Received: 6 January 2022 | Accepted: 3 June 2022

DOI: 10.1111/2041-210X.13937

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

The RadOrgMiner pipeline: Automated genotyping of organellar loci from RADseq data

Levente Laczkó[1,2,3,4] | Sándor Jordán[2,3,5] | Gábor Sramkó[1,2,3]

[1]MTA-DE "Lendület" Evolutionary Phylogenomics Research Group, Debrecen, Hungary

[2]Department of Botany, University of Debrecen, Debrecen, Hungary

[3]ELKH-DE Conservation Biology Research Group, Debrecen, Hungary

[4]Department of Metagenomics, University of Debrecen, Debrecen, Hungary

[5]Juhász-Nagy Pál Doctoral School, University of Debrecen, Debrecen, Hungary

Correspondence
Gábor Sramkó
Email: sramko.gabor@science.unideb.hu

Funding information
University of Debrecen; The 'OTKA' Young Researcher Excellence Program, Grant/Award Number: FK 137962; New National Excellence Program of the Hungarian Ministry of Innovation and Technology, Grant/Award Number: ÚNKP-20-4-I-DE-290

Handling Editor: Francisco Balao

Abstract

1. Different versions of Restriction-site-Associated DNA sequencing (RADseq) have become powerful and popular tools in molecular ecology. Although RADseq datasets are generally regarded as representative of the nuclear genome, reduced representation genomic libraries may also sample the organellar (i.e. the mitochondrial and, in the case of plants, the plastid) DNA. This cytoplasmic genetic variance provides a better understanding of evolutionary history by uncovering past hybridisation and identifying maternal or, rarely, paternal lineage due to rapid lineage sorting.
2. We developed a pipeline that is based on existing bioinformatic tools to automatically mine and genotype organellar loci from RADseq libraries. The efficacy 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 mitochondrial and plastid loci, which were subject to phylogenetic tree reconstruction.
3. In all cases, organellar phylogenies adequately supplemented the original studies by corroborating the large-scale picture based on RADseq or by bringing additional evidence on past or contemporary hybridisation. RADseq methods designed to achieve larger horizontal coverage (i.e. ddRAD, ezRAD) yielded longer organellar alignments, but sdRAD and GBS still provided valuable polymorphic organellar loci at no additional sequencing effort.
4. Our newly developed pipeline can be run under a Unix-like operating system and is freely accessible at https://doi.org/10.5281/zenodo.6619190

KEYWORDS
bioinformatics, cytonuclear discordance, hybridisation, phylogenetic incongruence, phylogeography, reduced representation library

1 | INTRODUCTION

Reduced complexity or reduced genomic representation library (RRL) approaches made the in-depth study of micro-evolutionary processes feasible giving rise to ecological or population genomics (Luikart et al., 2019; Narum et al., 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 et al., 2019) are from the group of Restriction-site-Associated DNA sequencing (RADseq) approaches (Andrews et al., 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). The RRL approaches that include a restriction enzyme to obtain DNA sequence information from a large set of loci at the genome level can be collectively called a RADseq group (Andrews et al., 2016). Later, numerous variants of RADseq protocols, relying on type II restriction enzymes to sample genomic diversity, have been developed (reviewed by Andrews et al., 2016; Rivera-Colón et al., 2021), with each of them having particular strengths for a given purpose or being disadvantageous in certain ways by providing different horizontal (genome coverage) and vertical coverage (read depth) (Davey et al., 2013; Elshire et al., 2011; Hohenlohe et al., 2019; Narum et al., 2013; Peterson et al., 2012; Puritz, Matz, et al., 2014; Toonen et al., 2013).

Typically, RADseq methods sample the whole genome anonymously (i.e. no a priori information is known regarding the origins of the genome-wide reads; Andrews et al., 2016; Hohenlohe et al., 2019). Nevertheless, irrespective of the bioinformatic pipeline used, the ascertained loci are typically treated as samples from the nuclear genome. Only a few pipelines can take the presence of organellar RAD tags into consideration and can be fine-tuned for haploid data such as ipyrad (Eaton & Overcast, 2020). However, in the reference genome-based analysis of D'Agostino et al. (2018), only 53.6% of the GBS tags aligned uniquely to the closely related nuclear reference genome. These authors hypothesised the unmapped sequences might contain organellar tags with high frequency, which partly, but not exclusively, could reduce the ratio of tags aligned to the reference genome used in their study. This suggests the frequency of organellar reads in RRLs to be more significant than previously thought, although their representation is limited by the cut site frequency of the given restriction enzyme in a particular organellar genome (Bentley et al., 2019). Nevertheless, previous studies have shown that even partially represented organellar genome obtained by RADseq can provide additional insight into the genetic composition of the studied organisms (Meger et al., 2019; Stobie et al., 2019).

Genetic information from the organelles has long been utilised in phylogenetics and phylogeography as sources of non-recombining, haploid, uniparentally inherited genomic compartments (Avise, 2000, 2004; Soltis & Soltis, 1998; Uncu et al., 2015) that often show a correlated structure with geography. Thus, organellar DNA has been the most important source of phylogeographical analyses until recently (Brito & Edwards, 2009; McCormack et al., 2013) owing to its one-fourth effective population size compared to the nuclear genome that leads to rapid lineage sorting (Schaal & Olsen, 2000). In addition, the 'cytonuclear discordance' (Rieseberg & Soltis, 1991) or 'mito-nuclear' discordance (Funk & Omland, 2003; Toews & Brelsford, 2012) can open a window into hybridisation via phylogenetic incongruence (Wendel & Doyle, 1998) between the nuclear and organellar genes.

Comparison of RADseq datasets and organellar datasets within the same study organism has gained some popularity and has usually uncovered hybridisation between lineages (Barnard-Kubow et al., 2015; Macher et al., 2015; Moura et al., 2015; Puckett et al., 2015; Streicher et al., 2014; Sutherland & Galloway, 2018; Uckele et al., 2021). In these studies, however, the RADseq dataset was regarded as a 'representative' of the nuclear genome, and the organellar dataset was obtained by additional sequencing experiments. Nevertheless, RRL approaches may also sample the organellar genome and can potentially be used to sort out organellar reads from nuclear ones (Forsman et al., 2017; Meger et al., 2019; Stobie et al., 2019; Terraneo et al., 2018). In this case, additional information from the RADseq dataset is provided by separating organellar reads from the nuclear ones without further sequencing effort. The utility of this approach, what we may term as 'organellar mining', was only addressed in a handful of studies (Clugston et al., 2019; Du et al., 2020; Feng et al., 2017; Forsman et al., 2017; McVay et al., 2017; Meger et al., 2019; Pujolar et al., 2014; Terraneo et al., 2018). All these studies either use existing software—such as Stacks (Rochette et al., 2019) as applied by Stobie et al. (2019) or ipyRAD (Eaton & Overcast, 2020) utilised by Du et al. (2020) or GATK (McKenna et al., 2010) as demonstrated by Meger et al. (2019)—to sort reads from the different genomes, which are sometimes proprietary solutions (e.g. Geneious as applied by Terraneo et al., 2018). Just a few of them take the unique properties of organellar tags into consideration (e.g. Stobie et al., 2019) and only Feng et al. (2017) proposed a ready-to-use tool designed explicitly for the assembly of organellar genomes using paired-end RADSeq data.

