Phylogenomics and biogeography of Torreya (Taxaceae)—Integrating data from three organelle genomes, morphology, and fossils and a practical method for reducing missing data from RAD-seq

Wenbin Zhou1*, AJ Harris2, and Qiu-Yun (Jenny) Xiang1*

1Department of Plant and Microbial Biology, North Carolina State University, Raleigh 27695 NC, USA
2Key Laboratory of Plant Resources Conservation and Sustainable Utilization, South China Botanical Garden, Chinese Academy of Sciences, Guangzhou 510650, China
*Authors for correspondence. Qiu-Yun (Jenny) Xiang. E-mail: jenny_xiang@ncsu.edu; Wenbin Zhou. E-mail: wzhou10@ncsu.edu and wenbin.evolution@gmail.com

Abstract Restriction site-associated DNA sequencing (RAD-seq) enables obtaining thousands of genetic markers for phylogenomic studies. However, RAD-seq data are subject to allele dropout (ADO) due to polymorphisms at enzyme cutting sites. We developed a new pipeline, RAD-seq Allele Dropout Remedy in our study of the gymnosperm genus, Torreya, to mitigate ADO in outgroups by recovering missing loci from previously published transcriptomes. By using RADADOR to supplement Rad-seq data in combination with plastome and mitochondrial gene sequences, morphology, and fossil records, we reconstructed the phylogenetic and biogeographic histories of the genus and tested hypotheses on anomalies of biodiversity of the eastern Asian–North American floristic disjunction. Our results showed that our pipeline recovered many loci missing from the outgroup, and the improved data yielded a more robust phylogeny for Torreya. Using the fossilized birth–death model and divergence–extinction–cladogenesis method, we resolved a detailed biogeographic history of Torreya that suggested a Jurassic origin spanning Laurasia and differential speciation and extinction among continents accounting for modern diversity, which is biased toward eastern Asia (EA). The biogeographic results also supported a vicariance origin of modern Torreya from a widespread ancestor in EA and North America (NA) in the mid–Eocene, and cross Beringian exchange in the early Paleogene before the vicariant isolation, in contrast to the “out of EA” pattern common to gymnosperms and to the “out of NA” hypothesis previously proposed for the genus. Furthermore, we observed phylogenetic discordance between the nuclear and plastid phylogenies for Torreya jackii, suggesting differential lineage sorting of plastid genomes among species of Torreya or plastid genome capture in T. jackii.

Key words: Angiosperms353, eastern Asian–North American disjunction, fossilized birth–death model, phylogenetic discordance, plastid genome sequences.

1 Introduction

The origin and evolution of the floristic similarity between eastern Asia (EA) and eastern North America (ENA) have long been a focal interest of botanists (Boufford & Spongberg, 1983; Tiffney, 1985; Xiang et al., 1998; Qian & Ricklefs, 1999, 2000, 2004; Wen 1999; Qian et al., 2005; Wen et al., 2010, 2016). Eastern Asia and ENA share many genera with closely related species confined to one of the two regions, and some genera also have close relatives in southeastern Europe, western Asia, and/or western North America (WNA) (Boufford & Spongberg, 1983; Wen, 1999). This phytogeographic pattern is most prominent in angiosperms, but it is also observed in ferns and gymnosperms (Wen et al., 2016). An outstanding phenomenon of the EA–ENA disjunct pattern is the diversity anomaly bias toward EA (Qian & Ricklefs, 1999, 2000). That is, in genera displaying the disjunct distribution, there are more species in EA than in ENA. The underlying evolutionary causes of this diversity bias have puzzled botanists, especially over the past two decades, and several factors have been considered to explain why EA has greater
species richness within genera compared to ENA. These included a longer time in EA for diversification, greater speciation rate in EA, and greater extinction in NA, and these alternatives are not mutually exclusive. At present, few studies have the necessary data to evaluate alternative hypotheses. Moreover, the limited data so far are conflicting regarding greater speciation in EA (e.g., Xiang et al., 2004; Lindelof et al., 2020; Melton et al., 2020), or a time/age effect due to "out of Asia" migration, which has been inferred in biogeographic syntheses across disjunct lineages (Donoghue & Smith, 2004; Harris et al., 2013). However, when considering gymnosperms only, the 15 genera displaying the disjunct pattern showed a dominant “out of North America (NA)” pattern (reviewed in Wang & Ran, 2014; Table 1), suggesting that we should observe a reverse pattern of diversity anomaly, with more species in NA than in EA, if time/age explains the pattern but greater richness in EA but other causes underlie it. Another major finding from studies of the disjunct pattern in the past two decades was that the Miocene represented a major period of divergence and migration via the Bering land bridge prior to the intercontinental isolation or divergence of disjunct taxa in EA and NA (Wen et al., 2010). These intercontinental isolation or divergence of disjunct taxa migration via the Bering land bridge prior to the Miocene represented a major period of divergence and underlie it. Another major the pattern but greater richness in EA but other causes

Here, we conduct such a study on the phylogenies and biogeographic histories of individual range evolution.

RAD-seq was chosen because it has been a cost-effective method to generate genome-wide molecular data for evolutionary studies (Arnold et al., 2013; Eaton & Ree, 2013; Leaché & Oaks, 2017; Fernández-Mazuecos et al., 2018; MacGuigan & Near, 2019). Many recent studies have shown that RAD-seq data are useful to address species relationships of lineages at various levels of phylogenetic divergence, including lineages that diverged in the late Cretaceous and early Paleogene (Cariou et al., 2013; Hou et al., 2016; Eaton et al., 2017; Leaché & Oaks, 2017; Lecaudey et al., 2018; MacGuigan & Near, 2019; Du et al., 2020; Mu et al., 2020). Despite the utility of RAD-seq in evolutionary studies, concerns remain about the method, especially regarding allele dropout (ADO), which is caused by mutations and evolutionary changes in restriction sites among distantly related taxa, ultimately leads to missing data in downstream alignments, and remains a challenge. These missing data are clustered within clades (i.e., non-randomly distributed across trees), and are, thus, often phylogenetically informative (DaCosta & Sorenson, 2016). However, ADO can be a significant issue for both phylogenetic (Leaché & Oaks, 2017) and population genetic studies (Davey et al., 2013; Gautier et al., 2013; Andrews et al., 2016; Cariou et al., 2016). In population genetics, many approaches have been developed for mitigating ADO (Díaz-Arce &
Table 1 Summary of phylogenetic and biogeographic studies of gymnosperm genera

| Family   | Genus     | Markers for phylogeny | Biogeographic methods | Fossils included in biogeographic analysis | Inferred origin | Current distribution | References                                      |
|----------|-----------|------------------------|-----------------------|------------------------------------------|-----------------|----------------------|------------------------------------------------|
| Pinaceae | Abies     | nrITS; nrITS, 3 cp, 2 mt; 10 n, 5 cp | Ad hoc from phylogeny; BBM, DEC; DEC + J, DIVALIKE + J | No | WNA | AF, AS, EU, NA | Semerikova et al. (2018); Xiang et al. (2009, 2015) |
|          | Larix     | 1 cp; nrITS            | Ad hoc from phylogeny | Not available | NA? | AS, EU, NA | Wei and Wang (2003, 2004) |
|          | Picea     | 2 cp; 2 cp; 1141 n     | MacClade, DIVA; SDIVA, DEC DEC + j | No | WNA | AS, EU, NA | Ran et al. (2006, 2015); Shao et al. (2019) |
|          | Pinus     | 4 cp; 1662 n           | DIVA; DIVALIKE + J   | No | Eurasian; Now uncertain | AS, EU, NA, NA | Eckert and Hall (2006); Jin et al. (2021) |
|          | Pseudotsuga | 5 cp, 1 n             | Ad hoc from phylogeny | Not available | NA | AS, NA | Wei et al. (2010) |
|          | Tsuga     | 3 cp; 881 n            | AReA, DIVA; DIVALIKE + j | No | Asia; Now NA | AS, NA | Feng et al. (2021); Havill et al. (2008) |
| Cupressaceae | Calocedrus | nrITS                 | None                  | Not available | EA-NA (wide) | AS, NA | Chen et al. (2009) |
|          | Chamaecyparis | 2 cp                 | Ad hoc from phylogeny | Not available | NA | AS, NA | Wang et al. (2003) |
|          | Cupressus  | 3 cp, nrITS, 1 n      | Ad hoc from phylogeny | Not available | Uncertain | AS, EU, NA | Little (2006) |
|          | Juniperus  | 9 cpDNA               | Bayes-DIVA            | No | EA | AS, EU, NA | Mao et al. (2010) |
|          | Thuja      | 1 n; cp genome        | DIVA; DEC             | No | NA or EA-NA (wide); EA? | AS, NA | Adelalul et al. (2020); Peng and Wang (2008) |
| Taxaceae | Taxus      | nrITS; 5 cp, 1 nrITS; 13 cp, 1 n | Ad hoc from phylogeny; DIVALIKE + j | No | NA | AS, EU, NA | Hao et al. (2008); Li et al. (2001a); Möller et al. (2020) |
|          | Torreya    | 5 cp, nrITS; complete cp; nrITS | DIVA, S-DEC           | No | Asia; EA-NA (wide) | AS, NA | Hao et al. (2008); Li et al. (2001b); Zhang et al. (2019) |
| Ephedraceae | Ephedra   | 9 cp, 2 nrITS, 2 n; 4 cp, 3 n | Ad hoc from phylogeny; DEC | No | Asia; Western Asia and Mediterranean | AF, AS, EU, NA, SA | Ickert-Bond et al. (2009, 2020) |

AF, Africa; AS, Asia; DEC, Dispersal–Extinction–Cladogenesis; DIVA, Dispersal Vicariance Analysis; EA, eastern Asia; EU, Europe; NA, North America; SA, South America; WNA, western North America. “cp” means chloroplast genes; “mt” means mitochondria genes; “n” means nuclear genes; and “nrITS” means Nuclear Ribosomal Internal Transcribed Spacer. “Markers used for phylogeny” refers to those used for the construction of the tree applied in biogeographic analysis.
Within phylogenetic studies, ADO is especially a problem in outgroup taxa. Many studies using RAD-seq revealed low numbers of loci recovered in the outgroup and a low number of loci shared between the ingroup and outgroup species, both of which are probably attributable to ADO (Cruaud et al., 2014; Hou et al., 2015; Díaz-Arce et al., 2016; País et al., 2017; Zhou et al., 2018; Lin et al., 2019; Du et al., 2020; Lindelof et al., 2020). Within outgroup taxa, a large amount of missing data might lead to an inaccurate phylogenetic inference, especially in cases of deep radiation (i.e., near the root). However, the effect of large amounts of missing data in outgroup taxa is usually either left unaddressed or mitigated by performing phylogenetic analyses on a reduced RAD-seq data matrix containing only loci present in the outgroup and ingroup taxa (e.g., Du et al., 2020). The latter approach results in the loss of a large amount of potentially phylogenetically informative loci within the ingroup. In this study, we developed a new pipeline, RAD-seq Allele Dropout Remedy (RADADOR), to mitigate the problem of ADO by mining transcriptome data in public databases and demonstrated its value via our phylogenomic study of Torreya. We developed the RADADOR pipeline in this study of Torreya because the genus is deeply diverged from its sister, Amentotaxus Pilg. (Leslie et al., 2012). Thus, it is a challenge for our molecular phylogenetic study using RAD-Seq, which otherwise represents an improvement over past studies of Torreya based on one or a few gene regions (Li et al., 2001b; Hao et al., 2008; Leslie et al., 2012; Wang et al., 2019; Zhang et al., 2019).

Our study had three objectives: (i) developing and assessing our pipeline, RADADOR, for retrieving sequences from publicly available transcriptomes to fill out missing loci in outgroups in the RAD-seq data, (ii) evaluating the impact of missing data in the outgroup by comparing the phylogenetic results from data including sequences obtained from the new pipeline with the results from RAD-Seq only, and (iii) elucidating the phylogeny and the biogeographic history of Torreya to gain insights into the origin and causes of the diversity anomaly of the EAA–NA disjunct pattern and to test the major “out of America” pattern inferred previously for gymnosperms.

