Seeing through the hedge

Seeing through the hedge: Phylogenomics of *Thuja* (Cupressaceae) reveals prominent incomplete lineage sorting and ancient introgression for Tertiary relict flora

Jialiang Li[1#], Yujiao Zhang[1#], Markus Ruhsam[2], Richard Ian Milne[3], Yi Wang[1], Dayu Wu[1], Shiyu Jia[1], Tongzhou Tao[1], Kangshan Mao[1,4*]

[1] Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, State Key Laboratory of Hydraulics and Mountain River Engineering, Sichuan University, Chengdu 610065, Sichuan, P.R. China;

[2] Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK;

[3] Institute of Molecular Plant Sciences, The University of Edinburgh, Edinburgh EH9 3JH, UK;

[4] College of Science, Tibet University, 850012 Lhasa, Xizang Autonomous Region, China

[*] Author for correspondence:

Kangshan Mao, Tel: +8613608091356; Email: maokangshan@163.com, maokangshan@scu.edu.cn

[#] Jialiang Li and Yujiao Zhang contributed equally to this work.

Running heads: Biogeographic history of *Thuja*

Abstract

The Eastern Asia (EA) - North America (NA) disjunction is a well-known biogeographic pattern of the Tertiary relict flora; however, few studies have investigated the evolutionary history of this disjunction using a phylogenomic approach. Here, we used 2,369 single copy nuclear genes and nearly full plastomes to reconstruct the evolutionary history of the small Tertiary relict genus *Thuja,* which consists of five disjunctly distributed species. The nuclear species tree strongly supported an EA clade *T. standishii*-*T. sutchuenensis* and a "disjunct clade", where western NA species *T. plicata* is sister to an EA-eastern NA disjunct *T. occidentalis*-*T. koraiensis* group. Our results suggested that the observed topological discordance among the gene trees as well as the cytonuclear discordance is mainly due to incomplete lineage sorting, probably facilitated by the fast diversification of *Thuja* around the Early Miocene and the large effective population sizes of ancestral lineages. Furthermore, ~20% of the *T. sutchuenensis* nuclear genome is derived from an unknown ancestral lineage of *Thuja*, which might explain the close resemblance of its cone morphology to that of an ancient fossil species. Overall, our study demonstrates that single genes may not resolve interspecific relationships for disjunct taxa, and that more reliable results will come from hundreds or thousands of loci, revealing a more complex evolutionary history. This will steadily improve our understanding of their origin and evolution.

Keywords: *Thuja*, disjunct distribution, eastern Asia, North America, incomplete lineage sorting, ghost introgression

1. Introduction

The eastern Asia (EA) and eastern North America (ENA) disjunction is one of the most well-known biogeographic patterns in the northern hemisphere, and the high level of similarity between these floras has been known since the time of Linnaeus (Gray, 1859; Graham, 1966; Davidse, 1983). Understanding the origin and evolution of this disjunction pattern has been a long-standing focus in biogeography and botany (Tiffney, 1985b; Wen, 1999; Donoghue et al., 2001; Milne and Abbott, 2002; Donoghue and Smith, 2004). This biogeographic disjunction is generally represented by relict lineages that were widely distributed in the Northern Hemisphere during the early to mid-Tertiary (Tiffney, 1985a; Tiffney and Manchester, 2001; Milne and Abbott, 2002). A commonly accepted explanation for the EA-ENA disjunct distribution is that members of a formerly widespread flora became extinct in western North America (WNA) and Europe due to a cooling climate and large-scale geological changes (orogenesis), while their congeners survived in both EA and ENA (Manchester, 1999; Wen, 1999; Wen et al., 2010). However, some studies have suggested that this intercontinental disjunction is unlikely to have been initiated by a single historical event (Tiffney, 1985b; Wang and Ran, 2014), and that more complex processes such as speciation, extinction, vicariance, and dispersal might have contributed to its origin (Wen, 1999; Wen et al., 2010; Feng et al., 2020; Zhang et al., 2021).

Large areas in ENA, WNA, EA, and Europe served as important refugia for a once more widespread Tertiary flora during cold periods (Milne and Abbott, 2002; Milne, 2006). If a formerly widespread taxon survived in multiple refugia, isolated at a similar time, it might undergo a radiative speciation event. Therefore, the relationships among extant lineages from different regions could be a result of random processes such as stochastic sorting of ancestral variation (Maddison and Knowles, 2006). This process is especially likely in lineages with large ancestral population sizes (Leache and Rannala, 2011; Wang et al., 2018), which could generate more complex evolutionary histories than a simple bifurcating

tree (Pease et al., 2016). Incomplete lineage sorting (ILS) can therefore be expected in formerly widespread Tertiary relict species, which now have an EA-ENA disjunction. Thus, it is possible that this remarkable biogeographic pattern is also partly the result of a random process during speciation.

Reconstruction of a robust phylogeny is required in order to understand the origin of biogeographic patterns. Until recently, biogeographic studies had to rely on often poorly resolved phylogenies of disjunct taxa due to the limited number of available molecular markers (Wen, 1999; Chan et al., 2020; Feng et al., 2020). So far, only a few phylogenies of disjunct taxa have been published which are based on hundreds or thousands of loci, e.g., *Picea* (Shao et al., 2019), *Acer* (Li et al., 2019), *Nyssa* (Zhou et al., 2020), *Corylus* (Zhao et al., 2020), and *Tsuga* (Feng et al., 2020). These phylogenomic studies showed that the disjunct taxa have more complex evolutionary histories than previously thought. Furthermore, both ILS and hybridization, which are the two great challenges in phylogenetic inference that contribute to gene tree heterogeneity (Dalquen et al., 2017; Morales-Briones et al., 2018), are commonly seen in some intercontinental disjunct lineages (Peng and Wang, 2008; Shao et al., 2019). Therefore, genome wide data are needed to resolve the phylogenetic relationships of disjunct taxa and reconstruct their complex evolutionary and biogeographic history.

The genus *Thuja* L. (Cupressaceae) provides an excellent opportunity to study the origin and evolution of intercontinental disjunct patterns with a complex history. *Thuja* is also well known for its hedging plants. The most widely cultivated species of this genus is *Thuja occidentalis* with hundreds of cultivars of varying stature, habit, foliage form and colour (Eckenwalder, 2009). *Thuja* comprises only five extant species which are disjunctly distributed in North America and eastern Asia (Fu et al., 1999). The three Asian species, *T. sutchuenensis* Franch., *T. koraiensis* Nakai and *T. standishii* (Gord.) Carr. are restricted to southwestern China, northeastern China plus the Korean Peninsula, and Japan, respectively,

and the two North American species, *T. occidentalis* L. and *T. plicata* D. Don, occur widely in eastern and western North America, respectively (Farjon, 2005). Even though there are only five species in *Thuja*, the interspecific relationships have been controversial. Based on fossil and extant seed cones, McIver and Basinger (1989) reconstructed the evolutionary history of *Thuja* and showed that *T. sutchuenensis* was more related to an ancestor similar to *T. ehrenswaerdii* (Heer) Schweitzer, while the other four species clustered together. Based on nuclear DNA ITS sequences, Li and Xiang (2005) proposed an EA origin of *Thuja* and reported two major clades. One clade contained (*T. occidentalis* (*T. standishii*, *T. sutchuenensis*)) while the other clade consisted of the two remaining species, *T. plicata* and *T. koraiensis*. Using both plastid and nuclear markers, Peng and Wang (2008) found considerable discordance among the plastid and nuclear gene trees, indicating the possibility of reticulate evolution in *Thuja*. Adelalu et al. (2020) inferred the interspecific phylogeny of *Thuja* using complete plastid genomes and obtained a different result with a (*T. standishii*, *T. koraiensis*) clade which was sister to a (*T. plicata*, (*T. occidentalis*, *T. sutchuenensis*)) clade. Overall, previous studies suggested that *Thuja* had a complex evolutionary history, and resolving the phylogenetic uncertainty among *Thuja* species and inferring their biogeographical history are challenging using limited data.

Here, we use more than 2,369 single copy nuclear loci and nearly full plastomes to reconstruct the evolutionary history of the intercontinental disjunct genus *Thuja*. Our goals are to (i) resolve the interspecific relationships within *Thuja*; (ii) reveal the contribution of hybridization and ILS to its complex evolutionary history; and (iii) understand the origin and evolution of the intercontinental disjunct pattern within *Thuja*.

Materials and Methods

Taxon Sampling and Target Enrichment Sequencing

We used targeted enrichment methods to capture, sequence the nuclear exome and the

nearly complete plastid genome to perform phylogenetic inferences for the genus *Thuja* (see Supplementary Materials for details). Thirteen individuals covering all currently recognized species in *Thuja*, plus one *Thujopsis dolabrata* individual as an outgroup, were sampled for target enrichment sequencing (Table S1). Total genomic DNA was extracted from silica-dried leaf tissue or herbarium material using the CTAB method (Doyle and Doyle, 1987), hybridized following the NimbleGen SeqCap EZ Library LR User's guide (Roche NimbleGen, Madison, Wisconsin), and sequenced on an Illumina HiSeq X Ten platform producing 150 bp paired end reads. Raw reads were filtered using the software Trimmomatic v 0.36 (Bolger et al., 2014) with the parameters set as "ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36".

Single-copy Orthologues Identification

Transcriptome assemblies of all five *Thuja* species plus one *Thujopsis dolabrata* accession were used to obtain single copy genes (SCGs, see Supplementary Materials). Contigs were assembled with default parameters using Trinity v 2.8.4 (Grabherr et al., 2011). Only the longest transcript was retained for each gene, and redundant contigs were further removed by CD-HIT. We used TransDecoder v 5.5.0 (Haas et al., 2013) to predict protein coding sequences. Peptide sequences of these six species were used in OrthoFinder v 2.3.11 (Emms and Kelly, 2015; Emms and Kelly, 2019) to perform the orthogroup search. Only single-copy orthologues with a minimum of 300 bp present in all individuals were selected for subsequent analyses. This resulted in 5,786 single-copy nuclear genes in total.

