A magnet to draw a bright needle out from the haystack – RADOrgMiner, an automated pipeline to genotype organellar reads from RADseq data

Levente Laczkó[1], Sándor Jordán[1], and Gábor Sramkó[1]

[1]University of Debrecen Faculty of Science and Technology

September 25, 2021

Abstract

Different versions of Restriction-site Associated DNA sequencing (RADseq) have become powerful and popular tools in molecular ecology. Although RADseq datasets are regarded as representative of the nuclear genome, reduced representation genomic libraries may also sample the organellar (mitochondrial and, in case of plants, plastid) DNA. Extraction of organellar loci from RADseq data can provide additional insights into the phylogenetics of the study group which comes at no additional sequencing effort. Cytoplasmic genetic variance can help better understand the evolutionary history by uncovering past hybridization and identifying the maternal (or, rarely, the paternal) lineage due to rapid lineage sorting. We developed a pipeline in bash that is based on existing bioinformatic tools to automatically mine and genotype organellar loci contained RADseq libraries. The utility of our pipeline is tested on eight, publicly available datasets spanning different phylogenetic levels (i.e. from family-level phylogenies to phylogeography) and RADseq methods (sdRAD, ddRAD, ezRAD, GBS) for genotyping both mitochondrial and plastid loci, which were subject to phylogenetic tree reconstruction. In all cases, organellar phylogenies adequately supplemented the original studies either by corroborating the large-scale picture based on RADseq or by bringing additional evidence on past or contemporary hybridization. RADseq methods designed to achieve a larger horizontal coverage (i.e. ddRAD, ezRAD) evidently yielded longer organellar alignments, but sdRAD and GBS still provided useful polymorphic loci found in the cytoplasmic DNA. Our newly developed pipeline for the above purpose can be run under a Unix-line operating system and is freely accessible at https://github.com/laczkol/RADOrgMiner

Title page

A magnet to draw a bright needle out from the haystack – RADOrgMiner, an automated pipeline to genotype organellar reads from RADseq data

Running title: RADseq Organellar DNA Miner and Genotyper

Levente Laczkó[1,2], Sándor Jordán[2,3], Gábor Sramkó[1,2]

[1]MTA-DE "Lendület" Evolutionary Phylogenomics Research Group, Egyetem tér 1, Debrecen, H-4032, Hungary

[2]Department of Botany, University of Debrecen, Egyetem tér 1, Debrecen, H-4032, Hungary

[3]Juhász-Nagy Pál Doctoral School, University of Debrecen, Egyetem tér 1, Debrecen, H-4032, Hungary

Corresponding author: sramko.gabor@science.unideb.hu

Abstract

Different versions of Restriction-site Associated DNA sequencing (RADseq) have become powerful and popular tools in molecular ecology. Although RADseq datasets are regarded as representative of the nuclear genome, reduced representation genomic libraries may also sample the organellar (mitochondrial and, in case of plants, plastid) DNA. Extraction of organellar loci from RADseq data can provide additional insights into the phylogenetics of the study group which comes at no additional sequencing effort. Cytoplasmic genetic variance can help better understand the evolutionary history by uncovering past hybridization and identifying the maternal (or, rarely, the paternal) lineage due to rapid lineage sorting. We developed a pipeline in bash that is based on existing bioinformatic tools to automatically mine and genotype organellar loci contained RADseq libraries. The utility of our pipeline is tested on eight, publicly available datasets spanning different phylogenetic levels (i.e. from family-level phylogenies to phylogeography) and RADseq methods (sdRAD, ddRAD, ezRAD, GBS) for genotyping both mitochondrial and plastid loci, which were subject to phylogenetic tree reconstruction. In all cases, organellar phylogenies adequately supplemented the original studies either by corroborating the large-scale picture based on RADseq or by bringing additional evidence on past or contemporary hybridization. RADseq methods designed to achieve a larger horizontal coverage (i.e. ddRAD, ezRAD) evidently yielded longer organellar alignments, but sdRAD and GBS still provided useful polymorphic loci found in the cytoplasmic DNA. Our newly developed pipeline for the above purpose can be run under a Unix-line operating system and is freely accessible at https://github.com/laczkol/RADOrgMiner.

Keywords

bioinformatics, cytonuclear discordance, phylogenetic incongruence, hybridization, phylogeography, reduced representation library

1 | Introduction

High-throughput sequencing (HTS) revolutionized ecological and evolutionary genetics by providing access to high-quality DNA sequence information for non-model organisms at the genomic level (Mardis, 2008; Metzker, 2010; Sboner, Mu, Greenbaum, Auerbach, & Gerstein, 2011; Schlötterer, 2004). Additionally, reduced complexity or reduced genomic representation library (RRL) approaches made the in-depth study of micro-evolutionary processes at the population level feasible by providing genomic-level data – usually genome-wide Single Nucleotide Polymorphism (SNP) – to characterize the genetic make-up of populations, which gave rise to population or ecological genomics (Luikart et al., 2019; Narum, Buerkle, Davey, Miller, & Hohenlohe, 2013). A growing number of publications rely on cost-effective SNP discovery made available through RRL approaches (Leaché & Oaks, 2017). Arguably, the most popular RRL methods nowadays (Holliday, Hallerman, & Haak, 2019) are from the group of Restriction-site Associated DNA-sequencing (RADseq) approaches (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). Originally, the term was meant to refer to a particular protocol used to obtain sequence information about a large number of loci (Baird et al., 2008), but was later adopted and fine-tuned to represent a wide range of techniques that rely on type II restriction enzymes to sample genomic diversity (Andrews et al., 2016; Rivera-Colón, Rochette, & Catchen, 2021). The original RADseq approach (Baird et al., 2008) – called also single-digest RADseq (sdRAD) – uses a single enzyme to initially cut the whole genomic DNA into fragments, which are then mechanically sheared and size-selected. In contrast, double-digest RADseq (ddRAD) (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012) makes a selection of markers by using two restriction enzymes at the initial step of the library preparation. A larger horizontal coverage can be achieved by applying an ezRAD protocol (Toonen et al., 2013), which uses a frequently cutting enzyme (or enzyme combination) to initially cut the genome up. Genotype-By-Sequencing (GBS) (Elshire et al., 2011) also uses (an) enzyme(s) to select markers at the beginning, but in contrast to the previously discussed methods, it then uses Polymerase Chain Reaction (PCR) to achieve size selection. Additional variants of the approaches listed above have been developed (reviewed by Andrews et al., 2016) with each of them having particular strengths for a given purpose or being disadvantageous in certain ways (Davey et al., 2013; Hohenlohe, Hand, Andrews, & Luikart, 2019; Narum et al., 2013; Puritz, Matz, et al., 2014). With all the variants included, the RRL approaches that include a restriction enzymes to obtain DNA sequence information from a large set of loci at the genome level can collectively be called RADseq (Andrews et al., 2016). A growing number of software solutions are available for the effective process of RADseq data (e.g. Catchen, Amores, Hohenlohe, Cresko, & Postlethwait, 2011; Davey et al., 2011; Eaton & Overcast, 2020; Puritz, Hollenbeck, & Gold, 2014; Rivera-Colón et al., 2021; Rochette, Rivera-Colón, & Catchen, 2019).

A common feature of all RADseq methods is that they sample the whole genome anonymously (i.e. no *a priori* information is disposable regarding the origins of the genome-wide reads) (Andrews et al., 2016; Hohenlohe et al., 2019). Nevertheless, irrespective of whether the RAD-loci were assembled *de novo* or mapped onto a reference genome, the filtered variants obtained by RADseq experiments are*de facto* treated as a set of biallelic markers originating solely from the nuclear genome. However, in a reference genome-based analysis only 53.6% of the GBS tags aligned to the closely related nuclear reference genome that was partly explained by the presence of tags originating from the cytoplasmic genome (D'Agostino et al., 2018). This highlights the presence of organellar reads in RRLs, although their representation can be poor and limited by the cut site frequency of the given restriction enzyme in a particular organellar genome (Bentley, Grauke, & Klein, 2019).

Genetic information from the organelles has long been utilized in phylogenetics and phylogeography as sources of haploid, uniparentally inherited genomic compartments (Avise, 2000; Avise, 2004; Soltis & Soltis, 1998; Uncu, Uncu, Celık, Doganlar, & Frary, 2015). Although the assembly of whole plastomes or mithochondria is unlikely to be derived from RRLs, some desirable characteristics of the organellar DNA, even if only partially represented, can provide additional insight into the evolutionary history of the studied organisms. Their molecular evolution is virtually independent from the nuclear genome and can open a window into contemporary and ancient hybridization events via the phenomenon called 'phylogenetic incongruence' (Wendel & Doyle, 1998) or 'cytonuclear discordance' (Rieseberg & Soltis, 1991). In such cases, the hybrid individual possesses organellar haplotypes different from the one(s) characteristic for their lineage in a nuclear phylogeny as a result of introgression from other (the hybridizing) lineages, thereby clearly indicating the direction of hybridization. This incongruence is frequently used to identify hybrid individuals or lineages (e.g. Daru et al., 2013; L. H. Rieseberg, Whitton, & Linder, 1996; Scheunert & Heubl, 2014; Seelanan, Schnabel, & Wendel, 1997; Sramkó et al., 2016) and thus is a useful source of information. In zoology, where the similar phenomenon is usually termed 'mito-nuclear' discordance, such incongruence is used in similar ways to address additional questions on the evolutionary history of the studied organisms (Funk & Omland, 2003; Toews & Brelsford, 2012). Although stringent analyses are needed to distinguish hybridization from incomplete lineage sorting (Lee-Yaw, Grassa, Joly, Andrew, & Rieseberg, 2019), the fourfold smaller effective population size of organellar DNA may lead to rapid lineage sorting (Zink & Barrowclough, 2008) and thus incongruence between organellar and nuclear DNA likely indicates hybridization in most cases. In addition to the above, organellar DNA has been the most important source of phylogeographic analyses until recently (Brito & Edwards, 2009; McCormack, Hird, Zellmer, Carstens, & Brumfield, 2013). This was due to some characteristics such as the lack of recombination, haploid nature, uniparental inheritance, and an often strong correlation with geography.