Although the number of phylogenetic and phylogeographical 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 we explicitly designed to sort out and genotype organellar loci found in datasets generated by the RADseq group of experiments. Our pipeline is a command line tool compatible with most UNIX-like operating systems and allows subsequent comparison of genetic information coming from the organellar and the nuclear genome without additional sequencing effort. We also demonstrate the performance of our software solution by re-analysing eight publicly available RADseq datasets (Table 1) that span different levels of phylogenetic divergence.

2 | MATERIALS AND METHODS

2.1 | RADOrgMiner pipeline description

Our pipeline uses existing bioinformatic tools to screen RADseq reads if they align well to a closely related organellar genome. It separates the organellar reads from the non-organellar ones, then genotypes loci using the sequence of the aligned reads (Figure 1). The pipeline uses two main steps. In the first step, we align all the reads to a closely related organellar reference genome using bwa 0.7.17 (Li, 2013) then separate reads that can be aligned with samtools 1.10.2 (Li et al., 2009). 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), the alignment of reads originating from loci in the inverted repeat can increase the number of ambiguous alignments (i.e. a read can be aligned to different genomic regions with the same mapping quality) inflating the ratio of missing data in the final dataset. Optional masking of one of the repeat regions can be done to reduce the number of ambiguous alignments in the dataset. Thus, the masking of the repeats makes the genotyping of loci located in the inverted repeat reliable if the sequences of repeats are assumed to be identical. Location of the inverted repeats is identified by self-blasting using blastn 2.10.1+ (Altschul et al., 1997). By self-blasting, we mean conducting a blast search of the query sequence against itself. The longest matching sequence will be the whole reference genome in this configuration. If an inverted repeat is present in the organellar (plastid) reference genome, the second and third highest scoring pairs (HSPs) should be the inverted repeat sequences. These repeats should have the same length and high sequence similarity (100% if the two repeat regions are identical). The automatic masking selects the second HSP and masks the genomic location of the subject sequence with N-s. At the second step, aligned reads originating from the organelle(s) are retained for haplotype calling, whereas unaligned (non-organellar) reads are saved and available for downstream analyses. To minimise 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 2). Nuclear sequences are expected to be present with a lower read depth relative to the organellar genome in genomic datasets (Ekblom et al., 2014). Similarly, organellar reads have a relatively high read depth compared to nuclear loci also in RAD libraries (e.g. Clugston et al., 2019). When filtering for the minimum depth of organellar loci, we assume that purely nuclear loci with a high sequence similarity to the organellar genome (i.e. NUMTs and NUPTs without their corresponding 'original' organellar sequence sampled) would have a lower read depth than 'true' organellar loci. Although the copy number of NUMTs and NUPTs can reach a few thousand copies (Richly & Leister, 2004a; Richly & Leister, 2004b), their effect on the read depth is supposedly lower than the effect of organellar genome copy number. The copy number of the

TABLE 1 Summary of the datasets collected from the literature to benchmark our pipeline to assemble organellar loci. NCBI accession numbers are given for the Bioproject (RADseq dataset) and the Nucleotide (cytoplasmic reference genome) collections

Study system	Bioproject	Scope	Protocol	Enzyme(s)	Sequencing type	Target genome	Reference genome(s)	Reference
g. *Paragorgia*	PRJNA317473	Phylogeny	sdRAD	*Pst*I	SE Illumina HiSeq 2000	Mitochondrion	KF785801	Herrera and Shank (2016)
Porites spp.	PRJNA380807	Hybridisation	ezRAD	*Mbo*I & *Sau3A*I	PE Illumina MiSeq	Mitochondrion	KU572435 & NC_027526	Forsman et al. (2017)
g. *Labeobarbus*	PRJNA493727	Phylogeny & phylogeography	ddRAD	*Mlu*CI & *Nla*III	PE Illumina HiSeq 2000	Mitochondrion	KX419437	Stobie et al. (2019)
Xylosandrus crassiusculus	PRJNA342041	Phylogeography	ddRAD	*Eco*RI & *Mse*I	SE Illumina NextSeq 500	Mitochondrion	NC_036284	Storer et al. (2017)
g. *Melicope*	PRJNA559258	Phylogeny	sdRAD	*Sbf*I	SE Illumina GAIIx	Plastid	MW046256	Paetzold et al. (2019)
g. *Helianthemum*	PRJNA573639	Phylogeny	GBS	*Ape*KI	PE Illumina HiSeq 2000	Plastid	MK776534	Martín-Hernanz et al. (2019)
Cycadales	PRJNA526348	Phylogeny	ezRAD	*Eco*RI & *Mse*I	PE Illumina NextSeq	Plastid	LC049069 & MT876215	Clugston et al. (2019)
g. *Stellaria*	PRJNA547948 & PRJNA473254	Phylogeny	ddRAD	*Eco*RI & *Mse*I	SE Illumina HiSeq 2000 & NextSeq 500	plastid	NC_044183	Sharples and Tripp (2019)