Assembly of Captured Sequence, Alignment and Filtering

We used HybPiper v 1.3.1 (Johnson et al., 2016) to assemble SCGs from capture sequenced quality-filtered reads. The sequences of the above identified 5,786 SCGs from each species

were used as target input file for HybPiper. The software MAFFT v 7.429 (Katoh and Standley, 2013) was used to align amino acid (AA) sequences, and the corresponding codon alignments were converted from the AA alignments using PAL2NAL v 14.0 (Suyama et al., 2006). Aligned loci with more than 20% missing data as well as individual DNA sequences with less than 300 bp or more than 50% gaps were removed. Only the filtered alignments which contained all individuals were retained.

Because recombination within loci might bias the inference of the species tree using coalescent methods (Morales-Briones et al., 2020), we further removed alignments showing a signal of recombination in the analyses using the coalescent model (i.e., ASTRAL, MP-EST, and BPP; see below). We used PhiPack v 1.1 (Bruen et al., 2006) to calculate the pairwise homoplasy index Φ for recombination, and a *P*-value of less than 0.05 was treated as significant.

Phylogeny Reconstruction of Nuclear Genes

We used a one-individual per species as well as a multi-individual per species data set to perform our analyses. The one individual dataset only included SCGs assembled from RNA-seq (five *Thuja* species plus one *Thujopsis* species ; six individuals/samples in total), while the multi-individual dataset included the sequences from the one individual dataset plus the SCGs from the HybPiper assembly, which sampled three to four accessions per species in *Thuja* and two individuals in *Thujopsis* (20 samples in total). The two data sets were used in different analyses, and if not stated otherwise, the one-individual dataset was adopted for most analyses.

Both maximum parsimony (MP) and maximum likelihood (ML) methods were used to reconstruct the phylogeny of *Thuja* based on concatenation approach. For the former, the SCGs were concatenated into a supermatrix, and the PAUP* v4.0a (Wilgenbusch and Swofford, 2003) was used to reconstruct a MP tree. Node supports were assessed by 1000

bootstrap replications. For the ML method, the best partitioning scheme for each codon position in each gene was established with the software ModelFinder (Kalyaanamoorthy et al., 2017), which was then used to reconstruct a ML tree in IQ-TREE v 2.0.4 (Nguyen et al., 2015). We used the ultrafast bootstrap approximation method (Hoang et al., 2018) to assess branch supports by resampling partitions and then sites within resampled partitions with 1,000 replicates (-p -B 1000 --sampling GENESITE; Gadagkar et al., 2005).

We further used the coalescent-based approach to estimate the species tree. Gene trees for each SCGs were generated via IQ-TREE with 1,000 ultrafast bootstraps and ModelFinder (Kalyaanamoorthy et al., 2017) implemented in IQ-TREE was used to select the best-fitting substitution model (-B 1000 -m MFP). We used ASTRAL v 5.6.3 (Zhang et al., 2018) to infer the species trees for both data before and after removal of recombinant loci from multi-individual dataset (Rabiee et al., 2019), measuring branch supports as local posterior probabilities (LPP; Sayyari and Mirarab, 2016).

Phylogeny Reconstruction of Plastid Genomes

We used the GetOrganelle v 1.7.4.1 (Jin et al., 2020) to get contigs for plastid genomes using six plastomes of *Thuja* and *Thujopsis* from NCBI GenBank (Qu et al., 2017; Adelalu et al., 2019; Adelalu et al., 2020) (Table S1) as the seed. Bowtie2 v 2.4.2 (Langmead and Salzberg, 2012) was used to map quality-filtered reads to the seed and recruit plastid-associated reads, and *de novo* assemblies were performed in SPAdes v 3.13.0 (Bankevich et al., 2012). Because tens of contigs were assembled for some individuals (Table S5), we reordered the assembled contigs in BWA-MEM algorithm v 0.7.17 (Li and Durbin, 2009) and extracted the consensus sequences in Geneious v 11.0.3 (Kearse et al., 2012) using the complete plastome of *T. plicata* (GenBank: KY290451; Adelalu et al., 2019) as the reference. The reordered contigs were aligned in MAFFT v 7.429 (Katoh and Standley, 2013). The graph-based clustering method performed in the software Divvier v 1.01 (Ali

et al., 2019) was used to address uncertainties and errors in the multiple sequence alignments. Comparing to other programs, Divvier can keep more informative sites and have a maximum number of true positive (Ali et al., 2019). The MP method was conducted in PAUP* v4.0a (Wilgenbusch and Swofford, 2003) with 1000 bootstrap replications, and the ML method was performed in IQ-TREE with 1000 ultrafast bootstraps. Two individuals ("*T. occidentalis* 5" and "*T. standishii* 4"; Table S1) downloaded from NCBI did not cluster with other individuals of the same species, which might due to hybridization, misidentification or other reasons. We removed the two samples for downstream analyses. Therefore, the final alignment of plastomes has 18 individuals from six species from the genera *Thuja* and *Thujopsis*, and each species contains 2–4 individuals (sequences). To measure concordance among individual sites for plastome data, we calculated the site concordance factors (sCF; Minh et al., 2020) implemented in IQ-TREE with 100 random quartets around each internal branch (--scf 100).

Species Network Analysis and Test of Hybridization

We used PhyloNet v 3.8.2 (Wen et al., 2018) to reconstruct phylogenetic networks from gene trees under a maximum pseudo-likelihood based on the multi-individual dataset. PhyloNet is used to infer species phylogenies while accounting not only for ILS but also for processes such as hybridization, taking the possibility of missing taxa due to extinction or incomplete sampling into account. This is important for groups like the Tertiary relicts which have a substantial probability of extinction events. Due to computational restrictions, the maximum number of allowed reticulation events was set to 1, 2, and 3, with 100 independent runs for each performed search to reach the global optimum of the likelihood (Cao et al., 2019). The optimum phylogenetic networks were visualized in Dendroscope (Huson and Scornavacca, 2012).

We then used the "CalcPopD" function in the R package "evobiR" v 1.3 (Blackmon and

225 Adams, 2015) to calculate Patterson's *D* (Green et al., 2010; Durand et al., 2011) and the
226 associated *Z*-scores for all possible four-taxon combinations in the same order as in the
227 ASTRAL species tree using the multi-individual dataset. The jackknife method was used
228 to calculate the statistical significance of Patterson's *D* for each combination with 100
229 replicates and a block size of 10,000 bp.
230
231 ***Concordance, ILS Simulations and Detecting the Anomaly Zone***
232 To test for concordance among gene trees and species trees, we first calculated the
233 percentage of quartets of the internal branches using ASTRAL (Mirarab et al., 2014) with
234 the parameter "-t 2". Individual gene trees were then mapped to the species tree estimated
235 in ASTRAL to count the number of gene trees supporting/conflicting each clade, and
236 estimated the "Internode Certainty All (ICA)" scores for each internode, using the software
237 phyparts v 0.01 (Smith et al., 2015). The ICA scores reflect the degree of certainty for a
238 given internode by considering the frequency of the bipartition defined by the internode in
239 a given set of trees in conjunction with that of all conflicting bipartitions in the same
240 underlying tree set (Salichos et al., 2014). ICA values near to 1 represent a strong
241 concordance for a given internode, while ICA value close to 0 indicate nearly equal
242 supports of one or more conflicting bipartitions. Negative ICA values indicate that the
243 conflicting bipartitions have higher frequencies. Finally, gene trees were converted to
244 ultrametric trees using the R package "ape" v 5.4 (Paradis et al., 2004), and visualized in
245 DensiTree v 2.2.5 (Bouckaert, 2010).
246 An anomaly zone is defined as a pair of internal branches in species trees that will
247 generate gene trees that are discordant with the species tree more often than gene trees that
248 are concordant (Degnan and Rosenberg, 2006). The anomaly zone is usually caused by
249 rapid speciation events in combination with large effective population sizes (Linkem et al.,
250 2016; Kapli et al., 2020). We calculated equation 4 from Degnan and Rosenberg (2006)

using the script provided by Linkem et al. (2016), to examine the anomaly zone in the ASTRAL species tree.

To evaluate ILS within *Thuja*, we used both DNA and protein sequences from the one-individual dataset. We conducted coalescent simulations to examine if ILS alone can explain the gene tree discordance and cytonuclear incongruence, using the pipelines of Mirarab et al. (2014) and Folk et al. (2017). We used MP-EST v 2.0 (Liu et al., 2010) to estimate species trees with branch lengths in coalescent units using both all and non-recombination loci. We first simulated gene trees in Dendropy v 4.4.0 (Sukumaran and Holder, 2010) using the "contained_coalescent_tree" function with the MP-EST trees as guide trees. A total of 100 simulations were performed, and each simulation produced the same number of estimated gene trees as did the observed gene tree in the one-individual dataset. We then calculated Robinson-Foulds (RF) distances between the species trees and each simulated or observed gene trees using the Python package ETE3 v 3.1.2 (Huerta-Cepas et al., 2016).

To infer if ILS is a source of cytonuclear discordance, we then simulated gene trees under the coalescence model of an organelle genome. We scaled branch lengths of the MP-EST trees by a factor two to account for organellar inheritance in monoecious plants (Rogalski et al., 2015) and generated 20,000 organellar gene trees under the coalescent model with Dendropy. If ILS is the main source of cytonuclear discordance, we can expect to find a high frequency of plastid-like topologies in the simulated data.

