UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae)

Cardueae Radiations Grp

2018-11
Academic Press Inc.
http://hdl.handle.net/10138/311395
Cardueae Radiations Grp 2018, 'Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae)', Molecular Phylogenetics and Evolution, vol. 128, pp. 69-87. https://doi.org/10.1016/j.ympev.2018.07.012

Downloaded from Helda, University of Helsinki institutional repository. https://helda.helsinki.fi
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.
Please cite the original version.

Exploring data processing strategies in NGS target enrichment to disentangle radiations in the tribe Cardueae (Compositae)

Sonia Herrando-Moraira[a,*] and The Cardueae Radiations Group (in alphabetical order: Juan Antonio Calleja[b], Pau Carnicero[b], Kazumi Fujikawa[c], Mercè Galbany-Casals[b], Núria Garcia-Jacas[a], Hyoung-Tak Im[d], Seung-Chul Kim[e], Jian-Quan Liu[f], Javier López-Alvarado[b], Jordi López-Pujol[a], Jennifer R. Mandel[g], Sergi Massó[a], Iraj Mehregan[h], Noemí Montes-Moreno[a], Elizaveta Pyak[i], Cristina Roquet[j], Llorenç Sáez[b], Alexander Sennikov[k], Alfonso Susanna[a], Roser Vilatersana[a])

a Botanic Institute of Barcelona (IBB, CSIC-ICUB), Pg. del Migdia, s. n., 08038 Barcelona, Spain

b Systematics and Evolution of Vascular Plants (UAB) – Associated Unit to CSIC. Departament de Biologia Animal, Biologia Vegetal i Ecologia, Facultat de Biociències, Universitat Autònoma de Barcelona, ES-08193 Bellaterra, Spain

c Kochi Prefectural Makino Botanical Garden, 4200-6, Godaisan, Kochi 781-8125, Japan

d Department of Biology, Chonnam National University, Gwangju 500 -757, Republic of Korea

e Department of Biological Sciences, Sungkyunkwan University, Gyeonggi-do 440-746, Republic of Korea

f Key Laboratory for Bio-Resources and Eco-Environment, College of Life Sciences, Sichuan University, Chengdu, China

g Department of Biological Sciences, University of Memphis, Memphis, Tennessee 38152, USA

h Department of Biology, Science and Research Branch, Islamic Azad University, Tehran, Iran

i Department of Botany, Inst. of Biology, Tomsk State University, RU-634050 Tomsk, Russia

j Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA (Laboratoire d'Ecologie Alpine), FR-38000 Grenoble, France

k Botanical Museum, Finnish Museum of Natural History, PO Box 7, FI-00014 University of Helsinki, Finland; and Herbarium, Komarov Botanical Institute of Russian Academy of Sciences, Prof. Popov str. 2, 197376 St. Petersburg, Russia

*Corresponding author at: Botanic Institute of Barcelona (IBB, CSIC-ICUB), Pg. del Migdia, s. n., ES-08038 Barcelona, Spain. *E-mail address*: sonia.herrando@gmail.com (S. Herrando-Moraira).

Abstract

Target enrichment is a cost-effective sequencing technique that holds promise to elucidate evolutionary relationships in fast-evolving lineages. However, potential biases and impact of bioinformatic sequence treatments in phylogenetic inference have not been thoroughly explored yet. Here, we investigate this issue with the final aim to shed light into a highly diversified group of Compositae (Asteraceae) constituted by four main genera: *Arctium*, *Cousinia*, *Saussurea*, and *Jurinea*. Specifically, we compared sequence data extraction methods implemented in two easy-to-use workflows, PHYLUCE and HybPiper, and assessed the impact of two filtering practices intended to reduce phylogenetic noise. In addition, we compared two phylogenetic inference methods: 1) the concatenation approach, in which all loci were concatenated in a supermatrix; and 2) the coalescence approach, in which gene trees were produced independently and then used to construct a species tree under coalescence assumptions. Here we confirm the usefulness of the set of 1061 COS targets (nuclear conserved orthology loci set developed for Compositae) across a variety of taxonomic levels. Intergeneric relationships were completely resolved: there are two sister groups, *Arctium-Cousinia* and *Saussurea-Jurinea*, which are in agreement with a morphological hypothesis. Intrageneric relationships among species of *Arctium*, *Cousinia*, and *Saussurea* are also well defined. Conversely, conflicting species relationships remain for *Jurinea*. Methodological choices significantly affected phylogenies in terms of topology, branch length, and support. Across all analyses, the phylogeny obtained using HybPiper and the strictest scheme of removing fast-evolving sites was estimated as the optimal. Regarding methodological choices, we conclude that: 1) trees obtained under the coalescence approach are more topologically congruent to each other than those inferred using the concatenation approach; 2) refining treatments only improved support values under the concatenation approach; and 3) branch support values are maximized when fast-evolving sites are removed for the concatenation approach, and when a higher number of loci is analyzed for the coalescence approach.

Keywords

Asteraceae
COS targets
HybPiper
NGS filtering strategies
Phylogenetic noise
PHYLUCE

1. Introduction

1.1. Target enrichment strategies

The advent of "target/hybrid enrichment" or "sequence capture" method has emerged in the last years as one of the most useful techniques in the field of phylogenomics and evolutionary studies (Cronn et al., 2012; Grover et al., 2012; Mamanova et al., 2009). This approach has provided significant advances, shedding light on previously unresolved evolutionary lineages analyzed using Sanger sequencing (Nicholls et al., 2015). This next generation sequencing (NGS) tool allows the recovery of hundreds to thousands of genetic markers from specific regions of the genome, even from degraded and ancient samples (Cronn et al., 2012). Remarkable advantages of this technique are: its reasonable sequencing cost, its power to resolve relationships at different taxonomic levels, and its reduced bioinformatic complexity compared to whole genome sequencing (Lemmon and Lemmon, 2013). The target DNA regions are enriched using probes or "baits". These can be specifically designed for the group of study via a known genome or transcriptome of a closely related species (e.g. Folk et al., 2015; García et al., 2017; Schmickl et al., 2016; Syring et al., 2016), or universally conserved loci (e.g., anchored hybrid enrichment, AHE) as for vertebrates (Lemmon et al., 2012) or angiosperms (Buddenhagen et al., 2016).

Concerning the Compositae or Asteraceae (both terms used to refer to the sunflower family; hereafter Compositae), Mandel et al. (2014) recently developed a target enrichment method, which uses the Hyb-Seq approach (Weitemier et al., 2014), comprising a probe set of 9678 baits targeting a total of 1061 conserved orthology loci (hereafter COS) in this family. These COS loci were identified from thousands of expressed sequence tags (EST) across three available genomes of the family (see Mandel et al., 2014). This method has already proven useful at varied taxonomical scales, from deep Compositae nodes to shallower ones (Mandel et al., 2014, 2015, 2017; Siniscalchi et al., in prep.). In addition, the method allows the recovery of plastome data captured from off-target sequenced reads (Mandel et al., 2015). Nevertheless, the analytical power of this approach to resolve species relationships of recently and rapidly radiated genera in the family remains untested. In addition, the above cited previous works using Compositae COS targets (Mandel et al., 2014, 2015, 2017) were performed following only one bioinformatics workflow for target sequences extraction, i.e. PHYLUCE (Faircloth, 2015).

The last point seems crucial, since it has not been thoroughly investigated yet whether different bioinformatics extraction approaches yield congruent phylogenetic results, and whether these

methodological choices could lead to bias in phylogenetic reconstruction. In recent years, a great number of easy-to-use workflows and automated pipelines are emerging to be used as target extraction procedures. The pipeline PHYLUCE (Faircloth, 2015) was initially designed for ultra-conserved elements (UCEs, Faircloth et al., 2012) and applied to a wide range of animal groups: birds (Hosner et al., 2015; Moyle et al., 2016), skinks (Bryson et al., 2017), ants (Ješovnik et al., 2017) and fishes (Burress et al., 2017; Longo et al., 2017). A bioinformatic approach for AHE was proposed in Prum et al. (2015) and used in several plant studies (Buddenhagen et al., 2016; Fragoso-Martínez et al., 2017; Mitchell et al., 2017; Wanke et al., 2017). Another method, HybPiper (Johnson et al., 2016) was designed specifically for Hyb-Seq data, implementing the ability to target exons and introns separately. The HybPiper workflow also offers the option to identify and separate paralogous copies. HybPiper has already been successfully applied to analyse data from captured target loci in plants (e.g. Crowl et al., 2017; Landis et al., 2017; Chau et al., 2018; Gernandt et al., 2018; Kates et al., 2018; Medina et al., 2018; Stubbs et al., 2018; Vatanparast et al., 2018). Other new and promising tools are aTRAM (Allen et al., 2015, 2017), HybPhyloMarker (Fér and Schmickly, 2018), and SECAPR (Andermann et al., 2018). Through all published pipelines, we selected for this study two of the most commonly used approaches (PHYLUCE and HybPiper) to explore the technical differences between them and asses the consequences in inferred phylogenies of choosing one or another.

1.2. Parsing phylogenetic signal from noise in NGS studies

Despite the large amount of DNA sequence characters generated with NGS, the true gene genealogy can be obscured by various kinds of "phylogenetic noise" (Straub et al., 2014; Townsend et al., 2012). Potential sources of noise in nucleotide sequences are: unusually fast-evolving sites, rich-indel regions, and ambiguous sequence calls; which altogether may lead to substitution saturation, i.e. convergence in nucleotide states (homoplasy) that contradicts the real phylogenetic signal and bias the ancestry character-state reconstructions (Rokas and Carroll, 2006). Additional noise may accumulate in all study phases due to sequencing errors, inaccurate assembly, or incorrect orthology assignment. Another possible source of error that should be taken into account with NGS data is the incorrect allele phasing in polyploid systems (Eriksson et al., 2018), in which phylogenetic trees can be often reconstructed from consensus sequences or chimeric consensus sequences rather than real allele sequences (Kates et al., 2018). Consequences of ignoring possible phylogenetic noise are well documented (Kostka et al., 2008; Straub et al., 2014; Townsend et al., 2012), and may lead to long-branch attraction artefacts, topological

differences among alternative reconstructions, or high support values for erroneous relationships (Dornburg et al., 2014; Jeffroy et al., 2006; Salichos and Rokas, 2013).

Part of this phylogenetic noise can be reduced with standard practices such as cleaning raw reads by quality scores and alignment trimming (i.e. removal of ambiguously aligned and indel-rich positions). However, final trimmed alignments commonly used to perform phylogenetic inferences may still contain considerable levels of noise. Nowadays, standard procedures to deal with this issue are not well established, and we still lack a widely applicable refining metric to minimize the negative effects of phylogenetic noise and maximize the likelihood of an accurate phylogenetic reconstruction. Many recent studies based on target enrichment incorporate diverse filtering strategies at different components of data matrices, such as species, positions, or even entire sets of loci (see Table 1). Among all these practices, the most commonly used is the exclusion of loci recovered for a low number of species, which aims to reduce the effects of missing data and systematic bias on tree inference (see Hosner et al., 2015 for further details on potential impacts of missing data).

1.3. Resolving radiations and the case of the groups Arctium-Cousinia *and* Jurinea-Saussurea *(tribe Cardueae)*

Explosive diversification events (referred here as radiations) represent events in which many species or lineages evolved from a common ancestor in a short time period (Wen et al., 2013, 2014), caused by geographic isolation, dispersal barriers, sexual selection, or in some cases by ecological divergence or acquisition of novel key traits (Givnish, 2015). These events may leave few genomic traces, yielding few nucleotide differences among species derived from a common radiation, and thus hindering the reconstruction of phylogenetic relationships. As a consequence, unresolved phylogenies with short internal branches or large polytomies have been often recovered with traditional Sanger sequenced markers in recently diverged genera, hampering the in-depth study of radiations. With the emergence of NGS techniques, researches focused on plant radiations are significantly increasing (*Heuchera* L., Folk et al., 2015; *Inga* Mill., Nicholls et al., 2015; *Cariceae-Dulichieae-Scirpeae* clade in Cyperaceae Juss., Léveillé-Bourret et al. 2016; order Zingiberales Griseb., Sass et al., 2016; *Salvia* L. subgenus *Calosphace* (Benth.) Epling, Fragoso-Martínez et al., 2017; *Protea* L., Mitchell et al., 2017; *Aristolochia* L., Wanke et al., 2017; "*Adenocalymma-Neojobertia*" clade from Bignoniaceae, Fonseca and Lohmann, 2018; Iochrominae clade from Solanaceae, Gates et al., 2018; *Pinus* subsection *Australes* Gernandt et al., 2018). Most of these studies obtained well resolved phylogenies, but they sampled only a small

proportion of their study group. However, such first step of method testing is essential before performing studies with more complete species sampling, a type of research that will probably rise in coming years.

The tribe Cardueae (Compositae) is one of the most species-rich of the family with more than 2500 species, which account for one tenth of Compositae (Susanna and Garcia-Jacas, 2007, 2009). Three of the complexes described within Cardueae rank among the largest radiations in the family: the *Arctium-Cousinia* group, with 600 species; the *Saussurea-Jurinea* group, involving ca. 550 species; and the *Carduus-Cirsium* group, with 350 species (Susanna and Garcia-Jacas, 2007). *Saussurea* DC. and *Jurinea* Cass. are especially interesting because they constitute two paradigmatic cases of mountain radiations. Previous molecular phylogenies of these genera resulted in large and undefined polytomies (*Saussurea*, Kita et al., 2004; Wang et al., 2009), as is usually the case with radiations. Another difficulty associated with the study of the radiations of *Saussurea* and *Jurinea* is the high number of satellite genera (up to 16) described within the complex (Susanna and Garcia-Jacas, 2009), considered at some point either *Saussurea* or *Jurinea*. A complete phylogenetic reconstruction of the whole group has never been performed and the taxonomic validity of the described genera remains unexplored with molecular data. In addition, species of both *Saussurea* and *Jurinea* always appeared entangled with the genera *Arctium* L. and *Cousinia* (Barres et al., 2013; Garcia-Jacas et al., 2002; Susanna et al., 2006). Thus, generic delimitation among these four genera is also unclear. Therefore, it is essential to obtain a well resolved phylogeny of these groups as a first step towards the improvement of the knowledge on the evolutionary processes that leaded to such diversified lineages.

Accordingly, we gathered for this study a representative sample of the four genera *Arctium*, *Cousinia*, *Saussurea*, and *Jurinea* together with several species within the tribe Cardueae and used the COS target enrichment approach with three main aims: 1) to evaluate the potential of COS loci for resolving relationships at inter- and intrageneric level of recently radiated genera in tribe Cardueae; 2) to elucidate the relationships among the genera *Arctium*, *Cousinia*, *Saussurea*, and *Jurinea*; 3) to test the differences between two extraction methods of target enriched data (PHYLUCE and HybPiper); and 4) to evaluate the effects of different filtering strategies on phylogenetic reconstruction and determine whether a widely applicable approach exists as a refining metric.

2. Materials and methods

2.1. Sampling strategy

In order to evaluate the usefulness of COS target enrichment methodology to resolve generic radiations in Compositae, we included several representatives of the four genera of interest: 11 species of *Arctium*, 22 species of *Cousinia*, 19 species of *Saussurea*, 24 species of *Jurinea*, and four species described under different genera within the *Saussurea-Jurinea* complex depending on the taxonomical treatment (see 4.1. for details). On the basis of previous phylogenetic studies of the tribe Cardueae (Barres et al., 2013; Garcia-Jacas et al., 2002; Susanna and Garcia-Jacas, 2007, 2009; Susanna et al., 2006), the following taxa were also added: *Alfredia acantholepis* Kar. & Kir., *Carduus pycnocephalus* L., *Cirsium sairamense* O.Fedtsch. & B.Fedtsch., *Olgaea petriprimi* B.A.Sharipova, and *Cynara cardunculus* L. For the last species, we directly incorporated into our bioinformatics workflow raw reads from Mandel et al. (2017). The information of location and voucher specimens of the 85 sampled species is summarized in Appendix, Table S1.

2.2. DNA extraction, library preparation, target enrichment, and sequencing

Dried leaf tissue was weighed to obtain a total amount of 200 mg per sample, which was later homogenized using Mixer Mill MM 301 (Retsch®, Haan, Germany). Genomic DNA was extracted using the DNeasy plant mini kit (Qiagen, Valencia, CA, USA) following manufacturer's specifications. The quantity of each extraction was checked with Qubit™ 3.0 Fluorometer (Thermo Scientific, Waltham, MA, USA). In order to obtain an average fragment size of 400–500 bp, approximately 1 µg in 70 µl per sample was sheared using Q800R2 Sonicator® machine (QSonica, Newtown, CT, USA). Sonication step was conducted with the following parameters: 3 min (with 10 s pulse on, and 10 s pulse off), and the amplitude set at 20%. To ensure that genomic DNA was sheared at approximately the selected fragment size, all samples were checked and evaluated on a 1.2% (w/v) agarose gel. After shearing, we prepared the barcoded sequencing libraries using the NEBNext Ultra II DNA Library Prep kit for Illumina (New England Biolabs, Ipswich, MA, USA), following the standard protocol provided by the manufacturer. We added 25 µl of AMPure XP beads (Beckman Coulter, La Brea, CA, USA) for the first step of size selection, and 10 µl for the second step. The PCR amplification was performed using 15 cycles and each library was barcoded employing a unique index primer using NEBNext Multiplex Oligos for Illumina. Library quantities were checked using the Qubit Fluorometer and then pooled in groups of four samples, aiming for quantity of 500 ng per group. Pools were evaporated in a speed vacuum centrifuge, and then were resuspended in 7 µl of dH_20. For sequence capture, we used MyBaits COS 1Kv1 (MYcroarray, Ann Arbor, MI, USA; http://www.mycroarray.com/mybaits/mybaits-

UCEs.html). We followed specifications in manufacturer's protocol with slight modifications, such as the time and temperature to allow baits to hybridize to their targets (40 h at 65ºC). A post-capture PCR reaction of 16 cycles was performed using KAPA® HiFi HotStart ReadyMix (Kapa Biosystems, USA) and "reamp" primers described in Meyer and Kircher (2010). To avoid adapter dimers problems, we added a supplementary cleanup magnetic bead-based step after the post-capture PCR reaction as specified in the NEBNext manual. Finally, target-enriched library pools were sent for sequencing to the DNA Sequencing Core CGRC/ICBR of the University of Florida or to Macrogen Co. (Seoul, South Korea) on one lane on a HiSeq 3000 sequencing platform (Illumina, USA) using 100 bp paired-end reads.

2.3. Raw data processing

A first quality control of raw sequence reads demultiplexed by sequencing cores was conducted in FastQC v.0.10.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Raw FASTQ data were then cleaned using ILUMIPROCESSOR (Faircloth, 2013), a wrapper program which incorporates Trimmomatic v.0.36 (Bolger et al., 2014) to remove Illumina adapters and to trim low-quality nucleotides of the reads. A trimming step was conducted with a sliding-window set to 5:20, cutting the start or the end of a read when the average of five terminal positions falls below 20 of the quality Phred+33 score. Cleaned reads finally retained were those with a minimum length of 36 bp and with both corresponding forward and reverse pair.

2.4. Extraction of target enrichment data: PHYLUCE and HybPiper pipelines

Two different orthology-detection methods were followed to extract and identify the sequence data that matched the 1061 target COS loci: the PHYLUCE pipeline package v.1.5 (Faircloth, 2015) and the HybPiper pipeline v.1.1 (Johnson et al., 2016). The main difference between both procedures is that the PHYLUCE pipeline begins with a *de novo* assembly of reads into contigs followed by a mapping step, aligning contigs back to the reference sequences. HybPiper first maps the reads against each target separately, and then assembles *de novo* the mapped reads into contigs, which are later mapped to targets (Fig. 1).

For the PHYLUCE method (Fig. 1), the trimmed reads were *de novo* assembled into contigs using the software SPAdes v.3.9.0 (Bankevich et al., 2012) testing several k-mer lengths: 21, 33, 55 and 77. Then, we mapped resultant contigs to the COS target sequences using LASTZ (Harris, 2007) with the python script "assembly_match_contigs_to_probes.py". This program ensures that matches are 80% identical in 80% of the total length, and also removes potential paralogs. These

potential paralogs are identified as assembled contigs that match with multiple loci, or different contigs that match the same COS locus. After COS identification, the "get_match_counts.py" script was used to generate a relational list of contig names, generated by the assembler, with the names of each COS target across taxa indicated in a "taxon-set" file. This relational database was used for the script "get_fastas_from_match_counts.py" to generate a monolithic FASTA-formatted file containing all loci recovered for all taxa specified. Separate files for each locus were obtained running "assembly_explode_get_fastas_file.py". For the final step of dataset creation, we used "align_remove_locus_name_from_nexus_lines.py" to remove locus name from the FASTA header line to only retain the taxon name as will be required for downstream analyses. The majority of raw extracted sequences were longer than the target length because reads were first assembled into contigs and then contigs were mapped to the reference targets. Therefore, extracted sequences could encompass part of non-coding regions outside the targets, which are derived from exonic regions.

For the HybPiper method, we used three sets of input files: the cleaned pair-end reads, a text-formatted list with the species names, and the target file that contains one or several orthologous sequences for each locus (see Mandel et al., 2014). We executed the entire pipeline with the script "reads_first.py" which, in a first phase, maps reads to each target gene using the BWA mapper (Li and Durbin, 2009), selecting the best target sequence as a reference according to mapping score. Secondly, reads mapped for each gene were *de novo* assembled into contigs with the best k-mer automatically detected by SPAdes assembler. In a third phase, "exonerate.py" was used to extract a unique longest contig that aligned to the reference sequence. If multiple equally long contigs coexisted for the same locus (potential paralogs), the contig with greater coverage depth (10 times more) or the one with greater similarity to the target was retained (for details see Johnson et al., 2016). As a rule, we extracted exons because our target set comes from EST (expressed sequence tags), but some contigs may contain an extension of flanking non-coding regions, and in these cases, contigs are usually longer than the target sequence. Finally, to retrieve sequences recovered for each species in a multi-fasta file for each gene, the "retrieve_sequences.py" script was executed.

In order to show the differences in recovery efficiency between PHYLUCE and HybPiper methods, "get_seq_lengths.py" from HybPiper package was applied with slight modifications to the individual unaligned loci.

2.5. Alignment, alignment trimming, loci concatenation, and summary statistics

For both PHYLUCE and HybPiper methods, the multi-fasta files generated were aligned, for each locus separately, using the *auto* setting of MAFFT v.7.266 (Katoh and Standley, 2013). The resulting alignments were trimmed with trimAl v.14 (applying the *automated1* flag) with the aim to remove positions ambiguously aligned (Capella-Gutiérrez et al., 2009). For subsequent phylogenetic inference based on supermatrix analysis, gene alignments were concatenated with FASconCAT-G v.1.02 (Kück and Longo, 2014), which also provides the necessary information of gene partitions for subsequent steps. Finally, summary statistics of concatenated matrices were computed with AMAS (Borowiec, 2016).

2.6. Phylogenetic analyses without filtering step

The phylogenetic reconstruction analyses were conducted twice: first, under the concatenation approach using a supermatrix for tree estimation (hereafter concatenation approach), and second, under coalescence assumptions, in which species tree is estimated based on individual gene trees resulting from phylogenetic analyses of each locus separately (hereafter coalescence approach).

Concerning the concatenation approach, we ran Maximum Likelihood (ML) analyses with the software RAxML v.8.2.9 (Stamatakis, 2014) implemented on XSEDE in the CIPRES Science Gateway v.3.1 (Miller et al., 2010). Specifically, we ran a simultaneous rapid bootstrapping and best ML tree search (Stamatakis et al., 2008), with 10 randomized maximum parsimony starting trees and a bootstrap resampling of 500 replicates to assess branch support values. We considered that only nodes with bootstrap (BS) support values > 70% were statistically supported (Hillis and Bull, 1993). In the RAxML analysis, each locus was treated as a unit partition, and the GTRGAMMA evolution model was applied as recommended in Stamatakis (2006). Resulting trees were visualized in FigTree v.1.4.3 (Rambaut, 2016).

