A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae

Citation for published version:

Digital Object Identifier (DOI):
10.1093/aob/mcy152

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Annals of Botany

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae

Thank you for agreeing to review this paper for Annals of Botany. The Annals of Botany aims to be among the very top of plant science journals and as we receive over 1000 submissions every year we need to be very selective in deciding which papers we can publish. In making your assessment of the manuscript's suitability for publication in the journal please consider the following points.

Scientific Scope

Annals of Botany welcomes papers in all areas of plant science. Papers may address questions at any level of biological organization ranging from molecular through cells and organs, to whole organisms, species, communities and ecosystems. Its scope extends to all flowering and non-flowering taxa, and to evolutionary and pathology research. Many questions are addressed using comparative studies, genetics, genomics, molecular tools, and modeling.

To merit publication in Annals of Botany, contributions should be substantial, concise, written in clear English and combine originality of content with potential general interest.

- We want to publish papers where our reviewers are enthusiastic about the science: is this a paper that you would keep for reference, or pass on to your colleagues? If the answer is “no” then please enter a low priority score when you submit your report.

- We want to publish papers with novel and original content that move the subject forward, not papers that report incremental advances or findings that are already well known in other species. Please consider this when you enter a score for originality when you submit your report.

Notes on categories of papers:

All review-type articles should be novel, rigorous, substantial and “make a difference” to plant science. The purpose is to summarise, clearly and succinctly, the “cutting edge” of the subject and how future research would best be directed. Reviews should be relevant to a broad audience and all should have a strong conclusion and illustrations including diagrams.

- Primary Research articles should report on original research relevant to the scope of the journal, demonstrating an important advance in the subject area, and the results should be clearly presented, novel and supported by appropriate experimental approaches. The Introduction should clearly set the context for the work and the Discussion should demonstrate the importance of the results within that context. Concise speculation, models and hypotheses are encouraged, but must be informed by the results and by the authors' expert knowledge of the subject.

- Reviews should place the subject in context, add significantly to previous reviews in the subject area and moving forward research in the subject area. Reviews should be selective, including the most important and best, up-to-date, references, not a blow-by-blow and exhaustive listing.

- Research in Context should combine a review/overview of a subject area with original research, often leading to new ideas or models; they present a hybrid of review and research. Typically a Research in Context article contains an extended Introduction that provides a general overview of the topic before incorporating new research results with a Discussion proposing general models and the impact of the research.

- Viewpoints are shorter reviews, presenting clear, concise and logical arguments supporting the authors’ opinions, and in doing so help to stimulate discussions within the topic.

- Botanical Briefings are concise, perhaps more specialised reviews and usually cover topical issues, maybe involving some controversy.
Article type: Original Article

Title:

A transcriptome-based resolution for a key taxonomic controversy in Cupressaceae

Kangshan Mao¹,†,*, Markus Ruhsam²,†, Yazhen Ma¹, Sean W. Graham³, Jianquan Liu¹, Philip Thomas², Richard I. Milne⁴, Peter M. Hollingsworth²

Affiliation:
¹Key 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
²Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh, EH3 5LR, UK
³Department of Botany, University of British Columbia, Vancouver, V6T 1Z4, Canada
⁴Institute of Molecular Plant Science, School of Biological Science, University of Edinburgh, Edinburgh, EH9 3BF, UK

†These authors contributed equally to this work

Running title: A transcriptome-based phylogeny for junipers and cypresses

*For correspondence: maokangshan@scu.edu.cn or maokangshan@163.com
ABSTRACT

Background and Aims Rapid evolutionary divergence and reticulate evolution may result in phylogenetic relationships that are difficult to resolve using small nucleotide sequence datasets. Next-generation sequencing methods can generate larger datasets that are better suited to solving these puzzles. One major and long-standing controversy in conifers concerns generic relationships within the subfamily Cupressoideae (105 species, ~1/6 of all conifers) of Cupressaceae, in particular the relationship between Juniperus, Cupressus and the Hesperocyparis-Callitropsis-Xanthocyparis (HCX) clade. Here we attempt to resolve this question using transcriptome-derived data.

Methods Transcriptome sequences of 20 species from Cupressoideae were collected. Using MarkerMiner, single copy nuclear (SCN) genes were extracted. These were applied to estimate phylogenies based on concatenated data, species trees, and a phylogenetic network. We further examined the effect of alternative backbone topologies on downstream analyses, including biogeographic inference and dating analysis.

Results Based on the 73 SCN genes (>200,000 bp total alignment length) we considered, all tree-building methods lent strong support for the relationship (HCX, (Juniperus, Cupressus)); however, strongly supported conflicts among individual gene trees was also detected. Molecular dating suggests that these three lineages shared a most recent common ancestor ~60 Mya, and that Juniperus and Cupressus diverged ~56 Mya. Ancestral area reconstructions (AARs) suggest an Asian origin for the entire clade, with subsequent dispersal to North America, Europe and Africa.

Conclusions Our analysis of SCN genes resolves a controversial phylogenetic relationship in the Cupressoideae, a major clade of conifers, and suggests that rapid evolutionary divergence and incomplete lineage sorting likely acted together as the source for conflicting phylogenetic inferences between gene trees and between our robust results and recently published studies. Our updated backbone topology has not substantially altered molecular dating estimates relative
to previous studies, however application of the latest AAR approaches has yielded a clearer picture of the biogeographic history of Cupressoideae.

Key words: single-copy nuclear genes, transcriptome, Cupressoideae, *Hesperocyparis*, *Cupressus*, *Juniperus*, *Xanthocyparis*, *Callitropsis*
INTRODUCTION

It can be challenging to accurately reconstruct deep phylogenetic relationships within groups that experienced rapid evolutionary divergence, incomplete lineage sorting and/or reticulate evolution, especially with small datasets (Maddison, 1997; Dunn et al., 2008; Jian et al., 2008; Zeng et al., 2014; Ruhsam et al., 2015). Rapid evolutionary divergence may lead to short internodal distances and soft polytomies (Weisrock et al., 2005; Whitfield & Lockhart, 2007; Jian et al., 2008; Pyron et al., 2014; Leaché et al., 2016). In addition, incomplete lineage sorting, which involves mis-sorting of ancestral polymorphisms relative to the species tree, or reticulate evolution, which involves the combination or transmission of genetic material between divergent evolutionary lineages due to hybridization and introgression, may both cause inaccurate or conflicting species-tree inference (Beiko et al., 2008; Sun et al., 2015).

Next-generation sequencing approaches, which generate large amounts of DNA sequence data from throughout the genome, are transforming phylogenetic inference (e.g. Dunn et al., 2008; Lee et al., 2011; Faircloth et al., 2012; Zeng et al., 2014). This is especially true where rapid evolutionary events resulted in few fixed substitutions between divergent species, yielding gene trees that are usually unresolved with respect to the true species tree, when only a few loci are used (Whitfield & Lockhart, 2007). A larger amount of sequence data is likely to capture such species-specific substitutions, potentially resulting in improved phylogenetic resolution (Jian et al., 2008; Zeng et al., 2014). In the case of incomplete lineage sorting, many independent gene trees from throughout the genome can be used to estimate a credible species tree by reconciling genealogical discordance between loci (Edwards, 2009; Lemmon and Lemmon, 2013). Therefore, phylogenetic estimation of species trees based on genomic datasets might resolve branches that were poorly supported in smaller datasets (Rokas et al., 2003; Dunn et al., 2008). For example, phylogenetic analyses using as few as 29 and 59 low-copy nuclear genes have resulted in well-resolved deep phylogenetic estimates for ferns (Rothfels et al., 2015) and flowering plants (Zeng et al., 2014), respectively.
Two main methods have recently been proposed to construct species trees from large datasets (Liu et al., 2009a, 2009b, 2015). One method uses the multiple-species coalescent model as implemented in the program *BEAST (Heled and Drummond, 2010), which estimates gene trees and the species tree at the same time. However, this method is computationally intensive (Edwards et al., 2007; Pyron et al., 2014), and may result in poor convergence if the dataset is large (O’Neill et al., 2013). The other method uses a two-step approach when estimating species trees. In the first step gene trees are generated using software such as RaxML (Stamatakis et al., 2014), and in the second step they are summarized under the coalescent model as implemented in the software MP-EST (Liu et al., 2010), STAR (Liu et al., 2009a). This method reduces computation time considerably when compared to analyses based on the multiple species coalescent model (Liu et al., 2009b). In addition, a recently developed two-step approach, ASTRAL-II (Mirarab et al., 2015; Mayyari and Mirarab, 2016), has been shown to run much faster and to be less sensitive than MP-EST to the effects of gene tree errors, when estimating a species tree based on large dataset (e.g. hundreds of taxa and thousands of genes). The accuracy of ASTRAL remains high when a small number of genes is adopted and a moderate level of incomplete lineage sorting is assumed, whereas its local posterior probabilities of quartet branches are conservative; this leads to very few false positives that have high support, at the cost of missing some true positives (Mayyari and Mirarab, 2016).

