# Comparative transcriptomics reveals divergence in pathogen response gene families amongst twenty forest tree species

[1*]Mengmeng Lu, [2]Min Cao, [2,3]Jie Yang, [1]Nathan G. Swenson

[1] Department of Biological Sciences, University of Notre Dame, 100 Galvin Life Sciences, Notre Dame, IN 46556, USA

[2] CAS Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan 666303, China.

[3]Center of Plant Ecology, Core Botanical Gardens, Chinese Academy of Sciences, Mengla, Yunnan 666303, China.

[*]Corresponding author: Department of Biological Sciences, University of Notre Dame, 100 Galvin Life Sciences, Notre Dame, IN 46556, USA. Email: mlu@nd.edu

**Abstract**

Forest trees provide critical ecosystem services for humanity that are under threat due to ongoing global change. Measuring and characterizing genetic diversity is key to understanding adaptive potential and developing strategies to mitigate negative consequences arising from climate change. In the area of forest genetic diversity, genetic divergence caused by large-scale changes at the chromosomal level has been largely understudied. In this study, we used the RNA-seq data of twenty co-occurring forest trees species from genera including *Acer*, *Alnus*, *Amelanchier*, *Betula*, *Cornus*, *Corylus*, *Dirca*, *Fraxinus*, *Ostrya*, *Populus*, *Prunus*, *Quercus*, *Ribes*, *Tilia*, and *Ulmus* sampled from Upper Peninsula of Michigan. These data were used to infer the origin and maintenance of gene family variation, species divergence time, as well as gene family expansion and contraction. We identified a signal of common whole genome duplication events shared by core eudicots. We also found rapid evolution, namely fast expansion or fast contraction of gene families, in plant-pathogen interaction genes amongst the studied diploid species. Finally, the results lay the foundation for further research on the genetic diversity and adaptive capacity of forest trees, which will inform forest management and conservation policies.

**Keywords:** angiosperms, gene duplication, gene family expansion and contraction, Ks plot, Ka/Ks, plant-pathogen interaction

## Introduction

Forest trees occupy nearly a third of the Earth's land surface and provide key ecological benefits and services such as water, air purification and carbon sequestration (Gamfeldt et al., 2013; Miura et al., 2015; FAO, 2020). They are critical components of terrestrial biodiversity in terms of their own diversity, but also with respect to the habitats and renewable materials such as wood, cellulose, and lignin (Cortés et al., 2020). It is estimated that there are ~73,000 tree species globally with most of being rare and regionally endemic, implying vulnerability of global tree species to anthropogenic and climate changes (Cazzolla Gatti et al., 2022). Indeed, climate change has already impacted forests and their species composition via range migration, wildfires, droughts, pest, and pathogens (Kirilenko and Sedjo, 2007; Allen et al., 2010; Rizvi et al., 2015; Hammond et al., 2022). The genetic composition and diversity of the tree species in forests is foundational not only to our understanding of the abiotic and biotic factors governing the distribution and dynamics of species populations, but also to how they will respond to current and future environmental change. Thus, measuring and characterizing intra- and inter-specific genetic diversity is the key for understanding the current structure and dynamics of forests and their adaptive potential and for developing strategies to mitigate the damages brought along with environmental change.

Gene duplication and gene loss are major sources of genetic diversity. Gene duplication, generated through either whole genome duplication (WGD) or small-scale duplication (SSD), is the primary source of new genetic material among flowering plants (Cui et al., 2006; Panchy et al., 2016). WGD contributes to the majority of duplicate genes (AMBORELLA GENOME PROJECT et al., 2013; Soltis and Soltis, 2016). The ancient WGDs can lead to competitive advantage, increased species diversity, and evolutionary innovations (Van de Peer et al., 2009). Most flowering species have a polyploid history either through autopolyploidy or allopolyploidy. Specifically, flowering species are either currently polyploid, or have a polyploid ancestor and have reverted to being diploid (Blanc and Wolfe, 2004). Earlier studies placed at least three WGD events during the radiation of angiosperms, the oldest WGD occurred before the monocot-eudicot divergence, a second WGD occurred specifically to eudicots, a more recent WGD, gamma ($\gamma$) triplication event, occurred before the separation of rosids and asterids (Tang et al., 2008; Jiao et al., 2012). Alternatively, pervasive SSD, including tandem duplication (Zhang, 2003), transposon-mediated duplication (Jiang et al., 2004), segmental duplication (Bailey et al., 2002b), and retroduplication (Brosius, 1991), also generates a great number of duplicate genes. While the majority of duplicate genes end up being silenced and eliminated (Lynch and Conery, 2000), a few preserved duplicate genes may be retained by selection due to neo-functionalization, relative dosage, absolute dosage or sub- functionalization (Conant et al., 2014), and play important role in generating genetic diversity and building up reproductive isolation barriers. Recent studies explained retention of duplicate genes as some paralogous genes depend on or compensate for their cognate copy (Diss et al., 2017), while some paralogous gene pairs are stuck in a functional and structural entanglement status over evolutionary time scale (Kuzmin et al., 2020).

Compared to gene duplication, the mechanism that underlie gene loss are not yet well understood. It is commonly assumed that gene loss is realized by deletion or pseudogenization, as a result of neutral or adaptive processes (Demuth and Hahn, 2009). Most nonsense mutations

are subject to drift to fixation. Some deleterious mutations could be accumulated and eventually become silenced (Johri et al., 2022). Gene loss might be selected to retain dosage balance, as evidenced in a set of single-copy genes with essential housekeeping functions, which are typically restored to become singletons after WGD or SSD (De Smet et al., 2013).

Differential duplication and loss of genes among evolutionary lineages results in gene family expansion and contraction (Demuth and Hahn, 2009). Gene families comprise a set of similar nucleotide or amino acid sequences. They arise as a result of duplication events and have similar cellular and biochemical functions. Gene family expansions and contractions related to metabolic, regulatory and signalling gene networks contribute to the adaptations of plants to various abiotic habitats and biotic interactions (Leebens-Mack et al., 2019). They are sources of plant functional evolution and innovation (Ohno, 1970; Force et al., 1999), exhibiting a clear functional basis, which are influenced by the mechanisms of duplication (Maere et al., 2005; Hanada et al., 2008). Comparisons of gene family content across species, therefore, may provide insight into the ecological interactions and evolutionary pressures that have shaped adaptation and diversity. This is particularly true when we conduct comparative transcriptomic analyses of species that currently co-occur and that, potentially have co-occurred for long periods of time (Swenson and Jones, 2017).

Genetic investigations of forest trees have, generally, focused on nucleotide substitutions in specific genes (Neale and Kremer, 2011; Isabel et al., 2020), while comparative analyses of genetic divergence caused by large-scale changes at the chromosomal levels have been largely understudied. However, the magnitude, rate, and distribution of changes in gene family size are key drivers of the functional differentiation and adaptation of species (Lynch and Conery, 2003; Demuth and Hahn, 2009). Large segmental duplication and changes in copy number are known to be associated with disease response and environmental adaptation (Bailey et al., 2002a; Reid et al., 2016). Therefore, gene family size change is as much important as nucleotide substitution to understanding genetic divergence and evolutionary forces responsible for these changes.

Such knowledge on genetic divergence and associated evolutionary forces in forest trees can help to promote breeding healthy trees (i.e. disease resistant trees) in a changing climate (Allen et al., 2010; Dodds and Rathjen, 2010). The increasing occurrences and severity of insect and disease outbreaks have threatened forest trees (Hurley et al., 2016; Linnakoski and Forbes, 2019), but we lack a thorough understanding to the disease related genes in forest trees. Trouern-Trend et al. (2020) revealed contraction in disease associated gene families in *Juglans* species. Lu et al. (2021) reported differential gene expression patterns between tolerant and susceptible lodgepole pine trees to a fungal pathogen. It would be interesting to explore the potential link between duplication events and disease response in expressed genes of forest trees, as the duplicate genes can quickly lead to changes in gene expression with the acquisition of increased resistance to different pathogens (Picart-Picolo et al., 2020).

In this study, we used the RNA-seq data for twenty co-occurring forest trees species to infer the origin and maintenance of gene family variation as well as species divergence time, gene expansion and contraction. We sought to address the following questions: What is the magnitude of gene duplication? What factors impact duplicate gene evolution? Are the gene gain and loss process random or controlled by natural selection? The answers to these questions are the key to

understanding the basis of adaptive divergence in forest tree species and will help to predict adaptive capacity of these species in a changing climate.

