Molecular Phylogenetics and Evolution 167 (2022) 107342

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

How challenging RADseq data turned out to favor coalescent-based species tree inference. A case study in Aichryson (Crassulaceae)

Philipp Hühn [a,b], Markus S. Dillenberger [b,c], Michael Gerschwitz-Eidt [b], Elvira Hörandl [d], Jessica A. Los [a,e], Thibaud F.E. Messerschmid [a,b,e], Claudia Paetzold [d,f], Benjamin Rieger [b], Gudrun Kadereit [e,*]

[a] Institute of Molecular Physiology (iMP), Johannes Gutenberg-University Mainz, Germany
[b] Institute of Organismic and Molecular Evolution (iOME), Johannes Gutenberg-University Mainz, Germany
[c] Institut für Biologie, Systematische Botanik und Pflanzengeographie, Freie Universität Berlin, Germany
[d] Department of Systematics, Biodiversity and Evolution of Plants, Georg-August-University Göttingen, Germany
[e] Prinzessin Therese von Bayern Lehrstuhl für Systematik, Biodiversität & Evolution der Pflanzen, Ludwig-Maximilians-Universität München, Germany
[f] Department of Botany and Molecular Evolution, Senckenberg Research Institute and Natural History Museum Frankfurt am Main, Germany

ARTICLE INFO

Keywords:
[clustering threshold selection
coalescent-based summary method
data bias
locus filtering
RADseq
species tree inference]

ABSTRACT

Analysing multiple genomic regions while incorporating detection and qualification of discordance among regions has become standard for understanding phylogenetic relationships. In plants, which usually have comparatively large genomes, this is feasible by the combination of reduced-representation library (RRL) methods and high-throughput sequencing enabling the cost effective acquisition of genomic data for thousands of loci from hundreds of samples. One popular RRL method is RADseq. A major disadvantage of established RADseq approaches is the rather short fragment and sequencing range, leading to loci of little individual phylogenetic information. This issue hampers the application of coalescent-based species tree inference. The modified RADseq protocol presented here targets ca. 5,000 loci of 300-600nt length, sequenced with the latest short-read-sequencing (SRS) technology, has the potential to overcome this drawback. To illustrate the advantages of this approach we use the study group *Aichryson* Webb & Berthelott (Crassulaceae), a plant genus that diversified on the Canary Islands. The data analysis approach used here aims at a careful quality control of the long loci dataset. It involves an informed selection of thresholds for accurate clustering, a thorough exploration of locus properties, such as locus length, coverage and variability, to identify potential biased data and a comparative phylogenetic inference of filtered datasets, accompanied by an evaluation of resulting BS support, gene and site concordance factor values, to improve overall resolution of the resulting phylogenetic trees. The final dataset contains variable loci with an average length of 373nt and facilitates species tree estimation using a coalescent-based summary approach. Additional improvements brought by the approach are critically discussed.

1. Introduction

Resolving phylogenetic relationships of recently and rapidly radiating species complexes is a challenge because first, standard markers using universal primers are too conserved and fail to provide sufficient information, and second, inferring relationships is often complicated due to incomplete lineage sorting (ILS), hybridization/introgression and gene duplication/loss events (Pamilo and Nei, 1988; Maddison, 1997;

Abbreviations: BSC, between-sample-clustering; CA-ML, maximum likelihood analysis of concatenated loci; CB-SM, coalescent-based summary method; CT, clustering threshold; gCF, gene concordance factor; GTEE, gene tree estimation error; HTS, high throughput sequencing; ILS, incomplete lineage sorting; ISC, in-sample-clustering; ML, maximum likelihood; MSC, multi-species coalescent (model); NPL, new polymorphic loci; PE, paired-end; PIC, parsimony informative character; PIS, parsimony informative site; RADseq, restriction site-associated DNA sequencing; REase, restriction endonuclease; RRL, reduced-representation library (methods); sCF, site concordance factor; SNP, single nucleotide polymorphism; SRS, short-read sequencing; SVD, SVDquartets; VAR, variable sites (sequence variation); var, variability (VAR/locus length/number of samples.

* Corresponding author at: Prinzessin Therese von Bayern Lehrstuhl für Systematik, Biodiversität & Evolution der Pflanzen, Ludwig-Maximilians-Universität München, Germany.
E-mail address: G.Kadereit@lmu.de (G. Kadereit).

https://doi.org/10.1016/j.ympev.2021.107342
Received 3 July 2020; Received in revised form 5 July 2021; Accepted 29 October 2021
Available online 14 November 2021

Maddison and Knowles, 2006; Kubatko and Degnan, 2007; Whitfield and Lockhart, 2007; Degnan et al., 2006, Degnan and Rosenberg, 2009; Heled and Drummond, 2009; Yang and Rannala, 2010; Rannala et al., 2020). Since different parts of the genome can have different evolutionary backgrounds, approaches analyzing multiple genomic regions have become the baseline for resolving such challenging lineages. The multi-species coalescent (MSC) model provides a natural framework for species tree inference that accounts for gene tree discordance due to ILS. However, full-coalescence approaches under the MSC are computationally very intensive when applied on large-scale genomic data and thus often not feasible (McCormack et al., 2013a; Smith et al., 2014; Zimmermann et al., 2014). Other approaches, such as maximum likelihood analysis of concatenated multi-locus data (de Queiroz et al., 1995; Yang 1996; de Queiroz and Gatesy 2007), coalescent-based summary methods that estimate species trees from independently inferred gene trees (here called "locus trees") (Mirarab et al., 2014a; Mirarab and Warnow, 2015; Rannala et al., 2020) or coalescent-based methods that use site patterns of assembled loci for species tree inference (Bryant et al., 2012; Chifman and Kubatko, 2014; Bryant and Hahn, 2020), became increasingly popular and widely used. Despite their popularity, these methods each have advantages and disadvantages and their correct application to modern high-throughput data, in particular approaches that generate short loci with high amounts of missing data such as RADseq, is highly controversial.

High-throughput sequencing (HTS) technologies and lab workflows for sample preparation improved enormously during the last decade and provide the opportunity to generate extensive datasets for phylogenetic inference (reviewed in Good, 2012; Reuter et al., 2015; Andrews et al., 2016; Mardis, 2017; McKain et al., 2018). Some of the most popular sample preparation protocols are grouped under the term reduced-representation library (RRL) preparation protocols, which are often combined with short-read sequencing (SRS). These methods target only a reduced subset of the studied genome for sequencing, therefore reducing computational complexity during assembly and analysis, facilitating a deeper sequencing depth per locus while increasing the number of samples included. The combination of both HTS and RRL enable simultaneous acquisition of genomic data of hundreds up to thousands of loci from dozens to hundreds of samples for systematic researchers and extend the questions and taxa that can be investigated tremendously. Widely used RRL approaches are hybridization capturing methods, e.g., on-array capture or in-solution capture (Mamanova et al., 2010), Hyb-Seq (Weitemier et al., 2014), targeted sequence capture (Grover et al., 2012) and restriction-site associated DNA sequencing (RADseq; Miller et al., 2007; Baird et al., 2008). The term RADseq comprises several methods that all rely on the enzymatic digestion of genomic DNA for complexity reduction, followed by adapter ligation, further reduction by size selection (either direct or indirect) and high-throughput sequencing (reviewed in Andrews et al., 2016). The cross-over approach hyRAD by Suchan et al. (2016) combines RADseq with capturing using either biotinylated DNA- or RNA-probes (Schmid et al., 2017; Suchan, 2018) obtained from the enzymatically fragmented DNA resources of the target group itself. Yet, the lab workflow is quite complex and time consuming. Thanks to the modular principle of RADseq, the individual wet lab steps, restriction endonucleases (REase/s) and adapters can be modified as required (see also McCormack et al. 2013b; Andrews et al., 2016; McKain et al., 2018; Parchman et al., 2018). This flexible toolbox of cheap, fast and individually scalable wet lab modules, as well as the fact that no prior genomic information is required, paved the way for the success of RADseq methods in various fields of evolutionary research, particularly in non-model organisms (e.g., Eaton and Ree, 2013; Escudero et al., 2014; Harvey et al., 2016; Herrera and Shank, 2016; Razkin et al., 2016; de Oca et al., 2017; Dillenberger and Kadereit, 2017; Hamon et al., 2017; Curto et al., 2018; Wagner et al., 2018; Gerschwitz-Eidt and Kadereit, 2019; Paetzold et al., 2019; Rancilhac et al., 2019; Hipp et al., 2020; Karbstein et al., 2020; Wagner et al., 2020; Buono et al., 2021).

Despite these obvious benefits of RADseq, the approach poses some inherent challenges regarding the wet lab workflow, sequence assembly, data set processing and the application of coalescent-based species tree inference. Characteristically, RADseq datasets comprise relatively short loci (typically 100–250nt) and a high proportion of missing data (Ree and Hipp, 2015; Andrews et al., 2016; Eaton et al., 2017; Lee et al., 2018; McKain et al., 2018). The average fragment length obtained (and locus length assembled) depends on the degree of genomic reduction, which in turn depends on the REase/s chosen, the selected size segregation window and the genome size of the study group. To some extent, missing data (absence of data or missingness) in RADseq data is inherently expected due to mutations of the REase-specific recognition sites (Rubin et al., 2012; Eaton et al., 2017; Lee et al., 2018). Technical causes for missingness include: varying DNA quantity and quality, size selection artifacts, PCR bias or low sequencing depth and quality. All of these factors influence the average information content per locus and the uniformity with which it is distributed across taxa, consequently limiting the applicability of inference methods (Gatesy and Springer, 2014; Xi et al., 2015; Xu and Yang, 2016; Eaton et al., 2017; Sayyari et al., 2017; Lee et al., 2018; Molloy and Warnow, 2018).

RADseq is particularly appealing for studying non-model taxa, as large genome-sized datasets can be generated quickly and cost-effectively and assembled without requiring a reference genome. However, *de novo* assembly and data processing can also be a major challenge. The bioinformatics effort related to RADseq data is often not straightforward and can heavily impact the assembly outcome regarding differentiation of orthologs and paralogs, as well as the quantity of recovered loci, sequence variation (VAR), single nucleotide polymorphisms (SNPs) and parsimony informative sites (PIS), respectively (Rubin et al., 2012; Ilut et al., 2014; Harvey et al., 2015; Shafer et al., 2017; Lee et al., 2018). To facilitate data processing, assembly pipelines such as Stacks (Catchen et al., 2013), dDocent (Puritz et al., 2014) and *ipyrad* (Eaton and Overcast, 2020) have been developed. These pipelines implement several main steps. 1) In-sample-clustering (ISC), in which reads within each sample are grouped by sequence similarity into putative loci. 2) Consensus calling of allele sequences from clustered reads. 3) Between-sample-clustering (BSC) of consensus sequences of all loci across all samples are clustered by sequence similarity to generate putatively homologous loci. 4) Data filtering based on given thresholds such as the number of samples per locus required (locus coverage) or the maximum proportion of shared heterozygous sites in a locus (detection of potential paralogs). To determine which reads represent the same genomic locus, a clustering threshold (CT) based on sequence similarity is used. Yet, genetic variation within the target genomes and across the studied taxa makes it difficult to find an appropriate CT (Rubin et al., 2012; Catchen et al., 2013; Hirsch and Buell, 2013; Ilut et al., 2014; Harvey et al., 2015; Ilut et al., 2014; Paris et al., 2017; Shafer et al., 2017; Lee et al., 2018; McCartney-Melstad et al., 2019). Both over- and undermerging are major issues in RADseq datasets, affecting ISC and BSC and therefore the resulting datasets. To ensure the homology of the assembled loci (Springer and Gatesy, 2018; McCartney-Melstad et al., 2019; Fernández et al., 2020; Simion et al., 2020), detailed evaluations of dataset metrics are used to find balanced dataset-specific CTs for ISC and BSC (e.g. Ilut et al., 2014; Mastretta-Yanes et al., 2015; McKinney et al., 2017; Paris et al., 2017; McCartney-Melstad et al., 2019). Approaches to facilitate this problem aim at the determination of suitable CTs for homology assessment by analyzing trends of several assembly metrics over a wide range of tested CTs (hereafter referred to as "CT selection approach"). This is accomplished by plotting the metrics as a function of the CT range and searching for a region that avoids over- and undermerging areas and that provides an accurate clustering for the majority of loci (hereafter referred to as "transition zone"). This transition zone is assumed to minimize the assembly of paralogs, to maximize the yield of sequence variation, and to form the smallest distance among taxa (Ilut et al., 2014; Mastretta-Yanes et al., 2015; McCartney-Melstad et al., 2019). In other words: an informed selection of dataset-

specific CTs yields maximum phylogenetic information with minimum missingness and least paralogs. Still, such CT selection approaches have to be taken with care because 1) the determined CT (for ISC and BSC) can never represent all taxa equally well and 2) all other chosen assembly parameters affect the outcome (Shafer et al., 2017; McCartney-Melstad et al., 2019).

Phylogenetic inference of assembled RADseq data presents the next challenge because the data properties often limit the choice of methods. Added to this is an ongoing, intense debate on the utilization of phylogenetic inference methods. The focus is mainly on: 1) the statistical consistency under the MSC, 2) the evolutionary framework to which the methods are applied (e.g. hybridization, horizontal gene transfer, ILS), and 3) the estimation accuracy under varying dataset conditions (e.g. linkage, phylogenetic information content, missingness, homology of data), leading to constant re-analyses and comparisons of simulated and empirical data to proof the diverging concepts (e.g. de Queiroz and Gatesy 2007; Edwards et al., 2007, 2016; Kubatko and Degnan, 2007; Degnan and Rosenberg, 2009; Leaché and Rannala, 2011; Song et al., 2012; Bayzid and Warnow, 2013; Wu et al., 2013; Gatesy and Springer, 2013, 2014; Springer and Gatesy 2014, 2016, 2018; Mirarab et al., 2014a, b, 2016; Chou et al., 2015; Roch and Steel 2015; Mendes and Hahn, 2018; Molloy and Warnow, 2018; Bryant and Hahn, 2020; Rannala et al., 2020). This somewhat amusing and abstruse debate, with sometimes remarkably tailored data for proof, complicates the search for appropriate phylogenetic inference methods for RRL-SRS data. Fact is that the locus properties are pivotal for selecting appropriate species tree inference methods. Due to the short fragment length, RADseq loci are generally assumed to lack sufficient phylogenetic information to generate locus trees as input for coalescent-based summary methods (Rubin et al., 2012; Gatesy and Springer, 2014; Xi et al., 2015; Hosner et al., 2016; Molloy and Warnow, 2018).

Gene-tree-based coalescent methods (summary methods; hereafter referred to as CB-SM) are a favorable choice for phylogenetic inference of rather long and informative loci (Mirarab et al., 2014a, 2016; Vachaspati and Warnow, 2015; Xu and Yang, 2016; Molloy and Warnow 2018; Rannala et al., 2020). CB-SM infer species trees by a two-step system: individual gene trees are estimated, and their summary statistics are then used as data input for species tree estimation. While CB-SM are becoming popular for their ability to handle large amounts of data in a short time, they are best known for their sensitivity to gene tree estimation error (GTEE). When applied to datasets composed of short loci of little individual phylogenetic information and a high proportion of missingness, as is characteristic of RADseq datasets, the effect on estimation accuracy can get quite severe (Chou et al., 2015; Roch and Warnow, 2015; Xi et al., 2015; Xu and Yang, 2016; Sayyari et al., 2017; Molloy and Warnow, 2018). Therefore, the focus on the effects of filtering loci for specific properties prior to gene and species tree estimation is becoming increasingly relevant (e.g. Lanier et al., 2014; Chen et al., 2015; Xi et al., 2015; Hosner et al., 2016; Huang and Knowles 2016; Simmons et al., 2016; Sayyari et al., 2017; Molloy and Warnow 2018).

Coalescent-based site-based methods are another option for species tree inference (Bryant et al., 2012; Chifman and Kubatko, 2014; Xu and Yang, 2016). Such approaches bypass the generation of locus trees by generating the species tree directly from all given site patterns, thus avoid the issue of GTEE. The sites are required to have individual histories or at least very little linkage. Violation of this assumption leads to a statistically inconsistent species tree estimate (Bryant et al., 2012; Chifman and Kubatko 2014; Xu and Yang, 2016). Under certain challenging data conditions, site-based methods were found to be more accurate than gene tree-based summary (Chou et al., 2015; Long and Kubatko, 2018; Molloy and Warnow, 2018).

RADseq data are most commonly analyzed using maximum likelihood analysis of a concatenated supermatrix (hereafter referred to as CA-ML) (Yang, 1996; de Queiroz and Gatesy, 2007; Rubin et al., 2012). In case of CA-ML, several thousand loci are treated as one locus that evolved under a single evolutionary history. This is violating the MSC and may theoretically lead to poorly resolved, incomplete, or positively misleading species tree estimates (Degnan et al., 2006, Degnan and Rosenberg, 2009; Kubatko and Degnan, 2007; Knowles, 2009; Roch and Steel, 2015, Xu and Yang, 2016; Mendes and Hahn, 2018; Rannala et al., 2020). In addition, bootstrapping is also commonly performed across the entire supermatrix, potentially resulting in spuriously high support values caused by the sheer dataset size (Kubatko and Degnan, 2007; Kumar et al., 2012; Rubin et al., 2012; Liu et al., 2015; Wang et al., 2017, Minh et al. 2020a). Still, it also has been shown that CA-ML can be comparably or more accurate than coalescent-based methods under various conditions of linkage, locus length, information content, missingness, ILS and GTEE (Mirarab et al., 2014a; Chou et al., 2015; Roch and Warnow, 2015; Mirarab et al., 2016; Springer and Gatesy, 2016; Long and Kubatko, 2018; Molloy and Warnow, 2018).

Despite the ongoing debate about the pros and cons of approaches to sequence generation, data assembly, phylogenetic inference, and, the assumption that RAD data do not favor coalescent-based summary methods, we think there is a need to take advantage of the significant methodical progress made in the last decade and explore their potential for practical use. Our objective is to test whether longer RADseq loci enable coalescent-based species tree inference, and to provide advice on how to handle and analyze challenging data.

We modified several modules of the RADseq toolbox to obtain a library containing a small number of fragments (ca 5,000 assembled loci), with lengths of ca. 300-600nt, sequenced with the latest SRS technology (Illumina MiSeq v3 kit, 300nt PE) and applied this protocol (Fig. 1) to the plant genus *Aichryson* Webb & Berthel. (Crassulaceae), a rapidly radiated yet relatively small genus distributed in Macaronesia, for which standard sanger sequenced markers failed to provide a resolved phylogeny (Fairfield et al., 2004). The data analysis (Fig. 2) included a CT selection approach to facilitate an informed choice of suitable CTs for ISC and BSC during *de novo* assembly (Fig. 3) and an exploratory approach to determine the properties of the assembled loci, with respect to locus coverage (missingness), locus variability (phylogenetic information) and locus length, and thus their suitability as input for CB-SM (Fig. 4). We compared the phylogenetic outcome of this assembly using CA-ML (RAxML by Stamatakis, 2014), CB-SM (ASTRAL III by Zhang et al., 2018) and put it in perspective to the site-based approach SVDquartets by Chifman and Kubatko (2014). To assess the phylogenetic results, we also evaluated the resulting BS support values relative to gene and site concordance factors that were calculated using IQ-TREE (Minh et al. 2020a, b).

2. Materials and methods

2.1. Study group, sampling and DNA extraction

Together with *Monanthes* Haw. and *Aeonium* Webb & Berthel., *Aichryson* belongs to the Macaronesian tribe Aeonieae of the Crassulaceae family (Eggli, 2008). The genus comprises 15 species with the centre of diversity on the Canary Islands (11 species; Bañares Baudet, 2002, 2015a, Baudet, 2017), three species on Madeira, and one species on the island of Santa Maria in the Azores (Moura et al., 2015). *Aichryson* is divided into two sections, sect. *Aichryson* and sect. *Macrobia* Webb & Berthel. Section *Macrobia* includes only *Aichryson tortuosum* (Aiton) Webb & Berthel., a perennial, small shrub endemic to Lanzarote (subsp. *tortuosum*) and Fuerteventura (subsp. *bethencourtianum* Botte & Bañares). All other species belong to sect. *Aichryson* and are monocarpic, mostly annual herbaceous plants (Bañares, 2015a). Within sect. *Aichryson* several natural hybrids are described (Bañares, 2015b). *Aichryson* proved to be monophyletic and likely sister to *Monanthes icterica* (Webb ex Bolle) Christ in molecular phylogenetic studies on Aeonieae based on cp markers and ITS (Mort et al., 2002; Fairfield et al., 2004). The genus comprises both diploid and tetraploid species (Uhl, 1961; Suda et al., 2005).

