Single-Copy Genes as Molecular Markers for Phylogenomic Studies in Seed Plants

Zhen Li[1,2,3], Amanda R. De La Torre[4,5], Lieven Sterck[1,2,3], Francisco M. Cánovas[6], Concepción Avila[6], Irene Merino[7], José Antonio Cabezas[8], María Teresa Cervera[8], Pär K. Ingvarsson[4,7], and Yves Van de Peer[1,2,3,9,*]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[2]Center for Plant Systems Biology, VIB, Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent, Belgium

[4]Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[5]Department of Plant Sciences, University of California-Davis, Davis, CA

[6]Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus Universitario de Teatinos s/n, Málaga, Spain

[7]Department of Plant Biology, Uppsala BioCenter, Swedish University of Agricultural Sciences, Uppsala, Sweden

[8]INIA, Center Forest Research (CIFOR), Madrid, Spain

[9]Genomics Research Institute, University of Pretoria, Hatfield Campus, Pretoria, South Africa

*Corresponding author: E-mail: yves.vandepeer@psb.vib-ugent.be.

Accepted: April 26, 2017

Abstract

Phylogenetic relationships among seed plant taxa, especially within the gymnosperms, remain contested. In contrast to angiosperms, for which several genomic, transcriptomic and phylogenetic resources are available, there are few, if any, molecular markers that allow broad comparisons among gymnosperm species. With few gymnosperm genomes available, recently obtained transcriptomes in gymnosperms are a great addition to identifying single-copy gene families as molecular markers for phylogenomic analysis in seed plants. Taking advantage of an increasing number of available genomes and transcriptomes, we identified single-copy genes in a broad collection of seed plants and used these to infer phylogenetic relationships between major seed plant taxa. This study aims at extending the current phylogenetic toolkit for seed plants, assessing its ability for resolving seed plant phylogeny, and discussing potential factors affecting phylogenetic reconstruction. In total, we identified 3,072 single-copy genes in 31 gymnosperms and 2,156 single-copy genes in 34 angiosperms. All studied seed plants shared 1,469 single-copy genes, which are generally involved in functions like DNA metabolism, cell cycle, and photosynthesis. A selected set of 106 single-copy genes provided good resolution for the seed plant phylogeny except for gnetophytes. Although some of our analyses support a sister relationship between gnetophytes and other gymnosperms, phylogenetic trees from concatenated alignments without 3rd codon positions and amino acid alignments under the CAT + GTR model, support gnetophytes as a sister group to Pinaceae. Our phylogenomic analyses demonstrate that, in general, single-copy genes can uncover both recent and deep divergences of seed plant phylogeny.

Key words: single-copy genes, gymnosperms, angiosperms, seed plants, phylogenomics.

Introduction

Seed plants originated ∼370 Ma, and probably comprise 260,000 to 310,000 extant species (Fiz-Palacios et al. 2011; Christenhusz and Byng 2016). Current seed plants consist of angiosperms (flowering plants) and gymnosperms, the latter of which are further subdivided into Cycadidae, Ginkgoidae, Gnetidae, and Pinidae (Chase and Reveal 2009). Both morphological and molecular studies have clearly shown that angiosperms and gymnosperms are two monophyletic groups (Chaw et al. 2000; Wang and Ran 2014), but the relationship between the different clades in gymnosperms is less clear than in angiosperms (Haston et al. 2009), despite great efforts in resolving the phylogeny with diverse sets of molecular markers (Zhong et al. 2010; Lee et al. 2011; Xi et al. 2013; Lu et al. 2014). Particularly, the exact phylogenetic position of gnetophytes, a morphologically unique clade with accelerated molecular evolution rates, remains elusive (Wang and Ran 2014). Morphological studies, historically, agree that gnetophytes are a sister group of angiosperms (anthophyte hypothesis) (reviewed in Doyle 1998), because of obviously similar characteristics, such as, the existence of vessel elements and the simple, unisexual, flower-like reproductive organs. However, this hypothesis was later questioned on the basis of a flood of molecular data, with some providing support for gnetophytes as sister to the other seed plants (Gnetales—other seed plant hypothesis) (Burleigh and Mathews 2004) and others providing support for a sister group relationship with the other gymnosperms (Gnetales—other gymnosperms hypothesis) (Cibrián-Jaramillo et al. 2010; Lee et al. 2011). Still others provided support, usually based on mitochondrial or plastid genes, for gnetophytes as a sister group to conifers (Gnetifer hypothesis) (Ran et al. 2010), to one clade of conifers, that is cupressophytes (Gnecup hypothesis) (Xi et al. 2013; Lu et al. 2014), or to the other conifer clade, that is Pinaceae (Gnepine hypothesis) (Zhong et al. 2010, 2011; Wu et al. 2011; Burleigh et al. 2012). Also different approaches and data treatments yielded different phylogenetic placements of gnetophytes within the gymnosperms (Zhong et al. 2010, 2011; Wickett et al. 2014). Besides the controversial systematic position of gnetophytes, *Ginkgo*, which is a monotypic genus of an ancient lineage that originated at least 270 Ma, also has an ambiguous placement among the gymnosperms (Wang and Ran 2014). Some studies suggest *Ginkgo* as a sister group to a clade comprising conifers and gnetophytes (Mathews 2009; Ran et al. 2010; Lu et al. 2014); whereas several recent phylogenomic analyses support a sister relationship between *Ginkgo* and cycads (Cibrián-Jaramillo et al. 2010; Wu et al. 2013; Xi et al. 2013; Wickett et al. 2014).

Increased species sampling could help resolving the evolutionary relationships within seed plants (Zwickl and Hillis 2002), but molecular markers for gymnosperms are still lacking to allow broad comparisons between taxa (Cibrián-Jaramillo et al. 2010; Lu et al. 2014). Single-copy gene families, or single-copy genes, have long been recognized as ideal molecular markers for inferring relationships of previously unresolved lineages (Levin et al. 2009; Duarte et al. 2010; Salas-Leiva et al. 2014). Some characteristics, such as the uniqueness and high sequence conservation across species, allow single-copy genes to be straightforwardly amplified and sequenced. As nuclear genes, single-copy genes have biparental inheritance, unlike organelle genes that are mostly uniparentally inherited, so they may be better suited when dealing with hybridization, speciation, and incomplete lineage sorting of closely related species (Duarte et al. 2010; Zhang et al. 2012). The use of multiple unlinked nuclear single-copy genes is more likely to reflect true species relationships and may solve incongruences between organelle genes (Zhang et al. 2012; Lu et al. 2014; Zeng et al. 2014).

Although widely applied to angiosperms (Wu et al. 2006; Zhang et al. 2012; Zeng et al. 2014), only a few single-copy genes have been used to resolve gymnosperm relationships (Xi et al. 2013; Lu et al. 2014; Salas-Leiva et al. 2014). In addition, current single-copy genes in gymnosperms were identified on the basis of those in angiosperms (Salas-Leiva et al. 2014; Wickett et al. 2014). Whole genome sequencing can facilitate the identification of single-copy genes (De Smet et al. 2013; Li et al. 2016) but the huge genome sizes of gymnosperms (20–30 Gb) have greatly complicated their de novo sequencing (De La Torre et al. 2014). As a consequence, only a few gymnosperm species have been sequenced so far (Birol et al. 2013; Nystedt et al. 2013; Neale et al. 2014; Warren et al. 2015). However, since single-copy genes are often more broadly expressed and at higher levels than nonsingle-copy genes (De Smet et al. 2013; De La Torre et al. 2015), single-copy genes can be relatively easily detected by transcriptome sequencing, thereby simplifying the procedure to identify suitable molecular markers. In this study, using previously and newly developed genomic and transcriptomic data in 31 gymnosperms and 34 angiosperms, we identified single-copy gene families to increase the number of phylogenetic markers shared between gymnosperms (and between gymnosperms and angiosperms) that could be used for phylogenetic and comparative studies in seed plants (De La Torre et al. 2017).

Materials and Methods

Plant Material and cDNA Libraries Construction

Pinus pinaster seeds from the Oria provenance (Southern Spain) were germinated and grown at 20/24 °C with a 16/8 h photoperiod. Germinating seeds were watered twice a week with distilled water. One-month-old seedlings were used for cryosectioning and 0.5-cm tissue sections were processed for laser capture microdissection (Cañas et al. 2014). Tissues of *P. pinaster* were collected from cortex of hypocotyl, cortex of developing root, cortex of root, developing needle, mesophyll of cotyledon, mesophyll of new needle, pith hypocotyl, root apical meristem, shoot apical meristem, and vascular tissues of cotyledon, developing root, root, hypocotyl, and new needle. Pooled samples from needles, roots and stems from Galicia 1056xOria6 F1 progenies grown under different stress and hormone treatments were also included (supplementary table S1, Supplementary Material online). RNA isolation, cDNA synthesis, and construction of normalized cDNA libraries were performed following the protocol described by (Cañas et al. 2014).

Pinus sylvestris tissues represent different developmental stages during the development of zygotic embryogenesis. Zygotic embryos (E) and megagametophyte (M) samples were collected from immature cones and sorted separately into four different stages: early embryos (E1, M1), embryos at the stage of cleavage (E2, M2), dominant and subordinate embryos (E3DO, E3SU, M3) and dominant embryos before cotyledon differentiation (E4, M4) (supplementary table S1, Supplementary Material online). Total RNA was isolated by using the RNAqueous-Micro RNA isolation kit (Ambion) and its quality was verified by an Agilent 2100 BioAnalyzer System (Agilent Technologies) following manufacturer's instructions. Double-strand cDNA libraries were constructed by using the Mint-2 cDNA synthesis kit (Evrogen), followed by a reamplification step to incorporate the 454 pyrosequencing specific primers.

Transcriptome Sequencing and De Novo Assembly

Transcriptome sequencing was performed using the GS-FLX+ platform with a GS-FLX Titanium kit, Roche Applied Sciences (Indianapolis, IN) as described by (Cañas et al. 2014) (supplementary table S1, Supplementary Material online). We assembled transcriptomes of *P. pinaster* and *P. sylvestris* from the 454 sequencing reads using the Newbler software (v2.8.1). Before feeding reads to Newbler, we removed adapter sequences and reads shorter than 75 base pairs (bp) by SeqClean. Newbler then assembled all the remaining reads for *P. pinaster* and for *P. sylvestris*, until overrepresented sequences were removed. CD-HIT-EST (Li and Godzik 2006) then clustered the Newbler assemblies in each isogroup, which represents a unique transcriptional locus. In the end, we selected the longest transcript (at least 150 bp) as a unique representative for each isogroup.

In order to integrate public transcriptomes, we built an integration pipeline. SeqClean first screened the public data against the NCBI UniVec resource and retained transcripts longer than or equal to 150 bp. Next, public data was compared with the Newbler assemblies described above by CD-HIT-EST-2D (Li and Godzik 2006) to add novel transcripts to our assemblies. Finally, CD-HIT-EST (Li and Godzik 2006) selected a representative sequence from the clusters formed by the novel transcripts and the Newbler assemblies with 90% identity to remove redundant transcripts. For *P. pinaster*, we integrated 15,648 PlantGDB-assembled Unique Transcripts (PUTs, based on GenBank release 177) (Duvick et al. 2008) and 210,513 unigenes from SustainPineDB (Canales et al. 2013). For *P. sylvestris*, we integrated 73,609 PUTs (based on GenBank release 187) and a set of 2,261 EST assemblies. With respect to *Picea glauca* and *Picea sitchensis*, only public transcriptomes are available, so we carried out CD-HIT-EST with 90% identity to construct nonredundant transcripts from 48,315 PUTs (based on GenBank release 175) and 27,660 FL-cDNAs (Rigault et al. 2011) in *P. glauca* as well as 31,087 PUTs (based on GenBank release 183) and 13,197 EST assemblies in *P. sitchensis* (Ralph et al. 2008).