Molecular Dating and Multispecies Coalescent Analysis

As there are only six species in the *Thuja-Thujopsis* clade, few fossils can be used to calibrate node ages, which could result in a biased estimation of molecular dates (Linder et al., 2005; Wang and Mao, 2016). Therefore, we extended our sampling scheme to be able to include more calibration fossils. The extended sampling covered 16 Cupressaceae

277 species (Table S1). The single-copy genes were identified using the same pipeline as
278 described above, and only the 1st and 2nd codon positions for nuclear genes were used in
279 this analysis. We selected the three fossil calibration points used in Mao et al. (2012), plus
280 one newly discovered fossil record of *Chamaecyparis* (Xu et al., 2018), and three
281 secondary calibration points (Table S2). We conducted dating analyses using the program
282 MCMCTree in PAML v 4.9i (Yang, 2007). The package BASEML was used to estimate
283 the overall substitution rate under the GTR model (model=7). The divergence time between
284 *Sequoiadendron giganteum* and *Thuja* was assumed as ~183 Ma (Mao et al., 2012), which
285 resulted in a substitution rate per time unit (100 Ma) of 0.0225. Therefore, the parameter
286 "rgene_gamma" was set as "G(1, 44.38)", and the parameter "sigma2_gamma" was set as
287 "G(1, 10, 1)". We applied a burn in of 20,000,000, and sampled 50,000,000 generations
288 with a sample frequency of every 2,000 generations. The effective sample size (ESS) for
289 each parameter was verified by ESS >200 using Tracer v 1.7.1 to make sure that the MCMC
290 have reached convergence (Rambaut et al., 2018).
291 We further employed the Bayesian program BPP v 4.2.9 (Flouri et al., 2018) to estimate
292 coalescent processes within *Thuja* using the multi-individual dataset as an additional
293 analysis of divergence times. Using a large data set of 2,369 loci for 20 individuals would
294 increase the computational cost in BPP. Therefore, we only used the no-recombination loci
295 for coalescent inference, which is more than sufficient to get a reliable inference.
296 Firstly, we used the multispecies coalescent (MSC) model to estimate relative node ages
297 (τ) and nucleotide diversity (θ) based on a fixed species phylogeny inferred by ASTRAL.
298 Secondly, we inferred cross-species gene flow in *Thuja* under the multispecies-coalescent-
299 with-introgression (MSci) model (Flouri et al., 2020) performed in BPP based on the
300 network inferred in PhyloNet when the reticulation event was set to 1. Using the MSci
301 model, we can calculate the number, timings, and intensities of introgression events, as
302 well as the current and ancestral genetic diversity. The full-likelihoods were calculated for

303 both the MSC and MSci models, and the likelihood ratio test (LRT) was used to compare
304 them. For both models, the divergence time between *Thuja* and *Thujopsis* was calibrated
305 based on the result of MCMCTree.
306 We also used MSC model performed in BPP to infer species divergence times and
307 population size parameters for plastid DNA. As evidences of plastome recombination had
308 been reported for some conifers (Marshall et al., 2001; Sullivan et al., 2017), different
309 plastome fragments of *Thuja* could have experienced different evolutionary histories.
310 Therefore, we first used the full sequences from the plastome alignment, by treating the
311 plastome as a single locus, to measure the divergence times and the genetic diversities.
312 Then, we divided the full plastome alignment into 68 plastome fragments with 2000 bp in
313 length and 75 plastid coding genes to infer plastome coalescent processes respectively,
314 assuming that each 2-kp plastome segment or coding gene experienced independent
315 evolutionary history. The parameters were set the same as the above.
316
317 ***Ancestral Area Reconstruction***
318 We used the BioGeoBEARS v 1.1.1 packages (Matzke, 2013) as implemented in RASP v
319 4.2 (Yu et al., 2020) to estimate the ancestral ranges and biogeographical history of *Thuja*.
320 We assigned three geographic areas to the tips of the tree according to distributions of
321 extant and fossil species: A, Asia; B, Western North America; C, Eastern North America
322 plus Greenland (Table S3). We tested the three models (DIVALIKE, DEC, and
323 BAYAREALIKE; Matzke, 2014) implemented in BioGeoBEARS, and the corrected
324 Akaike Information Criterion (AICc) was used to select the best model. Because the
325 founder event speciation (+*J* parameter) has been controversial (Ree and Sanmartín, 2018;
326 Matzke, 2021), we made model comparisons by using the +*J* parameter or without it.
327 According to the geological evidence (Tiffney and Manchester, 2001), the dispersal
328 probability matrix (Table S3) was coded for four time periods, 0–4.7, 4.7–45, 45–60, 60–

65 Ma, following Zhou et al. (2020). To better represent the ancestral biogeographical ranges, four fossil species were further incorporated in our biogeographical analysis (see Supplementary Materials). We used the R function "Fossil.graft" (https://github.com/evolucionario/fossilgraft; Claramunt and Cracraft, 2015) to add the fossil species to the time-calibrated tree, as terminal tips. The relationships between fossil and extant species were based on the phylogenetic relationships and morphological similarities reconstructed in previous studies (McIver and Basinger, 1989; LePage, 2003; Cui et al., 2015).

Results

Gene assembly and filtering

We first identified 5,786 single-copy genes (SCGs) using transcriptome data from five *Thuja* species and one *Thujopsis* accession (outgroup), where each species was represented by one individual (one-individual dataset). We then generated a multi-individual dataset (20 accessions, including 3 to 4 samples per *Thuja* species) by adding data from target enrichment sequencing. The number of quality-filtered reads per sample ranged from 13.05 million to 53.45 million with an average of 33.58 million, and more than 4,000 genes were assembled into contigs with sequences >25% of the target length (Table S4). After filtering, the one-individual dataset consisted of 5,663 single-copy genes, while 2,969 of them were retained after removal of recombinant loci. The multi-individual dataset consisted of 2,369 loci, and 1,145 of them were non-recombination loci.

Phylogenetic Inference

Based on the all 2,369 loci from the multi-individual dataset, we used both MP and ML approaches to reconstruct species trees using PAUP and IQ-TREE based on a concatenated supermatrix. The two approaches supported the same interspecific relationships within

Thuja (Figures 1a, S1–S3). We further reconstruct an ASTRAL species tree based on the coalescent-based approach using both all loci and only non-recombination loci. In all analyses, *T. sutchuenensis* and *T. standishii* formed a well-supported clade [MP bootstrap percentage (BP)=79, ML BP=100, local posterior probabilities (LPP)=1; Figures 1a, S1–S4], here termed the "EA clade". This clade was sister to a "disjunct clade", with a strongly supported *T. koraiensis*-*T. occidentalis* (EA-ENA; BP=100, LPP=1)) relationship which in turn was sister to *T. plicata* (WNA; BP=100, LPP=1; Figures 1a, S1–S4).

The conflict analyses showed a high level of gene tree discordance within *Thuja*. The gene tree quartet supports for the alternatives with the species-level branches are comparable to those in the main topologies (Figure 1a). The ICA scores also showed high discordance among individual gene trees. Of the 2,369 gene trees, only 245 supported the sister relationship between *T. occidentalis* and *T. koraiensis* (ICA = 0.097), 175 supported *T. plicata* as sister to *T. occidentalis*-*T. koraiensis* (ICA = -0.099), and 279 supported *T. standishii* and *T. sutchuenensis* clustering together (ICA = 0.104; Figure 1a).

Plastid Phylogeny of Thuja

The number of contigs assembled in GetOrganelle ranged from 1 to 17, and the assembled sizes range from 110,224 bp in "*T. plicata* 2" to 130,843 bp in "*T. occidentalis* 2" (Table S5). The final alignment contains 18 sequences (representing 6 species) with 131,017 columns, containing 3,689 parsimony-informative sites, 208 singleton sites, and 127,120 constant sites. The plastid phylogenies using both ML and MP methods differed from the nuclear analysis at one node: the WNA species *T. plicata* had a well-supported (ML BP=89, MP BP = 99; Figures 1b, S5–S6) sister relationship to *T. sutchuenensis*-*T. standishii* in the plastid tree, while a strongly supported (MP BP=100, ML BP=100, LPP=1) sister relationship to *T. occidentalis*-*T. koraiensis* was suggested in both the ASTRAL and the concatenated nuclear species trees. Most nodes represented high levels of concordances

between individual site and the plastid tree (gCF>60; Figure 1b). The plastomes of *T. plicata* showed chimeric DNA polymorphisms and a low site concordance factor. Only 45.6% of decisive alignment sites supporting the branch containing *T. plicata* and *T. sutchuenensis*-*T. standishii* (sCF=45.6; Figure 1b), and 36.37% supporting *T. plicata* sister to *T. occidentalis*-*T. koraiensis* clade (Figure 1b).

Network Analysis and Gene Flow

Up to three hybridization events among the clades of *Thuja* were examined in PhyloNet. One reticulation event, in which gene flow from a "ghost" ancestral *Thuja* lineage to the ancestors of *T. sutchuenensis*, was detected in all three examinations (Figures 2a–c and S7a–c), with *T. sutchuenensis* having an inheritance probability of ~4.9% from the that "ghost" lineage. None of the three possible networks supported introgression events between *T. plicata* and either *T. sutchuenensis* or *T. standishii* (Figures 2 and S7). In addition, the *D* statistics analysis, which tests for signals of gene flow, detected significant gene flow between *T. plicata* and *T. standishii*, but not between *T. plicata* and *T. sutchuenensis*. Taken overall, these results suggested that hybridization is unlikely to be the cause of the cytonuclear discordance. However, the *D*-statistics-based analyses also provided evidence of frequent gene flow in the genus *Thuja* (Figures 2d and S7d), which suggests that hybridization have contributed to a part of the phylogenetic discordance among nuclear gene trees.

Simulations of ILS and Tests of the Anomaly Zone

We first inferred an MP-EST tree, which recovered identical topologies to the ASTRAL tree, based on all 5,663 loci from one-individual data set, and used it as a guide to simulate gene trees under ILS. A total of 100 simulations were performed, and each simulation generated the same number of gene trees as in the real data (5,663 gene trees). The

distributions of the Robinson-Foulds (RF) distances of the simulated and observed gene trees compared to the species tree from the one-individual dataset largely overlapped (Figures 3c, S6 and S7), suggesting that ILS can account for most of the gene tree discordance (Wang et al., 2018). We also used the dataset after removal of recombination loci and rerun the ILS simulation, which conducted a similar result as the all loci did (Figures S8–S10; Table S7). A pair of internodes on the ASTRAL species tree was in the anomaly zone (orange and blue nodes in Figures 3e and 8e), indicating that these nodes might have experienced rapid speciation events.

Of the 20,000 simulated plastid gene trees, the two most common topologies were consistent with the observed nuclear species tree (1468 trees, 7.34%), followed by the same topology as the observed plastid tree (928 trees, 4.64%; Tables S6 and S7). In total, 1,965 trees contained a clade comprising *T. plicata*+*T. sutchuenensis*-*T. standishii*, with various tree topologies (Figures 3f and 8f). Reticulate evolution due to hybridization should not produce variation in the plastid tree topology, indicating that the inconsistency among organellar and species trees is likely to be due to the ILS.