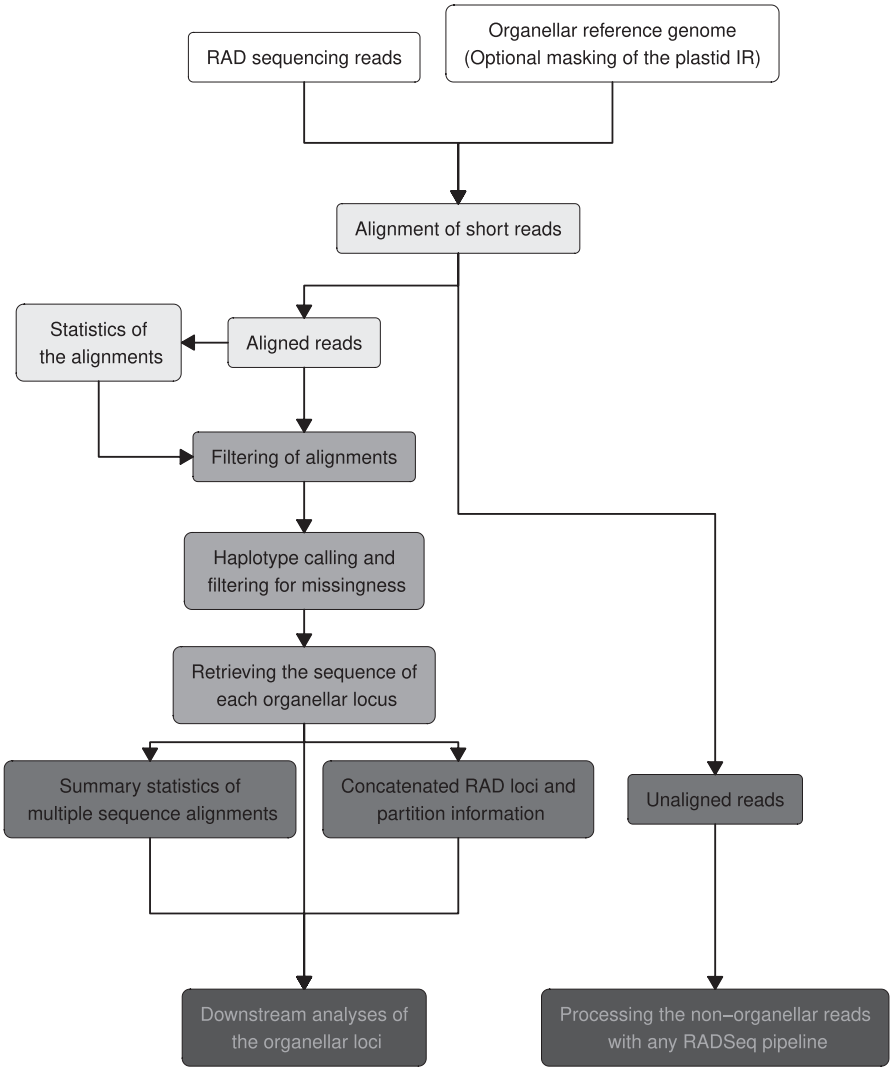

FIGURE 1 Schematic representation of the pipeline presented in this study. Boxes with different background colour correspond to different stages of the analyses. White boxes show the input data required by RADOrgMiner. Light grey boxes represent the alignment step of the pipeline, whereas middle grey boxes show the haplotype calling steps. Towards the bottom of the figure, increasingly darker colours show the output of RADOrgMiner and potential downstream analyses to be performed outside the presented pipeline.

whole plastid genome can range to ten thousands (Bendich, 1987) per cell depending on the organism and tissue type. A similar over-representation can also be observed for the mitochondria when energetically active tissue, such as muscle, is used for the DNA isolation (Ekblom & Wolf, 2014). Consequently, in most cases, the read depth of organellar RAD tags would be higher than the read depth of low-copy nuclear loci and NUMTs/NUPTs. Nevertheless, the copy number of NUPTs can be highly variable, with some organisms bearing only one plastid per cell (Richly & Leister, 2004b). For these cases, we introduced an option to filter for the maximum read depth of loci with a default value of one million. If the read depth of NUMTs or NUPTs is known to be high relative to the organellar genome, it is recommended to set this parameter to exclude nuclear RAD tags with a high copy number. Setting a threshold on the minimum and maximum read depth of loci prior to variant calling can help minimise the number of falsely aligned NUPTs and NUMTs and narrow the dataset to the loci most likely of cytoplasmic origin. 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 an individual locus. In light of organellar loci's supposedly high read depth, they might be visualised as 'spikes' along the reference genome as a function of read depth (Figure 2). The advantage of this approach is that haplotype calling of loci can be parallelised to drastically decrease the run time of this step.

We chose freebayes for its high and easy customisability 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 the 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. The constraint on the number of best alleles is an arbitrarily chosen threshold and aims to decrease the running time of the haplotype calling when a high read depth can be observed. It is worth highlighting this high actual value of the alternate allele frequency required in this step. If NUPTs and NUMTs can be found in the dataset with a nearly identical sequence to the organellar copy, they can be expected to be present with a lower frequency. The least diverged NUMTs and NUPTs are the youngest copies of organellar sequences transferred

TABLE 2 Summary statistics of organellar read alignments and organellar loci filtered from the re-analysed datasets using RADOrgMiner. Minimal coverage is the coverage of a locus set in any individual to be included in the analysis. For the phylogenetic reconstruction only, those loci were used that showed at least one informative site

Study system	Number of reads aligned to the reference (mean)	Proportion of organellar reads (%)	Minimal coverage	Mean read depth of organellar loci	Number of polymorphic (and informative) sites	Number of organellar loci (informative loci)	Percent of missing sites (%)	Length of loci in base pairs (mean)	Alignment length of organellar loci (reference genome coverage)
g. *Paragorgia*	492–35,494 (9051)	0.03–0.83 (0.26)	500	2.25–169.5	76 (57)	4 (4)	10.6	177–178 (177)	709 (4%)
Porites spp. (using *P. rus* as reference)	786–5332 (2384.48)	0.02–0.14 (0.07)	3	4.7–44.6	190 (98)	1 (1)	9.03	18,646 (18,646)	18,646 (100%)
Porites spp. (using *P. lobata* as reference)	786–5332 (2384.48)	0.02–0.14 (0.07)	3	4.7–44.6	188 (98)	1 (1)	9.03	18,646 (18,646)	18,646 (100%)
g. *Labeobarbus*	264–45,047 (8194,7)	0.006–0.59 (0.12)	10	0.92–232	363 (288)	7 (7)	8.34	285–3999 (1972.3)	13,806 (83.3%)
Xylosandrus crassiusculus	1006–14,203 (5132.75)	0.08–1.3 (0.52)	100	4.25–89.75	179 (174)	7 (7)	22.3	100–395 (251.3)	1508 (8.9%)
g. *Melicope*	9213–3,997,452 (609,756.4)	1.73–14 (5.36)	5000	4.93–2159	45 (31)	13 (11)	0.13	175–177 (176.9)	2300 (1.4%)
g. *Helianthemum*	7752–2,274,688 (560,541.4)	0.77–23.64 (7.9)	10,000	4.5–1564.9	385 (220)	12 (11)	9.5	112–884 (437.75)	5455 (4.36%)
Cycadales (using *Macrozamia mountperriensis* as reference)	19,551–189,194 (85,373.2)	0.48–3.99 (1.65)	100	11.4–119.92	11,541 (5369)	103 (102)	25.83	109–6675 (1037.93)	106,907 (64.2%)
Cycadales (using *Cycas shiwandashanica* as reference)	17,523–193,328 (81,480)	0.43–4.0 (1.58)	100	9.93–133.35	9166 (4893)	109 (108)	27.3	119–7382 (923)	100,617 (62.1%)
g. *Stellaria*	10,191–2,027,645 (444,062)	0.37–57.63 (11.62)	1000	4.78–1094.3	1397 (887)	46 (46)	15.2	101–413 (192)	8837 (5.9%)