Cupressaceae, also known as the cypress family, contains more than 160 species in 32 genera, of which 17 are monotypic (Farjon, 2005; Mao et al., 2012; Yang et al., 2012; Wang & Ran, 2014; Adams 2014). They occur in many different habitats on all continents except Antarctica (Farjon, 2005). Cupressoideae, which contains more than 100 species in 13 genera, is the largest of the seven subfamilies of Cupressaceae (Gadek et al., 2000; Mao et al., 2012; Yang et al., 2012). This subfamily occurs throughout the Northern Hemisphere and contains many ecologically important and dominant species especially in mountainous and arid or semi-arid regions (Farjon, 2005; Adams 2014). It contains many economically important timber species (e.g. *Calocedrus*, *Chamaecyparis*, *Cupressus* and *Thuja*) and ornamental trees (e.g.
Chamaecyparis, Juniperus, Platycladus and Thuja) (Farjon, 2005). Phylogenetic analyses suggest that this subfamily is monophyletic (Gadek et al., 2000; Mao et al., 2012; Yang et al., 2012) and comprises four clades (Gadek et al., 2000; Little 2006; Mao et al., 2012; Yang et al., 2012) which have been treated as separate tribes by some authors (Gadek et al., 2000). However, taxonomic treatment at the generic level and inter-generic relationships within the subfamily remains controversial (Little et al., 2004; Little 2006; Mill & Farjon, 2006; Rushforth, 2007; Adams et al., 2009; Christenhusz et al., 2011; Dörken et al., 2017), especially for Cupressus sensu lato (s.l.), which comprises 30 species (Little, 2006; Christenhusz et al., 2011; Dörken et al., 2017). Cupressus s.l. may be divided into four genera: Cupressus sensu stricto (s.s.) and Xanthocyparis s.s. in the Old World, and Hesperocyparis and Callitropsis s.s. in the New World (Adams et al., 2009; Mao et al., 2010; Christenhusz et al., 2011) (see Table 1 for a summary of taxonomic treatment history). Henceforth, if not stated otherwise, “Cupressus”, “Xanthocyparis”, and “Callitropsis” refer to Cupressus s.s., Xanthocyparis s.s. and Callitropsis s.s., respectively. Although the monophyly of Cupressus and the Hesperocyparis-Callitropsis-Xanthocyparis clade (the HCX clade; Mao et al., 2010) is well defined (Little et al., 2004; Little 2006; Mao et al., 2010, 2012; Yang et al., 2012), the phylogenetic relationship between Cupressus, the HCX clade and Juniperus remains uncertain. All possible phylogenetic topologies among these three clades have been supported by different studies with different datasets and analyses, as follows: (Cupressus, (Juniperus, HCX)) topology was recovered by Xiang & Li (2005), Adams et al. (2009) and Terry & Adams (2015); (Juniperus, (Cupressus, HCX)) topology by Mao et al. (2010); (HCX, (Cupressus, Juniperus)) topology by Little (2006) and Yang et al. (2012); and a trichotomy (HCX, Cupressus, Juniperus) by Mao et al. (2012). From here on, these topologies are referred to as Cu(HCX,Ju), Ju(Cu,HCX), HCX(Cu,Ju) and (HCX,Cu,Ju), respectively, for simplicity. A recent phylogenomic study based on the whole plastid genomes of 22 species of Cupressaceae and accounting for long branch attraction (e.g., Felsenstein, 1978; Hendy and Penny, 1989) supported the Ju(Cu,HCX) topology (Qu et al., 2017). However, all of these studies either used no more than four bi-parentally inherited nuclear loci (e.g. Little 2006; Adams
et al., 2009) or plastid DNA (ptDNA), the latter of which, despite the use of nine (Mao et al., 2010), 11 ptDNA regions (Terry & Adams, 2015) or even the whole plastid genome (Qu et al., 2017), can be considered to be a single locus due to its lack of recombination.

The aim of the current study is to resolve this long-standing controversy and to reconstruct the phylogenetic relationship between Cupressus, Juniperus and the HCX clade based on a number of single or low copy nuclear loci from transcriptome data using 17 species representing major lineages within these three clades, plus three outgroups. Specifically, we investigate (a) the evolutionary relationship between the three major lineages using a phylotranscriptomic approach, (b) compare and explain the discordance and agreement between the current species tree topology and phylogenetic topologies that were gained in previous studies, and characterize the impact of different topologies of the three major lineages on (c) molecular dating of this group and (d) the inference of its biogeographic history.

MATERIALS AND METHODS

Provenance of samples

Fresh leaf samples from a total of 18 species (including outgroup species Microbiota decussata) were collected for transcriptome sequencing. Fourteen samples were collected from the living collection of the Royal Botanic Garden Edinburgh (RBGE), three were collected in the field in Yunnan, China (Cupressus duclouxiana) and Xizang, China (Cupressus gigantea and Juniperus microsperma), and one (Cupressus funebris) was a cultivated individual from the campus of Sichuan University, Chengdu, China (Table 2). Additionally, we used transcriptome data for three outgroup species (Calocedrus decurrens, M. decussata, Thuja plicata) from the one thousand transcriptome project (‘1000 plant project,’ 1KP) (Table 2). All species were represented by a single accession apart from M. decussata (n=2).
Transcriptome sequencing, assembly, and alignment

Transcriptomes were either generated in Edinburgh/UK (RBGE, Table 2) or Chengdu/China (SZ, Table 2) apart from three downloaded from the 1KP project (http://www.onekp.com/samples/list.php; labelled as ‘1kp’ in Table 2). RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma-Aldrich, St. Louis, Missouri, USA) following protocol A with a few minor modifications (2-3 times the amount of lysis buffer, 750 µl binding buffer and three final washes). Library preparation and sequencing was outsourced to Edinburgh Genomics (Edinburgh, UK) and Novogene (Beijing, China) for RBGE and SZ samples, respectively (Table 2). Transcriptomes were sequenced on Illumina HiSeq platforms generating 2 × 100 bp paired-end reads. Raw reads were prepared for assembly using Trimmomatic (Bolger et al., 2014) with the parameters ‘LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36’ and cutadapt (Martin 2011) to remove adapters and low quality sequences. Reads for each taxon were then assembled into contigs with SOAPdenovo-trans (Xie et al., 2014) using SOAPdenovo-Trans-31mer with ‘-K 29 -L 100’. The programme Cd-hit (Li & Godzik 2006), software for clustering and comparing protein or nucleotide sequences, was used to retrieve only unique contigs from the SOAPdenovo-trans analysis with the command cd-hit-est and default values. The output of Cd-hit was then fed into MarkerMiner v1.0 (Chamala et al., 2015) with parameters ‘-singleCopyReference Athaliana -minTranscriptLen 900’. MarkerMiner identifies and aligns putative orthologous single or low copy nuclear genes in a set of transcriptome assemblies, using a reciprocal BLAST search against a reference database (Chamala et al., 2015). The alignments of genes included for further analyses were visually checked with minimal editing and trimming either side of the sequence where missing sites accounted for more than half of all available taxa. In subsequent analyses data from the two Microbiota accessions (Table 2) were amalgamated to represent one sample in order to minimise the amount of missing data for that species.
Phylogenetic analyses

