PLANT SCIENCES

10.1126/sciadv.1501084

Early genome duplications in conifers and other seed plants

Zheng Li,[1] Anthony E. Baniaga,[1] Emily B. Sessa,[2] Moira Scascitelli,[3] Sean W. Graham,[3] Loren H. Rieseberg,[3,4] Michael S. Barker[1]*

Polyploidy is a common mode of speciation and evolution in angiosperms (flowering plants). In contrast, there is little evidence to date that whole genome duplication (WGD) has played a significant role in the evolution of their putative extant sister lineage, the gymnosperms. Recent analyses of the spruce genome, the first published conifer genome, failed to detect evidence of WGDs in gene age distributions and attributed many aspects of conifer biology to a lack of WGDs. We present evidence for three ancient genome duplications during the evolution of gymnosperms, based on phylogenomic analyses of transcriptomes from 24 gymnosperms and 3 outgroups. We use a new algorithm to place these WGD events in phylogenetic context: two in the ancestry of major conifer clades (Pinaceae and cupressophyte conifers) and one in *Welwitschia* (Gnetales). We also confirm that a WGD hypothesized to be restricted to seed plants is indeed not shared with ferns and relatives (monilophytes), a result that was unclear in earlier studies. Contrary to previous genomic research that reported an absence of polyploidy in the ancestry of contemporary gymnosperms, our analyses indicate that polyploidy has contributed to the evolution of conifers and other gymnosperms. As in the flowering plants, the evolution of the large genome sizes of gymnosperms involved both polyploidy and repetitive element activity.

INTRODUCTION

Polyploidy, or whole genome duplication (WGD), is one of the most important forces in vascular plant evolution. Nearly 25% of vascular plants are recent polyploids (*1*), with approximately 15% of angiosperm and 31% of fern speciation events due to genome duplication (*2*). Ancient polyploidy is found in the ancestry of all extant seed and flowering plants (*3*), and many angiosperm lineages have experienced additional rounds of genome duplication (*4*–*10*). Changes in the rates of molecular evolution and turnover in genome content following polyploidy may have provided novel genetic variation that was important for the evolution of plant diversity (*3*, *8*, *11*–*16*).

Despite the prevalence of polyploidy in the history of flowering plants, the role of polyploidy in gymnosperm evolution is less clear. The extant gymnosperms appear to be the sister clade of angiosperms (*17*), and they diverged from their most recent common ancestor (MRCA) as much as 310 million years ago (*18*). Most evidence indicates that polyploid speciation is relatively rare among extant gymnosperms (*2*), although in some genera (for example, *Ephedra*), polyploidy is prevalent (*19*, *20*). Previous analyses of conifer genome sizes and chromosomes suggested that paleopolyploidy occurred in Pinaceae (*19*, *21*). Although there was evidence of an ancient polyploidy shared by all seed plants (*3*), no evidence of a gymnosperm or conifer ancient polyploidy was found in the genome of Norway spruce (*Picea abies*), the first published gymnosperm genome. However, this conclusion was based on only a single plot of the relative ages of duplicate genes, presumably because the genome assembly was not of high enough quality (N50 = 4.87 kb) for syntenic analyses. Based on the pattern of accumulation of paralogs seen in this plot, they suggested that the large genomes of conifers originated by mechanisms exclusive of WGD, in particular through proliferation of long terminal repeat retrotransposons (LTR-RTs). Given that paleopolyploidy has been repeatedly observed among flowering plants and is also hypothesized to occur among conifers (*19*, *21*), our goal was to test more thoroughly for evidence of ancient polyploidy in gymnosperms, using a phylogenetically diverse data set and a new phylogenomic method for determining the phylogenetic placement of WGDs.

We assembled transcriptomes for 24 gymnosperms and 3 outgroup species, including representatives of all major gymnosperm and vascular plant clades (table S1). Three of these transcriptomes—*Ophioglossum petiolatum*, *Gnetum gnemon*, and *Ephedra frustillata*—were newly sequenced to cover phylogenetic gaps in our data set. For each transcriptome, we used our DupPipe bioinformatic pipeline to generate age distributions of paralogs to identify shared bursts of gene duplication that are indicative of ancient WGD (*7*, *22*, *23*). We also introduce a newly developed algorithm, Multi-tAxon Paleopolyploidy Search (MAPS), to place inferred paleopolyploid events in phylogenetic context. For each node in a phylogeny, MAPS evaluates the percentage of gene duplications shared by all taxa descended from that node. Ancient WGDs are identified and located as peaks in plots of duplication events shared among a set of species (Materials and Methods; figs. S1 and S2). We used MAPS to confirm and locate genome duplication events in the history of the gymnosperms and seed plants.

RESULTS

Phylogenetic position of the ancient seed plant polyploidy

Most seed plant species contained evidence of a gene duplication peak consistent with previous evidence for a WGD in the ancestry of all seed plants (*3*). With the exception of the Gnetales taxa, each gymnosperm Ks plot (fig. S3) had a peak with a median Ks = 0.75 to 1.5, which, in some of these taxa, has previously been correlated with a WGD shared by all seed plants (*3*). Among the Gnetales, we only observed a peak with a median Ks = 1.05 in *Welwitschia mirabilis*, which is consistent with a *Welwitschia*-specific WGD (*4*). All three Gnetales taxa do not

[1]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA. [2]Department of Biology, University of Florida, Gainesville, FL 32611, USA. [3]Department of Botany, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [4]Department of Biology, Indiana University, Bloomington, IN 47405, USA.
*Corresponding author. E-mail: msbarker@email.arizona.edu

contain clear evidence of the putative seed plant WGD, perhaps due to elevated substitution or gene birth/death rates among these species.

To place this ancient WGD in the vascular plant phylogeny, we implemented a new multispecies paleopolyploid search tool, MAPS. Previous analyses found evidence for an ancient polyploidy in the ancestry of all extant seed plants, Jiao *et al.* (*3*). However, a major clade of vascular plants, the monilophytes (ferns), was not included in that analysis. It was therefore unclear if this WGD is shared among all euphyllophytes (seed plants and monilophytes) or restricted to only seed plants. To better place this WGD in the vascular plant phylogeny, we analyzed new transcriptome data from the eusporangiate fern *Ophioglossum* with data from *Araucaria* (gymnosperm), *Ginkgo* (gymnosperm), *Amborella* (angiosperm), and *Selaginella* (lycophyte, the sister lineage to euphyllophytes). Gene trees were constructed for 3235 gene families with at least one gene copy present in each species. Among these gene families, MAPS identified 544 subtrees that included the MRCA of *Araucaria*, *Ginkgo*, and *Amborella*, which were consistent with the species tree. Nearly 64% of these subtrees contained evidence for a shared duplication in the MRCA of the seed plants that was not shared with *Ophioglossum* (Fig. 1A, fig. S4A, and table S2). This result demonstrates that the unclearly delimited euphyllophyte genome duplication (*3*) is indeed limited to seed plants as a whole and not shared with ferns and other vascular plants (Fig. 2).