## Materials and methods

*Sample collection*

Leaves of each of twenty forest species were sampled in August, 2021, at the University of Notre Dame Environmental Research Center (UNDERC) near Land O' Lakes, Wisconsin, U.S.A. (Fig. 1). The UNDERC site is located at latitude: 46.23391, longitude: -89.537254, and encompasses approximately ~8,000 acres of mixed deciduous forest on both sides of the state line between Wisconsin and Michigan's upper peninsula. The twenty species used in this study are dominate components of the woody vegetation at the site and are: *Acer rubrum* (red maple), *Acer saccharum* (sugar maple), *Acer spicatum* (mountain maple), *Alnus incana* (Gary alder), *Amelanchier laevis* (smooth serviceberry), *Betula alleghaniensis* (yellow birch), *Betula papyrifera* (paper birch), *Cornus alternifolia* (pagoda dogwood), *Corylus cornuta* (beaked hazel), *Dirca palustris* (Eastern leatherwood), *Fraxinus nigra* (black ash), *Ostrya virgniana* (American hophornbeam), *Populus grandientata* (bigtooth aspen), *Populus tremuloides* (quaking aspen), *Prunus serotina* (black cherry), *Prunus virginiana* (chokecherry), *Quercus rubra* (Northern red oak), *Ribes cynosbati* (prickly gooseberry), *Tilia americana* (basswood), *Ulmus americana* (American elm). They are all angiosperms (flowering plants) and eudicots, belonging to the clades of asterids, saxifragles, and rosids. These species were chosen as they are the most dominant at the study site and in large national forests in region (i.e. Ottawa, Nicolet, and Hiawatha National Forests). Specifically, these species compose >95% of the tree individuals and above-ground biomass in these forests (Swenson et al., 2017). We focus on the most dominant species in order to provide insights into the ecological and evolutionary processes driving transcriptomic variation within and across populations and communities (Swenson and Jones, 2017). Upon collection, the samples were flash frozen in liquid nitrogen, then transported and stored in an ultracold freezer at the field station.

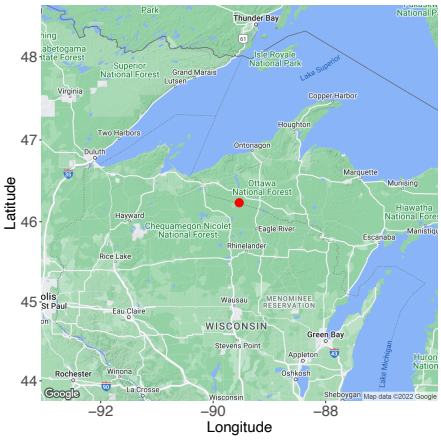


**Fig. 1. The location (red dot) of the University of Notre Dame Environmental Research Center (UNDERC) where all samples used in this study were collected.**

166 *RNA-sequencing, de novo transcriptome assembly, and annotation*
167 The samples were sent to Novogene corporation inc. (Chula Vista, CA, USA) for mRNA
168 extraction, library preparation, mRNA sequencing, and quality control. The samples were
169 sequenced on an Illumina NovaSeq 6000 platform to produce 150bp paired-end reads. Raw reads
170 filtering was performed by Novogene to get the clean reads including removing reads containing
171 adapters, removing reads containing N >10%, and removing reads containing low quality (Phred
172 quality score ≤5) base which is over 50% of the total base. An average of 6 Gb paired-end clean
173 reads were acquired for each sample and deposited in NCBI Sequence Read Archive (SRA
174 accession number: PRJNA896247). *De novo* transcriptome assembly was performed for each
175 sample using the software TransLiG v1.3 with default parameters (Liu et al., 2019). The clean
176 reads were then mapped to the assembled transcriptome using the software RSEM v1.2.25 with
177 default parameters (Li and Dewey, 2011). The longest transcript (isoform) with a transcript per
178 million (TPM) value ≥1.5 for each gene was retained. Redundancy was removed from the
179 retained transcripts using the software CD-HIT v4.8.1 with parameters of -c 0.95 -n 8 -T 32 -M
180 1540 (Fu et al., 2012). To remove potential contamination, the retained transcripts were aligned
181 against the non-redundant (nr) protein database (downloaded on January 22, 2022) using the
182 blastx function implemented by the software DIAMOND v2.0.13.151 (Buchfink et al., 2021)
183 with parameter of --outfmt 102 to assign the best known taxon ID to each query transcript. The
184 file that was used to convert accession numbers to taxonomy was downloaded from
185 https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/, and the tax ID file was downloaded
186 from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/. The query transcripts that could
187 be assigned to green plant (Viridiplantae) taxon IDs were retained. The filtered transcripts were
188 aligned to the nr database that only includes green plant subjects using the blastx function
189 implemented by the software DIAMOND v2.0.13.151 with default parameters. The aligned
190 transcripts were combined with those transcripts assigned with green plant taxon IDs to form the
191 final transcriptomes. The completeness of the final transcriptomes was examined using the 1,614
192 orthologues in the BUSCO (Benchmarking Universal Single Copy Orthologs) v5.2.2 set of
193 embryophyta_odb10 (Manni et al., 2021).
194
195 The transcriptomes were annotated with Gene Ontology (GO) and Kyoto Encyclopedia of Genes
196 and Genomes (KEGG) terms. The transcripts were aligned to the nr database that only includes
197 green plant subjects using the blastx function implemented by the software DIAMOND
198 v2.0.13.151 with parameters of -taxonlist 33090 --very-sensitive -f5. The .xml outputs were
199 used as input for mapping and annotation using the software Blast2GO (Götz et al., 2008). The
200 transcripts were translated to protein sequences using the software TransDecoder v5.5.0
201 (https://github.com/TransDecoder/TransDecoder/wiki) with default parameters to get the longest
202 ORF (open reading frame) per transcript, and then were annotated by the software InterProScan
203 v5.54-87.0 (Jones et al., 2014) with the parameters of -dra -f XML -goterms -pa. The final GO
204 terms were acquired by merging Blast2GO and InterProScan annotation outputs using the
205 software Blast2GO. To get the KEGG terms, the longest ORF per transcript were also annotated
206 using EggNOG-mapper v2.1.7 (Cantalapiedra et al., 2021) with parameters of -m diamond -
207 sensmode more-sensitive -tax_scope 33090 -tax_scope_mode 33090. The assembled
208 transcriptome and annotation were used in the downstream analyses as shown in the flowchart
209 (Supplementary Fig. S1).
210
211 *Ks analyses*

212 The level of synonymous substitutions (Ks) between two homologous sequences increases
213 approximately linearly with time. Thus pairs of paralogous genes within a species can be sorted
214 to estimate the relative ages of gene duplication (Blanc and Wolfe, 2004). To identify paralogous
215 gene pairs within a species, an all-versus-all blastp was performed within each species. The
216 peptide sequences (the longest ORF per transcript) of each species were used as input to run all-
217 versus-all blastp using the software DIAMOND v2.0.13.151 with parameters of -p 16 --max-
218 target-seqs 20 --evalue 1e-10 --outfmt 6 qseqid sseqid pident nident evalue. The blastp output
219 was filtered by removing self blastp results, keeping outputs with pident (percent of identical
220 matches) ≥20 and nident (number of identical matches) ≥50, and removing sequences with more
221 than 10 hits. The filtered paralogous gene pairs were aligned using the software MAFFT v7.490
222 (Katoh et al., 2002) with default parameters. The alignments were used as input for estimating
223 the number of synonymous substitutions per synonymous site (Ks) and the number of non-
224 synonymous substitutions per nonsynonymous site (Ka) using the program yn00 in the software
225 PAML v4.9j (Yang, 2007) with parameters of icode=0 weighting=0 commonf3×4=0. The Ks of
226 each paralogous gene pairs were acquired using the method by Yang and Nielsen (2000).
227
228 Ks can also be calculated using the orthologous gene pairs between species to estimate speciation
229 events (Blanc and Wolfe, 2004). To identify orthologous gene pairs between species, reciprocal
230 best hit blastp were run between the compared pairs. The peptide sequences (the longest ORF per
231 transcript) of the paired species were used as input to run blastp using the software DIAMOND
232 v2.0.13.151 with parameters of -p 16 --max-target-seqs 20 --evalue 1e-10 --outfmt 6 qseqid
233 sseqid pident nident evalue. The best hits between forward and reverse blastp for each compared
234 sequence pairs were retained, which were then used for sequence alignment and estimating Ka
235 and Ks as abovementioned.
236
237 The Ks distribution is proportional to the time lapse. Thus, the Ks distribution within species can
238 be used to infer the timing of WGD, and the Ks distribution between two different species can
239 infer time since speciation (Blanc and Wolfe, 2004). To plot the Ks distribution, the mixture
240 model was fit and plot for the Ks values ≥0.01 and ≤2 using the R package mixR (Yu, 2022).
241 The best model from a candidate of mixture models was selected using the "select" function
242 based on the information criterion BIC.
243
244 <u>*Estimating duplication events and divergence time*</u>
245 To infer duplication events, the species tree and gene trees were first inferred using STAG and
246 STRIDE algorithms implemented by the program OrthoFinder v2.5.4 (Emms and Kelly, 2017,
247 2018). The inferred species tree was similar to that expected given current knowledge of the
248 Angiosperm phylogeny with the exception of the placement of *Populus tremuloides* and *P.*
249 *grandidentata*. Thus, for downstream analysis and discussion, we removed these two species for
250 the species tree inference. The inferred species tree and gene trees were used to estimate
251 duplication events and divergence time. The gene duplication events on the species tree node
252 were reported if, at least 50% of the descendant species have retained both copies of the
253 duplicated gene. The divergence time was estimated using the software r8s v1.81 (Sanderson,
254 2003) with calibration points of 115 million years (MYA) between *Fraxinus nigra* and *Corylus*
255 *cornuta*, which were obtained from http://www.timetree.org/ (Kumar et al., 2017).
256
257 <u>*Gene families and their evolutionary rate*</u>