Alignments of putative single copy loci from MarkerMiner v1.0 (Chamala et al., 2015) were used to compile two sets of data, the first comprising individual genes, in which each locus is treated independently, and the second a concatenated dataset in which all chosen loci were combined into one ‘super locus’. First, we used three conventional methods, MP, ML and Bayesian Inference to infer phylogenetic trees based on the concatenated dataset: analyses based on such dataset could lead to species-tree mis-inference if there is sufficient conflict between gene trees, but these concatenation-based methods often recovers the same tree that other species-tree estimation methods recover (e.g. Wickett et al., 2014) and is commonly done to compare to other species-tree estimation methods, e.g. MP-EST (Liu et al., 2010), STAR (Liu et al., 2009a), ASTRAL (Mirarab et al., 2014). Maximum parsimony analysis (MP) was performed using PAUP*4.0b10 (Swofford 2003), with gaps treated as missing data and polymorphic states as uncertain. A ‘branch and bound’ search with MulTrees on was carried out for both datasets. Branch support was estimated via bootstrapping with 1000 bootstrap replicates using heuristic searches (Felsenstein, 1985). We also used RAxML v8 (Stamatakis 2014) to estimate a maximum likelihood (ML) tree and ML bootstrap values, by applying the parameters ‘-f a -m GTRGAMMA -p 12345 -x 12345 -# 1000’ where the GTRGAMMA model and 1000 bootstrap replicates were applied (see RAxML manual for detailed parameter settings). A Bayesian inference analysis (BI) was also performed using MrBayes v. 3.1.2 (Huelsenbeck & Ronquist 2001; Ronquist & Huelsenbeck 2003) with the GTR+I+G model, which was selected using MrModelTest v. 2.3 (Nylander 2004) under the Akaike Information Criterion. The analysis was run for 2 million generations with four MCMC chains in two independent parallel analyses, with one tree sampled every 500 generations. The average standard deviation of split frequencies was 0.00000 at the end of the run. TRACER v1.5 (Rambaut & Drummond 2009) was used to assess the quality of the MCMC simulations and suggested a high degree of convergence between runs. The effective sample size values (ESS), that is the number of effectively independent draws from the posterior, were >500 for all parameters, indicating that sufficient sampling occurred.
From the individual gene dataset, we constructed individual ML gene trees for each locus using the software RAxML v8 (Stamatakis 2014) applying the parameters as for the concatenated data set. The topology of each gene tree was then manually examined, looking in particular for well-supported alternative relationships that might indicate gene-tree conflict.

Then, based on these gene trees, we generated a species tree based on the multispecies coalescent model in ASTRAL 5.6.1 (Mirarab et al., 2014; Mayyari and Mirarab, 2016), which estimates species trees from unrooted gene trees, and maximizes the number of quartet trees shared between the gene trees and the species tree. ASTRAL-II estimates branch lengths for internal branches (not terminal branches) in coalescent units, and branch support values measure the support for a quadripartition (the four clusters around a branch) and not the bipartition, as is commonly done. The species tree was fully annotated using “-t 4” option, which calculates the measurements for each branch, including quartet support (q), total number of quartet trees in all the gene trees (f), and the local posterior probabilities (pp) for the main topology and the first, second alternatives, total number of quartets defined around each branch (QC), and the effective number of genes for the branch (EN).

Conservative pp scores cause some true positives to be overlooked in ASTRAL (Mayyari and Mirarab, 2016); moreover, average positive branch rates, which represent the proportion of the estimated species tree in which a certain branch is successfully recovered, may be lower in ASTRAL than in STAR and MP-EST (Liu et al., 2015). Therefore, we also conducted STAR and MP-EST analyses based on gene trees to reduce the chance of missing any true positives, and to improve the average positive branch rates. Hence the rooted ‘best tree’ RAxML output for each gene plus bootstrap values for each gene tree using 1000 replicates was then uploaded to ‘The Species TRee Analysis Webserver’ STRAW (Shaw et al., 2013) to estimate the species tree using STAR (Liu et al., 2009a) and MP-EST (Liu et al., 2010). Both programmes apply the multispecies coalescent model (Rannala & Yang 2003) to obtain estimates of the species tree from gene trees. STAR (Liu et al., 2009a) uses the average ranks of coalescences, whereas
MP-EST (Liu et al., 2010) uses a pseudo-likelihood function of the species tree, and both of them generate bootstrap support values using nonparametric bootstrap techniques (Liu et al., 2009a, 2010). Both methods are based on summary statistics calculated across all gene trees, with the effect that a small number of genes that significantly deviate from the coalescent model will have relatively little effect on the ability to accurately infer the species tree.

Because there was some well-supported conflict among gene trees (see Results), we conducted two additional analyses to investigate this further. First, we applied MulRF (Chaudhary et al., 2013, 2015) to estimate the best species tree, i.e. the one that minimizes the overall Robinson-Foulds (RF) distance between each candidate species tree and the individual gene trees. This software is also able to calculate the MulRF score of a given tree topology, which is the RF distance between this given tree and all gene trees. In a soft polytomy where relationship among three clades are difficult to resolve, this function may be used to compare the compatibility of each of the three dichotomy candidate species trees with all gene trees.

Finally, we used the NeighborNet method implemented in SplitsTree 4.11.3 (Huson and Bryant, 2006) to reconstruct phylogenetic networks based on the concatenated alignment of all 73 nuclear genes. For distance calculations, we excluded insertions/deletions (indels) and used the K2P model (Kimura, 1980). The relative robustness of the clades was estimated by performing 1000 bootstrap replicates, and a confidence network was generated with a 95% threshold (Huson and Bryant, 2006). This analysis can summarize how homoplasy that might include hybridization or incomplete lineage sorting might have affected the phylogenetic reconstruction.

Molecular dating

To investigate the impact of topological differences on the evolutionary divergence timescale in Cupressoideae, we conducted molecular dating analyses. We tried to adopt the eight fossil calibration points as in Mao et al. (2010), but only three of them could be used for the dating of our 20-taxon data set, whereas the remaining five could only be attached to apparently deeper
nodes, relative to those in the phylogeny of Mao et al. (2010). As too few calibration points
and/or assigning fossils to deeper nodes (due to sparse sampling) has been shown to bias the
estimates of node ages (e.g. Linder et al., 2005; Mao et al., 2012; Wang & Mao, 2016), we
adopted a hybrid strategy to reconstruct the evolutionary divergence timescale of Cupressoideae.
Hence dating was carried out on our previous ptDNA dataset comprising nine ptDNA fragments
from 84 species (Mao et al., 2010), but with the relationship between the three main clades
constrained to the topology from the current study based on transcriptome data (see below). The
original ptDNA dataset comprising 92 accessions was slightly reduced by removing multiple
accessions of six species, resulting in a final dataset of 86 accessions representing 84 species in
Cupressoideae (referred to as ‘86-accession data set’ from here on). Three parallel molecular
dating analyses were carried out, one constraining to the HCX(Cu,Ju) topology, another
constraining to the Cu(HCX,Ju) topology, and the third was unconstrained, allowing it to retain
the Ju(Cu,HCX) topology from Mao et al. (2010). We adopted eight calibration fossils from Mao
et al. (2010), seven of which were used as minimum age constraints with uniform priors, and one
was set as a fixed age constraint with a normal prior (see Table 1 in Mao et al., 2010 for details).
BEAST version 1.8.0 (Drummond & Rambaut, 2007) was used to simultaneously estimate
topology, substitution rates and node ages by employing a Bayesian MCMC chain. BEAST
parameter settings, including fossil calibration settings, were all the same as in Mao et al. (2010),
except that two independent MCMC analyses of 100 000 000 generations were conducted,
sampled every 2 000 generations, with 20% burn-in. The program Tracer 1.5.1 (Rambaut &
Drummond, 2007) was employed to check effective sample size, and the program TreeAnnotator
1.8.0 (part of the BEAST 1.8.0 package) was used to summarize the output results. Finally, a tree
with ages for each node and their 95% highest posterior density intervals (95%HPD), was
displayed and formatted in FigTree 1.3.1 (Rambaut, 2008).