Independent paleopolyploidies in Pinaceae and Cupressaceae

Most gymnosperm lineages only contained evidence for a single, ancient WGD, but some species had multiple signals. The Ks plots for most of the conifers contained a younger peak consistent with a WGD since the seed plant genome duplication (fig. S3). Among Pinaceae, we observed a younger peak with a median Ks = 0.2 to 0.4 for each taxon in our data set. Similarly, gene age distributions for taxa in Cephalotaxaceae, Cupressaceae, and Taxaceae contained a younger peak with a median Ks = 0.2 to 0.5. *Araucaria* was the only conifer in our data set without an unambiguous younger peak. Thus, the Ks plots suggest that there may have been one shared conifer WGD or independent WGDs in the history of different conifer families.

We conducted two different MAPS analyses to resolve the placement and number of WGDs among the conifers. For one analysis, we selected the transcriptomes of *Pinus*, *Larix*, and *Cedrus* to represent Pinaceae, and the transcriptome of *Taxus* to represent Taxaceae; we chose *Ginkgo*, *Ophioglossum*, and *Selaginella* as outgroups. We recovered 2175 gene family phylogenies with at least one gene copy from each taxon. MAPS identified 625 subtrees among these gene family phylogenies that included the MRCA of Pinaceae. More than 52% of the subtrees supported a shared duplication in the ancestry of Pinaceae (Fig. 1B, fig. S4B, and table S3). In contrast, only 9% of 535 subtrees supported a gene duplication shared between Pinaceae and Taxaceae. In the second analysis, we selected *Taxus* (Taxaceae), *Cephalotaxus* (Cephalotaxaceae), *Cryptomeria* (Cupressaceae), and *Pinus* (Pinaceae), with *Ginkgo*, *Ophioglossum*, and *Selaginella* as outgroups. Among 1886 gene family phylogenies for these taxa, MAPS identified 469 subtrees that included the MRCA of the cupressophytes. More than 42% of the subtrees supported a shared gene duplication in the MRCA of Cupressaceae and Taxaceae (Fig. 1C, fig. S4C, and table S4). Only 10% of the subtrees supported a duplication event shared by Pinaceae, Cupressaceae, and Taxaceae. We found similar results with MAPS using only gene trees with >50% bootstrap support for all branches (table S5). These results suggest that there are two ancient WGDs in the conifers: one shared by Cupressaceae and Taxaceae (the cupressophytes), and one in the ancestry of Pinaceae (Fig. 2).

Analyses of ortholog divergence corroborated our MAPS results and supported independent WGDs among the conifers. We identified 3266 orthologs by reciprocal best BLAST hit (*22*) from representatives

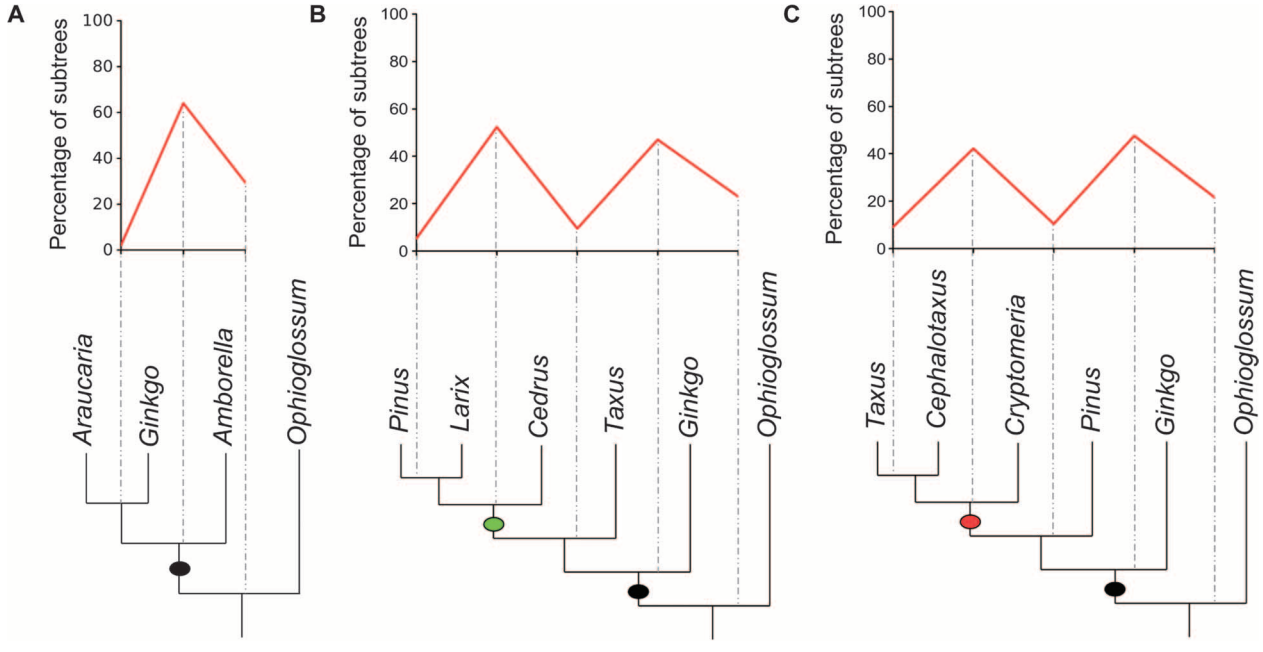

Fig. 1. MAPS results on the associated phylogeny. Percentage of subtrees that contained a gene duplication (red line) shared by descendant species at each node. Ovals correspond to inferred locations of WGD events. (**A**) Seed plant analysis: black oval, seed plant WGD. (**B**) Pinaceae analysis: black oval, seed plant WGD; green oval, Pinaceae WGD. (**C**) Cupressophyte analysis: black oval, seed plant WGD; red oval, cupressophyte WGD.