To infer gene families, the peptide sequences (the longest ORF per transcript) of the 20 studied transcriptomes were used to construct orthogroups (gene families) using the program OrthoFinder v2.5.4 with parameters of -S diamond -t 16 -a 16. Genes assigned to each gene family were counted for each species. The evolutionary rate of protein coding genes was represented by the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (Ka/Ks) with *Dillenia indica* (elephant apple) being the outgroup and the twenty studied species being the ingroup. *D. indica* is an eudicot species, belonging to the order Dilleniales and the RNA-seq data for this species were downloaded from https://trace.ncbi.nlm.nih.gov/Traces/index.html?view=run_browser&acc=ERR2040190&display=data-access (Leebens-Mack et al., 2019). The raw sequences were filtered using the program fastp with default parameters (Chen et al., 2018). The transcriptome assembly and protein sequence prediction were performed as abovementioned, yielding 22,753 transcripts and 19,345 protein sequences. Then the reciprocal best hit blastp was performed between *D.indica* and each of the twenty studied species to identify the orthologous pairs. The best hits between forward and reverse blastp for each compared sequence pairs were retained, which were then used for sequence alignment and estimating Ka/Ks as abovementioned. The mean Ka/Ks across genes for each gene family per species was used to represent the evolutionary rate for each gene family. Gene families with mean Ka/Ks >10 were filtered out. The graphs were plotted using the R package ggplot2 and ggpubr (Wickham, 2016; Kassambara, 2020). The size and evolutionary rate of gene families were correlated using scatter plot. All statistical analyses were performed in R (R Core Team, 2021). Gene families with a size ≥5 and Ka/Ks >1 were highlighted for further analysis. The enriched GO and KEGG terms were identified in these highlighted gene families. The enriched GO terms of each category (biological process, molecular function, cellular component) were separately detected using the "weight" algorithm with the Fisher ratio test implemented by the R package "topGO" (Alexa and Rahnenfuhrer, 2018). Enriched KEGG terms were detected using the hypergeometric method implemented by the R package "plyr" (Wickham, 2011). Heatmaps were plot using the R package pheatmap (Kolde, 2019).

*Estimating gene family expansion and contraction*

To further study gene family changes, species tree and gene family counts were predicted and used to estimate gene family expansion and contraction. All-versus-all blastp was performed to find the most similar sequences for each sequence in the studied species using the program DIAMOND (Buchfink et al., 2021). The program MCL (Enright et al., 2002) was used to find clusters of similar sequences (gene families). The species tree was estimated using the program FastTree (Price et al., 2009). The programs DIAMOND, MCL, and FastTree were implemented by the program OrthoFinder v2.5.4, default parameters were used. Then, the species tree was converted to a format for estimating ultrametric tree using the program CAFE5 (Mendes et al., 2020) with the command python prep_r8s.py -s 60164 -p 'Fraxinus_nigra,Corylus_cornuta' -c 120. "60164" was the number of sites in the alignment. "-c 120" represents a median of 120 MYA divergence time, it is the calibration points between *Fraxinus nigra* and *Corylus cornuta* checked from http://www.timetree.org/. Then the ultrametric tree was estimated using the program r8s (Sanderson, 2003) with default parameters. The ultrametric tree and gene family counts were used as input for the program CAFE5 (Mendes et al., 2020). Since CAFE5 was not recommended to analyze recent polyploid species, only diploid study species were retained for running CAFE5, including *Cornus alternifolia*, *Fraxinus nigra*, *Dirca palustris*, *Tilia americana*, *Acer saccharum*, *Acer spicatum*, *Amelanchier laevis*, *Quercus rubra*, *Betula papyrifera*, *Alnus*

*incana*, *Corylus cornuta*, *Ostrya virginiana*, and *Ribes cynosbati*. Since assembly errors might cause the observed number of gene copies in gene families to deviate from the true ones and overestimation of the birth-death parameter $\lambda$ (number of gene gains and losses per gene per million years), an error model was estimated first with the ultrametric tree and gene family counts using CAFE5 with the parameters of "-c 16 -p -e". This error model along with the ultrametric tree and gene family counts were used to estimate the birth-death parameter $\lambda$, which represents the probability that any gene will be gained or lost. Multiple $\lambda$ mode was used with three different rates of gene family evolution assigned as below:

((Cornus_alternifolia:3,Fraxinus_nigra:3):3,((((Dirca_palustris:2,Tilia_americana:2):2,(Acer_saccharum:2,Acer_spicatum:2):2):2,((((Betula_papyrifera:2,Alnus_incana:2):2,(Corylus_cornuta:2,Ostrya_virginiana:2):2):2,Quercus_rubra:2):2,Amelanchier_laevis:2):2):2,Ribes_cynosbati:1):1);

The numbers 1, 2, 3 represent three different gene family evolutionary rates that were shared by different clades. This tree file was used as input along with the ultrametric tree, gene family counts, and error model using CAFE5 with the parameters of -c 16 -p -k 3. Gene families or parent nodes with a significantly greater rate of evolution (family-wide p-value <0.01), in other words, rapidly evolving gene families, were counted for each species and each parent node. These gene families significantly violated a random gene birth and death model. Thus, the null hypothesis of neutral evolution was rejected and these families could be recognized as fast evolving gene families (Hahn et al., 2005). Enriched GO and KEGG terms were detected in the fast expanding and fast contracting gene families, respectively.

## Results

### *Ks analyses*

We acquired an average of 41 million raw reads for each sample. The total number of assembled transcripts ranged between 19,459 and 31,804 (Supplementary Table S1). We retrieved 76 ~ 88% of the BUSCOs in the assembled transcriptomes (Supplementary Fig. S2). Most species have an initial Ks peak around 0.02~0.04, except *Fraxinus nigra* and *Tilia americana*, which have the initial peak around 0.1 and 0.09, respectively (Fig. 2). Such initial peaks represent most recently duplicate genes. Another discernable separate peak shared by the studied species is around 0.5 ~ 0.7. Such shared peaks by all species may correspond to large-scale duplication events, perhaps the common WGDs across core eudicots. Some species have other discernable peaks, like *Populus grandidentata* and *Populus tremuloides*, which have discernable peaks around 0.18. Such unique peaks may represent ancient duplication events specific to certain species and different evolutionary patterns across species. The two *Prunus* species and the two *Populus* species have similar Ks shapes, suggestive of common WGDs before the two species diverged from the same ancestors. By examining the orthologous Ks between the two species, we found a high peak close to zero, which may represent the duplication events after the WGDs and divergence (Supplementary Fig. S3).