Comparison of RADseq datasets and organellar datasets within the same study organism has gained some popularity and has usually uncovered hybridization between lineages (Barnard-Kubow, Debban, & Galloway, 2015; Macher et al., 2015; Moura et al., 2015; Puckett, Etter, Johnson, & Eggert, 2015; Streicher et al., 2014; Sutherland & Galloway, 2018; Uckele, Adams, Schwarzbach, & Parchman, 2021). In these studies, however, the RADseq dataset was regarded as a 'representative' of the nuclear genome and the organellar dataset was obtained by Sanger-sequencing mitochondrial or plastid candidate-genes. Nevertheless, RRL approaches may also sample the organellar genome, and can potentially be used to sort out organellar reads from nuclear ones (Stobie, Cunningham, Oosthuizen, & Bloomer, 2019). This approach is different from assembling the organellar genome to achieve increased phylogenetic resolution using capture-based HTS (Mariac et al., 2014) or genome skimming (Cronn et al., 2008; Parks, Cronn, & Liston, 2009; Straub et al., 2012). In this case, the sole aim is to provide extra information from the RADseq dataset by separating organellar reads from those coming from the nuclear genome without additional sequencing effort. This latter approach, what we may term as 'organellar mining', was only applied in a handful of studies with an aim to address the potential utility of tags originating from the organellar genome (Clugston et al., 2019; Du, Harris, & Xiang, 2020; Feng, Xu, Feng, von Wettberg, & Kang, 2017; Forsman et al., 2017; McVay, Hipp, & Manos, 2017; Meger, Ulaszewski, Vendramin, & Burczyk, 2019; Rincon-Sandoval, Betancur-R, & Maldonado-Ocampo, 2019; Straub et al., 2012; Terraneo, Arrigoni, Benzoni, Forsman, & Berumen, 2018). All these studies above either use existing software solutions in a custom-modified way to sort reads from the different genomes or use an in-house script for this purpose.

Although the number of phylogenetic and phylogeographic studies that rely on SNP datasets is growing fast, the contrast between the genetic information from organellar loci and the nuclear genome can still be highly important. Here, we introduce a custom pipeline, RADOrgMiner, which is explicitly designed to automatically sort out organellar reads from nuclear reads in datasets generated from RADseq group of approaches (i.e. RRL approaches that rely on the restriction digest technique). Our user-friendly software allows subsequent comparison of genetic information coming from the organellar and the nuclear genome without additional sequencing effort.

2 | Materials and Methods

2.1 | Benchmarking

We demonstrate the power of our software solution for mining out organellar reads generated in a RADseq experiment by re-analyzing publicly available datasets (using various variants of this methodological group) originating from eight studies with a focus on different phylogenetic levels (i.e. from family-level phylogenies to phylogeography) and assessed the presence of mitochondrial and plastid loci in libraries (Table 1). Some datasets were also screened for cytoplasmic sequence tags by the original authors and we use those results to compare the output of the different analyses. In the cases of the Cycadales and the *Porites* datasets, we used two different reference genomes to assess the level of filtering robustness. For the*Stellaria* dataset, we only analyzed the samples that belong to the 'broad' taxonomic range (Sharples & Tripp, 2019) which consisted of fewer samples.

Datasets of the eight studies were downloaded and used as input to our pipeline with the references specified (Table 1). The genotyped organellar loci were used to generate phylogenetic trees, which are described below and their information content is compared to the original publications. We used cytoplasmic loci to reconstruct the phylogenetic relationships of the samples from the eight benchmark studies using IQtree 2.0.3 (Minh et al., 2020). Phylogenetic tree reconstruction was performed by turning on automatic model selection and setting each locus as individual partition that could be merged for a better model fit. We calculated the approximate likelihood ratio test (aLRT) (Anisimova & Gascuel, 2006) branch support values after 1000 replications. We defined statistical support for branch robustness as existing if aLRT[?]80%. We visualized the resulting trees, the read depth of loci, and the amount and proportion of reads used for the assemblies with R 3.6.3 (R Core, 2012), ggplot 2 3.3.4 (Wickham, 2016), and ggtree 2.0.4 (Yu, Smith, Zhu, Guan, & Lam, 2017) and further edited in Inkscape 0.92 (https://inkscape.org/) to improve readability.

2.2 | RADOrgMiner pipeline

Our pipeline uses existing bioinformatic tools to screen RADseq reads if they align well to a closely related organellar genome provided by the user, separates the organellar reads from those not coming from organelle, then genotypes loci from the aligned reads. The pipeline uses two main steps. In the first step, we align all the reads to a closely related reference genome using bwa 0.7.17. (Li, 2013) then separat reads that can be aligned with samtools 1.10.2 (Danecek et al., 2011). To decrease the number of chimeric reads in the resulting data, we require both ends to be aligned to the reference for paired-end reads. As the concerted evolution of plastid inverted repeats cannot be ruled out (e.g. Knox, 2014), we mask one copy of the repeat with N-s to minimize the number of ambiguous alignments. The location of the inverted repeats is identified by self-blasting using blastn 2.10.1+ (Altschul et al., 1997). At the second step, aligned reads (originating from the organelle) are stored in bam files, whereas unaligned reads (representing the non-organellar genome) are stored as fastq files and can be processed further by any pipeline designed for RADseq data processing. To minimize the amount of missing data and false alignments of nuclear plastid (NUPTs) and nuclear mitochondrial (NUMTs) DNA, an alignment interval is only processed further as an individual locus if the read depth is higher in any individual than the defined minimum value (as exemplified in Table 1). We inspected the mean and individual read depth for each case study and set this value to include the highest number of base pairs (bps) without increasing the amount of missing data. As nuclear sequences are expected to be present with a lower read depth relative to the organellar genome (Ekblom, Smeds, & Ellegren, 2014), setting this threshold can help minimizing falsely aligned NUPTs and NUMTs. We use the aligned reads to call haplotypes using freebayes 1.3.2 (Garrison & Marth, 2012) for which alignment intervals are created with bedtools 2.29.2 (Quinlan & Hall, 2010). An alignment interval defines a genomic region with continuously overlapping reads that we refer to as individual locus. In light of the supposedly high read depth of organellar loci, they might be visualized as ”spikes” along the reference genome as a function of read depth (Figure 1). The advantage of this approach is that haplotype calling of loci can be parallelized to decrease the run time of this step drastically. Moreover, base calling can be controlled and narrowed down to use only those loci most probably of cytoplasmic origin.

We chose freebayes for its high and easy customizability for a Bayesian haplotype calling using the aligned reads. For default settings of genotyping in the pipeline, we only consider reads with a mapping quality larger than 30 and bases with a quality larger than 20. Minimum coverage for base calling step that used the five most probable alleles is set to five, and, to exclude low-frequency mismatches from the base calls, a minimum of 40% of the total read depth is required for an alternate allele to be called. All the above settings can be changed from the command line allowing fine-tuning the genotyping for a given dataset. Species, or if multiple populations can be analyzed within a species, populations can be used as a prior. Freebayes is included in our pipeline to use mapping quality for likelihood calculation with clumping of haplotypes disabled, and HWE priors turned off. Binomial observations priors are turned off, and read placement probability, strand balance probability, and read position probability are used instead. As freebayes is capable of ploidy-aware base calls, ploidy is set to one. All sites are annotated, including the monomorphic ones, and are exported into a vcf file. Missing data, arising mainly at the sheared ends, are filtered with vcftools 0.1.16 (Danecek et al., 2011), allowing a maximum of 20% missingness across all individuals as default. Vcf files are converted to fasta with vcf2fasta from the vcflib 1.0 package (Garrison, Kronenberg, Dawson, Pedersen, & Prins, 2021) and aligned with muscle 3.8.1 (Edgar, 2004). As vcf2fasta uses the reference genome for vcf conversion, the reference is subsetted by the start and end coordinates specified in the filtered vcf files of each locus. This way, regions without reads aligned and sites with a high amount of missing data will not be included in the final dataset, and the total length of alignments (loci) can be included in downstream analyses. We calculate alignment statistics, including the alignment length, number of polymorphic and informative sites and concatenate the individual loci with a minimum length of 100 bp using the AMAS 1.0 python package (Borowiec, 2016).

The pipeline was written in bash and can be parametrized from the command line for easy and reproducible usability. All benchmark runs were conducted in a Debian 10.1 environment, but with dependencies correctly installed, the pipeline should run using most Unix-like operating system. The pipeline with the list of dependencies, documentation, and example run is available at https://github.com/laczkol/RADOrgMiner.

3 | Results

3.1 | *Paragorgia*dataset

Alignment of the reads of the *Paragorgia* dataset to the reference showed that individuals had 492–35,494 reads aligned to the reference (mean = 9051) that represented 0.03–0.83% (mean = 0.26%) of all reads (Figure 2). Mean read depth ranged from 2.25 to 169.5 across the reference genome. The analysis yielded four loci (Figure 1A) with 76 polymorphic sites in total, 57 of which appeared to be informative. The final alignment length was 709 bp long that covered 4% of the mitochondrial genome, and loci appeared to be 177–178 bp long (mean = 177). Phylogenetic reconstruction (Figure 2) placed *Anthomastus*and *Heteropolypus* on the same branch. The next branch separated*Corallium* and *Heterocorallium* from all the*Paragorgia* and *Sibogagorgia* samples. Within*Paragorgia,* a poorly supported branch separated *Paragorgia kaupeka* from *Sibogagorgia cauliflora.* Towards the end of the tree, although the separation of species was supported, the phylogenetic relationship could not be resolved as *Paragorgia coralloides, Paragorgia arborea* and *Paragorgia pacifica* and the rest of the samples formed a trichotomy.

3.2 | ***Porites*dataset**

The reanalysis of the *Porites* dataset using two different reference genomes yielded nearly identical results. The only difference, if using *Porites rus* as a reference, resulted in two more polymorphic, but not parsimony informative sites. Thus, we only present the results based on the other reference genome of *Porites lobata.* Samples had 786–5332 reads aligned to the reference (mean = 2384.48) that corresponded to 0.02–0.14 % of all reads (mean = 0.07%) (Figure 3). We recovered the entire reference genome as a continuous alignment (Figure 1B,C) with a minimum coverage of 4.7 and a maximum of 44.6 across the reference. The alignment of loci was 18,646 bp long (100% of the reference genome), showing 188 polymorphic and 98 informative sites. Phylogenetic analysis (Figure 3) separated*Porites superfusa* first with high certainty. *Porites rus*and *P. evermanni* were identified as sister taxa. *Porites lobata* clustered in two groups, one of them is sister to *P.* cf.*brighami,* whereas the other is mixed with *P. compressa* .

3.3 | ***Labeobarbus*dataset**

We identified a total of seven loci (Figure 1D) in the*Labeobarbus* dataset, of which all appeared to be polymorphic and yielded a total of 363 polymorphic and 288 informative sites. We were able to align 0.006–0.59% (mean = 0.12) of all reads that represented 264–45,047 (mean = 8194.7) reads in total (Figure 4). Loci were 285–3999 bp long (mean = 1972.3), yielding a total of 13,806 bp long alignment that covered 83.3% of the reference genome with a mean coverage of 0.93–232 within individuals. Phylogenetic reconstruction placed *Labeobarbus natalensis* on a distinct clade (Figure 4).*Labeobarbus aeneus* and *L. kimberleyensis* were identified as sister species, although some samples of *L. aeneus* bore haplotypes of *L. kimberleyensis.* Technical replicates from the study showed maximum one difference if not counting the missing data (La005, LnBL004, see Supplementary Material).