Regarding the coalescence approach, we first searched individual gene trees with RAxML applying the same search options specified above but running 200-bootstrap replicates. Species tree inference under coalescence approach was then performed using ASTRAL (Mirarab et al., 2014), which estimates the species tree that maximizes the number of quartets from a given input of unrooted gene trees under the assumption that all of them are correct. Branch support values were inferred through local posterior probabilities (LPP; Sayyari and Mirarab, 2016) calculated in ASTRAL-III v.5.5.3 (Zhang et al., 2018). The use of LPP as branch support metric has been proved to be more precise than multi-locus bootstrapping, especially when the error in estimating gene trees is low (Sayyari and Mirarab, 2016). Values of LPP > 0.95 were considered as strong branch

support with very high precision, although lower values (LPP = 0.7–0.9) also give high precision (Sayyari and Mirarab, 2016).

2.7. Phylogenetic informativeness and position filtering

As a filtering step recently recommended by Fragoso-Martínez et al. (2017), we evaluated the effect of eliminating the "phantom" spike positions (ambiguous, indel-rich positions, or positions with high substitution rates) that can add phylogenetic noise and bias phylogenetic reconstructions. To identify these fast-evolving sites in our alignments, a first Phylogenetic Informativeness (PI) analysis and net PI profiles were performed in the web application PhyDesign (López-Giráldez and Townsend, 2011), specifically calculating the substitution rates per site with the implemented program HyPhy (Kosakovsky Pond et al., 2005). For the rate calculations, we used two inputs: 1) the partitioned concatenated matrices, from both PHYLUCE and HybPiper methods; and 2) their respective ML trees, that were transformed to ultrametric with TreeEdit v.1.0a10 (Rambaut, 2002), applying the non-parametric rate smoothing algorithm (Sanderson, 1997) and scaled to a total height of 1.

To detect which positions exceeded the substitution rate (SR) values higher than three arbitrary pre-defined cut-off thresholds (5, 2.5, and 1), we imported the rate files per locus from PhyDesign to the R script “mmc3.R” developed by Fragoso-Martínez et al. (2017) and ran it in R v.3.1.2 (R Core Team, 2014) three times for PHYLUCE and HybPiper datasets. The resulting spreadsheet of each analysis contained specific positions to remove from each locus (spreadsheet available in Appendix A). The final filtered matrices were generated in RAxML with the command –E, using the lists of positions to be removed, original matrices, and partition information. Next, ML and PI analyses were performed with the six filtered matrices. We used the AMAS software to separate each locus and re-ran RAxML per gene to later perform the coalescence analysis with ASTRAL-III.

2.8. Selection of the most informative loci

In order to reduce phylogenetic noise, another filtering strategy based on the selection of the most informative loci according to several parameters was implemented as suggested by Borowiec et al. (2015). We used the script “gene_stats.R“ (available in Borowiec et al., 2015) for the locus selection procedure. As input, we used the individual ML gene trees and the trimmed alignments both with the PHYLUCE and the HybPiper unfiltered datasets. Then, the loci were scored for each of the following parameters: 1) the number of species recovered (accounting for taxon occupancy

and missing data); 2) the average BS support value of the ML gene tree obtained (depicting information content); 3) the R^2 of mutational saturation regression curves (Philippe and Forterre, 1999), obtained from the inferred substitution values based on ML gene tree branch lengths against the number of observed differences in sequences for a given pair of species (representing saturation); and 4) average branch length of the ML gene tree calculated from the division of total tree length by total tree nodes (characterizing the rate of molecular evolution).

For each parameter, we scored each locus with 0 or 1 points, depending on whether it exceeded the arbitrary thresholds defined here (0 for parameter value below the threshold and 1 above the threshold). The thresholds selected were the following for each parameter. For taxon occupancy, the loci recovered in at least 50% of taxa (43 species) in the PHYLUCE dataset, and 95% of taxa (81 species) in the HybPiper dataset, were scored with 1 point. For the average BS value, the loci that yielded a tree with at least 60% mean BS in the PHYLUCE dataset, and 40% in the HybPiper dataset, were scored with 1 point. For saturation, the 25% of loci with the highest R^2 of saturation curves were scored with 1 point. Finally, for evolution rates, the 25% of loci with lowest average branch length were scored with 1 point. Accordingly, a binary matrix with 0 and 1 points for each locus and each parameter was obtained. Finally, the selection of the best informative loci was performed in two steps: first, we calculated the points obtained for each locus, which ranged from 0 to 4, considering the four parameters together. And second, we selected those loci with at least had 2 points. Note that the four parameters were equally weighted, without any additional ponderation step, and that threshold values can be modified by the user depending on the parameter scores or the characteristics of the dataset analyzed.

The spreadsheets with parameters and scores are provided in Appendix A. After applying this locus filtering strategy, new datasets that only contained the selected loci were created accordingly, one comprising the concatenated matrix that was analyzed under concatenation approach, and the other with each locus in a separate file, analyzed under coalescence approach (both approaches described in section 2.6).

2.9. Topological comparisons

Differences in topology among all trees generated (unfiltered and filtered matrices, and in each case under concatenation and coalescence approaches) were estimated with the Robinson-Foulds distance (RF; Robinson and Foulds, 1981). First, we computed pairwise RF distances using PAUP v.4.0a (Swofford, 2003) and adjusted RF (RFadj) manually, which was estimated from RFajd = RF/(2n–6) being n the number of tree nodes (Mitchell et al., 2017; Steel and Penny, 1993), and

ranging from 0 (same topology) to 1 (completely discordant topology). Secondly, the RF distances were exported as a tree distance matrix in R to compare all trees in the same tree space using the multidimensional scaling approach (Hillis et al., 2005) implemented in R function "cmdscale" from the R package "stats".

3. Results

3.1. Target capture sequencing and efficiency

The average of raw pair-end reads was 4,263,196 per species. The outgroup *Carduus pycnocephalus* was the species sequenced here with the lowest number of reads (741,845), whereas *Saussurea davurica* had the highest number of reads (11,202,023).

From the 1061 targeted loci, we recovered a total of 675 loci (63.6%) with the PHYLUCE method and a total of 1055 loci (99.4%) with the HybPiper method (Table 2). Per species, the mean of on-target loci was 341 with the PHYLUCE method (lowest number of loci recovered for a species = 208, highest number of loci recovered for a species = 424), and 991 (510–1018) with the HybPiper method. In addition to this remarkable difference in the percentage of captured targets, our results show that the recovered loci per species are not equally distributed across the matrix in the PHYLUCE method (Fig. 2A–B). We pruned 48 loci recovered with PHYLUCE and four loci recovered with HybPiper because they were captured only in one or two species. Consequently, the final set of loci comprised 627 loci (59.1%) with the PHYLUCE method and 1051 (99.1%) with the HybPiper method. Only 9.2% of the loci selected with PHYLUCE were captured for 90% or more of the species sampled. In contrast, the taxa recovery was greater with HybPiper, with 89.6% of the selected loci captured in 90% or more of the sampled species. Despite these differences in missing data, the mean alignment length per locus was higher with the PHYLUCE method (823 bp; 139–3134 bp) than with the HybPiper method (317 bp; 63–1475 bp). Regarding the length of the captured loci in relation to the length of the respective reference target, we found that, in the case of PHYLUCE, 77.18% of the loci recovered were longer than the corresponding target (Fig. 2C); whereas with HybPiper, only 15.2% of the loci recovered were longer (Fig. 2D). The final aligned trimmed concatenated matrices were composed by 515,875 bp with the PHYLUCE method and 333,614 bp with the HybPiper method, with 48% and 36% of variable sites, respectively (Table 2).

3.2. Phylogeny estimation

The capture probes designed for Compositae targeting 1061 conserved ortholog loci have been useful to elucidate relationships among *Arctium*, *Cousinia*, *Saussurea*, and *Jurinea*, their generic delimitation, and also many of the relationships among closely related species. All the inferred phylogenies across the total 10 evaluated approximations (Fig. 1 and Table 3) support the monophyly of the four main genera (Supplementary Figs. S1–S4), unambiguously reflecting the four highly diversified lineages. The relationships as sister groups between *Arctium*-*Cousinia* and *Saussurea*-*Jurinea* were fully resolved with maximum support values also in all datasets analyzed (BS = 100 and LPP = 1). At lower taxonomical levels, shallow relationships were generally reconstructed with high support values for *Arctium*, *Cousinia*, and *Saussurea*, with only slight differences between analyses. In contrast, species relationships within *Jurinea* were not clearly outlined, presenting low-moderate support values (Supplementary Figs. S1–S4).

Across all the analyses under the concatenation approach, the best-resolved tree was that obtained with the HybPiper method and removing positions with SR > 1, with an average of 93.4% of BS internode support value and only five branches with BS < 70% (Fig. 3A and Table 3). The individual gene trees showed low average BS support values for both target extraction methods, although the values obtained with the PHYLUCE method (BS = 58.5) were considerably higher than the ones obtained with HybPiper (BS = 32.2), probably due to the longer loci and a few number of recovered species per locus (Table 2). A positive correlation was detected when the length of locus alignments and average BS support values of gene trees were compared for the PHYLUCE (Pearson's $r = 0.54$, $p < 0.0001$) and the HybPiper (Pearson's $r = 0.60$, $p < 0.0001$) methods. We also found another correlation, this time negative, between the number of taxa recovered and the average BS support value per locus obtained with the PHYLUCE unfiltered dataset (Pearson's $r = -0.616$, $p < 0.0001$) and the HybPiper unfiltered dataset (Pearson's $r = -0.161$, $p < 0.0001$). This indicated that gene trees performed with those loci with a lower number of taxa recovered tended to be more supported. The lack of support in individual gene trees or the incongruence among them was reflected in short internal branches in coalescence units in the trees inferred under the coalescence approach (Fig. 3 and Supplementary Figs. S1–S4). Also, support values of coalescence trees were lower than those of the trees inferred with the concatenation approach (Table 3).

3.3. Phylogenetic informativeness and position filtering results

The amount of positions selected as "fast-evolving" sites (and thus removed from the alignments) varied considerably depending on the target extraction method. For all thresholds tested, a greater number of fast-evolving sites were removed from the PHYLUCE dataset, which were distributed in a greater number of loci than in the HybPiper dataset (see Table 3 for more details). For example, for a given filtering scenario (SR > 1) the positions trimmed were 9244 in the PHYLUCE dataset and 1885 in the HybPiper dataset. When these values were corrected for the matrix length, the number of positions filtered in the PHYLUCE dataset remained higher than in the HybPiper dataset, representing 1.8% of the PHYLUCE dataset and 0.6% of the HybPiper dataset.

The net phylogenetic informativeness (PI) mean value was markedly higher for the unfiltered PHYLUCE alignment (193.55) than for the unfiltered HybPiper alignment (26.20). The maximum PI value, which is related to "phantom spikes", was also higher for the unfiltered PHYLUCE dataset (7832.29) than for the HybPiper dataset (370.51). Overall, the highest PI values were coincident with the divergence of the four genera and their respective subsequent linages, approximately at the time of 0.2–0.7 (PHYLUCE; Fig. 4A) and 0.3–0.8 (HybPiper; Fig. 4C). On the other hand, at earlier timing range (0–0.2), coincident with the main diversification of *Jurinea* and *Saussurea*, the PI profiles showed several peaks of loci visualized as "phantom spikes" that represent fast-evolving sites.

The removal of positions with substitution rates higher than 5 or 2.5 from unfiltered alignments did not improve the BS support values of trees (Supplementary Figs. S5 and S6). In contrast, with the strictest filtering scheme (SR > 1) of fast evolving sites removal, the number of resolved nodes in the concatenation approach notably increased (Fig. 4), and the curves were completely softened in the case of the HybPiper dataset (Fig. 4D). However, for the PHYLUCE matrix, some peaks close to zero, meaning towards the present and the shallowest clades, appeared in all three filtering schemes (Fig. 4B). This could indicate that a stricter threshold would probably be needed to remove very fast-evolving positions and produce more refined PI profiles in the case of the PHYLUCE dataset.

3.4. Selection of the most informative loci

In the loci filtering strategy (using measures of taxon sampling, information content, saturation, and the rate of evolution), we finally retained 304 loci (48% of the loci initially recovered; 234,118 bp) in the PHYLUCE dataset, and 570 loci (54% of the loci initially recovered; 200,632 bp) in the HybPiper dataset, which accounted for the highest phylogenetic signal (Table 3).

In the concatenation approach, the selection of the most informative loci resulted in significantly higher BS support values, decreasing the number of unsupported nodes from 17 to 9 in PHYLUCE and from 11 to 8 in HybPiper. In the coalescence approach, the selection of the best loci was not effective in terms of improving the LPP support values or the number of unsupported nodes, both values of these two metrics were even lower than the ones obtained with the corresponding unfiltered datasets (Figs. 5 and 6, and Table 3).

3.5. Comparison of tree topologies

Varied tree topologies were recovered in the global tree space among the approaches of concatenation/coalescence and the non-filtering/filtering strategies (Fig. 7). Discordant topologies between the concatenation and the coalescence approaches are well illustrated in the distant position that they occupy in the bidimensional tree space along both first and second dimensions (Fig. 7). On average, the RFadj between all the trees under concatenation vs. all the trees under coalescence was relatively high, with a 0.44 value (0.33–0.59; Supplementary Table S3). When topologies were compared among all those obtained under the concatenation approach and among all those obtained under the coalescence approach, we found that the topologies obtained with the coalescence approach were more similar to each other (RFadj = 0.26) than the topologies obtained with the concatenation approach (RFadj = 0.31).

In relation to the impact of filtering, in general the softest filtering strategies of fast-evolving sites removal (SR > 5 and 2.5) did not significantly alter the tree topologies with respect to those obtained with the corresponding unfiltered datasets, both in the PHYLUCE and the HybPiper methods, under concatenation (RFadj = 0–0.01) or under coalescence (RFadj = 0–0.15). In contrast, for the strictest threshold scheme (SR > 1), topologies were more variable between unfiltered and filtered (SR > 1) datasets (Fig. 7). Especially, this effect was remarkably evident for the trees inferred under concatenation, when the unfiltered PHYLUCE dataset was compared to the filtered (SR > 1) scheme (RFadj = 0.26). The selection of the most informative loci was the filtering strategy that resulted in most discordant topologies when compared to the unfiltered datasets (Fig. 4 and Supplementary Table S3). In particular, the tree based on the best loci selected for PHYLUCE under concatenation resulted in highly incongruent topologies compared to all the rest (mean of RFadj = 0.54), i.e. this dataset yielded the trees more distantly related to the other tree topologies inferred.

4. Discussion

4.1. COS loci resolve previously obscure generic relationships in Cardueae

The COS probe targets tested for deep nodes in Compositae (Mandel et al., 2014, 2015, 2017) are also useful to resolve close relationships at intergeneric levels. This evidence adds to previous studies (Mandel et al., 2014, 2015, 2017) and confirms the wide range of taxonomical applicability of COS loci for phylogenomic and evolutionary studies on the largest family of flowering plants (Stebbins, 1970). For the first time, we were able to recover almost the entire set of target loci (99%, 1051 from 1061) using the novel pipeline HybPiper. Conversely, the pipeline PHYLUCE (the one used in previous studies for the COS set) captured only 627 loci (59%), a similar amount to those obtained in other studies for shallow species range (694 in Siniscalchi et al., in prep.) and higher taxonomical levels (763 and 795 in Mandel et al., 2014, 2015, respectively).

Here, we confidently resolved the historically obscure relationships among *Arctium*, *Cousinia*, *Saussurea*, and *Jurinea*. All phylogenies inferred in this study supported the sister relationships between *Arctium-Cousinia* and *Saussurea-Jurinea*, forming two separate complexes, a result that is congruent with the morphological hypothesis proposed by Susanna and Garcia-Jacas (2007, 2009). None of the preceding phylogenies built on Sanger sequencing data had been able to resolve with statistical support the evolutionary relationships between these four genera (Barres et al., 2013; Garcia-Jacas et al., 2002; Kita et al., 2004; Raab-Straube, 2003; Wang et al., 2007, 2013). In some cases, the genera were correctly nested but without support (López-Vinyallonga et al., 2009; Susanna et al., 2003, 2006; Susanna and Garcia-Jacas, 2009; Wang et al., 2009). Our study illustrates that controversial plant complexes with cryptic backbone relationships can be resolved with NGS target enriched data. Indeed, this NGS approach represents one of the most promising methodologies to date in the field of systematics and evolutionary biology (Buddenhagen et al., 2016), allowing the disentangling of both deep and shallow relationships of complex plant groups (e.g. Nicholls et al., 2015).

Certainly, the generic delimitation obtained here represents the first step to deepen in the knowledge of the evolution of highly diversified genera of the tribe Cardueae. The infrageneric relationships of the *Arctium-Cousinia* complex have been extensively explored with Sanger sequencing (see López-Vinyallonga et al., 2009, 2011; Susanna et al., 2003), but a complete phylogenetic assessment of *Saussurea-Jurinea* including all of the 16 small satellite genera described is still missing. Despite our reduced sampling, we have been able to clarify four possible cases of problematic classifications in the *Saussurea-Jurinea* complex. The first case concerns

Saussurea leptophylla Hemsl., which is here sampled for the first time in a phylogenetic tree. This species had been considered either as belonging to *Saussurea* (Lipschitz, 1979) or *Jurinea* (as *Jurinea ancistrophylla* Boiss., cf. Boissier, 1888). Phylogenies inferred in the present study indicate that the species should be placed in *Jurinea* (Fig. 3). Second, the satellite genus *Lipschitziella* R. Kam. (included here under *Jurinea*), was described to accommodate *Saussurea carduicephala* and *Jurinea ceratocarpa* (Raab-Straube, 2003). Our results show that *Lipschitziella* groups with *Jurinea* (Fig. 3), matching previous phylogenies (Kita et al., 2004; Raab-Straube, 2003; Susanna et al., 2006; Wang et al. 2009). Third, we confirm that the monotypic genus *Outreya* Jaub. & Spach [included here as *Jurinea carduiformis* (Jaub. & Spach) Boiss., according to Garcia-Jacas et al., 2002] belongs to the *Jurinea* clade, as it has been shown previously (Garcia-Jacas et al., 2002; Susanna et al., 2006; Wang et al., 2013). Thus, its distinction as a separate genus is not supported with the present data (Fig. 3). The last case concerns *Modestia darwasica* (C.Winkl.) Kharadze & Tamamsch., which has been treated as a different genus within the complex. However, we found that this species was clearly nested in the *Jurinea* clade, as previously reported by Susanna et al. (2006). Despite these results, a more completely sampled phylogeny of the *Saussurea-Jurinea* complex should be conducted to confirm generic boundaries within the complex.

4.2. COS loci resolve species relationships within the radiated genera Arctium, Cousinia *and* Saussurea

Our study shows that COS loci are able to resolve the relationships among species at shallow taxonomic levels for the genera *Arctium*, *Cousinia*, and *Saussurea*. Previous studies on these genera based on chloroplast and nuclear conventional markers (e.g. for *Arctium-Cousinia* in López-Vinyallonga et al., 2009; for *Saussurea* in Wang et al., 2009) retrieved large polytomies, which hindered the phylogenetic assessment of subgeneric classifications. With the target enrichment technique, we have been able to recover dichotomous relationships highly supported in most clades, especially under the concatenation approach.

In general, species from the same section grouped together, which reflects congruence between molecular and morphological assemblies. It should be noted that the topology obtained with the coalescence approach matched the morphological sections in a higher number of cases than the tree inferred with the concatenation approach (Fig. 3). For example, the three species of *Arctium* sect. *Hypacanthodes* clustered together in the coalescence tree, whereas this section was paraphyletic in the concatenation based one. This was also the case for *Cousinia* and the two taxa

of sect. *Alpinae* (Fig. 3). This fact highlights the usefulness of exploring both concatenation and coalescence approaches in phylogenetic reconstructions as currently recommended for phylogenomic data.

The comparison of the *Arctium* and *Cousinia* species relationships yielded by our optimal estimated phylogenies (Fig. 3) with previously published ones (López-Vinyallonga et al., 2009, 2011; Mehregan and Assadi, 2016; Mehregan and Kadereit, 2009; Susanna et al., 2003) shows that they are congruent except for a few cases. The phylogenies here presented provide the following new findings: 1) *Arctium grandifolium* (sect. *Amberbopsis*) and *A. eriophorum* (sect. *Schmalhausenia*) are not nested within sect. *Arctium* as previously recovered with ITS and *rpS4-trnT-trnL* markers; 2) *Cousinia tenella* (sect. *Tenellae*) groups with other *Cousinia* species, in contrast with the unusual grouping at the base of the whole *Arctium-Cousinia* complex retrieved in previous papers; and 3) after the divergence of *C. tenella*, the clade composed by *C. pusilla* and *C. polytimetica* (both from sect. *Dichotomae*) is sister to the rest of *Cousinia*. The last two points are very interesting since it is observed that the annual species of *Cousinia* (*C. tenella*, *C. pusilla*, and *C. polytimetica*) are in separate lineages from all the other species (usually monocarpic and often biennial), which are grouped together in a different and much more diversified clade. These results suggest that a life strategy shift from annual to perennial would have allowed *Cousinia* to expand into new habitats, triggering higher diversification rates in similar way to that that reported for *Lupinus* L. in montane habitats (Drummond, 2008). In the case of *Cousinia*, we observed that when monocarpic clade begins to diverge, individual gene trees became fairly incongruent and the resultant coalescence species tree, at this part, was poorly-moderately supported (Fig. 3B). This pattern of heterogeneous gene trees could be caused by an ancient hybridization or polyploidization (Folk et al., 2018), but given that these processes are very rare in *Cousinia* (Mehregan and Kadereit, 2009; Watanabe, 2002), incomplete lineage sorting (ILS) may be more likely. However, this hypothesis needs further confirmation given that our taxon sampling is limited.

Concerning *Saussurea*, the high support values found for all the clades analyzed holds promise for future resolution of this radiation with a higher taxa sampling. Previous phylogenies, also with a reduced taxa sampling, retrieved poor-moderate resolution at the species level (Kita et al., 2004; Raab-Straube, 2003; Wang et al., 2009). Species relationships within sect. *Saussurea* were different in the trees obtained with the concatenation and the coalescence approaches (Fig. 3), causing a topological incongruence (see 4.3. for possible methodological tools to explore causes of incongruence). These differences could derive from fast and island-like radiation events in the

major diversity center of *Saussurea* located in China (Wang et al., 2009; Wen et al., 2014), where more than 150 endemic species of the sect. *Saussurea* are found (Shi et al., 2011).