We sampled a total of 29 individuals representing 14 species of *Aichryson* (only *A. santa-mariensis* M.Moura, Carine & M.Seq. is missing) and two accessions of *Monanthes icterica* as outgroup (Supplementary Table 1, "sampling"). For 20 samples we were able to assess the ploidy level on a CyFlow cytometer (PARTEC) using the isolation buffer "OTTO I" (2.1 g Citric-acid-1-hydrat, 10 ml 5% Triton X-100, 90 ml ddH_2O). FloMax v2.8.2 (QA GmbH, Münster, Germany) was used for the particle analysis and the measurement of the peaks (Table S1, "flow cytometry"). For the remaining samples, published ploidy levels were incorporated (Uhl, 1961; Suda et al., 2005).

DNA-extraction was conducted using the DNeasy Plant Mini-Kit (QIAGEN, Venlo, Netherlands) according to the manufacturer's protocol for "Purification of Total DNA from Plant Tissue (Mini Protocol)" with a number of modifications outlined in the online Appendix 1. The DNA concentration and quality were evaluated using a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA) and gel electrophoresis.

2.2. *In silico digestion, restriction enzyme choice and adapter design*

The search for suitable restriction enzymes for our approach was performed *in silico* and based on 1) the desired fragment length (300-600nt), 2) the number of samples per library (up to 50), 3) the expected sequencing output of the MiSeq v3 kit (up to 25 Million), and 4) the targeted sequencing depth (aimed at ~ 10 × per fragment), leading to the required fragment yield of 5,000 within the target length range. Initially we tested commonly used REases individually. However, the REases tested did not meet our requirements, thus we tested combinations of two REases each. For this, we have taken into account a minimum length of 6nt for the recognition site and the simultaneous applicability of two REases in a single reaction. The *in silico* digestion was performed using the software CLC genomics Workbench v9.5.5 (Qiagen) with its included "Restriction Site Analysis" for several genomes of various focal groups including *Beta vulgaris* L., Amaranthaceae (BioProject PRJNA41497) and *Kalanchoe fedtschenkoi* Raym.-Hamet & H.Perrier, Crassulaceae (BioProject PRJNA397334). The resulting restriction maps were evaluated with respect to fragments showing two cut sites within the desired length window of 300-600nt. Among other suitable REase combinations, the REases *BamHI* (G'GATCC) and *KpnI* (GGTAC'C) best met our criteria for a double digest (for excerpts of the REase selection, see also Supplementary Table S2, "*in silico* digest"). In case of *Aichryson*, the *in silico* digest of the distantly (yet closest) related *K. fedtschenkoi* genome (divergence to Aeonieae is roughly 58.60 [44.60–73.62] mya, Messerschmid et al., 2020), resulted in 61,692 fragments, of which 4,429 fragments fell in the targeted length range.

In contrast to widely established strategies (Elshire et al., 2011; Peterson et al., 2012; Andrews et al., 2016), we aimed at sequencing all generated fragment types, including fragments framed by identical restriction motifs. Thus, we designed the barcode and common adapters for both REases motifs (Table S2, "BamHI adapter", "KpnI adapter"). The barcode sequences were obtained from Elshire et al. (2011) and van Gurp (2017). Both barcode and common adapter fit to the overhang of the *BamHI* and *KpnI* cut sites (Fig. 1b). We were able to achieve the set aim with this design, however, we recommend a more flexible adapter/indexing strategy that accounts for technical bias during wet lab and sequencing (e.g. MacConaill et al., 2018; Bayona-Vásquez et al., 2019).

2.3. *RADseq*

The major changes compared to other RADseq approaches such as ddRADseq (Peterson et al., 2012) or Genotyping by Sequencing (GBS; Elshire et al., 2011) are: the usage of two rare cutter REases that produce c. 5,000 fragments within a target range of 300-600nt (Fig. 1a), adapters binding to all generated fragments (Fig. 1b), an extended size selection range (Fig. 1d) and an extra size selection step during the final purification (Fig. 1f). In particular the two size selections were important to fully exploit the sequencing range (see also Appendix 1, Fig. A1.6, A1.7). Since the RADseq toolbox includes many modifiable modules, various protocols might be capable of generating libraries/datasets of an extended length range and we encourage an impartial testing of this potential (see also: McCormack et al. 2013b; Andrews et al., 2016; McKain et al., 2018; Parchman et al., 2018). The following is a brief overview of the workflow. For the detailed protocol, see Appendix 1 and Supplementary Table S3.

2.3.1. *RADseq lab workflow*

We used 200 ng genomic DNA as input for the double digest reaction (Fig. 1a), which was followed by adapter ligation (Fig. 1b) in the same reaction tube. For thorough saturation of cut sites, 6 µl adapter working solution (0.5 ng/µl) containing equimolar amounts each motif pair were used. Reactions were incubated for 3 h at 37 °C, respectively. The libraries were multiplexed using 100 ng DNA each (Fig. 1c), followed by a column-based cleaning of the pool. Size selection (Fig. 1d) was performed using Pippin Prep (Sage Science, Beverly, MA, USA) with a segregation range of 350-720nt. The size-selected products were amplified using a low-cycle 2-step PCR protocol (Fig. 1e). Subsequently, PCR products were collected in three pools (Table S3), purified and quantified. Final purification, accompanied by the 2nd size segregation, was done using the NucleoMag NGS kit (Macherey-Nagel, Düren, Germany) with a ratio of 0.8 bead suspension to one part library. The purified library was resuspended in 25 µl Buffer AE for sequencing.

2.3.2. *Library quality assessment and sequencing*

Library quality was validated by measuring the DNA concentration by Qubit Fluorometer and assessing the fragment distribution by Bioanalyzer electropherogram (Appendix 1). Sequencing was performed on an Illumina MiSeq (San Diego, CA, USA; Reagent Kit v3 600-cycle) at StarSEQ (Mainz, Germany) producing 300nt PE reads in three different runs (Supplementary Table S4).

2.4. *Data assembly*

2.4.1. *Raw sequence treatment*

Raw data quality was assessed with FastQC 0.11.4 (Andrews, 2010; Fig. 2a; Table S4 "run I-III"). Raw reads were demultiplexed (Table S2 and S3) using *ipyrad* v0.9.52 (Eaton and Overcast, 2020) twice, once for each REase cut site (Fig. 2a). This two-fold demultiplexing was necessary due to the motifs occurring on both read directions. The fastq-files were combined and adapter sequences were removed with Cutadapt 1.18 (Martin, 2011). FastQC reports of the demultiplexed/adapter trimmed samples were combined using MultiQC v1.9 (Ewels et al., 2016; Table S4 "mean quality scores").

2.4.2. *ipyrad*

We used *ipyrad* v0.9.52 (Eaton and Overcast, 2020) for *de novo* RADseq assembly. Several filtering parameters of the *ipyrad* pipeline (v9 or above, Eaton and Overcast, 2020) represent percentages, allowing the application of the selected thresholds to variable read lengths and thus supporting clustering of datasets obtained by a broad sequencing range. We used default parameters, except for the ones outlined below.

2.4.3. *Assembly parameter settings*

The de-multiplexed samples were split into two groups according to ploidy level (di- or tetraploid; Table S1). The diploid dataset contained nine *Aichryson* samples, the tetraploid dataset contained 18 *Aichryson* and two *Monanthes icterica* samples. Parameter #18 (max_alleles_consens) was set to two and four, respectively (Supplementary Table S5). With respect to the extended read length, we allowed up to 24 indels per locus (parameter #23). We assumed increased gene flow and set parameter #24 to 0.7 (Bañares, 2015b; max_Hs_consens). Parameters #11 and #12, which give the minimum depth for statistical and majority

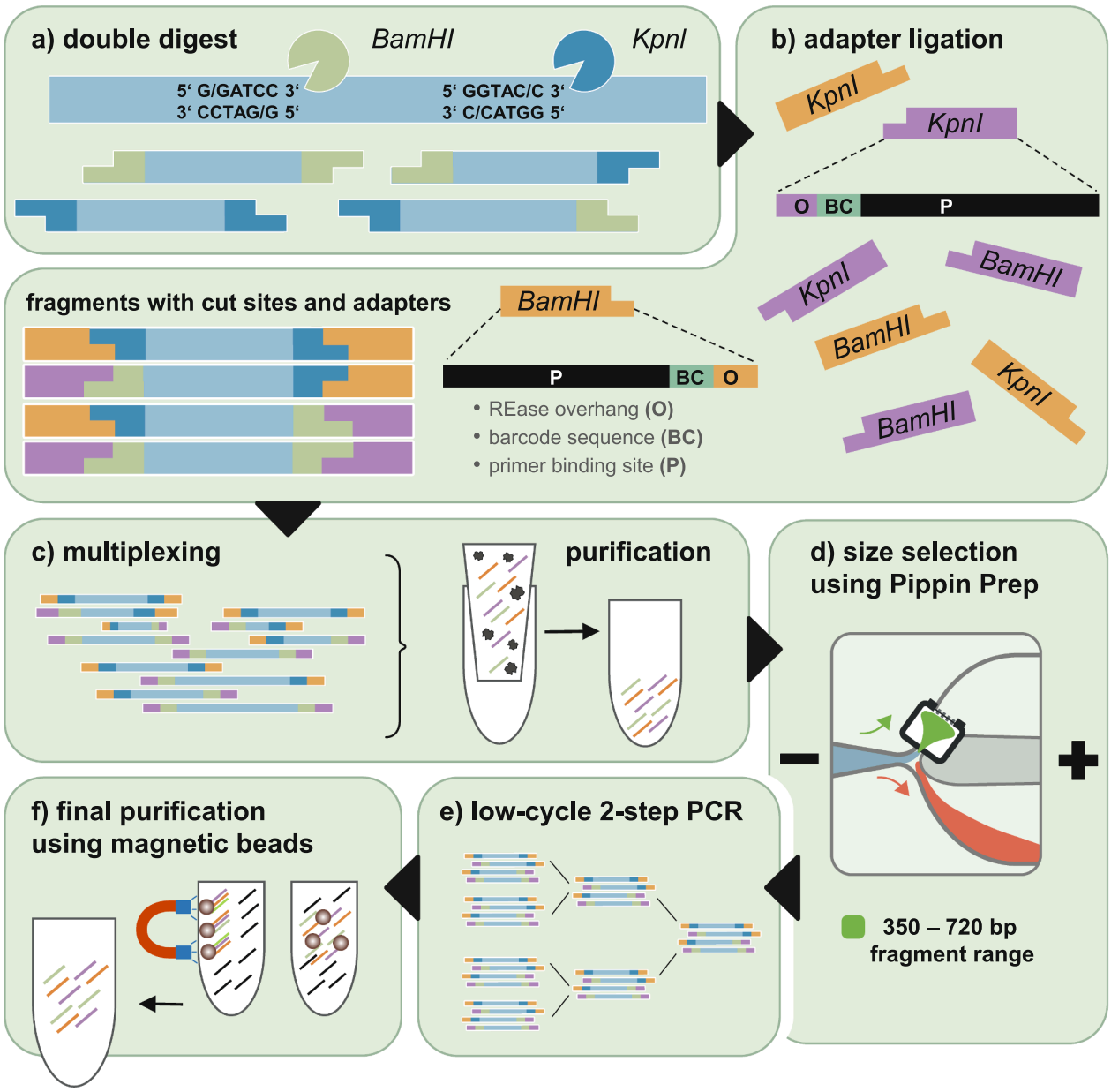


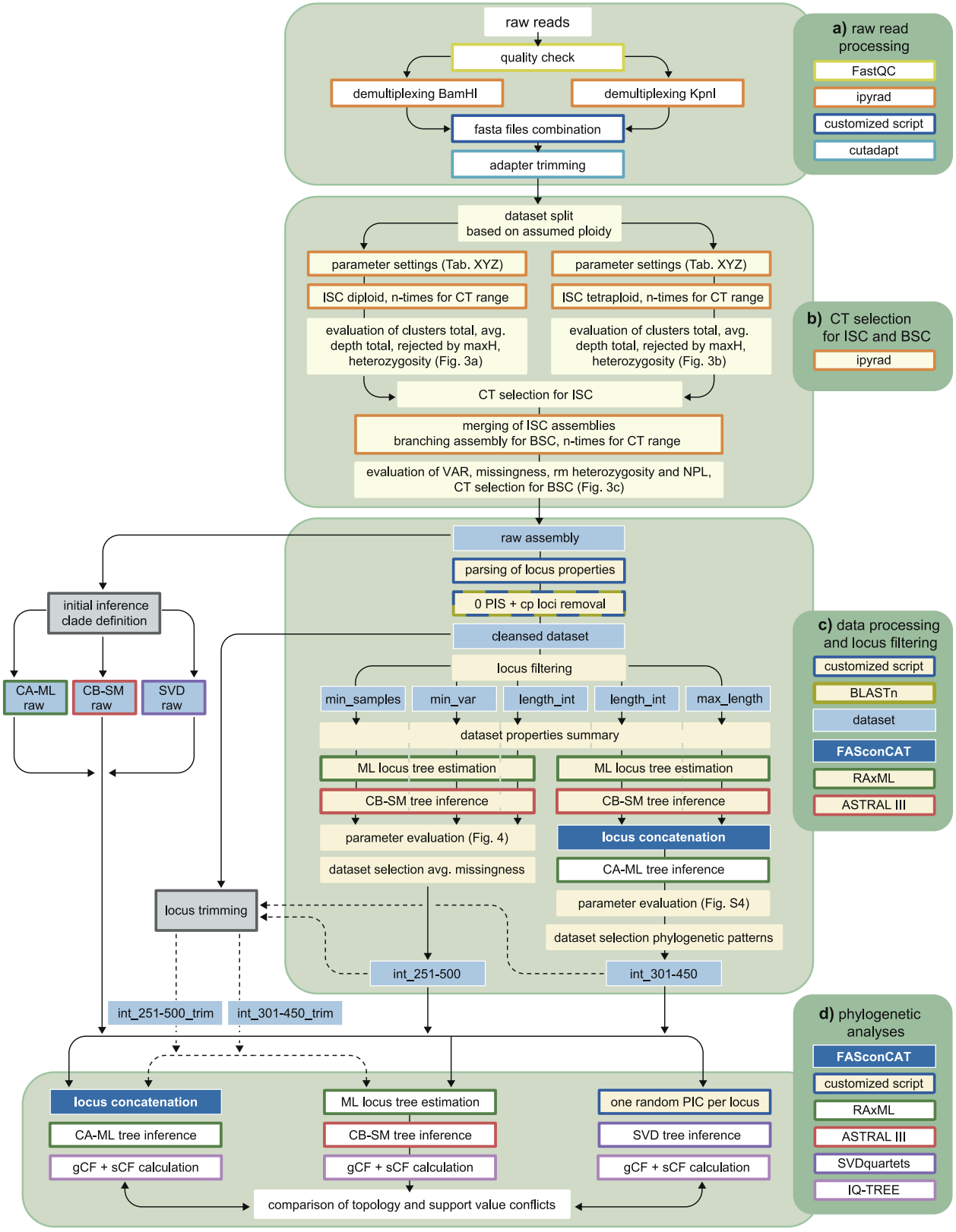
Fig. 1. The lab workflow of the modified RADseq protocol consists of six steps (a – f). a) Genomic DNA is digested simultaneously using the REases *BamHI* and *KpnI*. b) Barcode and common adapters are ligated to the fragments. c) The barcoded samples are multiplexed and purified. d) The pool is size selected to a 350 – 720 bp length range using Pippin Prep. e) The size selected pool is amplified using a low-cycle 2-step PCR. f) The final purification using magnetic beads removes PCR and size selection artifacts.

rule base calling, were set to 10. We aimed at an average cluster depth (avg_depth_mj) of > 20 × for statistical base calling (Pamilo et al., 2011; Eaton and Overcast, 2020).

2.4.4. Selection of suitable clustering thresholds for ISC and BSC

Avoiding both, over- and undermerging of putative loci is not trivial in high-throughput datasets. If the selected CT is too lax, paralogous reads will be incorrectly clustered and treated as orthologs (overmerging) and if the selected CT is too strict, reads belonging to an actual locus will incorrectly be split into several loci (undermerging) with low variability (Supplementary Figure S1.A). To determine suitable CTs for ISC and BSC, we used several CT selection approaches as guidance (Ilut et al., 2014; Mastretta-Yanes et al., 2015; Paris et al., 2017; McCartney-Melstad et al., 2019) and defined the assumptions to determine suitable CTs. 1) Over- and undermerging ranges have to be identified to avoid merging/splitting effects within these areas. 2) Overmerging is indicated by highly heterozygous clusters/alleles with a high proportion of filtered paralogs (Ilut et al., 2014; McCartney-Melstad et al., 2019). Hence, a suitable CT is expected in an increasing area of heterozygosity and a decreasing area of flagged paralogs, between the maxima of both metrics. 3) Undermerging of orthologs leads to an increased number of loci (and lower locus coverage in ISC, lower sample coverage per locus in BSC) while sequence divergence among taxa decreases (Mastretta-Yanes et al., 2015; McCartney-Melstad et al., 2019). Thus, sequence variation declines while missingness increases. A suitable CT is expected near a steep increase in number of clusters/loci and amount of missingness while heterozygosity is biologically realistic (ISC) and locus variability is high (BSC).

To prevent introducing a potential bias due to ploidy, we split the samples into two groups (di- and tetraploid) for ISC assembly (*ipyrad* assembly steps 1–5, Fig. 2b). Following ISC CT selection, all samples were merged for BSC (*ipyrad* assembly steps 6 and 7). A CT range of



(caption on next page)

Fig. 2. The schematic overview of the data analysis is split into four major parts (a-d, boxes on the right side). The boxes in light blue indicate sub-/datasets. Dashed arrows illustrate parameter applications between datasets. Colored box edges show the software used for the work step. During the raw read processing (a) the quality is assessed using FastQC, the reads are demultiplexed two times with respect to the REase cut sites and the sample specific barcodes, combined into sample fasta files, and adapter and cut sites are removed using cutadapt. For the clustering threshold (CT) selection approach (b), the data set is split based on the assumed ploidy and the *ipyrad* parameters are adjusted as required. For in-sample-clustering (ISC) a CT range of 0.81 – 0.99 is tested for both datasets and *ipyrad* outputs are evaluated with respect to the number of total clusters, total average read depth, clusters rejected by maxH (flagged paralogs) and heterozygosity (Fig. 3a and b). The selected ISC assemblies are merged and branched to test the CT range (see above) for between-sample-clustering (BSC). The resulting assemblies are evaluated with respect to the number of retained loci, the retained sequence variation (VAR), missingness and the number of new polymorphic loci (NPL, Fig. 3c). The selected "raw" assembly is used for initial phylogenetic inference and clade definition (c). The locus properties (locus ID, length, number of samples, number of SNPs, PIS and VAR) are parsed using a customized script. Loci showing no variation and chloroplast loci are removed. The loci of the "cleansed" dataset are filtered into several sub-datasets based on their properties. The first locus filtering approach, using a missingness threshold for dataset selection, resulted in the "int_251-500" dataset. The second filtering approach, using sub-dataset properties and resulting phylogenetic patterns for dataset selection, resulted in the "int_301-450" dataset. The truncated loci of the "raw" assembly were re-arranged based on the selected datasets of the locus filtering (locus truncation, dashed arrows). The datasets ("raw", "int_251-500", "int_301-450" and "short") are used for comparative phylogenetic inference (d). Individual loci are either concatenated using FASconCAT for CA-ML inference or used to calculate ML locus trees as input for CB-SM inference. The SVD datasets are created by picking a single randomly selected parsimony informative character (PIC) of each locus. To assess the resulting trees of the tested inference methods across datasets, we compared changes in BS support values and gene (gCF) and site concordance factor (sCF) values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

0.81–0.99 (in 0.01 increments) was tested. To assess the above-mentioned criteria for CT selection, we plotted a variety of metrics as a function of the tested CT range as box- and scatter plots (see also Figure S1.B and S1.C). For the ISC CT selection, we evaluated the number of clusters (clusters_total), the average read depth (avg_depth_total), the number of filtered paralogs (filtered_by_maxH) and the heterozygosity. For the BSC CT selection, we additionally evaluated the number of retained loci, sequence variation (VAR, SNPs and PIS) and proportion of missingness (sequences_missing). In addition, we calculated the "new polymorphic loci" (NPL) in order to detect the assembly containing most accurately clustered sequence variation, which is indicated by the so-called "hockey stick signal" (Paris et al., 2017). We expected the transition zone from over- to undermerging to be characterized by trend changes, e.g. prominent differences in the medians of adjacent CTs and compressions or expansions of the quartiles (in boxplots) or changes in the slope intensity (in scatter plots). Multiple suitable CTs within a transition zone of a metric and across metrics were averaged to determine a consensus CT.