We used TransDecoder (r20131117) to predict open reading frames (ORFs) in the transcripts of *P. pinaster*, *P. sylvestris*, *P. glauca* and *P. sitchensis* based on training sets built from protein-coding genes in *Picea abies* (Nystedt et al. 2013) and *Pinus taeda* (Neale et al. 2014). We queried the transcripts from *P. pinaster*, *P. sylvestris*, *P. glauca* and *P. sitchensis* against the proteins from *P. abies* and *P. taeda* by BLASTX (Camacho et al. 2009). For each transcript, the complete ORF found within one High Scoring Pair was retained in the training sets. TransDecoder then used the training sets to build a Markov model and to predict ORFs with default parameters. Pfam (27.0) domains in the predicted ORFs were identified by HMMER embedded in TransDecoder.

Retrieval and Integration of Transcriptome Data from Public Databases

We retrieved transcriptome data from another 25 gymnosperms that were stored in PlantGDB (Duvick et al. 2008), oneKP (Wickett et al. 2014), and TreeGenes (https://dendrome.ucdavis.edu/treegenes/; last accessed April 25, 2017). These data are fragmented and redundant, as they have been generated by different technologies and experiments (supplementary table S2, Supplementary Material online). To obtain a nonredundant set of transcripts for each species, we used SeqClean to remove NCBI UniVec vectors and poly-As from the downloaded transcripts. MIRA4 assembled ESTs into longer transcripts unless PUTs were available (Chevreux et al. 2004). Next, we clustered transcripts in each species with 90% identity by feeding MIRA assemblies or PUTs, cDNAs, 454 assemblies, Transcriptome Shotgun Assemblies (TSAs), and oneKP assemblies to CD-HIT-EST (Li and Godzik 2006), which produced a set of nonredundant representative sequences which were then further assembled by CAP3 into unigenes (Huang and Madan 1999). TransDecoder (r20131117) was applied to predict ORFs in a self-training mode, which used the 500 longest ORFs to train a Markov model for coding sequences. For angiosperms, we downloaded protein-coding genes for 34 angiosperms, one moss, and two green algae from PLAZA 3.0 (Proost et al. 2015). Green algae (*Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*) and moss (*Physcomitrella patens*) were used as outgroups in this study.

Identification of Single-Copy Gene Families

We started with building gene families in six conifers, that is *P. pinaster*, *P. sylvestris*, *P. taeda*, *P. abies*, *P. glauca*, and *P. sitchensis*, because they, compared with other gymnosperms, have abundant genomic or transcriptomic data of outstanding quality. For instance, genes from *P. taeda* and *P. abies* were predicted based on genomes (Nystedt et al. 2013; Neale et al. 2014; Zimin et al. 2014) and transcript sequences in *P. glauca* and *P. sitchens* were supplemented with Sanger reads based on BACs (Ralph et al. 2008; Rigault et al. 2011), while, because of their economic importance, high-coverage transcript data were generated for *P. pinaster* and *P. sylvestris*

(European ProCoGen project; see www.procogen.eu (last accessed April 25, 2017) for more information). Applying OrthoMCL (Li et al. 2003) to these datasets, we obtained 32,017 multi-gene gene families comprised of 147,782 of the 259,547 input proteins (56.9%). To narrow down the search space for single-copy genes, we selected 11,152 gene families that were conserved throughout, and had low-copy number, in the six conifers. Furthermore, these gene families needed to be present in at least four of the six conifers and could have maximum two copies in two species.

To assign proteins from other species to the 11,152 gene families, we first used HMMER (v3.1b1) (Eddy 2011) to build an HMM profile for each of the gene families based on a multiple sequence alignment created by ClustalW (v2.1) (Larkin et al. 2007) using parameters for amino acids as recommended by (Hall 2011). For every species, additional proteins were retrieved using a profile search against the HMM profiles with HMMSCAN. For each HMM profile, hits with *E* values $<1\times10^{-10}$ were retained and their bit-scores were used to infer a cumulative probability distribution. The hits were assigned to a gene family accounting for 95% of the cumulative distribution (supplementary fig. S1*A*, Supplementary Material online) (Wickett et al. 2014). Since the above-described approach might fail to assign genes with similar sequences to the assigned hit at the 95% border, we further assigned those genes to a gene family if their *E* values were similar enough (ΔE value $< 1\times10^{20}$) to the hit with the smallest bit-score (supplementary fig. S1*B*, Supplementary Material online).

After assigning additional genes to the initial gene families, we selected gene families according to species occurrence, that is gene families had to be present in >20 (out of 31) gymnosperms and >30 (out of 37) species in PLAZA 3.0 (Proost et al. 2015). Afterwards, we removed gene families for which the single-copy percentage was <80%, which was defined as the fraction of species with exactly one copy in a gene family (Li et al. 2016). In the end, if more than five genes in a gene family were assigned to other gene families, we removed the gene family from further analysis. When fewer than five genes were assigned to other gene families, we reassigned these genes to the proper gene families according to the lowest *E* value. Species occurrence and single-copy percentage were double checked for the modified gene families.

Gen Ontology Enrichment Analysis

Gene Ontology Slim (GOSlim) enrichment analyses were carried out by BiNGO (3.03) with a threshold of 0.01 for *P* values, which were corrected for multiple testing by Benjamini and Hochberg False Discovery Rate (Maere et al. 2005). We used the *A. thaliana* annotation from TAIR (release 06/03/2016) and the *P. pinaster* annotation predicted by InterProScan (v5.15-54). GO terms for both species were mapped to GO slim plant by Map2Slim in OWLTools.

Selection of Phylogenetic Markers

To remove paralogs and to increase sequence sampling for phylogenetic analysis, we used the following procedure to find reciprocal best hits to select phylogenetic markers. Because HMMSCAN uses proteins to find matching HMM profiles and HMMSEARCH uses HMM profiles to find matching proteins, we carried out both of them sequentially. A pair of protein and HMM profile was considered as each other's reciprocal best hit if they were the best match to each other. From the 1,469 single-copy genes in seed plants, we finally retained 106 such gene families that were present in 36 out of 37 species from PLAZA 3.0 and 30 out of 31 gymnosperms species for multiple sequence alignment. We used Muscle (v3.8.31) to align amino acid sequences (Edgar 2004) followed by trimal (v1.4) to remove low-quality alignment regions in a heuristic mode ("-automated1") and to back-translate the amino acid alignments into nucleotide sequence alignments (Capella-Gutiérrez et al. 2009).

Phylogenetic Analyses

We employed different substitution models and partitioning strategies to reconstruct the phylogeny of seed plants. We built five sets of concatenated nucleotide sequence alignments: one with all codon positions (NT123); one with only the first two codon positions (NT12); and another three with each codon position separately (NT1, NT2, and NT3). For the NT123 alignment, we partitioned it as: 1) one partition; 2) two partitions with 1st and 2nd codon positions as the first part, and 3rd codon positions as the second one; 3) three partitions with each codon positions; 4) 52 partitions by PartitionFinder (v1.1.1) given different genes and codon positions (Lanfear et al. 2012). Similarly, the NT12 alignment was partitioned as: 1) one partition; 2) two partitions with 1st and 2nd codon positions; 3) 37 partitions by PartitionFinder given different genes and codon positions. RAxML (v8.2) was used to infer maximum likelihood (ML) trees based on the above-described concatenated alignments with different partitioning strategies under the GTR + GAMMA model (Stamatakis 2014). The best ML tree was searched from optimizing every 5th bootstrap tree in 200 rapid bootstraps.

For the corresponding amino acid alignment of NT123, we first used ProtTest3 to select the best-fit model according to the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) score and the corrected AIC (AICc) (Darriba et al. 2011). The JTT + I+GAMMA + F model outperformed all the other models and was used in RAxML to search the ML tree with 200 rapid bootstrap analysis. For Bayesian reconstruction, we carried out PhyloBayes-MPI with the CAT and CAT-GTR model and a discrete gamma distribution with four rate categories. We ran two independent chains under each model and considered the chains to be converged when the "maxdiff" parameter was <0.1 and the effective size >300 (Lartillot et al. 2009). Due to limitations of computational resources, especially for the CAT + GTR model, the original amino acid alignment was trimmed by trimal with "-gt 0.9 –cons 10", followed by removing invariant sites and sequences from the two green algae.

In addition to the DNA and amino acid model, we selected the Goldman and Yang (GY) model (Goldman and Yang 1994) among several available codon models for the NT123 alignment, with codon frequency estimated by ML implemented in CodonPhyML (v1.0) (Gil et al. 2013). The ratios of nonsynonymous to synonymous substitutions were drawn from a discrete gamma distribution with four rate categories. The ML tree was estimated from a BioNJ tree optimized by Nearest Neighbor Interchange and Subtree Pruning and Regrafting. Branch support values were represented by the SH-like approximate likelihood-ratio test (Guindon et al. 2010) instead of traditional bootstrap values.

Two recently developed coalescent methods, that is Species Tree estimation using Average Ranks of coalescence (STAR) (Liu et al. 2009) and Accurate Species Tree ALgorithm II (ASTRAL-II) (Mirarab and Warnow 2015), were used to infer the species phylogeny. For both coalescent analyses, we constructed a gene tree for each of the 106 phylogenetic markers by RAxML with the GTR + GAMMA model and 200 rapid bootstraps. To test the effects of 3rd codon positions, we built two sets of gene trees, one with (GT123) and the other without 3rd codon positions (GT12), for the coalescent analyses. Then the 106 gene trees were fed to STAR in an R package "phybase" (v1.4) and ASTRAL-II (v4.10.0) to infer the species phylogeny under the multi-species coalescent model. To obtain branch support, we used bootstrap values that were obtained by bootstrapping both gene loci and the sequence alignments with 100 replicates and reconstructed 100 coalescent species trees for both analyses.

Saturation of Substitutions and Approximate Unbiased Test

We determined an entropy-based index of substitution saturation (I_{SS}) for nucleotides using DAMBE5 for NT123, NT12, NT1, NT2, and NT3 alignments (Xia et al. 2003; Xia 2013). Two hundred replicates were performed with gaps treated as unknown states. Approximate Unbiased (AU) tests (Shimodaira 2002) were carried out by CONSEL (v0.20) (Shimodaira and Hasegawa 2001) on both the NT123 and NT12 alignments with partitions by each codon position and partitions from PartitionFinder. RAxML was carried out to calculate per site log-likelihood values based on the GTR + GAMMA model (Stamatakis 2014).