4.3. Conflicting species relationships within Jurinea

Compared to the other genera, the branch support values of interspecific relationships within *Jurinea* was surprisingly low. Specifically, relationships and topologies recovered were highly variable among the different phylogenomic approaches (concatenation/coalescence, among target extraction methods, and among posterior filtering treatments; Supplementary Figs. S1–S4). Whereas the optimal phylogenetic tree inferred with the concatenation approach resulted in moderate-high supported branches (only 16.7% of the internodes were unsupported; Fig. 3A), almost all branches of the coalescence tree were unsupported (93.8%; Fig. 3B), revealing high incongruence among gene trees. In addition, branch lengths of the *Jurinea* group were shorter than the branches of the other genera. Overall, this topological structure could match mainly with two typical scenarios: 1) ILS (persistence of ancestral polymorphisms of genes after species splitting), which could be common in cases of rapid radiations, resulting in short internal branches that would correspond to a simultaneous species divergence (Oliver, 2013; Rokas and Caroll, 2006; Whitfield and Lockhart, 2007); or 2) introgression phenomenon or hybrid speciation, in which gene tree histories are discordant due to events of genetic admixture with other lineages (Folk et al., 2018). The limited taxon sampling of the present and previous studies on *Jurinea* (14–18 species with ISSR or ITS in Dogan et al., 2007, 2010; Salmerón-Sánchez et al., 2015) does not allow to discriminate between these two hypotheses. The high persistence of gene tree discordance found here for this group could be matching one of most common ILS effects, which is the occurrence of the inferred species tree in the “anomaly zone” (Degnan and Rosenberg, 2006; Linkem et al., 2016). This term was described to refer to a tree space area where the most likely gene tree topologies do not reflect the true species tree topology. Therefore, phylogenetic inference methods fail to reconstruct the true species tree, especially a critical effect under the concatenation approach (Mendes and Hahn, 2018). In future investigations, the relative influence of ILS and hybridization could be tested through multiple approaches recently proposed for Hyb-Seq data (see García et al., 2017; Kamneva et al., 2017; Mitchell et al., 2017; Simmons et al., 2016). The evolutionary role of polyploidization could be also explored as suggested by Crowl et al. (2017), Eriksson et al. (2018), or Grover et al. (2015). Although COS loci have been designed from low-copy nuclear genes, several possible paralog copies have been detected (Mandel et al., 2015) as we found here (see

section 4.4.), and as had been reported for AHE data (Buddenhagen et al., 2016). However, polyploidy seems to be as rare in *Jurinea* as it is in *Cousinia* (Watanabe, 2002).

Several strategies may be followed to shed light into the evolutionary history of rapidly diversified genera (e.g. *Helianthus* L., Stephens et al., 2015) in which gene tree discordance prevails, and is even magnified, with phylogenomic data. Certainly, the first step to improve branch support values is to obtain a complete sampling of species, which is essential for reconstructing well resolved phylogenies (Lecointre et al., 1993; Philippe et al., 2011). We included here a very small representation of *Jurinea* (26 out of the 200 described species; Susanna and Garcia-Jacas, 2007), so a broader representation is crucial to extract solid conclusions about its evolution. In agreement, we found that the position of species that are unique representatives of a section were the most variable cases in different phylogenetic analyses (Fig. 3). Another possible improvement in relation to sampling is the addition of several individuals per species (Kubatko and Degnan, 2007; Maddison and Knowles, 2006; McCormack et al., 2009), specially recommended for the coalescence approach when considerable levels of gene tree heterogeneity exist in the clade of interest.

Another option would be to increase gene alignment length by concatenating compatible loci through methods like naive or statistical binning (Bayzid and Warnow, 2013; Bayzid et al., 2015; Mirarab et al., 2014). In this way, the possible effect to incorporate gene trees derived from short alignments with a weak phylogenetic signal, which could lead to a poorly resolved species tree under coalescence approach, is minimized. However, disparate results have been found applying binning procedures, recovering well resolved coalescence trees in some cases (Blaimer et al., 2016) and poorly resolved phylogenies in others (Streicher et al., 2018). Another alternative in study cases with low ILS effect, which seems not appropriate for *Jurinea* case, would be to recover a higher number of variable positions such as those located in introns or flanking regions around the core of conserved probes set, as has been reported that variability within the alignments increases with increasing distance from the center of UCE anchored loci (Bossert et al., 2017; Faircloth et al., 2012; Gilbert et al., 2015; Van Dam et al., 2017). As an example, this strategy has proved useful to resolve species divergence in the radiated genus *Heuchera* (Folk et al., 2015).

Finally, other variants of high throughput sequencing, like restriction-site associated sequencing (RAD-seq; Baird et al., 2008) could help to clarify evolutionary relationships in rapidly diversified lineages, as it has been successfully achieved for other radiations (e.g. Darwell et al., 2016; Tripp et al., 2017; Wagner et al., 2013). Nevertheless, important drawbacks should be considered for this method: the short length of the loci captured (< 300 bp; Andrews et al., 2016) with an increase of uncertain homology in relation to time since species divergence (Wagner et

al., 2013), the difficulties to link data from different studies, and the problems already detected to resolve short internal branches (Leaché et al., 2015).

4.4. Evaluating differences between target extraction methods: PHYLUCE and HybPiper pipelines

This study represents the first evaluation of the impact in phylogenies of two target extraction methods implemented in the automated pipelines PHYLUCE (Faircloth, 2015) and HybPiper (Johnson et al., 2016). One of the notable differences observed between the two approaches is the length of the final matrix recovered: the PHYLUCE matrix was 35.3% longer than the HybPiper one (see Table 2 for details). At first sight, this result is quite surprising given that the total number of loci found was lower with PHYLUCE (627) than with HybPiper (1051). However, this is likely due to the fact that with PHYLUCE the reads are assembled into contigs before being mapped to the targets, which results in contigs that are longer than the targets themselves (Fig. 2C). In contrast, with HybPiper the reads are assembled after being mapped to the targets and thus the resulting contigs cannot be much longer than the targets, and actually tend to be shorter (Fig. 2D). This is also reflected in length differences of the individual locus alignments (on average of 823 bp per locus with PHYLUCE and 317 bp per locus with HybPiper).

Despite the fact that longer alignments are desirable for gene tree reconstructions, regions outside the core of the COS targets (identified from EST; Mandel et al., 2014) might include non-coding regions that are more variable and as consequence high number of positions could have anomalous high substitution rates. Thus, in this non-coding regions saturation and multiple hits effects may tend to be high, and accordingly the positional homology would be questionable. To this point, the phylogenetic informativeness analysis detected great amounts of "phantom spikes" in the PHYLUCE matrix, and even in the strictest scheme of fast-evolving positions removal (SR > 1, 9244 bp removed; Table 3), the curves of locus profiles were not smoothed sufficiently (Fig. 4B). However, this result could not only be due to the recovery of highly variable regions outside the conservative core of the EST regions, it could also be related to the lack of a target reference sequence to map the non-target sequences, thus resulting in a poorly aligned regions with considerable homoplasy problems. Overall, the conservative core of the EST regions (i.e. the COS targets) showed enough variation to infer robust phylogenetic relationships, as shown by HybPiper dataset. Therefore, largely variable sites located outside the target length could be decreasing phylogenetic signal-to-noise ratio instead of adding valuable phylogenetic information. However, it should be tested with other bioinformatics methods if for some more recent diversified lineages

with low ILS the inclusion of the COS flanking regions with great amounts of variation would provide valuable information to resolve entangled phylogenetic histories.

Another notable difference between the two methods is the different treatment of sequence variants (potentially paralogous copies or alleles; see section 2.4. for methodological details). Briefly, in PHYLUCE the retained loci are only those with a unique copy; in contrast, HybPiper retrieves multiple-copy loci, but only one of the copies (potential paralogs) is retained in the end based on the criteria described before (see section 2.4). In our sequence dataset, between 0 and 167 (144 on average) loci were flagged with paralog warnings in the HybPiper method, from a total of 1051 target loci. Such multiple copies could originate from different sources: real paralog coexistence, recent polyploidy, contamination, sequencing errors, or allelic variants. In conjunction, the species analyzed do not seem to be strongly affected by potential paralogs. However, flagged loci with paralog warnings detected with HybPiper should be further evaluated or removed from downstream analyses in a conservative framework given that small-scale duplications (segmental, tandem, and retro-duplications) have been shown to occur commonly in plant genomes (Hudson et al., 2011; Rensing, 2014).

In sum, how reads are assembled into contigs is probably the factor that contributes most to differences in the number of targets recovered between both analysis packages, rather than paralog treatment. This is evident from the fact that, with HybPiper, an average of 144 potential paralogs was detected, a number that is much lower to the difference in loci retrieved between PHYLUCE (675) and HybPiper (1055). Compared to other extraction pipelines like aTRAM (Allen et al., 2015, 2017) or the recently published HybPyloMarker (Fér and Schmickly, 2018), the predominant procedure and probably the best strategy to recover the target loci seems to be to perform assembly after the reads are mapped to the targets (see Table 2 in Fér and Schmickly, 2018).

Concerning their influence on phylogenetic results, we found that both reference-based extraction methods were successful in the resolution of backbone relationships among the evaluated genera. The high amounts of missing data per loci retrieved with PHYLUCE (only 9.2% of genes were recovered for 90% or more of the species) did not affect the branch support values of intergeneric relationships. This is in agreement with Hosner et al. (2015), who reported that missing positions in alignments could be more problematic than entire missing sequences of a given locus. At shallow taxonomic levels, both packages were also able to detect gene tree discordances in the same proportion, independently of the data analysis pipeline used (Fig. 6). However, topologies built under the concatenation approach and under coalescence with the PHYLUCE dataset were more different from each other than the ones obtained under the two

764 approaches with the HybPiper dataset (Fig. 7). Nonetheless, in the concatenation approach
765 analyses, considerable differences between the two extraction methods were found at species
766 relationships level. The PHYLUCE method failed to estimate with confidence species
767 relationships in *Jurinea*, resulting in an entangled topology with fairly low branch support values
768 compared to the results found with HybPiper (Supplementary Figs. S1A, S2A). A possible
769 explanation could be that the high number of missing loci for some species in the concatenated
770 dataset hindered ancestry state reconstructions in resampled data matrices when bootstrap
771 replicates were calculated under the concatenation approach.
772 As observed here and as García et al. (2017) reported with other extracting methods, the use of
773 different procedures of target extraction can lead to different estimates of topology and branch
774 lengths of tree reconstructions. Thus, it is evident that the choice of a given bioinformatic workflow
775 can have a critical impact on the results obtained. In summary, the PHYLUCE method seems to
776 present more limitations and introduces more phylogenetic noise than the HybPiper method.
777 However, in taxonomical groups with low-moderate degrees of ILS, hybridization and polyploidy,
778 PHYLUCE is more conservative in terms of avoiding potential paralog copies, more efficient in
779 computational time demanded, memory used, and number of files produced in comparison with
780 HybPiper.
781
782 *4.5. The coalescence approach yields higher topological robustness of phylogenetic trees*
783 High throughput sequencing has provided extensive genome-scale datasets and has been useful to
784 resolve many prior uncertain branches of the tree of life. However, incongruence between nuclear,
785 mitochondrial or chloroplast based phylogenies, and conflicting gene tree histories persist across
786 phylogenetic reconstructions (Jeffroy et al., 2006). This incongruence could be masked when gene
787 sequences recovered are concatenated as a single supergene unit (supermatrix or concatenation
788 approach). However, this analytical practice is currently under discussion in phylogenomics since
789 it tends to produce maximum bootstrap support values and completely resolved phylogenies even
790 when biological factors (like ILS, hybridization, horizontal gene transfer, recombination, and gene
791 duplication/loss), random biases, or systematic errors (compositional heterogeneity, long-branch
792 attraction, gene-tree discordances, and missing sequence data) are present in the input data
793 (Kubatko and Degnan, 2007; Liu et al., 2015b; Salichos and Rokas, 2013). In our study, we
794 obtained higher support values and almost fully resolved phylogenies applying the concatenation
795 approach (Supplementary Figs. S1 and S2), but the resulting trees showed considerable conflicting
796 topologies among the different extraction and filtering procedures (Fig. 7). These results support

the claim of previous researchers (Kubatko and Degnan, 2007; Salichos and Rokas 2013), who suggested avoiding the use of traditional bootstrap values as a metric to quantify tree certainty in the concatenation approach.

Alternatively, analyzing sequence data under the coalescence assumptions may aid in avoiding reconstruction artifacts, detect possible gene incongruences, and better integrate different gene histories (see review in Liu et al., 2015b). Here, it has been confirmed that our study group presents high gene-tree heterogeneity, which is reflected in the weakly supported internal branches of the coalescence tree (Fig. 3B). Causes of incongruence may be derived from several factors. One is the relatively short length of our gene alignments (average of 823 bp in PHYLUCE and 317 bp in HybPiper), which could result in insufficient phylogenetic signal yielding poorly resolved gene trees (average of bootstrap 58.7 in PHYLUCE and 32.2 in HybPiper). Indeed, we found positive correlations between loci lengths and mean BS support values in gene trees. In light of this observation, future studies should consider using a limited number of naive bins or a statistical binning approach (Mirarab et al., 2014) in order to improve gene trees reconciliation. It has also been proposed that high levels of missing data (missing samples per locus) could lead to low support and accuracy of coalescence trees (Gatesy and Springer, 2014). However, our phylogenetic analyses were resilient to the effects of this type of missing data, since no remarkable differences were observed between tree topologies obtained with the PHYLUCE and the HybPiper methods (Fig. 5) despite their significantly distinct proportion of missing data (Fig. 2A and 2B). Such resilience was also shown in the simulation study by Hovmöller et al. (2013). It is well documented that the coalescence approach can consistently yield trees closer to the correct species tree as the number of loci increases (Liu et al., 2015a). Concordantly, we observed that phylogenies estimated with a reduced loci dataset showed lower branch support values in our coalescence approach (see section 4.6. for details).

Despite the incongruence detected across coalescence trees (Supplementary Figs. S3 and S4) and their lower support values with respect to concatenation trees (Fig. 5 and Table 3), we detected that coalescence tree topologies obtained with alternative extraction and refining methods were more congruent or similar to each other than those obtained under the same conditions using the concatenation approach (Fig. 7). This pattern is in agreement with results reported by other researchers (Buddenhagen et al., 2016; Edwards et al., 2016; Mitchell et al., 2017), which highlighted the topological robustness of coalescence methods.

4.6. Impact of filtering target-enriched data

Recent target-enriched studies have added an additional step of sequence refining to minimize the impact of phylogenetic noise (Table 1). We explored the effectiveness of two types of dataset filtration: on the one hand removing positions with unusually high substitution rates (fast-evolving regions; Fragoso-Martínez et al., 2017), and on the other selecting the most informative loci under different criteria (Borowiec et al., 2015).

First, it should be noted that all coalescence analyses were unaffected by the application of any filtering scheme, indicating that gene-tree discordances cannot be attributed to phylogenetic noise derived from fast-evolving sites or the addition of uninformative loci (see section 4.5 for possible sources). In contrast, in the case of the concatenation approach, both strategies of filtering initial matrices before phylogenetic inference were in general effective (Table 3). This result is in agreement with similar findings reported by Xi et al. (2014), which showed that coalescence approaches were more robust in the presence of positions with high substitution rates compared to concatenation approaches.

The first strategy of position filtering proved more useful when the strictest threshold was applied (SR > 1), improving the bootstrap support values (Fig. 5) and increasing the number of supported nodes (Fig. 6) for both different target extraction pipelines. Previous works (Goremykin et al., 2010; Parks et al., 2012; Straub et al., 2014; Xi et al., 2014) and the first studies applying this filtering workflow (Fragoso-Martínez et al., 2017; Wanke et al., 2017) already suggested its benefits to reduce phylogenetic noise and saturation. In particular, the noise in our study was specially mitigated in *Jurinea*, in which the filtering strategies employed here resulted in resolving initially unsupported nodes, for instance varying from 13 to 3 in the PHYLUCE method (Figs. 4A and 4B). However, as previously highlighted, removing too many positions may lead to inappropriate exclusion of phylogenetically informative characters and consequently to loss of robustly supported clades (Drew et al., 2014; Streicher et al., 2018). This occurred in *Arctium*, for which node resolution decreased in greatest refining scenarios (filtering by positions SR > 1 in PHYLUCE; Fig. 6). For this reason, it would be desirable to test several thresholds of filtering positions and see which one fits better the entire tree or the particular clade of interest. Additionally, less restrictive cut-offs for position filtering can result in an increase of unresolved nodes, as we observed in HybPiper dataset for *Cousinia* and scheme SR > 2.5 (green square in Fig. 6).

Nowadays, one of the main questions in phylogenomics is how many loci are needed to produce robust phylogenies. The answer is complex, and there is an increasing number of studies evaluating the effects of prioritizing the quality (information-rich loci or loci recovered in high number of taxa) or the quantity (as many loci as possible) (e.g. Borowiec et al., 2015; Hosner et

al., 2015; Misof et al., 2013; Salichos and Rokas, 2013; Streicher et al., 2016). Here we observed that, in the concatenation approach, the use of less loci but those with the highest phylogenetic signal increased the resolution of entangled clades (Fig. 6), a trend observed in other works (Borowiec et al., 2015; Salichos and Rokas, 2013). However, in the coalescence approach, the retention of only the most informative loci (approximately half of them) resulted in low LPP support values and low phylogenetic resolution (Figs. 5 and 6 and Table 3). Accordingly, incongruence between gene trees persisted and former uncertain nodes of coalescence trees remained inconclusive after locus filtering, in agreement with Longo et al. (2017). Therefore, our outcomes suggest that in the coalescence approaches it seems preferable to keep all loci, rather than keeping only the most informative ones, as outlined by Liu et al. (2015a, b) and Streicher et al. (2016). Nonetheless, the strategy of eliminating relatively uninformative gene trees was successful in Hosner et al. (2015).

In sum, filtering by positions (in our case at threshold SR > 1) was the best refining strategy given the notable increase of tree resolution and the minimum topological differences in respect to the topologies recovered with unfiltered sequences (Fig. 7 and Supplementary Table 3). However, generalizing for future investigations, an optimal comprehensive filtration metric may not exist, given the different impacts of each filtering strategy depending on the clade of interest. The described methodologies of heat map (Buddenhagen et al., 2016) and internode certainty (Salichos and Rokas, 2013) could help to detect the most highly confident reconstructed clades and the more sensitive groups to particular data treatments. Additionally, trees inferred under concatenation and coalescence approaches benefit differently from sampling, filtering and post-processing strategies. In our case, it would be preferable to give priority to loci quality (removing fast-evolving positions or using the most informative ones) in the concatenation approach, and to maximize the number of loci in the coalescence approach.

Acknowledgements

Authors thank Carolina Siniscalchi, Maria Luisa Gutiérrez and Fernando Castro for providing technical support during the laboratory process. We also thank the herbaria that provided material for the study: BC, DUSH, E, ERE, FRU, GDA, LE, MJG, TK, and W. We would like to thank Stefan Wanke and two anonymous reviewers for their comments and suggestions that improved the manuscript. Financial support from the Spanish “Ministerio de Ciencia e Innovación” (Project

CGL2015-66703-P and Ph.D. grant to Sonia Herrando-Moraira) and the Catalan government ("Ajuts a grups consolidats" 2014/SGR/514 and 2017-SGR1116) is gratefully acknowledged.

Appendix A. Supplementary material

Supplementary Figures S1–S6 and Supplementary Tables S1–S3 can be found, in the online version, at http://dx.doi.org/XXXXXXXX. All alignments and tree files for each dataset are deposited in a dryad package (http://dx.doi.org/XXXXXX).

References

Allen, J.M., Boyd, B., Nguyen, N.-P., Vachaspati, P., Warnow, T., Huang, D.I., Grady, P. G., Bell, K.C., Cronk, Q.C., Mugisha, L., Pittendrigh, B.R., Leonardi, M.S., Reed, D.L. & Johnson, K.P. 2017. Phylogenomics from whole genome sequences using aTRAM. *Systematic Biology* 66: 786–798. https://doi.org/10.1093/sysbio/syw105

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: 98. https://doi.org/10.1186/s12859-015-0515-2

Andermann, T., Cano, A., Zizka, A., Bacon, C. & Antonelli, A. 2018. SECAPR-A bioinformatics pipeline for the rapid and user-friendly processing of Illumina sequences, from raw reads to alignments. *PeerJ* Preprints 6: e26477v3. https://doi.org/10.7287/peerj.preprints.26477v3

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G. & Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics* 17: 81–92. https://doi.org/10.1038/nrg.2015.28

Baird, N.A., Etter, P.D., Atwood, T.S., Currey, M.C., Shiver, A.L., Lewis, Z.A., Selker, E.U., Cresko, W.A. & Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* 3: e3376. https://doi.org/10.1371/journal.pone.0003376

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. & Pyshkin, A.V. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19: 455–477. https://doi.org/10.1089/cmb.2012.0021

Barres, L., Sanmartín, I., Anderson, C.L., Susanna, A., Buerki, S., Galbany-Casals, M. & Vilatersana, R. 2013. Reconstructing the evolution and biogeographic history of tribe Cardueae (Compositae). *American Journal of Botany* 100: 867–882. https://doi.org/10.3732/ajb.1200058

Bayzid, M.S. & Warnow, T. 2013. Naive binning improves phylogenomic analyses. *Bioinformatics* 29: 2277–2284. https://doi.org/10.1093/bioinformatics/btt394

Bayzid, M.S., Mirarab, S., Boussau, B. & Warnow, T. 2015. Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS ONE* 10: e0129183. https://doi.org/10.1371/journal.pone.0129183

Blaimer, B.B., LaPolla, J.S., Branstetter, M.G., Lloyd, M.W. & Brady, S.G. 2016. Phylogenomics, biogeography and diversification of obligate mealybug-tending ants in the genus *Acropyga*. *Molecular Phylogenetics and Evolution* 102: 20–29. https://doi.org/10.1016/j.ympev.2016.05.030

Boissier, P.E. 1888. *Flora Orientalis sive enumeratio plantarum in Oriente a Graecia et Aegypto ad Indiae fines hucusque observatarum*. H. Georg, Basilea.

Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borowiec, M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ* 4: e1660. https://doi.org/10.7717/peerj.1660

Borowiec, M.L., Lee, E.K., Chiu, J.C. & Plachetzki, D.C. 2015. Extracting phylogenetic signal and accounting for bias in whole-genome data sets supports the Ctenophora as sister to remaining Metazoa. *BMC Genomics* 16: 987. https://doi.org/10.1186/s12864-015-2146-4

Bossert, S., Murray, E.A., Blaimer, B.B. & Danforth, B.N. 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. *Molecular Phylogenetics and Evolution* 111: 149–157. https://doi.org/10.1016/j.ympev.2017.03.022

Branstetter, M.G., Ješovnik, A., Sosa-Calvo, J., Lloyd, M.W., Faircloth, B.C., Brady, S.G. & Schultz, T.R. 2017. Dry habitats were crucibles of domestication in the evolution of agriculture in ants. *Proceedings of the Royal Society B* 284: 20170095. https://doi.org/10.1098/rspb.2017.0095

964 Bryson, R.W., Linkem, C.W., Pavón-Vázquez, C.J., Nieto-Montes de Oca, A., Klicka, J. &
965 McCormack, J.E. 2017. A phylogenomic perspective on the biogeography of skinks in the
966 *Plestiodon brevirostris* group inferred from target enrichment of ultraconserved elements.
967 *Journal of Biogeography* 44: 2033–2044. https://doi.org/10.1111/jbi.12989
968 Buddenhagen, C., Lemmon, A.R., Lemmon E.M., Bruhl, J., Cappa, J., Clement, W.L., Donoghue,
969 M., Edwards, E.J., Hipp, A.L., Kortyna, M., Mitchell, N., Moore, A., Prychid, C.J.,
970 Segovia-Salcedo, M.C., Simmons, M.P., Soltis, P.S., Wanke, S. & Mast, A. 2016.
971 Anchored phylogenomics of angiosperms I: Assessing the robustness of phylogenetic
972 estimates. https://doi.org/10.1101/086298
973 Burress, E.D., Alda, F., Duarte, A., Loureiro, M., Armbruster, J.W. & Chakrabarty, P. 2017.
974 Phylogenomics of pike cichlids (Cichlidae: *Crenicichla*): the rapid ecological speciation of
975 an incipient species flock. *Journal of Evolutionary Biology* 31: 14–30.
976 https://doi.org/10.1111/jeb.13196
977 Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. 2009. trimAl: a tool for automated
978 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972–1973.
979 https://doi.org/10.1093/bioinformatics/btp348
980 Chau, J.H., Rahfeldt, W.A. & Olmstead, R.G. 2018. Comparison of taxon-specific versus general
981 locus sets for targeted sequence capture in plant phylogenomics. *Applications in Plant*
982 *Sciences* 6: e1032. https://doi.org/10.1002/aps3.1032
983 Cronn, R., Knaus, B.J., Liston, A., Maughan, P.J., Parks, M., Syring, J.V. & Udall, J. 2012.
984 Targeted enrichment strategies for next-generation plant biology. *American Journal of*
985 *Botany* 99: 291–311. https://doi.org/10.3732/ajb.1100356
986 Crowl, A.A., Myers, C. & Cellinese, N. 2017. Embracing discordance: Phylogenomic analyses
987 provide evidence for allopolyploidy leading to cryptic diversity in a Mediterranean
988 *Campanula* (Campanulaceae) clade. *Evolution* 71: 913–922.
989 https://doi.org/10.1111/evo.13203
990 Cummins, C.A. & McInerney, J.O. 2011. A method for inferring the rate of evolution of
991 homologous characters that can potentially improve phylogenetic inference, resolve deep
992 divergence and correct systematic biases. *Systematic Biology* 60: 833–844.
993 https://doi.org/10.1093/sysbio/syr064
994 Darwell, C.T., Rivers, D.M. & Althoff, D.M. 2016. RAD-seq phylogenomics recovers a well-
995 resolved phylogeny of a rapid radiation of mutualistic and antagonistic yucca moths.
996 *Systematic Entomology* 41: 672–682. https://doi.org/10.1111/syen.12185

997 Degnan J.H. & Rosenberg N.A. 2006. Discordance of species trees with their most likely gene
998 trees. *PLoS Genetics* 2: 762–768. https://doi.org/10.1371/journal.pgen.0020068
999 Dogan, B., Duran, A. & Hakki, E.E. 2007. Phylogenetic analysis of *Jurinea* (Asteraceae) species
1000 from Turkey based on ISSR amplification. *Annales Botanici Fennici* 44: 353–358.
1001 Dogan, B., Hakki, E.E. & Duran, A. 2010. A phylogenetic analysis of *Jurinea* (Compositae)
1002 species from Turkey based on ITS sequence data. *African Journal of Biotechnology* 9:
1003 1741–1745. https://doi.org/10.5897/AJB10.1525
1004 Dornburg, A., Townsend, J.P., Brooks, W., Spriggs, E., Eytan, R.I., Moore, J.A., Wainwright,
1005 P.C., Lemmon, A., Lemmon, E.M. & Near, T.J. 2017. New insights on the sister lineage
1006 of percomorph fishes with an anchored hybrid enrichment dataset. *Molecular*
1007 *Phylogenetics and Evolution* 110: 27–38. https://doi.org/10.1016/j.ympev.2017.02.017
1008 Dornburg, A., Townsend, J.P., Friedman, M. & Near, T.J. 2014. Phylogenetic informativeness
1009 reconciles ray-finned fish molecular divergence times. *BMC Evolutionary Biology* 14: 169.
1010 https://doi.org/10.1186/s12862-014-0169-0
1011 Drew, B.T., Ruhfel, B.R., Smith, S.A., Moore, M.J., Briggs, B.G., Gitzendanner, M.A., Soltis, P.
1012 & Soltis, D.E. 2014. Another look at the root of the angiosperms reveals a familiar tale.
1013 *Systematic Biology* 63: 368–382. https://doi.org/10.1093/sysbio/syt108
1014 Drummond, C.S. 2008. Diversification of *Lupinus* (Leguminosae) in the western New World:
1015 derived evolution of perennial life history and colonization of montane habitats. *Molecular*
1016 *Phylogenetics and Evolution* 48: 408–421. https://doi.org/10.1016/j.ympev.2008.03.009
1017 Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., Zhong, B., Wu,
1018 S., Lemmon, E.M. & Lemmon, A.R. 2016. Implementing and testing the multispecies
1019 coalescent model: a valuable paradigm for phylogenomics. *Molecular Phylogenetics and*
1020 *Evolution* 94: 447–462. https://doi.org/10.1016/j.ympev.2015.10.027
1021 Eriksson, J.S., de Sousa, F., Bertrand, Y.J., Antonelli, A., Oxelman, B. & Pfeil, B.E. 2018. Allele
1022 phasing is critical to revealing a shared allopolyploid origin of *Medicago arborea* and *M.*
1023 *strasseri* (Fabaceae). *BMC Evolutionary Biology* 18: 9. https://doi.org/10.1186/s12862-
1024 018-1127-z
1025 Faircloth, B.C. 2013. Illumiprocessor: A trimmomatic wrapper for parallel adapter and quality
1026 trimming. Available at https://github.com/faircloth-lab/illumiprocessor.
1027 http://dx.doi.org/10.6079/J9ILL
1028 Faircloth, B.C. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci.
1029 *Bioinformatics* 32: 786–788. https://doi.org/10.1093/bioinformatics/btv646

Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T. & Glenn T.C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology* 61: 717–726. https://doi.org/10.1093/sysbio/sys004

Fér, T. & Schmickl, R.E. 2018. HybPhyloMaker: Target enrichment data analysis from raw reads to species trees. *Evolutionary Bioinformatics* 14: 1–9. https://doi.org/10.1177/1176934317742613

Folk, R.A., Mandel, J.R. & Freudenstein, J.V. 2015. A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: a phylogenomic example from *Heuchera* (Saxifragaceae). *Applications in Plant Sciences* 3: 1500039. https://doi.org/10.3732/apps.1500039

Folk, R.A., Soltis, P.S., Soltis, D.E. & Guralnick, R. 2018, in press. New prospects in the detection and comparative analysis of hybridization in the tree of life. *American Journal of Botany*. https://doi.org/10.1002/ajb2.1018

Fonseca, L.H.M. & Lohmann, L.G. 2018. Combining high-throughput sequencing and targeted loci data to infer the phylogeny of the "*Adenocalymma-Neojobertia*" clade (Bignonieae, Bignoniaceae). *Molecular Phylogenetics and Evolution* 123: 1–15. https://doi.org/10.1016/j.ympev.2018.01.023

Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E.M., Lemmon, A.R., Sazatornil, F. & Mendoza, C.G. 2017. A pilot study applying the plant Anchored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae). *Molecular Phylogenetics and Evolution* 117: 124–134. https://doi.org/10.1016/j.ympev.2017.02.006

García, N., Folk, R.A., Meerow, A.W., Chamala, S., Gitzendanner, M.A., de Oliveira, R.S., Soltis, D.E. & Soltis, P.S. 2017. Deep reticulation and incomplete lineage sorting obscure the diploid phylogeny of rain-lilies and allies (Amaryllidaceae tribe Hippeastreae). *Molecular Phylogenetics and Evolution* 111: 231–247. https://doi.org/10.1016/j.ympev.2017.04.003

Garcia-Jacas, N., Garnatje, T., Susanna, A. & Vilatersana, R. 2002. Tribal and subtribal delimitation and phylogeny of the Cardueae (Asteraceae): A combined nuclear and chloroplast DNA analysis. *Molecular Phylogenetics and Evolution* 22: 51–64. https://doi.org/10.1006/mpev.2001.1038

Gates, D.J., Pilson, D. & Smith, S.D. 2018. Filtering of target sequence capture individuals facilitates species tree construction in the plant subtribe Iochrominae (Solanaceae).

Molecular Phylogenetics and Evolution 123: 26–34. https://doi.org/10.1016/j.ympev.2018.02.002

Gatesy, J. & Springer, M.S. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution* 80: 231–266. https://doi.org/10.1016/j.ympev.2014.08.013

Gernandt, D.S., Aguirre Dugua, X., Vázquez-Lobo, A., Willyard, A., Moreno Letelier, A., Pérez de la Rosa, J.A., Piñero, D. & Liston, A. 2018. Multi-locus phylogenetics, lineage sorting, and reticulation in *Pinus* subsection *Australes*. *American Journal of Botany* 105: 1–15. https://doi.org/10.1002/ajb2.1052

Gilbert, P.S., Chang, J., Pan, C., Sobel, E.M., Sinsheimer, J.S., Faircloth, B.C. & Alfaro, M.E. 2015. Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. *Molecular Phylogenetics and Evolution* 92: 140–146. https://doi.org/10.1016/j.ympev.2015.05.027

Givnish, T.J. 2015. Adaptive radiation versus "radiation" and "explosive diversification": why conceptual distinctions are fundamental to understanding evolution. *New Phytologist* 207: 297–303. https://doi.org/10.1111/nph.13482

Goremykin, V.V., Nikiforova, S.V. & Bininda-Emonds, O.R. 2010. Automated removal of noisy data in phylogenomic analyses. *Journal of Molecular Evolution*: 71: 319–331. https://doi.org/10.1007/s00239-010-9398-z

Grover, C.E., Gallagher, J.P., Jareczek, J.J., Page, J.T., Udall, J.A., Gore, M.A. & Wendel, J.F. 2015. Re-evaluating the phylogeny of allopolyploid *Gossypium* L. *Molecular Phylogenetics and Evolution* 92: 45–52. https://doi.org/10.1016/j.ympev.2015.05.023

Grover, C.E., Salmon, A. & Wendel, J.F. 2012. Targeted sequence capture as a powerful tool for evolutionary analysis 1. *American Journal of Botany* 99: 312–319. https://doi.org/10.3732/ajb.1100323

Harris, R.S. 2007. *Improved pairwise alignment of genomic DNA*. PhD Thesis, Pennsylvania State University, Pennsylvania.

Heibl, C. 2008. PHYLOCH: R language tree plotting tools and interfaces to diverse phylogenetic software packages. http://www.christophheibl.de/Rpackages.html

Hillis, D.M. & Bull, J.J. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42: 182–192. https://doi.org/10.1093/sysbio/42.2.182

Hillis, D.M., Heath, T.A. & John, K.S. 2005. Analysis and visualization of tree space. *Systematic Biology* 54: 471–482. https://doi.org/10.1080/10635150590946961

1097 Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L. & Kimball, R.T. 2015. Avoiding missing
1098 data biases in phylogenomic inference: an empirical study in the landfowl (Aves:
1099 Galliformes). *Molecular Biology and Evolution* 33: 1110–1125.
1100 https://doi.org/10.1093/molbev/msv347
1101 Hovmöller, R., Knowles, L.L. & Kubatko, L.S. 2013. Effects of missing data on species tree
1102 estimation under the coalescent. *Molecular Phylogenetics and Evolution* 69: 1057–1062.
1103 https://doi.org/10.1016/j.ympev.2013.06.004
1104 Hudson, C.M., Puckett, E.E., Bekaert, M., Pires, J.C. & Conant, G.C. 2011. Selection for higher
1105 gene copy number after different types of plant gene duplications. *Genome Biology and*
1106 *Evolution* 3: 1369–1380. https://doi.org/10.1093/gbe/evr115
1107 Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. 2006. Phylogenomics: the beginning of
1108 incongruence? *Trends in Genetics* 22: 225–231. https://doi.org/10.1016/j.tig.2006.02.003
1109 Ješovnik, A., Sosa-Calvo, J., Lloyd, M.W., Branstetter, M.G., Fernández, F. & Schultz, T.R. 2017.
1110 Phylogenomic species delimitation and host-symbiont coevolution in the fungus-farming
1111 ant genus *Sericomyrmex* Mayr (Hymenoptera: Formicidae): ultraconserved elements
1112 (UCEs) resolve a recent radiation. *Systematic Entomology* 42: 523–542.
1113 https://doi.org/10.1111/syen.12228
1114 Johnson, M.G., Gardner, E.M., Liu, Y., Medina, R., Goffinet, B., Shaw, A.J., Zerega, N.J.C. &
1115 Wickett, N.J. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics
1116 from high-throughput sequencing reads using target enrichment. *Applications in Plant*
1117 *Sciences* 4: 1600016. https://doi.org/10.3732/apps.1600016
1118 Junier, T. & Zdobnov, E.M. 2010. The Newick utilities: high-throughput phylogenetic tree
1119 processing in the UNIX shell. *Bioinformatics* 26: 1669–1670.
1120 https://doi.org/10.1093/bioinformatics/btq243
1121 Kamneva, O.K., Syring, J., Liston, A. & Rosenberg, N.A. 2017. Evaluating allopolyploid origins
1122 in strawberries (*Fragaria*) using haplotypes generated from target capture sequencing.
1123 *BMC Evolutionary Biology* 17: 180. https://doi.org/10.1186/s12862-017-1019-7
1124 Kates, H.R., Johnson, M.G., Gardner, E.M., Zerega, N.J. & Wickett, N.J. 2018. Allele phasing has
1125 minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a
1126 case study of *Artocarpus*. *American Journal of Botany* 105: 404–416.
1127 https://doi.org/10.1002/ajb2.1068
1128 Katoh, K. & Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7:
1129 improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–
1130 780. https://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649. https://doi.org/10.1093/bioinformatics/bts199.

Kita, Y., Fujikawa, K., Ito, M., Ohba, H. & Kato, M. 2004. Molecular phylogenetics analyses and systematics of the genus *Saussurea* and related genera (Asteraceae, Cardueae). *Taxon* 53: 679–690. https://doi.org/10.2307/4135443

Kosakovsky Pond, S.L.K., Frost, S.D.W. & Muse, S.V. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676–679. https://doi.org/10.1093/bioinformatics/bti079Kostka, M., Uzlikova, M., Cepicka, I. & Flegr, J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of *Blastocystis*. *BMC Bioinformatics* 9: 341. https://doi.org/10.1186/1471-2105-9-341

Kubatko, L.S. & Degnan, J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* 56: 17–24. https://doi.org/10.1080/10635150601146041

Kück, P. & Longo, G.C. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* 11: 81. https://doi.org/10.1186/s12983-014-0081-x

Landis, J.B., Soltis, D.E. & Soltis, P.S. 2017. Comparative transcriptomic analysis of the evolution and development of flower size in *Saltugilia* (Polemoniaceae). *BMC Genomics* 18: 475. https://doi.org/10.1186/s12864-017-3868-2

Leaché, A.D., Chavez, A.S., Jones, L.N., Grummer, J.A., Gottscho, A.D. & Linkem, C.W. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology and Evolution* 7: 706–719. https://doi.org/10.1093/gbe/evv026

Lecointre, G., Philippe, H., Vân Lê, H.L. & Le Guyader, H. 1993. Species sampling has a major impact on phylogenetic inference. *Molecular Phylogenetics and Evolution* 2: 205–224. https://doi.org/10.1006/mpev.1993.1021

Lemmon, A.R., Emme, S.A. & Lemmon, E.M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology* 61: 727–744. https://doi.org/10.1093/sysbio/sys049

Lemmon, E.M. & Lemmon, A.R. 2013. High-throughput genomic data in systematics and phylogenetics. *Annual Review of Ecology, Evolution and Systematics* 44: 99–121. https://doi.org/10.1146/annurev-ecolsys-110512-135822

Léveillé-Bourret, E., Starr, J.R., Ford, B.A., Lemmon, E.M. & Lemmon, A.R. 2016. Resolving rapid radiations within angiosperm families using anchored phylogenomics. *Systematic Biology* 67: 94–112. https://doi.org/10.1093/sysbio/syx050

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with burrowswheeler transform. *Bioinformatics* 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Linkem, C.W., Minin, V.N. & Leaché, A.D. 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae). *Systematic Biology* 65: 465–477. https://doi.org/10.1093/sysbio/syw001

Lipschitz, S.J. 1979. *Genus Saussurea DC (Asteraceae)* [in Russian]. Nauka, Leningrad.

Liu, L., Wu, S. & Yu, L. 2015a. Coalescent methods for estimating species trees from phylogenomic data. *Journal of Systematics and Evolution* 53: 380–390. https://doi.org/10.1111/jse.12160

Liu, L., Xi, Z., Wu, S., Davis, C.C. & Edwards, S.V. 2015b. Estimating phylogenetic trees from genome-scale data. *Annals of the New York Academy of Sciences* 1360: 36–53. https://doi.org/10.1111/nyas.12747

Longo, S.J., Faircloth, B.C., Meyer, A., Westneat, M.W., Alfaro, M.E. & Wainwright, P.C. 2017. Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. *Molecular Phylogenetics and Evolution* 113: 33–48. https://doi.org/10.1016/j.ympev.2017.05.002

López-Giráldez, F. & Townsend, J.P. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evolutionary Biology* 11: 152. https://doi.org/10.1186/1471-2148-11-152

López-Vinyallonga, S., Mehregan, I., Garcia-Jacas, N., Tscherneva, O., Susanna, A. & Kadereit, J.W. 2009. Phylogeny and evolution of the *Arctium-Cousinia* complex (Compositae, Cardueae-Carduinae). *Taxon* 58: 153–171.

López-Vinyallonga, S., Romaschenko, K., Susanna, A. & Garcia-Jacas, N. 2011. Systematics of the arctioid group: disentangling *Arctium* and *Cousinia* (Cardueae, Carduinae). *Taxon* 60: 539–554.

Maddison, W.P. & Knowles, L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* 55: 21–30. https://doi.org/10.1080/10635150500354928

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. & Turner, D.J. 2009. Target-enrichment strategies for next-generation sequencing. *Nature Methods* 7: 111–118. https://doi.org/10.1038/nmeth.1419

Mandel, J.R., Barker, M.S., Bayer, R.J., Dikow, R.B., Gao, T., Jones, K.E., Keely, S., Kilian, N., Ma, H., Siniscalchi, C.M., Susanna, A., Thapa, R., Watson, L. & Funk, V.A. 2017. The Compositae Tree of Life in the age of phylogenomics. *Journal of Systematics and Evolution* 55: 405–403. https://doi.org/10.1111/jse.12265

Mandel, J.R., Dikow, R.B. & Funk, V.A. 2015. Using phylogenomics to resolve mega-families: An example from Compositae. *Journal of Systematics and Evolution* 53: 391–402. https://doi.org/10.1111/jse.12167

Mandel, J.R., Dikow, R.B., Funk, V.A., Masalia, R.R., Staton, S.E., Kozik, A., Michelmore, R.W., Rieseberg, L.H. & Burke, J.M. 2014. A target enrichment method for gathering phylogenetic information from hundreds of loci: An example from the Compositae. *Applications in Plant Sciences* 2: 1300085. https://doi.org/10.3732/apps.1300085

McCormack, J.E., Huang, H. & Knowles, L.L. 2009. Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Systematic Biology* 58: 501–508. https://doi.org/10.1093/sysbio/syp045Medina, R., Johnson, M., Liu, Y., Wilding, N., Hedderson, T.A., Wickett, N. & Goffinet, B. 2018. Evolutionary dynamism in bryophytes: Phylogenomic inferences confirm rapid radiation in the moss family Funariaceae. *Molecular Phylogenetics and Evolution* 120: 240–247. https://doi.org/10.1016/j.ympev.2017.12.002

Mehregan, I. & Assadi, M. 2016. A synopsis of *Cousinia* sect. *Pseudactinia* (Cardueae, Asteraceae) including a new species from NE Iran. *Phytotaxa* 257: 271–279. https://doi.org/10.11646/phytotaxa.257.3.5

Mehregan, I. & Kadereit, J.W. 2009. The role of hybridization in the evolution of *Cousinia* s.str. (Asteraceae, Cardueae). *Willdenowia* 39: 35–47. https://doi.org/10.3372/wi.39.39102

Meiklejohn, K.A., Faircloth, B.C., Glenn, T.C., Kimball, R.T. & Braun, E.L. 2016. Analysis of a rapid evolutionary radiation using ultraconserved elements: evidence for a bias in some multispecies coalescent methods. *Systematic Biology* 65: 612–627. https://doi.org/10.1093/sysbio/syw014

Mendes, F.K. & Hahn, M.W. 2018. Why concatenation fails near the anomaly zone. *Systematic Biology* 67: 158–169. https://doi.org/10.1093/sysbio/syx063

Meyer, M. & Kircher, M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* 2010: pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Miller, M.A., Pfeiffer, W. & Schwartz, T. 2010. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010*. New Orleans: 1–8. https://doi.org/10.1109/GCE.2010.5676129

Mirarab, S., Bayzid, M.S., Boussau, B. & Warnow, T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346: 1250463. https://doi.org/10.1126/science.1250463

Misof, B., Meyer, B., von Reumont, B.M., Kück, P., Misof, K. & Meusemann, K. 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics* 14: 348. https://doi.org/10.1186/1471-2105-14-348

Mitchell, N., Lewis, P.O., Lemmon, E.M., Lemmon, A.R. & Holsinger, K.E. 2017. Anchored phylogenomics improves the resolution of evolutionary relationships in the rapid radiation of *Protea* L. *American Journal of Botany* 104: 102–115. https://doi.org/10.3732/ajb.1600227

Moyle, R.G., Oliveros, C.H., Andersen, M.J., Hosner, P.A., Benz, B.W., Manthey, J.D., Travers, S.L., Brown, R.M & Faircloth, B.C. 2016. Tectonic collision and uplift of Wallacea triggered the global songbird radiation. *Nature communications* 7: 12709. https://doi.org/10.1038/ncomms12709

Nicholls, J.A., Pennington, R.T., Koenen, E.J., Hughes, C.E., Hearn, J., Bunnefeld, L., Dexter, K.G., Stone, G.N. & Kidner, C.A. 2015. Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). *Frontiers in Plant Science* 6: 710. https://doi.org/10.3389/fpls.2015.00710

Oliver, J.C. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. *Evolution* 67: 1823-1830. https://doi.org/10.1111/evo.12047

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* 20: 289–290. https://doi.org/10.1093/bioinformatics/btg412

Parks, M., Cronn, R. & Liston, A. 2012. Separating the wheat from the chaff: mitigating the effects of noise in a plastome phylogenomic data set from *Pinus* L. (Pinaceae). *BMC Evolutionary Biology* 12: 100. https://doi.org/10.1186/1471-2148-12-100

Philippe, H. & Forterre, P. 1999. The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution* 49: 509–523. https://doi.org/10.1007/PL00006573

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G. & Baurain, D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biology* 9: e1000602. https://doi.org/10.1371/journal.pbio.1000602

Prum, R.O., Berv, J. S., Dornburg, A., Field, D.J., Townsend, J.P., Lemmon, E.M. & Lemmon, A.R. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526: 569. https://doi.org/10.1038/nature15697

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Raab-Straube, E.V. 2003. Phylogenetic relationships in *Saussurea* (Compositae, Cardueae) sensu lato, inferred from morphological, ITS and trn L-trn F sequence data, with a synopsis of *Himalaiella* gen. nov., *Lipschitziella* and *Frolovia*. *Willdenowia* 33: 379–402. https://doi.org/10.3372/wi.33.33214

Rambaut, A. 2002. *TreeEdit, Version 1.0a10*. Evolutionary Biology Group, University of Oxford, Oxford. Available at http://evolve.zoo.ox.ac.uk

Rambaut, A. 2016. *FigTree ver. 1.4.3*. Department of Zoology, University of Oxford, Oxford. Available at http://tree.bio.ed.ac.uk/software/figtree/

Rensing, S.A. 2014. Gene duplication as a driver of plant morphogenetic evolution. *Current opinion in Plant Biology* 17: 43–48. https://doi.org/10.1016/j.pbi.2013.11.002

Robinson, D.F. & Foulds, L.R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53: 131–147. https://doi.org/10.1016/0025-5564(81)90043-2

Rokas, A. & Carroll, S.B. 2006. Bushes in the tree of life. *PLoS Biology* 4: e352. https://doi.org/10.1371/journal.pbio.0040352

Salichos, L. & Rokas, A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497: 327–331. https://doi.org/10.1038/nature12130

Salmerón-Sánchez, E., Pérez-García, F.J., Medina-Cazorla, J.M., Martínez-Nieto, M.I., Martínez-Hernández, F., Garrido-Becerra, J.A., Mendoza-Fernández, A.J., Merlo Calvente, M.E. & Poveda, J.M. 2015. Genetic analysis based on plastidial and ribosomal sequences of the endemic bi-edaphic taxon *Jurinea pinnata* (Lag.) DC. (Compositae) in the Guadix-Baza Basin. *Plant Biosystems* 149: 922–932. https://doi.org/10.1080/11263504.2014.983203

Sanderson, M.J. 1997. A nonparametric approach of rate constancy to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* 14: 1218–1231.