2 Material and Methods

2.1 DNA preparation

Our sampling consisted of 28 accessions of Torreya representing all seven species, based on the taxonomic treatment in the Flora of China (Fu et al., 1999) and Flora of NA (Hils, 1996). For the outgroup, due to the unavailability of fresh materials or genome sequences for Amentotaxus Pilg., the genus most closely related to Torreya (Hao et al., 2008; Leslie et al., 2012), we sampled Taxus wallichiana var. chinensis (Pilg.) Florin based on the close phylogenetic position of Taxus to Torreya within a Taxaceae phylogeny reconstructed by Cheng et al. (2000). We collected leaf samples of Torreya and Taxus from plants growing in the field or arboreta/botanical gardens of the United States and China (Table S2) and dried the leaves in silica gel. Once the leaves were dry, they were stored at −20°C until used for DNA extraction. Pressed specimens were generated from collections as vouchers for all samples and deposited in herbaria (Table S2). We did not distinguish infraspecific taxa in the sampling as our objective was to resolve the relationships within Torreya at the species level. Therefore, we recognized Torreya grandis as a single taxonomic entity despite its many varieties, cultivars, and infraspecific taxonomic uncertainties owing to its wide cultivation in China for its edible seeds (Hu, 1927; Cheng & Fu, 1978; Kang & Tang, 1995).

We performed total genomic DNA extractions using a modified CTAB protocol of Doyle (1991) following Cullings (1992) and Xiang et al. (1998) and checked the concentration and quality of each purified DNA sample using a Nanodrop spectrophotometer (ThermoFisher), 1% agarose gel electrophoresis, and a PicoGreen fluorescent dye assay (Life Technologies, ThermoFisher).

2.2 Library preparation

We prepared the double-digest RAD-seq libraries following País et al. (2017) and Peterson et al. (2012) using the PstI-HF and MspI restriction enzymes. Following initial preparation of libraries and pooling, we selected fragments from 400 to 600 bp in size, enriched these using PCR, and sent the samples for single-end sequencing (150 bp) on an Illumina HiSeq 2500 platform in GSL at NCSU. We also used Hyb-Seq (target enrichment) in an attempt to capture up to 353 nuclear genes, which are putatively single copy in many angiosperms (Johnson et al., 2019), with the Angiosperms353 probe kit. Simultaneously, we sought to harvest and analyze off-target genes from the results; namely, from the chloroplast and mitochondrial genomes. Due to the relatively high cost and conserved nature of the Angiosperms353 genes, we performed Hyb-Seq for eight of the 29 samples analyzed with RAD-seq. This sampling mainly involved the removal of intraspecific duplicate accessions so that all seven species of Torreya and the outgroup taxon (Taxus wallichiana var. chinensis) were represented by one accession each (Table S2). The goal was to generate a minimal data set from the Angiosperms353 genes for phylogenetic analysis of Torreya for a comparison with the RAD-seq data-based phylogeny, while contributing Angiosperms353 gene data to the community for future global phylogenetic analyses of seed plants. All Hyb-Seq procedures were performed by the Rapid Genomics Lab (Gainesville, Florida, USA) using our extracted DNAs. Sequencing was conducted using Illumina MiSeq (Illumina, San-Diego, California, USA) to produce 2 × 150 bp paired end reads. More details are provided in the Supplementary Materials.

2.3 Analyses of sequence data

To analyze raw sequence reads from RAD-seq, we applied iPyRAD v.0.7.28 (Eaton & Overcast, 2020) generally using the same parameters as in our previous studies of Aesculus (Du et al., 2020) and Cornus (Lindelof et al., 2020). Briefly, we demultiplexed the reads and then trimmed them to remove specific barcodes and adapters. Following quality control, we removed DNA sequences representing the plastid genome and endophytic microorganisms using the plastome of Torreya fargesii (NC_029398, Tao et al., 2016) and the complete bacterial and fungal refseq databases (https://ftp.
ncbi.nlm.nih.gov/refseq/release/) from NCBI, respectively, as references. Following Eaton (2014), we used ipyrad to generate data matrices of orthologous loci for diploid organisms because no polyplody has been reported from Torreya (n = 11 or 8; Chromosome Counts DataBase: http://ccdb.tau.ac.il/Gymnosperms/Taxaceae/Torreya/). We also detected and removed mitochondrial genes in the RAD-seq matrices using our new RADADOR pipeline (see below) based on the mitochondrial genome from a closely related species, Taxus cuspidata Siebold & Zucc. (MN593023, Kan et al., 2020), as the reference.

To test the effects of missing data in phylogenetic inferences, we selected different thresholds for the representation of a gene locus, from at least 20% to at least 80% in increments of 10% for a total of seven sets of data matrices. We named the matrices based on the threshold; for example, M20 comprises loci present in at least 20% of all samples (see Results). All matrices have been uploaded to Dryad (https://doi.org/10.5061/dryad.08kprr52s).

### 2.3.1 Extended RAD-seq pipeline, RADADOR

We developed RADADOR especially to address the large amount of missing data that we observed in the outgroup, T. wallichiana var. chinensis, for which we sought to recover loci from existing, publicly available transcriptomic data (SRR5894321, Ran et al., 2018). We also sought to obtain loci for the more closely related taxon, Amentotaxus argotaenia (Hance) Pllg. (SRR8113475, Yu et al., 2021), by applying RADADOR to its available transcriptome data. Therefore, we downloaded these transcriptomes and assembled them using Trinity (Grabherr et al., 2011) with default parameter settings. We used the assembled transcriptomes as input for RADADOR, which extracts orthologs these additional extracted orthologs to the RAD-seq data set. Thus, we were able to add these additional extracted orthologs to the RAD-seq data matrices.

The RADADOR pipeline is fully available on Github (https://github.com/Bean061/RAD_Allele_Dropout_Remedy) and comprises the following steps (Fig. 1): (s1) Apply the customized function “Loci2partition_nex” to convert the consensus sequences of RAD-seq “loci” file generated by ipyrad into a “partition” nexus file; (s2) use the partition information in the nexus file to split a phylip file containing the consensus sequences of RAD-seq loci into individual gene alignments by applying the “Concatenated2gene_phylip” function; (s3) combine all gene sequences from all individuals into one master fasta file using the “combine2RADseq” function as the query file for nucleotide BLAST (BLASTN, Altschul et al., 1990) in the next step; (s4) apply the master fasta file of RAD-seq sequences to query a database of contigs from an assembled transcriptome (e.g., generated using Trinity in our case) using BLASTN via the “Blast_py” function. We required “-max_target_seqs. 1 -max_hsps 1” in BLAST to keep only the best matching sequences for each query. (s5) Use the BLASTN result to pull putative orthologs from the assembled transcriptome corresponding to loci represented in the RAD-seq data set and add these to the individual files representing the loci (e.g., used in s2) via the “addRNA2RAD” function. (s6) Realign each locus with the newly added sequences using the “mafft_genes” function.

### 2.3.2 Bioinformatics for Hyb-Seq data

We demultiplexed all samples from Hyb-Seq using Illumina’s BCLtobaseq by Arbor Biosciences, yielding the raw sequencing reads, and we cleaned the data using Trimmomatic v.0.38 (Bolger et al., 2014). Unfortunately, applying the Angiosperms353 probe set to Torreya yielded no gene sequences. This was likely due to failed gene capture because of the high levels of divergence of these gene sequences between angiosperms and gymnosperms as the 353 gene probes were designed based on gene sequences of angiosperm only. To capture off-target chloroplast genome sequences, we mapped reads to the reference sequence of T. fargesii (NC_029398; similar to cleaning the RAD-seq data). We used BWA (Li & Durbin, 2009) and samtools (view -f 3) (Li et al., 2009) to map the raw Hyb-Seq reads from each species to the reference and output the results in a bam format file. We then used Geneious v.2020.1.1 (Kearse et al., 2012) to generate a consensus sequence from the mapped reads with more than 5x depth. For species representing the outgroup, Amentotaxus formosana H.L. Li, A. argotaenia, and T. wallichiana var. chinensis, we obtained annotated chloroplast CDS from NCBI (accession numbers: NC_024945, Hsu et al., 2014; NC_027581, Li et al., 2016; KX431996, Jia & Liu, 2017; respectively) using a customized python script. To increase sampling of Torreya in the chloroplast gene phylogeny, we downloaded five additional chloroplast genomes representing five species of Torreya from previous studies (Table S2) (Hsu et al., 2014; Li et al., 2016; Tao et al., 2016; Zhang et al., 2019). We combined these whole-chloroplast genome sequences with those from Hyb-Seq and used MAFFT (Katoh & Standley, 2013) to obtain an aligned matrix.

To obtain off-target mitochondrial genes, we used the only available mitochondrial genome of Taxus, T. cuspidata (MN593023, Kan et al., 2020; also used for cleaning our RAD-seq data), as the reference, and it contained 46 annotated mitochondrial genes. We used the same approaches described above for the chloroplast genes to extract and align the mitochondrial genes from the Hyb-Seq data before the construction of a mitochondrial phylogeny of Torreya. For the mitochondrial phylogeny, T. wallichiana var. chinensis was used as the outgroup, the same taxon used as the outgroup in analyses of the RAD-seq data.

### 2.4 Phylogenetic analyses

#### 2.4.1 Concatenation-based TREE phylogeny

We conducted phylogenetic analyses of the concatenated RAD-seq data matrices M20 to M80 using the maximum likelihood (ML) method in RAxML v8.2.4 (Stamatakis, 2014) on the CIPRES Science Gateway Portal (Miller et al., 2011) (http://www.phylo.org/portal2/home.action) to assess the general congruence of the resulting topologies. For all ML analyses, we applied the GTR+I model of nucleotide substitution according to AIC tests in jModelTest v.2.1.4 (Darriba et al., 2012) without partitioning, and conducted 1000 rapid bootstraps to ascertain support. In general, we observed incongruence among M20, M30, and the rest of the matrices regarding species relationships, while the M50, M60, and M70 trees showed the overall strongest support values for nodes (see Results). Thus, we selected the threshold of 50% missing per locus as the most suitable for downstream analyses because it...
contains more loci than M60 and M70 while producing trees with strong support, although we continued to process all matrices in the evaluation of our RADADOR pipeline. We also used these preliminary phylogenetic results to further curate our data set, especially by removing several dubious taxa, taxa of *Torreya* with too few loci in otherwise robust genomic matrices, and sequences containing over 30% uncertain bases. To distinguish the raw RAD-seq matrices from these further filtered matrices (with some questionable samples removed), we refer to the filtered matrices as ipyrad matrices (e.g., ipyradM50 represents the modified M50). We also conducted phylogenetic analyses of the ipyradM20 to ipyradM80 using RAxML v8.2.4 (results were deposited in Dryad).

For ipyradM50, we performed phylogenetic analyses using the ML method in IQ-TREE v. 1.6.12 (Nguyen et al., 2015) with *T. wallichiana* var. *chinensis* as the outgroup and sequence partitioning by gene locus. We chose IQ-TREE over RAxML here is because the program implemented more substitution models and permitted more rigorous analyses with gene partitioning under the best-fit molecular model for each gene region. We allowed IQ-TREE to find the best model of evolution for each locus using its implementation of ModelFinder (Kalyaanamoorthy et al., 2017). To estimate tree support, we performed 1000 ultrafast bootstrap replicates (Hoang et al., 2018).