Divergence Dating and Multispecies Coalescent Analysis

According to the MCMCTree, the stem age of *Thuja* (divergence from *Thujopsis*) was estimated to be 62.68 million years ago [Ma; early Paleogene, 95% highest posterior density (HPD): 58.61–73.77 Ma; Figure 4], and the crown age was 23.96 Ma (95% HPD: 19.39–29.43Ma), which corresponds to the Paleogene-Neogene boundary. Furthermore, *T. sutchuenensis* diverged from *T. standishii* about 20.05 Ma (95% HPD: 15.6–25.35 Ma), and the crown age of the clade containing *T. plicata* and *T. occidentalis*-*T. koraiensis* was estimated to be 22.09 Ma (95% HPD: 17.61–27.44 Ma). The EA-ENA disjunct pair *T. occidentalis* and *T. koraiensis* was estimated to be 19.55 Ma (95% HPD: 15.19–24.79 Ma; Figure 4).

The result of the likelihood ratio test strongly favored the MSci model over the MSC model [$2(\ln L_1 - \ln L_0) = 152.24$; P-value<0.001; Figure 5], which supported a "ghost introgression" event. As both models yielded similar parameter estimates (Figure 5), we used the MSci estimates because of the higher likelihood of this model. The BPP analysis gave a $\tau = 0.004633$ (Tables S8 and S9) for the stem age of *Thuja*, corresponding to 62.68 Ma from the MCMCTree. The "ghost" ancestral *Thuja* lineage diverged from the common ancestor of all extant *Thuja* species about 54.82 Ma (95% HPD: 51.63–57.60 Ma; τ=0.004246], and the estimate of the introgression event was dated to about 19.50 Ma (95% HPD: 18.70–20.28 Ma; τ=0.001519). BPP and MCMCTree yielded very similar results in terms of the crown age for *Thuja* and the divergence time between *T. sutchuenensis* and *T. standishii* (Figures 4 and 5). The main difference between the two analyses is the age estimates of the disjunct EA-ENA clade: the crown age inferred by BPP (~15 Ma; Figure 5) was younger than the one inferred by MCMCTree (~22.09 Ma; Figure 4). Similarly, the divergence time of the EA-ENA disjunct *T. occidentalis*-*T. koraiensis* clade was estimated by BPP to have occurred about 14.82 Ma (95% HPD: 14.12–15.53 Ma; Figure 5b), which contrasts with the MCMCtree divergence age estimate of ~19.55 Ma (Figure 4).

The population size parameter (θ) of the extant species ranged from θ=0.00192 (95% HPD: 0.001812–0.002021; *T. plicata*) to 0.00492 (95% HPD: 0.004596–0.005265; *T. standishii*), with much higher estimates for the respective ancestral lineages, which was supported by the coalescent processes inferred by the plastomes (Figure S11). Specifically, θ of the ancestral population of the disjunct clade (*T. plicata* sister to *T. sutchuenensis*-*T. standishii*; θ=0.0313, 95% HPD: 0.024183–0.038557) was about 10 times higher than the current population size estimate (θ=0.00192–0.00295). The introgression probability was estimated to be 0.2 (95% HPD: 0.16–0.24; Figure 5), suggesting that ~20% of the nuclear genome of *T. sutchuenensis* is derived from a "ghost" basal *Thuja* lineage.

Ancestral Area Reconstruction

Without fossil taxa, model tests performed in BioGeoBEARS suggested that the DEC model was better than all other models (Tables S10 and S11). When including the fossil taxa, the DIVALIKE+*J* model was the best one among all six models, and the DEC model performed better than either the DIVALIKE or BAYAREALIKE model (Tables S12 and S13). From the DEC models, the distribution ranges of the most recent common ancestor (MRCA) of all living species of *Thuja* and the disjunct clade most likely occurred in East Asia + North America (ABC; Figures 6b and d). The results from the biogeographical analysis including the fossil species showed that the ancestral range of *Thuja* (comprising all fossil and living species) is likely to be the eastern North America (C; Figure 6c and d), suggesting a North American origin of the genus *Thuja*. Using the DIVALIKE+*J* model, the ancestor of extant *Thuja* species most likely originated in East Asia, and then dispersed to East Asia + western North America, with subsequent diversification due to vicariance (A->AB->A|B; Figure 6c).

Discussion

Thuja Phylogenomics and Discordance of Gene Trees

We used more than 2,000 loci to estimate a species-level phylogeny of the Tertiary relict genus *Thuja*. Our analyses strongly supported the sister relationship of the EA-ENA disjunct species pair *T. occidentalis*-*T. koraiensis*, with a WNA *T. plicata* as sister to this clade (disjunct clade; Figure 1a). The remaining two EA species *T. standishii* and *T. sutchenensis* were clustered in a separate clade (EA clade) which was sister to the disjunct clade. The EA clade was also recovered by previous phylogenies based on nrDNA ITS (Li and Xiang, 2005) and two different low-copy nuclear genes (Peng and Wang, 2008), however, with different disjunct clade topologies, where either a sister relationship between *T. plicata* and *T. koraiensis* (ITS and 4*CL*) or *T. plicata* and *T. occidentalis* (*LEAFY*) were supported. The main uncertainty in previous studies was the phylogenetic position of *T.*

485 *occidentalis*. The nrDNA ITS tree supported a sister relationship between *T. occidentalis*
486 and *T. standishii-T. sutchenensis* (Li and Xiang, 2005), and in the *LEAFY* gene tree (Peng
487 and Wang, 2008), *T. occidentalis* is a sister species to *T. plicata*, while the basal position of
488 *T. occidentalis* was supported in the 4*CL* gene tree (Peng and Wang, 2008). None of the
489 previous phylogenies supported the sister relationship of *T. occidentalis*-*T. koraiensis*,
490 which indicates that using only a few loci cannot resolve the phylogenetic relationship
491 within *Thuja*.

492 Our phylogenomic analyses showed that there was a very high level of discordance
493 among individual gene trees and the species trees. Gene tree heterogeneity is commonly
494 explained by deep coalescent processes such as incomplete lineage sorting (ILS) or
495 hybridization (Olave et al., 2018). We first examined hybridization as a possible cause of
496 discordance using a pseudo-likelihood approach performed in PhyloNet and the *D*-
497 statistics test. A strong signal of interspecific gene flow between most *Thuja* lineages was
498 detected (Figure 2), indicating that hybridization could have caused the gene tree and
499 species tree discordance. However, both the PhyloNet and *D*-Statistic test detected little
500 introgression in *T. occidentalis*, indicating that the uncertain placement of this species in
501 previous studies (Li and Xiang, 2005; Peng and Wang, 2008) is unlikely to be explained
502 by hybridization. The alternative explanation, ILS, is likely to apply to species that
503 diverged during rapid speciation events and/or had large population sizes (Flouri et al.,
504 2018). Our simulation analysis showed that the distribution of tree-to-tree distances of
505 simulated and observed gene trees to the species tree largely overlapped, indicating that
506 ILS alone could explain most of the gene tree discordance (Figures 3b–c, S8–S10). Testing
507 for anomalous zones in the species tree highlighted two internodes which generated gene
508 trees that are discordant with the species tree more often than gene trees that are concordant.
509 Three species, *T. plicata*, *T. occidentalis* and *T. koraiensis*, were involved in this anomaly
510 zone. Moreover, the multispecies coalescent analysis hinted at very high levels of DNA

polymorphism in the most recent common ancestor (MRCA) of these three species (Figure 5), which might have contributed to incomplete lineage sorting. Therefore, the most likely explanation for the inconsistent placement of *T. occidentalis* inferred from different loci in previous studies is ILS, which might have also been facilitated by large ancestral population sizes.

Cytonuclear Discordance as further Evidence for ILS

The reason for the inconsistent position of *T. plicata* in phylogenies based on nuclear and plastid data has long been debated. Peng and Wang (2008) found a high level of site discordance for the phylogenetic position of *T. plicata*. A total of 15 parsimony-informative sites were obtained from 5,099 bp plastid DNA alignment, and eight of them were shared between *T. plicata* and *T. sutchenensis*-*T. standishii* clade, while six sites shared between *T. plicata* and the clade containing *T. koraiensis*-*T. occidentalis* (Peng and Wang, 2008). Our cpDNA phylogeny based on plastome alignment resolved *T. plicata* as sister to the *T. sutchenensis*-*T. standishii* pair with a high level of individual site conflict (sCF=45.6; Figure 1b), and *T. plicata* has a chimeric plastome, confirming earlier results based on five cpDNA regions by Peng and Wang (2008).

In contrast, our nuclear analysis (ASTRAL tree) suggested a position of *T. plicata* as sister to the ENA-EA disjunct *T. occidentalis*-*T. koraiensis* group. Cytonuclear discordance could result from either ILS or hybridization (especially organellar introgression). However, only a few studies have provided evidence for ILS (Wang et al., 2018; Stull et al., 2020), suggesting that hybridization is the more common cause of cytonuclear discordance (Folk et al., 2017; Lee-Yaw et al., 2019; Li et al., 2020; Wang et al., 2021). The organelle genome is uniparentally inherited, therefore its effective population size is one-quarter in dioecious species and one half in monoecious species (like *Thuja*) of the nuclear autosomes (Rogalski et al., 2015). Haplotypes of plastid genes are therefore expected to have a higher rate of

genetic drift and a lower level of ILS compared to nuclear genes (Hamilton, 2009; Sloan et al., 2017).

Here, we tried to distinguish between hybridization and ILS, using a coalescent simulation under the model of an organellar gene tree. The simulations produced a large proportion of simulated organellar gene trees which were consistent with the observed plastid tree (4.64%, Tables S5 and S6). As relicts from the Tertiary, the ancestors of all living *Thuja* species were estimated to have large population sizes of either nuclear DNA (Figure 5) or plastid DNA (Figure S11), indicating that a phylogeny based on plastid genes might be greatly affected by the incomplete sorting of ancient polymorphism. Although a signature of hybridization was detected between *T. plicata* and *T. sutchenensis*, ILS alone can explain the observed cytonuclear discordance, suggesting that the effect of ILS on the organellar phylogeny is greater than previously thought.