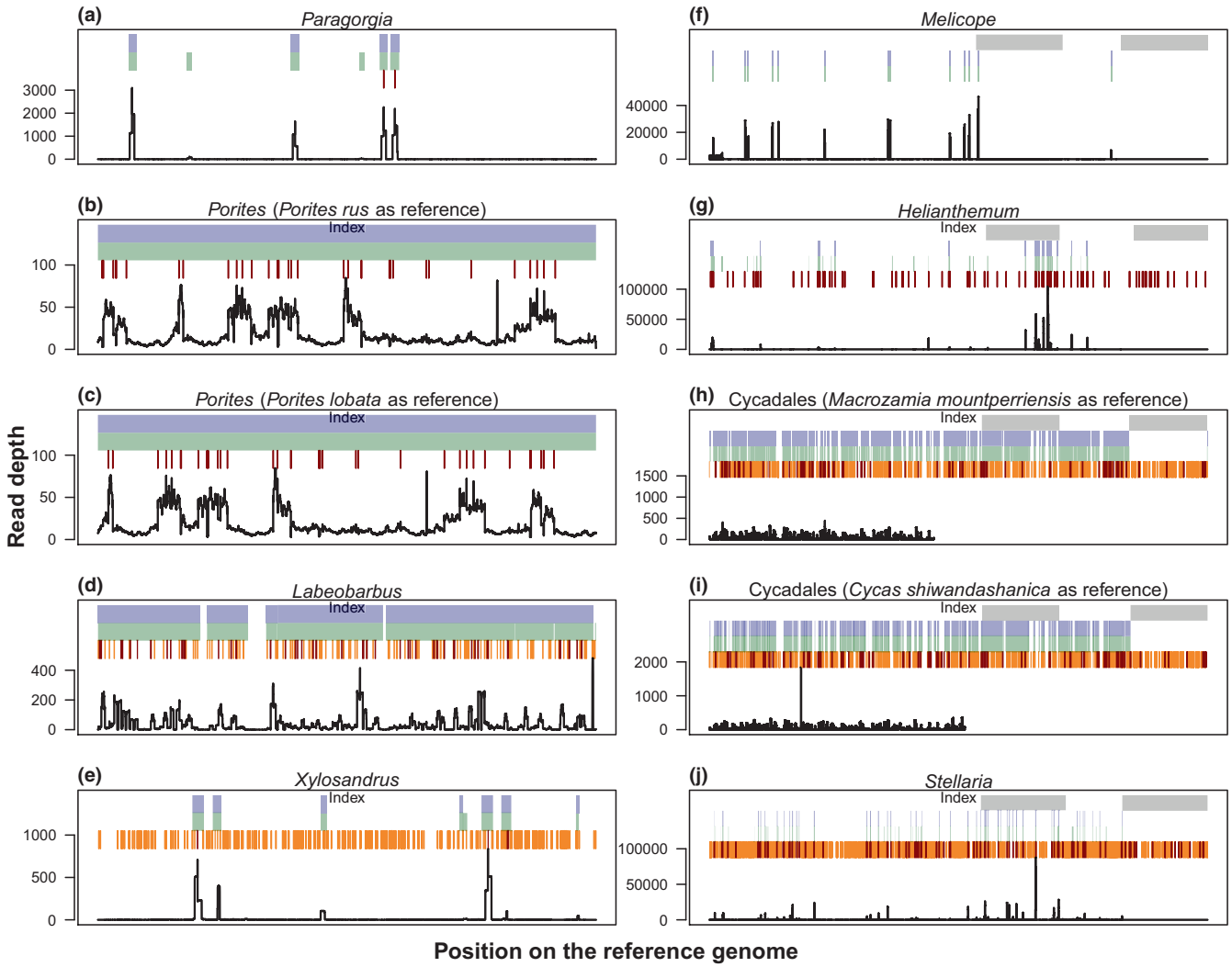

FIGURE 2 Mean read depth along the reference sequences of the benchmarking datasets: genus *Paragoria* (2a), genus *Porites* using the reference genome of *P. rus* (2b), genus *Porites* using the reference genome of *P. lobata* (2c), genus *Labeobarbus* (2d), species *Xylosandrus crassiusculus* (2e), genus *Melicope* (2f), genus *Helianthemum* (2g), order Cycadales using the reference genome of *Macrozamia mountperriensis* (2h), order Cycadales using the reference genome of *Cycas shiwandashanica* (2i), genus *Stellaria* (2j). 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. Dark red and orange bars show the location of restriction enzyme cut sites applied by the original authors. For the *Porites* dataset, where both enzymes have the same cut site, the rare and frequent cutter is not distinguished. For the other studies that used two restriction enzymes, the cut site of the rare cutter is marked with dark red, whereas the cut site of the frequent cutter is marked with orange. Light green bars show the location of organellar loci identified by minimum read depth and light blue bars represent loci that were filtered through our pipeline and could be used in downstream analyses. For plant species, a light grey bar show the location of the plastid inverted repeat. It is worth noting that no cut sites were found in silico for the *Melicope* dataset and all loci filtered out could be linked to the star-activity of *Sbf*I. Differences of cut site sequences compared to the recognition site of the enzyme could be found only at the 5′ or 3′ end of the cut site as revealed by the alignment of reads of the *Melicope* samples.

to the nuclear genome (Michalovova et al., 2013). After the incorporation of organellar sequences in the nuclear genome, the copy number of NUMTs and NUPTs increases by post-insertion duplication (Michalovova et al., 2013). In turn, young and least diverged NUMTs and NUPTs sharing a cut site with the organellar genome would be present with a low frequency if the organellar and nuclear sequences are sampled in the RAD library. Setting a minimum alternative allele frequency coupled with constraining the variant call to consider haploid data should effectively exclude the nuclear variants from the final dataset. Setting this value not only decreases the number of sequencing errors in the final dataset, but also helps to eliminate variants from the haplotype calls that originate from NUPTs and NUMTs. When NUMTs or NUPTs with a high copy number can be observed in the dataset, the value of alternate allele frequency could be set higher to effectively exclude the nuclear polymorphisms of nucleus-transferred genomic regions. All the above settings can be changed from the command line, allowing fine-tuning of the genotyping for a given dataset. Species, or if multiple populations can

be analysed within a species, populations can be used as a prior for the population-based Bayesian inference model that will be partitioned by the groups supplied for the pipeline. We set freebayes to use mapping quality for likelihood calculation with clumping of haplotypes disabled, priors on and Hardy–Weinberg equilibrium turned off. Binomial observation priors are turned off, and read placement probability, strand balance probability and read position probability are used instead. Since freebayes is capable of ploidy-aware base calls, ploidy is set to one by default. All sites are annotated, including the monomorphic ones, and are exported into a vcf file. Missing data, arising mainly at the sheared ends in certain RADseq experiments, are filtered with vcftools 0.1.16 (Danecek et al., 2011) by allowing a maximum of 20% missingness across all individuals as default. In case of a low number of overlapping loci, the analysis can be narrowed down to include only samples with a given minimum number of base-pairs and/or reads sequenced. Vcf files are converted to fasta with vcf2fasta from the vcflib 1.0 package (Garrison et al., 2021) and aligned with muscle 3.8.1 (Edgar, 2004). As vcf2fasta uses the reference genome for vcf conversion, the reference is subset 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 (i.e. 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 base pairs (bp) using the AMAS 1.0 PYTHON package (Borowiec, 2016). Removing partial alignments that failed to yield an organellar locus (i.e. shorter than the specified length) aims to exclude fragmented loci. NUPTs and NUMTs can be prone to fragmentation by transposable genetic elements and tend to decay over time (Michalovova et al., 2013). As a consequence of fragmentation, we can assume that some of these pseudo-organellar loci will have a shorter final alignment length than expected by the read length. Thus, this filtering option can be an additional approach to exclude falsely recovered nuclear loci.

As a proof of concept for excluding NUPTs and NUMTs, we set up an experiment that called the variants of the organellar genomes again without fixing ploidy at one and assuming pooled sequencing in the benchmark studies described below. All other analysis parameters were left at default or set as given in Table 2. In this experiment, we investigated the allele balance (the ratio of reads showing the reference allele to all reads; AB) over the whole dataset and evaluated the frequency of 'non-haploid' observations, then assessed the number of individuals a 'non-haploid' (i.e. multiple variants can be supported by the observations) variant occurs in. Here we assume that haploid loci of the organellar genomes should show only one alternative allele compared to the reference sequence. Without constraining the ploidy level, if more than one probable allele can be observed at a given site, we expect an AB value larger than 0 for that particular observation. Consequently, if the ratio of NUMTs and NUPTs after applying all the above filters remained high in the dataset, we would expect them to inflate the heterozygosity at given sites resulting in higher AB values. A drawback of this approach could be its inability to differentiate between NUMTs/NUPTs and heteroplasmy. To account for the frequency of 'non-haploid' variants over loci and individuals, AB values were calculated separately for each variant of each sample and treated as independent observations of AB values.