3.4 | ***Xylosandrus*dataset**

The first analysis of *Xylosandrus* samples yielded no loci with maximum 20% of missing data. After checking the read depth distribution of the samples, a large amount of samples with non-overlapping loci was noticeable. To analyze only those samples with overlapping loci, we set a threshold of 1000 on the minimum number of aligned reads in a sample to include it in the pipeline and increased maximum missingness to 50% in the second analysis. This constrain decreased the number of samples from 198 to 76. The second analysis yielded six loci in total (Figure 1E). All loci seemed to be variable and the total alignment showed 179 polymorphic and 174 informative sites on 1508 bp length (8.9% of the reference mitochondrial genome). Read depth across the entire mitochondrial genome ranged from 4.25 to 89.75. Loci appeared to be 100–395 bp long with a mean length of 251.3 bp. Samples had 1006–14,203 reads aligned (mean = 5132.75) which represents 0.08–1.3% of all reads (mean = 0.52%). IQtree placed the samples from Okinawa in two separate clusters (Figure 5). Five of them, clustered together with samples from Aichi, were placed on a poorly supported branch sister to most of the North American sampling sites. The other main clade clustered the samples from China and Taiwan as sister branches to the African sampling sites. These could be separated from two samples collected at Okinawa, although with low support value.

3.5 | *Melicope*dataset

Samples of the *Melicope* dataset had 9213–3,997,452 reads aligned (mean = 609,756.4). 1.73–14% of reads of individuals (mean = 5.36%) aligned to the reference genome (Figure 6). The mean read depth within samples across the entire reference ranged from 4.93 to 2159. In total, the analysis yielded 13 loci (Figure 1F) located on the plastid genome. The default value of missingness resulted in a scattered final alignment, we decreased the maximum amount of missingness to 10%. One locus was monomorphic, and two did not show parsimony informative sites. The final alignment length was 2300 bp long (1.4% of the plastome) that showed 45 polymorphic and 31 informative sites and consisted of 175–177 bp long loci (mean = 176.9). For the phylogenetic reconstruction, we only used the loci with at least one informative site. IQtree used a 1946 bp long (1.22% of the plastome) alignment and 43 polymorphic sites. The phylogenetic reconstruction (Figure 6) separated the outgroup sequences (*Melicope aneura, M. polyadenia, M. brassii, M. durifolia* , and *M. triphylla)* and the *Platydesma* group with high certainty. The next branching with moderate support placed the*Pelea* group sister to the rest of the samples. The two*Apocarpa* groups described by Paetzold, Wood, Eaton, Wagner, and Appelhans (2019) were mixed and placed as a sister lineage to the groups*Megacarpa* + *Cubicarpa* .

3.6 | *Helianthemum*dataset

We aligned 7752–2,274,688 reads (mean = 560,541.4) to the reference genome of the samples of the *Helianthemum* dataset that represents 0.77–23.64% (mean = 7.9%) of all reads with a mean read depth ranging from 4.5 to 1564.9 (Figure 7). The analysis yielded 12 loci in total (Figure 1G), of which one appeared to be monomorphic. In total, we discovered 385 variable and 220 informative sites on an alignment length of 5455 bp (43.6% of the reference). The length of loci ranged from 112 bp to 884 bp with a mean length of 437.75. The phylogenetic reconstruction (Figure 7) using the polymorphic loci separated the outgroup taxa (*Fumana thymifolia, Tuberaria macrosepala, Halmium lasianthum, Cistus ladanifer*) with high probability. The subgenera*Plectolobum* and *Helianthemum* were placed as sister lineages. Within *Plectolobum* , the section *Macularia* were separated first, and the sections *Caput-felis* and*Atlanthemum* were placed sister to *Pseudocistus.* We found two main clusters within the subgenus *Helianthemum* .*Pseudomacularia* appeared to be nested within *Eriocarpum.*The sections *Lavandulaceaum* and *Argyrolepis* were found on a relatively long branch. The other main clade consisted of the sections*Brachypetalum* and *Helianthemum* that were placed as sister lineages.

3.7 | Cycadales dataset

Aligning the samples of the Cycadales dataset to the reference plastome of *Cycas shiwandashanica* resulted in fewer polymorphic sites and shorter alignments than using the reference plastome of *Macrozamia mountperriensis.* However, the read depth along the reference genome sequence seemed similar (Figure 1H and Figure 1J). When we used*C. shiwandashanica* as a reference, samples had 17,523–193,328 (mean = 81,480) reads aligned that correspond to 0.43–4.0% (mean = 1.58%) of all reads. Within individual mean depth ranged from 9.93 to 133.35. This analysis yielded 109 loci, all of which appeared to be polymorphic with a mean length of 923 bp. Length of loci ranged from 119 to 7382 bp. We discovered 9166 polymorphic and 4893 informative sites on an alignment length of 100,617 bp (62.1% of the reference). When using*M. mountperriensis* as a reference, we aligned 19,551–189,194 (mean = 85,373.2) per sample that represented 0.48–3.99% (mean = 1.65%) of all reads. Individual mean read depth appeared to be between 11.4–119.92. The pipeline returned 103 loci with a minimum length of 109 and a maximum of 6675 (mean = 1037.93). The final alignment showed 11,541 polymorphic and 5369 informative sites on a 106,907 bp length. Phylogenetic analysis reconstructed the same phylogenetic tree with nearly equal support values; thus, we only present our results (Figure 8) using *M. mountperriensis* that was based on a longer alignment covering 64.26% of the reference with more polymorphic sites. The genus*Cycas* could be well separated and was found at a high genetic distance. Within the ingroup, *Dioon mejiae* seemed to diverge the earliest. The tribe Encephalarateae was placed sister to the rest of the samples. The family Stangeriaceae and the tribe Ceratozamieae appeared to be mixed. *Bowenia spectabilis* appeared to have diverged the earliest and *Ceratozamia kuesteriana* and

*Stangeria eriopus*were clustered as a sister lineage to the tribe Zamieae (*Microcycas calocoma* and *Zamia integrifolia*).

3.8 | *Stellaria*dataset

Samples of the *Stellaria* dataset had 10,191–2,027,645 (mean = 444,062) reads aligned to the reference, which represented 0.37–57.63% (mean = 11.62%) of the reads (Figure 9). Mean read depth within individuals ranged from 4.78 to 1094.3. We discovered 46 organellar loci (Figure 1K), of which all were polymorphic, with a minimum length of 101 bp and a maximum of 413 bp (mean = 192.1 bp). The final alignment showed 1397 polymorphic sites of which 887 appeared to be informative on an 8837 bp long alignment that covered 5.9% of the reference plastome. Phylogenetic reconstruction (Figure 9) separated the tribes Sileneae, Sagineae, Arenarieae, and Alsineae with high certainty. Within Alsineae,*Stellaria howardii* and *Stellaria antillana* formed a distinct cluster. The remaining samples of Alsineae could be divided into three main clusters. The first consisted of three *Stellaria*species, *Adenonema, Mesostemma* and *Pseudostellaria. Stellaria mannii* and *Stellaria monosperma* were clustered together, but *Stellaria americana* was grouped with*Pseudostellaria. Adenonema* was placed as a sister lineage to*Mesostemma* + *Pseudostellaria. Nubeleria* was identified as a monophyletic clade sister to the remaining Alsineae. The clade of*Holosteum* and *Cerastium* could be separated from the core*Stellaria* species. Within core *Stellaria* ,*Petiolares* and *Insignes* diverged earliest and appeared to be mixed. The next branching event separated *Plettkeae* from*Nitentes* , which we identified as the sister lineage of*Larbreae* .

4 | Discussion

Our study introduces the pipeline RADOrgMiner specifically designed to genotype organellar loci found in RADseq data. Even though the proportion of reads aligned to the reference could be very variable even within a given dataset, genotype calls were consistent according to the taxonomy of the samples (i.e. *a priori* similar groups could be clustered together at various taxonomic levels). Our pipeline evidently relies on the availability of a closely related reference organellar genome. However, as more and more such genomes become available in public databases, it is not expected to seriously hinder our pipeline's utility.

4.1 | *Paragorgia*dataset

The *Paragorgia* dataset yielded the shortest final alignment. Loci could be equivocally identified at four different parts of the mitochondrial genome. In the placement of the most distinct lineages, our phylogenetic tree reconstruction agreed with the analysis of Herrera and Shank (2016) based solely on the mitochondrial *mut* S gene but placed *Sibogagorgia cauliflora* and *Paragorgia kaupeka*differently. Moreover, especially towards the tip of the phylogenetic tree, we observed some unresolved branches, most probably due to the low mutation rate reported in this group (Herrera & Shank, 2016). Another cause for the low resolution could be the very short alignment length, although 10.7% of the alignment was polymorphic. Finer-scale phylogenetic resolution based on mitochondrial DNA could be heavily influenced by incomplete lineage sorting, which could also explain the different results of the two analyses.

4.2 | *Porites*dataset

We were able to confirm the results of Forsman et al. (2017) on the observed low mitochondrial polymorphism of *Porites* species, as only 1% of the alignment was polymorphic. This low level of polymorphism could explain why using two different reference genomes for read alignment yielded identical phylogenetic results. Our results are in line with those presented by Forsman et al. (2017), except in the placement of three *Porites compressa* and one *P. lobata*samples, which, in the analysis of the original authors, showed unique haplotypes within the mixed *P. compressa-P. lobata* group. This contradiction might be caused by the different pipelines applied to align and assemble the mitochondrial genome. Whereas the original authors used Geneious 8.0.2 (Biomatters Inc.) for all tasks, we aligned the reads with bwa using default options;

then, instead of generating a 0% majority consensus sequence, we used freebayes to annotate positions also performing a local re-alignment of indels to minimize false SNP calls and implements haploid base calling. The discovery of the same number of informative sites suggests consistency over the approaches (i.e. Geneious or RADOrgMiner) used.

4.3 | *Labeobarbus*dataset

Our pipeline could be applied well to the *Labeobarbus* dataset, and our results agreed with those presented by Stobie et al. (2019). Despite the presumed higher mutation rate, the polymorphism can be considered relatively low (2.62% of the alignment). Still, we identified the same three clades as the original authors, and the general structure within the main clades was concordant with the results of Stobie et al. (2019). Moreover, the number of individuals that bore the haplotype of the sister species was the same; thus, we identified the same number of hybrid individuals as the original authors. Given the low error rate of technical replicates, using this dataset, we consider our pipeline to be at least as accurate as the method presented by Stobie et al. (2019). Single SNP differences between the technical replicates were observed in our results, which did not otherwise influence phylogenetic results. The visual inspection of the alignment confirmed that these SNPs could result from misalignment by muscle as they were found around blocks of missing data.

4.4 | *Xylosandrus*dataset

The drastically different read depth distribution in the*Xylosandrus* dataset highlights the potential technical limitations in the application of our pipeline since overlapping loci could only be found in a smaller proportion of the individuals. Although it could not be tested explicitly, the high missingness of our first analysis could stem from the wet-lab protocol or the applied sequencing method (SE Illumina NextSeq). Still, the samples included in the final alignment showed a congruent picture with those results presented by Storer, Payton, McDaniel, Jordal, and Hulcr (2017). Despite the relatively short alignment length, owing to the observed 11.8% polymorphism, we could identify very similar phylogeographic clusters as the original authors both at the native (East-Asia) and introduced range (Africa and North-America) of the study species. Moreover, based on the branch lengths of the phylogenetic tree (Figure 5), samples of the native range showed a somewhat higher differentiation than the samples of the introduced populations, which could be expected from the analysis of mitochondrial DNA.