In the ipyradM50 matrix, over 98% of loci were missing in the outgroup, *T. wallichiana* var. *chinensis* (see Results). Thus, we applied our RADADOR pipeline to mine transcriptome data and supplement the outgroup with additional loci. Based on these results, we generated a new matrix, which we refer to as radadorM50. To determine if the missing data of the outgroups in the ipyradM50 affected phylogenetic results, we performed an ML analysis in IQ-TREE of the radadorM50 matrix for comparison as described for ipyradM50. We also compared all ipyrad matrices (ipyradM20-ipyramidM80) with the corresponding RADADOR...
matrices (radadorM20-radadorM80). Both the ipyrad and radador matrices incorporated only Taxus as the outgroup, the taxon whose data were obtained from our RAD-seq experiment. However, to test whether a closer outgroup influences the tree topology, we repeated the analyses of the radadorM50 matrix with both Taxus and A. argotaenia (data obtained from transcriptome mining) as the outgroup. We chose radadorM50 as the final matrix for downstream analysis (e.g., coalescent-based species tree reconstruction, divergence time dating, and biogeographic analysis) because it provided the highest support for species relationships that were consistent among all radador data sets (except radadorM30) (see Results).

Because relatively lower support was observed for relationships among the three subclades of Torreya in the concatenated tree and coalescence-based species tree (see Results), we further conducted an analysis for the radadorM50 matrix including T. wallichiana var. chinensis as the outgroup using the Neighbor-Net method (Bryant & Moulton, 2004) in SplitsTree4 (Huson & Bryant, 2006), which does not enforce a tree-like topology and, consequently, can help to determine if introgression was involved in the evolutionary history of Torreya.

2.4.2 Coalescent-based species tree
We constructed species trees by applying both SVDQuartets (Chifman & Kubatko, 2014) and ASTRAL-III (Zhang et al., 2018) to the ipyradM50 and radadorM50 matrices. We performed SVDQuartets analyses in PAUP* v4.0a166 (Swofford, 2002) and used all quartets with 1000 bootstrap replicates in each analysis. Applying the results from each matrix, we produced a summary tree with the quartet assembly method, QFM (Reaz et al., 2014), following Zhou et al. (2021). For the phylogeny, we removed two likely misidentifications (TF_17) and (TT_4), which were consistent among all radador data sets (except radadorM30) (see Results).

Before the analyses, we used all quartets with 1000 bootstrap replicates in each analysis. Applying the results from each matrix, we produced a summary tree with the quartet assembly method, QFM (Reaz et al., 2014), following Zhou et al. (2021). For the analyses in ASTRAL-III, we used the IQ-TREE to generate gene trees for each locus as the input and ran ASTRAL-III under default parameters. We visualized and edited the trees from SVDQuartets and ASTRAL-III in FigTree v.1.4.4 (Rambaut, 2012) and Adobe Illustrator 2020 (Adobe Systems, San Francisco, CA, USA).

2.4.3 Chloroplast and mitochondrial phylogenies
We reconstructed phylogenetic trees based on the chloroplast and mitochondrial genes obtained from off-target Hyb-Seq data independently. We performed the analyses using ML in IQ-TREE as described above for ipyrad- and radadorM50, but without any gene partitions. Before the final reconstruction of the chloroplast phylogeny, we removed two likely misidentified samples, Torreya taxifolia (TT_4) and T. fargesii (TF_17), which were also removed from our RAD-seq data (see above) (Table S2).

2.5 Introgression and ILS test
We tested whether incomplete lineage sorting (ILS) or introgression may best explain the incongruences that we observed among gene trees at nodes with relatively low support in the coalescent-based species trees and between the coalescent-based species trees and the chloroplast gene tree. We applied both the D-statistic (ABBA-BABA) method in Dsuite (Malinsky et al., 2021) and a phylogenetic network method in PhyloNetworks (Solís-Lemus et al., 2017) to detect introgression. The main conflict among gene trees involved the relationships among Torreya jackii and the two pairs of EA species, and the conflict between the coalescent-based species trees and the chloroplast gene tree involved only the placement of T. jackii (see Results). Thus, we ran the analysis in Dsuite for T. jackii, T. fargesii-Torreya yunnanensis, and T. grandis-Torreya nucifera with either Torreya californica or T. taxifolia as the outgroup (O) based on the phylogenetic results from the radadorM50 matrix. We also used the chloroplast gene tree as the backbone and T. jackii as the outgroup to test whether introgression events occurred among the three other major clades (three species pairs) of Torreya: T. fargesii + T. yunnanensis, T. grandis + T. nucifera, and T. californica + T. taxifolia. For the PhyloNetworks, we used the unlinked SNPs from radadorM20 (i.e., one SNP per locus) to include more data and ran ten independent analyses with zero, one, and two hypothesized reticulation events.

2.6 Biogeographic analysis
2.6.1 Constructing a total-evidence dated phylogeny including fossils
We performed divergence time dating using a total-evidence approach (Ronquist et al., 2012), following Gavryushkina et al. (2017), Lindelof et al. (2020), and Zhou et al. (2020). This approach incorporated data from the radadorM50 and the morphology and ages of fossils into phylogenetic reconstruction and divergence time estimation simultaneously using the fossilized birth-death (FBD) model implemented in BEAST v2.6.2 (Bouckaert et al., 2014). Within the FBD model, we used 23 fossils of Torreya ranging in age from mid-Jurassic to Neogene (Table S1). For each fossil and all of the living species of Torreya, we attempted to score 21 morphological characteristics that vary among species (Table S3). Following scoring, we appended the morphological data to the radadorM50 matrix. In the analysis, we implemented the GTR + Γ model for molecular data and the MK model (Lewis, 2001) for the morphological data and used Amentotaxus as the outgroup. We also applied an age constraint of 178-199 at the root of the phylogeny based on the earliest fossil of Taxaceae, Paleotaxus rediviva Nathorst (1908), from the early Jurassic and several fossils of Torreya from the mid-Jurassic. We ran the analysis using a relaxed, uncorrelated lognormal clock (Drummond et al., 2006), without any topological constraints on the fossils or extant species, for 600 million generations with sampling of trees every 10,000 generations. We assessed the quality of the run in Tracer v.1.7.1. (Rambaut et al., 2018) to make sure that the likelihood scores reached a plateau after burn-in, and most of the effective sample size (ESS) values were greater than 200. We generated a maximum clad credibility tree with median branch lengths in TreeAnnotator (Bouckaert et al., 2019) following a 20% burn-in.

2.6.2 Reconstructing the biogeographic history using DEC
We performed historical biogeographic reconstruction using the Dispersal–Extinction–Cladogenesis (DEC) method (Ree & Smith, 2008) based on the dated total evidence tree generated under the FBD model. We defined four geographic regions for Torreya in the DEC analysis to cover the distributions of all modern and extinct taxa: EA (“A”);ENA (“B”);WNA (“C”); and Europe (“E”) (Table S1), and we assessed dispersal capabilities between regions
during different times (Scotese, 1991, 2001; Graham, 1999b, 1999a, 2018; Manchester & Tiffney, 2001; Mann, 2007) and applied a matrix of relative dispersal rates (Table S3). The relative rates were as follows: 0.001 for time slices when the two areas were physically disconnected or isolated by environmental barriers (such as dry areas); 1.0 when the two areas were physically and ecologically connected; and 0.2, 0.5, or 0.75 when the two areas were partially connected by islands that were spaced at different distances (Table S4). For ancestral ranges, we excluded the probability of “AB” and “CE” because such ancestral distributions would require frequent intercontinental long-distance dispersal of pollen or propagules to prevent eventual allopatric speciation. We performed three analyses based on uncertainties in the ancestral range of the outgroup, Amentotaxus, which we coded as “E” (Harris, 1976) and “C” (Florin, 1963), based on fossil evidence, and “ACE”, which covers the entire range of all fossils and extant species.

More details of all analyses described above are available in the Supplementary Materials.

3 Results

3.1 Illumina sequence data: RAD-seq and Hyb-Seq data

We generated an average of 757,947 reads per accession of Torreya and had an average of 757,500 reads remaining following filtering for high quality in ipyrad and removal of reads from the chloroplast genome and endophytes (Tables 2, 7). When we compiled filtered reads based on missing data, we obtained a range of 768 loci for the M80 matrix to 5970 loci for the M20 matrix (Table S6). Our results from RAxML analyses of the M20–M80 matrices showed that the M50 and M60 matrices had the lowest cumulative frequencies of nodes with low support (equal or less than 60% bootstrap values) (Fig. S1). The results also showed that trees from M40 through M80 were congruent regarding species relationships, although varied in support values at some nodes, whereas trees from M20 and M30 differed from this tree topology (Figs. S2–S8). Among the congruent trees, the M50, M60, and M70 trees all showed high support for reconstructed species relationships. Our results from RAxML analyses of ipyradM20–ipyradM80 showed that all ipyradM30–ipyradM80 has consistent topologies with the M40–M80 of the raw data, but with low support values at the node connecting Torreya grandis–Torreya nucifera–T. jackii and T. fargesii–T. yunnanensis (the Asian clade) (Fig. S1). Thus, we selected the ipyradM50 for downstream analyses because it included more data than the other matrices that generated congruent trees with high support (2731 loci and 20 of 29 accessions, with less than 30% missing data) (Table S6).

Within our Hyb-Seq data for Torreya, we recovered none of the 353 target genes represented in the Angiosperms353 probe kit, and this was likely due to the high sequence divergence of these genes in gymnosperms compared to angiosperms. However, we successfully used the Hyb-Seq data to obtain off-target reads representing partial chloroplast genomes ranging in size from 12,190 bp to 87,357 bp among accessions (compared to approximately 137,000 bp on average for the complete chloroplast genomes of species Torreya [Zhang et al., 2019]) and partial mitochondrial genomes ranging from 4579 bp and 15,599 bp (compared to approximately 469,000 bp in the complete mitochondrial genome of Taxus; Table S7).

3.2 Missing loci recovered from transcriptome data

With our RADADOR pipeline, we retrieved 574 additional orthologous loci from the publicly available transcriptomic data for Amentotaxus argotaenia and 317 loci for Taxus wallichiana var. chinensis (Table S6). This is in contrast to the sequences originally obtained from RAD-seq, which represented only 47 loci for T. wallichiana var. chinensis in ipyradM50 (see Methods) and none for A. argotaenia, which we did not sequence. Thus, RADADOR can retrieve about seven times more loci for downstream analyses (Table S6).

We incorporated the sequences obtained using RADADOR into the radadorM50 matrix and all other radador matrices (see Methods). On comparing the RADADOR matrices to the ipyrad ones, we found that larger numbers of RAD-seq loci of Torreya (e.g., in ipyradM20 compared through ipyradM80) resulted in correspondingly higher numbers of recovered loci in outgroups from transcriptomes (Table S6).

3.3 Phylogenetic analyses

3.3.1 Concatenated species trees

The preliminary RAxML results based on the raw RAD-seq data sets (M20 through M80) revealed two differing phylogenies that varied in relationships, with relatively low support, among three, strongly supported major clades (Figs. S2–S8): Torreya fargesii–Torreya yunnanensis, T. grandis–T. nucifera, and Torreya californica–Torreya taxifolia (Figs. S2–S3 vs. Figs. S4–S8) and between Torreya jackii and the former two major clades. Torreya jackii was placed as either the sister of T. fargesii–T. yunnanensis (Figs. S2–S4) or of T. grandis–T. nucifera (Figs. S5–S8). The relationships of these lineages in the ipyradM50, ipyradM60, and ipyradM70 trees were identical (also observed in trees from ipyradM30, ipyradM40, ipyradM80, and M40–M80 of the raw data) and mostly highly supported among the trees (trees from ipyrad matrices deposited in Dryad). The IQ-TREE phylogeny of ipyradM50 (24 samples, Fig. S9) similarly showed T. jackii as the sister of T. grandis–T. nucifera with strong support (99 UFBS), as observed in the RAxML M50 tree (Fig. S5). However, it differed from the RAxML M50 tree in showing the three species being sister to the clade of the two North American species, T. californica–T. taxifolia, with moderate support (72 UFBS; Fig. 2A). In the RAxML M50 tree the three species were united with T. fargesii–T. yunnanensis (all from eastern Asia) (Fig. S9), suggesting the potential impact of molecular models and gene partitioning on phylogenetic analysis using the ML method.