The pipeline can be parameterised from the command line for easy and reproducible usability. All runs were conducted in a Debian 10.1 environment. We avoided inclusion of proprietary software to boost the transparency and reproducibility of this approach. The pipeline with the list of dependencies, installation instructions, documentation and example runs are available at: https://github.com/laczkol/RADOrgMiner.

2.2 | Demonstration of the pipeline

We demonstrate the power of our software solution for mining out organellar reads generated in a RADseq experiment by reanalysing publicly available datasets representing various variants of RADseq and 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 the 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. We randomly selected datasets from the literature to represent the mitochondrion and the chloroplast. We focused on representing different RAD flavours instead of the scope of the study. For plant study systems, we analysed only the plastid RAD tags as in botany most molecular phylogenetic studies rely on the variability of the plastome and there is much greater availability of reference plastomes than mitochondria.

Datasets of the eight studies were downloaded and used as input to our pipeline with the references specified (Table 1). In the cases of the Cycadales and the *Porites* datasets, we used two different reference genomes to assess the level of filtering robustness (Tables 1 and 2). For the *Stellaria* dataset, we only analysed samples that belong to the 'broad' taxonomic range (Sharples & Tripp, 2019), which consisted of fewer samples. We inspected the mean and individual read depth for each case study and set this value to output the highest number of base pairs without increasing the amount of missing data. The genotyped organellar loci were subject to phylogenetic tree reconstruction with IQtree 2.0.3 (Minh et al., 2020). The phylogenetic reconstruction used the whole sequences of loci (i.e. polymorphic and constant sites) and relied only on those loci that showed at least one informative site (Table 2). The initial partitioning scheme was set to consider all loci as a different data partition. We used the automatic model selection of IQtree (MFP+MERGE) to apply the optimal substitution model and partitioning scheme to the dataset. We calculated the approximate likelihood ratio test (aLRT) (Anisimova & Gascuel, 2006) branch support values after 1000 replications.

Results obtained by our analysis were interpreted in the light of the original authors' results. We defined statistical support for branch robustness as existing if aLRT ≥ 80%. We visualised 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 Team, 2012), ggplot 2 3.3.4 (Wickham, 2016) and ggtree 2.0.4 (Yu et al., 2017), and further edited in Inkscape 0.92 (https://inkscape.org/) to improve readability.

3 | RESULTS

3.1 | Organellar DNA content of RADseq libraries

All studies we used to demonstrate the utility of our pipeline contained organellar RAD tags. However, horizontal and vertical coverage seemed to be variable not only between different RAD flavours but even within datasets (Figure 2 and Table 2). Some parameter adjustments were applied to gain further insight into the performance of our pipeline. The first analysis of *Xylosandrus* samples yielded no loci with a maximum of 20% missing data (the default setting of our pipeline). After checking the read depth distribution of the samples, a large number of samples with non-overlapping loci was noticeable. To analyse 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 constraint decreased the number of samples from 198 to 76. The default value of missingness resulted in a scattered final alignment for the *Melicope* dataset; thus, we decreased the maximum amount of missingness to 10%.

In the case of the mitochondrial datasets, *Porites* (ezRAD) samples had the lowest proportion of reads aligned to the reference but had the highest horizontal coverage. *Xylosandrus* (ddRAD) samples showed the highest proportion of reads aligned to the reference yielding the second lowest coverage, whereas the *Paragorgia* (sdRAD) dataset covered the lowest proportion of the mitochondrion. Despite this, the *Paragorgia* dataset contained a higher number of mitochondrial reads than the *Labeobarbus* (ddRAD) dataset that covered the second highest proportion of the reference genome after the *Porites* dataset (Table 2).

The Cycadales (ezRAD) dataset covered the highest proportion of the reference plastome sequence but had the lowest proportion of organellar reads. The *Stellaria* (ddRAD) dataset covered the second highest proportion of the reference plastome and showed the highest number of organellar reads of the plastid datasets. In the *Helianthemum* (GBS) dataset, we discovered lower horizontal coverage and proportion of organellar reads compared to the *Stellaria* dataset. The *Melicope* dataset—that showed only loci linked to restriction enzyme star activity (Figure 2f), as the 5′ or 3′ end of the cut site was different from what was recognised by *Sbf*I—yielded the second lowest proportion of organellar reads and the lowest proportion of reference plastome coverage.

3.2 | Phylogenetic reconstruction and frequency of 'non-haploid' variants

All datasets yielded RADseq-derived organellar loci suitable for downstream analyses. We describe the re-analysis of the *Labeobarbus* and the Cycadales dataset here. For the detailed phylogenetic results of all re-analysed datasets reconstructed using organellar RAD loci, please consult the Supporting Information of this study.

3.2.1 | The *Labeobarbus* dataset

The phylogenetic reconstruction placed *Labeobarbus natalensis* on a distinct clade (Figure 3 and Figure S5). *L. 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 (La005, LnBL004) if not counting the missing data (Figure S5).

3.2.2 | The Cycadales dataset

The phylogenetic analysis reconstructed the same phylogenetic tree with nearly equal support values regardless of organellar reference genome used; thus, we only present our results (Figure 4 and Figure S12) using *M. mountperriensis* that was based on a longer alignment covering 64.26% of the reference plastome with more polymorphic sites (Table 2). 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*).

The frequency of 'non-haploid' variants was low in all cases (Figure S15). Relative to all independent observations of AB values, variants with multiple probable alleles could be discovered occasionally with a similar read depth to the reference allele. These variants with an AB > 0 occurred mostly in one individual and could be rarely discovered in a handful of samples and never dominated the dataset. Removal of 'non-haploid' variant positions from the first variant call did not change the outcome of the phylogenetic reconstructions (results not shown).

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 variable even within a given dataset (Table 2), genotype calls were consistent according to the taxonomy of the samples (i.e. a priori