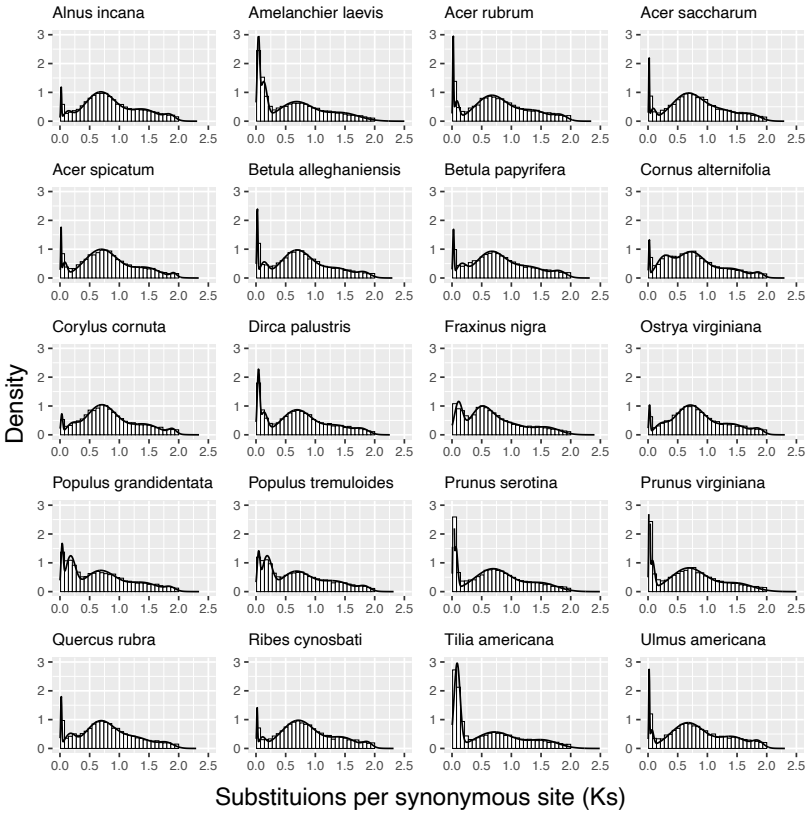


**Fig. 2. Distribution of Ks (the level of synonymous substitutions). A mixture model was fit and plot for the Ks values ≥0.01 and ≤2 of each species using the R package mixR.**

<u>*Duplication events and divergence time*</u>

Using a species tree with estimated branch lengths, we inferred divergence times and duplication events (with at least 50 % support rate, Fig. 3). Most duplication events (n=1843) occurred at the divergence of rosids and asterids, which may correspond to common WGD across eudicots species. The ancestor of *Populus tremuloides* and *Populus grandidentata* had 1,625 duplication events (Supplementary Fig. S4), which is consistent with the recent Ks peaks within these two species. The divergence time of the two asterids species *Cornus alternifolia* and *Fraxinus nigra* from a common ancestor was about 104 MYA (Fig. 3). The divergence time of the studied malvids and fabids from a common ancestor with Saxifragles was about 100 MYA.

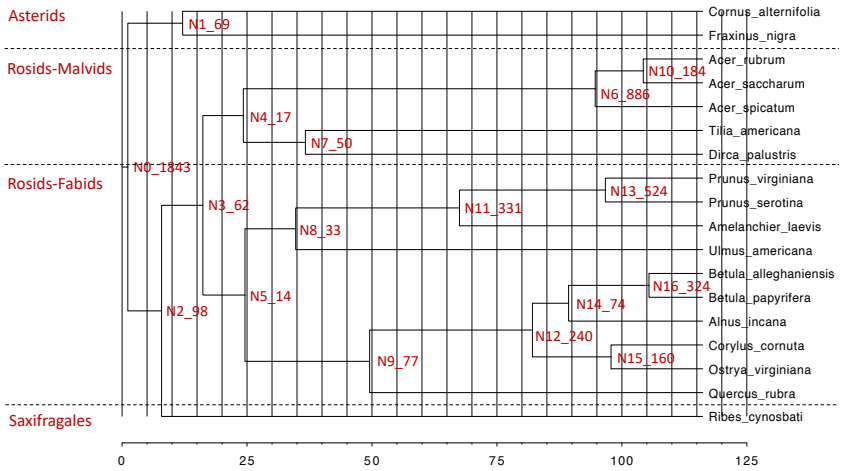


**Fig. 3. Divergence time in million years and gene duplication events with 50 % support rate. Node numbers were followed by number of gene duplication events.**

*Relationship between gene family size and evolutionary rate*

OrthoFinder assigned 393,796 genes (96.2 % of total) to 21,900 orthogroups (gene families). The mean and median gene family size is 17.3 and 12, respectively. We found that 3,060 families present in all species, whereas 19 consisted entirely of single-copy genes.

After identifying reciprocal best blastp hits with the outgroup *D.indica*, we retained a total of 6903 ~ 7302 gene families for the studied species to calculate the relationship between gene family size and evolutionary rate. Gene families with large sizes tended to have low Ka/Ks, yet there are a few exceptions. We highlighted the gene families with a size ≥5 and Ka/Ks >1 in red (Fig. 4a). We tested the enriched KEGG and GO terms in these highlighted gene families (Fig. 4b and Supplementary Fig. S5, see the whole list of GO enrichment in Supplementary Table S2). The most highly shared enriched KEGG and GO terms are related to plant-pathogen interactions, signaling and secondary metabolite biosynthesis.

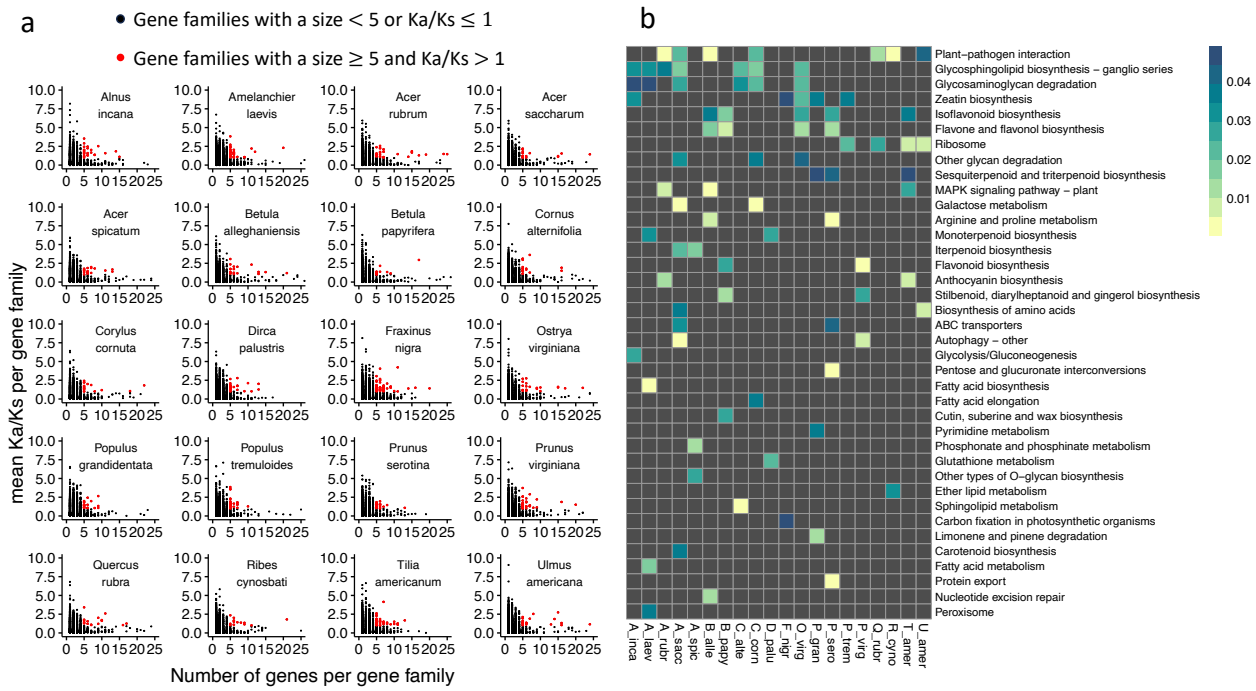


**Fig. 4. Scatter plot between gene size and Ka/Ks per gene family (a) and KEGG pathway terms enrichment (b) in gene families with a size ≥5 and Ka/Ks >1. The color bar in b) represents the adjusted p-values of KEGG enrichment test.**

*Gene family expansion and contraction*

We inferred the birth-death parameter $\lambda$ (number of gene gains and losses per gene per million years), which is specified in different clades, using the program CAFE5. $\lambda 1$ (saxifragales) =0.0034, $\lambda 2$ (rosids) =0.0042, $\lambda 3$ (asterids) =0.0027, which means that the lineages leading to rosids have the highest gene family turnover rates, followed by saxifragales, and then by asterids. The percentage and numbers of gene family expansion and contraction are shown in Fig. 5a and Supplementary Fig. S6. All of the studied species have more rapidly expanding than rapidly contracting gene families (family-wide p-value <0.01, Fig. 5b), which may contribute to their lineage-specific novelty.