2.4.5. Processing of the unfiltered ipyrad assembly

The *ipyrad* loci-file of the unfiltered "raw" assembly was parsed with a custom Perl script (available on GitHub https://github.com/philipphuehn/RADseq-locus-filtering) for the specific locus ID, the length, the number of samples, SNPs and PIS (VAR in total) and the proportion of missingness for each locus (Fig. 2c "parsing of locus properties"). We used BLAST + 2.7.1 (Camacho et al., 2009) to identify chloroplast loci by blasting all loci against four reference plastomes from the Crassulaceae (GenBank accessions: *Sedum uniflorum* subsp. *oryzifolium* (Makino) H.Ohba: NC_027837, *Sedum sarmentosum* Bunge: NC_023085, *Phedimus takesimensis* (Nakai) 't Hart: NC_026065, *Phedimus kamtschaticus* (Fisch. & C.A.Mey.) 't Hart: NC_037946). Loci of a plastid origin as well as loci showing no parsimony informative sites were removed (Fig. 2c, "0 PIS + cp loci removal"). In addition, this "raw" assembly was used for initial phylogenetic inference and clade definition to compare potentially different phylogenetic results from subsequently filtered datasets (see 3.4.1).

2.5. Locus filtering and dataset selection

In general, phylogenetic inference by CB-SM is very sensitive to GTEE, which most often is caused by loci showing little sequence variation, high missingness or fractional coverage (Chou et al., 2015; Roch and Warnow, 2015; Xi et al., 2015, 2016; Xu and Yang, 2016; Sayyari et al., 2017; Hosner et al., 2016; Lee et al., 2018; Molloy and Warnow, 2018). We filtered the here generated RADseq loci into several sub-datasets to test for a potential influence of locus properties on phylogenetic inference (Fig. 2c, "locus filtering"). First, we determined the impact of the locus properties on CB-SM reconstruction (see 2.5.1). This filtering approach suggested a potential impact of biased phylogenetic signal due to non-randomly distributed partial taxon coverage (Sanderson et al., 2010, 2011, 2015; Simmons, 2012; Xi et al., 2015; 2016; Hosner et al., 2016; Sayyari et al., 2017; Dobrin et al., 2018). This so-called "biased missingness" has been shown to cause high GTEE, thus results in conflicting, unsupported locus trees and consequently in a decline of species tree estimation performance (Xi et al, 2015, 2016; Hosner et al., 2016; Sayyari et al, 2017; Molloy and Warnow, 2018). We therefore performed a second locus filtering with respect to locus length and evaluated phylogenetic patterns of CB-SM and CA-ML reconstructions (see 2.5.2). The locus filtering scripts are available at GitHub (https://github.com/philipphuehn/RADseq-locus-filtering).

2.5.1. Locus filtering by coverage, variability, length intervals and dataset selection based on average missingness

The loci were filtered with respect to the average variability (var = VAR/locus length/number of samples; "min_var"), minimum number of samples per locus (number of samples/locus; "min_samples"), and locus length intervals ("length_int") and rearranged to new sub-datasets (Fig. 2c, "locus filtering", Supplementary Table S6). For the "min_var" sub-datasets, seven thresholds were used (0.01, 0.25, 0.50, 0.75, 1.0, 2.0, 3.0, "min_var_001" – "min_var_300"). Six thresholds by increments of four were used for the "min_samples" sub-datasets (4, 8, 12, 16, 20, 24, "min_samples_4" – "min_samples_24"). The locus length interval datasets were created based on eight intervals starting from the minimum length to 250nt, and then ranging by 50nt steps from 251nt to 550nt, and 551nt to the maximum length ("int_min-250" – "int_551-max"). Properties of these datasets, such as the total number of loci, VAR, SNPs, PIS (average per locus), sample coverage/missingness, and average locus length were recorded (Fig. 2c, "sub-dataset properties summary"). For each rearranged sub-dataset, ML locus trees were estimated and used for CB-SM inference (see 2.6.2). We recorded the bootstrap support values of all branches of each tree and assigned them to three categories: backbone, clade and within clade branch support values. Clade branches contained all samples of the defined clades (see 3.4.1 for clade definition). All support values within the defined clades were assigned to within clade branches. All other support values, spanning from the outgroup to the clade branches, were recorded as backbone support values. Topology changes and conflicts were not accounted for. Based on this and on recommendations by studies investigating the impact of locus filtering for summary methods (Xi et al., 2016; Sayyari et al., 2017; Molloy and Warnow, 2018), we selected an average missingness threshold to filter the locus sets (Fig. 2c, "dataset selection avg. missingness"). The resulting dataset was subsequently used for comparative phylogenetic inference (Fig. 2d).

2.5.2. Locus filtering by length and dataset selection based on sub-dataset properties and phylogenetic patterns

In order to narrow down the suspected dataset bias in terms of fractional, non-random locus and/or taxon coverage, we used phylogenetic patterns to assess sub-datasets filtered by length. CA-ML inference of datasets exhibiting this type of bias can result in unsupported or overly high supported polytomies resolved as a terraced topology (Sanderson et al., 2010, 2011, 2015; Simmons, 2012; Dobrin et al., 2018). Dobrin et al. (2018) have reported numerous empirical multilocus datasets to be impacted by this issue (e.g. Springer et al., 2012; Burleigh et al., 2015; Shi and Rabosky, 2015). Since we generated ML locus trees as input for species tree estimation with CB-SM, we assumed this terraced topology to also appear if the bias of the underlying data was strong. Besides, Hosner et al. (2016) and Sayyari et al. (2017) found that a high proportion of fragmentary data (biased incongruence of locus trees) can lead to a sharp drop of the resulting BS support values for CB-SM inference.

In addition to the length interval sub-datasets of the first filtering ("int_min-250" – "int_551-max"), we filtered the loci requiring an increasing, cumulative maximum length (Fig. 2c, "locus filtering", Supplementary Table S7). The eight maximum locus length sub-datasets were generated starting at a threshold of 250nt ("max_250", all loci up to 250nt length were included) increasing by 50nt increments up to the maximum locus length. Each sub-dataset was subjected to phylogenetic inference using CA-ML and CB-SM. The sub-dataset properties and resulting BS support values were recorded as described in 2.5.1.

While bootstrapping across a concatenated matrix almost automatically increases the resulting support values with increasing matrix size (Kubatko and Degnan, 2007; Liu et al., 2015; Minh et al. 2020a), the multi-locus bootstrapping used with CB-SM employs a 2-stage system that accounts for variations among loci by resampling during BS calculation (Seo, 2008) and thus reacts very sensitive to fragmentary data (Xi et al., 2015, 2016; Hosner et al., 2016; Sayyari et al., 2017). We expected the BS support values to collapse as soon as the ratio of biased to unbiased data (respecting a non-randomly distributed partial taxon coverage) became too high. For CA-ML, we expected a similar but less sensitive pattern, in particular for the sub-datasets of an increasing maximum locus length.

For the evaluation of a terrace-like topology pattern, the number of samples resolved on terraced branches was recorded. We defined that a terraced branch must either -originate from a dichotomous branch of the tree's backbone, - the clade's backbone containing that sample, - or must follow an individual branch within a clade, - but must not be included within a dichotomous constellation. For instance, phylogenetic inference of the "raw" dataset using CA-ML, CB-SM and SVD resulted in two, five and three terraced branches for clade 5, respectively (Supplementary Figure S2). The SVD tree contained another terraced branch in clade 4, but the CA-ML and CB-SM trees did not. By increasing the maximum locus length required, we expected the topology to switch from a terraced to a dichotomous tree pattern once the biased area has been passed or compensated (and vice versa). CB-SM was expected to react more sensitive than CA-ML due to the reduced amount of data, with individual gene trees as input (Xu and Yang, 2016). Therefore, the terraced pattern was assumed to be over-expressed once the amount of data became too small (in particular for the length interval sub-datasets), and likewise a larger portion of unbiased data would be needed for compensation (for the maximum length sub-datasets).

The dataset, which was intended to be a reasonable compromise for both methods, had to meet the following criteria: 1) relatively low average missingness, 2) relatively high ratio of PIS to SNPs, 3) relatively high BS support values for all tree sections, 4) relatively low number of samples resolved on terraced branches, 5) and had to avoid over- and under-represented assembly regions. The selected dataset was used for comparative phylogenetic inference (Fig. 2d).

2.5.3. Generating 'short' loci by locus truncation

The loci of the *ipyrad* "raw" assembly were truncated to one third of their original length to compare potential performance differences of the here generated loci to a RAD dataset obtained by assembly of 100nt PE reads. These shorter loci were intended to show less sequence variation and thus negatively affect phylogenetic inference. The truncated loci were re-arranged based on the selected datasets of the locus filtering (Table 1, Fig. 2c, "locus truncation").

2.6. Phylogenetic inference

We have chosen three commonly used approaches for phylogenetic inference of the generated main- and sub-datasets (Table 1, S6 and S7). CA-ML and CB-SM were used for inference during locus filtering. For the comparative phylogenetic inference, we additionally used SVDquartets as third inference approach. We decided not to test a full-coalescent method that uses co-estimation of locus trees and species trees such as implemented in BEST (Liu, 2008) or BEAST 2 (Bouckaert et al., 2014) because computation time and capacities required increase sharply with the number of loci and samples. Thus, full-coalescent methods are currently not practical for large-scale datasets with thousands of loci (e.g. Bayzid and Warnow, 2013; McCormack et al., 2013a; Zimmermann et al., 2014).

2.6.1. Phylogenetic inference with RAxML (CA-ML)

We used RAxML v8.2.12 (Stamatakis, 2014) to infer maximum likelihood phylogenies using GTRGAMMA as substitution model, 20 runs for BestML and 1,000 bootstrap replicates to assess statistical support of relationships. We used the unfiltered *ipyrad* supermatrix for inference of the "raw" assembly. For all other datasets, we concatenated individual loci to a supermatrix using FASconCAT v1.11 (Kück and Meusemann, 2010).

2.6.2. Species tree inference with ASTRAL-III (CB-SM)

ASTRAL-III v5.7.4 (Zhang et al., 2018) estimates species relationships based on gene/locus trees. To generate these locus trees, we used RAxML v8.2.12 (Stamatakis, 2014) under the GTRGAMMA model with 20 runs for BestML and 1,000 bootstrap replicates. ASTRAL was run in default mode using unrooted locus trees. We used multilocus bootstrapping (Seo, 2008) to compute branch support for the estimated species trees.

2.6.3. Svdquartets analysis (SVD)

SVDquartets (Chifman and Kubatko, 2014) is a quartet-based algorithm to compute species trees from SNP datasets. We used FASconCAT-G (Kück and Longo, 2014) to extract and concatenate the 25,320 parsimony informative characters (polymorphisms that are shared by at least two samples, PICs) of the 3,818 loci constituting the "raw" assembly. To meet the requirement for linkage of the dataset (sites must be unlinked), we randomly selected a single PIC of each informative locus for each dataset (Table 1, "unlinked PICs"). Analyses were run in SVDquartets as implemented in PAUP*4.0a168 (Swofford, 2003) with 1,000 bootstrap replicates under the multilocus bootstrap (Seo, 2008). The scripts for generating PIC datasets are available at GitHub (https://github.com/philipphuehn/RADseq-locus-filtering).

2.6.4. IQ-TREE analysis

We used IQ-TREE v2.1.2 (Minh et al. 2020a, b) to calculate the gene (gCF) and site concordance factors (sCF) of the resulting phylogenies, which give the percentage of decisive locus trees and alignment sites containing or supporting a specific branch in a given reference tree, respectively. Locus trees obtained with RAxML were used for gCF calculation. For sCF calculation, 1000 quartets were used to obtain stable estimations. To assess the resulting phylogenies with respect to a potential influence of biased data, we put the resulting topologies and BS support values in context with the gCF and sCF values and value

Table 1
The properties of the unfiltered "raw" assembly, the "cleansed" dataset, the datasets selected by locus filtering and their length truncated variants.

dataset	raw	cleansed	int_251-500	int_301-450	int_251-500_short	int_301-450_short
loci	3,818	3,225	2,788	1,599	2,788	1,599
VAR total	71,691	68,490	56,448	33,480	18,590	10,625
VAR per locus	18.78 (±16.69)	21.24 (±16.82)	20.24 (±15.70)	20.94 (±16.15)	6.67 (±5.65)	6.65 (±5.63)
SNPs total	36,413	33,261	26,533	15,673	8,779	5,040
SNPs per locus	9.54 (±5.25)	10.31 (±9.89)	9.51 (±8.66)	9.80 (±8.86)	3.15 (±3.17)	3.15 (±3.16)
PIS total	35,278	35,229	29,915	17,807	9,811	5,585
PIS per locus	9.24 (±10.73)	10.92 (±10.86)	10.73 (±10.67)	11.14 (±11.05)	3.52 (±3.93)	3.49 (±3.91)
unlinked PICs total	2,723		2,220	1,287		
locus coverage	8.86 (±5.25)	9.37 (±5.45)	9.67 (±5.57)	9.96 (+-5.62)	9.67 (±5.57)	9.96 (+-5.62)
sample coverage	1,166 (±467)		930 (±333)	549 (±204)		
missingness avg. [%]	69.79	67.69	66.66	65.64	66.66	65.64
locus length avg. [nt]	376 (±93)	379 (±93)	360 (±70)	373 (±43)	120 (±23)	123 (±18)

Given are the total number of loci (loci), the total and average values per locus (standard deviations in parentheses) for the number of variable sites (VAR), single nucleotide polymorphisms (SNPs), and parsimony informative sites (PIS), the total number of unlinked PICs as input for SVD inference, and the average locus coverage (samples per locus), sample coverage (loci per sample), the average proportion of missingness [%] and the average locus length [nt].

differences. In general, both concordance factors are expected to be similar if the phylogenetic signal is only impacted by discordant signal, e.g. due to ILS (Minh et al. 2020a, b). If other processes affect the dataset, such as limited information or a data bias, the gCF values can be a lot lower than the sCF values, resulting in large factor value differences. A large proportion of conflicting signal or a significant variation of sites in the dataset can lead to a completely random resolution, which is indicated by sCF values ~ 33%. The reasons are either true phylogenetic signal caused by ILS or biased signal caused by uneven coverage. Distinct factor value differences of alternative topologies may indicate non-phylogenetic signal.

3. Results

3.1. Final library and MiSeq output

The fragment distribution of the final library ranged from ca. 370-770nt. The majority of fragments outside the target range were successfully removed (Appendix 1, Fig. A1.4, A 1.5). The MiSeq runs generated a total of 6,870,208 paired raw reads for the 29 samples (Table S4, "samples"). Sequence quality decreased with increasing read length (Table S4, "run I-III"). The quality of the R2 reads started to decline below a Phred quality score of 20 from ca 260nt read length (Table S4, "mean quality scores"). The number of reads per sample ranged from 98,754 for *A. laxum* var. *latipetalum* Bañares & M.Marrero to 587,377 for *M. icterica* BG Bonn with an average of 236,903 reads per sample. Demultiplexed raw data is available at the NCBI Sequence Read Archive in BioProject PRJNA642981.

3.2. ISC and BSC threshold selection

In general, the plots of the selected metrics showed the expected trends and met the requirements (Fig. 3 and S1.B and S1.C). For the ISC metrics, however, the indicators were not as distinct as expected. The transition zones of the metrics were averaged to consensus CTs for the diploid and tetraploid samples, respectively (Supplementary Table S8).

For the ISC of diploid samples (Fig. 3a and S1.B, "ISC 2n"), the onset of the undermerging area was initiated by an abrupt increase in the number of clusters at CT 0.95, which was indicated by a compression of the third quartile (Q3) for the CTs 0.93 and 0.94 and a simultaneous increasing slope intensity in the scatter plots (Fig. 3a and S1.B, "clusters total", transition zone: 0.93–0.94). Allelic variation was highest in the transition zone of 0.92–0.95 and started to decrease strongly with increasing sample coverage (Fig. 3a and S1B, "heterozygosity"). The peak CT for heterozygosity was 0.92 (transition zone: 0.92–0.95) while the paralog peak was 0.88 (transition zone: 0.88–0.95). These maxima were preceded by irregular jumps of adjacent medians and an intensity change of the slopes (Fig. S1.B). This area was enclosed by the transition zone of the average read depth per cluster trend, which was indicated by an increasing Q3 compression and a steady slope shift (Fig. 3a, Fig. S1.B, "avg. depth total", transition zone: 0.92–0.95). The CTs within the described transition zones were averaged to a consensus CT of 0.93 (Table S8, "ISC consensus CT").

For the ISC of tetraploid samples (Fig. 3b and S1.C, "ISC 4n"), undermerging was initiated by a Q3 compression within the transition zone of the number of clusters and increased in slope from CT 0.94 on (Fig. 3b and S1C, "clusters total", transition zone: 0.92–0.93), while allelic variation also started to decline steeply with increasing CTs (Fig. 3b and S1.C, "heterozygosity", peak at 0.94, transition zone: 0.89–0.94). The transition zone of the average depth per cluster showed a steadily declining trend, a few slight median jumps and an increasing Q2 compression (Fig. 3b and S1.C, "avg. depth total", transition zone: 0.89–0.92). The transition zone of filtered paralogs showed a prominent median jump and a moderate slope decline towards the undermerging area (Fig. 3b and S1.C, "filtered by maxH", peak at 0.90, transition zone: 0.90–0.92). The averaged consensus CT was 0.91 (Table S8, "ISC consensus CT").

The scatter plots of the ISC metrics showed that some samples can have a larger effect on the overall trend of a metric than others. For instance, the sample "A_tort_RIII_A36_J49" (*A. tortuosum* subsp. *tortuosum*) showed one of the lowest average cluster depths ("avg. depth total") while a high number of clusters ("clusters total") was found (Fig. S1.B). It also showed the highest amount of filtered paralogs ("filtered by maxH") and a two times higher heterozygosity than the other diploid samples, although flow cytometry confirmed its diploid status (Table S1). The tetraploids also showed some samples that were clearly different from the others (Figure S1.C).

For the BSC threshold selection (Fig. 3c), the undermerging area was indicated by the steady increase in retained loci while the sequence variation (VAR) started to decrease at CT 0.92. At this point, the missingness of the assembly matrix was still low before it increased abruptly starting at CT 0.92. According to McCartney-Melstad et al. (2019) and Mastretta-Yanes et al., (2015), a suitable CT is right before the decrease in sequence variation and the steep increase in missingness while the sample coverage (retained loci) still increases, at CT 0.91. The hockey-stick signal was identified by the first positive swing of the "blade" following the NPL minimum (Fig. 3c, "new polymorphic loci", Paris et al., 2017). This upward swing was in the transition of the CTs 91/90 that corresponds to a CT of 0.91 (Table S8, "NPL") and thus supports the other requirements. We selected 0.91 as BSC threshold.