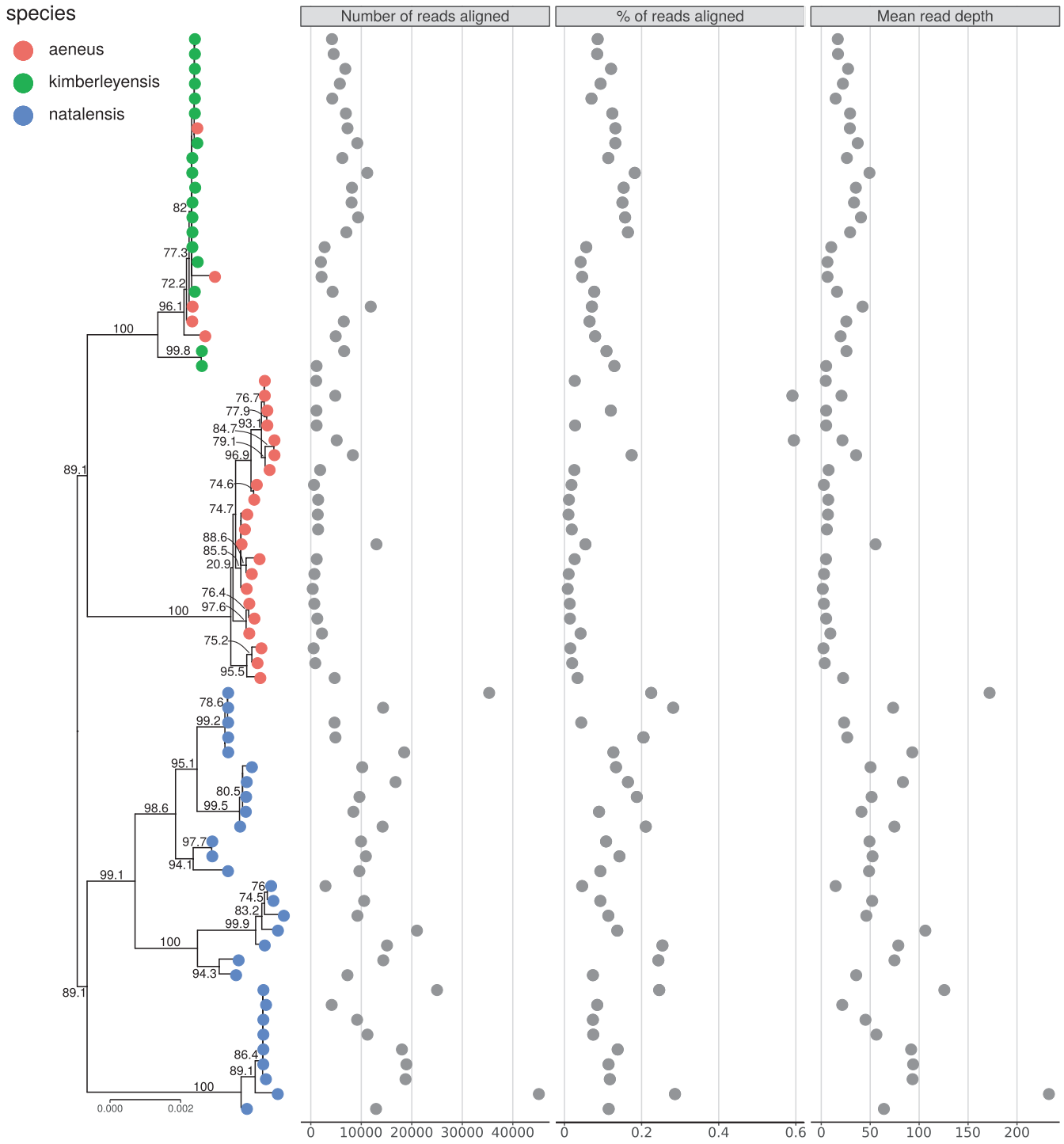

FIGURE 3 Phylogenetic tree reconstruction and the main alignment statistic of the *Labeobarbus* dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check Figure S5. Figure legend represents the population map used for base calling.

similar groups could be clustered together at various taxonomic levels). Unlike tools used by previous authors to mine for organellar loci from RADseq datasets (e.g. Forsman et al., 2017; Stobie et al., 2019), our pipeline is highly specialised and thus automatised for this task. Some RADseq analysis pipeline can be tailored to effectively assemble organellar RAD loci. To mention a few, ipyrad (Eaton & Overcast, 2020) can use a (organellar) reference genome to reconstruct the RAD loci and the ploidy of the dataset can also be constrained. Similarly, the dDocent (Puritz, Hollenbeck, et al., 2014) pipeline can be applied by specifying a reference genome sequence and, like RADOrgMiner, uses freebayes for the reconstruction of haplotypes. Stacks (Rochette et al., 2019), regardless of conducting a de novo or reference-based reconstruction of RAD loci, expects diploid data and is highly suitable for the

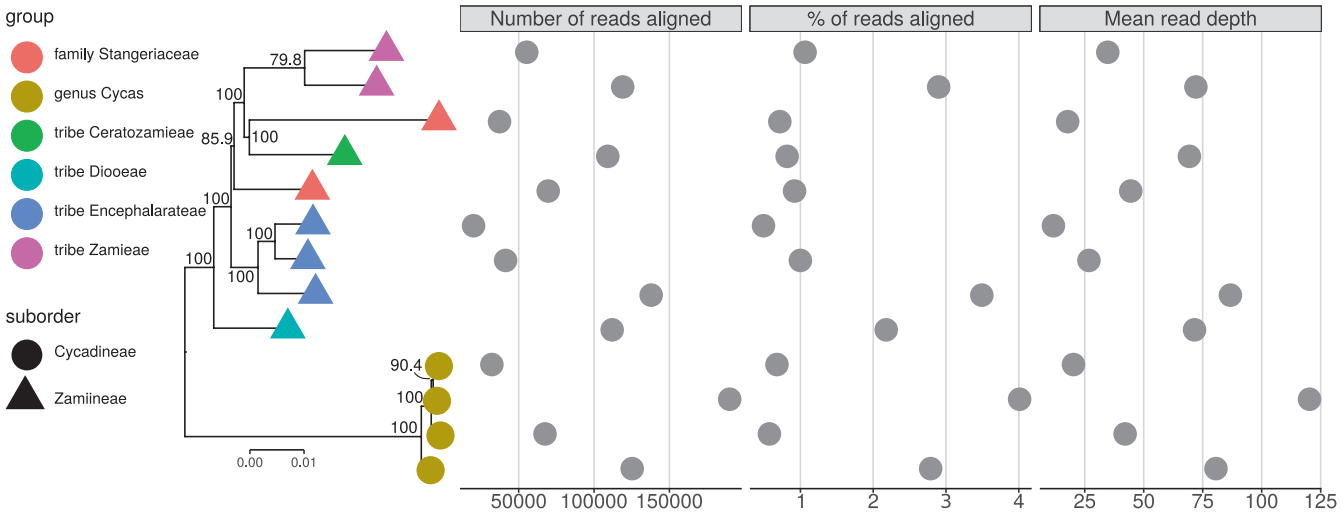

FIGURE 4 Phylogenetic tree reconstruction and the main alignment statistic of the Cycadales dataset. Some short branches are not annotated for better readability. For an unedited phylogenetic tree, please check Figure S12. Figure legend represents the population map used for base calling.

population genomic analysis of genome wide SNPs. Still, as shown by Stobie et al. (2019), the genotypes obtained by these pipelines can be a reliable representation of haplotypes, but the obtained haplotypes might need some manual data curation. Unlike the abovementioned tools, our pipeline aims to take the unique features of organellar RAD tags—such as read depth and the possible presence and filtering of NUMTs and NUPTs—into consideration and by providing statistics on the short read alignments an informed decision on the analysis parameters can be made, thus eliminating the need of manual data curation and increasing the reliability of organellar haplotypes mined from RADseq datasets.

4.1 | 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 (Figures 3 and 4; Figures S1–S14) adequately supplement the original studies' findings either by corroborating the large-scale picture based on the RADseq-derived results of the original authors or by bringing additional evidence on hybridisation. Our approach revealed some variants (Figure S15) that show more than one probable allele. Given that the alternative alleles were present in a similarly high read depth to the variant called when setting ploidy to 1, they are likely to originate from the organelles and may represent heteroplasmy, sequencing error or even chimerisation of RAD tags. We regard this as proof of our filtering approach removing NUPTs and NUMTs effectively.