Ancestral area reconstruction
We conducted an ancestral area reconstruction using the BioGeoBEARS packages as implemented in RASP 4.0 (Yu et al., 2015). Four operational geographic areas (A: North America, B: Africa, C: Asia, D: Europe), were defined for our analyses, following those in Mao et al. (2010). A total of 400 trees, which were resampled from the output trees of the BEAST analysis, and the BEAST summary tree, were imported into RASP 4.0, along with the distribution information of each species. BioGeoBEARS allows the testing of six models (DIVALIKE, DIVALIKE+J, DEC, DEC+j, BAYAREALIKE and BAYAREALIKE+J) (Matzke 2013, 2014). Of the six models, the DIVALIKE model (Ronquist, 1997) is an event-based approach that adopts a simple biogeographic model; it does not consider general area relationships or branch lengths of the input tree, and it applies different costs to vicariance, duplication, dispersal, and extinction to construct ancestral distributions (Ronquist, 1997; Yu et al., 2015). The DEC model (Ree and Smith, 2008) allows dispersal, extinction, and cladogenesis as fundamental processes, accommodates differing dispersal probabilities among areas across different time-periods, and can integrate branch lengths, divergence times, and geological information (Ree and Smith, 2008; Yu et al., 2015). In contrast to the former two models, which accept only bifurcating trees, the BAYAREALIKE model (Landis et al., 2013) allows polytomies. It considers distribution area to be a ‘trait’ of a species, and hence reconstructs ancestral ‘traits’ using Bayesian inference; furthermore, it does not define the dispersal rate, constrain the maximum number of areas at each node, or exclude widespread and unlikely ancestral areas before analysis (Landis et al., 2013; Yu et al., 2015). The other three models with the ‘+J’ suffix (i.e., DIVALIKE+J, DEC+j, BAYAREALIKE+J) allow founder-event speciation, in contrast to the three original models (Matzke, 2014).

We conducted model testing and two models (the best and the second-best model, as given in the Results section) were employed to reconstruct the ancestral area for every node in the phylogeny based on 100 trees that were randomly selected from 400 BEAST trees. At most two areas were allowed for any node in any tree. An among-area dispersal probability matrix, which is the same as in Mao et al. (2012), was coded to define different dispersal probabilities in five
time periods, 0-5 Mya, 5-30 Mya, 30-45 Mya, 45-70 Mya and 70-115 Mya. The ancestral area
reconstruction results were optimized in the Treeview window of the RASP program.

RESULTS

Phylogenetic analyses

A total of 581 putative single or low copy genes were detected by MarkerMiner. About 50% of
these were shared by five or fewer samples. To minimise the impact of missing data on the
ability to confidently resolve phylogenetic relationships, we only included genes that did not
have more than two ingroup and two outgroup taxa missing, for further phylogenetic analyses.
This resulted in 73 putatively single or low copy genes yielding an alignment of 208,484 nuclear
base pairs for 20 taxa (Supplementary Table S1). The DNA sequences have been deposited in
NCBI GenBank (accession numbers shown in Table S2).

For six of the 73 genes chosen by MarkerMiner (Chamala et al., 2015), a secondary transcript
that passed the BLAST filtering process was reported for one of the included taxa
(Supplementary Table S1). These secondary transcripts, defined as the one of the two from a
given taxon that received a lower BLAST score, may represent splice isoforms, putative
paralogs, or partially assembled transcripts (Chamala et al., 2015). However, as removing the
taxon from the particular alignment for which a secondary transcript was detected did not change
the phylogenetic results in any way, for these species (data not shown), the transcript for each
taxon with the higher BLAST score was included in all analyses.

In this matrix consisting of 73 genes, 183,022 (87.8%) characters were constant, 16,405
(7.9%) were variable but parsimony-uninformative and 9,057 (4.3%) were
parsimony-informative. The same topology was retrieved regardless of the bifurcate
tree-building method used (including MP, ML, BI, MP-EST, STAR, ASTRAL-II, MulRF), with
an HCX clade (comprising a monophyletic *Hesperocyparis*, plus *Callitropsis* and
Xanthocyparis) as the sister group of a clade consisting of Cupressus (monophyletic) and Juniperus (also monophyletic) (Figs 1, 2). Branch support was high for the MP, ML and BI analyses of the concatenated data set and for the MP-EST, STAR and ASTRAL-II analyses using a coalescent approach. Individual trees produced by Bayesian analysis and the three coalescent approaches were identical. The MP tree topology differed from these in only one respect: here (J. procera (J. indica, J. microsperma)) was sister to (J. flaccida, J. scopulorum) (Supplementary Fig. S1), whereas in the other analyses J. procera was sister to ((J. indica, J. microsperma) (J. flaccida, J. scopulorum)) (Figs 1, 2). However, two branches concerning this relationship were weakly or moderately supported by the bootstrap analysis in the MP analysis (BS=55% and 76%, respectively, Fig. S1). The Juniperus clade was also the only ingroup clade where some of the internal relationships did not receive maximum branch support; this was true for all six analyses methods used (MP, BI, ML, MP-EST, STAR and ASTRAL-II; Figs 1, 2).

The HCX(Cu,Ju) topology of our species tree based on single nuclear genes conflicts with the Ju(Cu,HCX) topology based on the ptDNA data from Mao et al. (2010) (Supplementary Fig. S2). Quartet support analyses in ASTRAL-II suggest that the HCX(Cu,Ju), Ju(Cu,HCX) and (Cu(HCX,Ju)) topologies are supported by 54.16%, 24.24% and 21.60% of the gene trees, respectively (Fig. 2). A manual check of gene trees (Fig. S3) which were generated in RAXML using maximum likelihood bootstrapping (MLBS) indicated similar proportions of gene trees supporting these three topologies (i.e., most supporting the HCX(Cu,Ju) topology, which is equivalent to the Cupressus-Juniperus sister topology in Table S1), except that a few MLBS gene trees are unresolved (Table S3). We also calculated the MulRF score (the total Robinson-Foulds topological distance of all 73 gene trees against the candidate species tree) for each of the above three topologies concerning Cupressus, Juniperus and HCX (other relationships remain the same). The HCX(Cu,Ju), Ju(Cu,HCX) and Cu(HCX,Ju) topologies received MulRF scores of 744 (= closest and therefore favored), 786 and 790 (= furthest), respectively.
Finally, NeighborNet analyses provide 100% bootstrap support for the quartet branch that links (*Juniperus, Cupressus*) and (*HCX, outgroups*), although the length of this branch is relatively short (Fig. 3A). Very few strongly supported relationships that might have suggested hybridization or incomplete lineage sorting (bootstrap support >95%) were recovered in the NeighborNet confidence network; these were mainly found within *Juniperus*, but also once within *Cupressus* (the branch leading to (*C. gigantea, C. duclouxiana*)), at the basal position of the HCX clade, and among the three outgroups (Fig. 3B).

**Molecular dating**

The BEAST analysis based on the two phylogenetic topologies, HCX(Cu,Ju) and Ju(Cu,HCX), yielded effective sample sizes that were well above 200 (> 800) for branch lengths, topology and clade posteriors and all other relevant parameters, indicating adequate sampling of the posterior distribution. However, the BEAST analysis based on the Cu(HCX,Ju) topology failed, despite being identical to other analyses in all but enforcement of topology, because every one of >20 attempts returned an error message that the log likelihood of the initial tree is negative Infinity.

Based on the HCX(Cu,Ju) topology that was supported by single-copy nuclear (SCN) genes, we estimate that the most recent common ancestor (MRCA) of *Cupressus, Juniperus* and the HCX clade diverged from the MRCA of *Platycladus, Microbiota, Calocedrus* and *Tetraclinis* 81.06 Mya [70.50-90.40] (from here on, shown in square brackets are the 95% HPD range of age estimation), the HCX clade diverged from the MRCA of *Cupressus* and *Juniperus* 59.80 Mya [48.45-71.74], and *Juniperus* diverged from *Cupressus* 56.33 [45.30-67.95] Mya. The crown ages of the HCX clade, *Cupressus* and *Juniperus* were estimated to be 37.45 Mya [23.54-52.30], 28.73 Mya [16.65-42.15], and 41.34 Mya [29.99-44.63], respectively (Fig. 4, Table 3).

Based on the Ju(Cu,HCX) topology that was supported by ptDNA tree, age estimation for all nodes overlapped with the above estimation (see Table 3) except that the HCX clade diverged from *Cupressus* 54.09 Mya (95% HPD: 41.29-67.03). A comparison of age estimation of major nodes in BEAST analyses based on each of the above two topologies are shown in Table 3.
**Ancestral area reconstruction**

Model tests in the BioGeoBEARS package, based on either the HCX(Cu,Ju) or the Ju(Cu,HCX) topology, suggested that DIVALIKE+J is the best-performing model (AICc_wt values: HCX(Cu,Ju) topology = 0.65, Ju(Cu,HCX) topology = 0.60), whereas DEC+J model is the second-best model (AICc_wt values: HCX(Cu,Ju) topology = 0.33, Ju(Cu,HCX) topology = 0.39).