4.5 | *Melicope*dataset

The example of the *Melicope* dataset proves that ancient hybridization events can be effectively detected by supplementing the nuclear SNPs with cytoplasmic sequence data. Despite the low polymorphism (1.9%) and relatively low phylogenetic support, we detected that the members of the two *Apocarpa* groups bear the same haplogroup. This result could be especially exciting when considering the results of Paetzold et al. (2019). Their results using nuclear SNPs divided *Apocarpa* into two groups and clustered them on two distinct clades. Although the analysis of the nuclear dataset clearly showed that an ancient hybridization event is possible between*Pelea* and *Apocarpa* , this result was not supplemented by organellar DNA, which could directly show introgression. In such cases, analysis of the organellar DNA can provide valuable additional evidence on the observed pattern of the nuclear SNPs.

4.6 | *Helianthemum*dataset

Similarly, the analysis of the *Helianthemum* dataset demonstrated that cytonuclear discordance might be evaluated using our pipeline. Martin-Hernanz et al. (2019) described subgenus *Helianthemum* as paraphyletic with the need of including of subgenus *Plectolobum*to achieve monophyly, although the main lineages are sitting on short branches (Figure 5. of Martin-Hernanz et al., 2019). Our results based on organellar loci show subgenus *Helianthemum* to be monophyletic (Figure 7), although the node connecting two main lineages of the subgenus *Helianthemum* could not be resolved with high support. Nevertheless, the genetic distance between the two *Helianthemum*clades seemed lower than the one leading to *Plectolobum* . The incongruent placement of one of the main clades of subgenus*Helianthemum* supports the hypothesis of the original authors, who concluded on the role of hybridization in the diversification of their studied taxa. However, the fine-scale pattern in our phylogenetic tree was not fully compatible with the SNP dataset, most possibly due to incomplete lineage sorting.

4.7 | Cycadales dataset

Although Clugston et al. (2019) described organellar DNA as suboptimal for phylogenetics, 10.7% of our organellar alignment was variable. However, this must be connected to the broad taxonomical focus of the dataset (a plant family), and the proportion of polymorphic sites is comparable to other datasets covering a narrower taxonomic scope. Still, a high proportion of the plastome could be assembled, and the phylogeny could be reconstructed. The phylogenetic relationships reconstructed for Zamiaceae are fully compatible with the results of Salas-Leiva et al. (2013) if only using the plastid genes *rbc* L and *mat* K. The same result suggests a high accuracy of our pipeline and an unequivocal phylogenetic signal across the plastome, further corroborating by the same phylogenetic results using two distantly related references. RADSeq could be sensitive to allele dropout (Andrews et al., 2016). Surprisingly, a high proportion of common loci could be assembled from all samples despite the high evolutionary distance between*Cycadaceae* and *Zamiaceae* (Salas-Leiva et al., 2013). Although considering the group's low sequence variability (see Clugston et al. 2019), the high amount of overlapping loci is less surprising. Another conclusion could be that at least for some, more conserved groups, these sequence data source could resolve relatively deeper splits in the phylogeny with a lower sensitivity to allele dropout. Conversely, achieving a finer scale resolution, in this case, can be hindered by the low mutation rate. A comprehensive study, similar to that of Gautier et al. (2013), could effectively assess the extent of usability of RRL derived organellar reads by comparing the allele dropout frequency of different parts of the whole genome (i.e. nuclear and organellar) given various conditions. Ideally, such an experiment should use real data and include Sanger re-sequencing to assess base calling accuracy. However, answering these questions is clearly out of the scope of this study.

4.8 | *Stellaria*dataset

The broad-scale phylogenetic reconstruction of the *Stellaria* dataset showed a very similar picture to the results presented by Sharples and Tripp (2019) and branches received high support. However, the structure within core *Stellaria* was not compatible with the results of the SNP dataset of the original authors. This incompatibility at the tip of the phylogenetic tree, similar to the *Helianthemum* dataset, could possibly be caused by incomplete lineage sorting, which, despite the relatively high polymorphism (15% across the whole dataset and 6% within core *Stellaria*), blurs phylogenetic resolution. Still, the major lineages, including core *Stellaria* , could be identified with high certainty using both SNPs (Sharples & Tripp, 2019) and the plastid sequence data obtained in our experiment.

4.9 | The utility of organellar loci mined from RADseq datasets

The results obtained by our pipeline show a great variety of phylogenetic resolution when applied to different datasets. In all cases, organellar phylogenies (Figure 2–9) adequately supplement the original studies' findings either by corroborating the large-scale picture based on RADseq or by bringing additional evidence on hybridization. Having said that we also report low phylogenetic resolution based solely on mined organellar reads, which may hinder drawing additional evidence from this source of information. The examples listed above testify the strong dependence of organellar phylogenetic resolution on the dynamics of molecular evolution of cytoplasmic DNA in the focal study group. Above all, as mined organellar loci does not come at a cost to the sequencing experiment, it seems reasonable to extract cytoplasmic loci from the RADseq dataset to check for potential additional phylogenetic information. Our newly devised pipeline is an excellent start to draw organellar loci out from the 'stack of DNA sequences' in a RADseq experiment at no additional sequencing effort.

Attention must, however, be paid to the applied wet-lab protocol as different methodology to obtain RADseq data has a significant impact on the sound application of our pipeline. Our comparison shows that the wet-lab protocols, including the cut frequency of restriction enzyme(s), strongly affect the length of the final organellar alignment. Not surprisingly, ezRAD (Toonen et al., 2013), which is known to provide higher horizontal coverage, produced longer alignments. Similarly, ddRAD (Peterson et al., 2012) could also provide high read depth and horizontal coverage if frequently cutting enzymes are used. The average organellar read depth obtained in different RADseq protocols indicates that the higher number of cut sites might need a higher total read number to capture enough data to assess sequence variation. If a higher mutation rate is assumed for the organellar DNA or the studied taxa are more distantly related, sdRAD (Baird et al., 2008) or GBS (Elshire et al., 2011), which tended to result in a lower horizontal coverage, could still yield sufficient amount of organellar sequence variation. Although using different reference genomes did not show pronounced effect on the outcome of the analysis, using more distantly related genomes as reference could result in fewer polymorphisms (Bohling, 2020); thus, if available, the usage of the most closely related reference could still be advised.

Consequently, both the molecular evolution of the studied group and the applied library preparation protocol are essential factors to mine for organellar loci in RADseq data. Suppose a closely related reference genome is available and recovery of organellar loci is an important factor. In that case the libraries should be assessed before the experiment to check for the abovementioned conditions by using dedicated bioinformatic tools (e.g. fragmatic (Chafin, Martin, Mussmann, Douglas, & Douglas, 2018), GBS-Pacecar (Melo & Hale, 2018) or RADinitio (Rivera-Colon et al., 2021)); see also Stobie et al. (2019).

In sum, we showed that RRLs might contain a significant amount of cytoplasmic DNA. Our pipeline can reliably genotype organellar loci from RADseq datasets and provides basic measurements to assess the read depth and variability of the loci. We showed that the analysis of organellar loci could effectively supplement the results of the nuclear SNP dataset at no additional sequencing effort. As expected by introducing restriction enzymes in RADseq, the total organellar genome could not be recovered in most cases, but this is also feasible under certain circumstances. In addition, given the ever growing number of available organellar genomes in public repositories (Tonti-Filippini, Nevill, Dixon, & Small, 2017), the mined organellar loci can help to put the focal taxa into a wider phylogenetic context, if the orthologus organellar regions can be identified by sequence matching against related sequences. Reduced genomic complexity sequencing is a valuable tool for SNP discovery, and, as already pointed by Stobie et al. (2019), the complementary analysis of the nuclear and the organellar genome sampled by RRLs can provide important information about cytonuclear discordance, the frequency, and directionality of hybridization, and phylogeography.

Acknowledgements

We would like to thank for Laszlo Bartha, Tamas Malkocs and Lajos Szatmari for their comments on the pipeline, and Nikoletta Nagy, Jelena Marinkov and Jacint Tokolyi for their advices on the first draft. Financial support was achieved via the New National Excellence Program of the Hungarian Ministry of Innovation and Technology (UNKP-20-4-I-DE-290) to LL, and the "OTKA" young researcher excellence program (FK_ 137962) to GS. SJ thanks the support of the Juhasz-Nagy Pal Doctoral School of Biology and Environmental Sciences (University of Debrecen).

Author contributions

LL and GS conceived the idea and designed the study; LL wrote the bioinformatics pipeline and SJ contributed; LL and SJ made analyses on the exemplary datasets; LL and GS evaluated final results; LL drafted the first version of the manuscript, while all authors have contributed to writing.

Data availability statement

All datasets used in this study are publicly available from NCBI GenBank (see Table 1). The newly devised pipeline is available from https://github.com/laczkol/RADOrgMiner

ORCID

Levente Laczko https://orcid.org/0000-0002-9379-7527

Sandor Jordan https://orcid.org/0000-0002-3556-4127

Gabor Sramko https://orcid.org/0000-0001-8588-6362

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research, 25* (17), 3389-3402. doi:10.1093/nar/25.17.3389

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics, 17* (2), 81-92. doi:10.1038/nrg.2015.28

Anisimova, M., & Gascuel, O. (2006). Approximate Likelihood-Ratio test for branches: a fast, accurate, and powerful alternative. *Systematic Biology, 55* (4), 539–552. doi:10.1080/10635150600755453

Avise, J. C. (2000).*Phylogeography: the history and formation of species* . Cambridge: Harvard University Press.

Avise, J. C. (2004). *Molecular markers, natural history, and evolution* (2nd ed.). Sunderland: Sinauer Associates Publisher.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., . . . Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE, 3* (10), e3376. doi:10.1371/journal.pone.0003376

Barnard-Kubow, K. B., Debban, C. L., & Galloway, L. F. (2015). Multiple glacial refugia lead to genetic structuring and the potential for reproductive isolation in a herbaceous plant. *American Journal of Botany, 102* (11), 1842-1853. doi:10.3732/ajb.1500267

Bentley, N., Grauke, L. J., & Klein, P. (2019). Genotyping by sequencing (GBS) and SNP marker analysis of diverse accessions of pecan (*Carya illinoinensis*). *Tree Genetics & Genomes, 15* (1), 8. doi:10.1007/s11295-018-1314-5

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution, 10* (14), 7585-7601. doi:10.1002/ece3.6483

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

Brito, P. H., & Edwards, S. V. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica, 135* (3), 439-455. doi:10.1007/s10709-008-9293-3

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: building and genotyping loci *de novo* from short-read sequences.*G3: Genes, Genomes, Genetics, 1* (3), 171–182. doi:10.1534/g3.111.000240

Chafin, T. K., Martin, B. T., Mussmann, S. M., Douglas, M. R., & Douglas, M. E. (2018). FRAGMATIC: in silico locus prediction and its utility in optimizing ddRADseq projects.*Conservation Genetics Resources, 10* (3), 325–328. doi:10.1007/s12686-017-0814-1

Clugston, J. A. R., Kenicer, G. J., Milne, R., Overcast, I., Wilson, T. C., & Nagalingum, N. S. (2019). RADseq as a valuable tool for plants with large genomes—A case study in cycads. *Molecular Ecology Resources, 19* (6), 1610-1622. doi:10.1111/1755-0998.13085

Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., & Mockler, T. (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Research, 36* (19), e122.