Below, we briefly evaluate the results from the re-analyses of the published datasets (Tables 1 and 2), which reflect the strengths and limitations of our pipeline.

Organellar DNA can show limited variability, as in the case of the *Paragorgia* dataset. This may influence the phylogenetic tree reconstruction, which placed *Sibogagorgia cauliflora* and *Paragorgia kaupeka* differently in our analysis. The low level of polymorphisms of the *Porites* dataset explains why using two reference genomes for read alignment yielded identical phylogenetic results. The different haplotypes of some samples (Figures S3 and S4) can be deduced to the application of different pipelines (i.e. Geneious [Biomatters Inc.] as applied by Forsman et al. (2017) or RADOrgMiner).

The re-analysis of the *Labeobarbus* dataset recovered the three main clades of the original authors, and the general structure within clades was concordant with the results of Stobie et al. (2019). Moreover, we identified the same number of hybrid individuals as the original authors. The low error rate of technical replicates—that did not influence the outcome of the phylogenetic analysis—suggests our pipeline to be at least as accurate as the method presented by Stobie et al. (2019).

The drastically different organellar read depth distribution within the *Xylosandrus* dataset highlights the potential technical limitations in the application of our pipeline: overlapping loci can only be mined from a smaller proportion of the individuals in such cases. This uneven distribution of read depth can 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 (Figures S6 and S7) with those presented by Storer et al. (2017).

The example of the *Melicope* dataset proves that ancient hybridisation events can be effectively detected by supplementing the nuclear SNPs with cytoplasmic sequence data. Paetzold et al. (2019), who used a standard RADseq pipeline, divided members of the *Apocarpa* group into two distinct clades and demonstrated an ancient introgression event between lineages based virtually on the nuclear genome. This introgression event is corroborated by our organellar results that cluster all samples of the *Apocarpa* group on the same clade. Similarly, Martín-Hernanz et al. (2019) concluded on the importance of hybridisation in the diversification of *Helianthemum*.

The monophyly of *Helianthemum* reconstructed by organellar RAD loci supports this observation (Figures S10 and S11).

The phylogenetic relationships reconstructed for Zamiaceae are fully compatible with the published results (Salas-Leiva et al., 2013), where only two plastid genes were used. This suggests the high accuracy of our pipeline and an unequivocal phylogenetic signal across the plastome. Despite the high evolutionary distance, this dataset was not sensitive to allele dropout (Andrews et al., 2016), which can be explained by the low sequence variability of the group.

Whereas the phylogenetic structure within core *Stellaria* showed different relationships at the tip of the tree, still, the major lineages, including core *Stellaria* and closely related tribes, could be identified with high certainty in the ddRAD-derived phylogenomic dataset of the original work (Sharples & Tripp, 2019) and the loci from plastid sequence data obtained in our experiment. The incongruent phylogenetic placements at the tips may show the effect of incomplete lineage sorting and/or recent hybridisation at shallow levels of divergence but can also be influenced by the different scales of phylogenetic data (ddRAD vs partial plastome).

In some cases, we report low phylogenetic resolution based on mined organellar reads, which may hinder drawing additional evidence from this source of information. The examples listed above testify to 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 do 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 proper 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. A similar effect outlined for different RAD flavours in Andrews et al. (2016) could be observed for organellar RAD loci. Not surprisingly, ezRAD, designed to achieve a larger horizontal coverage, resulted in the longest alignment of organellar loci. The RADseq protocols using more frequent cutters (e.g. ezRAD or ddRAD with relatively frequent cutters), besides a larger horizontal coverage of the nuclear genome, can yield longer alignments of organellar RAD loci. Similarly, ddRAD seemed to be particularly suitable to obtain organellar RAD loci, but showed a large variability in the length of alignments (Figure 2 and Table 2). In general, sdRAD and GBS yielded the shorter alignments, which in some cases were comparable to the alignment length obtained by ddRAD (Figure 2 and Table 2). As already described by Bentley et al. (2019), the frequency of organellar RAD tags is largely limited by the frequency of restriction cut sites. The cutting frequency could be the most critical factor of the actual RAD flavour applied if the analysis of organellar RAD tags is an objective of the experiment. In light of this, the trend observed here should not be blindly generalised to RAD protocols, rather the cut site frequency of the restriction enzyme applied should be evaluated for the focal study groups.

As a potential drawback, our pipeline relies on the availability of a closely related reference organellar genome. In this study, we did not experience pronounced differences in the analysis outcome when using different reference genomes for the *Porites* and Cycadales datasets. Still, since selection of a reference genome with a high divergence relative to the focal taxonomic group can decrease the reliability of short read alignment, thus the accuracy of the obtained variant (see Bohling, 2020), we advise to use the most closely related organellar reference genome possible. Although the accurate reference genome assembly (both nuclear and organellar) of species is still a work in progress (e.g. Blaxter et al., 2022), given the ever-growing number of available organellar genomes in public repositories (Tonti-Filippini et al., 2017), this is not expected to seriously hinder our pipeline's utility. If multiple probable (equally distantly related) organellar references are available, the mapping quality of the reads aligned to the organellar reference genome and the read depth of loci should be evaluated to assess the reliability of the haplotype calling, which information is output by RADOrgMiner in tabular format.

Consequently, the molecular evolution of the studied group and the applied library preparation protocol (with an emphasis on the cut site frequency) are essential factors to mine for organellar loci in RADseq data. The assembly of organellar genomes became a routine task with the application of genome skimming. We argue that the mining organellar reads from RADseq libraries can be a cost-effective and viable option. Some studies that investigate the variability of both the nuclear and organellar genomes, besides RADseq, often utilise additional sequencing efforts to capture the organellar genome(s). As also outlined by previous authors (Clugston et al., 2019; Stobie et al., 2019), if the organellar genome(s) are sampled with a sufficient coverage using RADseq, the application of additional sequencing efforts might not be strictly necessary. The examples above show that by careful planning, RADseq can sample both the nuclear and organellar genomes with sufficient coverage and the mining of organellar loci does not come at an additional cost of sequencing. Dedicated bioinformatics tools (e.g. fragmatic [Chafin et al., 2018], GBS-Pacecar [Melo & Hale, 2018] or RADinitio [Rivera-Colón et al., 2021]); see also Stobie et al. (2019) can aid the planning of RADseq studies to represent the variability of both the nuclear and organellar genomes with a sufficient read depth. Although not explicitly tested, the fine-tuning capability of our pipeline raises the opportunity to mine also for RAD tags originating from sex chromosomes or endosymbiont genomes.

5 | CONCLUSION

In sum, we suggest that RRLs contain a significant amount of cytoplasmic DNA critical for discerning cytonuclear discordance as well as the frequency, and directionality of hybridisation, and phylogeography. More specifically, our pipeline can reliably genotype

organellar loci from RADseq datasets by effectively decreasing the sampling of NUPTs and NUMTs. Organellar genotyping is aided by providing basic measurements to assess the read depth and variability of the loci mined. We showed that organellar loci could effectively supplement the results of the nuclear dataset with no additional sequencing effort. As expected, by introducing restriction enzymes in RADseq, the number of mined loci is limited by cut site frequency. 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 evolutionary history and current processes.