Measurement of Phylogenetic Incongruence

Internode Confidence (IC) and Internode Confidence All (ICA) were estimated by RAxML with the two sets of gene trees based on the 106 phylogenetic markers (Salichos and Rokas 2013; Salichos et al. 2014). The probabilistic and observed adjustment schemes were applied, because the gene trees contained both comprehensive and partial trees (Kobert et al. 2016). An IC or ICA value close to 1 means absence of conflicting bipartitions for a given internode, while a value close to zero suggests that incongruent bipartitions equally exist, and a value close to −1 indicates the lack of support for a given internode (Salichos et al. 2014). However, random gene trees always give (close-to) zero IC or ICA value due to the lack of phylogenetic information. To rule out possibility of the random effect, we simulated 1,000 random gene trees and compared the Robinson-Foulds distance between a species tree and the random gene trees, and the real gene trees, respectively. The gene trees of the 106 phylogenetic markers had significantly shorter Robinson-Foulds distances to the species tree than the random gene trees to the species tree (P value $< 2.2 \times 10^{-16}$, Wilcoxon rank sum test), indicating that any conflicting bipartition that exists in the real gene trees is from actual phylogenetic signal.

Table 1.
Transcriptome Assembly and Open Reading Frame (ORF) Predictions

Species	# Transcripts	# ORFs	# ORFs with Pfam Domains
Pinus pinaster	206,574	76,426	43,771 (57.3%)
Pinus sylvestris	121,938	36,106	22,355 (61.9%)
Picea glauca	39,229	28,909	19,708 (68.2%)
Picea sitchensis	28,030	20,434	13,989 (68.5%)

Results

Transcriptome Assembly and Data Integration

After assembly and removing redundant transcripts (see "Materials and Methods" section), we reconstructed 206,574 unigenes in *P. pinaster* and 121,938 unigenes in *P. sylvestris*, with an average length of 893 bp and 1,242 bp, respectively. For *P. glauca* and *P. sitchensis*, we integrated available public transcriptome data (see "Materials and Methods" section), which yielded 39,229 unigenes for *P. glauca* and 28,030 unigenes for *P. sitchensis*. TransDecoder predicted 20,434 to 76,426 ORFs in the four species with 57.3–68.5% of the ORFs having at least one Pfam domain (table 1). For *P. abies* and *P. taeda*, we collected 54,381 proteins and 43,959 proteins from the two published conifer genomes, respectively (Nystedt et al. 2013; Neale et al. 2014). Transcriptomes of another 25 gymnosperms were retrieved from public databases followed by removing redundant transcripts and predicting ORFs (supplementary table S2, Supplementary Material online).

Identification of Single-Copy Genes in Gymnosperms and Angiosperms

Using OrthoMCL (Li et al. 2003) and HMMER (Eddy 2011), we identified 3,072 single-copy genes in gymnosperms and

2,156 single-copy genes in angiosperms (see "Materials and Methods" section). Among these, 1,603 gene families were single-copy genes only found in gymnosperms, and 687 single-copy genes were specific to angiosperms. Additionally, 1,469 single-copy genes are shared between gymnosperms and angiosperms, so they are considered as the single-copy gene set representative for the seed plants.

Both missing data and whole genome duplications complicate the identification of single-copy genes. First, as single-copy genes are usually conserved genes present in all seed plants by definition, species with incomplete annotations hamper the identification of conserved gene families and thus single-copy genes. Second, recent whole genome duplications resulted in a burst of recent duplicates, which decreases the number of identified single-copy genes. To explore the effects of missing data and genome duplication on the delineation of single-copy gene families, we performed *k*-means bi-clustering on copy-number profiles of gymnosperms and angiosperms to cluster the species into two groups with similar profiles of copy numbers (fig. 1). Compared with angiosperms, we found that, in gymnosperms, the major factor affecting the identification of single-copy genes was missing data, as 10 of the 31 gymnosperms showed serious incompleteness of gene space in the copy number profile (fig. 1*A*). These ten species had fewer proteins than the rest of the gymnosperms (*P* value $= 3.78 \times 10^{-5}$, Wilcoxon rank sum test). In addition, for the 687 angiosperm specific single-copy genes, 586 of them were not conserved in gymnosperms according to our criterion (see "Materials and Methods" section), suggesting these conserved genes in angiosperms were either lost in some, if not all, gymnosperm lineages, or missed in their transcriptomes.

For the copy number profile of angiosperms, the *k*-means bi-clustering grouped species with recent whole genome duplications together, indicating that species that have undergone recent genome duplications still contain a large fraction of duplicated genes in the single-copy gene families (fig. 1*B*). For example, all seven species in the upper part of the copy-number profile, that is *Malus domestica*, *Glycine max*, *Brassica rapa*, *Gossypium raimondii*, *Populus trichocarpa, Eucalyptus grandis* and *Physcomitrella patens*, have undergone lineage-specific whole genome duplications (Tuskan et al. 2006; Rensing et al. 2008; Schmutz et al. 2010; Velasco et al. 2010; Wang et al. 2011, 2012; Myburg et al. 2014). On the contrary, the partial genome of *Lotus japonicus* and the small(er) proteome sizes of *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus* resulted in the absence of a large number of orthologous genes in these species.

Functional Enrichment of Single-Copy Genes

Single-copy genes are functionally biased toward certain conserved biological processes and organelle-related functions (Duarte et al. 2010; De Smet et al. 2013; Li et al. 2016). Since *A. thaliana* has been the most comprehensively annotated plant genome so far, we used *A. thaliana* genes to describe functions of single-copy genes for the angiosperms. GOSlim enrichment analysis revealed that the 2,156 single-copy gene families in angiosperms were often involved in photosynthesis, DNA metabolic processes, and cell cycle. Also, they were strikingly overrepresented in the plastid. On the other hand, single-copy genes of angiosperms were underrepresented in functional categories such as transcription factor activity, response to stimulus, and signal transduction (fig. 2). For the 3,072 single-copy gene families in gymnosperms, we used functionally annotated genes in *P. pinaster* to perform the GOSlim enrichment analysis, which, to some degree, suggested their similar functions as in angiosperms but with some exceptions, for example, lack of underrepresentation in response to stimulus, and extra overrepresentation in catabolic and lipid metabolic processes (fig. 2). We argue that the difference in the enrichment analyses between angiosperms and gymnosperms is largely due to the incompleteness of GOSlim annotations in *P. pinaster*, which only had 32,716 of the 76,426 (42.8%) genes that were annotated by at least one GOSlim term, whereas in *A. thaliana*, the percentage increased to 21,106 of 27,205 (77.6%) genes. A gene set with severely incomplete GO annotations could introduce systematic bias in the enrichment analysis. At last, the 1,469 single-copy gene families in seed plants were overrepresented or underrepresented in nearly identical functional categories as the ones in angiosperms, when using *A. thaliana* genes as representatives (fig. 2). The functions of single-copy genes in seed plants further confirm that these genes are involved in essential functions conserved across all seed plants and even throughout eukaryotes (Waterhouse et al. 2011; De Smet et al. 2013; Li et al. 2016).

Reconstructing Seed Plant Phylogeny

We used both tree construction based on concatenated sequence alignments and multi-species coalescent approaches to reconstruct the phylogeny of seed plants based on 106 phylogenetic markers selected from the 1,469 single-copy genes in seed plants (see "Materials and Methods" section). As 3rd codon positions have been known to affect the placement of gnetophytes (Wickett et al. 2014), we built two different concatenated nucleotide sequence alignments from the 106 genes, one with and one without 3rd codon positions, named "NT123" and "NT12", respectively. Species trees were then inferred from the two alignments under the GTR + GAMMA model with different partitioning strategies (see "Materials and Methods" section). All of the inferred phylogenetic trees support a monophyletic origin for both extant gymnosperms and angiosperms (100% bootstrap percentage, BP) (De La Torre-Bárcena et al. 2009; Lee et al. 2011; Xi et al. 2013; Wickett et al. 2014). The angiosperm

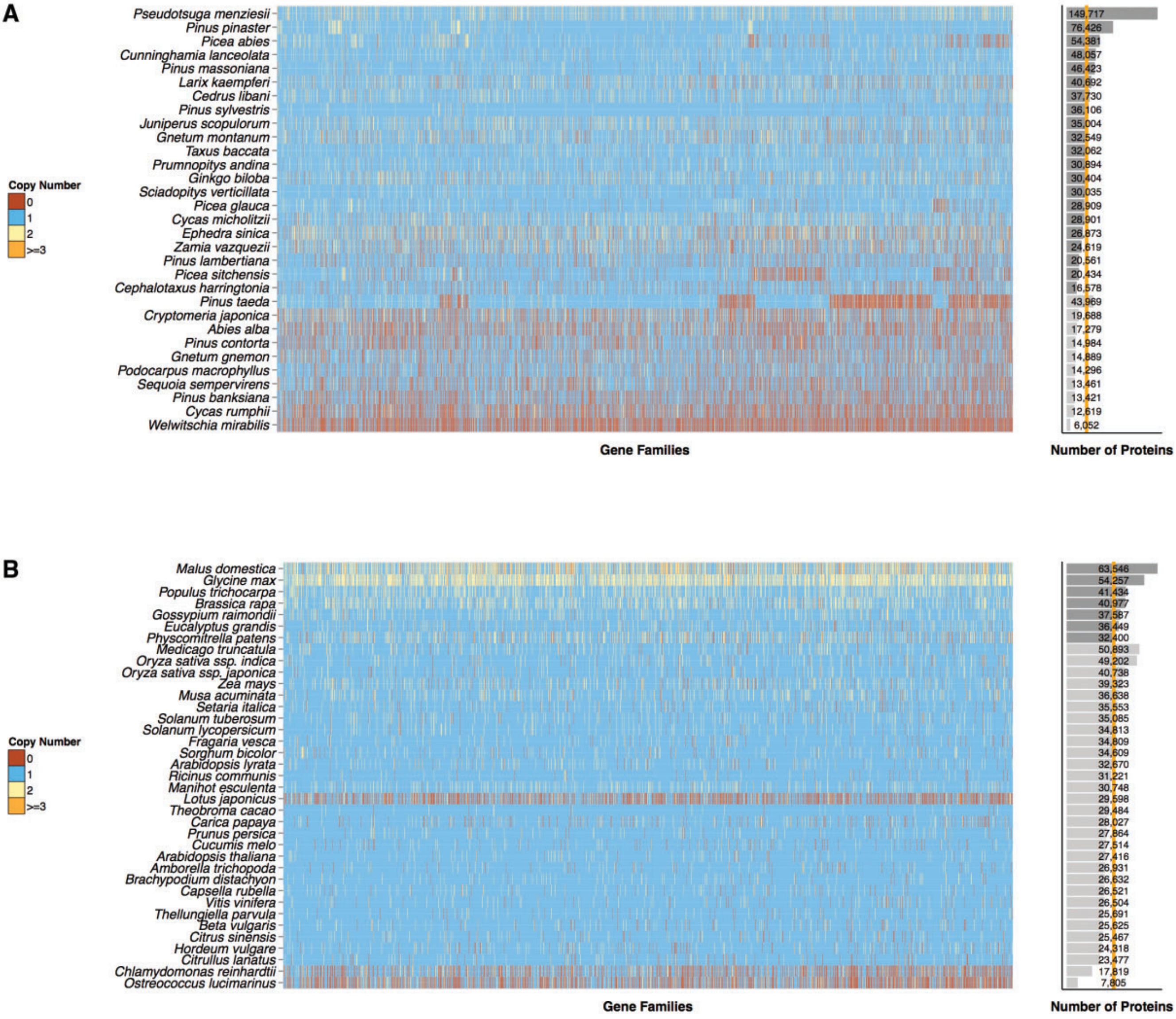

Fig. 1.—*k*-means bi-clustering of copy number profiles for single-copy genes in gymnosperms (*A*) and angiosperms (*B*). Rows represent species and columns represent gene families. In the copy number profiles, red denotes absence of genes in a gene family; blue denotes one copy; yellow denotes two copies; and orange denotes more than two copies in a gene family. The bar plot next to the copy number profile illustrates the number of proteins in each species with an orange line representing the average number of proteins. The dark and light gray bars distinguish the clusters identified by the *k*-means clustering.

phylogeny is largely congruent with the APGIII tree (Haston et al. 2009) with *Amborella* as a sister group to the monocots and dicots (figs. 3 and 4). The incongruence with respect to the position of the Malpighiales (i.e., *P. trichocarpa*, *Ricinus communis*, and *Manihot esculenta*) between our phylogeny and the APGIII tree has long been recognized (Zhu et al. 2007; Lee et al. 2011; Ruhfel et al. 2014). A hypothetical introgressive hybridization in the ancestral lineages of Fabidae and Malvidae has been proposed to explain a different ancestry of nuclear and chloroplast genes in extant Malpighiales (Sun et al. 2014).