3.3. ipyrad assembly output

The average total read depth (avg_depth_total) for the diploid and tetraploid samples was 6.21 (±2.17) at CT 0.93 and 5.55 (±1.80) at CT 0.91, respectively (Supplementary Table S9, "ISC 2n", "ISC 4n"). After

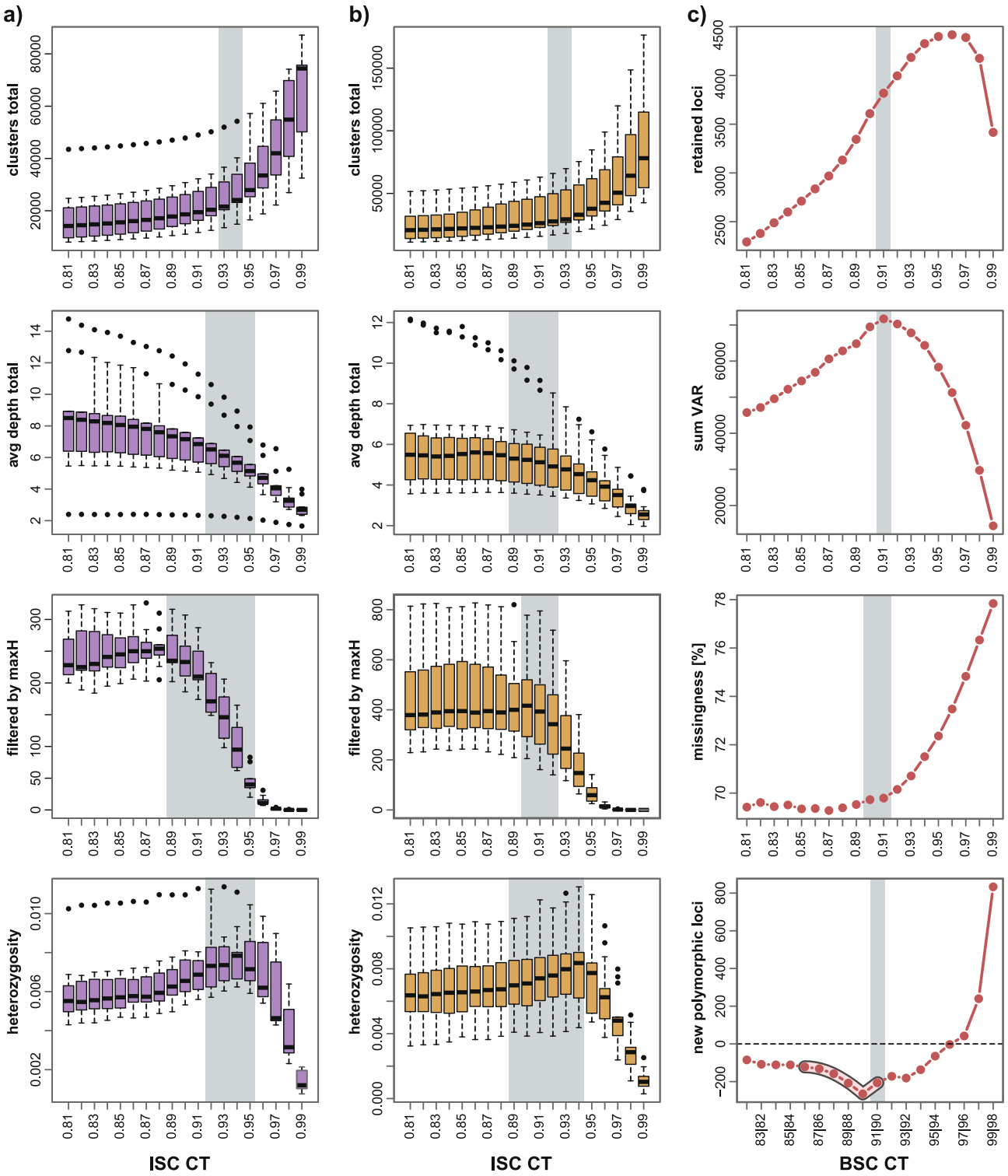


Fig. 3. To determine suitable thresholds for in-sample-clustering (ISC) and between-sample-clustering (BSC), trends of several metrics tested across a CT range of 0.81–0.99 were evaluated. For ISC threshold selection of the diploid (a) and tetraploid (b) samples, the number of clusters, the average read depth, flagged paralogs (filtered by maxH) and the allelic variation (heterozygosity) were recorded and plotted. Transition zones from the over- to the undermerging area containing several suitable CTs are shaded in grey. CTs within these zones were averaged to a consensus CT. To select a suitable threshold for clustering between samples of the merged ISC assemblies, the number of retained loci, the retained sequence variation (VAR), the missingness and the number of new polymorphic loci (NPL) were recorded (c). The "hockey stick signal" in the NPL plot, which indicates the assembly containing most accurately clustered sequence variation, is in line with the requirements for the other metrics.

applying the min_depth threshold of 10 for clustering, the majority read depth (avg_depth_mj) rose to 40.24 (±7.52) for the diploid and 39.22 (±17.10) for the tetraploid samples. On average, 26,280 (±11,873) clusters per individual were found for the diploids and 34,436 (±15,023) clusters per sample for the tetraploids. The average count of consensus reads was 2,635 (±692) for the diploid and 2,633 (±645) for the tetraploid samples.

The unfiltered assembly using a BSC threshold of 0.91 comprised 3,818 loci and 71,691 variable sites (Table 1, Fig. 2, "raw" assembly). Of these variable sites, 36,413 were unique SNPs and 35,278 were PIS. 92 loci showed no variation and 581 loci contained no PIS. The dataset included 69.79% missingness, on average 10 unique SNPs and 9 PIS per locus. The retained loci had an average length of 376nt (±93) with a maximum locus length of 618nt (including uncalled bases and gaps). The majority of retained loci ranged in length from 250 to 550nt (Table S9, "locus coverage"). The assembly length range > 500nt showed a prominent gap at ca. 540-580nt, after which a denser region with some samples of comparatively high coverage followed, at ca 590nt. Locus coverage per sample was fairly heterogeneous with an average of 1,242 (±385) and ranged from a minimum of 640 loci for *A. laxum* A29_J41 to a maximum of 2,092 loci for *A. roseum* A01_J02 (Table S9, "sample coverage"). The two outgroup samples contained 127 (*M. icterica* M30_N36) and 155 loci (*M. icterica* BG Bonn) in the final assembly. The BLAST results showed that our dataset contained 21 loci (118 SNPs and 66 PIS) with identities of 78.5–100% with the reference plastomes. After removing non-parsimony-informative loci and cp loci, the dataset contained 3,225 loci with an average of 67.69% missing data. Each locus contained on average 10 SNPs and 11 PIS and had an average length of 379nt (±93) (Table 1, "cleansed", Fig. 2c, "cp loci + 0-PIS loci removal").

3.4. Initial inference of the raw dataset and clade definition

Phylogenetic inference of the *ipyrad* "raw" assembly resulted in incongruent topologies (Table 2, Fig. S2). CA-ML (Fig. S2.A) and CB-SM (Fig. S2.B) yielded unsupported backbones, while the SVD reconstruction was fully supported (Fig. S2.C). All trees showed five well supported main clades: clade 1 comprised *A. laxum*, *A. pachycaulon* subsp. *parviflorum* and *A. palmense*, clade 2 included two subspecies of *A. pachycaulon,* subsp. *immaculatum* and subsp. *pachycaulon*, clade 3 was formed by three species from Madeira (*A. villosum*, *A. dumosum* and *A. divaricatum*), clade 4 comprised both subspecies of *A. tortuosum* and clade 5 comprised all remaining taxa (*A. roseum*, *A. punctatum*, *A. bituminosum*, *A. porphyrogenetos*, *A. brevipetalum*, *A. bollei* and *A. parlatorei*) as well as two subspecies of *A. pachycaulon,* i.e., *A. pachycaulon* subsp. *praetermissum* and subsp. *gonzalezhernandezii*. Relationships among clades was not resolved due to a lack of reliable BS support among reconstructions.

3.5. Locus filtering

The 3,225 loci of the "cleansed" dataset (Table 1, Fig. 2c) were first filtered respecting the locus coverage (minimum number of samples required), the locus variability (VAR/locus length/number of samples) and locus length intervals by 50nt steps. The properties of the resulting sub-datasets were recorded and phylogenies were inferred using CB-SM (see 3.5.1, Table S6, Supplementary Figure S3, all tree files available at Mendeley, https://doi.org/10.17632/yb6fd93dbw.1). For the second locus filtering, in addition to the length interval sub-datasets, the loci were filtered requiring an increasing, cumulative maximum length ("max length") and subjected to phylogenetic inference using CA-ML and CB-SM (see 3.5.2, Table S7, Supplementary Figure S4, all tree files available at Mendeley, https://doi.org/10.17632/yb6fd93dbw.1).

3.5.1. Locus filtering by coverage, variability, length intervals and dataset selection based on average missingness

We created six "min samples" sub-datasets by increments of four (Fig. 4a, Table S6, Fig. S3). The locus count and sequence variation (total) decreased as the number of samples increased (Fig. S3.A1 and S3.A4). The average number of SNPs per locus remained nearly constant across datasets, whereas the number of PIS per locus increased proportionately with VAR/locus until the "min_samples_16" dataset and then remained constant when increasing the parameter (Fig. S3.A4 and S3.B1). As expected, missingness declined with increasing number of samples (Fig. S3.B1 and S3.C1). The average locus length was constant across the datasets (Fig. S3.B1 and S3.C4). The branch support values of the CB-SM phylogenies showed a steady, slightly decreasing pattern across the datasets (Fig. S3.D1 and S3.D2). The backbone and within clade support values were around 80 and dropped by ca ten points with the "min_samples_24 dataset". The average clade branch support was close to 100 in all datasets.

Seven sub-datasets were filtered for the "min var" parameter (Fig. 4b, Table S6, Fig. S3). The number of loci and sequence variation (total) decreased with increasing minimum variability (Fig. S3.A2 and S3.A5).

Table 2
Bootstrap support values and concordance factor values and differences of the inferred datasets using CA-ML, CB-SM and SVD.

inference method	CA-ML			CB-SM			SVD		
dataset	raw	int_251-500	int_301-450	raw	int_251-500	int_301-450	raw	int_251-500	int_301-450
BS backbone branches	86.80	99.20	99.20	83.06	90.14	96.70	100	100	100
BS clade branches	94.80	100	99.60	99.92	99.98	99.08	100	100	100
BS within clade branches	93.71	95.29	94.41	80.25	83.92	83.94	100	100	100
BS all branches	92.63	96.89	96.26	84.41	88.05	89.10	100	100	100
CF clade 1	44.4; 69.2; 24.8	44.5; 68.8; 24.4	46.7; 69.4; 22.7	45.4; 69.1; 23.8	45.0; 68.0; 23.0	47.9; 68.7; 20.7	45.6; 70.0; 24.4	44.5; 69.0; 24.6	45.1; 61.3; 16.2
CF clade 2 + 3	48.1; 62.3; 14.2	48.7; 62.3; 13.6	50.0; 58.5; 8.5	42.6; 72.4; 29.8	43.1; 72.8; 29.7	44.6; 70.0; 25.4	43.2; 72.1; 28.8	41.8; 71.7; 29.9	43.1; 73.7; 30.6
CF clade 4	40.1; 64.1; 24.0	40.5; 64.5; 24.1	42.5; 66.1; 23.6	36.4; 60.1; 23.7	40.9; 65.4; 24.5	42.5; 65.8; 23.3	37.6; 54.4; 16.8	40.9; 63.8; 22.9	36.6; 59.1; 22.6
CF clade 5	17.4; 57.8; 40.5	17.5; 57.8; 40.3	17.4; 58.4; 41.0	16.1; 59.8; 43.7	18.9; 61.0; 42.1	19.9; 61.5; 41.6	18.7; 60.8; 42.1	18.4; 61.7; 43.3	17.3; 60.4; 43.1
CF clade branches	56.2; 83.3; 27.1	55.9; 83.2; 27.3	58.7; 81.7; 22.9	51.5; 80.8; 29.3	55.8; 81.5; 25.7	59.3; 80.2; 20.9	57.6; 83.2; 25.6	54.9; 81.7; 26.8	53.2; 79.2; 26.1
CF backbone branches	58.6; 75.9; 17.3	55.9; 69.5; 13.6	57.9; 71.2; 13.3	55.0; 68.1; 13.1	55.9; 69.6; 13.7	57.9; 71.6; 13.6	49.1; 62.7; 13.6	49.9; 66.7; 16.7	48.9; 64.8; 15.9

Given are the average BS support values (sectional and total) and the average gene (gCF) and site concordance factor (sCF) values (of the within clade branches, the clade branches and backbone branches) of the inferred datasets ("raw", "int_251-500″, "int_301-450″, "int_251-500_short", "int_301-450_short") using CA-ML, CB-SM and SVD. The average concordance factor (CF) values are shown in this order: gCF; sCF; gCF-sCF-difference.

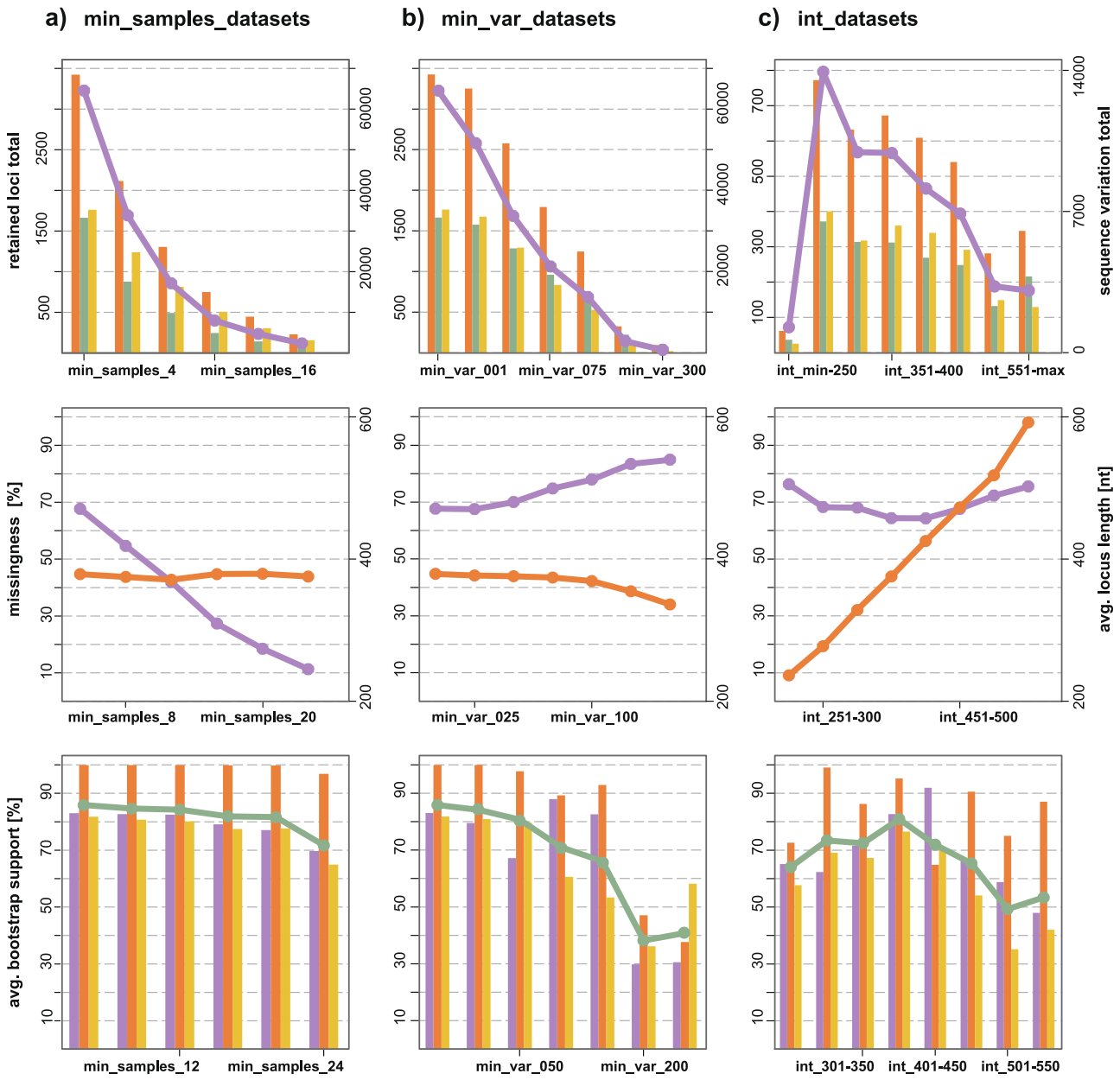


Fig. 4. The loci of the „cleansed" assembly were rearranged into sub-datasets based on the minimum number of samples required (a), the minimum variability required (b) and locus length intervals (c). For each sub-dataset, properties such as the number of retained loci (upper plots, purple line with data points), sequence variation (orange = VAR, green = SNPs, yellow = PIS), the average missingness (middle plots, purple line with data points) and average locus length (orange line) were recorded. The average BS support values of the resulting CB-SM trees are given in total (bottom plots, green line with data points) and for the three sections (purple = backbone branches, orange = clade branches, yellow = within clade branches). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In terms of the sequence variation (VAR) total and per locus, the ratio of SNPs to PIS shifted towards a higher SNPs proportion with increasing required minimum variability (Fig. S3.A5 and S3.B2) and missingness increased as well (Fig. S3.B2 and S3.C2). The average locus length decreased slightly with increasing variability required, with the "min_var_300" sub-dataset showing a clear shift towards shorter loci (Fig. S3.B5 and S3.C5). The BS support values showed a decreasing trend (Fig. S3.D2). The tree topologies received varying support across the sub-datasets. The backbone branches were supported highest for the "min_var_075" and "min_var_100" datasets, while the clade and within clade branches had highest support values in the "min_var_001, 025, 050" datasets. The average branch support decreased with increasing missingness (Fig. S3.D5).

The properties and resulting support values of the eight length interval datasets showed irregular trends (Fig. 4c, Table S6, Fig. S3). The amount of loci and sequence variation total (excluding the first sub-dataset containing only 72 loci) dropped from the highest value at "int_251-300" to the adjacent dataset, then rose and declined moderately until the next sharp decline from "int_451-500" to "int_501-550" (Fig. S3.A3 and S3.A6). The average sequence variation per locus was rising with increasing locus length. The proportions of SNPs and PIS in sequence variation (VAR) shifted towards a higher proportion of parsimony-uninformative sequence variation for the datasets "int_min-250" and "int_551-max", respectively (Fig. S3.B3). The missingness had a slightly convex trend with maxima for the flanking datasets (Fig. S3.B6 and S3.C3). The steadily increasing trend of the locus length showed unexpected averages for the two datasets containing the longest loci (Table S6, Fig. S3.B6 and S3.C6), matching the uneven locus length

distribution of the "raw" assembly (Table S9, "locus length distribution" and "locus coverage"). The resulting branch support values showed contrasting patterns (Fig. S3.D3 and S3.D6). The overall trend was shaped concavely. The backbone support initially increased to a maximum at "int_401-450" and then decreased with increasing locus length. The average clade support values were highest at "int_251-300", "int_351-400" and "int_451-500". The within clade branches were supported highest by the "int_351-400" sub-dataset, embedded in a descending trend towards the dataset edges.

Regarding the "min_samples" and the "min_var" datasets, the results were as expected and consistent with findings of previous studies (e.g. Chen et al., 2015; Huang and Knowles, 2016; Eaton et al., 2017; Molloy and Warnow, 2018,). For both parameters, the overall support decreased with increasing requirements, likely due to the simultaneous decline in number of loci and sequence variation. The irregular trends of the locus length interval datasets provided useful clues for subsequent dataset selection and further filtering (see 3.5.2). The trends observed here, together with the declining read quality (Table S4), the heterogeneous coverage of samples and loci, and the irregular assembly coverage respecting the over- and under-represented locus length ranges from ca. 250-280nt and ca. 540-580nt (Table S9), fit the definition of so-called "biased missingness" (Xi et al, 2015, 2016; Hosner et al., 2016; Sayyari et al, 2017; Molloy and Warnow, 2018). To reduce this impact, we selected the average proportion of missingness (69.58% for the length interval datasets) as threshold and discarded all datasets above this cut-off. The retained "int_251-500" dataset (Table 1, "int_251-500") consisted of 2,788 loci, containing in total 56,448 (20.24 ± 15.7 on average) VAR, 26,533 (9.51 ± 8.66) SNPs, 29,915 (10.73 ± 10.76) PIS, 66.66% missingness (9.67 samples/locus) and the average locus length was 360 (±70nt). The locus truncation to one third of the original length lead to a 2/3 reduction of sequence variation and locus length (Table 1, "int_251-500_short").

3.5.2. *Locus filtering by length intervals and increasing maximum length and dataset selection based on data qualities and phylogenetic patterns*

The conspicuous trends of the length interval datasets in terms of SNPs/PIS ratio and missingness/locus coverage relative to the resulting BS support values of the species tree sections motivated further filtering to narrow down the extent of potential biased missingness (Table S7 and Fig. S4).