AUTHORS' CONTRIBUTIONS

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

ACKNOWLEDGEMENTS

We thank László Bartha, Tamás Malkócs and Lajos Szatmári for their comments on the pipeline, Nikoletta A. Nagy, Jelena Marinkov and Jácint Tökölyi for advice on the first draft, and Isaac Overcast and three more anonymous reviewers for their comments and suggestions on the first version of this text. We greatly acknowledge the linguistic corrections generously provided by Susan M. Haig (Oregon State University). We are grateful to Mathew T. Sharples and Rafael G. Albaladejo for their feedback on the pipeline and comments on an earlier draft of the manuscript. We also appreciate issues opened at the GitHub repository of our pipeline that facilitated ease of use and compatibility of the script. Financial support was achieved via the New National Excellence Program of the Hungarian Ministry of Innovation and Technology (ÚNKP-20-4-I-DE-290) to L.L., and the 'OTKA' Young Researcher Excellence Program (FK 137962) to G.S. S.J. thanks the support of the Juhász-Nagy Pál Doctoral School of Biology and Environmental Sciences (University of Debrecen).

CONFLICT OF INTEREST

The authors have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13937.

DATA AVAILABILITY STATEMENT

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

The RADOrgMiner v0.9 script used in this study is achieved on Zenodo (Laczkó et al., 2022) and can be accessed at: https://doi.org/10.5281/zenodo.6619190

ORCID

Levente Laczkó https://orcid.org/0000-0002-9379-7527
Sándor Jordán https://orcid.org/0000-0002-3556-4127
Gábor Sramkó https://orcid.org/0000-0001-8588-6362

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, 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*, 3389–3402.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92.

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

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

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

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, *3*, e3376.

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*, 1842–1853.

Bendich, A. J. (1987). Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays*, *6*(6), 279–282.

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*, 8.

Blaxter, M., Archibald, J. M., Childers, A. K., Coddington, J. A., Crandall, K. A., Di Palma, F., Durbin, R., Edwards, S. V., Graves, J. A. M., Hackett, K. J., Hall, N., Jarvis, E. D., Johnson, R. N., Karlsson, E. K., Kress, W. J., Kuraku, S., Lawniczak, M. K. N., Lindblad-Toh, K., Lopez, J. V., … Lewin, H. A. (2022). Why sequence all eukaryotes? *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115636118.

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution*, *10*, 7585–7601.

Borowiec, M. L. (2016). AMAS: A fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660.

Brito, P. H., & Edwards, S. V. (2009). Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, *135*, 439–455.

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*, 325–328.

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*, 1610–1622.

D'Agostino, N., Taranto, F., Camposeo, S., Mangini, G., Fanelli, V., Gadaleta, S., Miazzi, M. M., Pavan, S., di Rienzo, V., Sabetta, W., Lombardo, L., Zelasco, S., Perri, E., Lotti, C., Ciani, E., & Montemurro, C. (2018). GBS-derived SNP catalogue unveiled wide genetic variability and geographical relationships of Italian olive cultivars. *Scientific Reports*, *8*, 15877.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean,

G., Durbin, R., & 1000 Genomes, P.A.G. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*, 2156–2158.

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*, 3151–3164.

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.

Eaton, D. A. R., & Overcast, I. (2020). ipyrad: Interactive assembly and analysis of RADseq datasets. *Bioinformatics*, *36*(8), 2592–2594.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797.

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

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042.

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*, e19379.

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*, 224.

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

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*, 397–423.

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.

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

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.

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). 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). 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 of the United States of America*, *111*, 11097–11102.

Laczkó, L., Jordán, S., & Sramkó, G. (2022). *laczkol/RADOrgMiner: RADOrgMiner v0.9* (v0.9) [Computer software]. *Zenodo*. https://doi.org/10.5281/ZENODO.6619190

Leaché, A. D., & Oaks, J. R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *48*, 69–84.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

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). 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*, 648–662.

Martín-Hernanz, S., Aparicio, A., Fernández-Mazuecos, M., Rubio, E., Reyes-Betancort, J. A., Santos-Guerra, A., Olangua-Corral, M., & 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*, e1416. https://doi.org/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*, 526–538.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303.

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*, 20170300.

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*, 7.

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*, 1783–1785.

Michalovova, M., Vyskot, B., & Kejnovsky, E. (2013). Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: Size, relative age and chromosomal localisation. *Heredity*, *111*, 314–320.

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*, 1530–1534.

Moura, A. E., Kenny, J. G., Chaudhuri, R. R., Hughes, M. A., Reisinger, R. R., de Bruyn, P. J. N., Dahlheim, M. E., Hall, N., & Hoelzel, A. R. (2015). Phylogenomics of the killer whale indicates ecotype divergence in sympatry. *Heredity*, *114*, 48–55.

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*, 2841–2847.

Paetzold, C., Wood, K. R., Eaton, D. A. R., Wagner, W. L., & Appelhans, M. S. (2019). Phylogeny of Hawaiian *Melicope* (Rutaceae): RADseq resolves species relationships and reveals ancient introgression. *Frontiers in Plant Science*, *10*, e1074. https://doi.org/10.3389/fpls.2019.01074

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*, e37135.

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 post-glacial admixture. *Molecular Biology and Evolution*, *32*, 2338–2350.

Pujolar, J. M., Jacobsen, M. W., Als, T. D., Frydenberg, J., Munch, K., Jónsson, B., Jian, J. B., Cheng, L., Maes, G. E., Bernatchez, L., & Hansen, M. M. (2014). Genome-wide single-generation signatures of local selection in the panmictic European eel. *Molecular Ecology*, *23*(10), 2514–2528.

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.

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*, 5937–5942.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842.

R Core Team. (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Richly, E., & Leister, D. (2004a). NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution*, *21*(6), 1081–1084.

Richly, E., & Leister, D. (2004b). NUPTs in sequenced eukaryotes and their genomic organisation in relation to NUMTs. *Molecular Biology and Evolution*, *21*(10), 1972–1980.

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

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. *Molecular Ecology Resources*, *21*, 363–378.

Rochette, N. C., Rivera-Colón, 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.

Salas-Leiva, D. E., Meerow, A. W., Calonje, M., Griffith, M. P., Francisco-Ortega, J., Nakamura, K., Stevenson, D. W., Lewis, C. E., & 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*, 1263–1278.

Schaal, B. A., & Olsen, K. M. (2000). Gene genealogies and population variation in plants. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(13), 7024–7029.

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*, 857–876.

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.

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*, 191–205.

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*, 10974–10986. https://doi.org/10.1002/ece3.362

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 hybridisation in polytypic barking frogs. *Molecular Ecology*, *23*, 3273–3291.

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*, 1760–1770.

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*, 173–174.

Toews, D. P. L., & Brelsford, A. (2012). The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, *21*, 3907–3930.

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

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.

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.

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*, 454–468.

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). Chapman & Hall.

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

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An r package for visualisation and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*, 28–36.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Laczkó, L., Jordán, S., & Sramkó, G. (2022). The RadOrgMiner pipeline: Automated genotyping of organellar loci from RADseq data. *Methods in Ecology and Evolution*, *13*, 1962–1975. https://doi.org/10.1111/2041-210X.13937