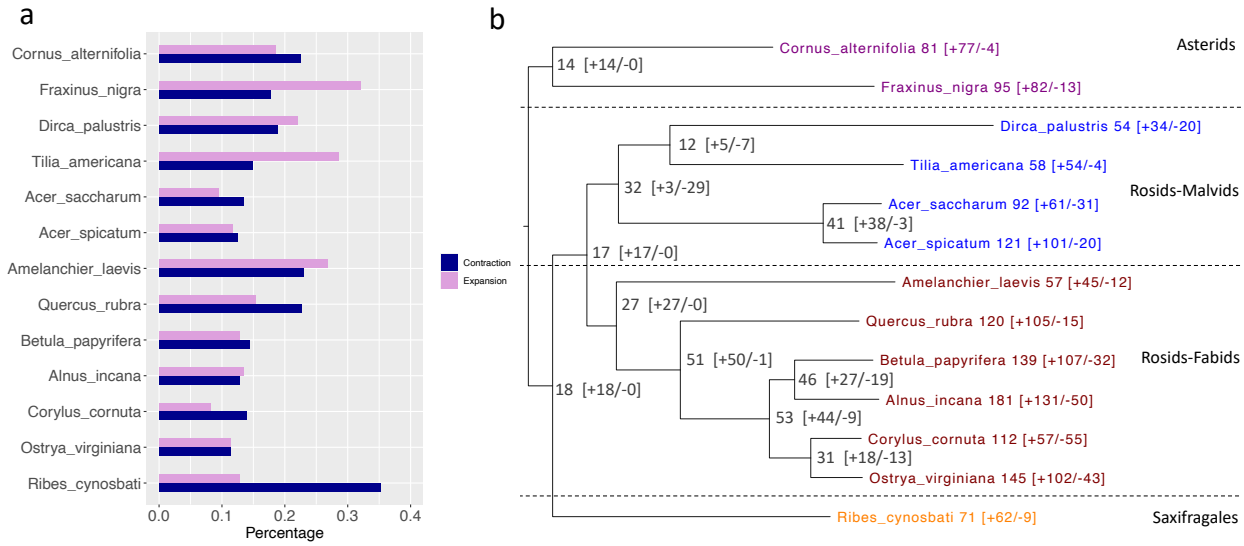


**Fig. 5. The percentage of gene family expansion and contraction (a) and the number of fast evolving gene families (b), with + representing fast expanding gene family and – representing fast contracting gene family.**

We tested the enriched GO and KEGG terms in the fast evolving (expanding or contracting) gene families. All of the studied species shared the enriched KEGG pathway term plant-pathogen interaction (Fig. 6). This term was detected in the fast-expanding gene families of *Cornus alternifolia*, *Tilia americana*, *Acer spicatum*, *Betula papyrifera*, *Quercus rubra*, *Amelanchier laevis*, *Ribes cynosbati*; and this term was detected in the fast-contracting gene families of *Fraxinus nigra*, *Dirca palustris*, *Alnus incana*, *Corylus cornuta*, *Ostrya virginiana*; also, this term was detected in both of the fast expanding and fast contracting gene families of *Acer saccharum*. The gene families assigned to the plant-pathogen interaction term encode proteins (Supplementary Fig. S7) related to disease resistance protein, leucine-rich repeat receptor like protein kinase and others. Some species shared the enriched KEGG pathway terms Tryptophan metabolism and MAPK signaling pathway (Fig. 6).

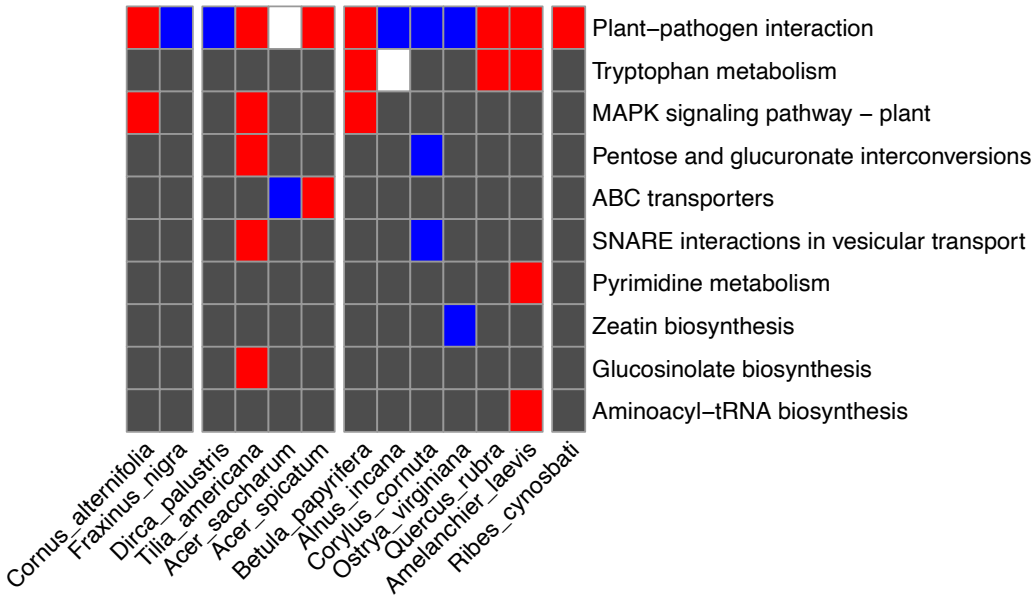


**Fig. 6. Enriched KEGG pathway terms in the studied diploid species. Red represents fast expanding gene families; blue represents fast contracting gene families; white represents both fast expanding and contracting gene families.**

## Discussion

Gene duplication events are prevalent in all organisms and provide genetic variation for evolutionary innovation and adaption to fluctuating conditions (Zhang, 2003; Van de Peer et al., 2009; Panchy et al., 2016). Research on the mechanisms of gene duplication as well as the origin, maintenance, and functional roles of genetic diversity in forest tree species are few due to the limited genome and transcriptome data resources. In the present study, we assembled twenty transcriptomes from forest trees that co-occur. We identified the signal of common WGDs shared by core eudicots, and the signal of recent duplication events specific to certain species. We found gene families involved in plant-pathogen interaction, signaling, and metabolite biosynthesis actively remain extensive variation. We also found fast expansion or fast contraction in plant-pathogen interaction genes amongst the studied diploid species. In the following sections, we will discuss these results in detail.

### *Tempo and mode of gene duplication*

The Ks distribution plots confirmed the widespread paleopolyploidy in plants, that's said, today's diploid species might be polyploid (paleopolyploid) at some points in evolutionary history (Blanc and Wolfe, 2004). The abundant duplication events provide resources for genetic variation. It is worth noting that Ks distribution is not suitable to infer very ancient events due to saturation with multiple substitutions (Blanc and Wolfe, 2004), moreover, different peaks might not be separable from each other, thus it is possible that we missed identifying some duplication events in the present study. Collinearity- or synteny-based methods along with chromosome-level reference genome assembly is alternative method to identify ancient WGD, yet high quality genome assembly is not easy to obtain (Zwaenepoel and Van de Peer, 2019).