"Ghost Introgression" into T. sutchuenensis

The phylogenetic position of *T. sutchuenensis* has puzzled taxonomists for a long time. The southwestern China endemic *T. sutchuenensis* had been listed as being extinct in the wild until it was rediscovered in 1999 (Xiang et al., 2002). In a phylogeny of fossil and extant species based on seed cone morphology, *T. sutchuenensis* was grouped in a clade with *T. ehrenswaerdii*, a fossil species known from the Paleocene sediments of Greenland (Schweitzer, 1974); this clade was sister to all other *Thuja* (McIver and Basinger, 1989). However, molecular studies did not suggest an ancestral position of *T. sutchuenensis* (Li and Xiang, 2005; Peng and Wang, 2008; Adelalu et al., 2020). We reconstructed the reticular evolutionary history of *Thuja* in PhyloNet, allowing for the existence of missing taxa due to incomplete sampling and/or extinction. Our results suggested gene flow from an ancestral *Thuja* "ghost lineage" into *T. sutchuenensis* (Figures 2a–c), indicating that the ancestor-like characters of *T. sutchuenensis* are most likely derived from an extinct

563 ancestral lineage via introgression ("ghost lineage"). This is supported by the results of the
564 BPP analysis which showed that ~20% (95% HPD: 16%–24%) of the nuclear genome of
565 *T. sutchuenensis* (Figure 5b) was derived from an ancient lineage of *Thuja* that is now
566 extinct. The analysis indicated that the "ghost lineage" originated in the late Paleocene
567 approximately 57.44 Ma (95% HPD: 54.26–60.47 Ma; Figure 5), and then hybridized with
568 the ancestor populations of *T. sutchuenensis* in the early Miocene (19.63–21.42 Ma; Figure
569 5), when the global climate was still warm and humid. The effective population size of both
570 the "ghost lineage" (θ_{sl}=0.001062; 95% HPD: 0.0053–0.01745; Figure 5b) and the
571 ancestral population of *T. sutchuenensis* (θ_{sr}=0.0185; Figure 5b) were relatively large.
572 Therefore, *T. sutchuenensis* was expected to have a wider distribution range in the past than
573 today, which might have increased the chances of contact and interbreeding. It is possible
574 that the genes coding for the unusual morphological traits of *T. sutchuenensis* were derived
575 from this "ghost lineage". This "ghost lineage" might be related to the fossil species *T.*
576 *ehrenswaerdii*, which was found in the Paleocene sediments of Greenland (Schweitzer,
577 1974).
578 Until recently, gene flow from extinct taxa could only be detected via extraction of DNA
579 from fossils, which has only been possible in a few groups such as hominids (Green et al.,
580 2010; Prüfer et al., 2014) and mammoths (van der Valk et al., 2021). In recent years, due
581 to the development of new molecular methods (Wang et al., 2018; Kuhlwilm et al., 2019),
582 a growing number of taxa such as *Phylloscopus* (Zhang et al., 2019), *Canis*
583 (Gopalakrishnan et al., 2018; Wang et al., 2020), *Pan* (Kuhlwilm et al., 2019), *Picea* (Ru
584 et al., 2018) and *Oxyria* (Luo et al., 2017), have been reported to show "ghost introgression"
585 using genomic data. It is therefore likely that "ghost introgression" is more common than
586 previously thought and may have played an important role in shaping the evolution of
587 extant species (Taylor and Larson, 2019; Zhang et al., 2019). Because Tertiary relict floras
588 are characterized by once extensive distributions subsequently contracted due to climate

589 change leading to local and regional extinctions (Milne and Abbott, 2002; Milne, 2006), it
590 is likely that some of the extant species coexisted with now-extinct lineages for a long time
591 during their evolutionary histories. These floras are hence strong candidates for "ghost
592 introgression" and the possibility should be tested in future biogeographic analysis of
593 Tertiary relict floras.
594
595 ***Biogeographic history of Thuja points to ILS***
596 The discovery of two unambiguous *Thuja* fossils with reproductive organs, i.e., *T. polaris*
597 and *T. ehrenswaerdii* (Schweitzer, 1974; McIver and Basinger, 1989), from the Paleocene
598 in the Canadian Arctic, suggests that the genus *Thuja* might have originated in higher
599 latitudes of North America. We included these two fossil species in our biogeographic
600 analysis which suggested that *Thuja* diverged from *Thujopsis* around 62.68 Ma and that
601 this genus originated in North America. An even earlier fossil from the late Cretaceous
602 discovered in Alaska (LePage, 2003) further supports an origin in northern North America.
603 After the Paleocene, climatic optima during the early and middle Eocene (Zachos et al.,
604 2008) supported the development of a circumboreal flora with warm temperate to tropical
605 elements (Azuma et al., 2001; Milne, 2006). The ancestral populations of *Thuja* probably
606 dispersed to large parts of East Asia and North America during this time period. The rich
607 fossil record of *Thuja* from the late Cretaceous to the Pleistocene in the Northern
608 Hemisphere further supports the hypothesis that this genus once had a wider distribution
609 (LePage, 2003; Taberlet and Luikart, 2008; Cui et al., 2015).
610 The crown age of extant *Thuja* species was estimated to be 23.96 Ma (95% HPD:
611 19.39–29.43 Ma; Figure 6) and its ancestral area was inferred to be widespread in East
612 Asia and North America (Figure 6). This suggests that *Thuja* has experienced further
613 diversification in the higher latitudes of the Northern Hemisphere during the late Paleogene
614 or early Neogene, when it might have formed two separate clades. One clade (EA clade)

occurred in eastern Asia where it diversified into two lineages (*T. standishii* and *T. sutchuenensis*) approximately 20.05 Ma (95% HPD: 15.6–25.35 Ma; Figure 5). An arid belt located in northern China constitutes an important barrier that impedes migration between northeastern and southeastern Asia (Milne and Abbott, 2002) and has been in existence from the Eocene on (Tiffney and Manchester, 2001; Guo et al., 2008). *Thuja standishii* and *T. sutchuenensis* occur on either side of that belt, and so it might have facilitated their divergence during the early Miocene

The other clade (disjunct clade) includes three species (*T. occidentalis*, *T. plicata* and *T. koraiensis*) that occur in ENA, WNA and EA, respectively, with a strongly supported EA-ENA disjunct sister species relationship of *T. occidentalis*-*T. koraiensis* (Figure 1a). Like the MRCA of *Thuja*, the ancestors of this clade had a widespread range across North America and Asia. While many EA-ENA disjunct pairs arose via extinction of widespread ancestors in WNA and Europe due to climatic cooling (Tiffney, 1985b; Tiffney, 1985a; Wen, 1999; Zhang et al., 2021), the *T. occidentalis*-*T. koraiensis* pair has an extant sister taxon in WNA (*T. plicata*). The EA-ENA disjunction may therefore reflect the sequence of diversification within this clade. Our results also indicated that the MRCA of this disjunct clade might have experienced a rapid radiation within a narrow time window of less than one million years that gave rise to the three species (Figure 5). Moreover, coalescent analysis indicates that the MRCA of the disjunct clade had a very large ancestral population size (Figure 5). Therefore, the close relationship of the EA-ENA disjunct pair might be the result of stochastic processes after a radiative speciation event in a relatively short time period (~0.8 Ma; Figure 5). Consistent with this, our coalescent analyses suggested that ILS is prominent in *Thuja* (Figure 3), especially within the disjunct clade, which also forms an anomaly zone (Figure 3e). The diversification of this clade occurred around (14.71–) 15.60–22.09 (–27.44) Ma depending on analysis method (Figures 4 and 5b), roughly corresponding to the Mid-Miocene Climatic Optimum (Zachos et al., 2008). This clade's

MRCA was likely widespread across North America and East Asia, and the Mid-Miocene Climatic Optimum might have facilitated the diversification due to new ecological niches on different continents, which in turn facilitated the rapid speciation of the disjunct clade. Alternatively, progressive cooling of the global climate from ~15 Ma onwards (Zachos et al., 2001; Milne and Abbott, 2002), might have forced speciation via formation of geographical/climatic barriers that separated EA, ENA and WNA (Azani et al., 2019). Both mechanisms might have contributed to this speciation event, separating one large continuous population into three smaller ones within a short timescale. This would allow for ILS, and therefore it is possible that the EA-ENA species pair, *T. occidentalis* and *T. koraiensis*, by chance became fixed for a similar set of genetic variation compared to the WNA species *T. plicata*, resulting in the EA-ENA disjunction.

Conclusion

Summarizing, we used more than 2,000 loci and integrated fossilized taxa in our analysis to reconstruct the evolutionary history of the small Tertiary relict genus *Thuja* which is well known for its EA-ENA disjunction. The most common ancestor of the genus diversified into five living species in a short time period of ca. 3.5 million years according to multispecies coalescent analysis, with the three members of the disjunct clade diversifying over a narrow time window of just ~0.8 million years. Multispecies coalescent and simulation studies revealed that ancient lineages of *Thuja* had large population sizes, which might have contributed, together with rapid divergence, to ILS, especially in the disjunct clade. This could be the underlying cause for much of the conflict among gene trees and the cytonuclear discordance which have puzzled systematists of this genus for a long time. However, “ghost introgression” from extinct species might also have contributed to the discordance among gene trees and could have left a signature in the morphology of *T. sutchuenensis*: we found that ~20% of the nuclear genome of *T. sutchuenensis* is derived

667 from a “ghost lineage” ancestral to *Thuja*, which might explain the close resemblance of
668 its cone morphology to that of an ancient fossil species. Overall, our study revealed a
669 complex evolutionary history of a small and disjunct Tertiary relict genus, involving ILS,
670 hybridization and extinction. It also demonstrates that phylogenies based on a few genes
671 might not be able to resolve the biogeographic history of disjunct taxa accurately. Genomic
672 data are therefore needed to reveal the complex history of intercontinental disjunct taxa.
673
674 **Acknowledgements**
675 We thank Prof. Yuanwen Duan for assistances with sample collection. We thank Royal
676 Botanic Garden Edinburgh for providing dried herbarium and living materials. This study
677 is financially supported by National Science Foundation of China (grant No. U20A2080,
678 31622015) and Fundamental Research Funds for the Central Universities (SCU2020D003,
679 SCU2021D006). The Royal Botanic Garden Edinburgh is supported by the Scottish
680 Government’s Rural and Environment Science and Analytical Services Division.
681
682 **Conflict of Interest**
683 The authors declared that they have no conflicts of interest to this work.
684
685 **Author contributions**
686 K.M. and J.L. conceived the research; M.R., J.L. and T.T. collected samples; J.L., Y.W.,
687 D.W., S.J. and Y.Z. collected and analyzed the data; J.L, M.R., R.M. and K.M. wrote the
688 manuscript; K.M., and M.R. revised the manuscript.
689
690 **Data Availability Statement**
691 Data available from the Dryad Digital Repository:
692 http://dx.doi.org/10.5061/dryad.44j0zpcd8.