The IQ-TREE trees from the RAD-seq matrices modified by the RADADOR pipeline (Figs. S10–S16) showed slightly different relationships among themselves, but they had improved support values compared to those from the ipyrad matrices. Among the analyses of the radadorM20 to radadorM80 matrices including two outgroup taxa, A. argotaenia and T. wallichiana var. chinensis, all results (except that of radadorM30; Fig. S11) showed consistent and generally well-supported relationships among three strongly supported major
clades: T. californica–T. taxifolia, T. jackii–T. nucifera–T. grandis, and T. fargesii–T. yunnanensis (Figs. S10–S16). The radadorM50 matrix resulted in a topology with the highest support values among different RADADOR matrices and placed the two NA species (T. californica as sister to T. taxifolia) as the sister to the remainder of the genus, T. jackii–T. nucifera–T. grandis and T. fargesii–T. yunnanensis from EA, different from the ipyradM50 topology from the IQ-TREE (Fig. S13 vs. Fig. S9; Figs. 2A–2C). The phylogenies resulting from analyses of the radadorM50 with Taxus alone or both Taxus and Amentotaxus as the outgroup were identical in topology and both were highly supported (Figs. 2B, 2C, S13). However, the IQ-TREE phylogeny using both Amentotaxus and Taxus as the outgroups had higher support for relationships among the three major clades (at the node connecting all EA species) (Figs. 2C, S13) than the phylogeny using only Taxus as the outgroup (Figs. 2B, S17) and the phylogeny from ipyradM50 (Fig. 2A). The SplitsTree network based on the Neighbor-Net method showed similar major lineages, except that T. jackii and T. grandis–T. nucifera are placed close to the outgroup taxa (Fig. 4).

Fig. 2. Phylogenies inferred from the restriction site-associated DNA sequencing (RAD-seq) ipyradM50 matrix and the improved RAD-seq radadorM50 matrix through the RADADOR pipeline using the IQ-TREE. A, Topology based on ipyradM50 using only Taxus as the outgroup (simplified from Fig. S9). B, C, Topologies based on radadorM50 using only Taxus and both Taxus and Amentotaxus as the outgroups, respectively. Tree topologies in B and C are simplified from Figs. S17 and 3A. Numbers on branches show ultrafast bootstrap values. Numbers in red represent the changed support values.

Fig. 3. Phylogenies of Torreya inferred in IQ-TREE using concatenated gene matrices from restriction site-associated DNA sequences (RAD-seq) and plastome sequences, respectively, with Amentotaxus and Taxus as the outgroup. The plastome data were recovered from Hyb-Seq using the Angiosperms353 probe kit for Torreya, which failed to generate any target nuclear genes. All RAD-seq loci of the outgroups were recovered by the RAD-seq Allele Dropout Remedy (RADADOR) pipeline. Numbers on branches of both trees show ultrafast bootstrap values from IQ-TREE. Dashed lines connect the same samples used in both RAD-seq and Hyb-Seq analyses. Data for the remaining samples in the plastome phylogeny were obtained from NCBI. Circles following a species name indicate the status of albumen (endosperm): black, ruminant; white, nonruminant.
3.3.2 Coalescent-based species tree
The results from analyses of radadorM50 using SVDQuartets and ASTRAL-III slightly differed from each other regarding the relationships among the three EA lineages (Fig. 5). The species trees from both methods (Fig. 5) were identical to the concatenation-based radadorM50 tree (Fig. 3A) in showing the NA clade being sister to all of the EA species. However, within the EA clade, the SVDQuartets tree showed a sister relationship between *T. jackii* and *T. grandis*–*T. nucifera* with 63% bootstrap support (Fig. 5A), while the ASTRAL-III tree showed that *T. jackii* was sister to the remaining four EA species, albeit with negligible support at only 0.22 PP (Fig. 5B; branch is too short to be visible on the tree). The SVDQuartets tree based on ipyradM50 showed another different relationship for *T. jackii*, placing the species as sister to the NA clade with low support (46.4) (Fig. 5C), whereas the ASTRAL-III tree of ipyradM50 placed *T. jackii* as the sister of *T. grandis*–*T. nucifera* (similar to the SVDQuartets tree of radadorM50), but with stronger support (0.85) (Fig. 5D). Additionally, the ASTRAL-III tree of ipyradM50 showed collapse of the monophyly of the NA species, with *T. californica* sister to the EA species (Fig. 5D).

3.3.3 Chloroplast gene tree
The chloroplast gene phylogeny (Fig. 3B) showed similar relationships among species and major clades as the concatenation-based trees of the radadorM50 (Fig. 3A), except for the position of *T. jackii*, which was sister to the rest of *Torreya*. Thus, *T. nucifera*–*T. grandis* was sister to *T. fargesii*–*T. yunnanensis* with 100% bootstrap support, and these two clades were sister to the NA clade with 97% bootstrap support (Fig. 3B).

3.3.4 Mitochondrial gene tree
Compared to the chloroplast phylogeny, the mitochondrial phylogeny showed negligible resolution at almost all nodes (Fig. S18). Only the crown node of *Torreya* had 100% UFBS support, while all other nodes had support of ≤70 UFBS (Fig. S18). This is likely because the sequences recovered for the mitochondrial genes contained insufficient informative sites compared to the chloroplast genes, and we obtained less data from the mitochondrial genome overall than the chloroplast genome (Table S7).

3.4 Introgression and ILS test
Based on the D-statistic tests, we detected one supported introgression event between *Torreya jackii* and *T. grandis* when using either *T. californica* (*D* = 0.2157, *P* = 0.0169) or the *T. californica* and *T. taxifolia* clade (*D* = 0.1717, *P* = 0.0388) as the outgroup. The PhyloNetworks based on unlinked SNP data suggested one introgression event between *Amentotaxus* and *T. yunnanensis* when one hybrid event was allowed and another introgression event between *T. jackii* and *T. grandis* when two hybrid events were allowed (Fig. S19). Within PhyloNetworks, *T. jackii* was sister to *T. grandis*, with which it introgressed in a biased manner, yielding *T. nucifera*, bearing 0.10 and 0.90 of the parental genomes, respectively (Fig. S19), suggesting a hybrid origin of *T. nucifera*.

3.5 Total-evidence dating
The results of the total evidence dating analysis suggested that *Torreya* diverged from a shared ancestor with *Amentotaxus* in the Lower Jurassic (186.4 Ma, 95% HPD: 175.9–196.7 Ma) and likely underwent early diversification into several lineages within the Jurassic (Fig. S20). The divergence time of the crown node of extant lineages...
began in the mid-Eocene (43.3 Ma, 95% HPD: 30.8–56.5 Ma) (Fig. S20). Within the North American clade of *Torreya*, the divergence time between *T. californica* and *T. taxifolia* dated to the Oligocene (26.4 Ma, 95% HPD: 20.4–35.7 Ma). In the eastern Asian clade, the divergence time between the *T. fargesii*–*T. yunnanensis* clade and the *T. jackii*–*T. grandis*–*T. nucifera* clade was estimated to be in the late Eocene (37.1 Ma, 95% HPD: 25.9–49.4 Ma). Within each eastern Asian clade, *T. fargesii* and *T. yunnanensis* diverged during the mid-Miocene (13.7 Ma, 95% HPD: 6.3–20.5 Ma), *T. jackii* diverged from *T. grandis*–*T. nucifera* in the late Eocene-early Oligocene (34.4 Ma, 95% HPD: 23.0–44.2 Ma), and, subsequently, *T. grandis* split from *T. nucifera* in the early Miocene (25.8 Ma, 95% HPD: 23.3–32.1 Ma). All results are shown in Fig. S20.

### 3.6 Biogeographic history

Results from the biogeographic analysis with the outgroup, *Amentotaxus*, coded in “E” suggested the ancestor of *Torreya* most likely widespread in the Northern Hemisphere, spanning all four areas (i.e., “ABCE”) (Fig. 6). During the Jurassic and Cretaceous periods, the widespread ancestor underwent several vicariance events and experienced extinctions in Eurasia and NA, as evidenced by the fossil taxa in the phylogeny. Following extinctions in Europe and Asia in the Jurassic, an NA lineage persisted through the Cretaceous. In the late Cretaceous and early Paleogene, the results suggest that a dispersal of this NA lineage into Asia re-established the genus there. This widespread ancestor (i.e., in “ABC”) became the progenitor of all extant lineages of *Torreya*. In the mid-Eocene, a vicariance event separated the EA and NA lineages, and a vicariance event in the late Paleogene isolated NA lineages in ENA and WNA. The genus diversified within each continent, resulting in seven lineages in EA, four lineages in WNA, and one lineage in ENA. Due to extinction events during the late Paleogene and Neogene periods, five are now surviving in EA, one in ENA (*T. taxifolia*), and one in WNA (*T. californica*). Within the EA clade, the results suggested two
dispersals to Europe during the late Paleogene and the late Neogene, respectively, but those European lineages did not persist to the present time. The results from analyses coding the ancestral range of *Amentotaxus* as “C” or “ACE” were largely congruent, except for minor variations in the probability of the most likely ranges at some nodes near the root (Figs. S21, S22).

4 Discussion

4.1 Utility of the RADADOR pipeline in reducing missing data in RAD-seq data

RAD-seq is widely regarded as an effective method in molecular phylogenomics, and our study supports previous work showing its utility in resolving both deep and shallow phylogenetic relationships (Cariou et al., 2013; Hou et al., 2016; Eaton et al., 2017; Leaché & Oaks, 2017; Lecaudef et al., 2018; MacGuigan & Near, 2019; Du et al., 2020; Mu et al., 2020). One highly discussed shortcoming of RAD-seq for deep phylogenetic resolution is that mutations in restriction sites within lineages under study lead to missing data, and the probability of mutation increases with time and evolutionary distance (Quek & Huang, 2021). Notably, the effects of missing data, or ADO, in RAD-seq may be limited for relatively closely related taxa (e.g., with divergences up to ca. 80 Ma [Herrera & Shank, 2016]), as shown in previous work on *Castanea* (Zhou et al., unpublished data), *Aesculus* (Du et al., 2020), *Hamamelis* (Zhou et al., unpublished data), and *Cornus* (Lindelof et al., 2020) from our research group.
However, in these studies, we found that missing data among more deeply diverged outgroup taxa were very high, ranging from 98.4% in Fagus in the study of Castanea (a total of 1290 loci from M50; Zhou et al., unpublished data) to 41.8% in Parrotiopsis in the study of Hamamelis (a total of 1595 loci from M50; Zhou et al., unpublished data). This limitation of RAD-seq, especially for outgroups, is also noted in other studies (e.g., Leaché & Oaks, 2017). In this study, we detected only 47 loci from Taxus in the ipyradM50 matrix, which contained a total of 2731 loci, meaning that 98.2% of loci were missing for the outgroup.

Our novel RADADOR pipeline allowed us to resolve the missing data problem and generate a more robust outgroup in two ways. First, we were able to increase the number of loci representing Taxus wallichiana var. chinensis from 47 (1.7%) to 317 (11.6%). Additionally, we added 574 (21.0%) loci from the more closely related outgroup species, Amentotaxus argotaenia, which we were unable to sequence, exclusively from publicly available data. Addition of these data from RADADOR to our ipyrad data resulted in different relationships among the three major clades, Torreya californica–Torreya taxifolia, T. yunnanensis–Torreya fargesii, and Torreya grandis–Torreya nucifera–Torreya jackii (Figs. 2, S10–S16), in the phylogeny and increased confidence in our phylogenetic results. Consequently, the increased confidence provided a stronger basis for the downstream analyses, especially for the inference of historical biogeography. These results also indicated a substantial negative impact of large amounts of missing data in outgroups on phylogenetic inference and highlight the value of the RADADOR pipeline for mitigating ADO.