D'Agostino, N., Taranto, F., Camposeo, S., Mangini, G., Fanelli, V., Gadaleta, S., . . . Montemurro, C. (2018). GBS-derived SNP catalogue unveiled wide genetic variability and geographical relationships of Italian olive cultivars.*Scientific Reports, 8* (1), 15877. doi:10.1038/s41598-018-34207-y

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . 1000 Genomes, P. A. G. (2011). The variant call format and VCFtools. *Bioinformatics, 27* (15), 2156–2158. doi:10.1093/bioinformatics/btr330

Daru, B. H., Manning, J. C., Boatwright, J. S., Maurin, O., Maclean, N., Schaefer, H., . . . van der Bank, M. (2013). Molecular and morphological analysis of subfamily Alooideae (Asphodelaceae) and the inclusion of *Chortolirion* in*Aloe* . *Taxon, 62* (1), 62–76. doi:10.1002/tax.621006

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013). Special features of RAD Sequencing data: implications for genotyping.*Molecular Ecology, 22* (11), 3151-3164. doi:10.1111/mec.12084

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics, 12* (7), 499-510. doi:10.1038/nrg3012

Du, Z.-Y., Harris, A. J., & Xiang, Q.-Y. (2020). Phylogenomics, co-evolution of ecological niche and morphology, and historical biogeography of buckeyes, horsechestnuts, and their relatives (Hippocastaneae, Sapindaceae) and the value of RAD-Seq for deep evolutionary inferences back to the Late Cretaceous.*Molecular Phylogenetics and Evolution, 145* , 106726. doi:10.1016/j.ympev.2019.106726

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

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32* (5), 1792-1797. doi:10.1093/nar/gkh340

Ekblom, R., Smeds, L., & Ellegren, H. (2014). Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics, 15* (1), 467. doi:10.1186/1471-2164-15-467

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS ONE, 6* (5), e19379. doi:10.1371/journal.pone.0019379

Feng, C., Xu, M., Feng, C., von Wettberg, E. J. B., & Kang, M. (2017). The complete chloroplast genome of Primulina and two novel strategies for development of high polymorphic loci for population genetic and phylogenetic studies.*BMC Evolutionary Biology, 17* (1), 224. doi:10.1186/s12862-017-1067-z

Forsman, Z. H., Knapp, I. S. S., Tisthammer, K., Eaton, D. A. R., Belcaid, M., & Toonen, R. J. (2017). Coral hybridization or phenotypic variation? Genomic data reveal gene flow between *Porites lobata* and *P. compressa* .*Molecular Phylogenetics and Evolution, 111* , 132-148. doi:10.1016/j.ympev.2017.03.023

Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics, 34* (1), 397-423. doi:10.1146/annurev.ecolsys.34.011802.132421

Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S., & Prins, P. (2021). Vcflib and tools for processing the VCF variant call format. *bioRxiv* , 2021.2005.2021.445151. doi:10.1101/2021.05.21.445151

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing.*arXiv preprint arXiv:1207.3907* .

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhue, C., Pudlo, P., . . . Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology, 22* (11), 3165-3178. doi:10.1111/mec.12089

Herrera, S., & Shank, T. M. (2016). RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa.*Molecular Phylogenetics and Evolution, 100* , 70-79. doi:10.1016/j.ympev.2016.03.010

Hohenlohe, P. A., Hand, B. K., Andrews, K. R., & Luikart, G. (2019). Population genomics provides key insights in ecology and evolution. In O. P. Rajora (Ed.),*Population Genomics: Concepts, Approaches and Applications* (pp. 483-510). Cham: Springer International Publishing.

Holliday, J. A., Hallerman, E. M., & Haak, D. C. (2019). Genotyping and sequencing technologies in population genetics and genomics. In O. P. Rajora (Ed.), *Population Genomics: Concepts, Approaches and Applications* (pp. 83-125). Cham: Springer International Publishing.

Knox, E. B. (2014). The dynamic history of plastid genomes in the Campanulaceae *sensu lato* is unique among angiosperms. *Proceedings of the National Academy of Sciences, 111* (30), 11097-11102. doi:10.1073/pnas.1403363111

Leache, A. D., & Oaks, J. R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics, 48* (1), 69–84. doi:10.1146/annurev-ecolsys-110316-022645

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., & Rieseberg, L. H. (2019). An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus). *New Phytologist, 221* (1), 515-526. doi:10.1111/nph.15386

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* , 1303.3997v1302.

Luikart, G., Kardos, M., Hand, B. K., Rajora, O. P., Aitken, S. N., & Hohenlohe, P. A. (2019). Population Genomics: advancing understanding of nature. In O. P. Rajora (Ed.),*Population Genomics: Concepts, Approaches and Applications* (pp. 3–79). Cham: Springer International Publishing.

Macher, J.-N., Rozenberg, A., Pauls, S. U., Tollrian, R., Wagner, R., & Leese, F. (2015). Assessing the phylogeographic history of the montane caddisfly *Thremma gallicum*using mitochondrial and restriction-site-associated DNA (RAD) markers.*Ecology and Evolution, 5* (3), 648-662. doi:10.1002/ece3.1366

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics, 24* (3), 133-141. doi:10.1016/j.tig.2007.12.007

Mariac, C., Scarcelli, N., Pouzadou, J., Barnaud, A., Billot, C., Faye, A., . . . Couvreur, T. L. P. (2014). Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources, 14* (6), 1103-1113. doi:10.1111/1755-0998.12258

Martin-Hernanz, S., Aparicio, A., Fernandez-Mazuecos, M., Rubio, E., Reyes-Betancort, J. A., Santos-Guerra, A., . . . Albaladejo, R. G. (2019). Maximize resolution or minimize error? Using genotyping-by-sequencing to investigate the recent diversification of *Helianthemum* (Cistaceae).*Frontiers in Plant Science, 10* (1416). doi:10.3389/fpls.2019.01416

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., & Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and Evolution, 66* (2), 526–538. doi:10.1016/j.ympev.2011.12.007

McVay, J. D., Hipp, A. L., & Manos, P. S. (2017). A genetic legacy of introgression confounds phylogeny and biogeography in oaks. *Proceedings of the Royal Society B: Biological Sciences, 284* (1854), 20170300. doi:10.1098/rspb.2017.0300

Meger, J., Ulaszewski, B., Vendramin, G. G., & Burczyk, J. (2019). Using reduced representation libraries sequencing methods to identify cpDNA polymorphisms in European beech (Fagus sylvatica L). *Tree Genetics & Genomes, 15* (1), 7. doi:10.1007/s11295-018-1313-6

Melo, A. T. O., & Hale, I. (2018). Expanded functionality, increased accuracy, and enhanced speed in the de novo genotyping-by-sequencing pipeline GBS-SNP-CROP.*Bioinformatics, 35* (10), 1783-1785. doi:10.1093/bioinformatics/bty873

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics, 11* (1), 31-46. doi:10.1038/nrg2626

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

Moura, A. E., Kenny, J. G., Chaudhuri, R. R., Hughes, M. A., Reisinger, R. R., de Bruyn, P. J. N., . . . Hoelzel, A. R. (2015). Phylogenomics of the killer whale indicates ecotype divergence in sympatry. *Heredity, 114* (1), 48-55. doi:10.1038/hdy.2014.67

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics.*Molecular Ecology, 22* (11), 2841–2847. doi:10.1111/mec.12350

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. *Frontiers in Plant Science, 10* (1074). doi:10.3389/fpls.2019.01074

Parks, M., Cronn, R. C., & Liston, A. (2009). Increasing phylogenetic resolution at low taxonomic levels using massively parallel sequencing of chloroplast genomes. *BMC Biology, 7* , 84. doi:10.1186/1741-7007-7-84

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (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. doi:10.1371/journal.pone.0037135

Puckett, E. E., Etter, P. D., Johnson, E. A., & Eggert, L. S. (2015). Phylogeographic Analyses of American Black Bears (Ursus americanus) Suggest Four Glacial Refugia and Complex Patterns of Postglacial Admixture. *Molecular Biology and Evolution, 32* (9), 2338-2350. doi:10.1093/molbev/msv114

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

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology, 23* (24), 5937–5942. doi:10.1111/mec.12965

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics, 26* (6), 841-842. doi:10.1093/bioinformatics/btq033

R Core, T. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rieseberg, L. H., & Soltis, D. E. (1991). Phylogenetic consequences of cytoplasmic gene flow in plants.*Evolutionary Trends in Plants, 5* (1), 65–84.

Rieseberg, L. H., Whitton, J., & Linder, C. R. (1996). Molecular marker incongruence in plant hybrid zones and phylogenetic trees. *Acta Botanica Neerlandica, 45* (3), 243-262.

Rincon-Sandoval, M., Betancur-R, R., & Maldonado-Ocampo, J. A. (2019). Comparative phylogeography of trans-Andean freshwater fishes based on genome-wide nuclear and mitochondrial markers. *Molecular Ecology, 28* (5), 1096-1115. doi:10.1111/mec.15036

Rivera-Colon, A. G., Rochette, N. C., & Catchen, J. M. (2021). Simulation with RADinitio improves RADseq experimental design and sheds light on sources of missing data.*Molecular Ecology Resources, 21* (2), 363-378. doi:10.1111/1755-0998.13163

Rochette, N. C., Rivera-Colon, A. G., & Catchen, J. M. (2019). Stacks 2: Analytical methods for paired-end sequencing improve RADseq-based population genomics. *Molecular Ecology, 28* (21), 4737-4754. doi:10.1111/mec.15253

Salas-Leiva, D. E., Meerow, A. W., Calonje, M., Griffith, M. P., Francisco-Ortega, J., Nakamura, K., . . . Namoff, S. (2013). Phylogeny of the cycads based on multiple single-copy nuclear genes: congruence of concatenated parsimony, likelihood and species tree inference methods. *Annals of Botany, 112* (7), 1263-1278. doi:10.1093/aob/mct192

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., & Gerstein, M. B. (2011). The real cost of sequencing: higher than you think! *Genome Biology, 12* (8), 125. doi:10.1186/gb-2011-12-8-125

Scheunert, A., & Heubl, G. (2014). Diversification of *Scrophularia* (Scrophulariaceae) in the Western Mediterranean and Macaronesia – Phylogenetic relationships, reticulate evolution and biogeographic patterns. *Molecular Phylogenetics and Evolution, 70* (0), 296–313. doi:10.1016/j.ympev.2013.09.023

Schlotterer, C. (2004). The evolution of molecular markers — just a matter of fashion? *Nature Reviews Genetics, 5* (1), 63-69. doi:10.1038/nrg1249

Seelanan, T., Schnabel, A., & Wendel, J. F. (1997). Congruence and consensus in the Cotton tribe (Malvaceae). *Systematic Botany, 22* (2), 259–290. doi:10.2307/2419457

Sharples, M. T., & Tripp, E. A. (2019). Phylogenetic relationships within and delimitation of the cosmopolitan flowering plant genus *Stellaria* L. (Caryophyllaceae): core stars and fallen stars. *Systematic Botany, 44* (4), 857-876. doi:10.1600/036364419X15710776741440

Soltis, D. E., & Soltis, P. S. (1998). Choosing an approach and an appropriate gene for phylogenetic analysis. In D. E. Soltis, P. S. Soltis, & J. J. Doyle (Eds.),*Molecular Systematics of Plants II* (pp. 1–42): Springer US.