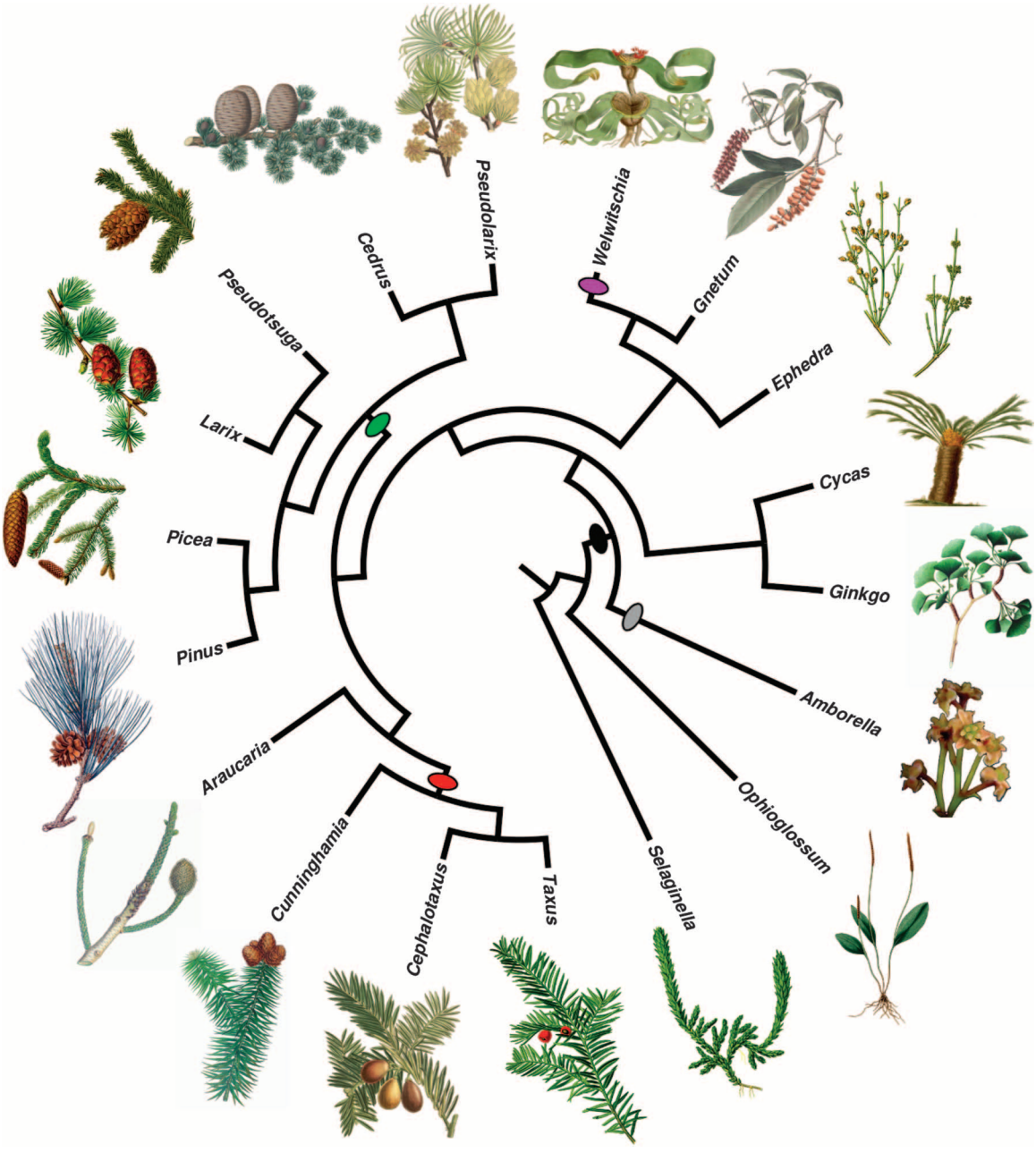

Fig. 2. Phylogenetic placement of WGDs in seed plant and gymnosperm history. Ovals correspond to inferred locations of WGD events; black, seed plant WGD; gray, angiosperm WGD; purple, *Welwitschia* WGD; green, Pinaceae WGD; red, cupressophyte WGD. All botanical illustrations are in the public domain. *Amborella* image adopted from *Amborella* Genome Project, 2013 (*46*). Other botanical illustrations are in the public domain (*59*–*75*).

of Pinaceae and Cupressaceae, *Picea glauca* and *Cryptomeria japonica*. Excluding poorly aligned orthologs with Ks >5, the median orthologous divergence between *P. glauca* and *C. japonica* was Ks = 0.78. In contrast, their most recent WGDs occurred at median Ks = 0.35 and 0.24, respectively (Fig. 3), much later than the divergence of their lineages. Orthologous divergence and phylogenomic approaches both support independent WGDs in Pinaceae and cupressophytes. Consistent with this interpretation is an absence of evidence for these WGDs in Araucariaceae (fig. S3). Overall, these results are consistent with previous analyses of chromosomes and genome sizes that hypothesized no paleopolyploidy in Araucariaceae, but likely ancient WGD in Pinaceae (*19*, *21*).

DISCUSSION

In contrast to the recently published study of the Norway spruce genome (*24*), our analyses find evidence for at least two independent WGDs in the ancestry of major conifer clades. Why did analyses of the spruce genome not recover similar evidence of this WGD? Visual

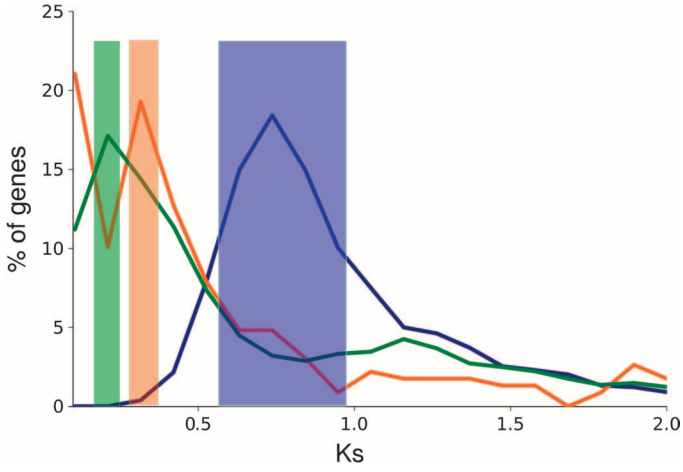

Fig. 3. Pinaceae-Cupressaceae ortholog divergence and independent WGDs. Combined Ks plot of the gene age distributions of *P. glauca* (Pinaceae; green) and *C. japonica* (Cupressaceae; orange), and their ortholog divergences (blue). The median peaks for these plots are highlighted. Analyses of ortholog divergence indicated that these two taxa diverged before their most recent WGDs.

evaluation of the age distribution of paralogs from that analysis [Supplementary Fig. 2.6 of Nystedt *et al.* (*24*)] suggests that there is in fact a peak consistent with a WGD near Ks ~0.25, similar to our results. Although it is not clear why this result was overlooked, the spruce genome results do appear to be fully consistent with our analyses. Our more extensive phylogenetic sampling provides additional support that this peak is likely a WGD because more than 50% of gene families in multiple Pinaceae species have paralogs from this event (Fig. 1, B and C, and fig. S4, B and C).