Sass, C., Iles, W.J., Barrett, C.F., Smith, S.Y. & Specht, C.D. 2016. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* 4: e1584. https://doi.org/10.7717/peerj.1584

Sayyari, E. & Mirarab, S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* 33: 1654–1668. https://doi.org/10.1093/molbev/msw079

Schmickl, R., Liston, A., Zeisek, V., Oberlander, K., Weitemier, K., Straub, S.C., Cronn, R.C., Dreyer, L.L. & Suda, J. 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* 16: 1124–1135. https://doi.org/10.1111/1755-0998.12487

Shi, Z., Raab-Straube, E.V., Greuter, W. & Martins, L. 2011. *Cardueae*. In: Wu, Z.Y., Raven, P.H. & Hong, D.Y. (eds.), *Flora of China*, vol. 20–21 (Asteraceae). Science Press and Missouri Botanical Garden Press, Beijing and St. Louis: 42–194.

Simmons, M.P., Sloan, D.B. & Gatesy, J. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Molecular Phylogenetics and Evolution* 97: 76–89. https://doi.org/10.1016/j.ympev.2015.12.013

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22: 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30: 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stamatakis, A., Hoover, P. & Rougemont, J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* 57: 758–771. https://doi.org/10.1080/10635150802429642

Stebbins, G.L. 1970. Adaptive radiation of reproductive characteristics in angiosperms, I: pollination mechanisms. *Annual Review of Ecology and Systematics* 1: 307–326.

Steel, M.A. & Penny, D. 1993. Distributions of tree comparison metrics—some new results. *Systematic Biology* 42: 126–141. https://doi.org/10.2307/2992536

Stephens, J.D., Rogers, W.L., Mason, C.M., Donovan, L.A. & Malmberg, R.L. 2015. Species tree estimation of diploid *Helianthus* (Asteraceae) using target enrichment. *American Journal of Botany* 102: 910–920. https://doi.org/10.3732/ajb.1500031

Straub, S.C., Moore, M.J., Soltis, P.S., Soltis, D.E., Liston, A. & Livshultz, T. 2014. Phylogenetic signal detection from an ancient rapid radiation: Effects of noise reduction, long-branch attraction, and model selection in crown clade Apocynaceae. *Molecular Phylogenetics and Evolution* 80: 169–185. https://doi.org/10.1016/j.ympev.2014.07.020

Streicher, J.W., Miller, E.C., Guerrero, P.C., Correa, C., Ortiz, J.C., Crawford, A.J., Pie, M.R. & Wiens, J.J. 2018. Evaluating methods for phylogenomic analyses, and a new phylogeny for a major frog clade (Hyloidea) based on 2214 loci. *Molecular Phylogenetics and Evolution* 119: 128–143. https://doi.org/10.1016/j.ympev.2017.10.013

Streicher, J.W., Schulte, J.A. & Wiens, J.J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Systematic Biology* 65: 128–145. https://doi.org/10.1093/sysbio/syv058

Struck, T.H. 2014. TreSpEx–-Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evolutionary Bioinformatics* 10: 51–67. https://doi.org/10.4137/EBO.S14239

Stubbs, R.L., Folk, R.A., Xiang, C.L., Soltis, D.E. & Cellinese, N. 2018. Pseudo-parallel patterns of disjunctions in an Arctic-alpine plant lineage. *Molecular Phylogenetics and Evolution* 123: 88–100. https://doi.org/10.1016/j.ympev.2018.02.016

Susanna, A. & Garcia-Jacas, N. 2007. Tribe Cardueae. In: Kadereit, J.W. & Jeffrey, C. (eds.), *Flowering Plants. Eudicots. Asterales*, vol. 8. In: Kubitzki, K. (ed.), *The families and genera of vascular plants*. Springer Verlag, Berlin: 123–146.

Susanna, A. & Garcia-Jacas, N. 2009. Cardueae (Carduoideae). In: Funk, V.A., Susanna, A., Stuessy, T.F. & Bayer, R.J. (eds.), *Systematics, evolution, and biogeography of Compositae*. IAPT, Vienna: 293–313.

Susanna, A., Garcia-Jacas, N., Hidalgo, O., Vilatersana, R. & Garnatje, T. 2006. The Cardueae (Compositae) revisited: Insights from ITS, *trn*L-*trn*F, and *mat*K nuclear and chloroplast DNA analysis. *Annals of the Missouri Botanical Garden* 93: 150–171. https://doi.org/10.3417/0026-6493(2006)93[150:TCCRIF]2.0.CO;2

Susanna, A., Garcia-Jacas, N., Vilatersana, R. & Garnatje, T. 2003. Generic boundaries and evolution of characters in the *Arctium* group: a nuclear and chloroplast DNA analysis. *Collectanea Botanica* 26: 101–118. https://doi.org/10.3989/collectbot.2003.v26.17

Swofford, D.L. 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0 b10. Sinauer, Sunderland.

Syring, J.V., Tennessen, J.A., Jennings, T.N., Wegrzyn, J., Scelfo-Dalbey, C. & Cronn, R. 2016. Targeted capture sequencing in whitebark pine reveals range-wide demographic and adaptive patterns despite challenges of a large, repetitive genome. *Frontiers in Plant Science* 7: 484. https://doi.org/10.3389/fpls.2016.00484

Townsend, J.P., López-Giráldez, F. & Friedman, R. 2008. The phylogenetic informativeness of nucleotide and amino acid sequences for reconstructing the vertebrate tree. *Journal of Molecular Evolution* 67: 437–447. https://doi.org/10.1007/s00239-008-9142-0

Townsend, J.P., Su, Z. & Tekle, Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. *Systematic Biology* 61: 835–849. https://doi.org/10.1093/sysbio/sys036

Tripp, E.A., Tsai, Y.H.E., Zhuang, Y. & Dexter, K.G. 2017. RADseq dataset with 90% missing data fully resolves recent radiation of *Petalidium* (Acanthaceae) in the ultra-arid deserts of Namibia. *Ecology and Evolution* 7: 7920–7936. https://doi.org/10.1002/ece3.3274

Tscherneva, O.V. 1997. Genus 1578. *Cousinia* Cass. [translated from Russian]. In: Schischkin, B.K. & Bobrov, E.G. (eds.), *Flora SSSR*. Bishen Singh Mahendra Pal Singh and Koeltz Scientific Books, Dehra Dun: 135–442.

Van Dam, M.H., Lam, A.W., Sagata, K., Gewa, B., Laufa, R., Balke, M., Faircloth, B.C. & Riedel, A. 2017. Ultraconserved elements (UCEs) resolve the phylogeny of Australasian smurf-weevils. *PLoS ONE* 12: e0188044. https://doi.org/10.1371/journal.pone.0188044

Vatanparast, M., Powell, A., Doyle, J.J. & Egan, A.N. 2018. Targeting legume loci: A comparison of three methods for target enrichment bait design in Leguminosae phylogenomics. *Applications in Plant Sciences* 6: e1036. https://doi.org/10.1002/aps3.1036

Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A. & Seehausen, O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology* 22: 787–798. https://doi.org/10.1111/mec.12023

Wang, Y. J., Raab-Straube, E.V., Susanna, A. & Liu, J.Q. 2013. *Shangwua* (Compositae), a new genus from the Qinghai-Tibetan Plateau and Himalayas. *Taxon* 62: 984–996. https://doi.org/10.12705/625.19

Wang, Y.J., Liu, J.Q. & Miehe, G. 2007. Phylogenetic origins of the Himalayan endemic *Dolomiaea*, *Diplazoptilon* and *Xanthopappus* (Asteraceae: Cardueae) based on three DNA regions. *Annals of Botany* 99: 311–322. https://doi.org/10.1093/aob/mcm284

Wang, Y.J., Susanna, A., Raab-Straube, E.V., Milne, R. & Liu, J.Q. 2009. Island-like radiation of *Saussurea* (Asteraceae: Cardueae) triggered by uplifts of the Qinghai–Tibetan Plateau. *Biological Journal of the Linnean Society* 97: 893–903. https://doi.org/10.1111/j.1095-8312.2009.01225.x

Wanke, S., Mendoza, C.G., Müller, S., Guillén, A.P., Neinhuis, C., Lemmon, A.R., Lemmon, E.M. & Samain, M.S. 2017. Recalcitrant deep and shallow nodes in *Aristolochia*

(Aristolochiaceae) illuminated using anchored hybrid enrichment. *Molecular Phylogenetics and Evolution* 117: 111–123. https://doi.org/10.1016/j.ympev.2017.05.014

Ward, P.S. & Branstetter, M.G. 2017. The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. *Proceedings of the Royal Society B* 284: 20162569. https://doi.org/10.1098/rspb.2016.2569

Watanabe, K. 2002. *Index to chromosome numbers in Asteraceae*. Available at http://www.lib.kobe-u.ac.jp/infolib/meta_pub/G0000003asteraceae_e

Weitemier, K., Straub, S.C., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. & Liston, A. 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. *Applications in Plant Sciences* 2: 1400042. https://doi.org/10.3732/apps.1400042

Wen, J., Ree, R.H., Ickert-Bond, S.M., Nie, Z. & Funk, V. 2013. Biogeography: Where do we go from here? *Taxon* 62: 912–927. https://doi.org/10.12705/625.15

Wen, J., Zhang, J.-Q., Nie, Z.-L., Zhong Y. & Sun, H. 2014. Evolutionary diversifications of plants on the Qinghai-Tibetan Plateau. *Frontiers in Genetics* 5: 4. https://doi.org/10.3389/fgene.2014.00004

Whitfield, J.B. & Lockhart, P.J. 2007. Deciphering ancient rapid radiations. *Trends in Ecology and Evolution* 22: 258–265. https://doi.org/10.1016/j.tree.2007.01.012

Xi, Z., Liu, L., Rest, J.S. & Davis, C.C. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. *Systematic Biology* 63: 919–932. https://doi.org/10.1093/sysbio/syu055

Zhang, C., Rabiee, M., Sayyari, E. & Mirarab, S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC bioinformatics* 19: 153. https://doi.org/10.1186/s12859-018-2129-y

TABLES

Table 1. Strategies of filtering target enrichment sequencing data before phylogenetic analyses.

Filter by	Criteria of filtering	Software that can be used	Examples of studies applying the filtering method
SPECIES	Exclusion of quickly evolving or unstable species	PHYLUCE[1] HybPhyloMarker[2]	Salichos and Rokas (2013) Ješovnik et al. (2017) Streicher et al. (2018) Gates et al. (2018)
POSITIONS	Exclusion of sites with high substitution rates	PhyDesign[3] OV[4] TIGER[5]	Dornburg et al. (2017) Fragoso-Martínez et al. (2017) Wanke et al. (2017) Streicher et al. (2018)
	Exclusion of sites containing gaps	trimAL[6]	Salichos and Rokas (2013)
	Inclusion of sites with high substitution rates	TIGER[5]	Streicher et al. (2018)
	Inclusion of positions with high read coverage	Custom scripts and ape[7]	Grover et al. (2015)
LOCI	Exclusion of loci with low taxa recovering	HybPhyloMarker[2] PHYLUCE[8]	Borowiec et al. (2015) Hosner et al. (2015) Streicher et al. (2016) Ješovnik et al. (2017) Longo et al. (2017) Mitchell et al. (2017) Streicher et al. (2018) Gernandt et al. (2018)
	Exclusion of loci detected as potential paralog	HybPiper[9]	Crowl et al. (2017) Chau et al. (2018) Gernandt et al. (2018) Vatanparast et al. (2018)
	Exclusion of loci of short length	Geneious[10]	Gernandt et al. (2018)
	Exclusion of highly variable loci	Geneious[10]	Gernandt et al. (2018)
	Exclusion of loci with high number of missing data	Geneious[10]	Gernandt et al. (2018)
	Exclusion of poorly aligned loci	Not specified	Salichos and Rokas (2013)
	Exclusion of loci with low long-branch score from long-branched species	HybPhyloMarker[2] TreSpEx[11] R script[12]	Borowiec et al. (2015)
	Inclusion of loci with strong phylogenetic signal (based on gene-trees with high bootstrap values average)	HybPhyloMarker[2] TreSpEx[11] R script[12] Newick utilities[13]	Salichos and Rokas (2013) Bossert et al. (2017) Branstetter et al. (2017) Ješovnik et al. (2017) Ward and Branstetter (2017)
	Inclusion of the most informative loci (high informative characters or parsimony informative sites)	HybPhyloMarker[2] PhyDesign[3] AMAS[14] Phyloch[15]	Hosner et al. (2015) Léveillé-Bourret et al. (2016) Meiklejohn et al. (2016) Longo et al. (2017)
	Inclusion of slowly evolving loci (based on smallest average of branch lengths)	HybPhyloMarker[2] TreSpEx[11] R script[12]	Salichos and Rokas (2013) Borowiec et al. (2015)
	Inclusion of less saturated loci	R script[12]	Borowiec et al. (2015)
	Inclusion of the most informative loci scored by some of the previous metrics	HybPhyloMarker[2] HybPiper[9] Geneious[10] TreSpEx[11] R script[12]	Borowiec et al. (2015) Gernandt et al. (2018)

[1]PHYLUCE (Faircloth, 2015) script "PHYLUCE_align_extract_taxa_from_alignments.py"; [2]HybPhyloMarker pipeline package (Fér and Schmickly, 2018); [3]PhyDesign online application (López-Giráldez and Townsend, 2011; http://phydesign.townsend.yale.edu/); [4]OV (observed variability) algorithm (Goremykin et al. 2010); [5]TIGER software (Cummins

and McInerey, 2011); [6]trimAL program (Capella-Gutiérrez et al., 2009); [7]R package "ape" (Paradis et al., 2004) and custom scripts (Grover et al., 2015) available at https://github.com/Wendellab/phylogenetics; [8]PHYLUCE (Faircloth, 2015) script "get_only_loci_with_min_taxa.py"; [9]HybPiper pipeline (Johnson et al., 2016), script "paralog_investigator.py"; [10]Geneious software (Kearse et al., 2012); [11]TreSpEx pipeline package (Struck, 2014); [12]R script "gene_stats.R" (Borowiec et al. 2015); [13]Newick utilities package (Junier and Zdobnov, 2010), function "nw_ed"; [14]AMAS software (Borowiec, 2016); [15]R package "phyloch" (Heibl, 2008), function pis.

Table 2. Comparison of the extraction performance of the 1061 COS targets (Mandel et al., 2014) of the methods compared, PHYLUCE (Faircloth, 2015) and HybPiper (Johnson et al., 2016), over the unfiltered datasets. The evaluated parameters from 3. to 10. were calculated based on the dataset specified in the parameter 2. (i.e. from the total recovered loci, removing those loci with only one or two species). Abbreviation used: max = maximum; min = minimum; Nº = number; sd = standard deviation.

Evaluated parameters	PHYLUCE unfiltered dataset	HybPiper unfiltered dataset
1. Nº of total recovered loci (%)	675 (63.3)	1055 (99.4)
2. Nº of total recovered loci removing those captured only for one or two species (%)	627 (59.1)	1051 (99.1)
3. Nº of captured loci for 90% or more species (%)	58 (9.2)	942 (89.6)
4. Average of recovered loci per species (sd; min–max)	340 (37; 208–410)	989 (68.2; 510–1061)
5. Average of number of species recovered per loci (sd; min–max)	46 (24.3; 4–84)	80 (12; 4–85)
6. Mean alignment length per locus in bp (sd; min–max)	823 (450; 139–3134)	317 (185; 63–1475)
7. Nº of loci longer than respective target length (%)	521 (77.2)	161 (15.2)
8. Length of final concatenated matrix in bp	515,875	333,614
9. Proportion of missing data in the concatenated matrix (%)	57.1	9.5
10. Proportion of variable sites in the final concatenated matrix (%)	48	36

Table 3. Characteristics of the datasets obtained with the two target extraction methods (PHYLUCE and HybPiper) under different strategies of matrix filtering: positions and loci. All concatenated matrices are included in Supplementary Material. Abbreviations used: bp = base pairs; BS = bootstrap support; LPP = local posterior probability; concat = concatenation approach; coalesc = coalescence approach; Nº = number; SR = substitution rate.

Dataset name	Nº of loci	Alignment length (bp)	Missing data (%)	Variable sites in bp (%)	Support mean (BS / LPP)	Number of unsupported nodes (concat / coalesc)	Description of data filtering
PHYLUCE_627	627	515,875	57.1	247,320 (48)	87.2 / 0.88	17 / 28	Unfiltered
PHYLUCE_675_5	627	514,460	57.0	245,905 (48)	87.0 / 0.89	16 / 28	Filtering SR > 5: removing 1415 (0.3%) characters from 93 (14.8%) loci
PHYLUCE_675_2.5	627	513,490	57.0	244,944 (48)	86.0 / 0.88	16 / 28	Filtering SR > 2.5: removing 2385 (0.5%) characters from 131 (20%) loci
PHYLUCE_675_1	627	506,631	56.7	238,076 (47)	94.7 / 0.88	7 / 26	Filtering SR > 1: removing 9244 (1.8%) characters from 252 (40.2%) loci
PHYLUCE_304	304	234,118	52.6	103,947 (44)	90.0 /0.85	9 / 33	48% of original loci scoring best in taxon occupancy, average bootstrap, saturation and evolution rate
HybPiper_1051	1051	333,614	9.5	118,542 (36)	91.2 / 0.87	11 / 28	Unfiltered
HybPiper_1051_5	1051	333,576	9.5	118,504 (36)	91.6 / 0.88	10 / 28	Filtering SR > 5: removing 38 (0.01%) characters from 9 (0.9%) loci
HybPiper_1051_2.5	1051	333,556	9.5	118,484 (36)	91.4 / 0.87	12 / 28	Filtering SR > 2.5: removing 58 (0.02%) characters from 18 (1.7%) loci

HybPiper_1051_1	1051	331,729	9.4	116,657 (35)	93.4 / 0.88	5 / 29	Filtering SR > 1: removing 1885 (0.6%) characters from 213 (20.3%) loci
HybPiper_570	570	200,632	7.5	70,774 (35)	92.6 / 0.86	8 / 31	54% of original loci scoring best in taxon occupancy, average bootstrap, saturation and evolution rate

1450

1451

FIGURE CAPTIONS

Fig. 1. Workflow representation of bioinformatic and phylogenetic analyses. The process followed consists of two different methods of target sequence extraction, PHYLUCE (Faircloth, 2015) and HybPiper (Johnson et al., 2016), and two approaches of sequence data refining: filtering by positions (Fragoso-Martínez et al., 2017) and filtering by loci (Borowiec et al., 2015). Squares in red and blue represent all the datasets analyzed (see Table 3 for details), in red showing analyses performed under the concatenation approach and in blue under the coalescence approach. The main programs used for the analyses are shown in brackets.

Fig. 2. Recovery efficiency for 1061 COS loci using two target extraction methods: (A) PHYLUCE (Faircloth, 2015), and (B) HybPiper (Johnson et al., 2016). Columns represent each target locus and rows the 85 sampled species. The cells of heat map in black represent loci on-target, and missing loci are showed in grey. Differences of length in base pairs (bp) between reference target and captured sequence (not aligned and trimmed) are represented for PHYLUCE dataset (C) and for HybPiper dataset (D). Blue bars represent loci shorter than the corresponding target and red bars represent loci that exceed the corresponding target in length. When target length is equal to captured locus length, value of y-axis is zero.

Fig. 3. Phylogenetic trees drawn opposite to each other (tanglegram) inferred from: (A) the concatenation approach from the HybPiper dataset filtering positions with substitution rates > 1, and (B) the coalescence approach from the unfiltered HybPiper dataset. Only bootstrap support values < 70 and local posterior probabilities < 0.95 are shown over branches. Continuous lines that link the species represent congruent positions between both trees and dashed lines represent incongruent topologies. The section where each species belongs is specified in brackets, except for *Jurinea* (see text for details). Taxonomic treatments followed are López-Vinyallonga et al. (2011) for *Arctium*, Tscherneva (1997) for *Cousinia*, and Lipschitz (1979) for *Saussurea*. The species with a superscript were originally described under a different genus within the *Saussurea-Jurinea* complex: [1]*Saussurea leptophylla* = *Jurinea ancistrophylla*; [2]*Saussurea carduicephala* = *Jurinea ceratocarpa* = *Lipschitziella*; [3]*Jurinea carduiformis* = *Outreya carduiformis*; [4]*Modestia darwasica* = *Jurinea sp*.

Fig. 4. Phylogenetic informativeness (PI) analyses showing ultrametric trees scaled to an arbritary scale of 1 (at the root) to 0 (at the tips) obtained with maximum likelihood analyses using the concatenation approach, and net phylogenetic informativeness profiles displaying curves for each locus in different colors. In this figure, we show the PI analyses performed to the PHYLUCE

dataset, in particular (A) to the unfiltered matrix and (B) to the same matrix after filtering positions with substitution rates (SR) > 1, and to the HybPiper dataset, in particular (C) to the unfiltered matrix, and (D) to the same matrix after filtering positions with SR > 1. For a threshold of SR > 5 and SR > 2.5 see Supplementary Figs. S5 and S6. Branches with low support values (bootstrap < 70) are marked and highlighted in red. The number of unsupported nodes is specified for each genus at the right of the tree to see the differences between unfiltered alignments and the best filtering positions scheme (SR > 1).

Fig. 5. Variation in support values across 50% of the less supported nodes, ranked from the minimum support to the maximum support obtained, according to different filtering treatments: unfiltered alignments, filtering positions with substitution rates (SR) higher than 5, 2.5 and 1, and selecting the most informative loci. Support values were extracted from trees obtained with the concatenation approach, using (A) the dataset obtained with the PHYLUCE extracting method and (B) the dataset obtained with the HybPiper extracting method, and from species trees obtained with the coalescence approach, using (C) the dataset obtained with the PHYLUCE extracting method and (D) the dataset obtained with the HybPiper extracting method. Abbreviations used: BS = bootstrap support; LPP = local posterior probability.

Fig. 6. Number of unsupported nodes represented in a color scale for all the executed analyses with the concatenation approach (considering nodes with bootstrap values < 70) and with the coalescence approach (local posterior probabilities < 0.95). For the two approaches, both types of target extraction methods are considered: PHYLUCE and HybPiper. Columns represent the four genera examined: *Arctium*, *Cousinia*, *Saussurea* and *Jurinea*, and rows the different analyzed datasets (see Fig. 1): alignments, filtering positions with substitution rates higher than 5, 2.5 and 1, and filtering loci following criteria of taxon occupancy, support content, saturation, and evolution rate.

Fig. 7. Tree space from a multidimensional scaling of Robinson-Foulds (RF) pairwise distance comparisons between all topologies of trees inferred. Trees obtained with the PHYLUCE extracting method are represented in squares and trees obtained with the HybPiper extracting method are represented in circles. Trees that resulted in equal topology to unfiltered alignments are displayed in orange, i.e. as the same color of unfiltered alignments. The same topologies or almost equivalent with RFadj = 0 or RFadj = 0.01 are, for the coalescence approach, between the PHYLUCE dataset among unfiltered and substitution rates (SR) > 5 and 2.5 and, for the concatenation approach, between unfiltered alignments and the smoothest filtering by positions (SR > 5 and 2.5) for both dataset cases of PHYLUCE and HybPiper.

Supplementary Tables

Table S1. Species sampled, authority and collection information.