Sramko, G., Molnar V, A., Toth, J. P., Laczko, L., Kalinka, A., Horvath, O., . . . Popiela, A. (2016). Molecular phylogenetics, seed morphometrics, chromosome number evolution and systematics of European Elatine L. (Elatinaceae) species.*PeerJ, 4* , e2800. doi:10.7717/peerj.2800

Stobie, C. S., Cunningham, M. J., Oosthuizen, C. J., & Bloomer, P. (2019). Finding stories in noise: Mitochondrial portraits from RAD data. *Molecular Ecology Resources, 19* (1), 191-205. doi:10.1111/1755-0998.12953

Storer, C., Payton, A., McDaniel, S., Jordal, B., & Hulcr, J. (2017). Cryptic genetic variation in an inbreeding and cosmopolitan pest, *Xylosandrus crassiusculus* , revealed using ddRADseq. *Ecology and Evolution, 7* (24), 10974-10986. doi:10.1002/ece3.3625

Straub, S. C. K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R. C., & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany, 99* (2), 349-364. doi:10.3732/ajb.1100335

Streicher, J. W., Devitt, T. J., Goldberg, C. S., Malone, J. H., Blackmon, H., & Fujita, M. K. (2014). Diversification and asymmetrical gene flow across time and space: lineage sorting and hybridization in polytypic barking frogs.*Molecular Ecology, 23* (13), 3273-3291. doi:10.1111/mec.12814

Sutherland, B. L., & Galloway, L. F. (2018). Effects of glaciation and whole genome duplication on the distribution of the *Campanula rotundifolia* polyploid complex.*American Journal of Botany, 105* (10), 1760-1770. doi:10.1002/ajb2.1162

Terraneo, T. I., Arrigoni, R., Benzoni, F., Forsman, Z. H., & Berumen, M. L. (2018). Using ezRAD to reconstruct the complete mitochondrial genome of *Porites fontanesii* (Cnidaria: Scleractinia). *Mitochondrial DNA Part B, 3* (1), 173-174. doi:10.1080/23802359.2018.1437805

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology, 21* (16), 3907-3930. doi:10.1111/j.1365-294X.2012.05664.x

Tonti-Filippini, J., Nevill, P. G., Dixon, K., & Small, I. (2017). What can we do with 1000 plastid genomes? *The Plant Journal, 90* (4), 808-818. doi:10.1111/tpj.13491

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ, 1* , e203. doi:10.7717/peerj.203

Uckele, K. A., Adams, R. P., Schwarzbach, A. E., & Parchman, T. L. (2021). Genome-wide RAD sequencing resolves the evolutionary history of serrate leaf*Juniperus* and reveals discordance with chloroplast phylogeny.*Molecular Phylogenetics and Evolution, 156* , 107022. doi:10.1016/j.ympev.2020.107022

Uncu, A. O., Uncu, A. T., Celık, İ., Doganlar, S., & Frary, A. (2015). A primer to molecular phylogenetic analysis in plants. *Critical Reviews in Plant Sciences, 34* (4), 454-468. doi:10.1080/07352689.2015.1047712

Wendel, J., & Doyle, J. (1998). Phylogenetic incongruence: Window into genome history and molecular evolution. In D. Soltis, P. Soltis, & J. Doyle (Eds.), *Molecular Systematics of Plants II* (pp. 265–296). London: Chapman & Hall.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis* . New York: Springer-Verlag.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution, 8* (1), 28-36. doi:10.1111/2041-210X.12628

Zink, R. M., & Barrowclough, G. F. (2008). Mitochondrial DNA under siege in avian phylogeography.*Molecular Ecology, 17* (9), 2107-2121. doi:10.1111/j.1365-294X.2008.03737.x

Supporting information

Additional supporting information may be found online in the Supporting Information section.

Tables and Figures

Table 1. Summary of the datasets collected from the literature to benchmark our pipeline to assemble organellar loci. NCBI accession numbers are given for Bioproject and Nucleotide Database (reference genome). Minimal coverage is the coverage of a locus set in any individual to be included in the analysis.

Study system	Bioproject	Scope	Protocol	Enzyme(s)
g. *Paragorgia*	PRJNA317473	phylogeny	sdRAD	PstI
Porites spp.	PRJNA380807	hybridization	ezRAD	MboI & Sau3AI
g. *Labeobarbus*	PRJNA493727	phylogeny & phylogeography	ddRAD	MluCI & NlaIII
Xylosandrus crassiusculus	PRJNA342041	phylogeography	ddRAD	EcoRI & MseI
g. *Melicope*	PRJNA559258	phylogeny	sdRAD	SbfI
g. *Helianthemum*	PRJNA573639	phylogeny	GBS	ApeKI
Cycadales	PRJNA526348	phylogeny	ezRAD	EcoRI & MseI
g. *Stellaria*	PRJNA547948 & PRJNA473254	phylogeny	ddRAD	EcoRI & MseI

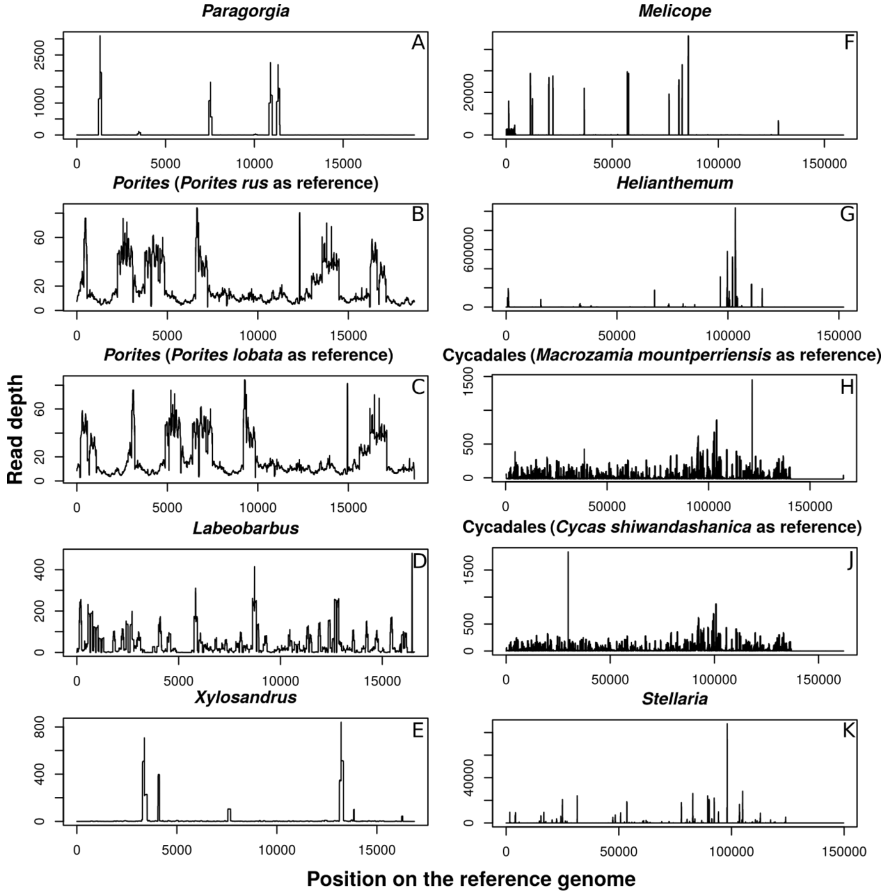

Figure 1. Mean read depth along the reference sequences of the benchmarking datasets. The horizontal axis represents the position on the reference genome, whereas the vertical axis shows the read depth at each site. Continuous alignments without read depth dropping to zero represent a locus.

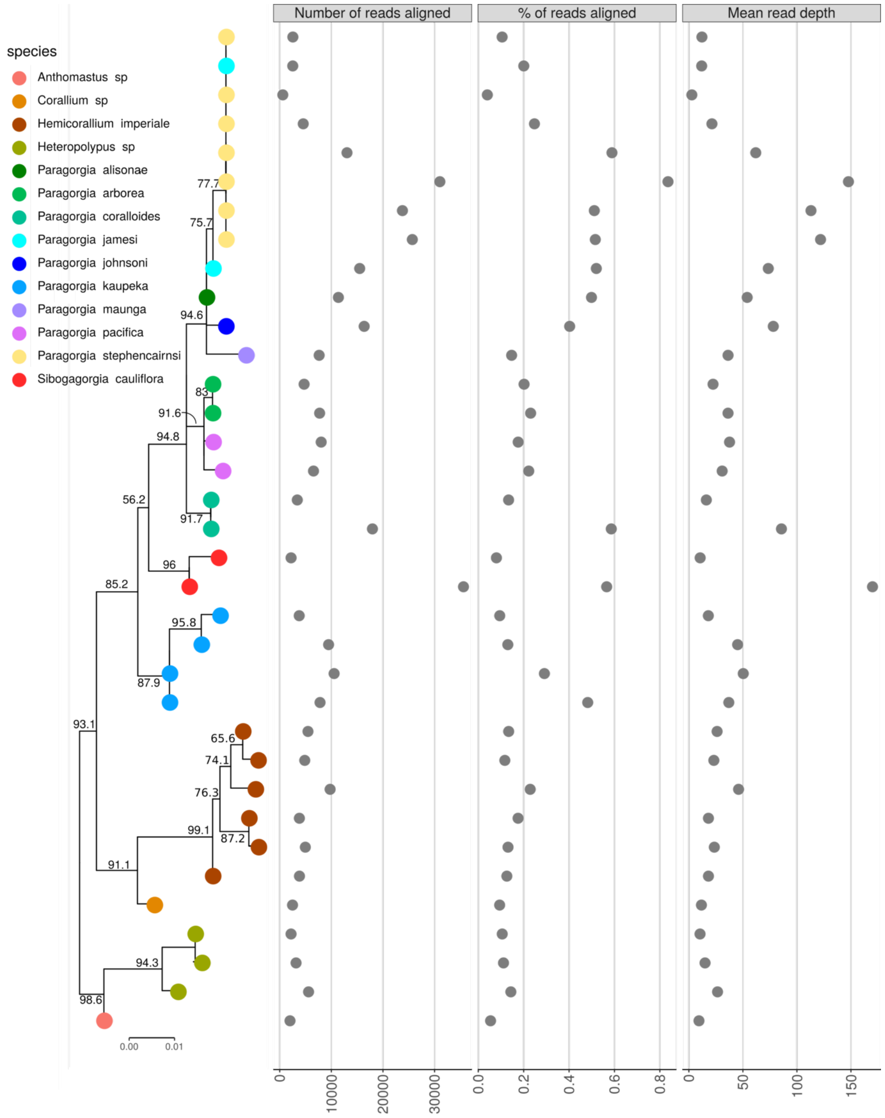

Figure 2. Phylogenetic tree reconstruction and the main alignment statistic of the *Paragorgia* dataset. For an unedited phylogenetic tree, please check **Figure S1** . Figure legend represents the population map used for base calling.