Scripts available from https://github.com/lijl459/Phylogenomics_for_Thuja.

Reference

Adelalu, K. F., Qu, X.-J., Sun, Y.-X., Deng, T., Sun, H., Wang, H.-C. 2019. Characterization of the complete plastome of western red cedar, *Thuja plicata* (Cupressaceae). Conserv. Genet. Resour. 11, 79–81.

Adelalu, K. F., Zhang, X., Qu, X., Landis, J. B., Shen, J., Sun, Y., Meng, A., Sun, H., Wang, H. 2020. Plastome Phylogenomic and Biogeographical Study on *Thuja* (Cupressaceae). Biomed Res. Int. 2020, 8426287.

Ali, R. H., Bogusz, M., Whelan, S. 2019. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. Mol. Biol. Evol. 36, 2340–2351.

Azani, N., Bruneau, A., Wojciechowski, M. F., Zarre, S. 2019. Miocene climate change as a driving force for multiple origins of annual species in *Astragalus* (Fabaceae, Papilionoideae). Mol. Phylogen. Evol. 137, 210–221.

Azuma, H., García-Franco, J. G., Rico-Gray, V., Thien, L. B. 2001. Molecular phylogeny of the Magnoliaceae: the biogeography of tropical and temperate disjunctions. Am. J. Bot. 88, 2275–2285.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D.*, et al.* 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19, 455–477.

Blackmon, H., Adams, R. A. 2015. EvobiR: Tools for comparative analyses and teaching evolutionary biology. https://doi.org/10.5281/zenodo.30938

Bolger, A. M., Lohse, M., Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bouckaert, R. R. 2010. DensiTree: making sense of sets of phylogenetic trees. Bioinformatics 26, 1372–1373.

Bruen, T. C., Philippe, H., Bryant, D. 2006. A Simple and Robust Statistical Test for Detecting the Presence of Recombination. Genetics 172, 2665.

Cao, Z., Liu, X., Ogilvie, H. A., Yan, Z., Nakhleh, L. 2019. Practical Aspects of Phylogenetic Network Analysis Using PhyloNet. bioRxiv, 746362.

Chan, K. O., Hutter, C. R., Wood, P. L., Jr., Grismer, L. L., Brown, R. M. 2020. Target-capture phylogenomics provide insights on gene and species tree discordances in Old World treefrogs (Anura: Rhacophoridae). Proc. R. Soc. B-Biol. Sci. 287, 20202102.

Claramunt, S., Cracraft, J. 2015. A new time tree reveals Earth history's imprint on the

evolution of modern birds. Sci. Adv. 1, e1501005.
Cui, Y. M., Sun, B., Wang, H. F., Ferguson, D. K., 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, e0138544. https://doi.org/10.1371/journal.pone.0138544
Dalquen, D. A., Zhu, T., Yang, Z. 2017. Maximum Likelihood Implementation of an Isolation-with-Migration Model for Three Species. Syst. Biol. 66, 379–398.
Davidse, G. 1983. Biogeographical relationships between temperate eastern Asia and temperate eastern North America: the twenty-ninth annual systematics symposium. Ann. Mo. Bot. Gard. 70, 421–422.
Degnan, J. H., Rosenberg, N. A. 2006. Discordance of Species Trees with Their Most Likely Gene Trees. PLoS Genet. 2, e68. https://doi.org/10.1371/journal.pgen.0020068
Donoghue, M. J., Bell, C. D., Li, J. H. 2001. Phylogenetic patterns in Northern Hemisphere plant geography. Int. J. Plant Sci. 162, S41–S52.
Donoghue, M. J., Smith, S. A. 2004. Patterns in the assembly of temperate forests around the Northern Hemisphere. Philos. Trans. R. Soc. B-Biol. Sci. 359, 1633–1644.
Doyle, J. J., Doyle, J. L. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochemical Bulletin 19, 11–15.
Durand, E. Y., Patterson, N., Reich, D., Slatkin, M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28, 2239–2252.
Eckenwalder, J. E. 2009. Conifers of the World: The Complete Reference. Timber Press, Portland.
Emms, D. M., Kelly, S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16, 157.
Emms, D. M., Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20, 238.
Farjon, A. 2005. Monograph of Cupressaceae and Sciadopitys. Royal Botanic Gardens, Kew, London.
Feng, Y.-Y., Shen, T.-T., Shao, C.-C., Du, H., Ran, J.-H., Wang, X.-Q. 2020. Phylotranscriptomics reveals the complex evolutionary and biogeographic history of the genus *Tsuga* with an East Asian-North American disjunct distribution. Mol. Phylogen. Evol. 157, 107066.
Flouri, T., Jiao, X., Rannala, B., Yang, Z. 2018. Species Tree Inference with BPP Using Genomic Sequences and the Multispecies Coalescent. Mol. Biol. Evol. 35, 2585–2593.
Folk, R. A., Mandel, J. R., Freudenstein, J. V. 2017. Ancestral Gene Flow and Parallel Organellar Genome Capture Result in Extreme Phylogenomic Discord in a Lineage

of Angiosperms. Syst. Biol. 66, 320–337.
Fu, L., Yu, Y., Farjon, A. 1999. Cupressaceae. In: Wu, Z., Raven, P. (Eds.), Flora of China. Science Press, Beijing, pp. 62–77.
Gadagkar, S. R., Rosenberg, M. S., Kumar, S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. J. Exp. Zool. Part B 304B, 64–74.
Gopalakrishnan, S., Sinding, M.-H. S., Ramos-Madrigal, J., Niemann, J., Samaniego Castruita, J. A., Vieira, F. G., Carøe, C., Montero, M. d. M., Kuderna, L., Serres, A., *et al.* 2018. Interspecific Gene Flow Shaped the Evolution of the Genus Canis. Curr. Biol. 28, 3441–3449.e3445.
Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652.
Graham, A. 1966. Plantae rariores camschatcenses: A translation of the dissertation of Jonas P. Halenius, 1750. Brittonia 18, 131–139.
Gray, A. 1859. Diagnostic characters of new species of phanerogamous plants collected in Japan by Charles Wright, botanist of the U.S. North Pacific Exploring Expedition. Mem. Am. Acad. Arts 6, 377–452.
Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H., *et al.* 2010. A draft sequence of the Neandertal genome. Science 328, 710–722.
Guo, Z. T., Sun, B., Zhang, Z. S., Peng, S. Z., Xiao, G. Q., Ge, J. Y., Hao, Q. Z., Qiao, Y. S., Liang, M. Y., Liu, J. F., *et al.* 2008. A major reorganization of Asian climate by the early Miocene. Clim. Past. 4, 153–174.
Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., *et al.* 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8, 1494–1512.
Hamilton, M. 2009. Population genetics. John Wiley & Sons, Chichester.
Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., Vinh, L. S. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35, 518–522.
Huerta-Cepas, J., Serra, F., Bork, P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. 33, 1635–1638.
Huson, D. H., Scornavacca, C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst. Biol. 61, 1061–1067.
Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., dePamphilis, C. W., Yi, T.-S., Li, D.-Z. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21, 241.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., Zerega, N. J. C., Wickett, N. J. 2016. HybPiper: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. Appl. Plant Sci. 4, 1600016.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., Jermiin, L. S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14, 587–589.

Kapli, P., Yang, Z., Telford, M. J. 2020. Phylogenetic tree building in the genomic age. Nat. Rev. Genet. 21, 428–444.

Katoh, K., Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C.*, et al.* 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.

Kuhlwilm, M., Han, S., Sousa, V. C., Excoffier, L., Marques-Bonet, T. 2019. Ancient admixture from an extinct ape lineage into bonobos. Nature Ecology & Evolution 3, 957–965.

Langmead, B., Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Leache, A. D., Rannala, B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60, 126–137.

Lee-Yaw, J. A., Grassa, C. J., Joly, S., Andrew, R. L., Rieseberg, L. H. 2019. An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (Helianthus). New Phytol. 221, 515–526.

LePage, B. A. 2003. A new species of *Thuja* (Cupressaceae) from the Late Cretaceous of Alaska: implications of being evergreen in a polar environment. Am. J. Bot. 90, 167–174.

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

Li, J., Milne, R. I., Ru, D., Miao, 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. Mol. Ecol. 29, 1250–1266.

Li, J., Stukel, M., Bussies, P., Skinner, K., Lemmon, A. R., Lemmon, E. M., Brown, K., Bekmetjev, A., Swenson, N. G. 2019. Maple phylogeny and biogeography inferred from phylogenomic data. J. Syst. Evol. 57, 594–606.

Li, J. H., Xiang, Q. P. 2005. Phylogeny and biogeography of *Thuja* L. (Cupressaceae), an eastern Asian and North American disjunct genus. J. Integr. Plant Biol. 47, 651–659.