Species	Authority	Location and voucher
Alfredia acantholepis	Kar. & Kir.	Kazakhstan: Almatinskaya oblast, Alatau mt. above Almaty, *Susanna 2092 et al.* (BC)
Arctium abolinii	(Kult. ex Tscherneva) S. López et al.	Kyrgyzstan: SW, Jalal Abad Oblast, Kara Saj Tal, Aksy Rayan, *Lazkov s.n.* (LE)
Arctium arctioides	(Schrenk.) Kuntze	Kazakhstan: Dzhezkazganskaya reg., Turgajskaya lowland, 49 km to SW from Dzhezkazgana, *Kamelin 6434* (LE)
Arctium aureum	(C. Winkl.) Kuntze	Tajikistan: Schtut, road to Penjikent, *Susanna 2514 et al.* (BC)
Arctium egregium	(Juz.) S. López et al.	Uzbekistan: Angren valley, rise to Kamchik pass, *Kamelin 420* (LE)
Arctium eriophorum	(Regel & Schmalh.) Kuntze	Kazakhstan: Almatinskaya oblast, Alatau mt., above Almaty, *Susanna 2088 et al.* (BC)
Arctium fedtschenkoanum	(Bornm.) S. López et al.	Tajikistan: *Romaschenko 632 & Susanna* (BC)
Arctium grandifolium	(Kult.) S. López et al.	Kazakhstan: Zambylskaya oblast, Talaski Alatau, 6 km W from Il Tai, *Susanna 2181 et al.* (BC)
Arctium karatavicum	(Regel & Schmalh.) Kuntze	Tadjikistan: *Romanshenko 607 & Susanna* (BC)
Arctium leiospermum	Juz. & Ye. V. Serg.	Kazakhstan: Zambylskaya oblast, Kurdai pass, *Susanna 2154 et al.* (BC)
Arctium minus	(Hill) Bernh.	Spain: Barcelona, Dosrius, Canyamars, dry riverbed of Canyamars, *Vilatersana 1100 & López-Vinyallonga* (BC)
Arctium umbrosum	Kuntze	Kazakhstan: Almatinskaya oblast, Alatau mt. above Almaty, *Susanna 2100 et al.* (BC)
Carduus pycnocephalus	L.	Spain: Barcelona, Montjuïc, near the Botanic Institute of Barcelona, *Garnatje & Susanna 1827* (BC)
Cirsium sairamense	O. Fedtsch. & B. Fedtsch.	Tadjikistan: Maijora canyon near Ziddi, *Susanna 2468 et al.* (BC)
Cousinia albertoregelia	C. Winkl.	Tadjikistan: Tujuntau mountains, west of Shaaftuz, *Botschantzev 166* (LE)
Cousinia armena	Takht.	Armenia: Kotayk, Abovian, *Vitek 03-07458 et al.* (BC)
Cousinia badghysi	Kult.	Turkmenistan: Badghys, Eroylanduz lake, SE "sopki" Kazan, *Kamelin 360* (LE)
Cousinia brachyptera	DC.	Armenia: Shirak province, Talin district, Arteni mountain, *Tamanian s.n.* (ERE)
Cousinia coerulea	Kult. ex Tscherneva	Tajikistan: Vorzov canyon, *Susanna 2459 et al.* (BC)
Cousinia fetissowii	C. Winkl.	Kyrgyzstan: Kirghizia, Kyrgyz Ala-Too, upper reaches of Nyldy River, *Sudnitsyna & Gorbunova s.n.* (FRU)
Cousinia franchetii	C.Winkl.	Tadjikistan: near kishlag Zimargh, *Susanna 2498 et al.* (BC)
Cousinia knorringiae	Bornm.	Kyrgyzstan: Bozbu-Too, 21.V.1970, *Sudnitsyna s.n.* (FRU, LE)
Cousinia macroptera	C. A. Mey. ex DC.	Armenia: Ararat province, Ashtarak distr., Kahtsrashen village, *Tamanian s.n.* (ERE)
Cousinia ninae	Juz.	Kyrgyzstan: Oshskaya, Torgulsky reg., Oitaya area north from Shoporovo vil., *Sultanova s.n.* (LE)
Cousinia onopordioides	Ledeb.	Iran: Tehran, between Firuzkuh and Semnan, *Susanna 1637 et al.* (BC)
Cousinia polytimetica	Tscherneva	Uzbekistan: Bukharskaya reg., Zeravshan river, to SE from Uzlishkent vil., *Kryakin s.n.* (LE)
Cousinia pusilla	C. Winkl.	Tajikistan: S Tajikistan, from Besharcha mts. to Babatag range, *Botschantzev 117* (LE)
Cousinia schischkinii	Juz.	Kyrgyzstan: Kara-Suu Lake, Nura River, 17.VI.1973, *Aydarova et al. s.n.* (FRU)
Cousinia serawschanica	C. Winkl.	Tajikistan: Voru, *Susanna 2516 et al.* (BC)
Cousinia sewertzowii	Regel	Kazakhstan: Aksu-Dzabagly reservation, *Susanna 2178 et al.* (BC)
Cousinia sogdiana	Bornm.	Uzbekistan: Karakalpakstan, 27 km from Nukus, *Kalibernova 5262 et al.* (LE)
Cousinia splendida	C. Winkl.	Tadjikistan: Fan mountains, road above Iskandar-Kul, *Susanna 2507 et al.* (BC)

Cousinia spryginii	Kult.	Uzbekistan: Kashkadarbinskaya reg., low mountains to SE of vil. Dekhanabad, *Botschantzev 46* (LE)
Cousinia strobilocephala	Tschern. & Vved.	Kyrgyzstan: Kirghizia, Qurama Range, Kayyndy-Say River, *Aydarova & Chypaev s.n.* (FRU)
Cousinia tenella	Fisch. & C. A. Mey.	Iran: Golestan Nat. Park, between Sharlegh and Cheshmeh Khan, *Akhani 243* (MJG)
Cousinia tianschanica	Kult.	Kazakhstan: Shimkientskaya oblast, Aksu Dzabagly reservation, Aksu canyon, *Susanna 2191 et al.* (BC)
Cynara cardunculus	L.	United States of America: Greenhouse grown seed, collected UW Medicinal Plant Garden, *Mandel s.n.* (GA 135)
Jurinea abramowii	Regel & Herder	Tadjikistan: Hissar Mt., *Smirnova 224 et al.* (DUSH)
Jurinea alata	(Desf.) Cass.	Culta in Horto Botanico Barcinonense (BC)
Jurinea algida	Iljin	Kyrgyzstan: Kok-Suu River, 16.VIII.2006, *Lazkov s.n.* (FRU)
Jurinea atropurpurea	C. Winkl. ex Iljin	Tadjikistan: sine loc, *Kotehkariova & Zhogolieva 16094* (DUSH)
Jurinea baldschuanica	C. Winkl.	Tadjikistan: mountains above Kara-Chuiráá, *Susanna 2561 et al.* (BC)
Jurinea bucarica	C. Winkl.	Sine loc. nec col., 22.IV.1975, 10387 (DUSH)
Jurinea caespitans	Iljin	Kyrgyzstan: north of Kara-Jygach village, 09.VII.2016, *Sennikov s.n.* (H)
Jurinea capusii	Franch.	Kyrgyzstan: Chapchyma-Say, 14.VII.2016, *Sennikov s.n.* (H)
Jurinea carduiformis	(Jaub. & Spach) Boiss.	Iran: Tehran, near Sorkhehesar, *Susanna 1631 et al.* (BC)
Jurinea ferganica	(Iljin) Iljin	Kyrgyzstan: near Kadamzhay village, 18.VII.2016, *Sennikov s.n.* (H)
Jurinea fontqueri	Cuatrec.	Spain: Jaén, cerro Cárceles, Mágina, *Martínez Lirola s.n.* (GDA 44615)
Jurinea kokanica	Iljin	Kyrgyzstan: 15 km E of Kosh-Bulak village, 09.V.2007, *Ganybaeva s.n.* (FRU)
Jurinea kyzylkyrensis	Kamelin & Tscherneva	Kyrgyzstan: left side of Naryn River, Kyzyl-Kyr, 12.VIII.1979, *Botschantzev et al. s.n.* (FRU)
Jurinea lanipes	Rupr.	Kyrgyzstan: Boom ravine, *Sennikov 428a* (H) [locus classicus of *Jurinea abolinii* Iljin]
Jurinea leptoloba	DC.	Iran: 30 km N from Tabriz, *Susanna 1654 et al.* (BC)
Jurinea macrocephala	DC.	Iran: 20 Km N of Qarabchaman, *Susanna 1650 et al.* (BC)
Jurinea narynensis	Kamelin & Tscherneva	Kyrgyzstan: 8 km from Tash-Kumyr to Jangi-Jol, *Lazkov & Omuralieva 11* (FRU)
Jurinea olgae	Ragel & Schmalh.	Tadjikistan: slopes over kishlag Voru, *Susanna 2517 et al.* (BC)
Jurinea orientalis	(Iljin) Iljin	Kyrgyzstan: near Shekoftar village, 13.VII.2016, *Sennikov s.n.* (H)
Jurinea pinnata	DC.	Morocco: Meknès-Tafilalt, Middle-Atlas, from Midelt to Timahdite, col du Zad, *Calleja & Hipold 20103091* (BC)
Jurinea popovii	Iljin	Tadjikistan: sine loc., *Chukavina et al. 163(86)* (DUSH)
Jurinea schachimardanica	Iljin	Kyrgyzstan: sine loc., 2016, *Sennikov s.n.* (H)
Jurinea stenophylla	Iljin	Kyrgyzstan: Kasan-Say River near Terek-Say village, 14.VI.1996, *Pimenov et al. s.n.* (FRU)
Jurinea stoechadifolia	(M. Bieb.) DC.	Ukraine: Crimea, *Romo 10321 et al.* (BC)
Jurinea suffruticosa	Regel	Kyrgyzstan: Kasan-Say River, 14.VII.2016, *Sennikov s.n.* (H)
Jurinea thianschanica	Regel & Schmalh.	Kyrgyzstan: between Kochkor and Ottuk, near Orto-Tokoy village, 03.VII.2016, *Sennikov s.n.* (H)
Jurinea trautvetteriana	Regel & Schmalh.	Tadjikistan: sine loc., *Ovczinnikov 16305 & Zaprjagaeva* (DUSH)
Jurinea winkleri	Iljin	Kyrgyzstan: east of Uch-Korgon village, 16.VII.2016, *Sennikov s.n.* (H)
Modestia darwasica	(C. Winkl.) Kharadze & Tamamsch.	Kyrgyzstan: 20 km NW of Samarkandyk, Kyzyl-Suu, 10.V.1978, *Aidarova & Ubukeeva s.n.* (FRU)
Jurinea xeranthemoides	Iljin	Kyrgyzstan: near Uch-Korgon village, 16.VII.2016, *Sennikov s.n.* (H)
Olgaea petriprimi	B. A. Sharipova	Tadjikistan: Kishlag Selandi, *Susanna 2539 et al.* (BC)
Saussurea carduicephala	(Iljin) Iljin	Tajikistan: Gorno-Badakhshan, Shughnon, Shughnonskii Ridge, *Semakov & Dengubenko s.n.* (LE 8428)

Saussurea controversa	DC.	Russia: Krasnoyarsk Krai, Sharypovsky, village Bolshoe ozero, *A. Pyak, E. Pyak & Cazzolla Gatti 10005* (TK t-01-2016)
Saussurea davurica	Adams	Russia: Altai, Kosh-Agachsky, Kuraiskiy Ridge, village Chagan-Usun, *A. Pyak & E. Pyak 11049* (TK a-067-2016)
Saussurea elegans	Ledeb.	Tadjikistan: Iskandar valley, Fan mountains, *Susanna 2505 et al.* (BC)
Saussurea foliosa	Ledeb.	Russia: Khakassia, Tashtypsky, Sayanskii Mountain Pass, *A. Pyak, E. Pyak & Cazzolla Gatti 10025* (TK t-30-2016)
Saussurea glacialis	Herder	Russia: Altai, Kosh-Agachsky, Kuraiskiy Ridge, *A. Pyak & E. Pyak 11021* (TK a-043-2016)
Saussurea jadrinzevii	Krylov	Russia: Altai, Ongudaysky, the Mount Belyy Bom, *A. Pyak & E. Pyak 11005* (TK a-023-2016)
Saussurea krylovii	Schischk. & Serg.	Russia: Altai, Kosh-Agachsky, Juzhno-Chuysky Ridge, the Jazator River Valley, *A. Pyak & E. Pyak 11079* (TK a-108-2016)
Saussurea larionowii	C. Winkl.	Kyrgyzstan: sine loc., *Ovczinnikov 16* (DUSH)
Saussurea latifolia	Ledeb.	Russia: Krasnoyarsk Krai, Yermakovsky, Ergaki Ridge, *A. Pyak & E. Pyak 10009* (TK t-02-2016)
Saussurea leptophylla	Hemsl.	Afghanistan: Kapisa, *Podlech 12500* (W)
Saussurea leucophylla	Schrenk	Russia: Altai, Kosh-Agachsky, northern spurs of the Mount Tjepliy Kljuch, *A. Pyak & E. Pyak 11073* (TK a-102-2016)
Saussurea manshurica	Kom.	Russia: Amur province, 02.VIII.1979, *Boyko & Starchenko s.n.* (LE)
Saussurea orgaadayi	Khanm. & Krasnob.	Russia: Altai, Kosh-Agachsky, Kuraiskiy Ridge, the Kokorja River Valley, *A. Pyak & E. Pyak 11083* (TK a-119-2016)
Saussurea sp.	N. D. Simpson	Russia: Altai, Kosh-Agachsky, Kuraiskiy Ridge, village Chagan-Usun, near Lake Balhash, *A. Pyak & E. Pyak 11044* (TK a-065-2016)
Saussurea pseudoalpina	Simps.	Russia: Altai, Kosh-Agachsky, Kuraiskiy Ridge, the Ortolyk River, *A. Pyak & E. Pyak 11032* (TK a-048-2016)
Saussurea salicifolia	(L.) DC.	Russia: Tyva, Kaa-Khemsky, the Mount Ondum, the Kaa-Khem River, *A. Pyak & E. Pyak 10014* (TK t-12-2016)
Saussurea salsa	(Pall. ex Pall.) Spreng.	Russia: Altai, Kosh-Agachsky, Chuya Steppe, village Aktal, *A. Pyak & E. Pyak 11087* (TK a-120-2016)
Saussurea schanginiana	Fisch. ex Herder	Russia: Khakassia, Tashtypsky, Sayanskii Ridge, Sayanskii Mountain Pass, *A. Pyak & E. Pyak 10057* (TK t-24-2016)
Saussurea stubendorffii	Herder	Russia: Tyva, Barun-Khemchiksky, Sayanskii Ridge, Ak-sug River Valley, *A. Pyak & E. Pyak 10057* (TK t-24-2016)
Saussurea subacaulis	(Ledeb.) Serg.	Russia: Altai, Kosh-Agachsky, Kuraiskiy Ridge, Ortolyk River, *A. Pyak & E. Pyak 11026* (TK a-046-2016)

Table S2. Species sampled and their corresponding number of raw reads, and number of informative and missing loci recovered with the PHYLUCE and the HybPiper methods.

		PHYLUCE method		**HybPiper method**	
Species	**Nº of raw reads**	**Nº of recovered COS loci**	**Nº of missing COS loci**	**Nº of recovered COS loci**	**Nº of missing COS loci**
Alfredia acantholepis	8,217,881	337	338	510	545
Arctium abolinii	1,853,731	300	375	1003	52
Arctium arctioides	4,754,092	350	325	1006	49
Arctium aureum	4,565,685	294	381	1018	37
Arctium egregium	1,570,769	342	333	1008	47
Arctium eriophorum	2,984,349	327	348	1008	47
Arctium fedtschenkoanum	3,550,984	289	386	1006	49
Arctium grandifolium	2,300,488	324	351	1004	51
Arctium karatavicum	2,976,529	276	399	1010	45
Arctium leiospermum	3,462,345	336	339	1006	49
Arctium minus	10,007,019	350	325	1008	47
Arctium umbrosum	4,663,613	324	351	1012	43
Carduus pycnocephalus	741,845	336	339	667	388
Cirsium sairamense	5,389,901	370	305	988	67
Cousinia albertoregelia	2,524,381	305	370	999	56
Cousinia armena	2,341,432	307	368	1003	52
Cousinia badghysi	2,650,319	208	467	1017	38
Cousinia brachyptera	2,620,531	310	365	1007	48
Cousinia coerulea	2,171,772	336	339	1004	51
Cousinia fetissowii	6,097,776	332	343	1009	46
Cousinia franchetii	3,678,564	303	372	1008	47
Cousinia knorringiae	3,129,866	338	337	1007	48

Cousinia macroptera	2,298,897	295	380	1013	42
Cousinia ninae	3,280,858	303	372	1006	49
Cousinia onopordioides	2,369,328	243	432	1008	47
Cousinia polytimetica	2,905,323	246	429	1002	53
Cousinia pusilla	3,383,791	264	411	1008	47
Cousinia schischkinii	2,694,939	296	379	1013	42
Cousinia serawschanica	4,345,972	266	409	1016	39
Cousinia sewertzowii	4,077,710	327	348	1006	49
Cousinia sogdiana	3,949,050	315	360	1000	55
Cousinia splendida	3,595,227	310	365	1013	42
Cousinia spryginii	3,305,209	337	338	1007	48
Cousinia strobilocephala	4,105,769	325	350	1014	41
Cousinia tenella	5,279,465	412	263	1005	50
Cousinia tianschanica	2,113,147	342	333	1003	52
Cynara cardunculus	454,885	424	251	796	259
Jurinea abramowii	4,803,672	381	294	993	62
Jurinea alata	5,069,639	386	289	1002	53
Jurinea algida	3,743,171	375	300	991	64
Jurinea atropurpurea	4,316,866	354	321	1002	53
Jurinea baldschuanica	5,113,980	351	324	999	56
Jurinea caespitans	4,407,313	351	324	992	63
Jurinea capusii	4,726,638	374	301	991	64
Jurinea carduiformis	5,200,789	383	292	983	72
Jurinea ferganica	5,170,117	348	327	999	56
Jurinea fontqueri	5,240,423	386	289	993	62
Jurinea kokanica	4,531,178	376	299	992	63
Jurinea kyzylkyrensis	5,561,006	362	313	998	57
Jurinea lanipes	4,601,775	375	300	995	60
Jurinea leptoloba	5,487,798	378	297	996	59
Jurinea macrocephala	4,093,061	374	301	985	70
Jurinea narynensis	4,564,064	374	301	989	66
Jurinea olgae	4,941,133	369	306	992	63
Jurinea orientalis	3,155,790	361	314	990	65
Jurinea pinnata	2,996,426	368	307	996	59
Jurinea popovii	3,304,462	367	308	999	56
Jurinea schachimardanica	3,568,519	368	307	994	61
Jurinea stenophylla	3,240,161	370	305	999	56
Jurinea stoechadifolia	4,403,856	323	352	1002	53
Jurinea suffruticosa	2,658,663	362	313	1000	55
Jurinea trautvetteriana	2,087,532	377	298	993	62
Modestia darwasica	5,083,617	380	295	993	62
Olgaea petriprimi	5,310,933	339	336	1001	54
Saussurea carduicephala	7,948,211	348	327	1016	39
Saussurea controversa	8,091,449	349	326	1013	42
Saussurea davurica	11,202,023	376	299	994	61
Saussurea elegans	2,784,084	359	316	997	58
Saussurea foliosa	4,089,960	340	335	1014	41
Saussurea glacialis	4,072,633	368	307	1006	49
Saussurea jadrinzevii	9,091,105	351	324	1010	45
Saussurea krylovii	3,576,809	356	319	1006	49
Saussurea larionowii	4,733,404	344	331	1001	54
Saussurea latifolia	5,065,459	335	340	1007	48
Saussurea leptophylla	6,055,256	366	309	1013	42
Saussurea leucophylla	5,597,695	352	323	1010	45
Saussurea manshurica	4,417,126	348	327	1010	45
Saussurea orgaadayi	3,578,510	378	297	1002	53
Saussurea sp.	3,884,640	364	311	998	57
Saussurea pseudoalpina	3,887,786	334	341	1012	43
Saussurea salicifolia	4,799,838	312	363	1000	55
Saussurea salsa	2,458,299	377	298	996	59
Saussurea schanginiana	4,568,611	366	309	1008	47
Saussurea stubendorffii	5,329,546	309	366	1016	39
Saussurea subacaulis	8,252,488	343	332	1013	42
Average (± standard deviation)	4,263,196 (± 1,822,355)	341.2 (± 37.4)	333.8 (± 37.4)	991.1 (± 67.9)	63.9 (± 67.9)

Table S3. Pairwise comparisons of tree topologies obtained from both the concatenation and coalescence approaches using the Robinson-Foulds (RF) distance among trees and the adjusted RF showed in brackets, ranging from 0 (identical topology) to 1 (completely discordant) calculated from RFadj = RF = RF/(2*n* – 6) being *n* the number of tree nodes.

		Concatenation approach										Coalescence approach									
		Phyluce_627	**Phyluce_627_5**	**Phyluce_627_2.5**	**Phyluce_627_1**	**Phyluce_304**	**HybPiper_1051**	**HybPiper_1051_5**	**HybPiper_1051_2.5**	**HybPiper_1051_1**	**HybPiper_570**	**Phyluce_627**	**Phyluce_627_5**	**Phyluce_627_2.5**	**Phyluce_627_1**	**Phyluce_304**	**HybPiper_1051**	**HybPiper_1051_5**	**HybPiper_1051_2.5**	**HybPiper_1051_1**	**HybPiper_570**
Concatenation approach	**Phyluce 627**	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Phyluce 627_5	0 (0)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Phyluce 627_2.5	2 (0.01)	2 (0.01)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Phyluce 627_1	42 (0.26)	42 (0.26)	42 (0.26)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	Phyluce 304	72 (0.44)	72 (0.44)	72 (0.44)	78 (0.48)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	HybPiper_1051	64 (0.40)	64 (0.40)	64 (0.40)	60 (0.37)	88 (0.54)	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–
	HybPiper_1051_5	64 (0.40)	64 (0.40)	64 (0.40)	60 (0.37)	88 (0.54)	0 (0)	0	–	–	–	–	–	–	–	–	–	–	–	–	–
	HybPiper_1051_2.5	64 (0.40)	64 (0.40)	64 (0.40)	60 (0.37)	88 (0.54)	0 (0)	0 (0)	0	–	–	–	–	–	–	–	–	–	–	–	–
	HybPiper_1051_1	64 (0.40)	64 (0.40)	64 (0.40)	58 (0.36)	88 (0.54)	2 (0.01)	2 (0.01)	2 (0.01)	0	–	–	–	–	–	–	–	–	–	–	–
	HybPiper_570	64 (0.40)	64 (0.40)	64 (0.40)	64 (0.40)	88 (0.54)	24 (0.15)	24 (0.15)	24 (0.15)	24 (0.15)	0	–	–	–	–	–	–	–	–	–	–
Coalescence approach	**Phyluce 627**	59 (0.36)	59 (0.36)	59 (0.36)	53 (0.33)	85 (0.52)	63 (0.39)	63 (0.39)	63 (0.39)	63 (0.39)	61 (0.38)	0	–	–	–	–	–	–	–	–	–
	Phyluce 627_5	59 (0.36)	59 (0.36)	59 (0.36)	53 (0.33)	85 (0.52)	63 (0.39)	63 (0.39)	63 (0.39)	63 (0.39)	61 (0.38)	0 (0)	0	–	–	–	–	–	–	–	–
	Phyluce 627_2.5	59 (0.36)	59 (0.36)	59 (0.36)	53 (0.33)	85 (0.52)	63 (0.39)	63 (0.39)	63 (0.39)	63 (0.39)	61 (0.38)	0 (0)	0 (0)	0	–	–	–	–	–	–	–
	Phyluce 627_1	59 (0.36)	59 (0.36)	59 (0.36)	53 (0.33)	85 (0.52)	65 (0.40)	65 (0.40)	65 (0.40)	65 (0.40)	61 (0.38)	4 (0.02)	4 (0.02)	4 (0.02)	0	–	–	–	–	–	–
	Phyluce 304	75 (0.46)	75 (0.46)	75 (0.46)	67 (0.41)	91 (0.56)	77 (0.48)	77 (0.48)	77 (0.48)	77 (0.48)	81 (0.50)	44 (0.27)	44 (0.27)	44 (0.27)	48 (0.30)	0	–	–	–	–	–
	HybPiper_1051	77 (0.48)	77 (0.48)	77 (0.48)	71 (0.44)	95 (0.59)	85 (0.52)	85 (0.52)	85 (0.52)	83 (0.51)	79 (0.49)	60 (0.37)	60 (0.37)	60 (0.37)	62 (0.38)	50 (0.31)	0	–	–	–	–
	HybPiper_1051_5	67 (0.41)	67 (0.41)	67 (0.41)	59 (0.36)	85 (0.52)	73 (0.45)	73 (0.45)	73 (0.45)	71 (0.44)	67 (0.41)	50 (0.31)	50 (0.31)	50 (0.31)	52 (0.32)	56 (0.35)	18 (0.11)	0	–	–	–
	HybPiper_1051_2.5	69 (0.43)	69 (0.43)	69 (0.43)	61 (0.38)	89 (0.55)	75 (0.46)	75 (0.46)	75 (0.46)	73 (0.45)	69 (0.43)	52 (0.32)	52 (0.32)	52 (0.32)	54 (0.33)	66 (0.41)	24 (0.15)	16 (0.10)	0	–	–
	HybPiper_1051_1	73 (0.45)	73 (0.45)	73 (0.45)	65 (0.40)	89 (0.55)	75 (0.46)	75 (0.46)	75 (0.46)	73 (0.45)	71 (0.44)	56 (0.35)	56 (0.35)	56 (0.35)	58 (0.36)	54 (0.33)	22 (0.14)	20 (0.12)	28 (0.17)	0	–
	HybPiper_570	79 (0.49)	79 (0.49)	79 (0.49)	71 (0.44)	95 (0.59)	83 (0.51)	83 (0.51)	83 (0.51)	83 (0.51)	73 (0.45)	60 (0.37)	60 (0.37)	60 (0.37)	62 (0.38)	66 (0.41)	48 (0.30)	46 (0.28)	54 (0.33)	44 (0.27)	0

Supplementary Figures

Fig. S1. Phylogenetic trees estimated from the sequences extracted with the PHYLUCE method and the concatenation approach (see main text for details). **(A)** Using the unfiltered dataset, **(B)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 5, **(C)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 2.5, **(D)** using the filtered dataset, filtering positions with substitution rates (SR) higher than 1, and **(E)** using the loci filtered dataset, containing only the best informative loci selected under the criteria explained in the main text.