In general, the effectiveness of any pipeline at discovering shared loci between RAD-seq and transcriptome data for species with missing data depends on the specific transcriptomes of the species to be queried, the RAD-seq dataset, and the bioinformatics methods (e.g., Bowtie2, Langmead & Salzberg, 2012 or BLAST, Altschul et al., 1990). For example, a previous attempt (Rodríguez et al., 2017) to mine RAD-seq loci from transcriptome data using Bowtie2 yielded only a few shared loci. However, in the RADADOR pipeline, we implemented balanced parameters for BLAST to detect matching results of any length in the transcriptomes, thus leading to the discovery of many more matching sequences in our RAD-seq data, itself generated using a careful experimental procedure targeting fragments 400–600 bp in size. These matching transcriptome data filled in substantial amounts of the missing data in our outgroup species (see Methods). Therefore, we believe that our simple RADADOR pipeline is valuable for supplementing RAD-seq data matrices in phylogenomic studies, especially when coupled with a robust approach to upstream laboratory and bioinformatics procedures. The RADADOR pipeline is also applicable for ingroup taxa, although we did not apply it to fill in missing data for species of Torreya (but see Methods for our efforts to apply it to extracting mitochondrial genes from our RAD-seq data). We focused on rescuing missing data in the outgroup taxa because of the prominent ADO problem associated with outgroups, and filing in missing data for a small number of loci in species of Torreya would not have a major impact on the phylogenetic reconstruction, given the large number of loci available, as shown in our results (see radador Matrices in Figs. S10–S16). In our case, it would have been difficult to decide which individuals of species of Torreya should receive the sequences of loci rescued from the public data. We did not think it was reasonable to add the same rescued sequences to all individuals of a species and did not want to reduce the sampling to one per species. Furthermore, very limited transcriptome and genome data were available for species of Torreya. For these reasons, we did not apply the pipeline for our ingroup taxa. Nonetheless, we do suggest applying the method for ingroup taxa as needed and when sources of data are available.

4.2 Species relationships, taxonomy, and conflicts among gene loci and between plastid and nuclear genomes

Species relationships within Torreya have been addressed in several previous molecular phylogenetic studies using one or a few genes and whole-plastid genomes (Li et al., 2001b; Hao et al., 2008; Leslie et al., 2012; Wang et al., 2019; Zhang et al., 2019). However, these studies yielded incongruent phylogenies with mostly low support. Our study based on genome-wide data from the chloroplast and nuclear genomes strongly supports a close relationship between T. californica and T. taxifolia from NA, between T. fargesii and T. yunnanensis, and between T. grandis and T. nucifera, with the latter two species pairs from EA forming a sister group (Figs. 3, 5A, 5B). The relationship of the fifth EA species, T. jackii, differed between the chloroplast and nuclear phylogenies as the sister to the remainder of Torreya or as the sister of T. grandis–T. nucifera, respectively (Figs. 3, 5A, 5B). Notably, our limited mitochondrial gene sequences obtained from the Hyb-Seq data showed little power for resolving species relationships in Torreya. The phylogenies reconstructed using our nuclear and chloroplast data are largely congruent with the result from Wang et al. (2019) based on ITS data and seven chloroplast regions, but, to some extent, different from those from other previous studies (Li et al., 2001b; Hao et al., 2008; Leslie et al., 2012; Zhang et al., 2019), where relationships were usually not well resolved or not well supported. The chloroplast phylogeny reconstructed in our study is different from the plastome phylogeny of Torreya based on the whole-genome sequences inferred in Zhang et al. (2019). Specifically, Zhang et al. (2019) showed that T. nucifera and T. californica were united with T. fargesii and T. grandis, respectively, while our study shows them with T. grandis and T. taxifolia, respectively. The placements of T. nucifera and T. californica found in our study were consistent between the nuclear phylogeny, which included multiple accessions for each species, and the chloroplast phylogeny. The samples of these two species used in Zhang et al. (2019) were from botanical gardens and have no clear information on origin. We suspect that the samples of both T. nucifera and T. californica in Zhang et al. (2019) were misidentified because we were able to examine a photo of the voucher for the T. nucifera and confirmed that it was not T. nucifera. Based on our phylogeny, the samples representing T. nucifera and T. californica in Zhang et al. (2019) were likely T. fargesii and T. nucifera, respectively.

Overall, our phylogenetic results do not support the classification of Torreya into the two traditionally recognized sections: sect. Torreya, including species with ruminant albumen, and sect. Nuciferae, including species with
non-ruminant albumen (Hu, 1927; Cheng & Fu, 1978; Kang & Tang, 1995). In the phylogenetic trees, neither species with ruminant albumen nor species with non-ruminant albumen formed monophyletic groups (Figs. 3A, 5A, 5B). Instead, the molecular phylogenies indicate that the rumination of albumen in Torreya is likely to be evolutionarily labile. This agrees with all previous phylogenetic studies (Li et al., 2001b; Hao et al., 2008; Leslie et al., 2012; Wang et al., 2019; Zhang et al., 2019).

Within the genus, there has been considerable taxonomic uncertainty regarding the status of both T. fargesii and T. yunnanensis. Some taxonomists have recognized both species (e.g., Cheng & Fu, 1978), while others included them within a more broadly defined T. grandis (Silba, 1984). Our data clearly support the recognition of T. fargesii and T. yunnanensis, and both were distinct from T. grandis. Based on the RAD-seq and chloroplast gene data (Figs. 3–5), samples referable to T. fargesii and T. yunnanensis formed mutually monophyletic groups diverging from each other 13.7 Ma in the Oligocene. These divergence times are almost assuredly long enough for species of Torreya to achieve speciation (Baker & Bradley, 2006), even though species of the genus are long-lived (Howard, 1992) and have lengthy generation times (Verdú, 2002). Furthermore, our data also show that T. grandis is more closely related to T. nucifera than it is to T. fargesii and T. yunnanensis (Figs. 3–5). These relationships were resolved differently without strong support in previous molecular studies. For example, in Hao et al. (2008), T. grandis was shown to be the sister of T. fargesii–T. yunnanensis based on data from five chloroplast regions but was shown to be the sister of all other Torreya species except T. jackii based on ITS sequences. In Li et al. (2001b), T. grandis was resolved as the sister to T. jackii, while T. nucifera was sister to T. fargesii–T. yunnanensis based on ITS sequences. Furthermore, Zhang et al. (2019) resolved T. grandis as the sister of T. californica, and T. nucifera as the sister of T. fargesii–T. yunnanensis, but this was likely due to misidentification of species as discussed above.

The conflicting, but strongly supported, placements of T. jackii between the plastid gene tree (Fig. 3B) and the concatenation-based tree from RAD-seq data (Fig. 3A) were also found in previous studies (Hao et al., 2008; Leslie et al., 2012, 2020; Wang et al., 2019; Zhang et al., 2019), and the conflict is difficult to explain. It could be a result of differential lineage sorting of a dimorphic ancestral chloroplast genome among species and clades of Torreya or the result of chloroplast capture by T. jackii from Amentotaxus (Fig. 6). The differential lineage-sorting hypothesis requires that both T. jackii and the rest of Torreya had dimorphic chloroplast genomes at the time of their divergence. Thereafter, one of the ancestral chloroplast types would have to become fixed in T. jackii, while the other type would have to become fixed in the other three clades independently. Such a biased frequency of fixation between dimorphic plastid genome types is unlikely during the random process of lineage sorting and would probably require specific selective pressures. However, we do not have evidence to invoke favorable selection for one type of plastid over the other. Unfortunately, an explanation involving hybridization may be equally difficult to reconcile. Notably, we did detect possible introgression between Amentotaxus and another species of Torreya, T. yunnanensis, based on PhylotNetworks, and inter-generic introgression has been reported in other gymnosperms (e.g., Garland & Moore, 2012). Thus, hybridization and introgression between Amentotaxus and T. jackii do not seem unreasonable on the grounds that they are too divergent. Moreover, Torreya jackii is endemic to N. Fujian, N. E. Jiangxi, and S. Zhejiang in eastern China, while two species of Amentotaxus occur within eastern and southeastern China. Specifically, A. argotaelia is found in Fujian, Guangdong, Guangxi, Jiangxi, Zhejiang, and Taiwan, and A. formosana is endemic to SE Taiwan. All three species occur in forested areas at roughly similar elevations ranging from 300 to 1300 m. The areas where T. jackii and two species of Amentotaxus occur are either presently overlapping or have experienced geological and floristic connection through time, particularly considering that Taiwan became separated from the Chinese mainland only ca. 1.5 Ma (Osozawa et al., 2012; Tang et al., 2013). Thus, T. jackii may have had the opportunity to hybridize with Amentotaxus. However, if the plastid genome capture did occur in T. jackii, it must have occurred in deep time (early age of formation of T. jackii), followed by subsequent long-term evolution of the captured genome, which must have evolved convergently with Torreya, but divergently with Amentotaxus. Consequently, the plastid genome in T. jackii has become substantially distinct from Amentotaxus and closer to Torreya, as shown in the unrooted phylogeny of the plastid genome (Fig. S23). Therefore, the chloroplast DNA capture hypothesis does not sound more plausible than the differential lineage sorting scenario, but nonetheless may be tested in future studies with more sampling from the two genera and applying new methods as they are developed.

The low support for the relationship of T. jackii to T. grandis–T. nucifera in the coalescent-based species trees also indicates conflicts among gene loci of the nuclear genome regarding the relationship among T. jackii, T. yunnanensis–T. fargesii, and T. grandis–T. nucifera (Figs. 5A, 5B). Our total evidence dating (Figs. 6, S20) revealed rapid divergence of the three lineages in the late Eocene, confirming that ILS was likely the cause for the phylogenetic conflicts among loci. However, our results from the D-statistic test (Table S8) and PhylotNetworks both support introgression between T. jackii and T. grandis (Fig. S19). The result from PhylotNetworks suggested a hybrid origin of T. nucifera from T. jackii and T. grandis with asymmetrical backcrosses biased toward T. grandis (90% and 10%, respectively). Torreya grandis and T. jackii have overlapping distributions along the southeastern coast of China, and their ability to hybridize is supported by reports of naturally occurring hybrids (possibly also including T. grandis var. jiulongshanensis, Kou et al., 2017). We speculate that T. nucifera might have originated in the late Oligocene (Figs. 6, S20) along the southeastern coast of China, within the range of its parental species. During this time, Japan, where the species occurs in modern times, was still connected to the Asian continent (Barnes, 2003). Thus, the species later became extinct on the mainland and restricted to Japan.

The inconsistent phylogenetic positions of T. jackii among data sets and among phylogenetic methods highlights the
limitation of using data from a single genome for phylogenetic analysis or relying exclusively on a concatenated approach to generating a species tree. It further demonstrates the value of rigorous phylogenetic analyses that include coalescent-based methods of analysis and utilize both nuclear and plastid genomes (including mitochondrial genomes; e.g., Semerikova & Semerikov, 2014; Ran et al., 2015) to obtain a more accurate inference of the evolutionary history of plant lineages. Results from our study support an important role of ancient interspecific gene introgression and reticulate evolution in the diversification of gymnosperms. Overall, hybridization and introgression in gymnosperms appear to be rare based on the limited number of polyploids within most lineages, notwithstanding Ephedra, Juniperus, and Ginkgo (Koshoo, 1961; Wang & Ran, 2014; Ickert-Bond et al., 2020; Ohri, 2021). Nevertheless, many gymnosperm lineages contain several naturally occurring hybrids or establish hybrid zones with congers (e.g., Torreya as discussed above; Picea in Sun et al., 2018; Cupressus Li et al., 2020; Pinus in Delgado et al., 2007; Watano et al., 2004, and Williard et al., 2009; and Zamia in Pérez-Farrera et al., 2012). Thus, while hybridization and introgression may play less of a role in shaping the evolutionary history of gymnosperms than angiosperms (Ickert-Bond et al., 2020), its evolutionary footprint is still likely present in some of these relatively ancient lineages.