What are the implications of these results for our understanding of conifer genome evolution? First, Nystedt *et al.* (*24*) proposed a model of conifer genome evolution that must be revised in light of our results. Their model suggests that in the absence of polyploidy, 12 ancestral conifer chromosomes expanded at a slow and steady rate owing solely to the activity of a diverse set of LTR transposable elements. Although conifer chromosome numbers cluster near $n = 12$ (*25*), our discovery of WGDs in the ancestry of two major conifer clades (Pinaceae and cupressophytes) indicates that these numbers must have fluctuated rather than remained completely static over time. Our analyses do not contradict evidence that the expansion of repetitive DNA is the major contributor to conifer genome size evolution. However, the dynamics of conifer genome evolution clearly did involve WGDs, and genome duplication events have played a role in generating some of the largest genomes among conifers (for example, Pinaceae). It is notable that the genome sizes of paleopolyploid Cupressaceae and Taxaceae are not substantially larger on average than that of non-paleopolyploid Araucariaceae (*26*, *27*). This finding suggests that an insight from angiosperm genome evolution also holds true for gymnosperms; differences in turnover rates of genome content likely contribute more to genome size variation than a single paleopolyploidy (*12*, *28*, *29*).

Nystedt *et al.* (*24*) also suggests that conserved synteny across Pinaceae (*30*) results from an absence of paleopolyploidy. Analyses of angiosperm genomes indicate that the degree of synteny conservation following paleopolyploidy varies widely (*12*, *31*–*33*). The composition of parental genomes, in particular differences in transposon load, may establish genome dominance that leads to the biased retention and loss of genes (*33*). If most fractionation and genome rearrangements occur quickly after polyploidy, descendant polyploids may also inherit a largely common synteny (*34*, *35*). The lack of reciprocal genome rearrangements following WGDs, such as in Poaceae (*36*), would also reduce syntenic diversity in descendant lineages. For decades, the broad ancestry of polyploidy in the flowering plants was undetected in linkage mapping studies. Thus, relatively conserved synteny, especially from linkage map data, is not evidence against a paleopolyploidy in Pinaceae.

One of the most intriguing evolutionary questions raised by our analyses is, why are there so few polyploid species among extant conifers and other gymnosperms? Our analyses indicate that polyploid speciation contributed to their diversity. Perhaps these WGDs thrived at a climatically favorable time for polyploid species, as was proposed to explain the apparent clustering of angiosperm WGDs near the K-Pg mass extinction event (*37*). Based on our phylogenetic placements of WGDs and existing estimates for the ages of gymnosperm lineages (*38*), the conifer WGDs occurred ca. 210 to 275 million years ago (Cupressaceae + Taxaceae) and ca. 200 to 342 million years ago (Pinaceae). Many major events in Earth's history occurred during this time frame, including Earth's most severe mass extinction event, the Permian-Triassic extinction. Did polyploid conifers survive the end-Permian event better than did their diploid contemporaries? Given that many of these conifer clades originated during this period, these WGDs may have uniquely contributed to the morphological and biological diversity of these lineages. Polyploidy may differentially influence the evolution of dosage-sensitive genes and pathways (*16*, *39*–*41*) or generate novelty by sub- or neofunctionalization (*42*). Examining further data sets to more precisely pinpoint these WGDs in the conifer phylogeny and to explore the effects of duplication on specific gene families will be critical to further answer how polyploidy has contributed to conifer evolution.

MATERIALS AND METHODS

Sampling and sequencing

Leaf material of *O. petiolatum* (PRJNA257107), *G. gnemon* (PRJNA283231), and *E. frustillata* (PRJNA283230) was collected in liquid nitrogen from the University of British Columbia (UBC) Botanical Gardens and Greenhouse and then stored in a −80°C freezer (table S1). We extracted total RNA using the TRIzol reagent (Invitrogen)/RNeasy (Qiagen) approach as described by Lai *et al.* (*43*). For 454 sequencing (454 Life Sciences), we used modified oligo-dT primers for complementary DNA (cDNA) synthesis to reduce the length of mononucleotide runs associated with the polyadenylate [poly(A)] tail of mRNA. We used a "broken chain" short oligo-dT primer to prime the poly(A) tail of mRNA during first-strand cDNA synthesis (*44*). cDNA was amplified and normalized with the TRIMMER-DIRECT cDNA Normalization Kit. After normalization, we fragmented the cDNA to 500–800 base pair fragments by either sonication or nebulization and removed small fragments through size selection using AMPure SPRI beads (Angencourt). Then, the fragmented ends were polished and ligated with adaptors. The optimal ligation products were selectively amplified and subjected to two rounds of size selection by gel electrophoresis and AMPure SPRI bead purification (*45*). Normalized cDNA was prepared for sequencing following the standard genomic DNA shotgun protocol recommended by 454 Life Sciences.

Additional data sets were downloaded from the GenBank Sequence Read Archive (SRA) (table S1). These included Sanger and Illumina data from 22 species. Data sets were selected to provide broad phylogenetic coverage of the gymnosperms. We also obtained the annotated coding DNA sequences of *Amborella trichopoda* (*46*) and *Selaginella moellendorffii* (*47*) from Phytozome (www.phytozome.net/).

Transcriptome assembly

Raw read quality filtering and trimming were performed by SnoWhite (*48*) before assembly. Three different assembly strategies were used for our three different data types. Sanger expressed sequence tags (EST) were cleaned using the SeqClean pipeline and assembled using TGICL. For 454 data, we used a combination of MIRA and CAP3 to assemble contigs. We used MIRA version 3.2.1 (*49*) using the "accurate.est.denovo.454" assembly mode. Because MIRA may split up high-coverage contigs into multiple contigs, we used CAP3 at 94% identity to further assemble the MIRA contigs and singletons (*50*). SOAPdenovo-Trans (*51*) was used to assemble Illumina sequenced transcriptomes using a *k*-mer of about $^{2}/_{3}$ read length. All other parameters were set to default. Assembly statistics for the 26 assemblies are given in table S1.