Based on the HCX(Cu,Ju) topology, the DIVALIKE+J model and the 86-accession data set, *Cupressus*, *Juniperus* and the HCX clade share a common ancestor whose ancestral distribution area is probably Asia (ca. 0.96), whereas *Cupressus* and *Juniperus* shared a common ancestor whose ancestral distribution area is likely to be Asia (ca. 0.82) or less likely Europe (ca. 0.16). The ancestral area for the MRCA of the HCX clade is inferred to be Asia (ca. 0.54), North America (ca. 0.33) or a combination of these two (ca. 0.13). Within this clade, the common ancestor of all New World cypresses (*Callitropsis* plus *Hesperocyparis*) most likely migrated to and diversified in North America later (Fig. 5). The ancestral area for *Cupressus* is probably Asia (ca. 0.99), and *Cupressus semperivens* and its close allies dispersed to Europe around the middle Miocene (Fig. 5). Furthermore, the ancestral area for *Juniperus* is inferred to be Europe (ca. 0.72), or possibly Asia (ca. 0.23). The common ancestor of sect. *Juniperus* was inferred to be in Europe (ca. 0.56), Asia (ca. 0.36) or a combination of both, whereas that of sect. *Sabina* was probably in Asia (ca. 0.65) and possibly in Europe (ca. 0.25) or Africa (ca. 0.05); overall, *Juniperus* is most likely to have diversified within Eurasia, with three separate dispersal events to North America, and one to Africa. BioGeoBEARS analysis based on the DEC+J model yielded highly similar results (not shown), especially concerning major nodes in the phylogeny.

Based on the Ju(Cu,HCX) topology, either the DIVALIKE+J or the DEC+J model and the 86-accession data set, the ancestral area for nearly all nodes are highly similar to the HCX(Cu,Ju) topology, except for the node of the common ancestor of *Cupressus* and the HCX clade, which
does not exist in the HCX(Cu,Ju) topology. The ancestral area for this node in the Ju(Cu,HCX) topology is likely to be Asia (DIVALIKE+J: ca. 0.97; DEC+J: 0.95) (Supplementary Fig. S4).

**DISCUSSION**

**Rapid evolutionary divergence and inference of phylogenetic relationships among the three major clades in Cupressoideae**

The main aim of this paper is to resolve and explain the long-standing controversy of generic and inter-generic relationships between the three major lineages in Cupressoideae, *Cupressus*, the *Hesperocyparis-Callitropsis-Xanthocyparis* (HCX) clade, and *Juniperus*. Our phylogenetic analyses using maximum parsimony (MP), maximum likelihood (ML), Bayesian inference (BI) analyses of concatenated data and species tree analyses (MP-EST, STAR, and ASTRAL), based on 73 putative single copy nuclear genes totaling more than 200,000 base pairs, all show a maximally supported sister relationship of *Cupressus* and *Juniperus*, and that their common ancestor is sister to the HCX clade (Figs 1, 2). Although only weakly or moderately supported, the Ju(Cu,HCX) topology based on ptDNA (Supplementary Fig. S2; Mao *et al.*, 2010) conflicts with the HCX(Cu,Ju) topology here, as well as several published phylogenies (Xiang & Li, 2005; Little *et al.*, 2006; Adams *et al.*, 2009; Yang *et al.*, 2012; Terry & Adams, 2015). This incongruence may have been caused by incomplete lineage sorting due to rapid evolutionary divergence and/or hybridization and introgression between the three clades during their early evolutionary history. Yang *et al.* (2012) constructed a reticulate network using two nuclear loci, and because relationships of the three subclades were incongruent among different datasets, suggested that *Cupressus* “might have originated through hybridization between *Juniperus* and the ancestor of *Hesperocyparis–Callitropsis–Xanthocyparis*” (Yang *et al.*, 2012; p462). However, although hybridization and introgression during earlier history is a possibility, the main cause of the phylogenetic pattern among the three clades appears to be a combination of rapid evolutionary divergence and incomplete lineage sorting.
First, all phylogenetic analyses we conducted based on 73 loci support the HCX(Cu,Ju) topology. As we have shown above, the species tree constructed using MP-EST, STAR, ASTRAL or trees built using MP, ML, BI based on concatenated data show 100% support for the HCX(Cu,Ju) topology. The species tree estimate from MulRF also supports the HCX(Cu,Ju) topology: in particular, the RF distance between all gene trees to the HCX(Cu,Ju) topology is closer than to either the Ju(Cu,HCX) topology or the Cu(HCX,Ju) topology. In addition, Neighbor-Net tree based on concatenated dataset also support the HCX(Cu,Ju) topology with 100% bootstrap support (Fig. 3A), and the “reticulate” pattern among the three clades that Yang et al. (2012) reported was not detected (Fig. 3).

Second, the gene tree topology frequency we found here may fit better with incomplete lineage sorting as an explanation of conflicting gene trees. The maximum support value for nodes in the species tree does not necessarily mean that there is no conflict between the 73 gene trees. In our ASTRAL analyses, for example, the HCX(Cu,Ju), Ju(Cu,HCX) and Cu(HCX,Ju) topologies, received quartet support values of 54.16%, 24.24% and 21.60%, respectively (equivalent to 39.54, 17.69, 15.77 gene trees, respectively). We further checked the MLBS tree for each of the 73 genes and found that 38, 15, 14 gene trees support the above three topologies, respectively; if only MLBS values above 70% are considered (corresponding to a moderately well supported branch), then 31, 11 and 7 gene trees supported the three topologies, respectively. If conflict between gene trees is caused by incomplete lineage sorting, which is always a close companion of rapid evolutionary divergence (e.g., Maddison, 1997), then we would expect one high frequency topology and two lower frequency topologies. Conversely, if the conflict between gene trees is caused by hybridization and introgression (e.g. the hypothesis that Cupressus is a hybrid clade between Juniperus and HCX that Yang et al. (2012) have put forward), one might expect two major (and equivalent) frequency gene tree topologies (e.g., if the branch was a result of hybrid speciation) or some other set of frequencies (e.g., if a subset of the genome introgressed at this point). To conclude, the pattern of gene tree topology frequencies we found
above is more consistent with the scenario of incomplete lineage sorting than the hybridization and introgression scenario.

Third, the internode branch lengths between the three clades are consistently short in both our species tree (Fig. 2) and trees based on concatenated data (Fig. 1), and the molecular dating suggest that the interval between the MRCA of *HCX-Juniperus-Cupressus* (~59.8 Mya) and the MRCA of *Juniperus-Cupressus* (~56.3 Mya) is also relatively short (~3.5 Myr), consistent with rapid evolutionary divergence (and presumably a substantial chance of retaining some conflicting ancestral polymorphisms, as documented for our individual gene trees, Fig. S3). This has also been the case in previous phylogenies of Cupressoideae. For example, using the whole plastid genome the inferred internode branch length is short, regardless of whether the *HCX(Cu,Ju)* or *Ju(Cu,HCX)* topology is recovered (Qu *et al.*, 2017), and based on six ptDNA regions the phylogenetic relationship of these three clades remained unresolved (Mao *et al.*, 2012). However, there is one exception to the pattern, which is that the internode branch length between the MRCA of the three clades and the MRCA of *Cupressus* and *Juniperus* is relatively long based on the nuclear gene NEEDLY (Yang *et al.*, 2012).

Thus, although we cannot exclude reticulate evolution in shaping the current phylogenetic pattern of the three clades within Cupressoideae, rapid evolutionary divergence better explains the pattern we found. This inference is different from another case in this subfamily where reticulate evolution is clearly indicated among *Thuja* species (Peng and Dan, 2008).

**Transcriptomic data provide strong support for a four-genus taxonomic treatment in *Cupressus* s.l.**

Previous studies suggested four possible phylogenetic topologies concerning these three clades. Phylogenetic analyses based on either three or six ptDNA markers show that these three clades are part of a trichotomy (Little *et al.*, 2006; Mao *et al.*, 2012), whereas nine ptDNA markers provide moderate support for the *Ju(Cu,HCX)* topology (Mao *et al.*, 2010). A recent study based
on whole plastid genomes supported the clustering of *Juniperus* and *Cupressus*, while a filtered dataset, which was meant to reduce or elucidate long branch artifacts, supported the clustering of *Cupressus* and the HCX clade (Qu *et al*., 2017). The Cu(HCX,Ju) topology was supported by a series of studies: based on nrITS region alone (Xiang & Li, 2005), a combined dataset that included nrITS, two ptDNA markers and 56 morphological characters (Little *et al*., 2004); a combined dataset that included one ptDNA region and three nuclear regions (nrITS, ABI3, 4CL) (Adams *et al*., 2009); and a combined dataset that included 11 ptDNA regions and two nuclear regions (nrITS and NEEDLY) (Terry & Adams *et al*., 2015). Phylogenetic analyses based on a single nuclear region, NEEDLY (MP support value: 100%), and a combined dataset that included three ptDNA regions, two nuclear regions (ITS, NEEDLY) and 88 organismal characters (MP support value: 100%; Little, 2006) supported a HCX(Cu,Ju) topology, in agreement with our SCN results (Fig. 1).