4.3 Biogeography of Torreya and insights into the EA–ENA floristic disjunction

Our results suggest a highly dynamic, ancient biogeographic history of Torreya and support a Paleogene origin of the disjunction of the modern species in EA, ENA, and WNA via vicariance. A wide distribution of the ancestors of the modern Torreya across Asia and NA during the late Cretaceous to the mid-Eocene suggests an important role of the Bering land bridge as a corridor for connecting populations of Torreya between the two continents during this geological period. The Bering land bridge was available as a land route up to approximately 4.7 Mya (Graham, 2018) and may have been important for many gymnosperm lineages (Wang & Ran, 2014). Occurrences of fossils of Torreya within the Beringian subcontinent during the Eocene and more recently (Graham, 2018) lend additional support for the role of the Bering land bridge as a corridor for migration (see Edwards et al., 2018). Overall, our biogeographic and divergence time dating analyses reveal a very complicated history of Torreya including vicariance, dispersals, and extensive extinctions in all continental regions of the Northern Hemisphere throughout the history of the genus (Fig. 6).

Previous studies of genera of gymnosperms that have the EA–NA disjunction found that they more often originated in NA, while many such angiosperm lineages originated in EA (Table 1) (Wang & Ran, 2014). Thus, “out of NA” as a major pattern for gymnosperms has been proposed as a hypothesis, in contrast to the “out of Asia” hypothesis for angiosperms (Donoghue & Smith, 2004; Wen et al., 2010, 2016). However, our study of Torreya, which robustly integrates fossil species, does not represent either “out of NA” or “out of Asia”. Instead, the genus was widespread across the Northern Hemisphere (“ABCE”) before it began diversification in the Jurassic, and the modern disjunct distribution was the result of vicariance of a widespread ancestor (“ABC”) in the Paleogene (Fig. 6). These results for Torreya are consistent with recent studies of a few woody genera of angiosperms for which fossils were integrated into analyses (Castanea and Hamamelis [Zhou et al., unpublished data]; Nyssa [Zhou et al., 2020]; and two lineages of Cornus [Lindelof et al., 2020]). However, they differ from previous results for Torreya that considered only the ancestral origins of the extant crown clade (i.e., not including fossils) and found that it was present either only in EA (i.e., “out of Asia”) (Hao et al., 2008) or was widespread between Asia and WNA (Zhang et al., 2019). Therefore, our results highlight that inclusion of fossils is critical for addressing long-standing biogeographic questions and generating new and more realistic hypotheses.

The ancient, widespread range and diversification of Torreya in the Jurassic and Cretaceous found in our study (Fig. 6) agree with previous inferences from the fossil record that gymnosperm lineages underwent considerable diversification during the Jurassic or early Cretaceous coupled with global range expansions (Crisp & Cook, 2011; Wang & Ran, 2014; Birks, 2020). However, it is widely believed that angiosperms later effectively outcompeted gymnosperms in many types of habitats and, thus, led to increasingly restrictive geographic ranges for gymnosperms (Biffin et al., 2012; Brodribb et al., 2012). The extensive extinction of early lineages of Torreya may be partly attributed to competition from angiosperms that began radiation in the early Cretaceous (Crisp & Cook, 2011).

A major characteristic of the EA–NA floristic disjunction is the species diversity anomaly, which is biased toward EA. Specifically, disjunct genera occurring in the two areas have twice as many species in EA as in NA (Qian & Ricklefs, 1999, 2000). Greater speciation rates and longer times to diversify in EA have been the two major hypotheses invoked to explain this pattern (Guo & Ricklefs, 2000; Qian & Ricklefs, 2000; Donoghue & Smith, 2004; Xiang et al., 2004). There has been limited evidence from previous studies supporting both hypotheses. In Torreya, EA has five extant species, while NA has two. Based on our phylogenetic results with complete sampling of extant species and inclusion of all reliable fossils, greater species diversity in EA today is likely a combined result of greater speciation and less extinction within that region (Fig. 6).

5 Conclusions

In this study, we demonstrated the impacts of ADO from RAD-seq on the phylogenomic study of Torreya and developed a pipeline to mitigate the problem using existing data in public resources. Using our pipeline, RADADOR, we were able to obtain 600% more loci (317 vs. 47) for the outgroup taxon, Taxus wallichiana var. chinensis, and add 574 loci for the outgroup taxon, Amentotaxus argotaenia, for which we had no RAD-seq data. Our rigorous phylogenetic analyses resolved robust nuclear and plastid phylogenies of the genus and revealed conflicting placements for T. jackii. The conflict was probably the result of introgression, which is supported by a D-statistic test and PhyloNetworks, though we cannot rule out incomplete lineage sorting. Our biogeographic analysis including fossils of Torreya yielded a
considerably more complex pattern of range evolution than previous studies and supports a vicariant origin of the disjunct distribution of Torreya in EA and NA, although “out of NA” is more common among gymnosperms. Our results highlight the critical importance of integrating fossil data into phylogenies and historical biogeographic analyses to test long-standing hypotheses on the origins of intercontinental disjuncts. Further, our results highlight the value of the RADADOR pipeline to reduce missing data in phylogenomic studies using RAD-seq, and its application may extend to other types of data generated using high-throughput sequencing methods.

Acknowledgements

The study was supported by an NSF grant from the United States DEB—1442161 to Q. Y. (J.) Xiang. This work also benefited from the USDA National Institute of Food and Agriculture, Hatch project 02718, and a Shiu-Ying Hu Student/Post-Doctoral Exchange Award to W. Zhou from the Arnold Arboretum. We are very grateful to Solts’ Lab and M. Sun at University of Florida for assistance in the Hyb-Seq experiments. We are also grateful to H. Wang of the Wuhan Botanical Garden, Chinese Academy of Sciences, and to D. Knotts, P. Brownless, and M. Gardner of the Royal Botanical Garden, Chinese Academy of Sciences, and to D. Knotts, P. Brownless, and M. Gardner of the Royal Botanical Garden at Edinburgh for their assistance with accessing specimens used as sources for DNA in a prior study. Further, we thank the Solts lab at the Florida Museum of Natural History, Y. P. Zhao, P. Li, and R. S. Lu at Zhejiang University, Cao lab at the Kunming Institute of Botany, the Chinese Academy of Sciences, C. Chen from Hangzhou Botanical Garden, S. Sakaguchi from Kyoto University, J. C. Raulston Arboretum, Arnold Arboretum, University of Washington Botanic Gardens, P. Jones at Sarah Duke Garden, and the UNC herbarium for providing leaf tissues. In addition, we are grateful to J. Thorne and J. Sohigian from North Carolina State University for discussion of the RADADOR pipeline.

Conflict of Interests

No author had any conflicts of interest associated with any aspect of the manuscript.

Data Availability Statement

The data sets generated and analyzed during the current study are available in the Dryad repository (https://doi.org/10.5061/dryad.08kpr525). Sequence data that support the findings of this study have been deposited in the Short Read Archive (SRA) with the BioProject accession number: PRJNA747021.

References

Adelalou KF, Zhang X, Qu X, Landis JB, Shen J, Sun Y, Meng A, Sun H, Wang H. 2020. Plastome phylogenomic and biogeographical study on Thuja (Cupressaceae). BioMed Research International 2020: 1–13.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3): 403–410.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics 17(3): 81–92.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Molecular Ecology 22(11): 3179–3190.

Baker RJ, Bradley RD. 2006. Speciation in mammals and the genetic species concept. Journal of Mammalogy 87(4): 643–662.

Barnes GL. 2003. Origins of the Japanese Islands: The new “Big Picture”. Japan Review 15: 3–50.

Biffin E, Brodribb TJ, Hill RS, Thomas P, Lowe AJ. 2012. Leaf evolution in Southern Hemisphere conifers tracks the angiosperm ecological radiation. Proceedings of the Royal Society B: Biological Sciences 279(1727): 341–348.

Birks HJB. 2020. Angiosperms versus gymnosperms in the Cretaceous. Proceedings of the National Academy of Sciences USA 117(49): 30879–30881.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for illumina sequence data. Bioinformatics 30(15): 2114–2120.

Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLoS Computational Biology 10(4): e1003537.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Oglivie HA, du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siverson I, Drummond AJ. 2019. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Computational Biology 15(4): e1006650.

Boufford DE, Spongberg SA. 1983. Eastern Asian-eastern North American phytogeographical relationships—a history from the time of Linnaeus to the twentieth century. Annals of the Missouri Botanical Garden 70(3): 423–439.

Bresadola L, Link V, Buerkle CA, Lexer C, Wegmann D. 2020. NeighborNet, SC and LPA: New methods for the construction of phylogenetic networks. Molecular Biology and Evolution 25(12): 255–265.

Burke JC. 1975. Human use of the California nutmeg tree, Torreya californica, and of other members of the genus. Economic Botany 29(2): 127–140.

Cariou M, Duret L, Charlat S. 2013. Is RAD-seq suitable for phylogenetic inference? An in silico assessment and optimization. Ecology and Evolution 3(4): 846–852.

Cariou M, Duret L, Charlat S. 2016. How and how much does RAD-seq bias genetic diversity estimates? BMC Evolutionary Biology 16(1): 240.

Chen C-H, Huang J-P, Tsai C-C, Chaw S-M. 2009. Phylogeny of Calocedrus (Cupressaceae), an eastern Asian and western North American disjunct gymnosperm genus, inferred from nuclear ribosomal nrITS sequences. Botanical Studies 50: 9.

Cheng WC, Fu KL. 1978. Taxaceae. In: Flora Reipublicae Popularis Sinicae. Beijing: Science Press. 7: 437.
Cheng Y, Nicolson RG, Tripp K, Chaw S-M. 2000. Phylogeny of Taxaceae and Cephalotaxaceae genera inferred from chloroplast matk gene and nuclear rDNA ITS region. Molecular Phylogenetics and Evolution 14(3): 353–365.

Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30(23): 3317–3324.

Crisp MD, Cook LG. 2011. Cenozoic extinctions account for the low diversity of extant gymnosperms compared with angiosperms. New Phytologist 192(4): 997–1009.

Cruaud A, Gautier M, Galan M, Foucaud J, Sauné L, Genson G, Dubois E, Nidelet S, Deuve T, Rasplus J-Y. 2014. Empirical assessment of RAD sequencing for interspecific phylogeny. Molecular Biology and Evolution 31(5): 1272–1274.

Cui Y-M, Sun B, Wang H-F, Ferguson DK, Wang Y-F, Li C-S, Yang J, Ma Q-W. 2015. Exploring the formation of a disjunctive pattern between eastern Asia and North America based on fossil evidence from Thuja (Cupressaceae). PLoS One 10(6): e0138544.

Culling KW. 1992. Design and testing of a plant-specific PCR primer for ecological and evolutionary studies. Molecular Ecology 1(4): 233–240.

DaCosta JM, Sorenson MD. 2016. ddRAD-seq phylogenetics based on nucleotide, indel, and presence–absence polymorphisms: Analyses of two avian genera with contrasting histories. Molecular Phylogenetics and Evolution 94: 122–135.

Darriba D, Taboada GL, Daoilo R, Posada D. 2012. jModelTest 2: More models, new heuristics and parallel computing. Nature Methods 9(8): 772.

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML. 2013. Special features of RAD sequencing data: Implications for genotyping. Molecular Ecology 22(11): 3151–3164.

Delgado P, Salas-Lizana R, Vázquez-Lobo A, Wegier A, Anzidei M, Alvarez-Buylla ER, Vendramin GG, Piñero D. 2007. Introgressive hybridization in Pinus montezumae Lamb and Pinus pseudostrobus Lindl. (Pinaceae): Morphological and molecular (cpSSR) evidence. International Journal of Plant Sciences 168(6): 861–875.