Age distribution of paralogs

For each species data set, we used our DupPipe pipeline to construct gene families and estimate the age of gene duplications (*7*, *22*, *23*, *47*, *52*). Translations and reading frames were estimated by Genewise alignment to the best hit protein from a collection of proteins from 25 plant genomes on Phytozome. As in other DupPipe runs, we used protein-guided DNA alignments to align our nucleic acids while maintaining the reading frame. For each node in our gene family phylogenies, we estimated synonymous divergence (Ks) using PAML with the F3X4 model (*53*). Summary plots of the age distribution of gene duplications were evaluated for each gymnosperm species for peaks of gene duplication as evidence of ancient WGDs. Taxa with peaks suggesting ancient WGDs were further analyzed using a multispecies approach (described below) to assess what fraction of gene families show a shared gene duplication and simultaneously place potential WGDs in phylogenetic context.

Estimating the orthologous divergence of Pinaceae and Cupressaceae

To estimate the average ortholog divergence of conifer taxa and compare it to observed paleopolyploid peaks, we used our previously described RBH Ortholog pipeline (*22*). Briefly, we identified orthologs as reciprocal best blast hits in the transcriptomes of *P. glauca* (Pinaceae) and *C. japonica* (Cupressaceae). Using protein-guided DNA alignments, we estimated the pairwise synonymous (Ks) divergence for each pair of orthologs using PAML with the F3X4 model (*53*). We plotted the distribution of ortholog divergences and compared the median divergence against the synonymous divergence of paralogs from inferred WGDs in these lineages.

Inference of gene family phylogenies

Each transcriptome was translated into amino acid sequences using the TransPipe pipeline (*22*). We performed reciprocal protein BLAST (blastp) searches of selected transcriptomes with an e-value of 10^{-5} as a cutoff. Gene families were clustered from these BLAST results using OrthoMCL v2.0 with default parameters (*54*). Using a custom perl script, we filtered for gene families that contained at least one gene copy from each taxon and discarded the remaining OrthoMCL clusters. SATé was used for automatic alignment and phylogeny reconstruction of gene families (*55*). For each gene family phylogeny, we ran SATé until five iterations without an improvement in score using a centroid breaking strategy. MAFFT was used for alignments (*56*), Opal for mergers (*57*), and RAXML for tree estimation (*58*). The best SATé tree for each gene family was used to infer and locate WGDs by our MAPS algorithm.

Multi-tAxon Paleopolyploidy Search (MAPS)

To infer and locate ancient WGDs in our data sets, we developed a gene tree sorting and counting algorithm, MAPS. This algorithm uses a given species tree to filter for subtrees within complex gene trees consistent with relationships at each node in the species tree. For each node of the species tree, MAPS parses the species tree into subtrees with a sister species and an outgroup, for example, ((A,B),C). MAPS iteratively searches for each of these subtrees in the gene tree and will ignore subtrees that do not have the expected relationship. In-paralogs are collapsed by MAPS to simplify the search. We filter for these substrees, rather than filtering on entire topologies, because ancient WGDs may yield phylogenies with many nested and/or orthologous clades. Filtering for a simple gene tree that matches the species tree would eliminate many of the trees that support WGDs. By filtering for subtrees of the species tree, MAPS captures the evidence for polyploidy in complex gene family topologies. Using this filtered set of gene trees, MAPS records the number of subtrees that support a gene duplication at a particular node in the species tree (fig. S1). To infer and locate a potential WGD in the species tree, we plot the percentage of gene duplications shared by descendant taxa by node (fig. S2). A WGD will produce a large burst of shared duplications across taxa and gene trees. This burst of duplication will appear as an increase in the percentage of shared gene duplications in our MAPS analyses.

To evaluate if a WGD occurred before the divergence of taxa A and B, MAPS requires gene trees with at least a sister group A and B and an outgroup C (fig. S1). The basic algorithm of MAPS has two steps. In step 1, MAPS collapses in-paralogs that evolved after the divergence of A and B to a single copy in each gene tree (fig. S1). In step 2, MAPS counts subtrees from all gene trees that are consistent with a duplication event in the MRCA of A and B. In our ABC example, subtrees with a topology consistent with duplication before the divergence of A and B [for example, (((A,B),(A,B)),C)] will be recorded as a duplication at their MRCA node (fig. S1, 1.6). Additionally, subtrees with a topology consistent with duplication before the divergence of A and B followed by independent gene loss [for example, ((A,~),(A,B)),C) or (((A,B),(~,B),C)] will also be recorded as a duplication at their MRCA node (fig. S1, 1.7 to 1.10). If gene trees do not have a topology consistent with any gene duplication among the ingroup taxa, then no duplications will be recorded at the internal nodes (fig. S1, 1.1 to 1.5). When searching for ancient WGDs in a collection of gene trees that contain more than three taxa, MAPS will repeat the same algorithm on each node of the tree (fig. S2). WGDs are inferred by searching for evidence of a large number of shared duplications at a particular node(s) of the species tree (fig. S2).

To evaluate the phylogenetic placement of the putative "seed plant" WGD, we used MAPS to analyze gene families from representatives of each vascular plant lineage (Fig. 1A and fig. S4A). We selected *Araucaria angustifolia* and *Ginkgo biloba* to represent gymnosperms because our Ks plots suggest that they only experienced the seed plant WGD. We also analyzed the *Amborella* genome to represent angiosperms (*46*). The

newly sequenced *O. petiolatum* transcriptome and the *S. moellendorffii* genome (*47*) were chosen to represent ferns and lycophytes, respectively.

We conducted two MAPS analyses to evaluate numbers and placements of WGDs among conifers (Fig. 1, B and C, and fig. S4, B and C). Two analyses were conducted instead of one because the MAPS algorithm works best with simple, ladderized species trees. To maximize the numbers of gene trees in the MAPS analysis and have good coverage of the Pinaceae phylogeny, we selected the transcriptomes of *Pinus monticola*, *Larix gmelinii*, and *Cedrus atlantica* to represent Pinaceae. We also selected *Taxus mairei* to represent the cupressophytes. Likewise, we chose *T. mairei*, *Cephalotaxus hainanensis*, and *C. japonica* to represent cupressophytes, and *P. monticola* to represent Pinaceae. For both Pinaceae and cupressophyte analyses, the transcriptomes of *G. biloba* and *O. petiolatum* as well as the *S. moellendorffii* genome were selected as outgroups.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/1/10/e1501084/DC1

Fig. S1. Example topologies processed by MAPS to identify a gene duplication (red star) or not (black dot) in a given gene family phylogeny.