847 Linder, H. P., Hardy, C. R., Rutschmann, F. 2005. Taxon sampling effects in molecular
848 clock dating: An example from the African Restionaceae. Mol. Phylogen. Evol. 35,
849 569–582.
850 Linkem, C. W., Minin, V. N., Leaché, A. D. 2016. Detecting the Anomaly Zone in Species
851 Trees and Evidence for a Misleading Signal in Higher-Level Skink Phylogeny
852 (Squamata: Scincidae). Syst. Biol. 65, 465–477.
853 Liu, L., Yu, L., Edwards, S. V. 2010. A maximum pseudo-likelihood approach for
854 estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302–319.
855 Luo, X., Hu, Q., Zhou, P., Zhang, D., Wang, Q., Abbott, R. J., Liu, J. 2017. Chasing ghosts:
856 allopolyploid origin of *Oxyria sinensis* (Polygonaceae) from its only diploid
857 congener and an unknown ancestor. Mol. Ecol. 26, 3037–3049.
858 Maddison, W. P., Knowles, L. L. 2006. Inferring Phylogeny Despite Incomplete Lineage
859 Sorting. Syst. Biol. 55, 21–30.
860 Manchester, S. R. 1999. Biogeographical Relationships of North American Tertiary Floras.
861 Ann. Mo. Bot. Gard. 86, 472–522.
862 Mao, K., Milne, R. I., Zhang, L., Peng, Y., Liu, J., Thomas, P., Mill, R. R., S. Renner, S.
863 2012. Distribution of living Cupressaceae reflects the breakup of Pangea. Natl. Acad.
864 Sci. U. S. A. 109, 7793.
865 Marshall, H. D., Newton, C., Ritland, K. 2001. Sequence-Repeat Polymorphisms Exhibit
866 the Signature of Recombination in Lodgepole Pine Chloroplast DNA. Mol. Biol.
867 Evol. 18, 2136–2138.
868 Matzke, N. J. 2013. Probabilistic historical biogeography: new models for founder-event
869 speciation, imperfect detection, and fossils allow improved accuracy and model-
870 testing. Front. Biogeogr. 5, 242–248.
871 Matzke, N. J. 2014. Model Selection in Historical Biogeography Reveals that Founder-
872 Event Speciation Is a Crucial Process in Island Clades. Syst. Biol. 63, 951–970.
873 McIver, E., Basinger, J. 1989. The morphology and relationships of *Thuja polaris* sp.nov.
874 (Cupressaceae) from the early Tertiary, Ellesmere Island, Arctic Canada. Can. J. Plant
875 Sci. 67, 1903–1915.
876 Milne, R. I. 2006. Northern hemisphere plant disjunctions: A window on tertiary land
877 bridges and climate change? Ann. Bot. 98, 465–472.
878 Milne, R. I., Abbott, R. J. 2002. The origin and evolution of tertiary relict floras. Adv. Bot.
879 Res. Academic Press, pp. 281–314.
880 Minh, B. Q., Hahn, M. W., Lanfear, R. 2020. New Methods to Calculate Concordance
881 Factors for Phylogenomic Datasets. Mol. Biol. Evol. 37, 2727–2733.
882 Mirarab, S., Bayzid, M. S., Boussau, B., Warnow, T. 2014. Statistical binning enables an
883 accurate coalescent-based estimation of the avian tree. Science 346, 1250463.
884 Morales-Briones, D. F., Kadereit, G., Tefarikis, D. T., Moore, M. J., Smith, S. A.,
885 Brockington, S. F., Timoneda, A., Yim, W. C., Cushman, J. C., Yang, Y. 2020.

886 Disentangling Sources of Gene Tree Discordance in Phylogenomic Datasets: Testing
887 Ancient Hybridizations in Amaranthaceae s.l. Syst. Biol. 70, 219–235.
888 Morales-Briones, D. F., Liston, A., Tank, D. C. 2018. Phylogenomic analyses reveal a deep
889 history of hybridization and polyploidy in the Neotropical genus *Lachemilla*
890 (Rosaceae). New Phytol. 218, 1668–1684.
891 Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., Minh, B. Q. 2015. IQ-TREE: A Fast and
892 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
893 Mol. Biol. Evol. 32, 268–274.
894 Olave, M., Avila, L. J., Sites, J. W., Morando, M. 2018. Detecting hybridization by
895 likelihood calculation of gene tree extra lineages given explicit models. Methods
896 Ecol. Evol. 9, 121–133.
897 Paradis, E., Claude, J., Strimmer, K. 2004. APE: Analyses of Phylogenetics and Evolution
898 in R language. Bioinformatics 20, 289–290.
899 Pease, J. B., Haak, D. C., Hahn, M. W., Moyle, L. C. 2016. Phylogenomics Reveals Three
900 Sources of Adaptive Variation during a Rapid Radiation. PLoS Biol. 14, e1002379.
901 Peng, D., Wang, X.-Q. 2008. Reticulate evolution in *Thuja* inferred from multiple gene
902 sequences: Implications for the study of biogeographical disjunction between eastern
903 Asia and North America. Mol. Phylogen. Evol. 47, 1190–1202.
904 Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A.,
905 Renaud, G., Sudmant, P. H., de Filippo, C.*, et al.* 2014. The complete genome
906 sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49.
907 Qu, X.-J., Wu, C.-S., Chaw, S.-M., Yi, T.-S. 2017. Insights into the Existence of Isomeric
908 Plastomes in Cupressoideae (Cupressaceae). Genome Biol. Evol. 9, 1110–1119.
909 Rabiee, M., Sayyari, E., Mirarab, S. 2019. Multi-allele species reconstruction using
910 ASTRAL. Mol. Phylogen. Evol. 130, 286–296.
911 Rambaut, A., Drummond, A. J., Xie, D., Baele, G., Suchard, M. A. 2018. Posterior
912 Summarization in Bayesian Phylogenetics Using Tracer 1.7. Syst. Biol. 67, 901–904.
913 Ree, R. H., Sanmartín, I. 2018. Conceptual and statistical problems with the DEC+J model
914 of founder-event speciation and its comparison with DEC via model selection. J.
915 Biogeogr. 45, 741–749.
916 Rogalski, M., do Nascimento Vieira, L., Fraga, H. P., Guerra, M. P. 2015. Plastid genomics
917 in horticultural species: importance and applications for plant population genetics,
918 evolution, and biotechnology. Front. Plant Sci. 6, 586.
919 Ru, D., Sun, Y., Wang, D., Chen, Y., Wang, T., Hu, Q., Abbott, R. J., Liu, J. 2018. Population
920 genomic analysis reveals that homoploid hybrid speciation can be a lengthy process.
921 Mol. Ecol. 27, 4875–4887.
922 Salichos, L., Stamatakis, A., Rokas, A. 2014. Novel Information Theory-Based Measures
923 for Quantifying Incongruence among Phylogenetic Trees. Mol. Biol. Evol. 31, 1261–
924 1271.

925 Sayyari, E., Mirarab, S. 2016. Fast Coalescent-Based Computation of Local Branch
926 Support from Quartet Frequencies. Mol. Biol. Evol. 33, 1654–1668.
927 Schweitzer, H. 1974. Die Tertiaren Koniferen Spitzbergens. Palaeontogr. Abt. B-
928 Palaophytol. 149, 1–89.
929 Shao, C. C., Shen, T. T., Jin, W. T., Mao, H. J., Ran, J. H., Wang, X. Q. 2019.
930 Phylotranscriptomics resolves interspecific relationships and indicates multiple
931 historical out-of-North America dispersals through the Bering Land Bridge for the
932 genus *Picea* (Pinaceae). Mol. Phylogen. Evol. 141, 106610.
933 Sloan, D. B., Havird, J. C., Sharbrough, J. 2017. The on-again, off-again relationship
934 between mitochondrial genomes and species boundaries. Mol. Ecol. 26, 2212–2236.
935 Smith, S. A., Moore, M. J., Brown, J. W., Yang, Y. 2015. Analysis of phylogenomic datasets
936 reveals conflict, concordance, and gene duplications with examples from animals and
937 plants. BMC Evol. Biol. 15, 150.
938 Stull, G. W., Soltis, P. S., Soltis, D. E., Gitzendanner, M. A., Smith, S. A. 2020. Nuclear
939 phylogenomic analyses of asterids conflict with plastome trees and support novel
940 relationships among major lineages. Am. J. Bot. 107, 790–805.
941 Sukumaran, J., Holder, M. T. 2010. DendroPy: a Python library for phylogenetic computing.
942 Bioinformatics 26, 1569–1571.
943 Sullivan, A. R., Schiffthaler, B., Thompson, S. L., Street, N. R., Wang, X.-R. 2017.
944 Interspecific Plastome Recombination Reflects Ancient Reticulate Evolution in
945 *Picea* (Pinaceae). Mol. Biol. Evol. 34, 1689–1701.
946 Suyama, M., Torrents, D., Bork, P. 2006. PAL2NAL: robust conversion of protein sequence
947 alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609–
948 W612.
949 Taberlet, P., Luikart, G. 2008. Non-invasive genetic sampling and individual identification.
950 Biol. J. Linn. Soc. 68, 41–55.
951 Taylor, S. A., Larson, E. L. 2019. Insights from genomes into the evolutionary importance
952 and prevalence of hybridization in nature. Nature Ecology & Evolution 3, 170–177.
953 Tiffney, B. H. 1985a. The Eocene North Atlantic land bridge: its importance in Tertiary and
954 modern phytogeography of the northern hemisphere. J. Arnold Arbor. 66, 243–273.
955 Tiffney, B. H. 1985b. Perspectives on the origin of the floristic similarity between Eastern
956 Asia and Eastern North America. J. Arnold Arbor. 66, 73–94.
957 Tiffney, B. H., Manchester, S. R. 2001. The use of geological and paleontological evidence
958 in evaluating plant phylogeographic hypotheses in the Northern Hemisphere Tertiary.
959 Int. J. Plant Sci. 162, S3–S17.
960 van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J.,
961 Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E.*, et al.*
962 2021. Million-year-old DNA sheds light on the genomic history of mammoths.
963 Nature 591, 265–269.