One important implication of our results is that *Cupressus* s.l. is paraphyletic, and should be divided into four genera (see Fig. 1 and Table 1 for a clarification of taxon names). Nearly all published molecular phylogenetic analyses support the monophyly of both *Cupressus* s.s. and the HCX clade, yet the sister relationship between them is rarely supported (Mao *et al*., 2010). Hence, Little (2006) proposed to call the HCX clade *Callitropsis* s.l., where *Xanthocyparis* s.l. was merged into *Callitropsis* s.l., yet such a treatment is not universally accepted. Considering a proposal of Mill and Farjon (2006) to conserve the genus name *Xanthocyparis*, which was ratified by the International Botanical Congress in 2011 (Barrie, 2011), and that *Xanthocyparis* s.l. is not monophyletic, Adams *et al*. (2009) proposed to place all New World cypresses (*Cupressus* sensu Farjon species in North America) in the new genus *Hesperocyparis* and keep both *Xanthocyparis* s.s. and *Callitropsis* s.s. as monotypic genera. Our results support this, showing that both *Cupressus* s.l. and *Xanthocyparis* s.l. are paraphyletic, while each of *Cupressus*, the HCX clade, and *Hesperocyparis* is monophyletic. Hence, our nuclear-based results strongly support the division of *Cupressus* s.l. into four genera: *Cupressus*, *Hesperocyparis*, *Xanthocyparis* s.s. and *Callitropsis* s.s. (Adams *et al*., 2009; Mao *et al*., 2010)
and rejects both the combination of these four genera under *Cupressus* s.l. (Christenhusz et al., 2011; Table 1) and the combination of *Xanthocyparis* s.s. and *Callitropsis* s.s. under either *Xanthocyparis* s.l. or *Callitropsis* sensu Little (2004). Our data, as well as many others (e.g. Little, 2006; Mao et al., 2010), would also be consistent with combining *Xanthocyparis* s.s. and *Callitropsis* s.s. under either *Xanthocyparis* s.l. or *Callitropsis* sensu Little (2004). Our data, as well as many others (e.g. Little, 2006; Mao et al., 2010), would also be consistent with combining *Xanthocyparis* s.s. and *Callitropsis* s.s. under either *Xanthocyparis* s.l. or *Callitropsis* sensu Little (2004).

**An updated evolutionary divergence timescale and biogeographic history of *Cupressus*, the HCX clade and *Juniperus***

Rerunning the molecular dating on the ptDNA dataset from Mao et al. (2010) while constraining it with the nuclear species tree (HCX(Cu,Ju)) topology of our results, suggests that the split between the *Cupressus-Juniperus* clade and the HCX clade occurred (48.45-) 59.80 (71.74) Mya, with a split of the former clade into *Cupressus* and *Juniperus* happening only 3.47 Myr later, (45.30-) 56.33 (-67.95) Mya. Comparing this to the Ju(Cu,HCX) topology that was supported by ptDNA data (Mao et al., 2010), the only difference is that *Juniperus* diverges first (47.54-) 59.44 (-71.24) Mya, followed by the divergence of *Cupressus* from HCX (41.29-) 54.09 (-67.03) Mya, in the Ju(Cu,HCX) topology. All other nodes occur in both topologies and differ in age between topologies by no more than 1.04 Myr, a difference dwarfed by HPD error ranges (Fig. 4, Table 3). This indicates that, in our case, a single topological difference, even in the deep nodes in a phylogeny, had very limited effect on node age estimates. A possible reason for this may be that this particular topological difference did not alter the phylogenetic position of fossil calibration points, and barely affected the total length between any given node and the root of the tree (Sauquet et al., 2012; Wang & Mao, 2016).

We also reran the ancestral area reconstruction analyses for both the HCX(Cu,Ju) and Ju(Cu,HCX) topologies using BioGeoBEARS, and four parallel analyses were conducted for each of the two topologies based on two different models (DIVALIKE+J, DEC+J). Apart from the MRCA of *Cupressus* and *Juniperus*, and the MRCA of *Cupressus* and the HCX clade, that
are specific to the HCX(Cu,Ju) and Ju(Cu,HCX) topologies, respectively, the relative probability
of the ancestral area for all other nodes in all four parallel analyses are highly similar. We
therefore discuss the reconstructed biogeographic history of the HCX clade, *Cupressus* and
*Juniperus* based on the HCX(Cu,Ju) topology and the best model (DIVALIKE+J model). Our
ancestral area reconstruction (AAR) analysis inferred that both the MRCA of the HCX clade,
*Cupressus* and *Juniperus*, and the MRCA of *Cupressus* and *Juniperus*, most likely originated in
Asia. Likewise, the HCX clade most likely originated in Asia and then dispersed once to North
America and diversified there (Fig. 5). This fits a pattern of directional migration from the
northwest to the southeast in North America in New World Cypresses (*Callitropsis* and
*Hesperocyparis*), which may have been caused by climate cooling and aridification in the latter
half of the Cenozoic (Terry et al., 2016). *Cupressus* probably originated in Asia, and then
dispersed to Europe (and northern Africa) around the middle Miocene (Fig. 5). The genus
*Juniperus* and sect. *Juniperus* most likely originated in Europe, whereas sect. *Sabina* originated
in Asia; three independent migrations from Eurasia to North America and one migration from
Eurasia to Africa were inferred (Fig. 5).

Comparing these results to the previous AAR analysis based on S-DIVA (Mao et al., 2010),
the AAR analysis based on BioGeoBEARS yielded a clearer resolution, especially concerning
the ancestral area of the MRCA of the HCX clade, the MRCA of *Juniperus*, the MRCA of
*Juniperus* sect. *Juniperus* and sect. *Caryocedrus*, the MRCA of *Juniperus* sect. *Juniperus*, and
the MRCA of Clade I (*Juniperus pseudosabina* plus all Himalayan/Qinghai-Tibet Plateau
species except *J. microsperma* and *J. gausenii*) and Clade II (serrate-leaved junipers of North
America) (Mao et al., 2010). BioGeoBEARS tends to infer a single area as the ancestral area,
whereas the S-DIVA usually infers the combination of two disjunct areas as the ancestral area.
The integration of dispersal probability among areas during different time periods in the past,
and the use of a model test to seek the best-performing model are likely to have improved the
resolution of AAR in BioGeoBEARS compared to S-DIVA.
Conclusion

Phylogenetic relationships among Cupressus, Hesperocyparis-Callitropsis-Xanthocyparis (HCX) and Juniperus have been a contentious issue since the discovery of the Golden Vietnamese Cypress, Xanthocyparis vietnamensis. Our species tree based on 73 nuclear loci yielded 100% support for a (HCX, (Cupressus, Juniperus)) topology which is in agreement with previous phylogenies based on two nuclear loci (LEAFY and NEEDLY; Yang et al., 2012) and a combined dataset including both morphological characters and molecular dataset (Little, 2006), but contradicts many others. This indicates that Cupressus s.l. (Christenhusz et al., 2011; Table 1) is paraphyletic, and can be considered instead as two monophyletic genera, Cupressus (s.s.) and Hesperocyparis, and two monotypic genera, Callitropsis (s.s.) and Xanthocyparis (s.s.). Rapid evolutionary divergence and incomplete lineage sorting may have been the major cause for the minor conflicts observed among gene trees. Molecular dating based on the nuclear species tree (HCX(Cu,Ju)) topology suggests that the three clades underwent two evolutionary splits in a time period as short as ca. 3.47 Myr. The split between Cupressus+Juniperus and the HCX occurred ca. 59.80 Mya (95%HPD: 48.45-71.74 Mya), and the split between Cupressus and Juniperus occurred ca. 56.33 Mya (95%HPD: 45.30-67.95 Mya). Ancestral area reconstruction analyses suggest that the MRCA of Juniperus probably occurred in Europe, whereas the MRCAs of HCX, Cupressus, Cupressus+Juniperus, and HCX+Cupressus+Juniperus all most likely occurred in Asia. Therefore, the common ancestor of these three clades most likely originated in Asia and then diversified and dispersed to Europe, North America and Africa. Our study shows that combining low copy nuclear genes collected using next generation sequencing and coalescent-based species tree estimation methods is a powerful approach that provides more refined phylogenetic estimates of deep nodes in conifer phylogeny that were controversial based on small datasets.
Acknowledgements

We are grateful to Dr. Jun Wen for her constructive suggestions in early thoughts of this work. This study was funded by the National Natural Science Foundation of China (grant nos. 31590821, 31622015, 31370261), Department of Science and Technology of Sichuan Province (grant 2015JQ0018) and Sichuan University. The Royal Botanic Garden Edinburgh is supported by the Scottish Government’s Rural and Environment Science and Analytical Services Division.