For the locus length interval datasets, CA-ML showed the lowest and highest average BS support values for the "int_0-250" and "int_251-300" datasets, respectively (Fig. S4.C2 and S4.D1). The average branch support decreased steadily with increasing locus length. The three branch sections were irregularly supported by different sub-datasets. The highest count of terraced branches was found for both CA-ML and CB-SM for the "int_0-250" dataset (Fig. S4.D2). The second highest counts were recorded for the sub-datasets "int_501-550" and "int_551-max", respectively. CA-ML resolved the fewest terraced branches for the "int_301-350" and "int_351-400" sub-datasets. The CB-SM trees showed the smallest counts for the "int_251-300" and "int_351-400" datasets, with the latter having the highest average BS support value (Fig. S4.D1 and S4.D3).

For the maximum length datasets, CA-ML showed the lowest sectional and total average BS support values for the first two sub-datasets (Fig. S4.C4 and S4.D2). Then, the BS support raised sharply for the "max_350" sub-dataset and increased steadily up to the maximum for the "max_500" sub-dataset. Beyond this point, there was no gain in branch support. The CB-SM branch support values were lowest for the "max_250" sub-dataset, increased slightly until the "max_350" sub-dataset, showed a strong gain for the "max_400" and a maximum value for the "max_450" sub-dataset (Fig. S4.C2 and S4.D2). Then, the average BS support decreased with increasing maximum locus length, in particular the backbone section lost support. CA-ML and CB-SM resolved the highest terraced branch count for the first sub-dataset (Fig. S4.D4). The number of terraced branches decreased to a minimum of two for the CA-ML trees with increasing maximum length required. CB-SM resolved the fewest terraced branches for the "max_400" sub-dataset. With the addition of loci up to 500nt length ("max_500") the terraced branch count increased strongly and remained high up to the maximum locus length ("max_length").

For the final dataset selection, we classified all recorded locus properties of the sub-datasets and the phylogenetic patterns of the resulting trees into three categories, respectively (Fig. S4.E). The two extreme datasets of both assembly edges were either over- or under-represented (Table S9, "locus length distribution" and "locus coverage"). Those sub-datasets showed also a higher or almost equal ratio of SNPs to PIS relative to the average VAR per locus (Fig. 4, "int_datasets", Table S7, Fig. S4.A1-A6). The average missingness was highest for the filtering parameter edges and decreased towards the inner medium parameters (Fig. S4.B1-B4). The expected average locus lengths were met by the inner filtering parameters, while the values of the sub-datasets increasingly diverged towards the assembly edges (Fig. 4, "int_datasets", Table S7, Fig. S4.B5 and S4.B6). Both CA-ML and CB-SM showed the highest sectional and total BS support values for the inner filtered sub-datasets, with the highest gain for the "max_350" and "max_450" sub-datasets (Fig. S4.C1-C4 and S4.D1-D2). The BS support values of the backbone section profited most within this locus length range. Both approaches resolved the highest number of terraced branches for the filtering parameter edges (Fig. S4.D3-4). The terraced branch count decreased with increasing maximum locus length and increased again strongly beyond a locus length of 450nt for the CB-SM trees. With this locus length also the BS support values started decreasing steadily (Fig. S4.D2 and S4.D4). In summary, the locus properties and phylogenetic patterns associated with non-randomly distributed missingness or biased data were strongest at the filtering parameter edges, while the length ranges from 300 to 450nt appeared to be less affected (Fig. S4.E). The selected "int_301-450" dataset (Table 1) consisted of 1,599 loci of an average length of 373nt (±43nt), containing 15,673 SNPs (avg. 9.82), 17,808 PIS (avg. 11.24) and 65.56% missingness. Truncation resulted in a 2/3 reduction of locus properties (Table 1, "int_301-450_short").

3.6. *Phylogenetic inference*

We used three datasets for comparative phylogenetic inference (Table 1, Fig. 2c and 2d). The 3,818 loci of the "raw" assembly were used for initial inference and clade definition (see 3.4). We removed both cp and non-informative loci from this dataset. The retained 3,225 loci of the "cleansed" dataset were the input for the locus filtering approach (see 2.5 and 3.5). The first locus filtering by coverage, variability and length intervals resulted in the "int_251-500" dataset (see 2.5.1 and 3.5.1). The second locus filtering was intended to reduce the presumed biased phylogenetic signal by using phylogenetic patterns relative to the underlying sub-dataset qualities to detect impacted assembly areas (see 2.5.2 and 3.5.2). This approach yielded the "int_301-450" dataset. The filtering steps reduced the number of loci by 58% and the amount of PIS by 50% (Table 1, "raw" compared to "int_301-450"). Sequence variation and locus coverage increased slightly while the average missingness decreased by 4%. Loci-per-sample coverage decreased from an average of 1,166 to 549 loci while sample-per-locus coverage became more homogenous (Table S9, "sample coverage"). Hence, we assumed the "raw" assembly to contain the most, the "int_251-500" dataset to contain less, and the "int_301-450" dataset to contain the least biased phylogenetic signal. Filtering parsimony informative characters (unlinked PICs) resulted in three datasets for the SVD analyses (Table 1). The loci of the "raw" assembly were truncated to one third of their original length, rearranged respecting the locus filtering results and species relationships were inferred with CA-ML and CB-SM to compare potential performance differences in terms of locus length (Table 1, "short", Fig. 2c and 2d).

3.6.1. Comparative phylogenetic inference of the un-/filtered datasets

For the "raw" datasets, CA-ML (Fig. S2.A) and CB-SM (Fig. S2.B) resolved incongruent and weakly supported backbone topologies. The CA-ML tree showed an unresolved relationship between the clades 2, 3 and 4. CB-SM inference resulted in an unresolved relationship of clade 1 to clades 2, 3 and 4, with low support and low concordance factor values. The SVD tree (Fig. S2.C) showed full support for a third topology. However, the concordance factor values for the relationship of clade 1 to clade 5 were low. The within clade topology differed among all reconstructions.

For the "int_251-500" dataset, CA-ML (Supplementary Figure S5.A) and CB-SM inferences (Fig. S5.B) resolved congruent backbone topologies, however, for CB-SM the relationships of clades 2 + 3 + 4 to clade 5 lacked support. The concordance factor values increased compared to the "raw" dataset. The SVD tree (Fig. S5.C) showed a maximally supported conflicting topology with low concordance factor values for the relationship of clade 2 to clades 1 + 3 + 4. The within clade topology differed among all reconstructions.

For the "int_301-450" dataset, CA-ML (Supplementary Figure S6.A) and CB-SM (Fig. S6.B) inference resulted in a well-supported, congruent backbone topology (Fig. 5). Concordance factor values for the backbone and clade branches were similar. Again, the SVD tree (Fig. S6.C) showed a maximally supported conflicting topology but low concordance factor values for the relationship of clade 2 to clades 1 + 3 + 4.

3.6.2. gCF and sCF values obtained with IQ-Tree

Dataset reduction with respect to the exclusion of potentially biased assembly areas, clearly showed an improvement regarding the concordance factor values and differences for the CB-SM reconstructions (Table 2). The factor difference decreased for all within clade branches and clade branches. The factor values of the clades 1, 2, 3, and 5 decreased stronger compared to clade 4. The gCF value of the clade branches increased by more than 8% compared to the unreduced dataset, while the sCF value decreased slightly. Interestingly, the factor values of the backbone branches increased slightly while the difference increased slightly as well.

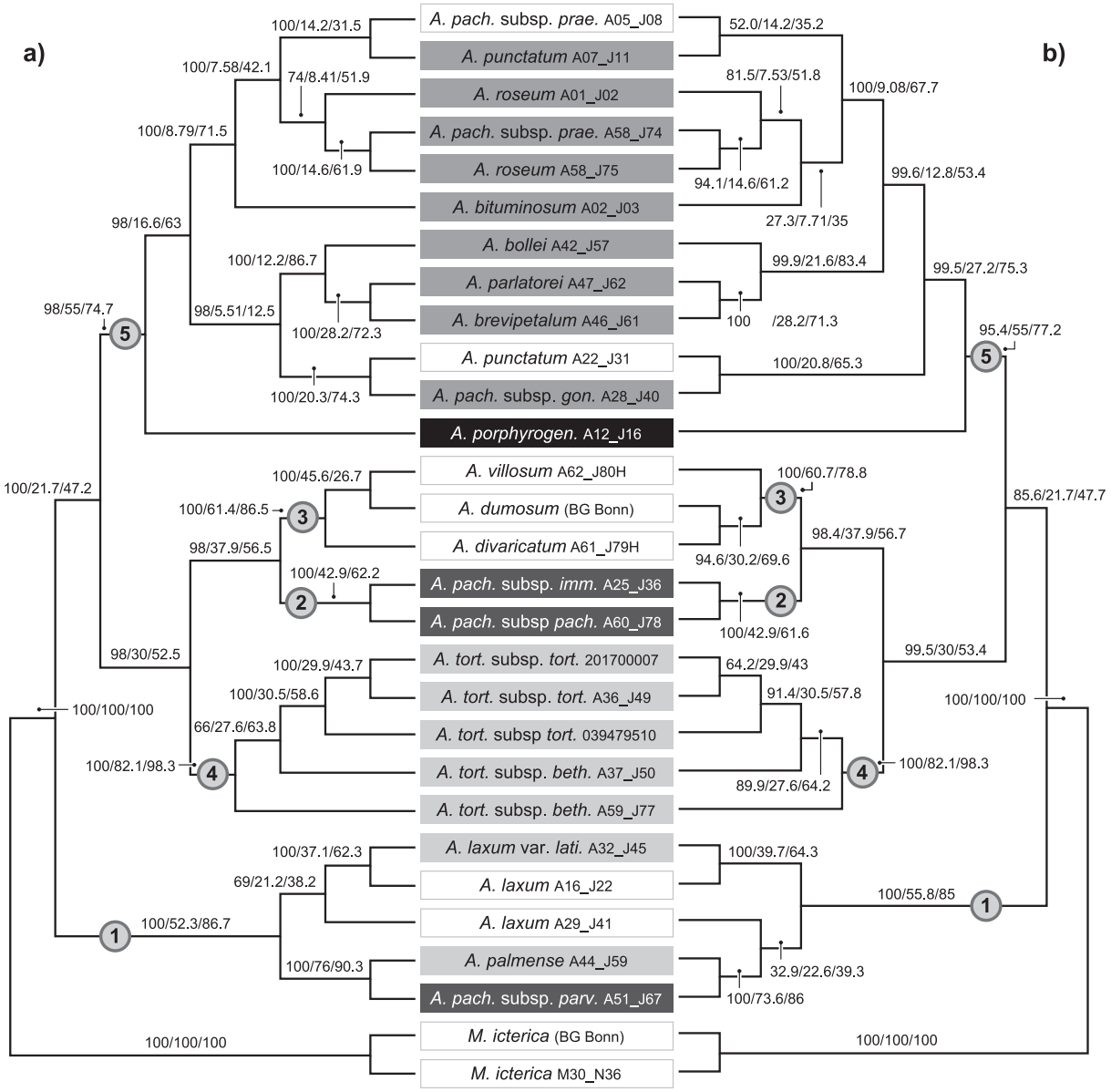


Fig. 5. The CA-ML (a) and CB-SM (b) phylogenies of the "int_301-450" dataset. Bootstrap support, gene and site concordance factor values are given above branches. Clades are indicated by the encircled numbers 1–5. Boxes shaded in light and dark gray indicate diploid and tetraploid samples, respectively. The sample *A. porphyrogennetos* A12_J16 showed an intermediate genome size and was treated as tetraploid (black box).

Concordance factors of CA-ML inference showed a similar pattern compared to CB-SM. Overall, the factor values increased with increasing dataset reduction. However, the effect was less pronounced compared to CB-SM and clade 5 even showed an increased factor difference. Notably, the effect of the factor differences for the clade branches was smaller while for the backbone branches it was larger, compared to the CB-SM reconstruction.

In general, the factor effects of the SVD reconstructions were in strong contrast to CB-SM and CA-ML. The SVD factor values were lower compared to CB-SM and CA-ML, and the factor differences raised for the clade and the backbone branches. For the within ancestral branches of clade 2 + 3 and all descendant relationships, the factor difference decreased strongly.

In terms of the resulting BS support values, data reduction had the strongest effect on the backbone branches with an increase in support by ~ 13% for CA-ML and ~ 16% for CB-SM (Table 2). Still, the gCF and sCF values suggest alternative topologies for the relationship of clades 2 + 3 + 4 to clade 5.

3.6.3. Phylogenetic inference of the truncated locus datasets

Inference of the truncated datasets using CA-ML (Supplementary Figure S7.A and S7.B) and CB-SM (Fig. S7.C and S7.D) resulted in alternative topologies compared to the full-length datasets, while also exhibiting distinctly lower concordance factor values and larger factor differences (Supplementary Table S10) or insufficient BS support for the backbone section. The BS support values decreased with decreasing locus length and the decrease was strongest in the backbone branches. The concordance factor values were mostly lower compared to the full-length datasets and the factor difference for the clade and backbone branches increased clearly for all reconstructions.

4. Discussion

Modification of several modules of the RADseq toolbox, inspired by GBS (Elshire et al., 2011) and ddRADseq (Peterson et al., 2012), has enabled a strong reduction of the number of targeted fragments. In addition, employing the maximum capacity for sequencing resulted in an extended locus length of up to 618nt. The CT selection approach enabled an informed selection of ISC/BSC thresholds for homology assessment of assembled loci. The locus filtering approach, based on properties known to affect phylogenetic inference, provided the opportunity to observe dataset-specific trends and identify potential adverse properties of the sub-datasets. Additional filtering using phylogenetic patterns for bias detection turned out to improve overall resolution, in particular for CB-SM inference. Besides these positive outcomes, there were also many challenges whose critical consideration led to suggestions for further improvements.

4.1. Lab workflow

Compared to other studies employing a RADseq approaches for sample preparation (e.g. Escudero et al., 2014; de Oca et al., 2017; Dillenberger and Kadereit, 2017; Hamon et al., 2017; Wagner et al., 2018; Gerschwitz-Eidt and Kadereit, 2019; Paetzold et al., 2019; Rancilhac et al., 2019; Hipp et al., 2020; Karbstein et al., 2020) we increased the fragment length range and thus the length of assembled loci clearly by shifting the size selection window and fully exploiting the sequencing range of 300nt PE. However, the raw reads varied strongly both in quantity and quality across the samples, which led to a loss of locus and sample coverage, in particular within the higher length range targeted (Supplementary Table S9). This biased distribution of phylogenetic information represented a substantial challenge to data evaluation.

Our lab workflow aims at long RAD loci and has been modified in three aspects: First, we included a specific size selection window ranging from 300 to 600nt for the resulting fragments of the utilized REases BamHI and KpnI. Second, barcode and common adapters were designed for both REase motifs to sequence all generated fragment types in contrast to the classic ddRADseq approach (compare Peterson et al. 2012). Third, the lab protocol contained two size selection steps to ensure complete removal of fragments outside the target range.

4.1.1. Employed REases

The flexible RADseq toolbox allows the use of various REases of a wide range of qualities for complexity reduction (see also: Andrews et al., 2016; McKain et al., 2018; Parchman et al., 2018). Testing and comparing single and dual enzyme strategies with respect to the desired degree of reduction, or in particular a reduced fragment number and an extended length range, either *in silico* or by sequencing a trial library when there is no reference available, can certainly reduce mutation-based locus dropout and ease library prep and adapter design (see also: Lepais and Weir, 2014; Mora-Márquez et al. 2017; Rivera-Colón et al., 2021). Double-digest approaches, using two REases for digestion (e.g. Peterson et al., 2012), are more prone to restriction site mutation disruption than single-digest protocols (e.g. Elshire et al., 2011). Hence, they tend to yield fewer fragments than single-digest approaches which are therefore more easily sequenced to sufficient depth (Andrews et al., 2016; Harvey et al., 2016; Eaton et al., 2017; McKain et al., 2018). Using the *K. fedtschenkoi* genome for *in silico* double-digest using *BamHI* and *KpnI*, we calculated about 4,400 fragments (see 2.2) and received about 3,800 assembled loci (Table 1, "raw"). The difference of ca. 600 fragments may be due to the loss of loci in the assembly range above 500nt (Table S9). Compared to capturing approaches, which typically produce loci of up to thousands of base pairs in length (e.g., McCormack et al. 2013a; Nicholls et al., 2015) the herein obtained locus length of Ø 376nt and 618nt at most may seem short. Still, the resulting loci showed sufficient sequence variation per locus as input for species tree estimation using CB-SM and were in line with approaches targeting similar length ranges (e.g. Hosner et al., 2016; Blom et al., 2017).

4.1.2. Adapter design

The design of adapters herein was based on the original GBS protocol to include and sequence all generated fragments (see Elshire et al., 2011). However, this approach proved not satisfactory. It did not account for potential chimera formation and index hopping (see also: Van der Valk et al., 2020) and the identical flow cell binding motifs meant a potential reduction in sequencing yield. While in general the sequencing output was not influenced, the second sequencing run, containing the majority of samples, yielded only 50% of the maximum sequencing output of the MiSeq v3 kit (Table S4, "run III"). In addition, the reads flanked by identical cut sites introduced a further step in data processing and locus assembly that could be avoided as the raw data had to be demultiplexed twice. Considering these hurdles, we recommend to design each adapter type for one cutsite motif only and to use an indexing approach that accounts for technical bias (e.g. MacConaill et al., 2018; Bayona-Vásquez et al., 2019).

4.1.3. Size selection window and fragment/locus length distribution

The use of coalescent-based summary methods for phylogenetic inference requires a relatively high quality content of sequence variation per locus to reduce GTEE (Chou et al., 2015; Liu et al., 2015; Mirarab et al., 2016; Xu and Yang, 2016; Molloy and Warnow, 2018). Because the average amount of phylogenetic information in a neutrally evolving locus generally correlates with its length (Blom et al., 2017; Mirarab et al., 2016; Chou et al., 2015; Molloy and Warnow, 2018), we chose a size selection window of approximately 300-600nt (ca. 380-720nt segregation range including the adapter and primer length) to obtain longer fragments and thus more informative loci (Fig. 1, Appendix 1). The 2nd size selection using a ratio of 0.8 parts magnetic bead suspension to one part library suspension is particularly important as it removes fragment artifacts from automated fragment segregation and PCR (Fig. 1f, Appendix 1). Compared to a library prepared with the same protocol but without final purification, the precision of the fragment

length segregation was clearly improved (Appendix 1). The length distribution of the final assembly was overall in the range targeted by the lab protocol. However, the strong decline in sequencing quality of R2 reads (Table S4, "run I-III", "mean quality scores") has resulted in a large degree of missingness in the length range of 500-600nt of assembled loci (Table S9, "locus length distribution"). Moreover, the quality filtering thresholds were set quite strictly (Table S5; Eaton and Overcast, 2020). This prevents assembly of erroneous sequences by discarding reads below a specified threshold for base and overall quality. In our dataset this applied especially to the R2 reads, starting at ca 260nt. Thus, a lot of information was lost by excluding high quality partners of low quality mates. Tan et al. (2019) found that declining base quality and higher error rates of fragments above 500nt are a general issue with multiple Illumina sequencing platforms and kits.

The descriptive analysis of the filtered sub-datasets showed that phylogenetic information across the length intervals provided varying support for different sections of the resulting species trees (see 2.5.2 and 3.5.2, Fig. 4c, Fig. S4, "length interval datasets"). Maximum support for all sections was covered by a locus length range of 300-450nt. Considering this and the decreasing quality of R2 reads, we recommend a size selection window of 300-500nt (ca. 380-620nt segregation range including the adapter and primer length). This might avoid locus loss due to the decreasing sequencing quality of the R2 reads and thereby achieve a more uniform assembly and evenly distributed phylogenetic information. However, other focal groups than *Aichryson* might require longer loci, as the retained variation per locus depends on the taxonomic level of interest and is very group specific.