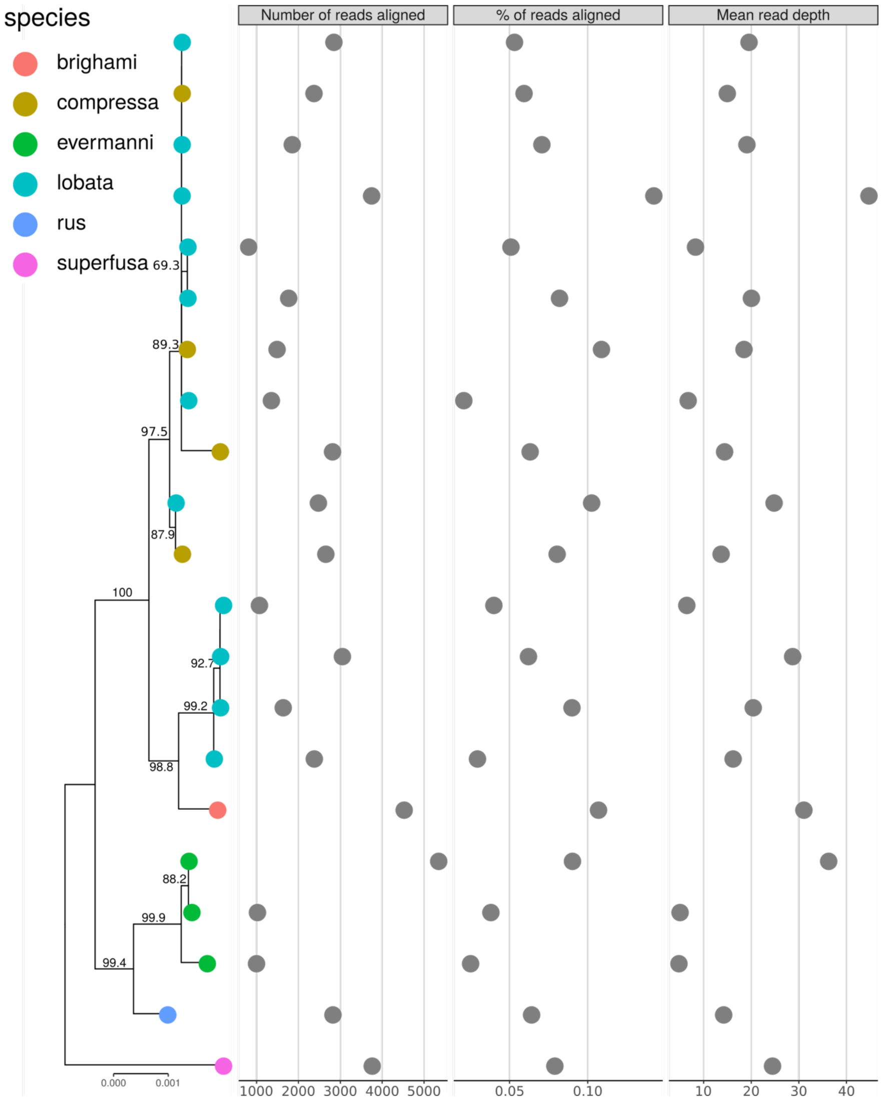

Figure 3. Phylogenetic tree reconstruction and the main alignment statistic of the *Porites* dataset. For an unedited phylogenetic tree, please check **Figure S2** . Figure legend represents the population map used for base calling.

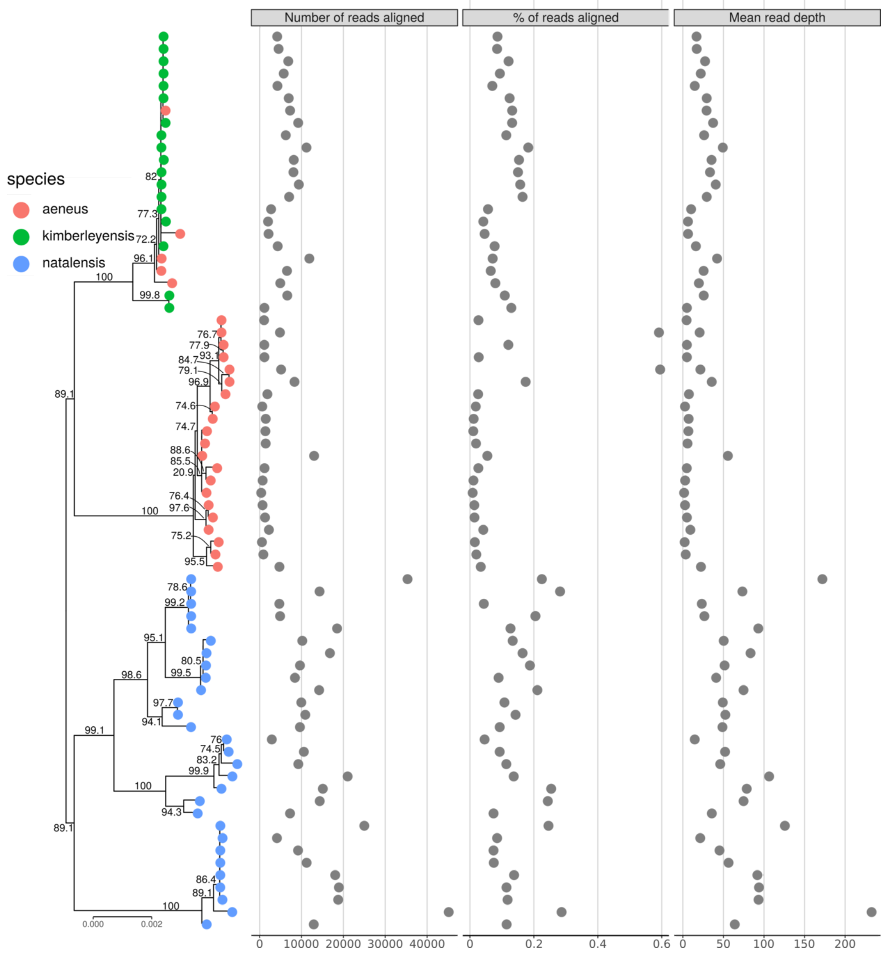

Figure 4. Phylogenetic tree reconstruction and the main alignment statistic of the *Labeobarbus* dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check**Figure S3** . Figure legend represents the population map used for base calling.

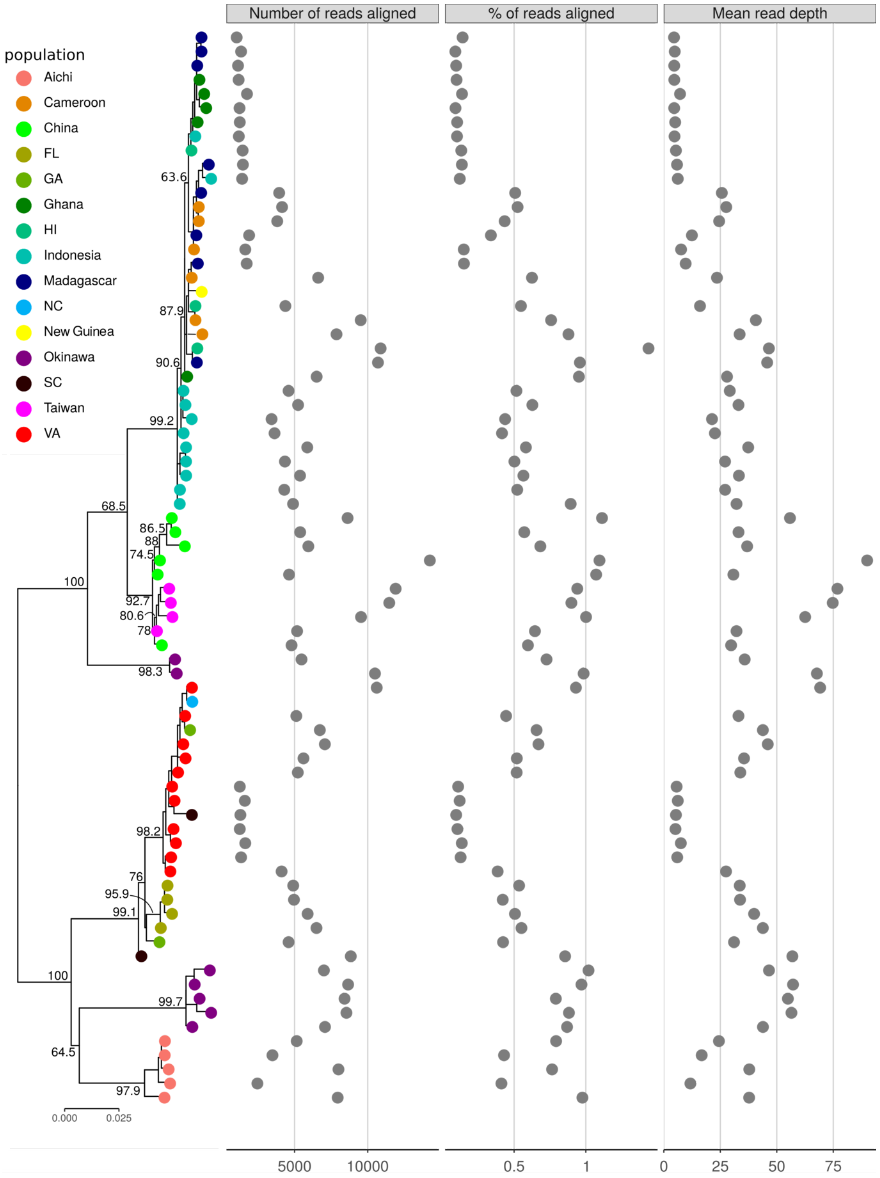

Figure 5. Phylogenetic tree reconstruction and the main alignment statistic of the *Xylosandrus* dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check **Figure S4** . Figure legend represents the population map used for base calling.