For gymnosperms, the species trees inferred from NT123 and NT12 were largely similar except for some of the relationships within Pinaceae and cycads, and particularly the position of gnetophytes (figs. 3 and 4, and supplementary figs. S2–S6, Supplementary Material online). For Pinaceae, the only difference concerned the genus *Pinus*. The NT123 alignment clearly distinguished between the two subgenera of *Pinus*, that is subgenus *Strobus* (*Pinus lambertiana*) and subgenus *Pinus* (100% BP). The subgenus *Pinus* consists of the sections *Trifoliae* (i.e., *P. taeda*, *Pinus contorta*, and *Pinus banksiana*) and *Pinus* (i.e., *P. pinaster*, *P. sylvestris*, and *Pinus massoniana*)

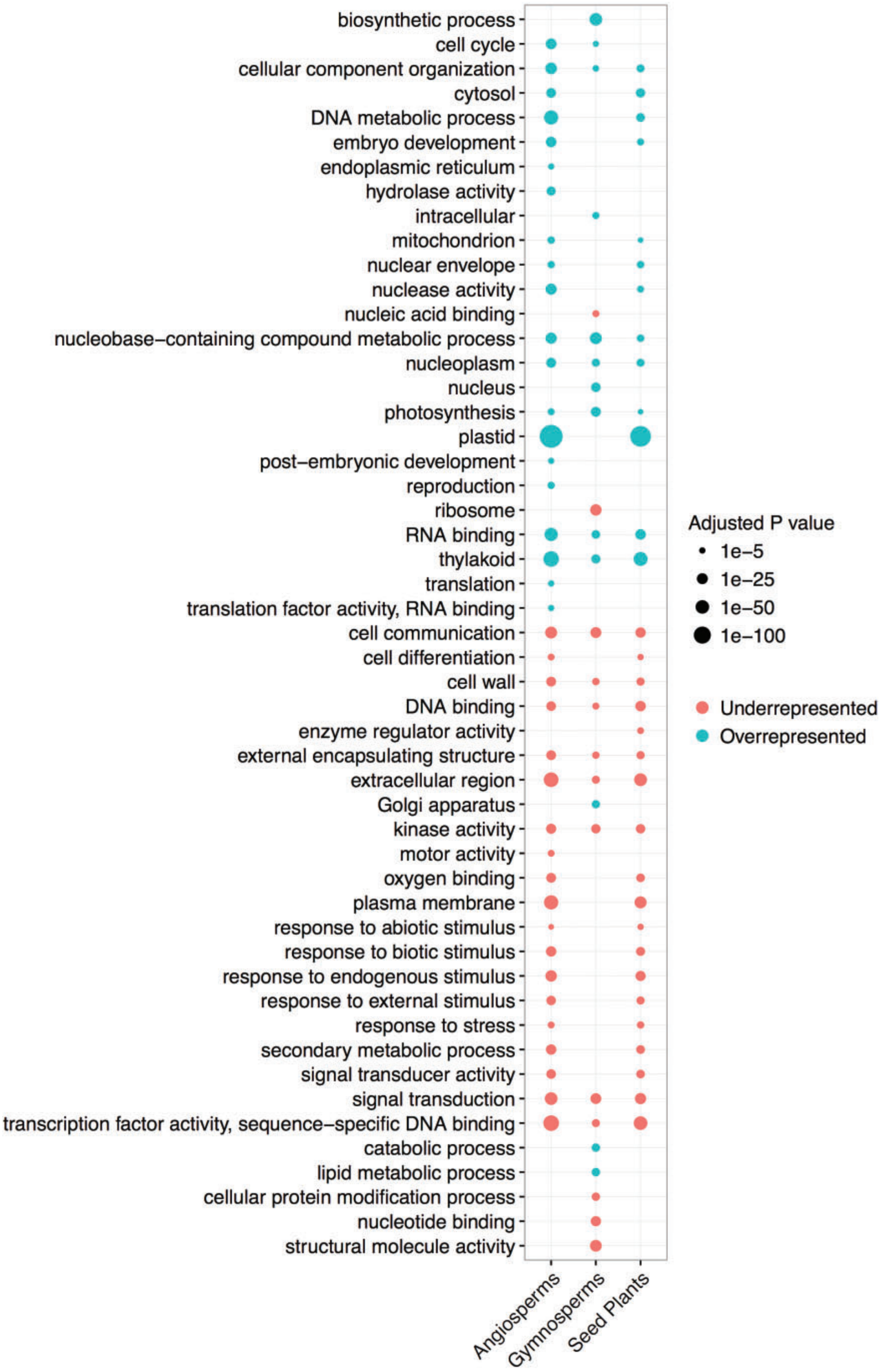

Fig. 2.—Gene ontology slim (GOSlim) enrichment analysis for single-copy genes in angiosperms, gymnosperms, and seed plants. Dot size is representative for the statistical significance of overrepresented (green) and underrepresented (red) GOSlim terms. *P* values were corrected for multiple tests by Benjamini and Hochberg False Discovery Rate.

(100% BP) as also observed in previous studies (Gernandt et al. 2005; Palmé et al. 2009). Trees inferred from the NT12 alignment had low bootstrap values for the genus *Pinus*, and incorrectly placed *Abies alba* (fig. 4), which was grouped with *Cedrus libani* as a sister to the other Pinaceae by the NT123 alignment (fig. 3), as expected based on morphological and molecular studies (Lin et al. 2010). Both alignments show *Larix* and *Pseudotsuga* to form a clade with *Pinus* and *Picea* as a sister clade.

For cupressophytes, all topologies suggest that *Podocarpaceae* diverged first, followed by *Sciadopityaceae*, and then *Taxaceae—Cephalotaxaceae* as a sister to *Cupressaceae*. For *Ginkgo*, our phylogenetic analyses suggest that it belongs to a sister group of cycads (100% BP), in accordance with recent phylogenomic analyses (Cibrián-Jaramillo et al. 2010; Wu et al. 2013; Xi et al. 2013), but in contrast to previous studies that support cycads as the sister lineage to the other gymnosperms (Mathews 2009; Ran et al. 2010; Lu et al. 2014).

The Phylogenetic Position of Gnetophytes

Regarding the phylogenetic position of gnetophytes, NT123 and NT12 alignments gave contradictory results. In all species trees based on the NT123 alignment (fig. 3 and supplementary figs. S2–S4, Supplementary Material online), gnetophytes were placed as a sister clade to the other gymnosperms (100% BP) in support of the "Gnetales—other gymnosperms" hypothesis. Species trees based on the NT12 alignment, however, clustered gnetophytes with Pinaceae thus supporting the "Gnepine" hypothesis ($\geqslant$73% BP, fig. 4 and supplementary figs. S5 and S6, Supplementary Material online). To obtain extra statistic support for the two alternative topologies instead of bootstrap values, we performed AU tests by CONSEL (Shimodaira 2002). Based on per site log likelihoods for the two topologies, the NT123 alignment significantly rejected the "Gnepine" topology (*P* value $= 2\times10^{-69}$ for three partitions by each codon position and *P* value $= 6\times10^{-36}$ for 52 partitions from PartitionFinder); notwithstanding, the NT12 alignment also rejected the "Gnetales-other gymnosperms" topology (*P* value = 0.014 for two partitions by each codon position and *P* value = 0.028 for 37 partitions from PartitionFinder). We further inferred the species phylogenies based on the concatenated alignments of each codon position, named "NT1", "NT2", and "NT3", to explore their contributions to the phylogenetic position of gnetophytes, independently. Interestingly, the NT3 alignment gave the same topology as the one based on the NT123 alignment and supported "Gnetales—other gymnosperms" hypothesis with 100% BP (supplementary fig. S7, Supplementary Material online). The NT1 and NT2 alignments both resulted in topologies similar to the ones obtained from the NT12 alignment by supporting the "Gnepine" hypothesis with 95% and 51% BP, respectively (supplementary figs. S8 and S9, Supplementary Material online). Our observations confirm that the inclusion of 3rd codon positions in the concatenated alignment indeed influences the phylogenetic position of gnetophytes in seed plant phylogeny as shown in previous phylogenomic studies (Wickett et al. 2014).

For nucleotide sequences of protein-coding genes, most sites from 3rd codon positions are synonymous sites due to codon degeneracy. It has been acknowledged that 3rd codon positions not only can contribute to phylogenetic signal (Seo and Kishino 2008; Ruhfel et al. 2014), but can also add noise to phylogenetic analysis because they quickly become saturated (Nei and Kumar 2000). This might lead to problems when using stationary time reversible models, especially when dealing with deep phylogenetic relationships (Wu et al. 2011; Zhong et al. 2011; Cooper 2014). Therefore, we further investigated base compositional heterogeneity and lineage specific changes of evolutionary rates on different codon positions in the five concatenated alignments of nucleotide sequences. The GC content of the 106 phylogenetic markers at different codon positions were dissimilar in different species, and in particular the 3rd codon positions were more variable compared with the 1st and the 2nd codon positions (supplementary fig. S10, Supplementary Material online). Pairwise comparisons of GC content among different species in the NT123, NT1, NT2, and NT3 alignments indicated that the NT123 and NT3 alignments exhibited significant compositional heterogeneity among different species ($P < 0.001$, Wilcoxon rank sum test with Bonferroni correction). The differences were most outspoken in two sets of groups, that is between the outgroup (two green algae and moss) and all seed plants, as well as between some angiosperms (especially Poaceae) and gymnosperms (fig. 5). However, significant differences in GC content in the NT1 and NT2 alignments almost only exist between the outgroup and seed plants. The pattern observed above still holds true after removing aligned codons that encode the same amino acids in the NT123 alignment (supplementary figs. S11 and S12, Supplementary Material online), suggesting that 3rd codon positions substantially contribute to the compositional heterogeneity in the NT123 alignment, while the base compositions of 1st and 2nd codon positions are in general very similar.

