelan, S., Irisarri, I., & Burki, F. (2018). PREQUAL: Detecting nonhomologous characters in sets of unaligned homologous sequences. *Bioinformatics*, *34*(22), 3929–3930. https://doi.org/10.1093/bioinformatics/bty448

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*(12), 1660–1666. https://doi.org/10.1093/bioinformatics/btu077

Xu, J.-H., Liu, Q., Hu, W., Wang, T., Xue, Q., & Messing, J. (2015). Dynamics of chloroplast genomes in green plants. *Genomics*, *106*(4), 221–231. https://doi.org/10.1016/j.ygeno.2015.07.004

Xue, J.-Y., Dong, S.-S., Wang, M.-Q., Song, T.-Q., Zhou, G.-C., Li, Z., Hang, Y.-Y. (2020). Mitochondrial genes from 18 angiosperms fill sampling gaps for phylogenomic inferences of the early diversification of flowering plants. *Journal of Systematics and Evolution*, https://doi.org/10.1111/jse.12708

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, C., Zhang, T., Luebert, F., Xiang, Y., Huang, C.-H., Hu, Y. I., Rees, M., Frohlich, M. W., Qi, J. I., Weigend, M., & Ma, H. (2020). Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications. *Molecular Biology and Evolution*, *37*(11), 3188–3210. https://doi.org/10.1093/molbev/msaa160

Zhang, F., Ding, Y., Zhu, C., Zhou, X., Orr, M. C., Scheu, S., & Luan, Y. (2019). Phylogenomics from low-coverage whole-genome sequencing. *Methods in Ecology and Evolution*, *10*(4), 507–517. https://doi.org/10.1111/2041-210X.13145

Zhang, N., Wen, J., & Zimmer, E. A. (2015). Congruent deep relationships in the grape family (Vitaceae) based on sequences of chloroplast

genomes and mitochondrial genes via genome skimming. *PLoS One*, *10*(12), e0144701. https://doi.org/10.1371/journal.pone.0144701

Zhang, R., Jin, J.-J., Moore, M. J., & Yi, T.-S. (2019). Assembly and comparative analyses of the mitochondrial genome of Castanospermum australe (Papilionoideae, Leguminosae). *Australian Systematic Botany*. *32*(6), 484–494. https://doi.org/10.1071/SB19014

Zhou, W., Soghigian, J., & Xiang, Q.-Y. (2021). A new pipeline for removing paralogs in target enrichment data. *Systematic Biology*. *syab044*. https://doi.org/10.1093/sysbio/syab044

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Pouchon, C., Boyer, F., Roquet, C., Denoeud, F., Chave, J., Coissac, E., Alsos, I. G., & Lavergne, S.; The PhyloAlps Consortium; The PhyloNorway Consortium (2022). ORTHOSKIM: In silico sequence capture from genomic and transcriptomic libraries for phylogenomic and barcoding applications. *Molecular Ecology Resources*, 22, 2018–2037. https://doi.org/10.1111/1755-0998.13584