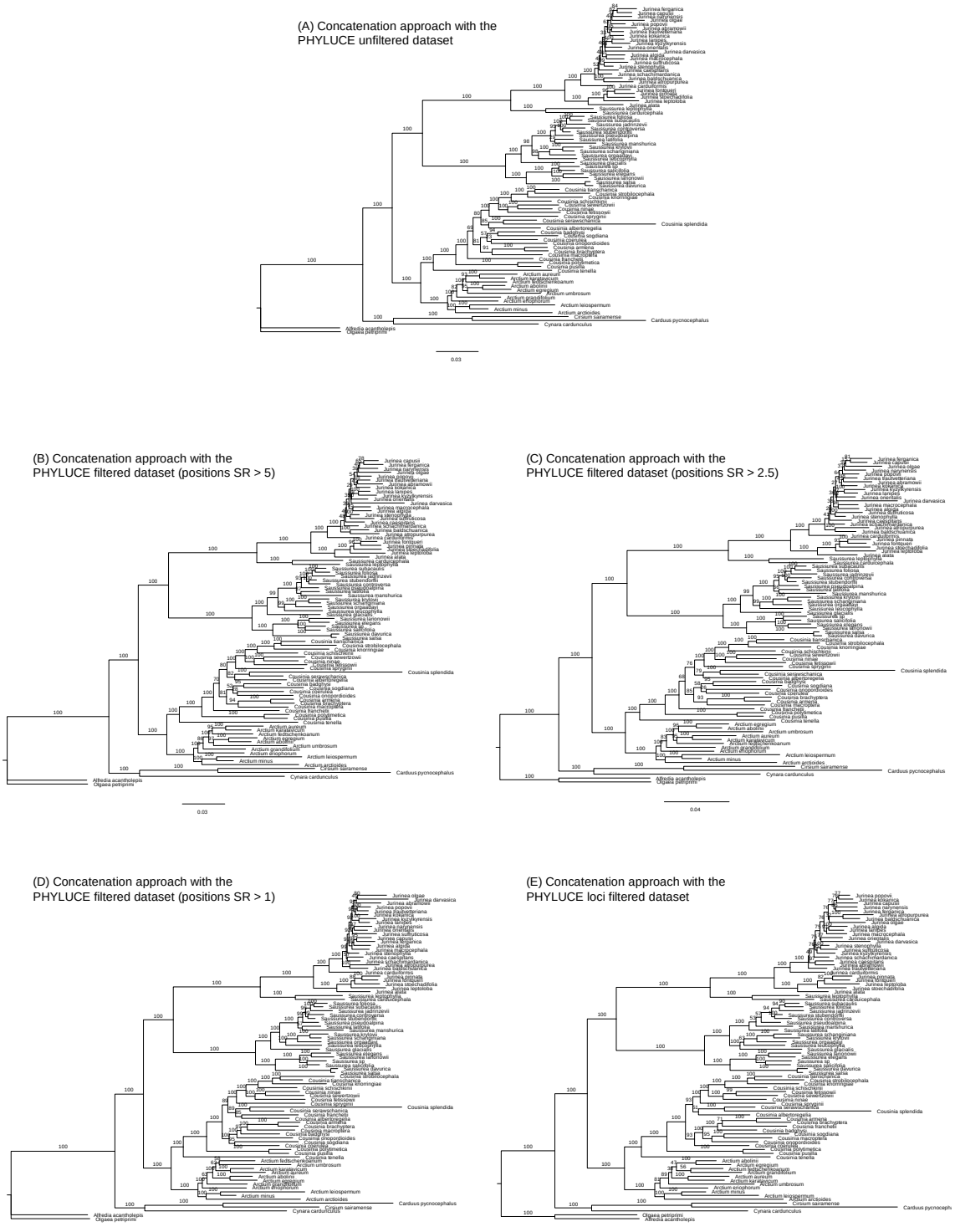


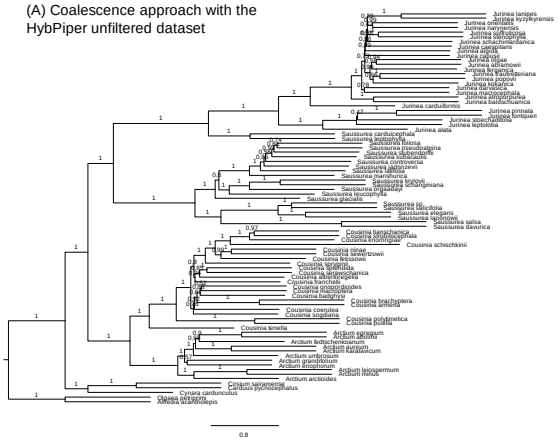
Fig. S2. Phylogenetic trees estimated from the sequences extracted with the HybPiper method and the concatenation approach (see main text for details). **(A)** sing the unfiltered dataset, **(B)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 5, **(C)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 2.5, **(D)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 1, and **(E)** using the loci filtered dataset, containing only the best informative loci selected under the criteria explained in the main text.

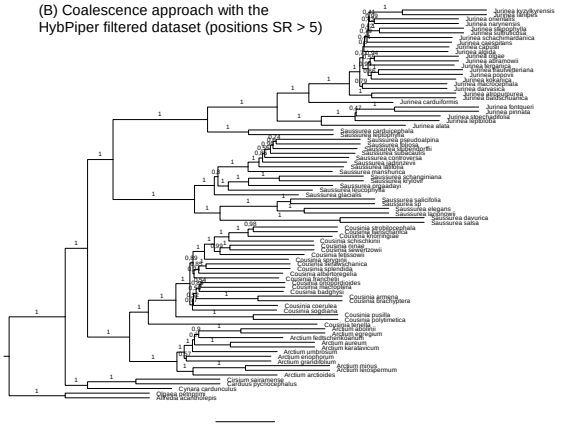


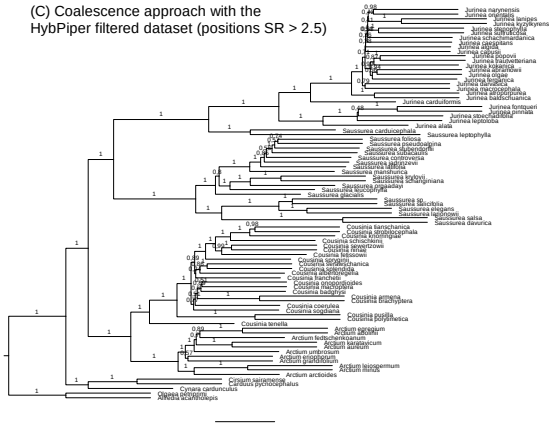
Fig. S3. Phylogenetic trees estimated from the sequences extracted with the PHYLUCE method and the coalescence approach (see main text for details). **(A)** Using the unfiltered dataset, **(B)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 5, **(C)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 2.5, **(D)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 1, and **(E)** using the loci filtered dataset, containing only the best informative loci selected under the criteria explained in the main text.

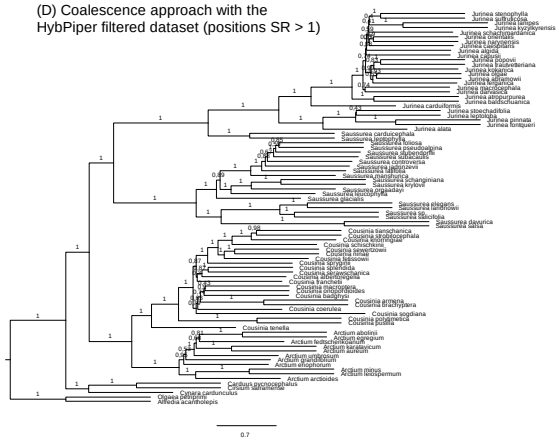


Fig. S4. Phylogenetic trees estimated from the sequences extracted with the HybPiper method and the coalescence approach (see main text for details). **(A)** Using the unfiltered dataset, **(B)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 5, **(C)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 2.5, **(D)** using the filtered dataset, filtering the positions with substitution rates (SR) higher than 1, and **(E)** using the loci filtered dataset, containing only the best informative loci selected under the criteria explained in the main text.









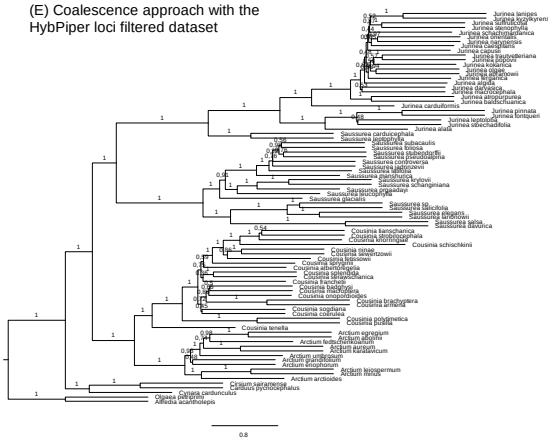


Fig. S5. Phylogenetic Informativeness analyses showing the ultrametric trees rescaled from 1 to 0 obtained from the maximum likelihood analyses obtained with the PHYLUCE dataset and the concatenation approach, and net phylogenetic informativeness curves representing the profiles for each locus displayed in different colors. The analyses of PI were done with **(A)** the unfiltered dataset, **(B)** the filtered dataset, removing the positions with substitution rates (SR) > 5, **(C)** the filtered dataset, removing the positions with SR > 2.5, and **(D)** the filtered dataset, removing the positions with SR > 1. Branches with bootstrap support values below 70 are outlined in red. For each of the four genera, the number unsupported nodes are shown at the right of the trees.

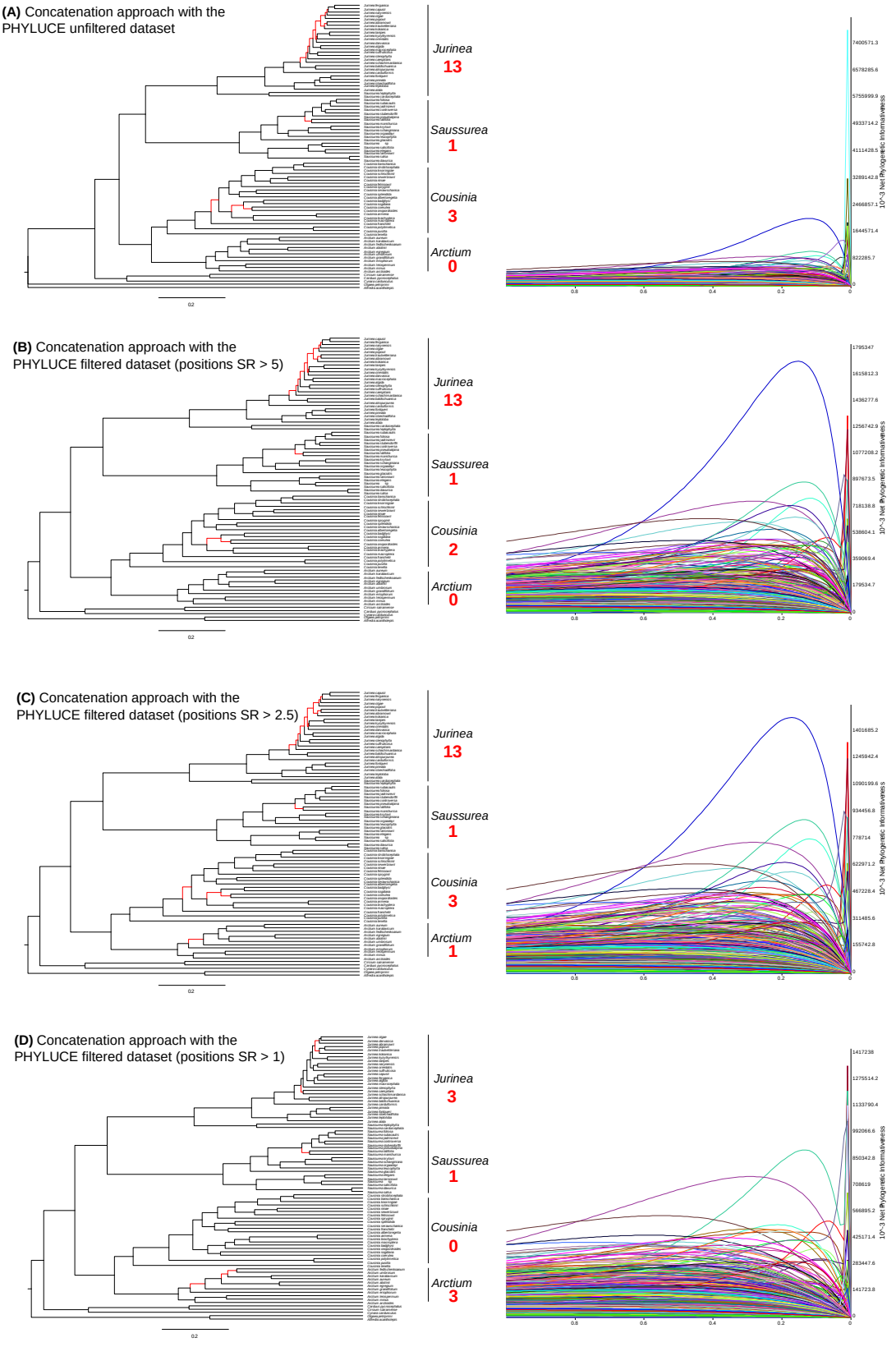
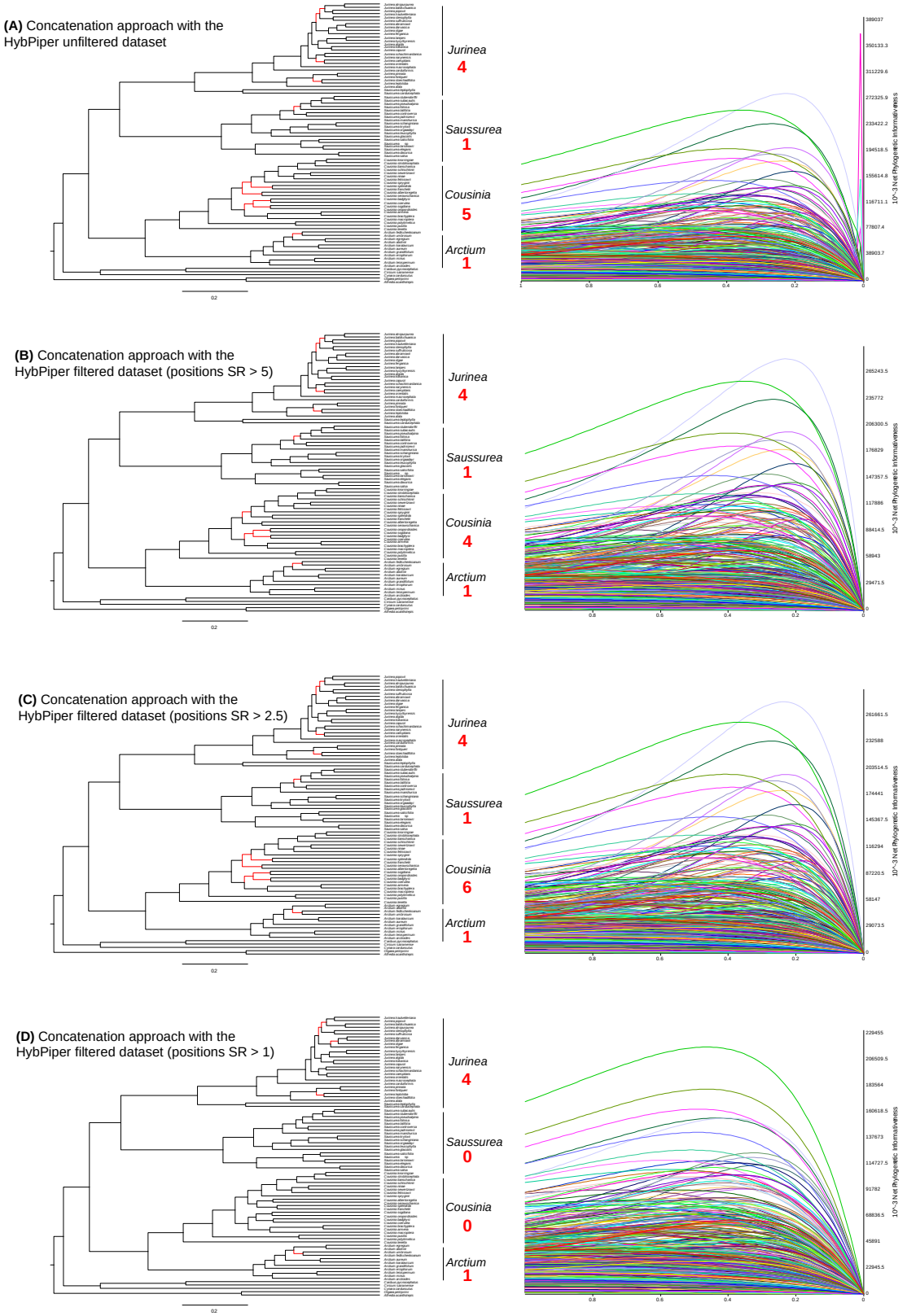
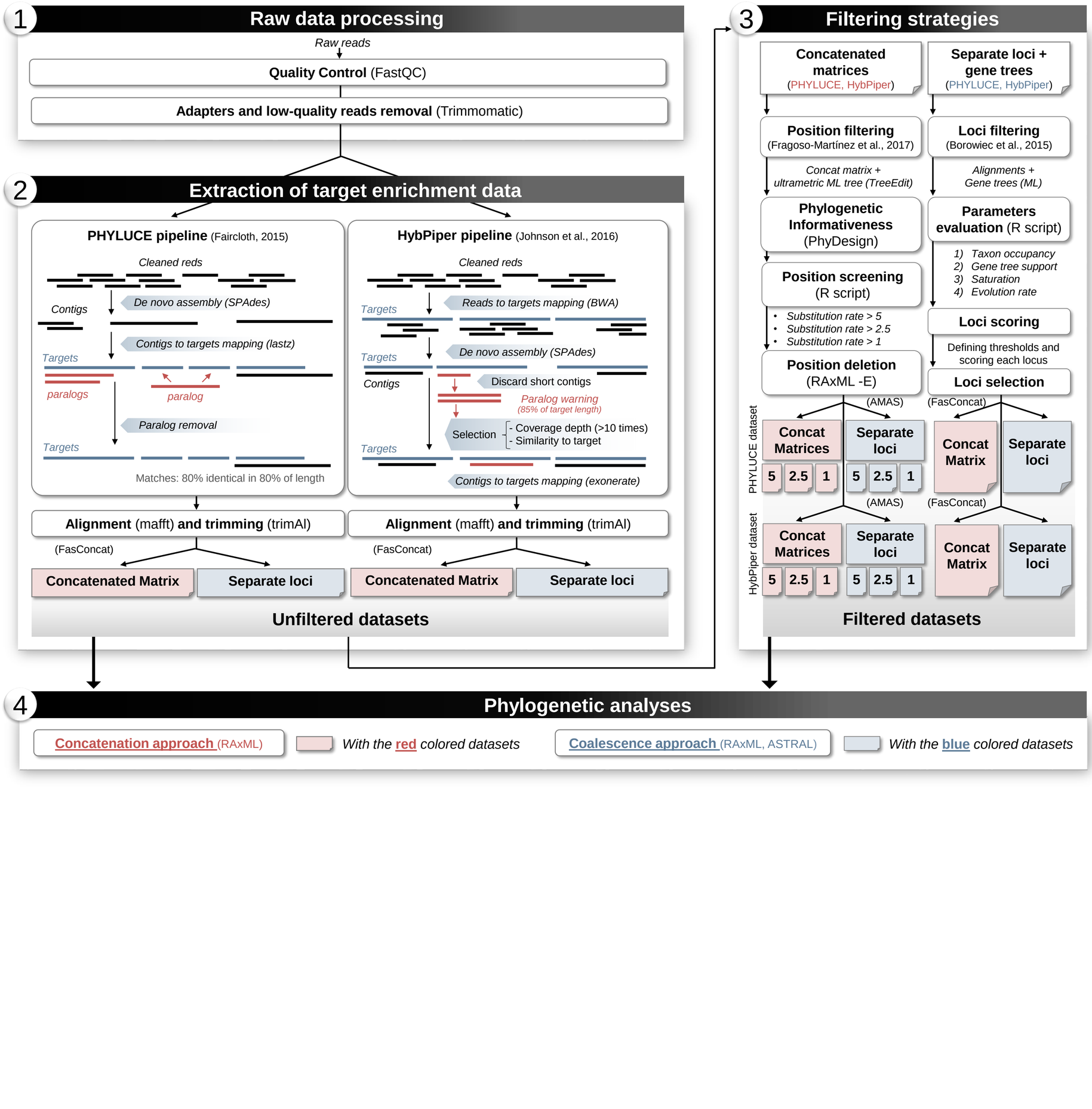
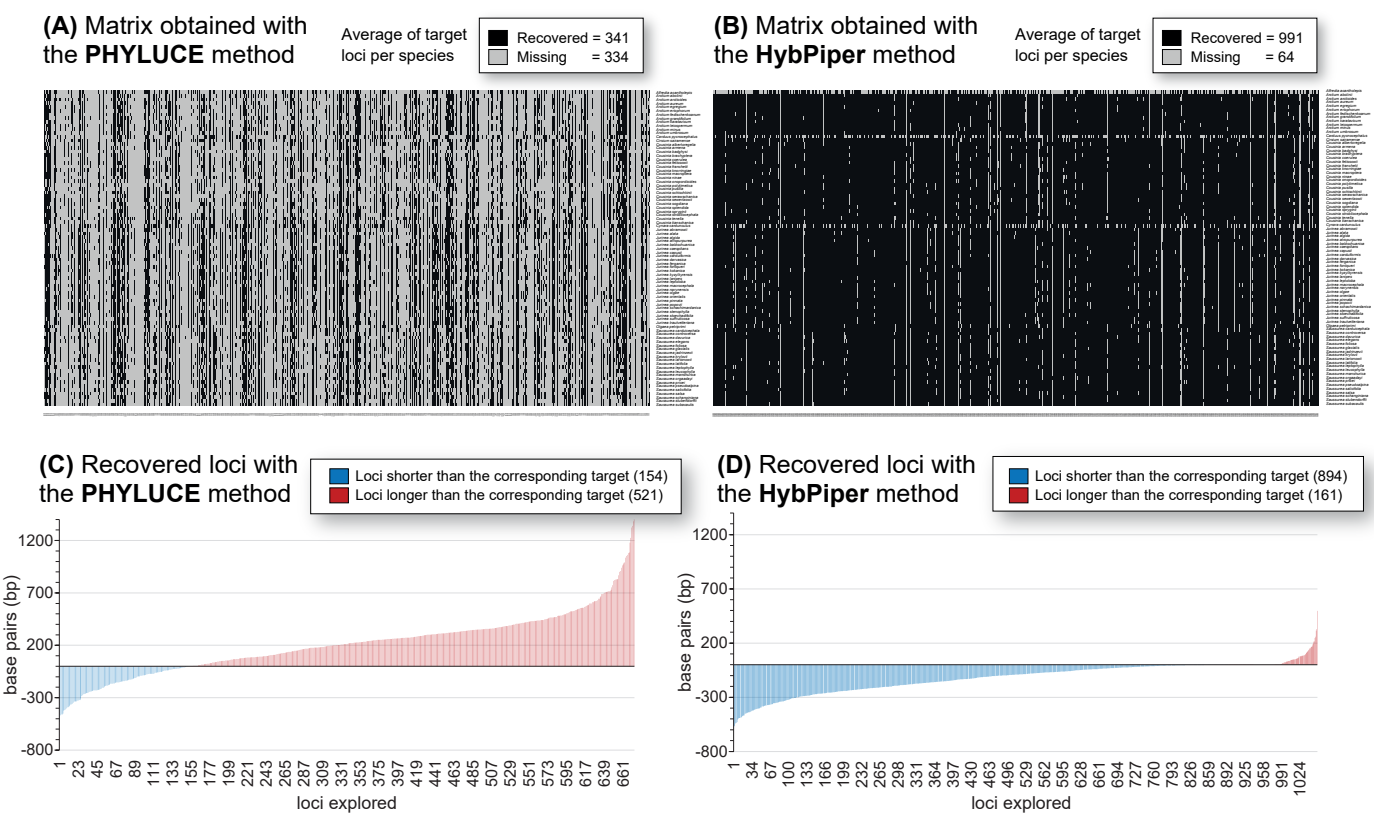
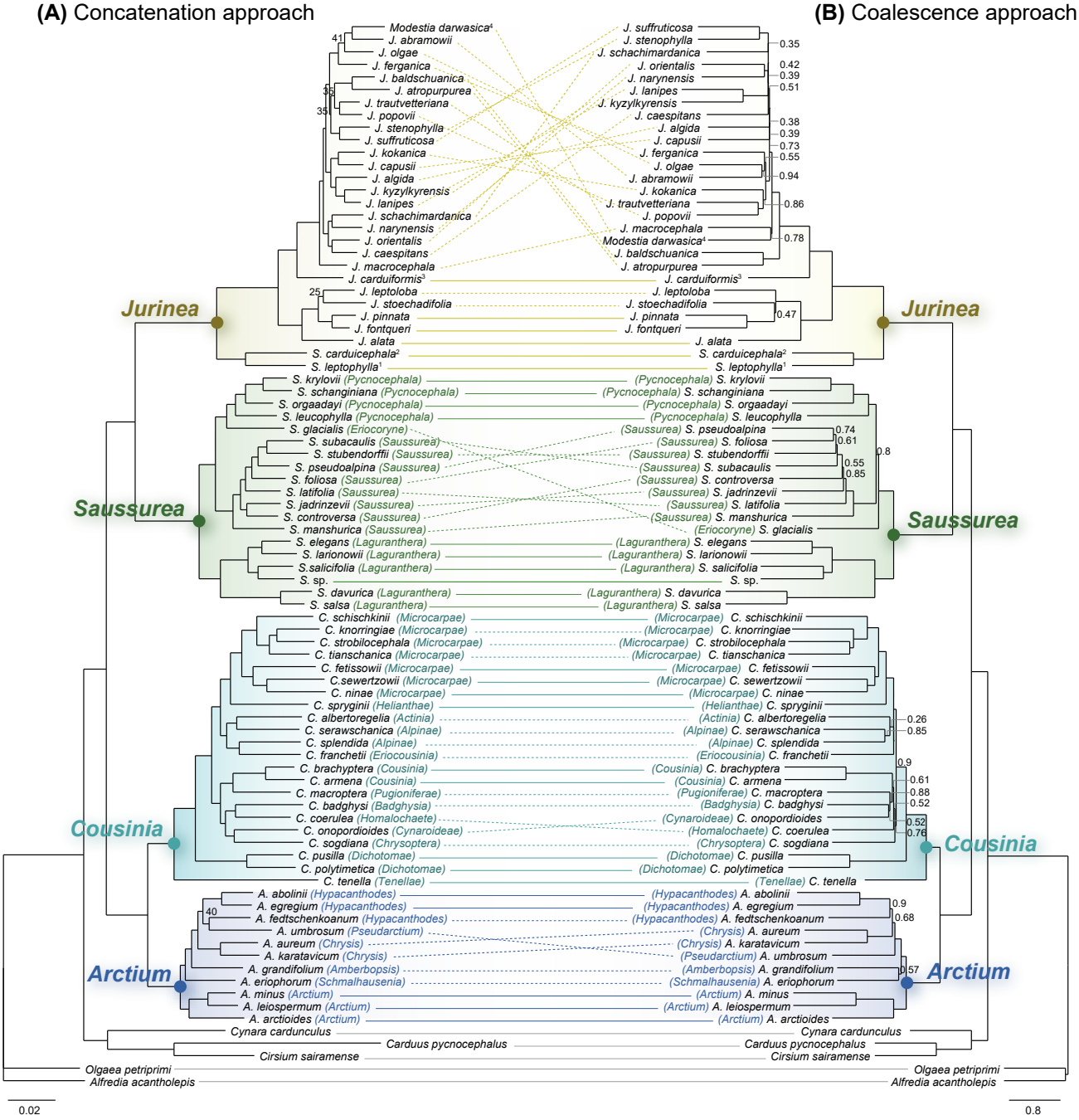