Author contributions

K.M., M.R., J.L. designed research; K.M., M.R., Y.M., S.G., J.L., P.T., R.I.M., and P.M.H. performed research; M.R., K.M., Y.M. analyzed data; K.M., M.R., S.G., P.T., and P.M.H. procured specimens; K.M., M.R., R.I.M. wrote the manuscript, all authors revised the manuscript, and K.M., M.R., S.G., R.I.M. finalized the manuscript.

References


Leache AD, Banbury BL, Linkem CW, de Oca ANM. 2016. Phylogenomics of a rapid radiation: is chromosomal evolution linked to increased diversification in north american spiny lizards (Genus *Sceloporus*)? *BMC Evolutionary Biology* 16: 63.


Qu XJ, Jin JJ, Chaw SM, Li DZ, Yi TS. 2017. Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of Cupressoideae (Cupressaceae). Scientific Reports 7: 41005.


Table 1. A brief summary of alternative taxonomic treatments of *Juniperus*, *Cupressus*, *Hesperocyparis*, *Callitropsis* and *Xanthocyparis* since the description of *Xanthocyparis vietnamensis* in 2002. Underlined taxa are either monophyletic or monotypic. The abbreviations in brackets after common names are in accordance with Fig. 1.

<table>
<thead>
<tr>
<th>(A)</th>
<th>Common names</th>
<th>Junipers</th>
<th>Old world cypresses (OWC)</th>
<th>New world cypresses (NWC)</th>
<th>Alaska cedar (A.)</th>
<th>Vietnamese golden cypress (V.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B)</td>
<td>This study; Adams <em>et al.</em> (2009); Mao <em>et al.</em> (2010, 2012)</td>
<td><em>Juniperus</em></td>
<td><em>Cupressus</em> (s.s.)</td>
<td><em>Hesperocyparis</em></td>
<td><em>Callitropsis</em> (s.s.)</td>
<td><em>Xanthocyparis</em> (s.s.)</td>
</tr>
<tr>
<td>(C)</td>
<td>Farjon <em>et al.</em> (2002); Farjon (2005)</td>
<td><em>Juniperus</em></td>
<td><em>Cupressus</em> sensu Farjon</td>
<td></td>
<td><em>Xanthocyparis</em> s. l.</td>
<td></td>
</tr>
<tr>
<td>(E)</td>
<td>Little (2006)</td>
<td>N/A</td>
<td><em>Cupressus</em> (s.s.)</td>
<td></td>
<td><em>Callitropsis</em> s.l.</td>
<td></td>
</tr>
<tr>
<td>(F)</td>
<td>Christenhusz <em>et al.</em> (2011)</td>
<td><em>Juniperus</em></td>
<td></td>
<td><em>Cupressus</em> s.l.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Accessions used for RNA extraction, transcriptome assembly and subsequent phylogenetic analyses. (RBGE) and (SZ) refer to material collected from the wild held at the Royal Botanic Garden Edinburgh and Sichuan University, respectively; (1kp) refers to transcriptome data downloaded from the ‘1000 plant project’ (http://www.onekp.com/samples/list.php) with vouchers held at the University of British Columbia (UBC).

<table>
<thead>
<tr>
<th>Species</th>
<th>Collecting number (Identifier)</th>
<th>Lat/Long</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Callitropsis nootkatensis</td>
<td>19941704B (RBGE)</td>
<td>49°24'N/123°11'W</td>
<td>Canada</td>
</tr>
<tr>
<td>Calocedrus decurrens</td>
<td>FRPM (1kp)</td>
<td>n/a</td>
<td>cultivated (UBC)</td>
</tr>
<tr>
<td>Cupressus duclouxiana</td>
<td>MSZ-49-01 (SZ)</td>
<td>27°00'N/100°14'E</td>
<td>China</td>
</tr>
<tr>
<td>C. funebris</td>
<td>Mao-CF (SZ)</td>
<td>n/a</td>
<td>cultivated (SZ)</td>
</tr>
<tr>
<td>Calocedrus gigantea</td>
<td>MSZ-24-03 (SZ)</td>
<td>29°00'N/93°14'E</td>
<td>China</td>
</tr>
<tr>
<td>Calocedrus sempervirens</td>
<td>19752308 (RBGE)</td>
<td>45°12'N/13°36'E</td>
<td>Croatia</td>
</tr>
<tr>
<td>Hesperocyparis arizonica</td>
<td>19921324*C (RBGE)</td>
<td>30°50'N/115°16'W</td>
<td>Mexico</td>
</tr>
<tr>
<td>H. bakeri</td>
<td>19851378*B (RBGE)</td>
<td>41°57'N/123°18'W</td>
<td>USA</td>
</tr>
<tr>
<td>H. macrocarpa</td>
<td>20090071 (RBGE)</td>
<td>36°31'N/121°56'W</td>
<td>USA</td>
</tr>
<tr>
<td>Juniperus drupacea</td>
<td>20100261 (RBGE)</td>
<td>37°55'N/36°34'E</td>
<td>Turkey</td>
</tr>
<tr>
<td>J. flaccida</td>
<td>19922158*C (RBGE)</td>
<td>25°17'N/100°26'W</td>
<td>Mexico</td>
</tr>
<tr>
<td>J. indica</td>
<td>19790193*A (RBGE)</td>
<td>27°13'N/88°02'E</td>
<td>India</td>
</tr>
<tr>
<td>J. microsperma</td>
<td>MSZ-11 (SZ)</td>
<td>29°37'N/96°20'E</td>
<td>China</td>
</tr>
<tr>
<td>J. oxycedrus</td>
<td>19921237A (RBGE)</td>
<td>37°54'N/2°52'W</td>
<td>Spain</td>
</tr>
<tr>
<td>J. phoenicea</td>
<td>19921233*A (RBGE)</td>
<td>37°54'N/2°52'W</td>
<td>Spain</td>
</tr>
<tr>
<td>J. procera</td>
<td>20080832*J (RBGE)</td>
<td>00°19'N/36°58'E</td>
<td>Kenya</td>
</tr>
<tr>
<td>J. scopulorum</td>
<td>20081601 (RBGE)</td>
<td>39°39'N/105°12'W</td>
<td>USA</td>
</tr>
<tr>
<td>Microbiota decussata</td>
<td>19881678*A (RBGE)</td>
<td>n/a</td>
<td>cultivated (RBGE)</td>
</tr>
<tr>
<td>M. decussata</td>
<td>XQSG (1kp)</td>
<td>n/a</td>
<td>cultivated (UBC)</td>
</tr>
<tr>
<td>Species</td>
<td>Location</td>
<td>Coordinates</td>
<td>Country</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Thuja plicata</td>
<td>VFYZ (1kp)</td>
<td>n/a</td>
<td>cultivated (UBC)</td>
</tr>
<tr>
<td>Xanthocyparis vietnamensis</td>
<td>20030523 (RBGE)</td>
<td>23°06'N/105°01'E</td>
<td>Vietnam</td>
</tr>
</tbody>
</table>
Table 3. Estimates for the divergence times for nodes within *Juniperus* and *Cupressus* (s.s.) and the *Hesperocyparis-Callitropsis-Xanthocyparis* clade (HCX), based on the ptDNA data set of Mao *et al.* (2010) using the constraint of the nuclear species tree topology from this study (HCX(Cu,Ju) topology) or without any constraint (i.e. ptDNA tree topology, Ju(Cu,HCX) topology) employing a relaxed molecular dating approach in BEAST.