453 The predicted species tree (Fig. 3) generally agrees with current phylogenetic hypotheses for the
454 angiosperms with the exception of the two *Populus* species, which is probably caused by the
455 incomplete gene sets obtained from one leaf RNA-seq sample and/or interspecific gene flow.
456 The duplication events at the root (Fig. 3) of the tree is likely to be related to the gamma WGD
457 before the divergence of rosids and asterids. Such large-scale duplication may contribute to
458 genome structure variation, organismal complexity, and rapid speciation in angiosperm species.
459 WGD events followed by different evolutionary events might also occur in the common
460 ancestors for *Acer*, *Prunus*, and *Populus* species. For *Acer*, it is likely that *Acer spicatum*
461 experienced a diploidization event to revert to diploid status, while *Acer rubrum* maintained a
462 polyploid status.
463
464 Divergence time mostly agrees with the previous reports, which stated rosids and asterids
465 diverged about 125 MYA, malvids and fabids diverged about 110 MYA (Zeng et al., 2017).
466 However, divergence estimate in the present study is rather rough. Further studies need to
467 consider using dense taxon sampling, diverse markers like nuclear internal transcribed spacer or
468 chloroplast DNA fragments, and molecular dating with fossil corrections methods to better
469 resolve the evolutionary relationships.
470
471 *Sustained large size gene families*
472 Though most duplicate genes were eliminated or pseudonized (Panchy et al., 2016), a small
473 portion of gene families can actively maintain duplicate genes as selection favors the fitness
474 related neofunctionalization. The main functions of these sustained large size gene families are
475 mainly involved in plant-pathogen interaction, signaling, and metabolite biosynthesis (Fig. 4b &
476 Supplementary Fig. S5). This is consistent with previous report that genes encoding functions
477 like signal transduction and stress response tend to have paralogs (Panchy et al., 2016), and
478 genes related to the function like transcription regulation and signal transduction had higher
479 possibility to survive in multiple rounds of genome duplication (Seoighe and Gehring, 2004).
480 Plants constantly combat with pathogens and pests. The major class of plant disease resistance
481 (R) genes contain a nucleotide-binding site (NBS) and C-terminal leucin-rich repeats (LRRs).
482 Tandem and segmental gene duplication explain much of the diversification among R genes
483 (Meyers et al., 2003; Leister, 2004; Andersen et al., 2020), possibly because they have
484 experienced frequent deletion and duplication throughout evolution, and thus are highly variable
485 (Hanada et al., 2008; Gao et al., 2018). Plant secondary metabolites, such as monoterpenoid,
486 flavonoid, benzoxazinoid, play important roles in defense against predation and pathogens
487 (Falcone Ferreyra et al., 2012; de Bruijn et al., 2018; Bergman et al., 2020). The regulation of
488 plant secondary metabolism requires the intimation of signal transduction network that leads to
489 activation of biosynthetic genes (Zhao et al., 2005).
490
491 Gene duplication and deletion give rise to gene family expansion and contraction. The high gene
492 gain rate likely reflects the occurrence of multiple WGDs and frequent SSDs (Carretero-Paulet et
493 al., 2015). While for the other species with contraction outnumbered expansion, the pattern is
494 mainly caused by gene loss and retention of specific gene families followed the common WGDs
495 preceded their common ancestor (Supplementary Fig. S4). We found a high gene family turnover
496 rate in rosids, but we must caution that the unbalanced taxon sampling in our study species might
497 cause this high turnover value. Moreover, future studies should consider distinguishing the

branches with or without recent WGDs, as the branches with recent WGDs tend to have a high gene turnover rate (Casola and Koralewski, 2018).

*Rapid evolution in plant-pathogen interaction genes*

We found rapid evolution of genes involved in plant-pathogen interaction, tryptophan metabolism and MAPK signaling pathway (Fig. 6) in the adaptive expansion and adaptive contraction gene families across the studied species. The specific expanding or contracting gene families involved in plant-pathogen interaction in each species differed from each other, implying lineage-specificity in changed gene families (Supplementary Fig. S7). This result is consistent with previous studies, which found that fast evolving gene families were mostly groups with functions related with immune defense/stress response, metabolism, cell signaling, chemoreception, and reproduction (Demuth and Hahn, 2009). Additionally, a study on *Juglans* (walnuts) found notable contractions in gene families that are involved in disease resistance and other abiotic stress response (Trouern-Trend et al., 2020). Both metabolism and MAPK signaling pathway are associated with plant defense. MAPK cascades sense pathogen/microbe-associated molecular patters and pathogen effectors, and thus play pivotal roles in signaling plant defense against pathogen attack (Meng and Zhang, 2013). Plant secondary metabolites contribute to plant immunity as antibiotics, or to control callose deposition and programmed cell death (Piasecka et al., 2015). The biotic stress exerted by pathogen or pests is a major source of selection pressure on the studied trees, which drive the rapid radiation of diverse and large plant immune receptor gene repertoires. Expansion and/or contraction of cell-surface receptor genes correlate with the expansion and/or contraction of intracellular receptor genes (Ngou et al., 2022).

## Conclusions

In summary, we found the common WGDs shared by all the studied species as well as signals of recent duplication events specific to certain species, which are suggestive of different evolutionary patterns. Gene families involved in plant-pathogen interaction, signaling, and metabolite biosynthesis actively remain extensive variation and are fast expanding or contracting amongst the studied species. This knowledge calls for an emphasis of gene family size variation, i.e. copy number variation, when studying forest genetic diversity, which will inform forest adaptive strategy. Future studies may focus on classification of duplications in forest tree species, and investigate how the duplications give rise to species-specific innovation and adaption.

## Data availability

The raw RNA-seq data were deposited in NCBI SRA (accession number: PRJNA896247; http://www.ncbi.nlm.nih.gov/sra). The assembled transcriptomes and their annotation were deposited in Dryad https://doi.org/10.5061/dryad.sj3tx968b. The scripts used for analyses were deposited in https://github.com/Mengmeng-Lu/Gene-duplication-and-gene-expansion-analyses-in-twenty-tree-species.

## Acknowledgements

We thank the Center for Research Computing at University of Notre Dame for high performance computing support. We specially acknowledge the assistance of Dr. Dodi Heryadi, Dr. Scott Hampton, and Júlia Simon. The research was performed at the University of

Notre Dame Environmental Research Center that has been generously supported by the Bernard J. Hank Family Endowment. The research was permitted under UNDERC research permit number: UNDERC-2021-2.

**Funding**

MM and NGS were supported by the US National Science Foundation Dimensions of Biodiversity grant (DEB-2124466). JY and MC were supported by a NSFC China-US Dimensions of Biodiversity Grant (DEB-32061123003).

**Conflicts of interest**

The authors declare no conflict of interest.

**Literature cited**

Alexa, A., and Rahnenfuhrer, J. 2018. topGO: Enrichment Analysis for Gene Ontology. R package version 2.34.0.

Allen, C.D., Macalady, A.K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., Kitzberger, T., Rigling, A., Breshears, D.D., Hogg, E.H.T., Gonzalez, P., Fensham, R., Zhang, Z., Castro, J., Demidova, N., Lim, J.-H., Allard, G., Running, S.W., Semerci, A., and Cobb, N. 2010. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. For Ecol Manage 259:660-684.

AMBORELLA GENOME PROJECT, Albert, V.A., Barbazuk, W.B., dePamphilis, C.W., Der, J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C., Soltis, D.E., Soltis, P.S., Wessler, S.R., Wing, R.A., Albert, V.A., Ammiraju, J.S.S., Barbazuk, W.B., Chamala, S., Chanderbali, A.S., dePamphilis, C.W., Der, J.P., Determann, R., Leebens-Mack, J., Ma, H., Ralph, P., Rounsley, S., Schuster, S.C., Soltis, D.E., Soltis, P.S., Talag, J., Tomsho, L., Walts, B., Wanke, S., Wing, R.A., Albert, V.A., Barbazuk, W.B., Chamala, S., Chanderbali, A.S., Chang, T.-H., Determann, R., Lan, T., Soltis, D.E., Soltis, P.S., Arikit, S., Axtell, M.J., Ayyampalayam, S., Barbazuk, W.B., Burnette, J.M., Chamala, S., De Paoli, E., dePamphilis, C.W., Der, J.P., Estill, J.C., Farrell, N.P., Harkess, A., Jiao, Y., Leebens-Mack, J., Liu, K., Mei, W., Meyers, B.C., Shahid, S., Wafula, E., Walts, B., Wessler, S.R., Zhai, J., Zhang, X., Albert, V.A., Carretero-Paulet, L., dePamphilis, C.W., Der, J.P., Jiao, Y., Leebens-Mack, J., Lyons, E., Sankoff, D., Tang, H., Wafula, E., Zheng, C., Albert, V.A., Altman, N.S., Barbazuk, W.B., Carretero-Paulet, L., dePamphilis, C.W., Der, J.P., Estill, J.C., Jiao, Y., Leebens-Mack, J., Liu, K., Mei, W., Wafula, E., Altman, N.S., Arikit, S., Axtell, M.J., Chamala, S., Chanderbali, A.S., Chen, F., Chen, J.-Q., Chiang, V., De Paoli, E., dePamphilis, C.W., Der, J.P., Determann, R., Fogliani, B., Guo, C., Harholt, J., Harkess, A., Job, C., Job, D., Kim, S., Kong, H., Leebens-Mack, J., Li, G., Li, L., Liu, J., Ma, H., Meyers, B.C., Park, J., Qi, X., Rajjou, L., Burtet-Sarramegna, V., Sederoff, R., Shahid, S., Soltis, D.E., Soltis, P.S., Sun, Y.-H., Ulvskov, P., Villegente, M., Xue, J.-Y., Yeh, T.-F., Yu, X., Zhai, J., Acosta, J.J., Albert, V.A., Barbazuk, W.B., Bruenn, R.A., Chamala, S., de Kochko, A.,