964 Wang, K., Lenstra, J. A., Liu, L., Hu, Q., Ma, T., Qiu, Q., Liu, J. 2018. Incomplete lineage
965 sorting rather than hybridization explains the inconsistent phylogeny of the wisent.
966 Commun. Biol. 1, 169.
967 Wang, M. S., Wang, S., Li, Y., Jhala, Y., Thakur, M., Otecko, N. O., Si, J. F., Chen, H. M.,
968 Shapiro, B., Nielsen, R., *et al.* 2020. Ancient hybridization with an unknown
969 population facilitated high altitude adaptation of canids. Mol. Biol. Evol. 37, 2616–
970 2629.
971 Wang, Q., Mao, K.-S. 2016. Puzzling rocks and complicated clocks: how to optimize
972 molecular dating approaches in historical phytogeography. New Phytol. 209, 1353–
973 1358.
974 Wang, X.-Q., Ran, J.-H. 2014. Evolution and biogeography of gymnosperms. Mol.
975 Phylogen. Evol. 75, 24–40.
976 Wang, Z., Jiang, Y., Bi, H., Lu, Z., Ma, Y., Yang, X., Chen, N., Tian, B., Liu, B., Mao, X.,
977 *et al.* 2021. Hybrid speciation via inheritance of alternate alleles of parental isolating
978 genes. Mol. Plant. 14, 208–222.
979 Wen, D., Yu, Y., Zhu, J., Nakhleh, L. 2018. Inferring Phylogenetic Networks Using
980 PhyloNet. Syst. Biol. 67, 735–740.
981 Wen, J. 1999. Evolution of eastern Asian and eastern North American disjunct distributions
982 in flowering plants. Annu. Rev. Ecol. Syst. 30, 421–455.
983 Wen, J., Ickert-Bond, S. M., Nie, Z.-L., Li, R. 2010. Timing and Modes of Evolution of
984 Eastern Asian - North American Biogeographic Disjunctions in Seed Plants. In: Long,
985 M., H., G., Z., Z. (Eds.), Darwin's heritage today: Proceedings of the Darwin 200
986 Beijing International Conference. Higher Education Press, Beijing, pp. 252–269.
987 Wilgenbusch, J. C., Swofford, D. 2003. Inferring Evolutionary Trees with PAUP*. Current
988 Protocols in Bioinformatics 00, 6.4.1–6.4.28.
989 Xiang, Q.-P., Farjon, A., Li, Z.-Y., Fu, L.-K., Liu, Z.-Y. 2002. *Thuja sutchuenensis*: a
990 rediscovered species of the Cupressaceae. Bot. J. Linn. Soc. 140, 93–94.
991 Xu, X.-H., Yang, L.-Y., Sun, B.-N., Yuan, J.-D., Dong, C., Wang, Y.-D. 2018. A new
992 discovery of Chamaecyparis from the Lower Cretaceous of Inner Mongolia, North
993 China and its significance. Rev. Palaeobot. Palynol. 257, 64–76.
994 Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol.
995 24, 1586–1591.
996 Yu, Y., Blair, C., He, X. 2020. RASP 4: Ancestral State Reconstruction Tool for Multiple
997 Genes and Characters. Mol. Biol. Evol. 37, 604–606.
998 Zachos, J., Pagani, M., Sloan, L., Thomas, E., Billups, K. 2001. Trends, Rhythms, and
999 Aberrations in Global Climate 65 Ma to Present. Science 292, 686.
1000 Zachos, J. C., Dickens, G. R., Zeebe, R. E. 2008. An early Cenozoic perspective on
1001 greenhouse warming and carbon-cycle dynamics. Nature 451, 279–283.
1002 Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S. 2018. ASTRAL-III: polynomial time

species tree reconstruction from partially resolved gene trees. BMC Bioinform. 19, 153.
Zhang, D., Tang, L., Cheng, Y., Hao, Y., Xiong, Y., Song, G., Qu, Y., Rheindt, F. E., Alström, P., Jia, C.*, et al.* 2019. "Ghost Introgression" As a Cause of Deep Mitochondrial Divergence in a Bird Species Complex. Mol. Biol. Evol. 36, 2375–2386.
Zhang, Q., Ree, R. H., Salamin, N., Xing, Y., Silvestro, D. 2021. Fossil-Informed Models Reveal a Boreotropical Origin and Divergent Evolutionary Trajectories in the Walnut Family (Juglandaceae). Syst. Biol. https://doi.org/10.1093/sysbio/syab030
Zhao, T., Wang, G., Ma, Q., Liang, L., Yang, Z. 2020. Multilocus data reveal deep phylogenetic relationships and intercontinental biogeography of the Eurasian-North American genus *Corylus* (Betulaceae). Mol. Phylogen. Evol. 142, 106658.
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). J. Syst. Evol. 58, 571–603.

Supporting Information

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

Table S1 Sample information.

Table S2 Fossil calibrations for divergence time estimation.

Table S3 The dispersal probability matrix.

Table S4 Summary of gene recovery efficiency for assemblies via HybPiper.

Table S5 Information on plastid genome assemblies.

Table S6 All possible 105 topologies for *Thuja* with their frequencies in observed nuclear gene trees, simulated nuclear gene trees, and simulated plastid trees based on all 5,663 loci of the one-individual data set.

Table S7 All possible 105 topologies for *Thuja* with their frequencies in observed nuclear gene trees, simulated nuclear gene trees, and simulated plastid trees based on 2,969 no-recombination loci of the one-individual data set.

Table S8 Parameter estimates under multispecies coalescent model using the software BPP.

Table S9 Parameter estimates under multispecies-coalescent-with-introgression model

using the software BPP.

Table S10 Result of model test performed in BioGeoBEARS including the +*J* parameter using extant species only.

Table S11 Result of model test performed in BioGeoBEARS not including the +*J* parameter using extant species only.

Table S12 Result of model test performed in BioGeoBEARS including the +*J* parameter using both extant and fossil *Thuja* species.

Table S13 Result of model test performed in BioGeoBEARS not including the +J parameter using both extant and fossil *Thuja* species.

Figure S1 Concatenated tree inferred from PUAP using maximum parsimony method based on all loci of the multi-individual dataset.

Figure S2 Concatenated tree inferred from IQ-TREE using maximum likelihood method based on all loci of the multi-individual dataset.

Figure S3 Astral tree with branch lengths in coalescent unit based on all loci of the multi-individual dataset.

Figure S4 Astral tree with branch lengths in coalescent unit based on non-recombinant loci of the multi-individual dataset.

Figure S5 Maximum parsimony tree based on nearly complete plastid genomic alignment.

Figure S6 Maximum likelihood tree based on nearly complete plastid genomic alignment.

Figure S7 Species network analysis and test of hybridization based on 1,145 non-recombination loci of the multi-individual dataset.

Figure S8 ILS simulations based on 2,969 non-recombination loci from the one-individual data set.

Figure S9 (a) Distribution of Robinson-Foulds (RF) distances of the simulated (blue violin plot) and true (orange numbers and points) gene trees to the species tree using protein

sequences. Violin plots are from 100 replicated simulations (each containing 2,969 gene trees). (b) A "cloudogram" of 2,969 gene trees using protein sequences for the one-individual dataset. (c) An MP-EST tree with branch in coalescent unit based on protein sequences.

Figure S10 Topology frequencies of simulated gene trees and observed gene trees based on (a) DNA sequences, and (b) protein sequences.

Figure S11 Species and plastid trees inferred under multispecies coalescent model based on (a) 1,145 nuclear CDS genes (non-recombinant loci of the multi-individual dataset), (b) full plastome sequences, (c) 68 plastome fragments with 2000bp in length, (d) plastid CDS genes, assuming that each 2-kp or coding gene experience independent evolutionary history.

Figure Legends

Figure 1 Phylogenetic relationship within *Thuja* based on the multi-individual dataset. (a) species tree inferred with ASTRAL based on 2,369 single copy nuclear genes, which had the same species-level topologies as recovered from PAUP and IQ-TREE based on maximum parsimony (MP) and maximum likelihood (ML) approaches, respectively. The 'internode certainty all scores' are shown below the branches. Number of gene trees concordant/conflicting with the shown node are depicted next to the nodes. Pie charts of the nodes denote the proportion of gene trees that support the shown topology (blue), support the main alternative topology (orange), and support the remaining alternatives (grey). (b) Maximum likelihood tree inferred from IQ-TREE using the full plastome sequence alignment. Pie charts of the nodes denote the site concordance factor averaged over 100 quartets (sCF; blue), site discordance factor for alternative quartet 1 (sDF1; orange), and site discordance factor for alternative quartet 2 (sDF2; grey). The MP/ML bootstrap values (/ASTRAL local posterior probabilities) are shown above the branches. The "*" denotes the branch supported with 100% bootstrap values (and a local posterior

probability of 1).

Figure 2 (a-d) Phylogenetic networks inferred from PhyloNet pseudolikelihood analyses with one (a), two (b), and three (c) hybridization events based on all loci of the multi-individual dataset. The major and minor edges of hybrid nodes are shown as blue and orange branches, respectively. (d) Patterson's *D* tests of all possible ten topologies within *Thuja*. The arrows denote gene flow between distantly related populations. Patterson's *D* and Z scores are shown above and under the arrows, respectively. Tdo: *Thujopsis dolabrata*; Toc: *T. occidentalis*, Tko: *T. koraiensis*; Tpl: *T. plicata*; Tst: *T. standishii*; Tsu: *T. sutchuenensis*; and ln*L*: log-likelihood.

Figure 3 (a) Conflict among gene trees for the all loci of one-individual dataset. Numbers above branches indicate the 'internode certainty all scores' of that node. The number of gene trees which are in concordance/conflict with the shown node is stated next to the nodes. Pie charts denote the proportion of gene trees that support the shown topology (blue), support the main alternative topology (orange), or support the remaining alternatives (grey). (b) Distributions of topology frequencies of observed and simulated gene trees based on all 5,663 loci of one-individual dataset. (c) Distribution of Robinson-Foulds (RF) distances of the simulated (blue violin plot) and true (orange numbers and points) gene trees to the species tree. Violin plots are from 100 replicated simulations (each containing 5,663 gene trees). (d) Coalescent model showing that *T. plicata* fixed a different plastid genome. (e) An astral tree with branch length in coalescent units. The branch lengths are inferred from the multi-individual dataset. The internodes that fall in the anomaly zone are marked in blue and orange. (f) Concordance of simulated plastid gene trees and observed plastid phylogeny. Numbers after nodes represent the number of genes trees which support the shown clades.

Figure 4 A time calibrated phylogeny of five *Thuja* species and 11 other Cupressaceae

species. The times were inferred by MCMCTree based on a concatenated set of 1811 nuclear single copy genes. The divergence times are shown behind the nodes, and the 95% highest posterior densities are represented as light-grey bars.

Figure 5 Species trees of all five extant *Thuja* species and the outgroup *Thujopsis dolabrata* including the parameter estimates based on (a) the multispecies coalescent model and (b) the multispecies-coalescent-with-introgression model using the software BPP. The absolute divergence times were calculated from the posterior mean branch lengths (τ) by calibrating the stem age of *Thuja* to 62.68 Ma (as inferred by MCMCTree). The posterior mean of population sizes (θ) and introgression probability (φ) are shown. All parameter estimates are based on the multi-individual dataset after removal of recombinant loci, and the 95% highest posterior densities for the divergence times are represented as light-grey bars.

Figure 6 Ancestral area reconstructions of *Thuja*. (a) Biogeographic regions defined in this study. A: eastern Asia; B: western North America; C: eastern North America. (b) Ancestral ranges inferred from the species tree without fossil taxa based on the DEC model. (c) Ancestral ranges inferred from the species tree including fossil taxa based on the DIVALIKE+J model. (d) Ancestral ranges inferred from the species tree including fossil taxa based on the DEC model.