<table>
<thead>
<tr>
<th>Node No.</th>
<th>Description of Node:</th>
<th>HCX(Cu,Ju) topology</th>
<th>Ju(Cu,HCX) topology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Stem of the MRCA of the three clades</td>
<td>81.06 (70.45-90.40)</td>
<td>80.96 (71.07-89.75)</td>
</tr>
<tr>
<td>2</td>
<td>Crown of the MRCA of the three clades</td>
<td>59.80 (48.45-71.74)</td>
<td>59.44 (47.54-71.24)</td>
</tr>
<tr>
<td>3</td>
<td>Split between <em>Cupressus</em> and <em>Juniperus</em></td>
<td>56.33 (45.30-67.95)</td>
<td>Equal to Node 2</td>
</tr>
<tr>
<td></td>
<td>Split between <em>Cupressus</em> and the HCX clade</td>
<td></td>
<td>54.09 (41.29-67.03)</td>
</tr>
<tr>
<td>5</td>
<td>Crown of <em>Cupressus</em></td>
<td>28.73 (11.65-42.15)</td>
<td>28.44 (16.57-42.03)</td>
</tr>
<tr>
<td>6</td>
<td>Crown of the HCX clade</td>
<td>37.45 (23.54-52.30)</td>
<td>36.41 (22.73-50.81)</td>
</tr>
<tr>
<td></td>
<td>Split between <em>Callitropsis</em> (s.s.) and <em>Hesperocyparis</em></td>
<td>32.30 (19.40-45.86)</td>
<td>31.58 (19.13-44.84)</td>
</tr>
<tr>
<td>8</td>
<td>Crown of genus <em>Juniperus</em></td>
<td>41.34 (33.90-49.45)</td>
<td>41.79 (33.91-50.80)</td>
</tr>
<tr>
<td>9</td>
<td>Split: sects. <em>Juniperus-Caryocedrus</em></td>
<td>33.80 (19.68-45.58)</td>
<td>34.12 (20.28-46.87)</td>
</tr>
<tr>
<td>10</td>
<td>Crown of sect. <em>Juniperus</em></td>
<td>17.20 (8.63-27.41)</td>
<td>17.16 (8.34-27.05)</td>
</tr>
<tr>
<td>11</td>
<td>Crown of sect. <em>Sabina</em></td>
<td>36.50 (29.99-44.53)</td>
<td>36.80 (29.49-44.86)</td>
</tr>
</tbody>
</table>
Figure 1. Bayesian tree based on 73 concatenated nuclear genes (208.484 bp). Numbers or asterisks above branches are statistical support values for maximum parsimony/maximum likelihood/Bayesian inference analyses, respectively, with * denoting maximum support in all three analyses. Colour and grey scale bars to the right of the cladogram illustrate (A) common names of all included taxa (OWC = Old World cypresses; NWC = New World cypresses; A. = Alaska cedar; V. = Vietnamese golden cypress; C. = Callitropsis; X. = Xanthocyparis; sl/s.l. = sensu lato; sL = sensu Little (2004)), taxonomic treatments adopted (B) in this study, Adams et al. (2009) and Mao et al. (2010, 2012), (C) by Farjon et al. (2002) and Farjon (2005), (D) by Little et al. (2004), (E) by Little (2006) and (F) by Christenhusz et al. (2011). Scale bar indicates the estimated number of mutations per site.

Figure 2. Species tree generated using ASTRAL-II based on 73 nuclear genes (208.484 bp). Numbers or asterisks above branches are branch support values for MP-EST, STAR and ASTRAL-II analyses, respectively, with asterisks denoting maximum support in all three analyses. ASTRAL-II measures branch lengths in coalescent units (scale bar shown corresponds to two coalescent unit) for internal branches and NOT terminal branches (branch lengths of terminal branches are therefore arbitrary and meaningless). The pie chart shows respective quartet support for the main topology, the first and the second alternative topology (as shown in the inset). Note that in the inset, the tree formulas in parentheses were presented in the sense of an unrooted quadripartition (four-taxon) tree, where the central piece of the tree is an internode branch between the two pairs of partitions. The first case shown ((Cupressus, Juniperus),(HCX, outgroups)) is consistent with HCX being sister to (Cupressus, Juniperus) in an outgroup-rooted tree.
**Figure 3.** Neighbour-Net networks based on 73 concatenated nuclear genes (208.484 bp) using SplitsTree. (A) A Neighbour-Net network with bootstrap support values, and (B) a consensus Neighbour-Net network using a 95% threshold based on 1,000 bootstrap replicates. In (A), numbers next to “branches” are bootstrap support values. Note that in both (A) and (B) the branches lead to the three outgroup taxa are truncated so as to show more details of ingroup relationships.

**Figure 4.** Evolutionary divergence timescale of Cupressoideae based on the ptDNA data set (86 accessions from Mao *et al.*, 2010) with the imposed constraint of the nuclear species tree (HCX(Cu,Ju)) topology from this study using BEAST. Blue bars represent 95% HPD (highest posterior density) for each node, and white triangular with black outline represent compressed clades. Letters in black circles represent fossil calibration points (corresponding to those in Table 1 in Mao *et al.*, 2010), and numbers in black squares indicate numbers for nodes of interest (see Table 3). ‘HCX’ stands for the *Hesperocyparis-Callitropsis-Xanthocyparis* clade.

**Figure 5.** Ancestral area reconstruction based on BEAST trees constrained using nuclear species tree (HCX(Cu,Ju)) topology and the DIVALIKE+J model in BioGeoBEARS, as implemented in RASP 4.0. The inset shows the division of the distribution area of the five genera into four operational areas (North America, Africa, Asia and Europe). The pie at each node represents the reconstructed ancestral area; different colours of circular sector in a pie represent the relative probabilities of different ancestral areas at a node.
Supporting Information

**Supplementary Table S1.** Detailed information for the 73 single/low copy nuclear genes used in phylogenetic analyses. The 'missing taxa' column highlights which taxa were not included due to missing data for a particular gene. Missing taxa in bold are from the ingroup; a list of acronyms and corresponding species names is given below the table.

**Supplementary Table S2.** NCBI GenBank accession numbers for the 73 genes of 20 taxa that were used in phylogenetic analyses.

**Supplementary Table S3.** A summary of the topology and the maximum likelihood (ML) bootstrap support value (MLBS) for each of the 73 nuclear genes used to construct the species tree. Different topologies showing the sister relationship between the three major clades, *Juniperus*, *Cupressus* and HCX, are shown. 'Other topology' refers to a situation where two or all of the three groups are polyphyletic and the topology does not fall into any of the first three categories. 'Polyphyletic group(s)' refers to a situation where one or more taxa are polyphyletic. 'MLBS value' refers to the ML bootstrap support value for the sister relationship between the indicated clade.

**Supplementary Figure S1.** Maximum parsimony (MP) tree based on 73 concatenated nuclear genes (208.484 bp). MP analysis resulted in one tree of 31,272 steps with CI = 0.86 and RI = 0.82. Scale bar indicates number of nucleotide changes. Numbers above or below branches are bootstrap support values based on 1000 bootstrap replicates.

**Supplementary Figure S2.** Phylogenetic relationships and posterior probability of major clades in Cupressoideae that were derived from BEAST using nine ptDNA fragments (86 accessions
from Mao et al., 2010) and eight fossil calibrations (a) without any constraint (i.e. ptDNA tree, Ju(Cu,HCX) topology) and (b) with a constraint using the nuclear species tree (HCX(Cu,Ju)) topology concerning the relationship of Cupressus, Juniperus and the HCX clade (only) that was obtained in this study (see Fig. 1). Triangle tips represent clades that comprising more than two species; for full list of species see Figs 5 and S4, respectively. Posterior probabilities of clades shown to the right or upper-left of each node.

**Supplementary Figure S3.** Maximum likelihood bootstrap tree for each of the 73 nuclear genes as constructed in RaxML based on 1000 bootstrap replicates. Gene names are shown in the top left of each subfigures. Numbers close to each node represent the bootstrap support value for the branch lead to a node.

**Supplementary Figure S4.** Ancestral area reconstruction based on ptDNA tree (Ju(Cu,HCX)) topology and the DIVALIKE+J model in BioGeoBEARS as implemented in RASP 4.0. The inset shows the division of the distribution area of the five genera into four operational areas, North America, Africa, Asia and Europe. The pie at each node represents the reconstructed ancestral area, different colours of circular sectors in a pie represent the relative probabilities of different ancestral areas at a node.