588 dePamphilis, C.W., Der, J.P., Herrera-Estrella, L.R., Ibarra-Laclette, E., Kirst, M., Leebens-
589 Mack, J., Pissis, S.P., Poncet, V., Schuster, S.C., Soltis, D.E., Soltis, P.S., and Tomsho, L.
590 2013. The Amborella Genome and the Evolution of Flowering Plants. Science
591 342:1241089.
592 Andersen, E.J., Nepal, M.P., Purintun, J.M., Nelson, D., Mermigka, G., and Sarris, P.F. 2020.
593 Wheat Disease Resistance Genes and Their Diversification Through Integrated Domain
594 Fusions. Front Genet 11:898.
595 Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers,
596 E.W., Li, P.W., and Eichler, E.E. 2002a. Recent segmental duplications in the human
597 genome. Science 297:1003-1007.
598 Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers,
599 E.W., Li, P.W., and Eichler, E.E. 2002b. Recent Segmental Duplications in the Human
600 Genome. Science 297:1003-1007.
601 Bergman, M.E., Chávez, Á., Ferrer, A., and Phillips, M.A. 2020. Distinct metabolic pathways drive
602 monoterpenoid biosynthesis in a natural population of Pelargonium graveolens. Journal
603 of Experimental Botany 71:258-271.
604 Blanc, G., and Wolfe, K.H. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred
605 from Age Distributions of Duplicate Genes[W]. The Plant Cell 16:1667-1678.
606 Brosius, J. 1991. Retroposons—Seeds of Evolution. Science 251:753-753.
607 Buchfink, B., Reuter, K., and Drost, H.-G. 2021. Sensitive protein alignments at tree-of-life scale
608 using DIAMOND. Nature Methods 18:366-368.
609 Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. 2021.
610 eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain
611 Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825-5829.
612 Carretero-Paulet, L., Librado, P., Chang, T.-H., Ibarra-Laclette, E., Herrera-Estrella, L., Rozas, J.,
613 and Albert, V.A. 2015. High Gene Family Turnover Rates and Gene Space Adaptation in
614 the Compact Genome of the Carnivorous Plant Utricularia gibba. Molecular Biology and
615 Evolution 32:1284-1295.
616 Casola, C., and Koralewski, T.E. 2018. Pinaceae show elevated rates of gene turnover that are
617 robust to incomplete gene annotation. Plant J.
618 Cazzolla Gatti, R., Reich, P.B., Gamarra, J.G.P., Crowther, T., Hui, C., Morera, A., Bastin, J.-F., de-
619 Miguel, S., Nabuurs, G.-J., Svenning, J.-C., Serra-Diaz, J.M., Merow, C., Enquist, B.,
620 Kamenetsky, M., Lee, J., Zhu, J., Fang, J., Jacobs, D.F., Pijanowski, B., Banerjee, A.,
621 Giaquinto, R.A., Alberti, G., Almeyda Zambrano, A.M., Alvarez-Davila, E., Araujo-
622 Murakami, A., Avitabile, V., Aymard, G.A., Balazy, R., Baraloto, C., Barroso, J.G., Bastian,
623 M.L., Birnbaum, P., Bitariho, R., Bogaert, J., Bongers, F., Bouriaud, O., Brancalion, P.H.S.,
624 Brearley, F.Q., Broadbent, E.N., Bussotti, F., Castro da Silva, W., César, R.G., Češljar, G.,
625 Chama Moscoso, V., Chen, H.Y.H., Cienciala, E., Clark, C.J., Coomes, D.A., Dayanandan,
626 S., Decuyper, M., Dee, L.E., Del Aguila Pasquel, J., Derroire, G., Djuikouo, M.N.K., Van Do,
627 T., Dolezal, J., Đorđević, I.Đ., Engel, J., Fayle, T.M., Feldpausch, T.R., Fridman, J.K., Harris,
628 D.J., Hemp, A., Hengeveld, G., Herault, B., Herold, M., Ibanez, T., Jagodzinski, A.M.,
629 Jaroszewicz, B., Jeffery, K.J., Johannsen, V.K., Jucker, T., Kangur, A., Karminov, V.N.,
630 Kartawinata, K., Kennard, D.K., Kepfer-Rojas, S., Keppel, G., Khan, M.L., Khare, P.K.,
631 Kileen, T.J., Kim, H.S., Korjus, H., Kumar, A., Kumar, A., Laarmann, D., Labrière, N., Lang,

632 M., Lewis, S.L., Lukina, N., Maitner, B.S., Malhi, Y., Marshall, A.R., Martynenko, O.V.,
633 Monteagudo Mendoza, A.L., Ontikov, P.V., Ortiz-Malavasi, E., Pallqui Camacho, N.C.,
634 Paquette, A., Park, M., Parthasarathy, N., Peri, P.L., Petronelli, P., Pfautsch, S., Phillips,
635 O.L., Picard, N., Piotto, D., Poorter, L., Poulsen, J.R., Pretzsch, H., Ramírez-Angulo, H.,
636 Restrepo Correa, Z., Rodeghiero, M., Rojas Gonzáles, R.D.P., Rolim, S.G., Rovero, F.,
637 Rutishauser, E., Saikia, P., Salas-Eljatib, C., Schepaschenko, D., Scherer-Lorenzen, M.,
638 Šebeň, V., Silveira, M., Slik, F., Sonké, B., Souza, A.F., Stereńczak, K.J., Svoboda, M.,
639 Taedoumg, H., Tchebakova, N., Terborgh, J., Tikhonova, E., Torres-Lezama, A., van der
640 Plas, F., Vásquez, R., Viana, H., Vibrans, A.C., Vilanova, E., Vos, V.A., Wang, H.-F.,
641 Westerlund, B., White, L.J.T., Wiser, S.K., Zawiła-Niedźwiecki, T., Zemagho, L., Zhu, Z.-X.,
642 Zo-Bi, I.C., and Liang, J. 2022. The number of tree species on Earth. Proceedings of the
643 National Academy of Sciences 119:e2115329119.
644 Chen, S., Zhou, Y., Chen, Y., and Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor.
645 Bioinformatics 34:i884-i890.
646 Conant, G.C., Birchler, J.A., and Pires, J.C. 2014. Dosage, duplication, and diploidization:
647 clarifying the interplay of multiple models for duplicate gene evolution over time.
648 Current Opinion in Plant Biology 19:91-98.
649 Cortés, A.J., Restrepo-Montoya, M., and Bedoya-Canas, L.E. 2020. Modern Strategies to Assess
650 and Breed Forest Tree Adaptation to Changing Climate. Front Plant Sci 11:583323.
651 Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson,
652 J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis, C.W. 2006.
653 Widespread genome duplications throughout the history of flowering plants. Genome
654 Res 16:738-749.
655 de Bruijn, W.J.C., Gruppen, H., and Vincken, J.-P. 2018. Structure and biosynthesis of
656 benzoxazinoids: Plant defence metabolites with potential as antimicrobial scaffolds.
657 Phytochemistry 155:233-243.
658 De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van de Peer, Y.
659 2013. Convergent gene loss following gene and genome duplications creates single-copy
660 families in flowering plants. Proceedings of the National Academy of Sciences 110:2898-
661 2903.
662 Demuth, J.P., and Hahn, M.W. 2009. The life and death of gene families. Bioessays 31:29-39.
663 Diss, G., Gagnon-Arsenault, I., Dion-Coté, A.-M., Vignaud, H., Ascencio, D.I., Berger, C.M., and
664 Landry, C.R. 2017. Gene duplication can impart fragility, not robustness, in the yeast
665 protein interaction network. Science 355:630-634.
666 Dodds, P.N., and Rathjen, J.P. 2010. Plant immunity: towards an integrated view of plant–
667 pathogen interactions. Nature Reviews Genetics 11:539-548.
668 Emms, D.M., and Kelly, S. 2017. STRIDE: Species Tree Root Inference from Gene Duplication
669 Events. Mol Biol Evol 34:3267-3278.
670 Emms, D.M., and Kelly, S. 2018. STAG: Species Tree Inference from All Genes (bioRxiv).
671 Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale
672 detection of protein families. Nucleic Acids Research 30:1575-1584.
673 Falcone Ferreyra, M.L., Rius, S., and Casati, P. 2012. Flavonoids: biosynthesis, biological
674 functions, and biotechnological applications. Frontiers in Plant Science 3.
675 FAO. 2020. Global Forest Resources Assessment 2020: Main report. Rome.