Disparate evolutionary rates of different sites among lineages, known as heterotachy, violate the assumption of one set of branch lengths for all sites in the homogeneous models (Wu et al. 2011; Zhong et al. 2011). Using the ML phylogenies inferred from NT1, NT2, and NT3, we measured branch lengths from the most recent common ancestor for each of the five monophyletic groups (i.e., angiosperms, gnetophytes, cycads and *Ginkgo*, cupressophytes, and Pinaceae) to every species in each group. As expected, the branch lengths were shorter for the trees inferred from 1st and 2nd codon positions than for the tree based on 3rd codon positions (fig. 6). An outspoken feature of the changes of branch lengths was

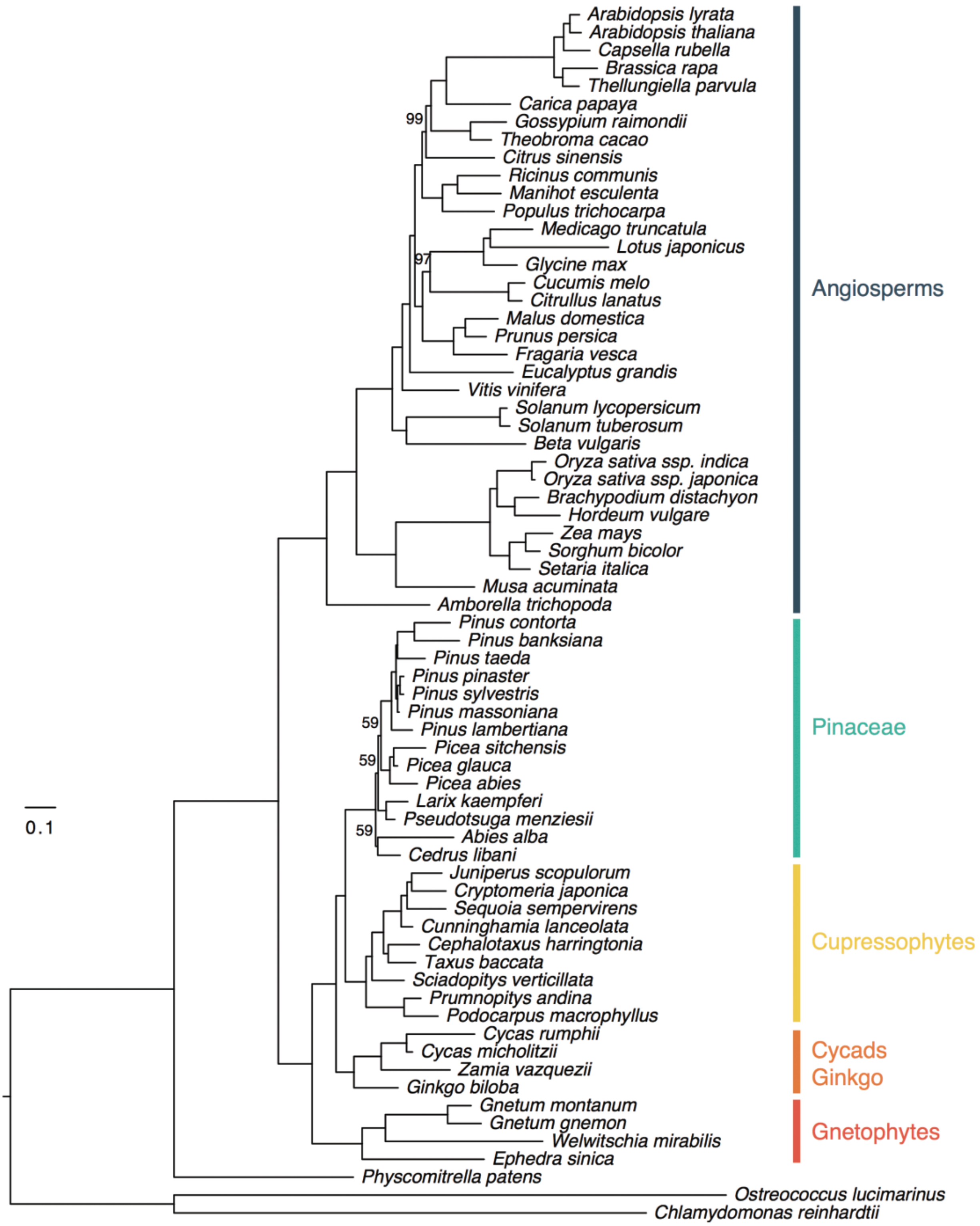

Fig. 3.—Maximum likelihood tree inferred from a concatenated alignment of 106 single-copy genes in seed plants including 3rd codon positions, partitioned by PartitionFinder. Bootstrap values <100% are shown on the specific branches. See supplementary figures S2–S4, Supplementary Material online, for maximum likelihood trees inferred from partitions based on codon positions.

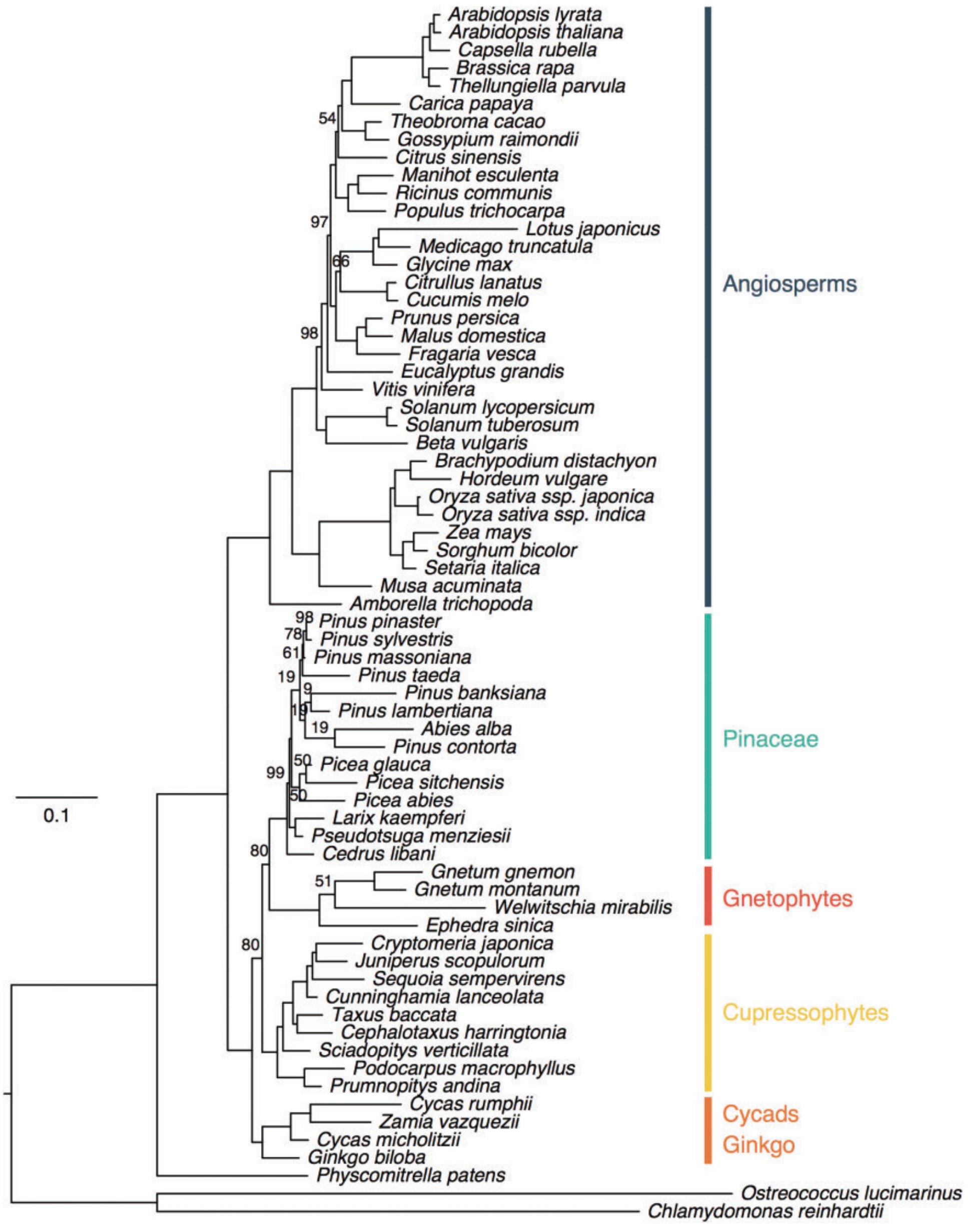

Fig. 4.—Maximum likelihood tree inferred from a concatenated alignment of 1st and 2nd codon positions for 106 single-copy genes in seed plants partitioned by PartitionFinder. Bootstrap values <100% are shown on the specific branches. See supplementary figures S5 and S6, Supplementary Material online, for maximum likelihood trees inferred from partitions based on codon positions.

their disproportional increase from 1st and 2nd codon positions to 3rd codon positions in the five lineages, from angiosperms as the fastest clade, followed by gnetophytes, cycads and *Ginkgo*, cupressophytes, to Pinaceae as the slowest. The drastic increase of branch lengths of the tree based on 3rd codon positions for angiosperms and gnetophytes, compared with the relatively stable alteration in Pinaceae, indicate distinctive various evolutionary rates among codon positions in the five clades, which is a characteristic signal of heterotachy.

The elevated evolutionary rates of 3rd codon positions might suggest substitution saturation, so we used I_{SS} to characterize substitution saturation in the nucleotide alignments. If I_{SS} is close to 1 or greater than a critical I_{SS} ($I_{SS.C}$), the alignment is considered to exhibit substantial saturation (Xia et al. 2003). Given its dependence on tree topologies, $I_{SS.C}$ is estimated under an extremely symmetrical ($I_{SS.C.Sym}$) as well as asymmetrical topology ($I_{SS.C.Asym}$). For the first two codon positions, either combined (NT12) or separate (NT1 and NT2), the I_{SS} values were significantly smaller than both $I_{SS.C,Sym}$ and $I_{SS.C,Asym}$ (P value $< 1 \times 10^{-4}$, two-tailed *t*-test, table 2), showing little evidence of substitution saturation on these sites. Nevertheless, for both alignments including