Díaz-Arce N, Arribazalaga H, Murua H, Irigoien X, Rodríguez-Espeleta N. 2016. RAD-seq derived genome-wide nuclear markers resolve the phylogeny of tunas. Molecular Phylogenetics and Evolution 102: 202–207.

Díaz-Arce N, Rodríguez-Espeleta N. 2019. Selecting RAD-seq data analysis parameters for population genetics: The more the better? Frontiers in Genetics 10: 533.

Donoghue MJ, Smith SA. 2004. Patterns in the assembly of temperate forests around the Northern Hemisphere. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359(1450): 1633–1644.

Doyle J. 1991. DNA protocols for plants. In: Hewitt GM, Johnston AWB, Young JPW eds. Molecular techniques in taxonomy. NATO ASI series (Series H: Cell biology). Berlin, Heidelberg: Springer. 57. https://doi.org/10.1007/978-3-642-83962-7_18

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4(5): e88.

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

Eaton DAR. 2014. PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. Bioinformatics 30(13): 1844–1849.

Eaton DAR, Overcast I. 2020. ipyrad: Interactive assembly and analysis of RADseq datasets. Bioinformatics 36(8): 2592–2594.

Eaton DAR, Ree RH. 2013. Inferring phylogeny and introgression using RADseq data: An example from flowering plants (Pedicularis: Orobanchaceae). Systematic Biology 62(5): 689–706.

Eaton DAR, Spriggs EL, Park B, Donoghue MJ. 2017. Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. Systematic Biology 66(3): 399–412.

Eckert AJ, Hall BD. 2006. Phylogeny, historical biogeography, and patterns of diversification for Pinus (Pinaceae): Phylogenetic tests of fossil-based hypotheses. Molecular Phylogenetics and Evolution 40(1): 166–182.

Edwards ME, Lloyd A, Armbruster WS. 2018. Assembly of Alaska-Yukon boreal steppe communities: Testing biogeographic hypotheses via modern ecological distributions: Alaska-Yukon boreal steppe assembly. Journal of Systematics and Evolution 56(5): 466–475.

Feng Y-Y, Shen T-T, Shao C-C, Du H, Han J-H, Wang X-Q. 2021. Phylotranscriptomics reveals the complex evolutionary and biogeographic history of the genus Tsuga with an East Asian-North American disjunct distribution. Molecular Phylogenetics and Evolution 157: 107066.

Fernández-Mazuecos M, Mellors G, Vigalondo B, Sáez L, Vargas P, Glover BJ. 2018. Resolving recent plant radiations: Power and robustness of genotyping-by-sequencing. Systematic Biology 67(2): 250–268.

Florin R. 1963. The distribution of conifer and taxad genera in time and space. Acta Horti Bergiani 20(4): 121–312.

Fu LG, Li N, Mill RR. 1999. Taxa ice. In: eFloras (2008). Published on the Internet http://www.efloras.org [accessed 12 April 2021]. Missouri Botanical Garden, St. Louis, MO & Harvard University Herbaria, Cambridge, MA.

Garland MA, Moore G. 2012. Hesperotropism, a new nothogenus for intergeneric crosses between Hesperocyparis and Callitropsis (Cupressaceae), and a review of the complicated nomenclatural history of the Leyland cypress. Taxon 61(3): 667–670.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, 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.

Gavryushkina A, Heath TA, Ksepkka DT, Stadler T, Welch D, Drummond AJ. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic Biology 66(1): 57–73.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q. 2011. Trinity: Reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology 29(7): 644.

Graham A. 1999a. Late Cretaceous and Cenozoic history of North American vegetation. New York: Oxford University Press.

Graham A. 1999b. The Tertiary history of the northern temperate element in the northern Latin American biota. American Journal of Botany 86(1): 32–38.

Graham A. 2018. The role of land bridges, ancient environments, and migrations in the assembly of the North American flora: Land bridges and North American floristic assembly. Journal of Systematics and Evolution 56(5): 405–429.

Guo Q, Ricklefs RE. 2000. Species richness in plant genera disjunct between temperate eastern Asia and North America. Botanical Journal of the Linnean Society 134(3): 401–423.
Hao DC, Xiao PG, Huang B, Ge GB, Yang L. 2008. Interspecific relationships and origins of Taxaceae and Cephalotaxaceae revealed by partitioned Bayesian analyses of chloroplast and nuclear DNA sequences. Plant Systematics and Evolution 276(1-2): 89–104.

Harris A, Wen J, Xiang Q-Y(J). 2013. Inferring the biogeographic origins of inter-continental endemic taxa using a Bayes-DIVA approach: Biogeography of continental endemic species. Journal of Systematics and Evolution 51(2): 117–133.

Harris TM. 1976. Two neglected aspects of fossil conifers. American Journal of Botany 63(6): 902–910.

Havill NP, Campbell CS, Vining TF, LePage B, Bayer RJ, Donoghue MJ. 2008. Phylogeny and biogeography of Tsuga (Pinaceae) inferred from nuclear ribosomal ITS and chloroplast DNA sequence data. Systematic Botany 33(3): 478–489.

Herrera S, Shank TM. 2016. RAD sequencing enables unprecedented phylogenetic resolution and objective species delimitation in recalcitrant divergent taxa. Molecular Phylogenetics and Evolution 100: 70–79.

Hils MH. 1996. Taxaceae. In: Flora of North America Editorial Committee eds. Flora of North America north of Mexico [Online]. 21. New York and Oxford. Vol. 2. Available from http://www.efloras.org/flora taxon.aspx?flora_id=1126taxon_id=10871 [accessed 08 June 2021]

Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. 2018. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nature Methods 14(6): 587–589.

Kan S-L, Shen T-T, Gong P, Ran J-H, Wang X-Q. 2020. The complete mitochondrial genome of Taxus cuspidata (Taxaceae): Eight protein-coding genes have transferred to the nuclear genome. BMC Evolutionary Biology 20(1): 10.

Kang N, Tang Z. 1995. Studies on the taxonomy of the genus Torreya. Bulletin of Botanical Research 15(3): 349–362.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30(4): 772–780.

Kearse M, Moir R, Wilson A, Stones–England T, Stein L, simmer P, Duran R, Thew J, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649.

Kimura T. 1987. Recent knowledge of Jurassic and Early Cretaceous floras in Japan and phytogeography of this time in East Asia. Tokyo Gakugei Daigaku Kiyo, Dai 4-Bu: Sugakuka, Shizen Kagaku 39: 87–115.

Klymuik AA, Stockey RA. 2012. A lower Cretaceous (Valanginian) seed cone provides the earliest fossil record for Picea (Pinaceae). American Journal of Botany 99(6): 1069–1082.

Kossho TN. 1961. Chromosome numbers in gymnosperms. Silvae Genetica 69(1): 13–19.

Kou Y-X, Xiao K, Lai X-R, Wang Y-J, Zhang Z-Y. 2017. Natural hybridization between Torreya jackii and T. grandis (Taxaceae) in southeast China: Natural hybridization in Torreya. Journal of Systematics and Evolution 55(1): 25–33.

Kang N, Tang Z. 1995. Studies on the taxonomy of the genus Torreya. Bulletin of Botanical Research 15(3): 349–362.

Kato K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30(4): 772–780.

Kearse M, Moir R, Wilson A, Stones–England T, Stein L, simmer P, Duran R, Thew J, Ashton B, Meintjes P, Drummond A. 2012. Geneious basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649.

Kimura T. 1987. Recent knowledge of Jurassic and Early Cretaceous floras in Japan and phytogeography of this time in East Asia. Tokyo Gakugei Daigaku Kiyo, Dai 4-Bu: Sugakuka, Shizen Kagaku 39: 87–115.

Klymuik AA, Stockey RA. 2012. A lower Cretaceous (Valanginian) seed cone provides the earliest fossil record for Picea (Pinaceae). American Journal of Botany 99(6): 1069–1082.

Kossho TN. 1961. Chromosome numbers in gymnosperms. Silvae Genetica 69(1): 13–19.

Kou Y-X, Xiao K, Lai X-R, Wang Y-J, Zhang Z-Y. 2017. Natural hybridization between Torreya jackii and T. grandis (Taxaceae) in southeast China: Natural hybridization in Torreya. Journal of Systematics and Evolution 55(1): 25–33.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9(4): 357–359.

Leaché AD, Oaks JR. 2017. The utility of single nucleotide polymorphism (SNP) data in phylogenetics. Annual Review of Ecology, Evolution, and Systematics 48(1): 69–84.

Lecaudery LA, Schiwenken UF, Osinov AG, Taylor EB, Bernatchez L, Weiss SJ. 2018. Inferring phylogenetic structure, hybridization and divergence times within Salomoniae (Teleootei: Salomoniae) using RAD-sequencing. Molecular Phylogenetics and Evolution 124: 82–99.

Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S. 2012. Hemispheric-scale differences in conifer evolutionary dynamics. Proceedings of the National Academy of Sciences USA 109(40): 16217–16221.

Lewis PO. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50(6): 913–925.

Li J, Davis CC, Del Tredici P, Donoghue MJ. 2001a. Phylogeny and biogeography of Taxus (Taxaceae) inferred from sequences of the internal transcribed spacer region of nuclear ribosomal DNA. Harvard Papers in Botany 6(1): 267–274.

Li J, Davis CC, Donoghue MJ, Kelley S, Tredici PD. 2001b. Phylogenetic relationships of Torreya (Taxaceae) inferred from evolution of global pines. Proceedings of the National Academy of Sciences USA 118(20): e2002230218.
sequences of nuclear ribosomal DNA its region. Harvard Papers in Botany. 6(1): 275–281.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

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

Li J, Gao L, Tao K, Su Y, Wang T. 2016. The complete chloroplast genome sequence of Amentotaxus argotaenia (Taxaceae). Mitochondrial DNA Part A 27(4): 2919–2920.

Li J, Milne RI, Ru D, Mao J, Tao W, Zhang L, Xu J, Liu J, Mao K. 2020. Allopatric divergence and hybridization within Cupressus chengiana (Cupressaceae), a threatened conifer in the northern Hengduan Mountains of western China. Molecular Ecology 29(7): 1250–1266.

Lin H-Y, Hao Y-J, Li J-H, Fu C-X, Soltis PS, Soltis DE, Zhao Y-P. 2019. Phylogenomic conflict resulting from ancient introgression following species diversification in Stewardia s.l. (Theaceae). Molecular Phylogenetics and Evolution 135: 1–11.

Lindelof K, Lindo JA, Zhou W, Ji X, Xiang Q-Y(J). 2020. Phylogenomics, biogeography, and evolution of the blue- or white-fruited dogwoods (Cornus)—Insights into morphological and ecological niche divergence following intercontinental geographic isolation. Journal of Systematics and Evolution 58(3): 604–645.

Little DP. 2006. Evolution and circumscription of the true cypresses (Cupressaceae: Cupressus). Systematic Botany 31(3): 461–480.

MacGuigan DJ, Near TJ. 2019. Phylogenetic signatures of ancient introgression in a rogue lineage of darters (Teleostei: Percidae). Systematic Biology 68(2): 329–346.

Malinsky M, Matschiner M, Svardal H. 2021. DSuite—Fast D-statistics and related admixture evidence from VCF files. Molecular Ecology Resources 21(2): 584–595.

Manchester SR, Tiffney BH. 2001. Integration of paleobotanical and neobotanical data in the assessment of phyogeographic history of holartic angiosperm clades. International Journal of Plant Sciences 162(S6): S19–S27.

Mann P. 2007. Overview of the tectonic history of northern Central America. Geological Society of America Special Papers 428: 1–19.

Mao K, Hao G, Liu J, Adams RP, Milne RI. 2010. Diversification and biogeography of Juniperus (Cupressaceae): Variable diversification rates and multiple intercontinental dispersals. New Phytologist 188(1): 254–272.