4.2. Data analysis

Assembly and analysis of RADseq data is often challenged by various factors depending on the selected library prep and bioinformatics approach, and, of course, the study group itself. The *Aichryson* data shown here united just about every conceivable challenge known to RADseq data. The samples had varying DNA qualities and were sequenced in three different libraries. The output of the three sequencing runs differed in terms of quantity and quality. The R2 reads showed an unevenly distributed drop in quality starting at about 260nt sequencing length (Table S4). And it turned out that this dataset had not only a high proportion of missing data, but also of biased missingness across the assembly length range, impacting sample and locus coverage (Table S9). Despite these unfavorable circumstances, or maybe because of them, the detailed analyses (Fig. 2), including a CT selection and a locus filtering approach, provided detailed insights into the data properties and their impact on phylogenetic inference.

4.2.1. CT selection approach

Clustering threshold selection approaches aim at determining balanced CTs to establish homology while avoiding clustering of paralogous RADseq loci (e.g., Ilut et al., 2014; Mastretta-Yanes et al., 2015; McKinney et al., 2017; Paris et al., 2017; McCartney-Melstad et al., 2019). For this purpose, assembly metrics are compared across a range of CTs to identify values that meet specified requirements. Application of such methods is becoming increasingly popular (e.g. Herrera and Shank, 2016; Razkin et al., 2016; Paetzold et al., 2019; Rancilhac et al., 2019; Karbstein et al., 2020; Wagner et al., 2020) to ensure the assembly of homologous loci (Shafer et al., 2017; Springer and Gatesy, 2018; McCartney-Melstad et al., 2019; Fernández et al., 2020; Simion et al., 2020). Following these previously proposed criteris, we were able to identify areas that met the requirements in terms of 1) the onset of the undermerging area, in which true orthologs are separated into paralogs (McCartney-Melstad et al., 2019), 2) an area of high heterozygosity with decreased clustering of paralogs (Ilut et al., 2014), 3) a maximized sequence variation count while missingness is minimized (Mastretta-Yanes et al., 2015), and 4) an increasing number of new polymorphic loci (NPL) indicated by the hockey stick signal (Paris et al., 2017). This procedure resulted in an assembly comprising 3,818 loci, of which ~ 84% contained parsimony informative sites (Table 1). The loci showed on average ~ 19–21 variable sites, of which ~ 9–11 were parsimony informative. Since these loci were found to be useful for CB-SM inference, we consider the here selected metrics and CT selection approaches in general as promising tools for an informed selection of thresholds during *de novo* assembly. Still, there are some issues that need to be considered: 1) The results shown herein and assumptions arising from them provide more empirical evidence on previous studies, however, are highly specific to our study group and do not constitute proof in general. Hence, simulation studies with known characteristics and focusing on each of these aspects are urgently required. 2) We selected only a few out of many more possible metrics that can be utilized to evaluate dataset-specific trends, such as the pairwise data missingness and genetic dissimilarity (McCartney-Melstad et al., 2019), the proportion of heterozygous loci in a sample and allelic ratios at each locus (McKinney et al., 2017) or the fraction of sequence variation shared by specific proportions of all individuals (Paris et al., 2017). 3) The selected CTs for ISC and BSC are an adequate representation of a majority of loci but one CT cannot appropriately characterize the entire sequence divergence within and across samples. Various causes of sequence divergence among genomic regions (e.g., coding or non-coding regions, thus degree of sequence conservation, and biological processes such as hybridization, horizontal gene-transfer and ILS) lead to a normalization within a range of suitable CTs, which we here referred to as the "transition zone". 4) Polyploid loci composed of greater allele numbers can show greater heterozygosity than loci composed of lower number of alleles presumably containing less sequence variation across orthologous alleles (Hirsch and Buell, 2013; Karbstein et al., 2021), and thus require different CTs for accurate clustering. Hence, merging of ISC samples of varying ploidy for BSC across all taxa leads to a clustering bias. 5) The resulting data, whether used for metric evaluation or inferences of population structure or species relationships, are heavily impacted by all other parameters chosen, depend on numerous properties of the study system (e.g.: taxonomic level, genomic variation, utilized lab protocols, quality and quantity of data) and will affect downstream analysis (e.g. Huang and Knowles 2016; Eaton et al., 2017; Shafer et al., 2017; Crotti et al., 2019; McCartney-Melstad et al., 2019). 6) Metric trends can be affected by heterogeneous read quality and quantity, as well as biological factors, such as genome size or repetitive regions. This presumably leads to different metric trends of individual samples, as seen in the scatter plots for the ISC threshold selection (paragraph 3.2, Supplementary Figure S1). As a consequence, the selection of potential CTs gets less precise. This problem may be improved by re-splitting samples into groups that show similar trend intensities and using specific CTs for each group. Simulation studies focusing on potential impacts of heterogeneous sample qualities on the CT selection and the resulting assembly are required. Nevertheless, we consider a thorough evaluation of assembly metrics, as shown in this and other studies (e.g. Paris et al., 2017; Paetzold et al., 2019; Rancilhac et al., 2019; McCartney-Melstad et al., 2019; Karbstein et al., 2020; Wagner et al., 2020), to be an improvement over simply using default settings.

4.2.2. Locus filtering

The impact of filtering loci regarding specific properties, such as length, sequence variation or missingness, prior to phylogenetic inference has been investigated by numerous studies (e.g. Chou et al., 2015; Liu et al., 2015; Xi et al., 2015, 2016; Hosner et al., 2016; Mirarab et al., 2016; Huang and Knowles 2016; Sayyari et al., 2017; Molloy and Warnow, 2018). We confirm general trends previously observed regarding locus coverage and sequence variation (see 2.5.1 and 3.5.1, Table S6, Fig. S3). As the minimum requirements increased, the number of loci and sequence variation decreased (Huang and Knowles, 2016; Eaton et al., 2017). This information loss resulted in sharply decreasing BS support values of the resulting species tree estimates. This is likely a result of higher locus dropout in more rapidly evolving loci (for the "min

var" datasets). The more conserved loci are less variable but also less prone to mutation-induced cut-site disruption and thus show a higher sample coverage (for the "min samples" datasets). An interesting point is that the two datasets with the highest minimum variability required (Table S6, "min_var_200" and "min_var_300") also showed a trend toward biased locus lengths. In addition, these loci contained on average more missing data and a higher portion of variable sites was parsimony un-informative. The negative impact of this constellation of locus properties on the accuracy of species tree estimation has been demonstrated by Xi et al. (2015), Hosner et al. (2016) and Lee et al. (2018). This constellation was also evident for the length interval datasets containing the shortest and longest loci at the assembly edges (Table S6 and S7, Fig. S3 and S4). For these assembly regions, we assume that the declining sequencing quality of R2 reads led to biased sample and locus coverage, which was reflected by the prominent gap between 500 and 600nt as well as the high number of loci in the 250-300nt length range of the assembly (Table S9). This kind of data bias causes high GTEE and artificial phylogenetic conflicts among taxa and clades, which negatively affects the species tree estimation performance (Sanderson et al., 2010, 2011, 2015; Simmons, 2012; Hosner et al., 2016; Xi et al., 2016; Sayyari et al., 2017; Dobrin et al., 2018).

To reduce this effect, we first chose a controversial approach and filtered the loci based on average missingness, which resulted in the "int_251-500" dataset. Locus filtering based on missingness is generally not recommended because it can lead to a significant loss of information and thus to a performance decline of phylogenetic inference (Huang and Knowles 2016; Eaton et al., 2017; Molloy and Warnow, 2018; Crotti et al., 2019). However, it can lead to an improvement in estimation accuracy if the extent of biased, non-randomly distributed phylogenetic signal is also reduced (Xi et al., 2015, 2016; Sayyari et al., 2017; Molloy and Warnow, 2018). Although this first filtering and dataset selection resulted in a slight improvement of the data quality and the resulting BS support and concordance factor values, it did not yield the required data quality for a successful CB-SM inference. Simply choosing the average missingness as a cutoff value may improve the quality of loci containing evenly distributed phylogenetic information, but not if the bias is unevenly distributed across the assembly.

To further reduce the extent of the biased assembly area, we binned the loci based on length, inferred CA-ML and CB-SM phylogenies for each sub-dataset and put resulting phylogenetic patterns in relation to sub-dataset properties to detect biased locus length ranges (see 2.5.2 and 3.5.2, Table S7, Fig. S4). This approach turned out beneficial with regard to the selection of less biased assembly areas, suitable for CB-SM inference. The typical responses of BS support values and reconstruction of terraced branches confirmed the assembly's edge regions as particularly biased. In these locus length regions of the assembly, either the BS support values collapsed or the number of terraced branches of the resulting topology was high. Consequently, we selected the remaining, presumably less biased, assembly range of 301-450nt length served as third dataset for comparative phylogenetic inference. While this second filtering and dataset selection procedure represented a drastic reduction of overall data quantity, it also increased data quality as indicated by the average sequence variation per locus, locus coverage/missingness and sample coverage (Table 1, Table S9).

The second filtering approach used here to examine the influence of locus properties on the resulting phylogenetic reconstructions resulted in a dataset favorable for CB-SM inference. However, the process was quite tedious, and at times somewhat crude, which indicates a number of opportunities for further refinement in the future. 1) Loci of certain properties within the excluded assembly ranges are likely to be also well suited for CB-SM inference. We filtered the loci by their relative sequence variation including SNPs and PIS (see 2.5.1). However, the notable PIS/SNPs ratio along with the average locus coverage evident in the locus length filtering (Fig. S3 and S4) may be a clue to filter loci by information quality (Xi et al., 2015; Hosner et al., 2016; Lee et al., 2018). 2) The bin sizes chosen for filtering locus properties might be smaller to enable a more accurate detection of potential trend changes respecting phylogenetic outcomes. 3) We calculated only one reconstruction per inference approach for each sub-dataset. Multiple replicates may be generated to identify and statistically assess potential variations. 4) We found overall matching trends of locus properties relative to the resulting phylogenetic patterns of CA-ML and CB-SM used for bias detection. Considering the presumably strongly biased signal scattered across taxa, the relative influence of technical errors and true biological conditions (e.g. ILS) remain difficult to assess. 5) Instead of multi-locus bootstrapping (Seo, 2008), the branch support might be assessed using Local Posterior Probability, which was shown to perform more accurate on locus trees with relatively high error (Sayyari and Mirarab, 2016) or quartet based methods to identify non-informativeness (Pease et al., 2018). 6) Counting the terrace-like branches in the resulting trees helped to identify biased assembly areas but did not provide insight into the actual underlying conflicts among taxa and clades. Besides, terraced branches can also represent the true topology (Sanderson et al., 2011). To account for artificial conflicts in the data, terrace-aware phylogenetic inference tools can be used (Sanderson et al., 2011, 2015; Chernomor et al., 2016; Dobrin et al., 2018). 7) Further approaches may be tested comparatively to allow for a more accurate data quality assessment, such as filtering for fragmentary data to achieve uniform taxon coverage (Xi et al., 2016; Sayyari et al., 2017) or subsampling specific loci to establish congruence across the dataset (Chen et al., 2015; Simmons et al., 2016). For future projects, an automated pipeline that filters loci based on multiple criteria, records the properties of these bins, and evaluates the resulting phylogenetic patterns, thus simplifying the tedious filtering process, would be of great value.

4.3. Phylogenetic inference

Previous attempts at resolving phylogenetic relationships in *Aichryson* were mainly hampered by lack of variability in the employed regions (Mort et al., 2002; Fairfield et al., 2004 which failed to resolve relationships at shallow taxonomic levels (e.g., Miller et al., 2003; Abeysinghe et al., 2009; Duan et al., 2015). The application of a modified RADseq approach together with detailed data processing, analysis of filtered sub-datasets and comparative phylogenetic inference resulted in the first well-supported phylogeny for *Aichryson*. Moreover, we gained further insight into the performance of the tested inference methods with respect to underlying data properties.

4.3.1. General trends of the CA-ML and CB-SM inference during locus filtering

During locus filtering, we initially filtered the loci by variability, locus coverage and length intervals (see 2.5.1 and 3.5.1). Contrary to our expectation, we were not able to reconstruct a well-supported CB-SM phylogeny using this approach. Instead, we found that the BS support values of the three species tree sections responded differently to the underlying locus length interval datasets (Fig. 4, Table S6, Fig. S3). The related locus properties in terms of sequence variation and missingness, as well as the distribution of data across the assembly, loci, and samples (Table S9), indicated a data bias (Sanderson et al., 2010; Hosner et al., 2016; Xi et al., 2016; Sayyari et al., 2017; Lee et al., 2018; Molloy and Warnow, 2018).

Subsequently, we used phylogenetic patterns yielded by CA-ML and CB-SM inference of locus length sub-datasets to detect potentially biased assembly areas (see 2.5.2 and 3.5.2, Table S7, Fig. S4). CB-SM resolved more terraced branches than CA-ML across the tested sub-datasets, in particular when the datasets were small (Xi et al., 2016; Fig. S4, "length interval" datasets). This is likely due to the information loss inherent to the method, using only summary statistics of the inferred gene trees as input for species tree estimation (Xu and Yang, 2016). Along with this come the clearly lower resulting support values of the multi-locus bootstrapping (Seo, 2008) when applied to fragmentary data (Xi et al., 2015, 2016; Hosner et al., 2016; Sayyari et al., 2017). The overall higher

and steadily increasing BS support values with increasing dataset size confirm prior observations regarding CA-ML inference (Kubatko and Degnan, 2007; Liu et al., 2015; Minh et al. 2020a). CA-ML inference of the length sub-datasets seemed less sensitive or more robust to data bias (Xi et al., 2016; Molloy and Warnow, 2018). Still, bootstrapping over the concatenated matrix showed quite similar trends compared to the multi-locus bootstrapping employed with CB-SM.

4.3.2. Comparative phylogenetic inference of the un-/filtered datasets

The filtering steps meant a maximum reduction of 58% for the number of loci and 50% for the number of PIS, while the average sequence variation and coverage per locus raised, average missingness declined and sample coverage became more evenly distributed (Table 1, "raw" compared to "int_301-450", Table S9, "sample coverage").

For CA-ML and CB-SM, the exclusion of presumably biased assembly areas, resulted in increasing statistical support while the concordance factor value differences decreased (Table 2, Table S10, Fig. S2, S5 and S6). These trends were stronger for the CB-SM inferences. The concordance factor values and differences of the within clade branches benefited slightly while those of the clade branches benefited most from reduction. This was accompanied by improved factor values and differences of the backbone branches. We suggest that the overall higher locus coverage and the more evenly distributed information across taxa (sample coverage) of the retained assembly area caused less artificial conflicts among clades and thus favored resolution and support of the backbone section (Sanderson et al., 2010, 2011; Xi et al., 2015, 2016; Hosner et al., 2016; Sayyari et al., 2017; Dobrin et al., 2018; Molloy and Warnow, 2018; Minh et al. 2020a, b). This increasing statistical support coincides with an increase in the number of terraced branches. For instance, the CA-ML and CB-SM inferences of the "raw" dataset reconstructed a dichotomous topology for the taxa of clade 4, but there was insufficient statistical support for the backbone sections (Fig. S2). The backbone topology of the strongly reduced "int_301-450" dataset was well supported, but in exchange the taxa of clade 4 were reconstructed on terraced branches (Fig. 5 and S6).

Phylogenetic inference of the datasets using SVD showed some contradictions. The lower factor values of the backbone branches for the alternative topologies and compared to the CA-ML and CB-SM inferences (Fig. S2, S5 and S6), increasing concordance factor value differences with increasing extent of reduction (Table 2), as well as the consistent maximum BS support values, suggest a random resolution due to limited and unevenly distributed information (Long and Kubatko, 2018; Minh et al. 2020a, b). This is certainly in part due to the selection of individual PICs per locus, which we performed to meet the methods requirements in terms of linkage (Bryant et al., 2012; Chiffman and Kubatko 2014; Xu and Yang, 2016). In addition, studies comparing the performance of inference methods under challenging data conditions showed that SVD is often less accurate than CA-ML and CB-SM (Chou et al., 2015; Molloy and Warnow, 2018). Still, the SVD inferences illustrated potentially conflicting topological alternatives.

In summary, phylogenetic inference of the three datasets ("raw", "int_251-500", and "int_301-450") showed positive trends in terms of the resulting BS support values and concordance factor values with increasing degree of dataset reduction for CA-ML and CB-SM. The resulting SVD reconstructions, however, appeared to be impeded by information limitation and data bias.

4.3.3. Phylogenetic inference of the truncated locus datasets

In general, increasing locus length is associated with increasing phylogenetic information, lower GTEE and thus an increased accuracy of species tree estimation (e.g. Mirarab et al., 2014, 2016; Xi et al., 2015; Chou et al., 2015; Hosner et al., 2016; Xu and Yang, 2016; Blom et al., 2017; Molloy and Warnow, 2018). We expected a decrease in locus length to decrease the total and average phylogenetic information per locus, and consequently to negatively affect performance. To test this, the "raw" assembly loci were truncated and used as input for CA-ML (Supplementary Figure S7 A and B) and CB-SM inference (Supplementary Figure S7 C and D).

The truncated datasets showed a 2/3 reduction in phylogenetic information (Table 1, "int_251-500_short" and "int_301-450_short"), resulted incongruently resolved tree topologies (Fig. S7), and yielded decreased estimated BS support and concordance factor values, while the factor value differences of the clade and backbone branches increased strongly compared to the original datasets (Table S10). Therefore, we conclude that the locus length reduction had a substantially negative impact on the phylogenetic inference. This is in line with findings by studies comparing the inference performance over varying locus lengths and information contents (e.g. Mirarab et al., 2014a, b, 2016; Xi et al., 2015; Chou et al., 2015; Xu and Yang, 2016; Molloy and Warnow, 2018).

However, we performed a drastic locus length reduction by 2/3, which resulted in an average locus length of 120/123nt (Table 1). As we found during locus filtering (see 2.5) and phylogenetic inference of the resulting datasets, an average locus length of 373nt (±43nt) in an assembly range of 300-450nt yielded sufficient phylogenetic information per locus and in total for successful CB-SM inference. Other empirical studies using similar or even shorter length ranges also achieved a successful CB-SM inference of the assembled data (e.g. Curto et al., 2018; Rancilhac et al., 2019). Based on our results, and as found by numerous studies (e.g., Gatesy and Springer, 2014; Lanier et al., 2014; Liu et al., 2015; Xi et al., 2015; Hosner et al., 2016; Huang and Knowles, 2016; Blom et al., 2017; Sayyari et al., 2017; Xu and Yang, 2016; Lee et al., 2018), we suggest that locus quality in terms of the information content and its distribution across the assembly and taxa is of greater importance than mere locus length. Yet, this also strongly depends on the taxonomic level, i.e. sequence divergence, of the study group.

4.3.4. On the accuracy of the Aichryson phylogeny

The accuracy of the phylogenetic outcome is the suggested by the emerging congruence of the CA-ML and CB-SM reconstructions with increasing data quality. Inference of the "int_301-450" dataset yielded overall congruent, similarly well-supported topologies as well as similar concordance factor values and differences. In addition, the phylogenetic pattern matches the species distributions. For instance, the species occurring on Madeira (*A. divaricatum*, *A. dumosum*, *A. villosum*) and the two *A. tortuosum* subspecies occurring on the eastern Canary Islands, Lanzarote (subsp. *tortuosum*) and Fuerteventura (subsp. *bethencourtianum*), each form a monophyletic group. The polyphyletic status of the *A. pachycaulon* subspecies is also consistent with previous studies (Mort et al., 2002; Fairfield et al., 2004).