Fig. S6. Phylogenetic Informativeness analyses showing the ultrametric trees rescaled from 1 to 0 obtained from maximum likelihood analyses obtained with the HybPiper dataset and the concatenation approach, and net phylogenetic informativeness curves representing the profiles for each locus displayed in different colors. The analyses of PI were done with **(A)** the unfiltered dataset, **(B)** the filtered dataset, removing the positions with substitution rates (SR) > 5, **(C)** the filtered dataset, removing the positions with SR > 2.5, and **(D)** the filtered dataset, removing the positions with SR > 1. Branches with bootstrap support values below 70 are outlined in red. For each of the four genera, the number unsupported nodes are shown at the right of the trees.

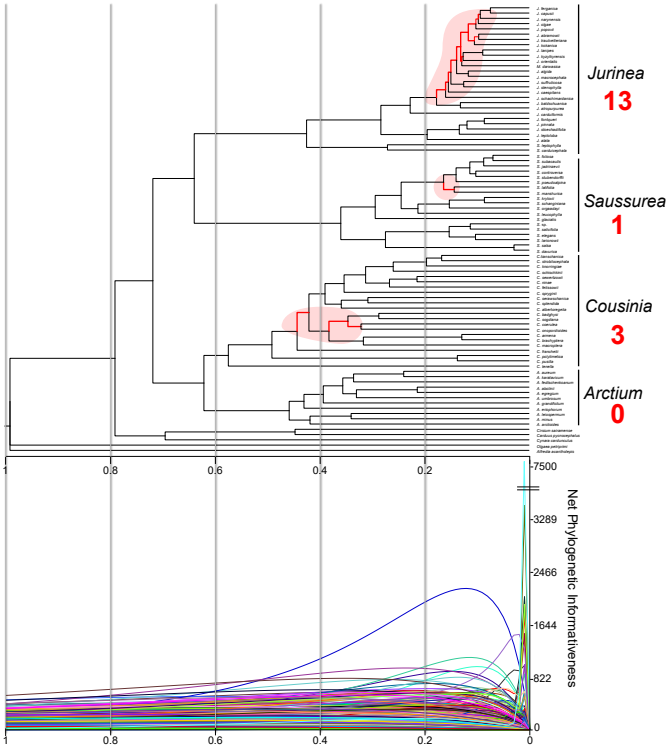



1
Raw data processing
Raw reads
Quality Control (FastQC)
Adapters and low-quality reads removal (Trimmomatic)
2
Extraction of target enrichment data
PHYLUCE pipeline (Faircloth, 2015)
Cleaned reds
Contigs
De novo assembly (SPAdes)
Contigs to targets mapping (lastz)
Targets
paralogs
paralog
Paralog removal
Targets
Matches: 80% identical in 80% of length
HybPiper pipeline (Johnson et al., 2016)
Cleaned reds
Targets
Reads to targets mapping (BWA)
Targets
De novo assembly (SPAdes)
Contigs
Discard short contigs
Paralog warning
(85% of target length)
Selection
- Coverage depth (>10 times)
- Similarity to target
Targets
Contigs to targets mapping (exonerate)
Alignment (mafft) and trimming (trimAl)
(FasConcat)
Concatenated Matrix
Separate loci
Alignment (mafft) and trimming (trimAl)
(FasConcat)
Concatenated Matrix
Separate loci
Unfiltered datasets
3
Filtering strategies
Concatenated matrices
(PHYLUCE, HybPiper)
Position filtering
(Fragoso-Martínez et al., 2017)
Concat matrix +
ultrametric ML tree (TreeEdit)
Phylogenetic Informativeness
(PhyDesign)
Position screening
(R script)
Substitution rate > 5
Substitution rate > 2.5
Substitution rate > 1
Position deletion
(RAxML -E)
(AMAS)
Concat Matrices
Separate loci
5
2.5
1
5
2.5
1
(AMAS)
Concat Matrices
Separate loci
5
2.5
1
5
2.5
1
PHYLUCE dataset
HybPiper dataset
Separate loci + gene trees
(PHYLUCE, HybPiper)
Loci filtering
(Borowiec et al., 2015)
Alignments +
Gene trees (ML)
Parameters evaluation (R script)
1) Taxon occupancy
2) Gene tree support
3) Saturation
4) Evolution rate
Loci scoring
Defining thresholds and scoring each locus
Loci selection
(FasConcat)
Concat Matrix
Separate loci
(FasConcat)
Concat Matrix
Separate loci
Filtered datasets
4
Phylogenetic analyses
Concatenation approach (RAxML)
With the red colored datasets
Coalescence approach (RAxML, ASTRAL)
With the blue colored datasets

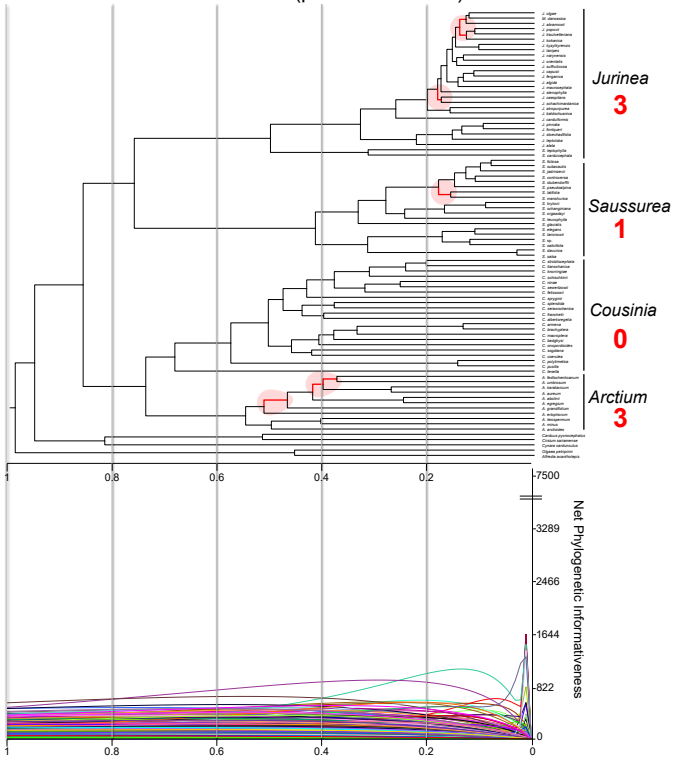

(A) Matrix obtained with the PHYLUCE method
Average of target loci per species
Recovered = 341
Missing = 334
(B) Matrix obtained with the HybPiper method
Average of target loci per species
Recovered = 991
Missing = 64
(C) Recovered loci with the PHYLUCE method
Loci shorter than the corresponding target (154)
Loci longer than the corresponding target (521)
base pairs (bp)
1200
700
200
-300
-800
1
23
45
67
89
111
133
155
177
199
221
243
265
287
309
331
353
375
397
419
441
463
485
507
529
551
573
595
617
639
661
loci explored
(D) Recovered loci with the HybPiper method
Loci shorter than the corresponding target (894)
Loci longer than the corresponding target (161)
base pairs (bp)
1200
700
200
-300
-800
1
34
67
100
133
166
199
232
265
298
331
364
397
430
463
496
529
562
595
628
661
694
727
760
793
826
859
892
925
958
991
1024
loci explored


(A) Concatenation approach
(B) Coalescence approach
Jurinea
Saussurea
Cousinia
Arctium
Modestia darwasica⁴
J. abramowii
J. olgae
J. ferganica
J. baldschuanica
J. atropurpurea
J. trautvetteriana
J. popovii
J. stenophylla
J. suffruticosa
J. kokanica
J. capusii
J. algida
J. kyzylkyrensis
J. lanipes
J. schachimardanica
J. narynensis
J. orientalis
J. caespitans
J. macrocephala
J. carduiformis³
J. leptoloba
J. stoechadifolia
J. pinnata
J. fontqueri
J. alata
S. carduicephala²
S. leptophylla¹
S. krylovii (Pycnocephala)
S. schanginiana (Pycnocephala)
S. orgaadayi (Pycnocephala)
S. leucophylla (Pycnocephala)
S. glacialis (Eriocoryne)
S. subacaulis (Saussurea)
S. stubendorffii (Saussurea)
S. pseudoalpina (Saussurea)
S. foliosa (Saussurea)
S. latifolia (Saussurea)
S. jadrinzevii (Saussurea)
S. controversa (Saussurea)
S. manshurica (Saussurea)
S. elegans (Lagurranthera)
S. larionowii (Lagurranthera)
S.salicifolia (Lagurranthera)
S. sp.
S. davurica (Lagurranthera)
S. salsa (Lagurranthera)
C. schischkinii (Microcarpae)
C. knorringiae (Microcarpae)
C. strobilocephala (Microcarpae)
C. tianschanica (Microcarpae)
C. fetissowii (Microcarpae)
C.sewertzowii (Microcarpae)
C. ninae (Microcarpae)
C. spryginii (Helianthae)
C. albertoregelia (Actinia)
C. serawschanica (Alpinae)
C. splendida (Alpinae)
C. franchetii (Eriocousinia)
C. brachyptera (Cousinia)
C. armena (Cousinia)
C. macroptera (Pugioniferae)
C. badghysi (Badghysia)
C. coerulea (Homalochaete)
C. onopordioides (Cynaroideae)
C. sogdiana (Chrysoptera)
C. pusilla (Dichotomae)
C. polytimetica (Dichotomae)
C. tenella (Tenellae)
A. abolinii (Hypacanthodes)
A. egregium (Hypacanthodes)
A. fedtschenkoanum (Hypacanthodes)
A. umbrosum (Pseudarctium)
A. aureum (Chrysis)
A. karatavicum (Chrysis)
A. grandifolium (Amberbopsis)
A. eriophorum (Schmalhausenia)
A. minus (Arctium)
A. leiospermum (Arctium)
A. arctioides (Arctium)
Cynara cardunculus
Carduus pycnocephalus
Cirsium sairamense
Olgaea petriprimi
Alfredia acantholepis
0.02
0.8

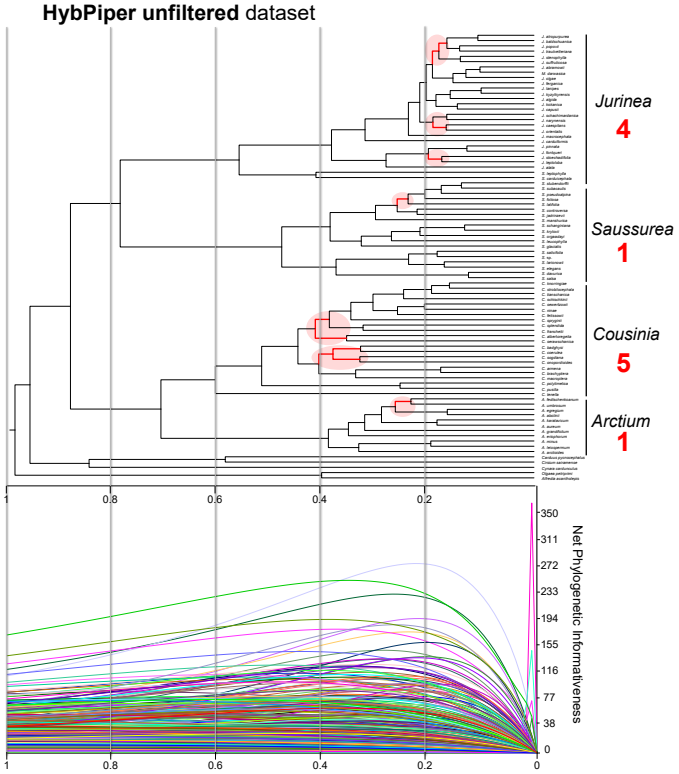
(A) Concatenation approach with the **PHYLUCE unfiltered** dataset



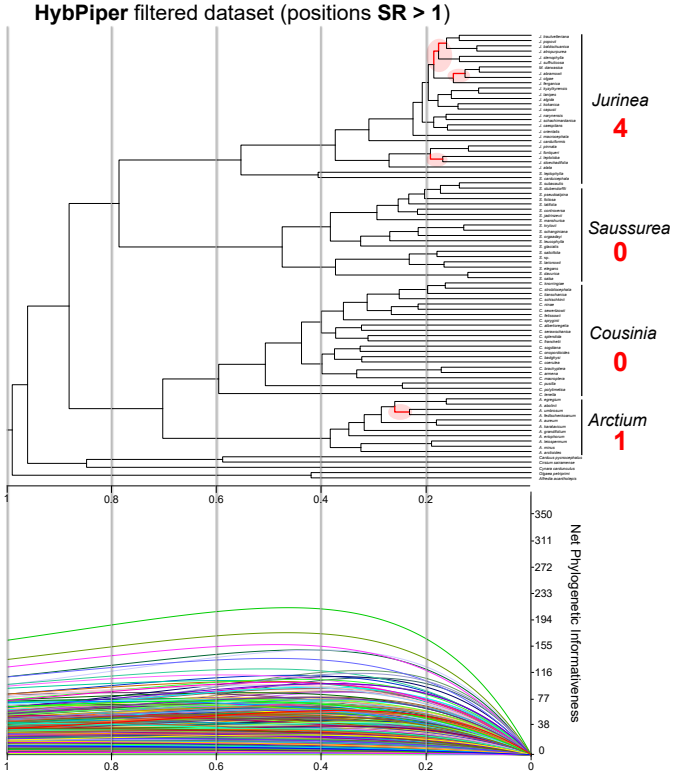
(B) Concatenation approach with the **PHYLUCE** filtered dataset (positions **SR > 1**)

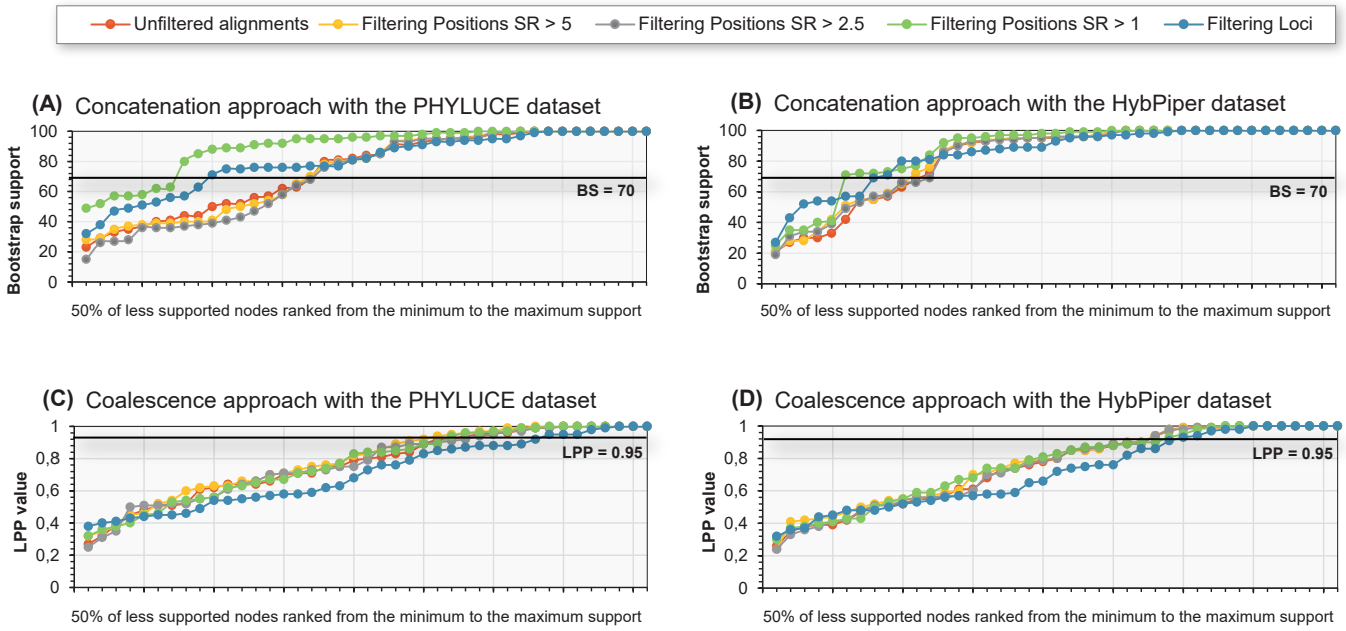


(C) Concatenation approach with the **HybPiper unfiltered** dataset



(D) Concatenation approach with the **HybPiper** filtered dataset (positions **SR > 1**)




Unfiltered alignments
Filtering Positions SR > 5
Filtering Positions SR > 2.5
Filtering Positions SR > 1
Filtering Loci
(A) Concatenation approach with the PHYLUCE dataset
Bootstrap support
BS = 70
50% of less supported nodes ranked from the minimum to the maximum support
(B) Concatenation approach with the HybPiper dataset
Bootstrap support
BS = 70
50% of less supported nodes ranked from the minimum to the maximum support
(C) Coalescence approach with the PHYLUCE dataset
LPP value
LPP = 0.95
50% of less supported nodes ranked from the minimum to the maximum support
(D) Coalescence approach with the HybPiper dataset
LPP value
LPP = 0.95
50% of less supported nodes ranked from the minimum to the maximum support