Fig. S2. Example MAPS summary results for a four-taxon phylogeny.

Fig. S3. Histograms of the age distribution of gene duplications from 24 gymnosperm transcriptomes.

Fig. S4. Numerical summary of MAPS results.

Table S1. Assembly statistics and accession numbers for 25 transcriptomes and 2 genomes.

Table S2. Number of gene subtrees that fit the expected species tree support shared duplication in seed plant analysis.

Table S3. Number of gene subtrees that fit the expected species tree support shared duplication in Pinaceae analysis.

Table S4. Number of gene subtrees that fit the expected species tree support shared duplication in cupressophyte analysis.

Table S5. Number of gene subtrees that fit the expected species tree support shared duplication in cupressophyte analysis using only trees with >50% bootstrap support for each branch.

REFERENCES AND NOTES

1. M. S. Barker, N. Arrigo, A. E. Baniaga, Z. Li, D. A. Levin. On the relative abundance of auto- and allopolyploids. *New Phytol.* 10.1111/nph.13698 (2015).
2. T. E. Wood, N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, L. H. Rieseberg., The frequency of polyploid speciation in vascular plants. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 13875–13879 (2009).
3. Y. Jiao, N. J. Wickett, S. Ayyampalayam, A. S. Chanderbali, L. Landherr, P. E. Ralph, L. P. Tomsho, Y. Hu, H. Liang, P. S. Soltis, D. E. Soltis, S. W. Clifton, S. E. Schlarbaum, S. C. Schuster, H. Ma, J. Leebens-Mack, C. W. dePamphilis, Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**, 97–100 (2011).
4. L. Cui, P. K. Wall, J. H. Leebens-Mack, B. G. Lindsay, D. E. Soltis, J. J. Doyle, P. S. Soltis, J. E. Carlson, K. Arumuganathan, A. Barakat, V. A. Albert, H. Ma, C. W. dePamphilis, Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).
5. D. E. Soltis, C. J. Visger, P. S. Soltis, The polyploidy revolution then…and now: Stebbins revisited. *Am. J. Bot.* **101**, 1057–1078 (2014) (available at www.amjbot.org/content/101/7/1057.short).
6. Y. Jiao, J. Li, H. Tang, A. H. Paterson. Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots. *Plant Cell* **26**, 2792–2802 (2014).
7. M. S. Barker, N. C. Kane, M. Matvienko, A. Kozik, R. W. Michelmore, S. J. Knapp, L. H. Rieseberg. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol. Biol. Evol.* **25**, 2445–2455 (2008).
8. N. Arrigo, M. S. Barker. Rarely successful polyploids and their legacy in plant genomes. *Curr. Opin. Plant Biol.* **15**, 140–146 (2012).
9. Steven B. Cannon, Michael R. McKain, Alex Harkess, Matthew N. Nelson, Sudhansu Dash, Michael K. Deyholos, Yanhui Peng, Blake Joyce, Charles N. Stewart Jr., Megan Rolf, Toni Kutchan, Xuemei Tan, Cui Chen, Yong Zhang, Eric Carpenter, Gane Ka-Shu Wong, Jeff J. Doyle, Jim Leebens-Mack, Multiple polyploidy events in the early radiation of nodulating and non-nodulating legumes. *Mol. Biol. Evol.* **32**, 193–210 (2015).
10. M. R. McKain, N. Wickett, Y. Zhang, S. Ayyampalayam, W. R. McCombie, M. W. Chase, J. C. Pires, C.W. dePamphilis, J. Leebens-Mack. Phylogenomic analysis of transcriptome data elucidates co-occurrence of a paleopolyploid event and the origin of bimodal karyotypes in Agavoideae (Asparagaceae). *Am. J. Bot.* **99**, 397–406 (2012).
11. Y. Van de Peer, S. Maere, A. Meyer. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).
12. M. S. Barker, G. J. Baute, S.-L. Liu, in *Plant Genome Diversity* (Springer, Vienna, 2012), vol. 1, pp. 155–169.
13. M. E. Schranz, S. Mohammadin, P. P. Edger. Ancient whole genome duplications, novelty and diversification: The WGD radiation lag-time model. *Curr. Opin. Plant Biol.* **15**, 147–153 (2012).
14. A. M. Selmecki, Y. E. Maruvka, P. A. Richmond, M. Guillet, N. Shoresh, A. L. Sorenson, S. De, R. Kishony, F. Michor, R. Dowell, D. Pellman, Polyploidy can drive rapid adaptation in yeast. *Nature* **519**, 349–352 (2015).
15. S. A. Rensing. Gene duplication as a driver of plant morphogenetic evolution. *Curr. Opin. Plant Biol.* **17**, 43–48 (2014).
16. M. Freeling, B. C. Thomas. Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res.* **16**, 805–814 (2006).
17. N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, E. Carpenter, N. Matasci, S. Ayyampalayam, M. S. Barker, J. G. Burleigh, M. A. Gitzendanner, B. R. Ruhfel, E. Wafula, J. P. Der, S. W. Graham, S. Mathews, M. Melkonian, D. E. Soltis, P. S. Soltis, N. W. Miles, C. J. Rothfels, L. Pokorny, A. J. Shaw, L. DeGironimo, D. W. Stevenson, B. Surek, J. C. Villarreal, B. Roure, H. Philippe, C. W. dePamphilis, T. Chen, M. K. Deyholos, R. S. Baucom, T. M. Kutchan, M. M. Augustin, J. Wang, Y. Zhang, Z. Tian, Z. Yan, X. Wu, X. Sun, G. K.-S. Wong, J. Leebens-Mack, Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E4859–E4868 (2014).
18. H. Schneider, E. Schuettpelz, K. M. Pryer, R. Cranfill, S. Magallon, R. Lupia. Ferns diversified in the shadow of angiosperms. *Nature* **428**, 553–557 (2004).
19. M. R. Ahuja. Polyploidy in gymnosperms: Revisited. *Silvae Genet.* **54**, 59–69 (2005).
20. S. M. Ickert-Bond, M. F. Wojciechowski. Phylogenetic relationships in *Ephedra* (Gnetales): Evidence from nuclear and chloroplast DNA sequence data. *Syst. Bot.* **29**, 834–849 (2004).
21. A. Drewry. The G-banded karyotype of *Pinus resinosa* Ait. *Silvae Genet.* **37**, 5–6 (1988) (available at www.sauerlaender-verlag.com/fileadmin/content/dokument/archiv/silvaegenetica/37_1988/37-5-6-218.pdf).
22. M. S. Barker, K. M. Dlugosch, L. Dinh, R. S. Challa, N. C. Kane, M. G. King, L. H. Rieseberg. EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. *Evol. Bioinform. Online* **6**, 143–149 (2010).
23. M. S. Barker, H. Vogel, M. E. Schranz. Paleopolyploidy in the Brassicales: Analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. *Genome Biol. Evol.* **1**, 391–399 (2009).
24. B. Nystedt, N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin, D. G. Scofield, F. Vezzi, N. Delhomme, S. Giacomello, A. Alexeyenko, R. Vicedomini, K. Sahlin, E. Sherwood, M. Elfstrand, L. Gramzow, K. Holmberg, J. Hällman, O. Keech, L. Klasson, M. Koriabine, M. Kucukoglu, M. Käller, J. Luthman, F. Lysholm, T. Niittylä, Å. Olson, N. Rilakovic, C. Ritland, J. A. Rosselló, J. Sena, T. Svensson, C. Talavera-López, G. Theißen, H. Tuominen, K. Vanneste, Z.-Q. Wu, B. Zhang, P. Zerbe, L. Arvestad, R. Bhalerao, J. Bohlmann, J. Bousquet, R. Garcia Gil, T. R. Hvidsten, P. de Jong, J. MacKay, M. Morgante, K. Ritland, B. Sundberg, S. L. Thompson, Y. Van de Peer, B. Andersson, O. Nilsson, P. K. Ingvarsson, J. Lundeberg, S. Jansson. The Norway spruce genome sequence and conifer genome evolution. *Nature* **497**, 579–584 (2013).
25. A. Rice, L. Glick, S. Abadi, M. Einhorn, N. M. Kopelman, A. Salman-Minkov, J. Mayzel, O. Chay, I. Mayrose. The Chromosome Counts Database (CCDB)—A community resource of plant chromosome numbers. *New Phytol.* **206**, 19–26 (2015).
26. S. Garcia, I. J. Leitch, A. Anadon-Rosell, M. Á. Canela, F. Gálvez, T. Garnatje, A. Gras, O. Higalgo, E. Johnston, G. Mas de Xaxars, J. Pellicer, S. Siljak-Yakovlev, J. Vallès, D. Vitales, M. D. Bennett. Recent updates and developments to plant genome size databases. *Nucleic Acids Res.* **42**, D1159–D1166 (2014).
27. J. G. Burleigh, W. B. Barbazuk, J. M. Davis, A. M. Morse, P. S. Soltis. Exploring diversification and genome size evolution in extant gymnosperms through phylogenetic synthesis. *J. Bot.* **2012**, 1–6 (**2012**).
28. A. R. Leitch, I. J. Leitch. Genomic plasticity and the diversity of polyploid plants. *Science* **320**, 481–483 (2008).
29. L. Bromham, X. Hua, R. Lanfear, P. F. Cowman. Exploring the relationships between mutation rates, life history, genome size, environment, and species richness in flowering plants. *Am. Nat.* **185**, 507–524 (2015).
30. N. Pavy, B. Pelgas, J. Laroche, P. Rigault, N. Isabel, J. Bousquet. A spruce gene map infers ancient plant genome reshuffling and subsequent slow evolution in the gymnosperm lineage leading to extant conifers. *BMC Biol.* **10**, 84 (2012).
31. H. Tang, X. Wang, J. E. Bowers, R. Ming, M. Alam, A. H. Paterson. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.* **18**, 1944–1954 (2008).
32. F. Murat, Y. Van de Peer, J. Salse. Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol. Evol.* **4**, 917–928 (2012).