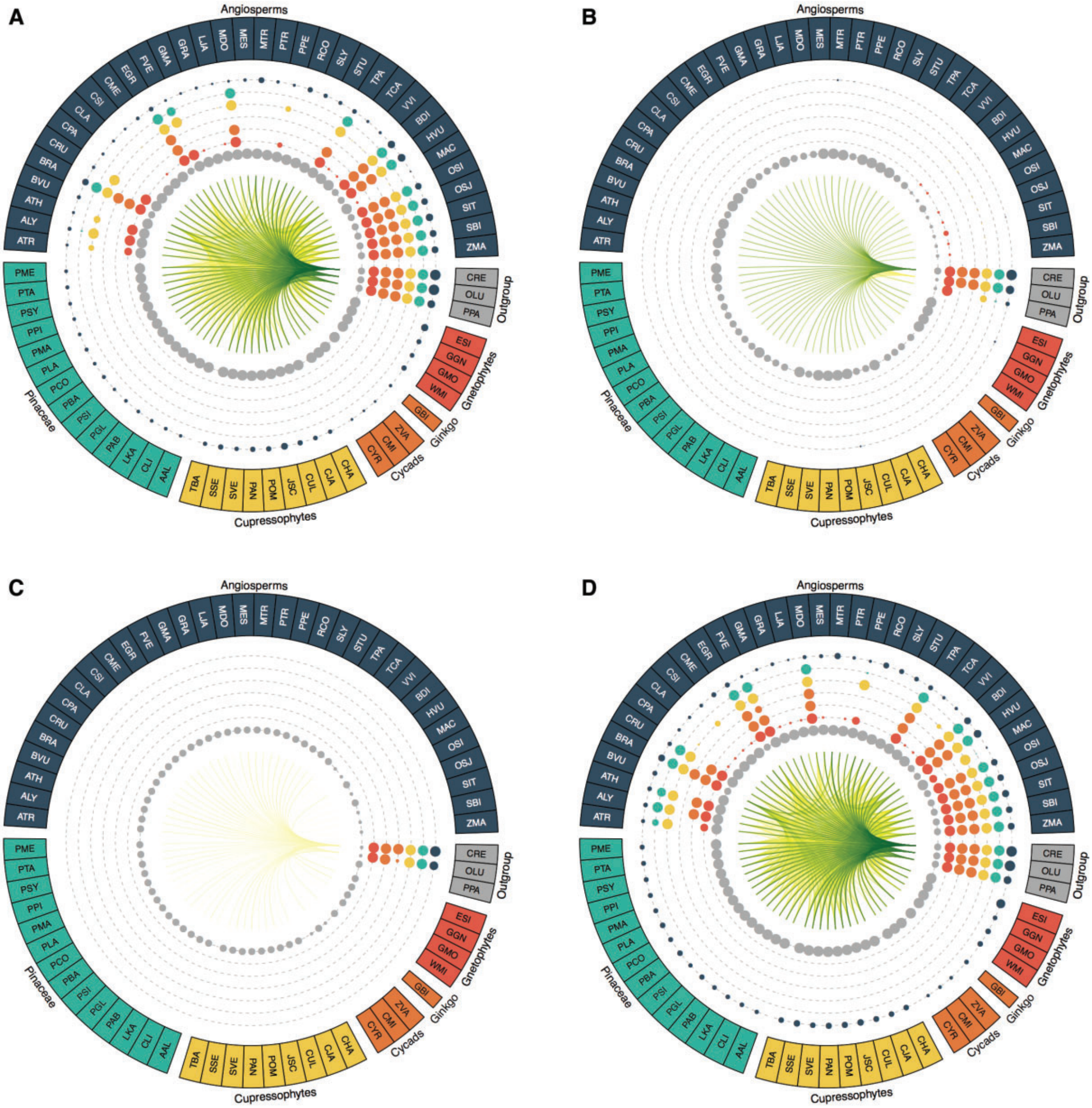

Fig. 5.—Comparison of GC content in the concatenated alignment (*A*) and at each codon position (*B*, *C*, and *D*) from 106 genes in 68 species. Dot size correlates with the number of species in each lineage (group) that have a significantly different GC% (Wilcox test, $P < 1 \times 10^{-3}$) with the species compared with (colors of dots correspond to the compared lineages). Lines connecting any two species represent significant difference in GC content, with most significant in green and weakest in yellow (1×10^{-3}). The full names for the species can be found in supplementary table S3, Supplementary Material online.

3rd codon positions (NT123 and NT3) I_{SS} were greater than $I_{SS.C,Asym}$ (P value $< 1 \times 10^{-4}$, two-tailed t-test, table 2), suggesting that sites from 3rd codon positions experienced substantially higher levels of substitution saturation than did sites from the 1st and 2nd codon positions. As values of I_{SS} for NT123 and NT3 were smaller than $I_{SS.C,Sym}$, they may be only useful when the real topology is extremely symmetrical, but the real topology of the sampled species in this study is somewhere in between a symmetrical and an asymmetrical tree.

The above results clearly illustrate that sites from the 3rd codon positions have features typically found in fast evolving sites, which are distinguishable from sites of the first two

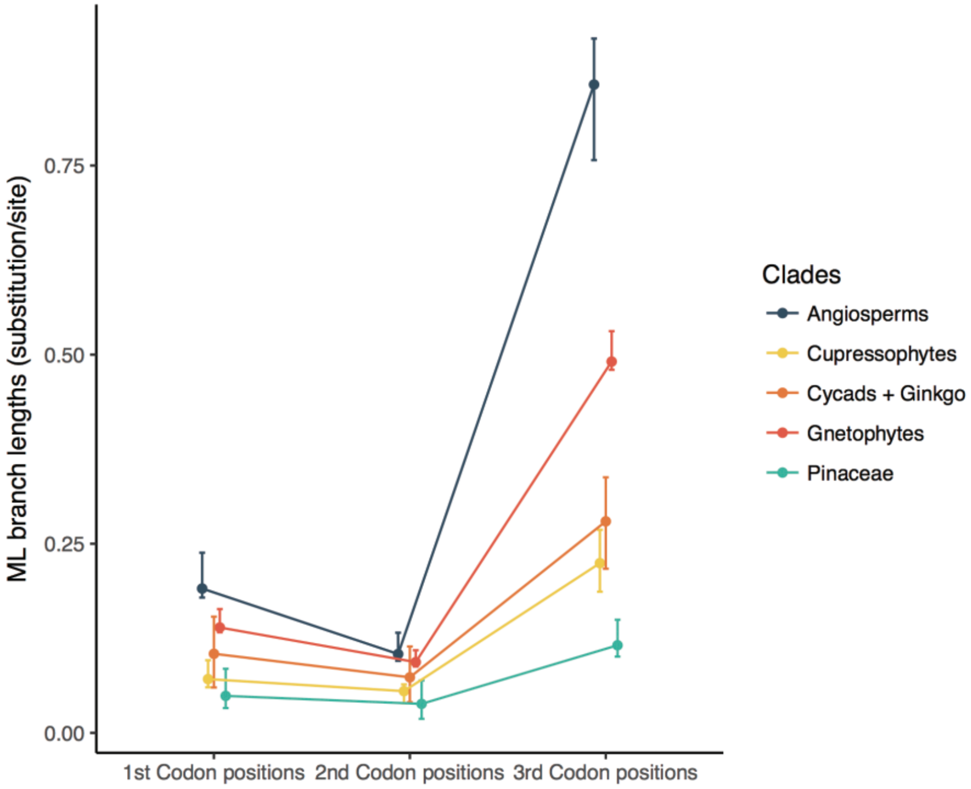

Fig. 6.—Lineage specific branch length estimates from each species to the most recent common ancestor of the five monophyletic groups (angiosperms, cupressophytes, cycads and Ginkgo, Gnetophytes, and Pinaceae), in trees inferred from sites at 1st, 2nd, and 3rd codon positions. See text for details.

Table 2.
The Index of Substitution Saturation (I_{SS}) on Concatenated Nucleotide Alignments and Alignments of Each Codon Position

Dataset	# Sites	I_{SS}	$I_{SS.C.Sym}$	$I_{SS.C.Asym}$
Alignment with 3rd codon positions (NT123)	149,679	0.612	0.820*	0.605*
Alignment with 1st and 2nd codon positions (NT12)	99,786	0.521	0.819*	0.603*
Alignment of 1st codon positions (NT1)	49,893	0.551	0.818*	0.598*
Alignment of 2nd codon positions (NT2)	49,893	0.494	0.818*	0.598*
Alignment of 3rd codon positions (NT3)	49,893	0.796	0.818*	0.598*

*P value $< 1\times10^{-4}$, two-tailed t-test.

codon positions. Since using 3rd codon positions solely can produce nearly identical phylogenies as those based on the NT123 alignment (fig. 3 and supplementary fig. S7, Supplementary Material online), it is plausible to assume that inclusion of the 3rd codon positions in the concatenated alignment of nucleotide sequences leads to systematic bias in the phylogenetic analysis of seed plants, which constantly placed gnetophytes as a sister group to the other gymnosperms.

We further tested whether codon and amino acid substitution models are robust to the potential bias introduced by the 3rd codon positions. Unlike DNA substitution models, codon substitution models can explicitly describe synonymous and nonsynonymous substitutions and realistically estimate natural selection acting on protein-coding sequences. By separating the two types of substitutions with different rates, they are supposed to reflect both recent and early divergences (Ren et al. 2005; Gil et al. 2013). Protein sequences, as the translated products of coding sequences, have been shown to be less affected by substitution saturation than nucleotide sequences (Wickett et al. 2014), as they record nonsynonymous substitutions but ignore synonymous substitutions that may hamper phylogenetic inference due to substitution saturation (Seo and Kishino 2008). As mostly synonymous sites, sites at 3rd codon positions may negligibly influence the phylogenetic placement of gnetophytes under the codon and amino acid substitution models. Therefore, trees built under the codon and amino acid models were expected to be congruent with those inferred from NT12 alignments and the GTR + GAMMA model. Surprisingly, the codon model and amino acid model both gave nearly identical ML trees as the topologies inferred from the NT123 alignment under the GTR + GAMMA model, highly supporting the "Gnetales—other gymnosperms" hypothesis (supplementary figs. S13 and S14, Supplementary Material online). A similar topology has been suggested by Lee et al. (2011) based on a concatenated amino acid matrix of nuclear genes, although all amino acid substitution matrices in Wickett et al. (2014) strongly support a closer relationship between gnetophytes and conifers.

Since the propensities of amino acids play an important role in the evolutionary rates across sites, an effect not modeled by the discrete GAMMA distribution in our ML analysis, we used the CAT and CAT + GTR model implemented in PhyloBayes-MPI to infer the phylogeny based on single-copy genes (Pagel and Meade 2004; Lartillot et al. 2009, 2013). For computational reasons, the original amino acid alignment consisting of 49,893 sites was reduced to a shorter alignment with 7,562 sites (see "Material and Methods" section). The reduced alignment resulted in a similar ML topology as the original amino acid alignment under the JTT + I+GAMMA + F model (supplementary fig. S15, Supplementary Material online). Interestingly, the CAT model supported the "Gnetales—other gymnosperms" hypothesis (posterior probability = 0.98, supplementary fig. S16, Supplementary Material online), while the CAT + GTR model supported the "Gnepine" hypothesis (posterior probability = 0.86, supplementary fig. S17, Supplementary Material online). Because the CAT model uses flat exchange rates that are not actually realistic, the CAT + GTR model is more appropriate for real biological data and is virtually always the model with the highest fit in PhyloBayes (Lartillot et al. 2009). Amino acid compositions also exhibited compositional heterogeneity in a few species distributed across the phylogeny, as "ppred" in PhyloBayes-MPI pointed out. *Physcomitrella patens*, *Medicago truncatula*, *Musa acuminata*, *Oryza sativa*, *Pinus taeda*, *Pinus banksiana*, and *Gnetum Montanum* rejected compositional homogeneity under the CAT + GTR model (posterior predictive $P < 0.05$). In summary, as the sites at 3rd codon positions were included in the "codon" alignment and GC content is correlated with specific amino acid residues (Ruhfel et al. 2014), the above results suggest that the codon model (GY) and the amino acid model (JTT + I+GAMMA + F and CAT) may fail to accommodate the systematic bias introduced by the 3rd codon positions, except for the CAT + GTR model.