However, as Goethe put it: „We know accurately only when we know little; with knowledge, doubt increases" (von Goethe, 2012, published postum). 1) *Aichryson* is not a model group and lacks comparable studies in terms of data properties (locus length, sequence variation, missingness), data analysis (data assembly, locus filtering) and phylogenetic inference. 2) We did not statistically assess potential variation in phylogenetic inference of the filtered datasets using multiple replicates. 3) The extent to which phylogenetic inference may be impacted by terraces due to artificial conflicts among clades arising from the data structure herein is unclear (Sanderson et al., 2010, 2011, 2015; Simmons, 2012; Dobrin et al., 2018). 4) Although locus properties gained quality and sample coverage became more even, the low concordance factor values of some backbone branches representing the relationships of clades 2 + 3 + 4 to clade 5 and high concordance factor value differences of the within clade branches of clade 5 suggest a strong conflict among clades and taxa, respectively (Minh et al. 2020a, b). However, we cannot assess whether this incongruence of information among locus trees is a true biological signal due to reticulate evolution or an artifact of the data structure. 5) In addition, the ongoing, sometimes heated debate over the most accurate application, analysis, and inference of a variety of RRL/SRS-based approaches, along with a series of comparisons of divergent concepts and opinions, further complicate the

interpretation of the results (e.g. de Queiroz and Gatesy 2007; Edwards et al., 2007, 2016; Kubatko and Degnan, 2007; Degnan and Rosenberg, 2009; Knowles, 2009; Leaché and Rannala, 2011; Song et al., 2012; Gatesy and Springer, 2013, 2014; Springer and Gatesy 2014, 2016, 2018; Mirarab et al., 2014b, 2016; Chou et al., 2015; Roch and Steel 2015; Mirarab and Warnow 2015; Solís-Lemus et al., 2016; Mendes and Hahn, 2018; Molloy and Warnow, 2018; Bryant and Hahn, 2020; Rannala et al., 2020). In particular, the inference accuracy of CA-ML in the presence of gene tree-species tree discordance (Degnan et al., 2006, Degnan and Rosenberg, 2009; Kubatko and Degnan, 2007; Knowles, 2009; Roch and Steel, 2015; Solís-Lemus et al., 2016; Mendes and Hahn, 2018; Bryant and Hahn, 2020) and the performance of CB-SM under conditions of GTEE (Springer and Gatesy, 2014, 2016; Roch and Warnow, 2015; Xi et al., 2015, 2016; Solís-Lemus et al., 2016; Xu and Yang, 2016; Sayyari et al., 2017; Molloy and Warnow, 2018) raise concerns.

In general, CA-ML and CB-SM are expected to yield congruent results under less challenging conditions of gene tree-species tree discordance (Edwards et al., 2007; Kubatko and Degnan, 2007; Leaché and Rannala, 2011). Comparative studies showed that CA-ML and CB-SM performed equally under various levels of ILS, with CA-ML performing more accurate under challenging GTEE conditions (Chou et al., 2015; Xi et al., 2015, 2016; Mirarab et al., 2016; Sayyari et al., 2017; Molloy and Warnow, 2018). Moreover, inference of empirical data using both approaches generally yielded congruent results (e.g. Chiari et al., 2012; Hosner et al., 2016; Blom et al., 2017; Sayyari et al., 2017; Curto et al., 2018; Rancilhac et al., 2019). The bottom line is that we cannot ultimately assess the accuracy of the species tree for *Aichryson*, still, we construe the overall congruence as supporting the accuracy of the phylogenetic outcome.

4.4. Conclusion

The methodology presented in this study successfully led to a coalescent-based inference of our focal group *Aichryson*. For some, however, the series of approaches tested by us may be equivalent to a butcher making "phylogenetic sausage" (for the definition of a "phylogenetic sausage" see: Gatesy and Springer, 2014; see further: Springer and Gatesy, 2016, 2018; Bryant and Hahn, 2020; Fernández et al., 2020; Rannala et al., 2020). Admittedly, all methodological components could be modified and improved in many ways. The resulting data were also quite demanding to analyze. Still, particularly the challenging data structure provided the opportunity to gain further valuable insights to drive the development of fast and reliable RRL-SRS approaches. 1) Minor modifications of the RADseq toolbox regarding fragment size selection and sequencing range yielded a strongly reduced locus set of extended length. 2) Evaluation of a few metrics enabled an informed selection of clustering thresholds for data assembly within and across samples. 3) Simple descriptive statistics of the resulting assembly were useful for an initial assessment of the data structure. 4) Locus filtering greatly assisted to identify assembly areas of presumably biased locus and taxon coverage. 5) Comparative evaluation of phylogenetic patterns, such as terrace-like branches, BS support values and concordance factor values highlighted the importance of data quality over mere quantity, in particular for the coalescent-based summary method.

We are convinced that the combination of highly flexible RRL-SRS laboratory, data analysis, and inference approaches is crucial for a fast and reliable biodiversity exploration. Hence, we highly encourage the community to: 1) modify the extensive RADseq toolbox regarding an extended fragment length and sequencing range, 2) reduce the data quantity in favor of data quality, 3) utilize approaches guiding an informed threshold selection for accurate clustering, 4) thoroughly analyze and test the resulting assembly and locus properties for potential biases, 5) and to compare and evaluate the resulting phylogenetic trends using multiple inference approaches.

Funding

This work benefitted from the sharing of expertise on RADseq data in various groups of organisms at various taxonomic levels within the Deutsche Forschungsgemeinschaft DFG priority program SPP 1991 TaxonOMICS and from financial support from DFG KA 1816/11–1.

CRediT authorship contribution statement

Philipp Hühn: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization, Project administration, Supervision. **Markus S. Dillenberger:** Methodology, Software, Formal analysis, Writing – review & editing. **Michael Gerschwitz-Eidt:** Methodology, Software, Formal analysis, Writing – review & editing. **Elvira Hörandl:** Resources, Writing – review & editing. **Jessica A. Los:** Investigation, Resources. **Thibaud F.E. Messerschmid:** Investigation, Resources. **Claudia Paetzold:** Methodology, Writing – review & editing. **Benjamin Rieger:** Methodology, Software. **Gudrun Kadereit:** Conceptualization, Resources, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

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

Acknowledgements

We thank the following people and institutions for providing materials for this study: Ángel Bañares Baudet (Tenerife), Stephan Scholz (Lanzarote), Stefan Abrahamczyk (Bonn) and Nadine Bobon (Mainz). We thank Hans Zischler (Mainz) and Dirk Albach (Oldenburg) for providing access to their lab facilities. We thank Ursula Martiné (Mainz) and Silvia Wienken (Munich) for lab assistance. We thank Oliver Hawlitschek and three further reviewers for helpful comments on the manuscript. We are grateful to Doris Franke and Maria Geyer (Mainz) for their assistance with the figure design, and to Christopher Wild (Mainz) for taking care of the living collection of Crassulaceae at the Botanical Garden Mainz.

This work used computing infrastructure of: -the Scientific Compute Cluster at the Göttingen Society for Scientific Data Processing (GWDG), as part of the joint data center of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen, -the supercomputer Mogon at Johannes Gutenberg University Mainz, which is a member of the Alliance for High Performance Computing in Rhineland Palatinate (AHRP) and the Gauss Alliance e.V., and -the Center for Genome Research and Biocomputing at the Oregon State University. We gratefully acknowledge the computing time granted. This research was funded by the German Science Foundation (DFG grant KA1816/11-1) within the priority programme 1991 Taxon-OMICS.

Sequence data

Demultiplexed raw data is available at the NCBI Sequence Read Archive (www.ncbi.nlm.nih.gov/sra/PRJNA642981), BioProject PRJNA642981 (www.ncbi.nlm.nih.gov/bioproject/PRJNA642981).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2021.107342.

References

Abeysinghe, P.D., Wijesinghe, K.G.G., Tachida, H., Yoshda, T., 2009. Molecular characterization of Cinnamon (*Cinnamomum verum* Presl) accessions and evaluation of genetic relatedness of Cinnamon species in Sri Lanka based on *trn*L intron region, intergenic spacers between *trn*T-*trn*L, *trn*L-*trn*F, *trn*H-*psb*A and nuclear ITS. Res. J. Agric. Biol. Sci. 5 (6), 1079–1088.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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. Nat. Rev. Genet. 17 (2), 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., Fay, J.C., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3 (10), e3376. https://doi.org/10.1371/journal.pone.0003376.

Bañares Baudet, Á., 2002. On some poorly known taxa of *Aichryson* sect. *Aichryson* and *A. bituminosum* sp. nova (Crassulaceae). Willdenowia 32 (2), 221–230. https://doi.org/10.3372/wi.32.32204.

Bañares Baudet, Á., 2015. Las plantas suculentas (Crassulaceae) endémicas de las Islas Canarias. Publicaciones Turquesa, Santa Cruz de Tenerife.

Bañares Baudet, Á., 2015b. Híbridos de la familia Crassulaceae en las islas Canarias. V. Vieraea 43, 189–206.

Bañares Baudet, Á., 2017. Typification of Aichryson pachycaulon subsp. praetermissum and description of *A. roseum* sp. nov. (Crassulaceae) from Gran Canaria, Canary Islands, Spain. Willdenowia 47 (2), 127–134. https://doi.org/10.3372/wi.47.47204.

Bayona-Vásquez, N.J., Glenn, T.C., Kieran, T.J., Pierson, T.W., Hoffberg, S.L., Scott, P.A., Bentley, K.E., Finger, J.W., Louha, S., Troendle, N. and Díaz-Jaimes, P., Mauricio, R., Faircloth, B.C., 2019. Adapterama III: Quadruple-indexed, double/triple-enzyme RADseq libraries (2RAD/3RAD). PeerJ 7:e7724. doi: 10.7717/peerj.7724.

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

Blom, M.P.K., Bragg, J.G., Potter, S., Moritz, C., 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst. Biol. 66, 352–366. https://doi.org/10.1093/sysbio/syw089.

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M.A., Rambaut, A., Drummond, A.J., Prlic, A., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comp. Biol. 10 (4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537.

Bryant, D., Hahn, M.W., 2020. The Concatenation Question. In: Scornavacca, C., Delsuc, F., Galtier, N., editors, Phylogenetics in the Genomic Era, chapter No. 3.4, pp. 3.4:1–3.4:23. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE. HAL Id: hal-02535651.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., RoyChoudhury, A., 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molec. Biol. Evol. 29 (8), 1917–1932. https://doi.org/10.1093/molbev/mss086.

Buono, D., Khan, G., von Hagen, K.B., Kosachev, P.A., Mayland-Quellhorst, E., Mosyakin, S.L., Albach, D.C., 2021. Comparative Phylogeography of Veronica spicata and V. longifolia (Plantaginaceae) Across Europe: Integrating Hybridization and Polyploidy in Phylogeography. Front. Plant. Sci. 11 https://doi.org/10.3389/fpls.2020.588354.

Burleigh, J.G., Kimball, R.T., Braun, E.L., 2015. Building the avian tree of life using a large-scale, sparse supermatrix. Mol. Phylogenet. Evol. 84, 53–63. https://doi.org/10.1016/j.ympev.2014.12.003.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinf. 10, 421. https://doi.org/10.1186/1471-2105-10-421.

Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., Cresko, W.A., 2013. Stacks: an analysis tool set for population genomics. Molec. Ecol. 22 (11), 3124–3140. https://doi.org/10.1111/mec.12354.

Chen, M.Y., Liang, D., Zhang, P., 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64 (6), 1104–1120. https://doi.org/10.1093/sysbio/syv059.

Chernomor, O., Von Haeseler, A., Minh, B.Q., 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65 (6), 997–1008. https://doi.org/10.1093/sysbio/syw037.

Chiari, Y., Cahais, V., Galtier, N., Delsuc, F., 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10 (1), 1–15. https://doi.org/10.1186/1741-7007-10-65.

Chifman, J., Kubatko, L., 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30 (23), 3317–3324. https://doi.org/10.1093/bioinformatics/btu530.

Chou, J., Gupta, A., Yaduvanshi, S., Davidson, R., Nute, M., Mirarab, S., Warnow, T., 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genomics 16, S2. https://doi.org/10.1186/1471-2164-16-S10-S2.

Crotti, M., Barratt, C.D., Loader, S.P., Gower, D.J., Streicher, J.W., 2019. Causes and analytical impacts of missing data in RADseq phylogenetics: insights from an African frog (*Afrixalus*). Zool. Scripta 48 (2), 157–167. https://doi.org/10.1111/zsc.2019.48.issue-210.1111/zsc.12335.

Curto, M., Schachtler, C., Puppo, P., Meimberg, H., 2018. Using a new RAD-sequencing approach to study the evolution of *Micromeria* in the Canary islands. Molec. Phylogen. Evol. 119, 160–169. https://doi.org/10.1016/j.ympev.2017.11.005.

de Oca, A.N.M., Barley, A.J., Meza-Lázaro, R.N., García-Vázquez, U.O., Zamora-Abrego, J.G., Thomson, R.C., Leaché, A.D., 2017. Phylogenomics and species delimitation in the knob-scaled lizards of the genus Xenosaurus (Squamata: Xenosauridae) using ddRADseq data reveal a substantial underestimation of diversity. Molec. Phylogen. Evol. 106, 241–253. https://doi.org/10.1016/j.ympev.2016.09.001.

de Queiroz, A., Donoghue, M.J., Kim, J., 1995. Separate versus combined analysis of phylogenetic evidence. Stable URL: Annu. Rev. Ecol. Syst. 26 (1), 657–681 https://www.jstor.org/stable/2097223.

de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol 22 (1), 34–41. https://doi.org/10.1016/j.tree.2006.10.002.

Degnan, J.H., Rosenberg, N.A., Wakeley, J., 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2 (5), e68. https://doi.org/10.1371/journal.pgen.0020068.

Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24 (6), 332–340. https://doi.org/10.1016/j.tree.2009.01.009.

Dillenberger, M.S., Kadereit, J.W., 2017. Simultaneous speciation in the European high mountain flowering plant genus *Facchinia (Minuartia* s.l., Caryophyllaceae) revealed by genotyping-by-sequencing. Molec. Phylogen. Evol. 112, 23–35. https://doi.org/10.1016/j.ympev.2017.04.016.

Dobrin, B.H., Zwickl, D.J., Sanderson, M.J., 2018. The prevalence of terraced treescapes in analyses of phylogenetic data sets. BMC Evol. Biol. 18 (1), 1–16. https://doi.org/10.1186/s12862-018-1162-9.

Duan, L., Wen, J., Yang, X., Liu, P.L., Arslan, E., Ertuğrul, K., Chang, Z.Y., 2015. Phylogeny of *Hedysarum* and tribe Hedysareae (Leguminosae: Papilionoideae) inferred from sequence data of ITS, *matK*, *trnL-F* and *psbA-trnH*. Taxon 64 (1), 49–64. https://doi.org/10.12705/641.26.

Eaton, D.A., Overcast, I., 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics, 36(8): 2592-2594. doi: 10.1093/bioinformatics/btz966.

Eaton, D.A.R., Ree, R.H., 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). Syst. Biol. 62 (5), 689–706. https://doi.org/10.1093/sysbio/syt032.

Eaton, D.A.R., Spriggs, E.L., Park, B., Donoghue, M.J., 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Syst. Biol. 66 (3), 399–412. https://doi.org/10.1093/sysbio/syw092.

Edwards, S.V., Liu, L., Pearl, D.K., 2007. High-resolution species trees without concatenation. P. Nati. A. Sci. USA 104 (14), 5936–5941. https://doi.org/10.1073/pnas.0607004104.

Edwards, S.V., Xi, Z., Janke, A., Faircloth, B.C., McCormack, J.E., Glenn, T.C., Zhong, B., Wu, S., Lemmon, E.M., Lemmon, A.R., Leaché, A.D., Liu, L., Davis, C.C., 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94, 447–462. https://doi.org/10.1016/j.ympev.2015.10.027.

Eggli, U., 2008. Sukkulenten, 2nd Edition. Eugen Ulmer KG, Stuttgart.

Elshire, R.J., Glaubitz, J.C., Sun, Q.i., Poland, J.A., Kawamoto, K., Buckler, E.S., Mitchell, S.E., Orban, L., 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6 (5), e19379. https://doi.org/10.1371/journal.pone.0019379.

Escudero, M., Eaton, D.A.R., Hahn, M., Hipp, A.L., 2014. Genotyping-by-sequencing as a tool to infer phylogeny and ancestral hybridization: a case study in *Carex* (Cyperaceae). Molec. Phylogen. Evol. 79, 359–367. https://doi.org/10.1016/j.ympev.2014.06.026.

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32 (19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

Fairfield, K.N., Mort, M.E., Santos-Guerra, A., 2004. Phylogenetics and evolution of the Macaronesian members of the genus *Aichryson* (Crassulaceae) inferred from nuclear and chloroplast sequence data. Pl. Syst. Evol. 248, 71–83. https://doi.org/10.1007/s00606-004-0190-7.

Fernández, R., Gabaldón, T., Dessimoz, C., 2020. Orthology: definitions, inference, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, Phylogenetics in the Genomic Era, chapter No. 2.4, pp. 2.4:1–2.4:14. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE. HAL Id: hal-02535414.

Gatesy, J., Springer, M.S., 2013. Concatenation versus coalescence versus "concatalescence". P. Natl. Acad. Sci. USA 110 (13). https://doi.org/10.1073/pnas.1221121110.

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

Gerschwitz-Eidt, M.A., Kadereit, J.W., 2019. Genotyping-by-sequencing (GBS), ITS and cpDNA phylogenies reveal the existence of a distinct Pyrenean/Cantabrian lineage in the European high mountain genus *Homogyne* (Asteraceae) and imply dual westward migration of the genus. Alp. Botany 129 (1), 21–31. https://doi.org/10.1007/s00035-018-0212-7.

Good, J.M., 2012. Reduced representation methods for subgenomic enrichment and next-generation sequencing. In: Orgogozo, V., Rockman, M.V. (Eds.), Methods in Molecular Biology Vol. 772: Molecular Methods for Evolutionary Genetics. Humana Press, New York, pp. 85–103. https://doi.org/10.1007/978-1-61779-228-1_5.

Grover, C.E., Salmon, A., Wendel, J.F., 2012. Targeted sequence capture as a powerful tool for evolutionary analysis1. Am. J. Botany. 99 (2), 312–319. https://doi.org/10.3732/ajb.1100323.

Hamon, P., Grover, C.E., Davis, A.P., Rakotomalala, J.-J., Raharimalala, N.E., Albert, V. A., Sreenath, H.L., Stoffelen, P., Mitchell, S.E., Couturon, E., Hamon, S., de

Kochko, A., Crouzillat, D., Rigoreau, M., Sumirat, U., Akaffou, S., Guyot, R., 2017. Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. Molec. Phylogen. Evol. 109, 351–361. https://doi.org/10.1016/j.ympev.2017.02.009.

Harvey, M.G., Judy, C.D., Seeholzer, G.F., Maley, J.M., Graves, G.R., Brumfield, R.T., 2015. Similarity thresholds used in DNA sequence assembly from short reads can reduce the comparability of population histories across species. PeerJ 3:e895. doi: 10.7717/peerj.895.

Harvey, M.G., Smith, B.T., Glenn, T.C., Faircloth, B.C., Brumfield, R.T., 2016. Sequence capture versus restriction site associated DNA sequencing for shallow systematics. Syst. Biol. 65 (5), 910–924. https://doi.org/10.1093/sysbio/syw036.

Heled, J., Drummond, A.J., 2009. Bayesian inference of species trees from multilocus data. Molec. Bio. Evol. 27 (3), 570–580. https://doi.org/10.1093/molbev/msp274.

Herrera, S., Shank, T.M., 2016. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. Molec. Phylogen. Evol. 100, 70–79. https://doi.org/10.1016/j.ympev.2016.03.010.

Hipp, A.L., Manos, P.S., Hahn, M., Avishai, M., Bodénès, C., Cavender-Bares, J., Crowl, A. A., Deng, M., Denk, T., Fitz-Gibbon, S., Gailing, O., González-Elizondo, M.S., González-Rodríguez, A., Grimm, G.W., Jiang, X.-L., Kremer, A., Lesur, I., McVay, J. D., Plomion, C., Rodríguez-Correa, H., Schulze, E.-D., Simeone, M.C., Sork, V.L., Valencia-Avalos, S., 2020. Genomic landscape of the global oak phylogeny. New Phytol. 226 (4), 1198–1212. https://doi.org/10.1111/nph.v226.410.1111/nph.16162.

Hirsch, C.N., Robin Buell, C., 2013. Tapping the promise of genomics in species with complex, nonmodel genomes. Annual Rev. Pl. Biol. 64 (1), 89–110. https://doi.org/10.1146/arplant.2013.64.issue-110.1146/annurev-arplant-050312-120237.

Hosner, P.A., Faircloth, B.C., Glenn, T.C., Braun, E.L., Kimball, R.T., 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33 (4), 1110–1125. https://doi.org/10.1093/molbev/msv347.

Huang, H., Knowles, L.L., 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. 65 (3), 357–365. https://doi.org/10.1093/sysbio/syu046.

Ilut, D.C., Nydam, M.L., Hare, M.P., 2014. Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. BioMed Res. Int. 2014, 1–9. https://doi.org/10.1155/2014/675158.