33. M. R. Woodhouse, F. Cheng, J. C. Pires, D. Lisch, M. Freeling, X. Wang. Origin, inheritance, and gene regulatory consequences of genome dominance in polyploids. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5283–5288 (2014).
34. R. J. A. Buggs, S. Chamala, W. Wu, J. A. Tate, P. S. Schnable, D. E. Soltis, P. S. Soltis, B. W. Barbazuk. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.* **22**, 248–252 (2012).
35. Z. Xiong, R. T. Gaeta, J. C. Pires. Homoeologous shuffling and chromosome compensation maintain genome balance in resynthesized allopolyploid *Brassica napus*. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 7908–7913 (2011).
36. J. C. Schnable, M. Freeling, E. Lyons. Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.* **4**, 265–277 (2012).
37. K. Vanneste, G. Baele, S. Maere, Y. Van de Peer. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous–Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).
38. Y. Lu, J.-H. Ran, D.-M. Guo, Z.-Y. Yang, X.-Q. Wang. Phylogeny and divergence times of gymnosperms inferred from single-copy nuclear genes. *PLOS One* **9**, e107679 (2014).
39. G. C. Conant, J. A. Birchler, J. C. Pires. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. *Curr. Opin. Plant Biol.* **19**, 91–98 (2014).
40. M. Freeling. Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60**, 433–453 (2009).
41. M. Bekaert, P. P. Edger, J. C. Pires, G. C. Conant. Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell* **23**, 1719–1728 (2011).
42. P. P. Edger, H. M. Heidel-Fischer, M. Bekaert, J. Rota, G. Glöckner, A. E. Platts, D. G. Heckel, J. P. Der, E. K. Wafula, M. Tang, J. A. Hofberger, A. Smithson, J. C. Hall, M. Blanchette, T. E. Bureau, S. I. Wright, C. W. dePamphilis, M. E. Schranz, M. S. Barker, G. C. Conant, N. Wahlberg, H. Vogel, J. C. Pires, C. W. Wheat. The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 8362–8366 (2015).
43. Z. Lai, B. L. Gross, Y. Zou, J. Andrews, L. H. Rieseberg. Microarray analysis reveals differential gene expression in hybrid sunflower species. *Mol. Ecol.* **15**, 1213–1227 (2006).
44. E. Meyer, G. V. Aglyamova, S. Wang, J. Buchanan-Carter, D. Abrego, J. K. Colbourne, B. L. Willis, M. V. Matz. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**, 219 (2009).
45. Z. Lai, Y. Zou, N. C. Kane, J.-H. Choi, X. Wang, L. H. Rieseberg, Preparation of normalized cDNA libraries for 454 Titanium transcriptome sequencing. *Methods Mol. Biol.* **888**, 119–133 (2012).
46. Amborella Genome Project, The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).
47. J. A. Banks, T. Nishiyama, M. Hasebe, J. L. Bowman, M. Gribskov, C. dePamphilis, V. A. Albert, N. Aono, T. Aoyama, B. A. Ambrose, N. W. Ashton, M. J. Axtell, E. Barker, M. S. Barker, J. L. Bennetzen, N. D. Bonawitz, C. Chapple, C. Cheng, L. G. Correa, M. Dacre, J. DeBarry, I. Dreyer, M. Elias, E. M. Engstrom, M. Estelle, L. Feng, C. Finet, S. K. Floyd, W. B. Frommer, T. Fujita, L. Gramzow, M. Gutensohn, J. Harholt, M. Hattori, A. Heyl, T. Hirai, Y. Hiwatashi, M. Ishikawa, M. Iwata, K. G. Karol, B. Koehler, U. Kolukisaoglu, M. Kubo, T. Kurata, S. Lalonde, K. Li, Y. Li, A. Litt, E. Lyons, G. Manning, T. Maruyama, T. P. Michael, K. Mikami, S. Miyazaki, S. Morinaga, T. Murata, B. Mueller-Roeber, D. R. Nelson, M. Obara, Y. Oguri, R. G. Olmstead, N. Onodera, B. L. Petersen, B. Pils, M. Prigge, S. A. Rensing, D. M. Riaño-Pachón, A. W. Roberts, Y. Sato, H. V. Scheller, B. Schulz, C. Schulz, E. V. Shakirov, N. Shibagaki, N. Shinohara, D. E. Shippen, I. Sørensen, R. Sottoka, N. Sugimoto, M. Sugita, N. Sumikawa, M. Tanurdzic, G. Theissen, P. Ulvskov, S. Wakazuki, J. K. Weng, W. W. Willats, D. Wipf, P. G. Wolf, L. Yang, A. D. Zimmer, Q. Zhu, T. Mitros, U. Hellsten, D. Loqué, R. Otillar, A. Salamov, J. Schmutz, H. Shapiro, E. Lindquist, S. Lucas, D. Rokhsar, I. V. Grigoriev. The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960–963 (2011).
48. K. M. Dlugosch, Z. Lai, A. Bonin, J. Hierro, L. H. Rieseberg, Allele identification for transcriptome-based population genomics in the invasive plant *Centaurea solstitialis*. *G3* **3**, 359–367 (2013).
49. B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. G. Müller, T. Wetter, S. Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14**, 1147–1159 (2004).
50. X. Huang, A. Madan. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
51. Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T.-W. Lam, Y. Li, X. Xu, G. K.-S. Wong, J. Wang, SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660–1666 (2014).
52. T. Shi, H. Huang, M. S. Barker. Ancient genome duplications during the evolution of kiwifruit (*Actinidia*) and related Ericales. *Ann. Bot.* **106**, 497–504 (2010).
53. Z. Yang. Phylogenetic analysis by maximum likelihood (PAML). http://abacus.gene.ucl.ac.uk/software/paml.html (2000).
54. L. Li, C. J. Stoeckert, D. S. Roos. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
55. K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, T. Warnow. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **324**, 1561–1564 (2009).
56. K. Katoh, K. Misawa, K.-I. Kuma, T. Miyata. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
57. T. J. Wheeler, J. D. Kececioglu. Multiple alignment by aligning alignments. *Bioinformatics* **23**, i559–i568 (2007).
58. A. Stamatakis. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
59. *Araucaria*, https://commons.wikimedia.org/wiki/File:Araucaria_cunninghamii_SZ139.png
60. *Cedrus*, http://biodiversitylibrary.org/page/31681271#page/138/mode/1up
61. *Cephalotaxus*, www.botanicus.org/page/469775
62. *Cunninghamia*, https://commons.wikimedia.org/wiki/File:Cunninghamia_sinensis_SZ104.png
63. *Cycas*, www.botanicus.org/page/494873
64. *Ephedra*, https://commons.wikimedia.org/wiki/File:Illustration_Ephedra_distachya0.jpg
65. *Ginkgo*, https://commons.wikimedia.org/wiki/File:Ginkgo_biloba_SZ136.png
66. *Gnetum*, www.botanicus.org/page/1944419
67. *Larix*, https://commons.wikimedia.org/wiki/File:Illustration_Larix_decudua0.jpg
68. *Ophioglossum*, www.botanicus.org/page/858718
69. *Picea*, https://commons.wikimedia.org/wiki/File:Picea_abies_K%C3%B6hler%E2%80%93s_Medizinal-Pflanzen-105.jpg
70. *Pinus*, https://commons.wikimedia.org/wiki/File:Pinus_massoniana_SZ114.png
71. *Pseudolarix*, www.botanicus.org/page/455355
72. *Pseudotsuga*, http://archive.org/stream/traitdesarbres03moui#page/81/mode/1up
73. *Selaginella*, https://commons.wikimedia.org/wiki/File:Illustration_Selaginella_selaginoides0.jpg
74. *Taxus*, https://commons.wikimedia.org/wiki/File:Illustration_Taxus_baccata0.jpg
75. *Welwitschia*, www.botanicus.org/page/438968