Mastretta-Yanes A, Arrigo N, Alvarez N, Jorgensen TH, Piñero D, Emerson BC. 2015. Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Molecular Ecology Resources 15(1): 28–41.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66(2): 526–538.

Melton AE, Chen S, Zhao Y, Fu C, Xiang Q(Jenny), Cheng S, Wong GK-S, Soltis PS, Soltis DE, Gitzendanner MA. 2020. Genetic insights into the evolution of genera with the eastern Asia–eastern North America floristic disjunction: A transcriptomics analysis. American Journal of Botany 107(12): 1736–1748.

Miller MA, Pfeiffer W, Schwartz T. 2011. The CIPRES science gateway: A community resource for phylogenetic analyses. Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery. 1–8.

Möller M, Liu J, Li Y, Li J-H, Ye L-J, Mill R, Thomas P, Li D-Z, Gao L-M. 2020. Repeated intercontinental migrations and recurring hybridizations characterise the evolutionary history of yew (Taxus L.). Molecular Phylogenetics and Evolution 153: 106952.

Mu X-Y, Tong L, Sun M, Zhu Y-X, Wen J, Lin Q-W, Liu B. 2020. Phylogeny and divergence time estimation of the walnut family (Juglandaceae) based on nuclear RAD-Seq and chloroplast genome data. Molecular Phylogenetics and Evolution 147: 106802.

Nathorst AG. 1908. Palaeobotanische Mittheilungen. 7. Über Palysyslla. Stachyotaxus und Paleotaxus 43(8): 1–16.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32(1): 268–274.

Ohri D. 2021. Polyploidy in gymnosperms—A reappraisal. Silvae Genetica 70(1): 22–38.

Osozawa S, Shinjo R, Armid A, Watanabe Y, Horiguchi T, Wakabayashi J. 2012. Palaeogeographic reconstruction of the 155 Ma synchronous isolation of the Ryukyu Islands, Japan, and Taiwan and inflow of the Kuroshio warm current. International Geology Review 54(12): 1369–1388.

Pais AL, Whetten RW, Xiang Q-Y(J). 2017. Ecological genomics of local adaptation in Cornus florida L. by genotyping by sequencing. Ecology and Evolution 7(1): 441–465.

Peng D, Wang X-Q. 2008. Reticulate evolution in Thuja inferred from multiple gene sequences: Implications for the study of biogeographical disjunction between eastern Asia and North America. Molecular Phylogenetics and Evolution 47(3): 1190–1202.

Pérez-Farrera MÁ, Vovides AP, Martínez-Camilo R, Martínez-Meléndez N, Gómez-Domínguez H, Galicia-Castellanos S. 2012. Zamia grijalvensis sp. Nov. (Zamiaceae, Cycadales) from Chiapas, Mexico with notes on hybridization and karyology. Nordic Journal of Botany 30(5): 565–570.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 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.

Qian H, Ricklefs RE. 1999. A comparison of the taxonomic richness of vascular plants in China and the United States. The American Naturalist 154(2): 160–181.

Qian H, Ricklefs RE. 2000. Large-scale processes and the Asian bias in species diversity of temperate plants. Nature 407: 180–182.

Qian H, Ricklefs RE. 2004. Geographical distribution and ecological conservatism of disjunct genera of vascular plants in eastern Asia and eastern North America. Journal of Ecology 92(2): 253–265.

Qian H, Ricklefs RE, White PS. 2005. Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. Ecology Letters 8(1): 15–22.

Quek ZBR, Huang D. 2021. Application of phylogenomic tools to unravel anthozoan evolution. Coral Reefs 2021: 3. https://doi.org/10.1007/s00338-021-02072-3
Phylogenomics and biogeography of Torreya

Xiang Q-P, Xiang Q-Y, Guo Y-Y, Zhang X-C. 2009. Phylogeny of Abies (Pinaceae) inferred from nrITS sequence data. *Taxon* 58(1): 141–152.

Xiang Q-Y, Crawford DJ, Wolfe AD, Tang Y-C, DePamphilis CW. 1998. Origin and biogeography of Aesculus L. (Hippocastanaceae): A molecular phylogenetic perspective. *Evolution* 52(4): 988–997.

Yang ZY, Ran JH, Wang XQ. 2012. Three genome phylogenies of Cupressaceae s.l.: Further evidence for the evolution of gymnosperms and Southern Hemisphere biogeography. *Molecular Phylogenetics and Evolution* 64(3): 452–470.

Yu C, Zhan X, Zhang C, Xu X, Huang J, Feng S, Shen C, Wang H. 2021. Comparative metabolomic analyses revealed the differential accumulation of taxoids, flavonoids and hormones among six Taxaceae trees. *Science Horticulture* 285: 110196.

Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19(6): 153.

Zhang X, Zhang H, Landis JB, Deng T, Meng A, Sun H, Peng Y, Wang H, Sun Y. 2019. Plastome phylogenomic analysis of Torreya (Taxaceae). *Journal of Systematics and Evolution* 57(6): 607–615.

Zhou W, Ji X, Obata S, Pais A, Dong Y, Peet R, Xiang Q(Y). 2018. Resolving relationships and phylogenetic history of the Nyssa sylvatica complex using data from RAD-seq and species distribution modeling. *Molecular Phylogenetics and Evolution* 126: 1–16.

Zhou W, Soghigian J, Xiang Q(Y). 2021. A new pipeline for removing paralogs in target enrichment data. *Systematic Biology* 71(2): 410–425.

Zhou W, Xiang Q(Y), Wen J. 2020. Phylogenomics, biogeography, and evolution of morphology and ecological niche of the eastern Asian–eastern North American Nyssa (Nyssaceae). *Journal of Systematics and Evolution* 58(5): 571–603.

**Supplementary Material**

The following supplementary material is available online for this article at http://onlinelibrary.wiley.com/doi/10.1111/jse.12716/supinfo:

**Fig. S1.** Cumulative plot of RAD-seq matrices from ipyradM20 to ipyradM80. The figure shows that the ipyradM50 tree, the ipyradM60 tree, and the ipyradM70 tree have lower cumulative frequencies (y-axis values) of nodes with low (<60) support values (x-axis) than other trees (all near the tips and within species; Figs. S5, S6), but similar or greater cumulative frequencies of nodes with higher support values (both lower and upper nodes) than other trees. The figure also shows that the ipyradM80 tree has the highest cumulative frequencies of support values greater than 90, but greater cumulative frequencies of nodes with low support values (<60) (involving backbones; Fig. S8) than the ipyradM50, ipyradM60, and ipyradM70 trees. **Figs. S2–S8.** Concatenation-based trees of Torreya from analyses of different raw RAD-seq data sets M20 to M80 including all samples using RAxML. Topologies vary among different matrices regarding relationships among four clades, *T. californica-T. taxifolia, T. fargesii-T. yunnanensis, T. grandis-T. nudifera,* and *T. jackii,* but become stable after M40. All trees also reveal unexpected positions of three samples (T. fargesii [TF_17], T. grandis [TG_14], and T. taxifolia [TT_4]), indicating misidentification or potential genetic introgression. These samples were removed from the matrices in subsequent analyses. Sample 23 (*Cephalotaxus lanceolata, as C. lanceolate CL_23 in the trees*) was also removed from the matrix in further analysis due to too much missing data. Numbers on branches indicate bootstrap support values.

**Fig. S9.** Concatenation-based tree of Torreya reconstructed from the ipyradM50 matrix using IQ-TREE program. Taxus is the outgroup. Numbers on branches are the values of ultrafast bootstrap support.

**Figs. S10–S16.** Concatenation-based tree of Torreya reconstructed from the radadorM20–radadorM80 (generated from the RADADOR pipeline) using the IQ-TREE, with both *Amentotaxus (A. argoteania)* and *Taxus (T. wallichiana var. chinensis)* as the outgroup. All trees show consistent relationships among the four clades, *T. californica-T. taxifolia, T. fargesii-T. yunnanensis, T. grandis-T. nudifera,* and *T. jackii,* but vary in support values, with the tree from radadorM50 showing the highest support. Numbers on branches show ultrafast bootstrap support.

**Fig. S17.** Concatenation-based tree of Torreya reconstructed from the radadorM50 matrix via IQ-TREE, using *Taxus* as the outgroup. All loci in the outgroup taxa were recovered from transcriptomic data using the RADADOR pipeline. Numbers on branches show ultrafast bootstrap support.

**Fig. S18.** Mitochondrial gene phylogeny of *Torreya* based on data available from our angiosperm 355 Hyb-Seq experiment and those downloaded from NCBI. The phylogeny was reconstructed using IQ-TREE with *Taxus wallichiana var. chinensis* as the outgroup. Numbers on branches indicate ultrafast bootstrap from IQ-TREE.

**Fig. S19.** Phylogenetic networks of Torreya revealed by PhyloNetworks analyses based on unlinked SNPs of radadorM20. The reticulation event was set as 0 in (a), 1 in (b), and 2 in (c). *Amentotaxus* was used as an outgroup. TC represents *Torreya californica;* TT represents *Torreya taxifolia;* TG represents *Torreya grandis;* TN represents *Torreya nudifera;* TJ represents *Torreya jackii;* TF represents *Torreya fargesii;* TY represents *Torreya yunnanensis;* and *Amento* represents *Amentotaxus.*

**Fig. S20.** Dated total evidence tree of Torreya based on RAD-seq data and morphological data reconstructed using the fossilized birth–death model in BEAST2. The median values of divergence times of nodes are indicated. Blue bars indicate the 95% HDP. Numbers on the geological time scale are in million years. Living species are in bold faces and fossils taxa are in gray color.

**Fig. S21.** Ancestral ranges reconstructed from biogeographic analyses of Torreya using the DEC model based on the dated total-evidence phylogeny with the distribution of *Amentotaxus* coded as western North America “C”. Letters at each node indicate the most likely ancestral distributions inferred from the analysis, whose probabilities are shown by non-black colors in the pie charts. The black slice in the pie chart indicates the total proportion (likelihood) of all other alternative ancestral ranges in combine. Inferred intercontinental dispersals are indicated by red arrows. Fossil taxa are shown in gray text. A: eastern Asia; B: eastern North America; C: western North America; E: Europe.

**Fig. S22.** Ancestral ranges reconstructed from biogeographic analyses of Torreya using the DEC model based on the dated total-evidence phylogeny, with the distribution of *Amentotaxus* coded as widespread in Eurasia and western North America.
America “ACE.” Letters at each node indicate the most likely ancestral distributions inferred from the analysis, whose probabilities are shown by non-black colors in the pie charts. The black slice in the pie chart indicates the total proportion (likelihood) of all other alternative ancestral ranges in combination. Inferred intercontinental dispersals are indicated by red arrows. Fossil taxa are labeled in gray text. A: eastern Asia; B: eastern North America; C: western North America; E: Europe.

Fig. S23. Unrooted plastome phylogeny of *Torreya* inferred from analyses using IQ-TREE. The plastome data were recovered from Hyb-Seq using the Angiosperms353 probe kit for *Torreya*, which failed to generate any target nuclear gene sequences.

**Table S1.** Fossils records of *Torreya* and its outgroup.
**Table S2.** Sampling information of taxa included in the study. The table includes the ddRAD ID numbers and Plastome accession numbers from both Hyb-Seq in our study and published data, voucher information, and origin source of each voucher.

**Table S3.** Morphology matrix of all *Torreya* species including fossils.
**Table S4.** Dispersal rate between geographic areas for DEC model.
**Table S5.** Reads data from Torreya RAD-seq M₅₀.
**Table S6.** Data information on Torreya RAD-seq matrices, showing comparisons of data among original matrixes (M#), modified matrixes (ipyradM#), and those improved using RADADOR (radadorM#).
**Table S7.** Chloroplast and mitochondria sequence information from Hyb-Seq off-target sequences.
**Table S8.** Results from D-statistic analyses of RAD-seq radadorM₅₀ data.