Karbstein, K., Tomasello, S., Hodač, L., Dunkel, F.G., Daubert, M., Hörandl, E., 2020. Phylogenomics supported by geometric morphometrics reveals delimitation of sexual species within the polyploid apomictic Ranunculus auricomus complex (Ranunculaceae). Taxon 69 (6), 1191–1220. https://doi.org/10.1002/tax.v69.610.1002/tax.12365.

Karbstein, K., Tomasello, S., Hodač, L., Lorberg, E., Daubert, M., Hörandl, E., 2021. Moving beyond assumptions: Polyploidy and environmental effects explain a geographical parthenogenesis scenario in European plants. Mol. Ecol. 30 (11), 2659–2675. https://doi.org/10.1111/mec.v30.1110.1111/mec.15919.

Knowles, L.L., 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. Syst. Biol. 58 (5), 463–467. https://doi.org/10.1093/sysbio/syp061.

Kubatko, L.S., Degnan J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56(1):17–24. doi: 10.1080/10635150601146041.

Kumar, S., Filipski, A.J., Battistuzzi, F.U., Kosakovsky Pond, S.L., Tamura, K., 2012. Statistics and truth in phylogenomics. Molec. Biol. Evol. 29 (2), 457–472. https://doi.org/10.1093/molbev/msr202.

Kück, P., Meusemann, K., 2010. FASconCAT: convenient handling of data matrices. Molec. Phylogenet. Evol. 56 (3), 1115–1118. https://doi.org/10.1016/j.ympev.2010.04.024.

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

Lanier, H.C., Huang, H., Knowles, L.L., 2014. How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. Mol. Phylogenet. Evol. 70, 112–119. https://doi.org/10.1016/j.ympev.2013.09.006.

Leaché, A.D., Rannala, B., 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60 (2), 126–137. https://doi.org/10.1093/sysbio/syq073.

Lee, K.M., Kivelä, S.M., Ivanov, V., Hausmann, A., Kaila, L., Wahlberg, N., Mutanen, M., 2018. Information dropout patterns in restriction site associated DNA phylogenomics and a comparison with multilocus Sanger data in a species-rich moth genus. Syst. Biol. 67(6):925–939. doi: 10.1093/sysbio/syy029.

Lepais, O., Weir, J.T., 2014. Sim RAD: an R package for simulation-based prediction of the number of loci expected in RAD seq and similar genotyping by sequencing approaches. Mol. Ecol. Resour. 14 (6), 1314–1321. https://doi.org/10.1111/men.2014.14.issue-610.1111/1755-0998.12273.

Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24 (21), 2542–2543. https://doi.org/10.1093/bioinformatics/btn484.

Liu, L., Xi, Z., Wu, S., Davis, C., Edwards, S.V., 2015. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. 1360, 36–53. https://doi.org/10.1111/nyas.12747.

Long, C., Kubatko, L., 2018. The effect of gene flow on coalescent-based species-tree inference. Syst. Biol. 67(5):770–785. doi: 10.1093/sysbio/syy020.

MacConaill, L.E., Burns, R.T., Nag, A., Coleman, H.A., Slevin, M.K., Giorda, K., Light, M., Lai, K., Jarosz, M., McNeill, M.S., Ducar, M.D., Meyerson, M., Thorner, A.R., 2018. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics 19 (1). https://doi.org/10.1186/s12864-017-4428-5.

Maddison, W. P., 1997. Gene trees in species trees. Syst, Biol. 46(3): 523-536. doi: 10.1093/sysbio/46.3.523.

Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55(1):21-30. doi: 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., 2010. Target-enrichment strategies for next-generation sequencing. Nat. Methods. 7 (2), 111–118. https://doi.org/10.1038/nmeth.1419.

Mardis, E.R., 2017. DNA sequencing technologies: 2006–2016. Nat. Protoc. 12 (2), 213–218. https://doi.org/10.1038/nprot.2016.182.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17 (1), 10. https://doi.org/10.14806/ej.17.110.14806/ej.17.1.200.

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T.H., Piñero, D., Emerson, B.C., 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Mol. Ecol. Resour. 15 (1), 28–41. https://doi.org/10.1111/men.2014.15.issue-110.1111/1755-0998.12291.

McCartney-Melstad, E., Gidiş, M., Shaffer, H.B., 2019. An empirical pipeline for choosing the optimal clustering threshold in RADseq studies. Molec. Ecol. Resour. 19 (5), 1195–1204. https://doi.org/10.5281/zenodo.2540263.

McCormack, J.E., Harvey, M.G., Faircloth, B.C., Crawford, N.G., Glenn, T.C., Brumfield, R.T., Alvarez, N., 2013a. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS ONE 8 (1), e54848. https://doi.org/10.1371/journal.pone.0054848.

McCormack, J.E., Hird, S.M., Zellmer, A.J., Carstens, B.C., Brumfield, R.T., 2013b. Applications of next-generation sequencing to phylogeography and phylogenetics. Molec. Phylogen. Evol. 66 (2), 526–538. https://doi.org/10.1016/j.ympev.2011.12.007.

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

McKinney, G.J., Waples, R.K., Seeb, L.W., Seeb, J.E., 2017. Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. Mol. Ecol. Resour. 17 (4), 656–669. https://doi.org/10.1111/men.2017.17.issue-410.1111/1755-0998.12613.

Mendes, F.K., Hahn, M.W., 2018. Why concatenation fails near the anomaly zone. Syst. Biol. 67 (1), 158–169. https://doi.org/10.1093/sysbio/syx063.

Messerschmid, T.F.E., Klein, J.T., Kadereit, G., Kadereit, J.W., 2020. Linnaeus' folly – phylogeny, evolution and classification of Sedum (Crassulaceae) and Crassulaceae subfamily Sempervivoideae. Taxon 69 (5), 892–926. https://doi.org/10.1002/tax.12316.

Miller, J.T., Grimes, J.W., Murphy, D.J., Bayer, R.J., Ladiges, P.Y., 2003. A phylogenetic analysis of the Acacieae and Ingeae (Mimosoideae: Fabaceae) based on *trnK*, *matK*, *psbA-trnH*, and *trnL/trnF* sequence data. Syst. Bot. 28 (3), 558–566. https://doi.org/10.1043/02-48.1.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., Johnson, E.A., 2007. Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17 (2), 240–248.

Minh, B.Q., Hahn, M.W., Lanfear, R., 2020a. New methods to calculate concordance factors for phylogenomic datasets. Mol. Biol. Evol. 37(9): 2727-2733. doi: 10.1093/molbev/msaa106.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R., 2020b. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37(5): 1530-1534. doi: 10.1093/molbev/msaa015.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014a. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30 (17), i541–i548. https://doi.org/10.1093/bioinformatics/btu462.

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

Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31 (12), i44–i52. https://doi.org/10.1093/bioinformatics/btv234.

Mirarab, S., Bayzid, M.S., Warnow, T., 2016. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65 (3), 366–380. https://doi.org/10.1093/sysbio/syu063.

Molloy, E.K., Warnow, T., 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67 (2), 285–303. https://doi.org/10.1093/sysbio/syx077.

Mora-Márquez, F., García-Olivares, V., Emerson, B.C., López de Heredia, U., 2017. ddradseqtools: a software package for in silico simulation and testing of double-digest RAD seq experiments. Mol. Ecol. Resour. 17 (2), 230–246. https://doi.org/10.1111/1755-0998.12550.

Mort, M.E., Soltis, D.E., Soltis, P.S., Francisco-Ortega, J., Santos-Guerra, A., 2002. Phylogenetics and evolution of the Macaronesian clade of Crassulaceae inferred from nuclear and chloroplast sequence data. Syst. Bot. 27 (2), 271–288. https://doi.org/10.1043/0363-6445-27.2.271.

Moura, M., Carine, M., De Sequeira, M.M., 2015. *Aichryson santamariensis* (Crassulaceae): a new species endemic to Santa Maria in the Azores. Phytotaxa 234 (1), 37–50. https://doi.org/10.11646/phytotaxa.234.1.2.

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). Front. Plant Sci. 6, 710. https://doi.org/10.3389/fpls.2015.00710.

Paetzold, C., Wood, K.R., Eaton, D.A.R., Wagner, W.L., Appelhans, M.S., 2019. Phylogeny of Hawaiian Melicope (Rutaceae): RAD-seq resolves species relationships and reveals ancient introgression. Front. Plant. Sci. 10 https://doi.org/10.3389/fpls.2019.01074.

Parchman, T.L., Jahner, J.P., Uckele, K.A., Galland, L.M., Eckert, A.J., 2018. RADseq approaches and applications for forest tree genetics. Tree Genet. Genomes. 14 (3), 39. https://doi.org/10.1007/s11295-018-1251-3.

Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5 (5), 568–583. https://doi.org/10.1093/oxfordjournals.molbev.a040517.

Paris, J.R., Stevens, J.R., Catchen, J.M., 2017. Lost in parameter space: a road map for stacks. Methods Ecol. Evol. 8 (10), 1360–1373. https://doi.org/10.1111/2041-210X.12775.

Pease, J.B., Brown, J.W., Walker, J.F., Hinchliff, C.E., Smith, S.A., 2018. Quartet Sampling distinguishes lack of support from conflicting support in the green plant tree of life. Am. J. Bot. 105 (3), 385–403. https://doi.org/10.1002/ajb2.2018.105.issue-310.1002/ajb2.1016.

Peterson, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., Hoekstra, H.E., Orlando, L., 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7 (5), e37135. https://doi.org/10.1371/journal.pone.0037135.

Puritz, J.B., Hollenbeck, C.M., Gold, J.R., 2014. dDocent: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2:e431. doi: 10.7717/peerj.431.

Rancilhac, L., Goudarzi, F., Gehara, M., Hemami, M.R., Elmer, K.R., Vences, M., Steinfarz, S., 2019. Phylogeny and species delimitation of near Eastern Neurergus newts (Salamandridae) based on genome-wide RADseq data analysis. Mol. Phylogenet. Evol. 133, 189–197. https://doi.org/10.1016/j.ympev.2019.01.003.

Rannala, B., Edwards, S.V., Leaché, A., Yang, Z., 2020. The Multispecies Coalescent Model and Species Tree Inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, Phylogenetics in the Genomic Era, chapter No. 3.3, pp. 3.3:1–3.3:21. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE. HAL Id: hal-02535622.

Razkin, O., Sonet, G., Breugelmans, K., Madeira, M.J., Gómez-Moliner, B.J., Backeljau, T., 2016. Species limits, interspecific hybridization and phylogeny in the cryptic land snail complex Pyramidula: the power of RADseq data. Mol. Phylogenet. Evol. 101, 267–278. https://doi.org/10.1016/j.ympev.2016.05.002.

Ree, R.H., Hipp, A.L., 2015. Inferring phylogenetic history from restriction site associated DNA (RADseq). In: Hörandl, E., Appelhans, M.S. (Eds.), Regnum Vegetabile Vol. 158: Next-Generation Sequencing in Plant Systematics. Koeltz Scientific Books, Oberreifenberg, pp. 181–204.

Reuter, J.A., Spacek, D.V., Snyder, M.P., 2015. High-throughput sequencing technologies. Molec. Cell 58 (4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004.

Rivera-Colón, A.G., Rochette, N.C., Catchen, J.M., 2021. Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data. Mol. Ecol. Resour. 21 (2), 363–378. https://doi.org/10.1111/men.v21.210.1111/1755-0998.13163.

Roch, S., Steel, M., 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Populat. Biol. 100, 56–62. https://doi.org/10.1016/j.tpb.2014.12.005.

Roch, S., Warnow, T., 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64 (4), 663–676. https://doi.org/10.1093/sysbio/syv016.

Rubin, B.E.R., Ree, R.H., Moreau, C.S., Kolokotronis, S.-O., 2012. Inferring phylogenies from RAD sequence data. PLoS ONE 7 (4), e33394. https://doi.org/10.1371/journal.pone.0033394.

Sanderson, M.J., McMahon, M.M., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10 (1), 1–13. https://doi.org/10.1186/1471-2148-10-155.

Sanderson, M.J., McMahon, M.M., Steel, M., 2011. Terraces in phylogenetic tree space. Science 333 (6041), 448–450. https://doi.org/10.1126/science.1206357.

Sanderson, M.J., McMahon, M.M., Stamatakis, A., Zwickl, D.J., Steel, M., 2015. Impacts of terraces on phylogenetic inference. Syst. Biol. 64 (5), 709–726. https://doi.org/10.1093/sysbio/syv024.

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Molec. Biol. Evol. 33 (7), 1654–1668. https://doi.org/10.1093/molbev/msw079.

Sayyari, E., Whitfield, J.B., Mirarab, S., 2017. Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. Molec. Biol. Evol. 34 (12), 3279–3291. https://doi.org/10.1093/molbev/msx261.

Schmid, S., Genevest, R., Gobet, E., Suchan, T., Sperisen, C., Tinner, W., Alvarez, N., 2017. Hy RAD-X, a versatile method combining exome capture and RAD sequencing to extract genomic information from ancient DNA. Methods Ecol. Evol. 8 (10), 1374–1388. https://doi.org/10.1111/2041-210X.12785.

Seo, T.K., 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Molec. Biol. Evol. 25 (5), 960–971. https://doi.org/10.1093/molbev/msn043.

Shafer, A.B.A., Peart, C.R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C.W., Wolf, J.B.W., Gilbert, M., 2017. Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. Methods Ecol. Evol. 8 (8), 907–917. https://doi.org/10.1111/mee3.2017.8.issue-810.1111/2041-210X.12700.

Shi, J.J., Rabosky, D.L., 2015. Speciation dynamics during the global radiation of extant bats. Evolution 69 (6), 1528–1545. https://doi.org/10.1111/evo.12681.

Simion, P., Delsuc, F., Philippe, H., 2020. To What Extent Current Limits of Phylogenomics Can Be Overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, Phylogenetics in the Genomic Era, chapter No. 2.1, pp. 2.1:1–2.1:34. No commercial publisher | Authors open access book. The book is freely available at https://hal.inria.fr/PGE. HAL Id: hal-02535366.

Simmons, M.P., 2012. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28 (2), 208–222. https://doi.org/10.1111/j.1096-0031.2011.00375.x.

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

Smith, B.T., Harvey, M.G., Faircloth, B.C., Glenn, T.C., Brumfield, R.T., 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol. 63 (1), 83–95. https://doi.org/10.1093/sysbio/syt061.

Solís-Lemus, C., Yang, M., Ané, C., 2016. Inconsistency of species tree methods under gene flow. Syst. Biol. 65 (5), 843–851. https://doi.org/10.1093/sysbio/syw030.

Song, S., Liu, L., Edwards, S.V., Wu, S., 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. P. Natl. Acad. Sci. USA 109 (37), 14942–14947. https://doi.org/10.1073/pnas.1211733109.

Springer, M.S., Meredith, R.W., Gatesy, J., Emerling, C.A., Park, J., Rabosky, D.L., Stadler, T., Steiner, C., Ryder, O.A., Janečka, J.E., Fisher, C.A., Murphy, W.J., Stanyon, R., 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7 (11), e49521. https://doi.org/10.1371/journal.pone.0049521.

Springer, M.S., Gatesy, J., 2014. Land plant origins and coalescence confusion. Trends Plant. Sci. 19 (5), 267–269. https://doi.org/10.1016/j.tplants.2014.02.012.

Springer, M.S., Gatesy, J., 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94, 1–33. https://doi.org/10.1016/j.ympev.2015.07.018.

Springer, M.S., Gatesy, J., 2018. On the importance of homology in the age of phylogenomics. Syst. Biodivers. 16 (3), 210–228. https://doi.org/10.1080/14772000.2017.1401016.

Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30 (9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

Suchan, T., Pitteloud, C., Gerasimova, N.S., Kostikova, A., Schmid, S., Arrigo, N., Pajkovic, M., Ronikier, M., Alvarez, N., Orlando, L., 2016. Hybridization capture using RAD probes (hyRAD), a new tool for performing genomic analyses on collection specimens. PLoS ONE 11 (3), e0151651. https://doi.org/10.1371/journal.pone.0151651.

Suchan, T., 2018. hyRAD RNA probes preparation and capture. Lab protocol available at: protocols.io, ID 14096, https://protocols.io/view/hyrad-rna-probes-preparation-and-capture-rzqd75w.

Suda, J., Kyncl, T., Jarolímová, V., 2005. Genome size variation in Macaronesian angiosperms: forty percent of the Canarian endemic flora completed. Pl. Syst. Evol. 252 (3–4), 215–238. https://doi.org/10.1007/s00606-004-0280-6.

Swofford, D.L., 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0a168. Sinauer Associates, Sunderland, Massachusetts, USA.

Tan, G., Opitz, L., Schlapbach, R., Rehrauer, H., 2019. Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci. Rep. 9 (1), 1–7. https://doi.org/10.1038/s41598-019-39076-7.

Uhl, C.H., 1961. The chromosomes of the Sempervivoideae (Crassulaceae). Amer. J. Bot. 48 (2), 114–123. https://doi.org/10.1002/ajb2.1961.48.issue-210.1002/j.1537-2197.1961.tb11612.x.

Vachaspati, P., Warnow, T., 2015. ASTRID: accurate species trees from internode distances. BMC Genomics 16 (10), 1–13. http://www.biomedcentral.com/1471-2164/16/S10/S3.

van der Valk, T., Vezzi, F., Ormestad, M., Dalén, L., Guschanski, K., 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol. Resour. 20 (5), 1171–1181. https://doi.org/10.1111/men.v20.510.1111/1755-0998.13009.

van Gurp, T.P., 2017. GBS Barcode Generator. http://www.deenabio.com/services/gbs-adapter (accessed January 2017).

von Goethe, J.W., 2012. Maximen und reflexionen. Jazzybee Verlag Jürgen Beck, Altenmüster, Germany.

Wagner, N.D., Gramlich, S., Hörandl, E., 2018. RAD sequencing resolved phylogenetic relationships in European shrub willows (*Salix* L. subg. *Chamaetia* and subg. *Vetrix*) and revealed multiple evolution of dwarf shrubs. Ecol Evol. 8 (16), 8243–8255. https://doi.org/10.1002/ece3.2018.8.issue-1610.1002/ece3.4360.

Wagner, F., Ott, T., Schall, M., Lautenschlager, U., Vogt, R., Oberprieler, C., 2020. Taming the Red Bastards: Hybridisation and species delimitation in the Rhodanthemum arundanum-group (Compositae, Anthemideae). Mol. Phylogenet. Evol. 144, 106702. https://doi.org/10.1016/j.ympev.2019.106702.

Wang, X., Ye, X., Zhao, L., Li, D., Guo, Z., Zhuang, H., 2017. Genome-wide RAD sequencing data provide unprecedented resolution of the phylogeny of temperate bamboos (Poaceae: Bambusoideae). Sci. Rep. 7, 11546. https://doi.org/10.1038/s41598-017-11367-x.

Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A., Liston, A., 2014. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Appl. Plant Sci. 2 (9), 1400042. https://doi.org/10.3732/apps.140004210.3732/apps.1400042.s1.

Whitfield, J.B., Lockhart, P.J., 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22 (5), 258–265. https://doi.org/10.1016/j.tree.2007.01.012.

Wu, S., Song, S., Liu, L., Edwards, S.V., 2013. Reply to Gatesy and Springer: the multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. P. Natl. Acad. Sci. USA 110 (13). https://doi.org/10.1073/pnas.1300129110.

Xi, Z., Liu, L., Davis, C.C., 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Mol. Phylogenet. Evol. 92, 63–71. https://doi.org/10.1016/j.ympev.2015.06.009.

Xi, Z., Liu, L., Davis, C.C., 2016. The impact of missing data on species tree estimation. Molec. Biol. Evol. 33 (3), 838–860. https://doi.org/10.1093/molbev/msv266.

Xu, B., Yang, Z., 2016. Challenges in species tree estimation under the multispecies coalescent model. Genetics 204 (4), 1353–1368. https://doi.org/10.1534/genetics.116.190173.

Yang, Z., Rannala, B., 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107 (20), 9264–9269. https://doi.org/10.1073/pnas.0913022107.

Yang, Z., 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42 (5), 587–596. https://doi.org/10.1007/BF02352289.

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

Zimmermann, T., Mirarab, S., Warnow, T., 2014. BBCA: Improving the scalability of* BEAST using random binning. BMC Genomics 15 (6), S11. https://doi.org/10.1186/1471-2164-15-S6-S11.