Acknowledgments: We thank K. Dlugosch, S. Jorgensen, and X. Qi for discussion. Hosting infrastructure and services were provided by the Biotechnology Computing Facility (BCF) at the University of Arizona. **Funding:** M.S.B. was supported by NSF-IOS-1339156. **Author contributions:** M.S.B. designed the research; M.S. and M.S.B. collected data; Z.L., A.E.B., E.B.S., S.W.G., L.H.R., and M.S.B. conducted analyses and interpreted results; Z.L. and M.S.B. wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Raw reads for the newly sequenced transcriptomes of *O. petiolatum* (PRJNA257107), *G. gnemon* (PRJNA283231), and *E. frustillata* (PRJNA283230) are deposited in the National Center for Biotechnology Information (NCBI) SRA. Additional data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 12 August 2015
Accepted 14 October 2015
Published 20 November 2015
10.1126/sciadv.1501084

Citation: Z. Li, A. E. Baniaga, E. B. Sessa, M. Scascitelli, S. W. Graham, L. H. Rieseberg, M. S. Barker, Early genome duplications in conifers and other seed plants. *Sci. Adv.* **1**, e1501084 (2015).

Early genome duplications in conifers and other seed plants

Zheng Li, Anthony E. Baniaga, Emily B. Sessa, Moira Scascitelli, Sean W. Graham, Loren H. Rieseberg and Michael S. Barker

Sci Adv **1** (10), e1501084.
DOI: 10.1126/sciadv.1501084