Phylogeny Based on Multi-Species Coalescent Model

Except for the analyses based on concatenated alignments, we also applied recently developed coalescent approaches implemented in STAR (Liu et al. 2009) and in ASTRAL-II (Mirarab and Warnow 2015), taking into account incomplete lineage sorting in gene trees. To further assess the effects of 3rd codon positions on the placement of gnetophytes, we built gene trees of the 106 different phylogenetic markers based on alignments with and without 3rd codon positions. The two sets of gene trees were named as "GT123" and "GT12", respectively. Coalescent analyses on GT123 from both STAR and ASTRAL-II were largely congruent with the ML phylogenies inferred from the NT123 alignment with both the DNA model, codon model, and amino acid model, hence in support of the "Gnetales-other gymnosperms" hypothesis (100% BP, supplementary figs. S18 and S19, Supplementary Material online). Nevertheless, GT12 resulted in two different topologies with respect to gnetophytes. STAR fully supported the "Gnetales-other gymnosperms" hypothesis (100% BP, supplementary fig. S20, Supplementary Material online), but ASTRAL-II supported the "Gnetifer" hypothesis (60% BP), which placed gnetophytes as a sister group to all conifers (supplementary fig. S21, Supplementary Material online). However, the "Gnetifer" topology was accepted by neither the NT123 alignment (*P* value $= 2\times10^{-11}$ for three partitions by each codon position and *P* value $= 3\times10^{-103}$ for 52 partitions by PartitionFinder) nor the NT12 alignment (*P* value $= 1\times10^{-47}$ for two partitions by each codon position, and *P* value $= 0.001$ for 37 partitions by PartitionFinder).

The phylogenetic signal in the two sets of gene trees was further measured by IC and ICA, which account for existed topological bipartitions in gene trees to estimate incongruence of phylogenetic signal (Salichos and Rokas 2013; Salichos et al. 2014; Kobert et al. 2016). We used IC and ICA to determine the incongruence in both GT123 and GT12 trees with respect to the three alternative topologies obtained from the phylogenomic analyses described above (fig. 7). Interestingly, both sets of gene trees have no prevalent bipartitions to support either cupressophytes (fig. 7*A* and *C*) or gnetophytes (fig. 7*B*) as a sister group to Pinaceae, since the values of IC and ICA were extremely close to zero. However, there was a slight phylogenetic signal to group gnetophytes within or with conifers from the GT12 gene trees inferred without 3rd codon positions (fig. 7*B* and *C*, respectively). In contrast to the incompatible phylogenetic signals for the position of gnetophytes, both sets of gene trees exhibited a strong phylogenetic signal for *Ginkgo* as a sister group to cycads independent of the position of gnetophytes (fig. 7).

Discussion

Single-Copy Genes Resolve the Phylogeny of Seed Plants

Resolving the exact phylogeny of seed plants is fundamental to our understanding of the evolution, diversification, and colonization of major plant groups on Earth. Despite recent advances in sequencing technologies and great efforts to use diverse sets of molecular markers, the phylogenetic relationships among the five main seed plant lineages remain contested. Here, we have identified a set of 1,469 single-copy genes that are shared among 65 species comprising five seed plant lineages. This data set represents one of the most comprehensive comparative studies including gymnosperm species. With such a broad taxonomic sampling that includes all

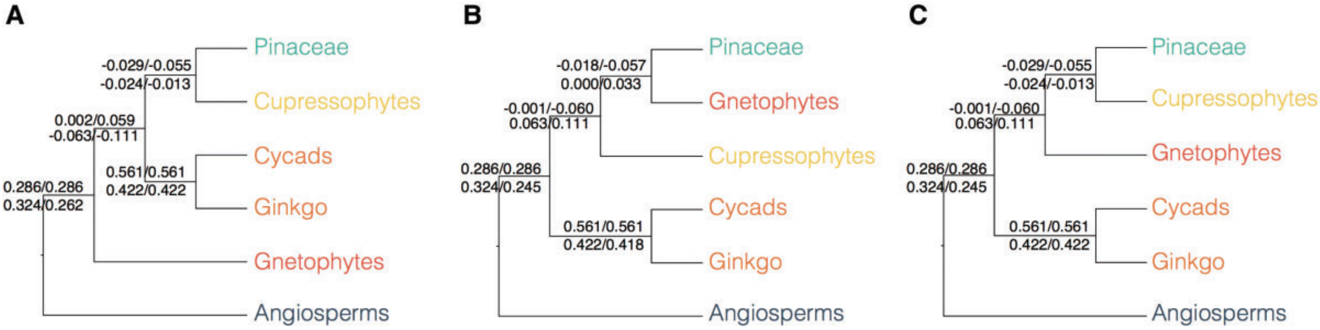

Fig. 7.—Internode certainty (IC) and internode certainty all (ICA) estimated from gene trees of 106 phylogenetic markers for the deep divergence of seed plants. (*A*) The "Gnetales—other gymnosperms" hypothesis; (*B*) the "Gnepine" hypothesis; and (*C*) the "Gnetifer" hypothesis. Numbers above branches represent IC and ICA estimated from the gene trees based on alignments with 3rd codon positions; numbers below branches represent IC and ICA estimated from the gene trees based on alignments without 3rd codon positions.

conifers (except Araucariaceae), cycads, *Ginkgo*, gnetophytes and angiosperms, our markers have the potential to unlock phylogenetic and evolutionary relationships in seed plants.

The phylogenetic markers developed here are effective markers for phylogenetic analyses in each lineage of seed plants. With different partitioning strategies and multispecies coalescent methods, the markers give clear phylogenetic relationships within angiosperms, Pinaceae, cupressophytes, cycads, and gnetophytes. The phylogenies, for instance, inferred from the NT123 alignment partitioned by PartitionFinder based on the GTR + GAMMA model (fig. 3), based on the codon substitution model (supplementary fig. S13, Supplementary Material online), and based on the multi-species coalescent models with GT123 (supplementary figs. S18 and S19, Supplementary Material online), all provide excellent examples of the applications of the 106 phylogenetic markers in all lineages of seed plants. It is also interesting to note that 3rd codon positions of the phylogenetic markers have limited effects on such phylogenetic relationships within each clade. Although the position of *A. alba* in Pinaceae changes in a small fraction of the phylogenetic trees, this is probably due to the lack of species available in closely related genera to *Abies*, for example *Keteleeria*, *Pseudolarix*, *Nothotsuga*, and *Tsuga*.

Our phylogenetic markers have the further potential to resolve the deep divergence of seed plants. The only conflicting clade in this study remains the gnetophytes, which is notorious in almost all current phylogenomic analyses (Zhong et al. 2010, 2011; Xi et al. 2013; Wang and Ran 2014; Wickett et al. 2014). Some of our topologies, including the ones inferred from the NT123 alignment with the substitution models of DNA, codons, and amino acids, as well as the coalescent based methods with exception of one ASTRAL-II analysis, all support the "Gnetales—other gymnosperms" hypothesis with high bootstrap values. The "Gnepine" topology is obtained by the amino acid alignment under the CAT + GTR model and the concatenated alignments of nucleotide sequences without 3rd codon positions (NT12, NT1, and NT2). The "Gnetifer" hypothesis is only supported—with low bootstrap values—by one ASTRAL-II analysis based on GT12 and is rejected by AU tests accounting for the NT123 and NT12 alignments.

Removing 3rd codon positions in nuclear genes can change the position of gnetophytes as shown in this study and in Wickett et al. (2014), and we found further evidence to argue that 3rd codon positions contribute to most of the compositional heterogeneity in the NT123 alignment and exhibit increase of evolutionary rates to different extents in different lineages of seed plants. Therefore, including 3rd codon positions in alignments of nuclear genes is most likely unfit for the GTR + GAMMA model and adds phylogenetic noise when dealing with the deep divergence of seed plants. Such noise may also pose problems for phylogenetic inference based on the amino acid and codon substitution models, which may explain the different observations reported by Lee et al. (2011) and Wickett et al. (2014). It is worth noting that although it is computationally intensive, the CAT + GTR model is still among one of the most robust amino acid models when it comes to dealing with various phylogenetic noise. Last but not least, gene trees of the 106 phylogenetic markers indicate an inconsistent mixture of disparate phylogenetic signals on the related internode with respect to the positions of gnetophytes (fig. 7). The heterogeneous phylogenetic signals for the exact phylogenetic position of gnetophytes are consistent with the evolutionary history of gymnosperms, which endured several extinctions and recent radiations (Crisp and Cook 2011; Nagalingum et al. 2011; Wang and Ran 2014). The lack of ancient diverged lineages in gymnosperms as well as the lack of exhaustive samples from fossil lineages may mislead current systematic studies.

With respect to the "Gnetales—other gymnosperms" hypothesis, the "Gnepine" hypothesis has been widely accepted when considering other molecular evidence except for molecular sequences. For example, both gnetophytes and Pinaceae lost some homologous genes in the chloroplast, such as the *rps16* gene and two introns of *clpP* (Wu et al. 2009). Alternatively, the loss of nonhomologous inverted repeats in

Pinaceae and cupressophytes is not against the “Gnepine” hypothesis (Wu et al. 2011). Among those lost genes, the most striking example is the loss of all 11 plastid *ndh* genes in gnetophytes and Pinaceae, which is usually interpreted as a major synapomorphy for gnetophytes and *Pinaceae* (Braukmann et al. 2009). Like other plastid protein complexes, the NDH complex requires subunits encoded in both the plastid and the nucleus, so related genes would get lost coordinately. However, the pattern of loss of nuclear-encoded *ndh* genes is different in gnetophytes and Pinaceae, particularly for the retained *ndhS* gene in *Pinaceae* (Ruhlman et al. 2015). Also, the loss of all plastid *ndh* genes is less likely an immediate but a continuous process, as many pseudogenes of *ndh* still exist in the chloroplast genome in extant Pinaceae (Wakasugi et al. 1994). Furthermore, convergent loss of *ndh* genes is not rare among seed plants. Several lineages in Orchidaceae and Geraniales also lost plastid and nuclear *ndh* genes, coordinately (Ruhlman et al. 2015). Therefore, the loss of *ndh* genes could be interpreted as compatible with both the “Gnepine” or “Gnetales—other gymnosperms” hypothesis.

Our results also confirm that *Ginkgo* and cycads form a monophyletic group, which is strongly supported by all phylogenomic topologies estimated in this study. Compared to previous studies, in which the sister relationship of *Ginkgo* and cycads depended on the presence or absence of gnetophytes and tree-building approaches used (Wu et al. 2013), our phylogenetic placement of *Ginkgo* is exceptionally solid. The gene trees of the 106 phylogenetic markers also show a definite preference for the monophyly, which is consistent with morphological traits such as haustorial pollen tube and motile sperm (Lee et al. 2011; Wang and Ran 2014).

Limits and Perspectives

We are well aware of the limitations of using draft genome assemblies and transcriptome data for the identification of single-copy genes. Single-copy gene families may suffer from the biased estimation of copy numbers due to gene predictions from draft assemblies (Denton et al. 2014) as well as artifacts of transcriptome assembly. Although transcriptome sequencing has considerably expanded our knowledge on the physiology and evolution of gymnosperms (Ralph et al. 2008; Rigault et al. 2011; Chen et al. 2012; Hodgins et al. 2016), they still often result in partial or redundant allelic transcripts, which may lead to erroneous copy number estimations because of the flawed construction of gene families. In fact, this is a more serious issue in gymnosperms than in angiosperms, because gymnosperms tend to have high heterozygosity (Wang and Ran 2014), which could fail De Bruijn Graph-based assembly algorithms and leads to partial or redundant allelic transcripts (Ruttink et al. 2013).