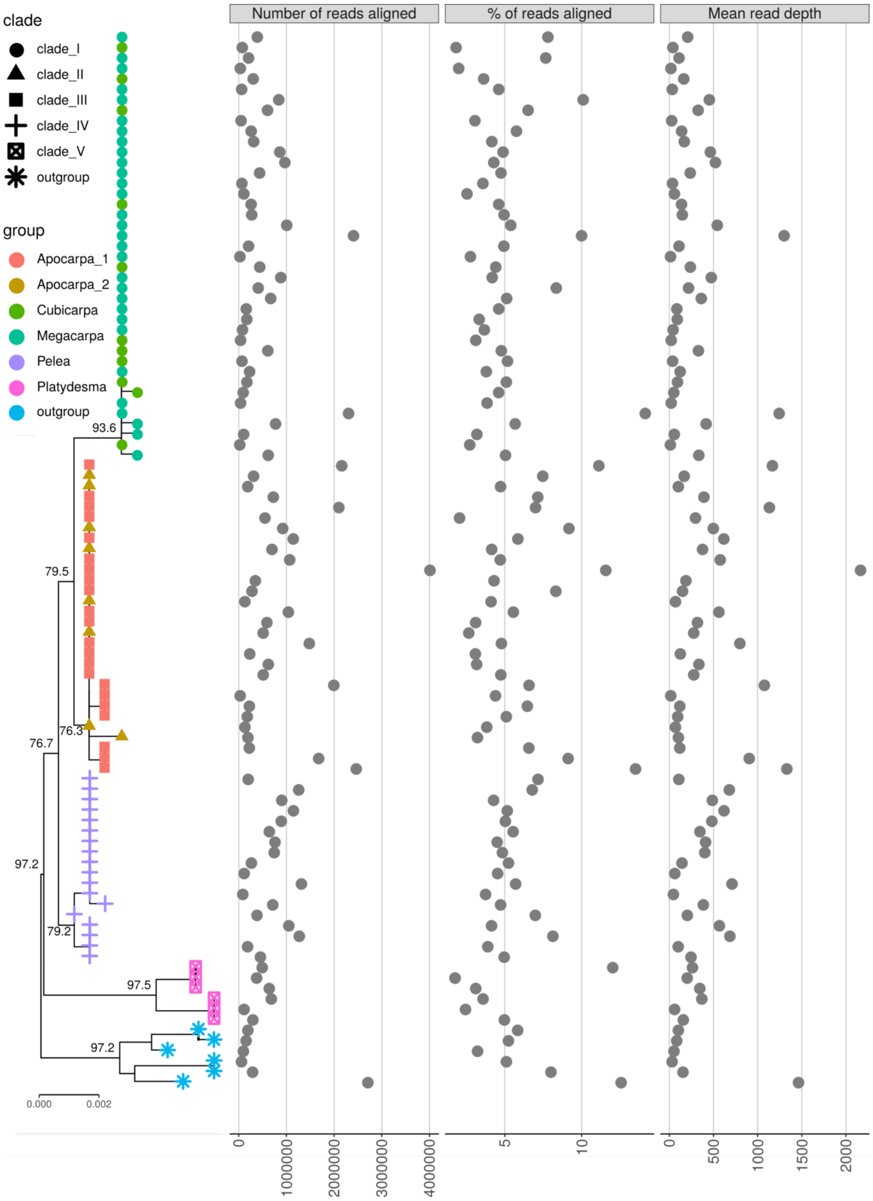

Figure 6. Phylogenetic tree reconstruction and the main alignment statistic of the *Melicope* dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check **Figure S5** . Figure legend represents the population map used for base calling.

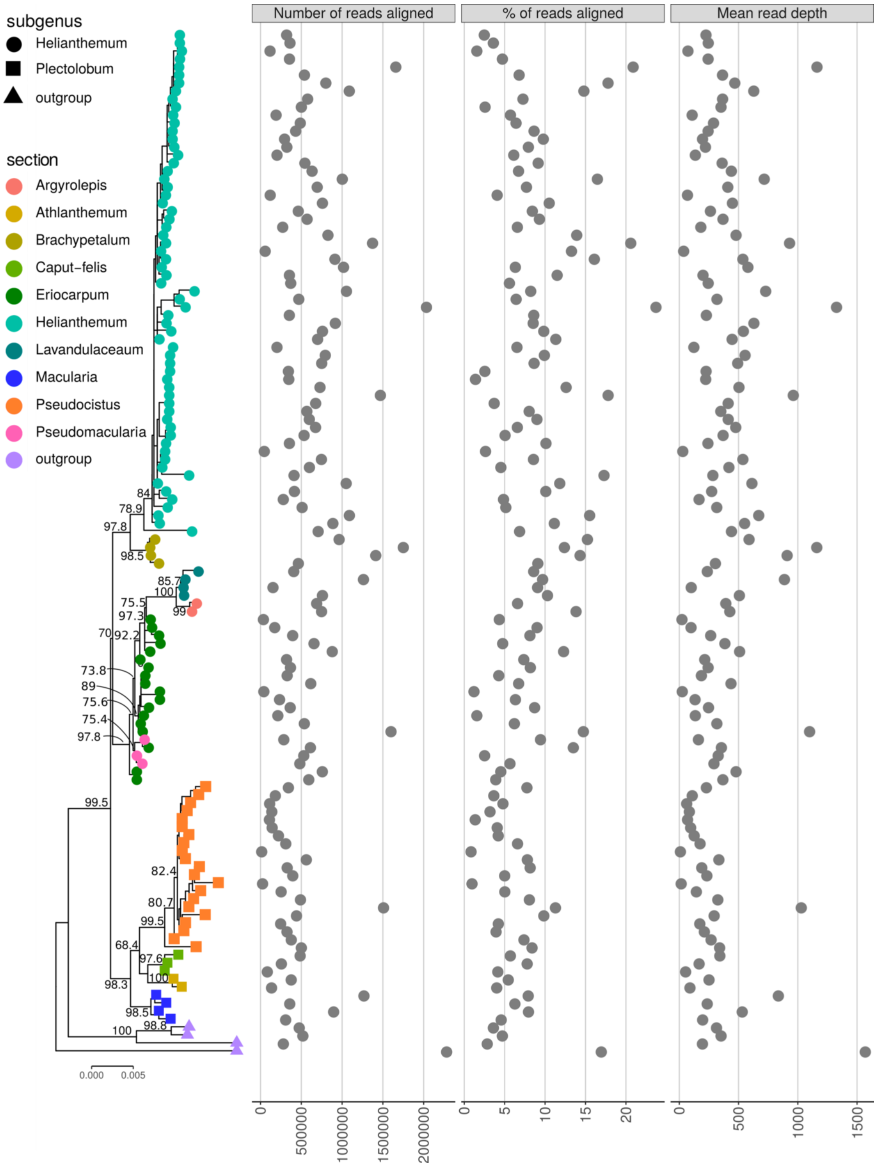

Figure 7. Phylogenetic tree reconstruction and the main alignment statistic of the *Helianthemum* dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check **Figure S6** . Figure legend represents the population map used for base calling.

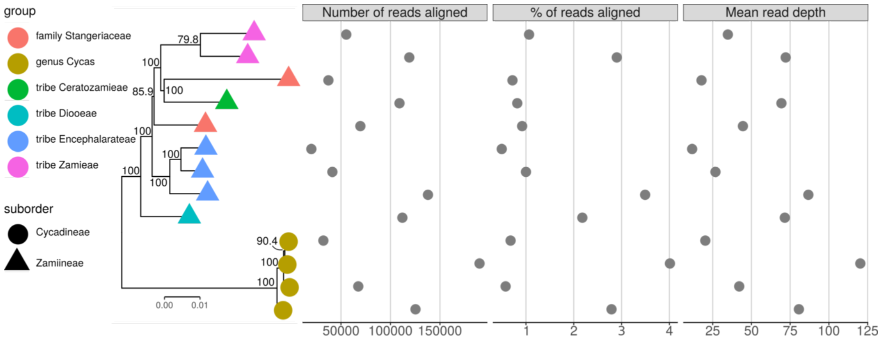

Figure 8. Phylogenetic tree reconstruction and the main alignment statistic of the cycads dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check **Figure S7** . Figure legend represents the population map used for base calling.

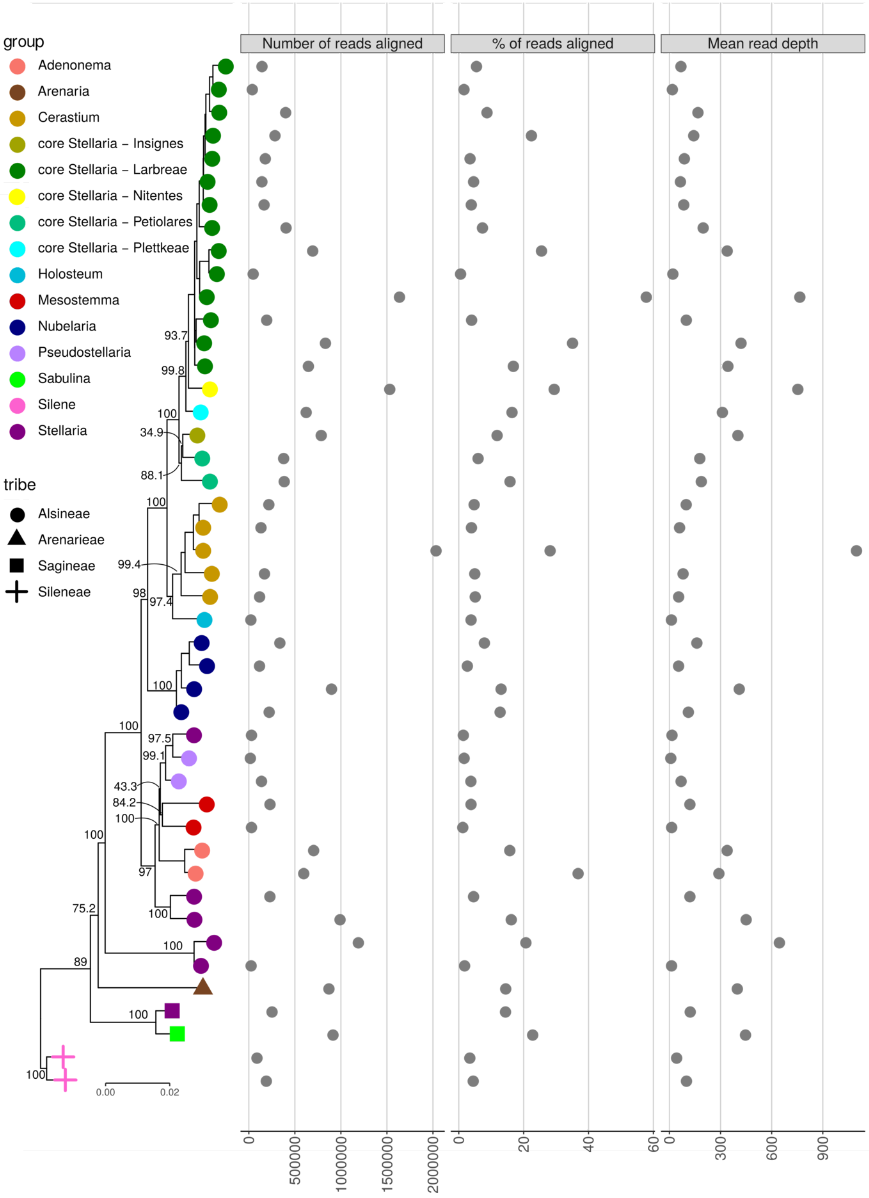

Figure 9. Phylogenetic tree reconstruction and the main alignment statistic of the broad *Stellaria* dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check **Figure S8** . Figure legend represents the population map used for base calling.