Besides, the integration pipeline we used to remove redundancy can also bias copy number estimation through elimination of some recently duplicated genes. Because CD-HIT-EST collapsed transcript sequences with similarities higher than 90%, not only different isoforms and allelic transcripts were removed, but possibly also some duplicated genes with high sequence similarity. However, a stringent cut-off of similarity may fail to deal with high allelic variation in gymnosperm sequences (Wang and Ran 2014) and data from different samples. To a certain degree, the functional analysis of single-copy genes in seed plants resulted in similar functional categories as the single-copy genes in angiosperms (De Smet et al. 2013; Li et al. 2016) and other eukaryotes (Waterhouse et al. 2011), suggesting the loose cut-off used here had only negligible effects.

The optimal solution to the problems described above are of course well-assembled gymnosperm genomes, but recently released conifer genomes are still extremely fragmented (Birol et al. 2013; Nystedt et al. 2013; Neale et al. 2014; Zimin et al. 2014; Warren et al. 2015). While the sequencing of some new gymnosperm genomes is in progress, the published ones are continuously being improved using more sophisticated assembly strategies or novel technologies, which yield longer reads and better genome assemblies (Warren et al. 2015). All these efforts would further improve our knowledge on seed plant phylogeny, diversification, and their evolutionary history.

Supplementary Material

Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments

This work was supported by the European Union Seventh Framework Programme (FP7/2007-2013) under the ProCoGen (Promoting Conifer Genomic Resources) project [FP7-KBBE-289841]; the Multidisciplinary Research Partnership “Bioinformatics: from nucleotides to networks” Project of Ghent University [01MR0310W]; and the Spanish Ministry of Economy and Competitiveness [PinCoxSeq-AGL2012-35175 to M.T.C]. The computational resources (Stevin Supercomputer Infrastructure) and services used for PhyloBayes-MPI were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government—Department EWI.

Literature Cited

Birol I, et al. 2013. Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497.

Braukmann TWA, Kuzmina M, Stefanović S. 2009. Loss of all plastid ndh genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr Genet. 55:323–337.

Burleigh JG, Barbazuk WB, Davis JM, Morse AM, Soltis PS. 2012. Exploring diversification and genome size evolution in extant gymnosperms through phylogenetic synthesis. J Bot. 2012:1–6.

Burleigh JG, Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am J Bot. 91:1599–1613.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Canales J, et al. 2013. De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology. Plant Biotechnol J. 12:286–299.

Cañas RA, Canales J, Gómez-Maldonado J, Avila C, Cánovas FM. 2014. Transcriptome analysis in maritime pine using laser capture microdissection and 454 pyrosequencing. Tree Physiol. 34:1278–1288.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Chase MW, Reveal JL. 2009. A phylogenetic classification of the land plants to accompany APG III. Bot J Linn Soc. 161:122–127.

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci U S A. 97:4086–4091.

Chen J, et al. 2012. Sequencing of the needle transcriptome from Norway spruce (*Picea abies* Karst L.) reveals lower substitution rates, but similar selective constraints in gymnosperms and angiosperms. BMC Genomics 13:589.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14:1147–1159.

Christenhusz M, Byng JW. 2016. The number of known plants species in the world and its annual increase. Phytotaxa. 261:201–217.

Cibrián-Jaramillo A, et al. 2010. Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. Genome Biol Evol. 2:225–239.

Cooper ED. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends Plant Sci. 19:576–582.

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

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

De La Torre AR, et al. 2014. Insights into conifer giga-genomes. Plant Physiol. 166:1724–1732.

De La Torre AR, Lin Y-C, Van de Peer Y, Ingvarsson PK. 2015. Genome-wide analysis reveals diverged patterns of codon bias, gene expression, and rates of sequence evolution in picea gene families. Genome Biol Evol. 7:1002–1015.

De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. 2017. Contrasting Rates of Molecular Evolution and Patterns of Selection among Gymnosperms and Flowering Plants. Mol Biol Evol. doi: 10.1093/molbev/msx069.

De La Torre-Bárcena JE, et al. 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome-wide EST data. PLoS One 4:e5764.

De Smet R, et al. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci. 110:2898–2903.

Denton JF, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 10:e1003998.

Doyle JA. 1998. Phylogeny of vascular plants. Annu Rev Ecol Syst. 29:567–599.

Duarte JM, et al. 2010. Identification of shared single copy nuclear genes in *Arabidopsis*, *Populus*, *Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. BMC Evol Biol. 10:61.

Duvick J, et al. 2008. PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res. 36:D959–D965.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol. 7:e1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fiz-Palacios O, Schneider H, Heinrichs J, Savolainen V. 2011. Diversification of land plants: insights from a family-level phylogenetic analysis. BMC Evol Biol. 11:341.

Gernandt DS, López GG, García SO, Liston A. 2005. Phylogeny and classification of *Pinus*. Taxon. 54:29–42.

Gil M, Zanetti MS, Zoller S, Anisimova M. 2013. CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models. Mol Biol Evol. 30:1270–1280.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 11:725–736.

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hall BG. 2011. Phylogenetic Trees Made Easy. Sunderland, MA: Sinauer Associates.

Haston E, Richardson JE, Stevens PF, Chase MW, Harris DJ. 2009. The Linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. Bot J Linn Soc. 161:128–131.

Hodgins KA, Yeaman S, Nurkowski KA, Rieseberg LH, Aitken SN. 2016. Expression divergence is correlated with sequence evolution but not positive selection in conifers. Mol Biol Evol. 33:1502–1516.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. Genome Res. 9:868–877.

Kobert K, Salichos L, Rokas A, Stamatakis A. 2016. Computing the internode certainty and related measures from partial gene trees. Mol Biol Evol. 33:1606–1617.

Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 29:1695–1701.

Larkin MA, et al. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25:2286–2288.

Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst Biol. 62:611–615.

Lee EK, et al. 2011. A functional phylogenomic view of the seed plants. PLoS Genet. 7:e1002411.

Levin RA, Whelan A, Miller JS. 2009. The utility of nuclear conserved ortholog set II (COSII) genomic regions for species-level phylogenetic inference in *Lycium* (Solanaceae). Mol Phylogenet Evol. 53:881–890.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.

Li Z, et al. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell 28:326–344.

Lin C-P, Huang J-P, Wu C-S, Hsu C-Y, Chaw S-M. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biol Evol. 2:504–517.

Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. Syst Biol. 58:468–477.

Lu Y, Ran J-H, Guo D-M, Yang Z-Y, Wang X-Q. 2014. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. PLoS One 9:e107679.

Maere S, Heymans K, Kuiper M. 2005. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21:3448–3449.

Mathews S. 2009. Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. American J Bot. 96:228–236.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Myburg AA, et al. 2014. The genome of *Eucalyptus grandis*. Nature 509:356–362.

Nagalingum NS, et al. 2011. Recent synchronous radiation of a living fossil. Science 334:796–799.

Neale DB, et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol. 15:R59.

Nei M, Kumar S. 2000. Molecular Evolution and Phylogenetics. Oxford: University Press.

Nystedt B, et al. 2013. The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584.

Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol. 53:571–581.

Palmé AE, Pyhäjärvi T, Wachowiak W, Savolainen O. 2009. Selection on nuclear genes in a Pinus phylogeny. Mol Biol Evol. 26:893–905.

Proost S, et al. 2015. PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res. 43:D974–D981.

Ralph SG, et al. 2008. A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for *Sitka spruce* (*Picea sitchensis*). BMC Genomics 9:484.

Ran J-H, Gao H, Wang X-Q. 2010. Fast evolution of the retroprocessed mitochondrial rps3 gene in Conifer II and further evidence for the phylogeny of gymnosperms. Mol Phylogenet Evol. 54:136–149.

Ren F, Tanaka H, Yang Z. 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst Biol. 54:808–818.

Rensing SA, et al. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69.

Rigault P, et al. 2011. A white spruce gene catalog for conifer genome analyses. Plant Physiol. 157:14–28.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms-inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol. 14:23.

Ruhlman TA, et al. 2015. NDH expression marks major transitions in plant evolution and reveals coordinate intracellular gene loss. BMC Plant Biol. 15:100.

Ruttink T, et al. 2013. Orthology Guided Assembly in highly heterozygous crops: creating a reference transcriptome to uncover genetic diversity in *Lolium perenne*. Plant Biotechnol J. 11:605–617.

Salas-Leiva DE, et al. 2014. Conserved genetic regions across angiosperms as tools to develop single-copy nuclear markers in gymnosperms: an example using cycads. Mol Ecol Resour. 14:831–845.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Salichos L, Stamatakis A, Rokas A. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol Biol Evol. 31:1261–1271.

Schmutz J, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.

Seo T-K, Kishino H. 2008. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol. 57:367–377.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51:492–508.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Sun M, et al. 2014. Deep phylogenetic incongruence in the angiosperm clade Rosidae. Mol Phylogenet Evol. 83C:156–166.

Tuskan GA, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604.

Velasco R, et al. 2010. The genome of the domesticated apple (Malus × domestica Borkh.). Nat Genet. 42:833–839.

Wakasugi T, et al. 1994. Loss of all ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. Proc Natl Acad Sci U S A. 91:9794–9798.

Wang K, et al. 2012. The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet. 44:1098–1103.

Wang X, et al. 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet. 43:1035–1039.

Wang X-Q, Ran J-H. 2014. Evolution and biogeography of gymnosperms. Mol Phylogenet Evol. 75:24–40.

Warren RL, et al. 2015. Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. Plant J. 83:189–212.

Waterhouse RM, Zdobnov EM, Kriventseva EV. 2011. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol Evol. 3:75–86.

Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci. 111:E4859–E4868.

Wu C-S, Chaw S-M, Huang Y-Y. 2013. Chloroplast phylogenomics indicates that Ginkgo biloba is sister to cycads. Genome Biol Evol. 5:243–254.

Wu C-S, Lai Y-T, Lin C-P, Wang Y-N, Chaw S-M. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. Mol Phylogenet Evol. 52:115–124.

Wu C-S, Wang Y-N, Hsu C-Y, Lin C-P, Chaw S-M. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. 3:1284–1295.

Wu F, Mueller LA, Crouzillat D, Pétiard V, Tanksley SD. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174:1407–1420.

Xi Z, Rest JS, Davis CC. 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. PLoS One 8:e80870.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol Phylogenet Evol. 26:1–7.

Xia X. 2013. DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol. 30:1720–1728.

Zeng L, et al. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun. 5:4956.

Zhang N, Zeng L, Shan H, Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195:923–937.

Zhong B, et al. 2011. Systematic error in seed plant phylogenomics. Genome Biol Evol. 3:1340–1348.

Zhong B, Yonezawa T, Zhong Y, Hasegawa M. 2010. The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol. 27:2855–2863.

Zhu X-Y, et al. 2007. Mitochondrial matR sequences help to resolve deep phylogenetic relationships in rosids. BMC Evol Biol. 7:217.

Zimin A, et al. 2014. Sequencing and assembly of the 22-gb loblolly pine genome. Genetics 196:875–890.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51:588–598.

Associate editor: Bill Martin