676 Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation
677 of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-
678 1545.
679 Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-
680 generation sequencing data. Bioinformatics 28:3150-3152.
681 Gamfeldt, L., Snäll, T., Bagchi, R., Jonsson, M., Gustafsson, L., Kjellander, P., Ruiz-Jaen, M.C.,
682 Fröberg, M., Stendahl, J., Philipson, C.D., Mikusiński, G., Andersson, E., Westerlund, B.,
683 Andrén, H., Moberg, F., Moen, J., and Bengtsson, J. 2013. Higher levels of multiple
684 ecosystem services are found in forests with more tree species. Nature Communications
685 4:1340.
686 Gao, Y., Wang, W., Zhang, T., Gong, Z., Zhao, H., and Han, G.-Z. 2018. Out of Water: The Origin
687 and Early Diversification of Plant R-Genes. Plant Physiology 177:82-89.
688 Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M.,
689 Talón, M., Dopazo, J., and Conesa, A. 2008. High-throughput functional annotation and
690 data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420-3435.
691 Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C., and Cristianini, N. 2005. Estimating the tempo
692 and mode of gene family evolution from comparative genomic data. Genome Res
693 15:1153-1160.
694 Hammond, W.M., Williams, A.P., Abatzoglou, J.T., Adams, H.D., Klein, T., López, R., Sáenz-
695 Romero, C., Hartmann, H., Breshears, D.D., and Allen, C.D. 2022. Global field
696 observations of tree die-off reveal hotter-drought fingerprint for Earth's forests. Nature
697 Communications 13:1761.
698 Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S.H. 2008. Importance of lineage-
699 specific expansion of plant tandem duplicates in the adaptive response to
700 environmental stimuli. Plant Physiol 148:993-1003.
701 Hurley, B.P., Garnas, J., Wingfield, M.J., Branco, M., Richardson, D.M., and Slippers, B. 2016.
702 Increasing numbers and intercontinental spread of invasive insects on eucalypts.
703 Biological Invasions 18:921-933.
704 Isabel, N., Holliday, J.A., and Aitken, S.N. 2020. Forest genomics: Advancing climate adaptation,
705 forest health, productivity, and conservation. Evolutionary Applications 13:3-10.
706 Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable
707 elements mediate gene evolution in plants. Nature 431:569-573.
708 Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M.,
709 Ruzicka, D.R., Wafula, E., Wickett, N.J., Wu, X., Zhang, Y., Wang, J., Zhang, Y., Carpenter,
710 E.J., Deyholos, M.K., Kutchan, T.M., Chanderbali, A.S., Soltis, P.S., Stevenson, D.W.,
711 McCombie, R., Pires, J.C., Wong, G.K., Soltis, D.E., and Depamphilis, C.W. 2012. A
712 genome triplication associated with early diversification of the core eudicots. Genome
713 Biol 13:R3.
714 Johri, P., Gout, J.-F., Doak, T.G., and Lynch, M. 2022. A Population-Genetic Lens into the Process
715 of Gene Loss Following Whole-Genome Duplication. Molecular Biology and Evolution
716 39:msac118.
717 Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J.,
718 Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M.,

Yong, S.Y., Lopez, R., and Hunter, S. 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30:1236-1240.

Kassambara, A. 2020. ggpubr: 'ggplot2' Based Publication Ready Plots. https://CRAN.R-project.org/package=ggpubr.

Katoh, K., Misawa, K., Kuma, K.i., and Miyata, T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30:3059-3066.

Kirilenko, A.P., and Sedjo, R.A. 2007. Climate change impacts on forestry. Proceedings of the National Academy of Sciences 104:19697-19702.

Kolde, R. 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12, https://CRAN.R-project.org/package=pheatmap.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution 34:1812-1819.

Kuzmin, E., VanderSluis, B., Nguyen Ba, A.N., Wang, W., Koch, E.N., Usaj, M., Khmelinskii, A., Usaj, M.M., van Leeuwen, J., Kraus, O., Tresenrider, A., Pryszlak, M., Hu, M.-C., Varriano, B., Costanzo, M., Knop, M., Moses, A., Myers, C.L., Andrews, B.J., and Boone, C. 2020. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science 368:eaaz5667.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M., Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J., DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B., Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M., Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington, S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L., Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D., Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P., Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J., Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden, G.T., Goh, F., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K., Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B., Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K., Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Lowe, A.J., Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T., McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W., Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.-M., Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C., Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A., Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S.G., Wenzel, M., Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D., Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen, L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z., Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y., Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N.-p., Matasci, N., Nelson, D.R., Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E., Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H., Lu-

763 Irving, P., Marx, H.E., Qi, X., Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu,
764 R., Zhan, S., Gramzow, L., Theißen, G., Wong, G.K.-S., and One Thousand Plant
765 Transcriptomes, I. 2019. One thousand plant transcriptomes and the phylogenomics of
766 green plants. Nature 574:679-685.
767 Leister, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of
768 plant disease resistance gene. Trends Genet 20:116-122.
769 Li, B., and Dewey, C.N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with
770 or without a reference genome. BMC Bioinformatics 12:323.
771 Linnakoski, R., and Forbes, K.M. 2019. Pathogens—The Hidden Face of Forest Invasions by
772 Wood-Boring Insect Pests. Frontiers in Plant Science 10.
773 Liu, J., Yu, T., Mu, Z., and Li, G. 2019. TransLiG: a de novo transcriptome assembler that uses line
774 graph iteration. Genome Biology 20:81.
775 Lu, M., Feau, N., Vidakovic, D.O., Ukrainetz, N., Wong, B., Aitken, S.N., Hamelin, R.C., and
776 Yeaman, S. 2021. Comparative Gene Expression Analysis Reveals Mechanism of Pinus
777 contorta Response to the Fungal Pathogen Dothistroma septosporum. Mol Plant
778 Microbe Interact:Mpmi10200282r.
779 Lynch, M., and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes.
780 Science 290:1151-1155.
781 Lynch, M., and Conery, J.S. 2003. The evolutionary demography of duplicate genes. Journal of
782 Structural and Functional Genomics 3:35-44.
783 Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y.
784 2005. Modeling gene and genome duplications in eukaryotes. Proceedings of the
785 National Academy of Sciences 102:5454-5459.
786 Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. 2021. BUSCO Update:
787 Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic
788 Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology
789 and Evolution 38:4647-4654.
790 Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. 2020. CAFE 5 models variation in
791 evolutionary rates among gene families. Bioinformatics 36:5516-5518.
792 Meng, X., and Zhang, S. 2013. MAPK cascades in plant disease resistance signaling. Annu Rev
793 Phytopathol 51:245-266.
794 Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. 2003. Genome-Wide
795 Analysis of NBS-LRR–Encoding Genes in Arabidopsis[W]. The Plant Cell 15:809-834.
796 Miura, S., Amacher, M., Hofer, T., San-Miguel-Ayanz, J., Ernawati, and Thackway, R. 2015.
797 Protective functions and ecosystem services of global forests in the past quarter-
798 century. Forest Ecology and Management 352:35-46.
799 Neale, D.B., and Kremer, A. 2011. Forest tree genomics: growing resources and applications.
800 Nat Rev Genet 12:111-122.
801 Ngou, B.P.M., Heal, R., Wyler, M., Schmid, M.W., and Jones, J.D.G. 2022. Concerted expansion
802 and contraction of immune receptor gene repertoires in plant genomes. Nature Plants
803 8:1146-1152.
804 Ohno, S. 1970. Evolution by gene duplication (Springer-Verlag).
805 Panchy, N., Lehti-Shiu, M., and Shiu, S.H. 2016. Evolution of Gene Duplication in Plants. Plant
806 Physiol 171:2294-2316.

Piasecka, A., Jedrzejczak-Rey, N., and Bednarek, P. 2015. Secondary metabolites in plant innate immunity: conserved function of divergent chemicals. New Phytologist 206:948-964.

Picart-Picolo, A., Grob, S., Picault, N., Franek, M., Llauro, C., Halter, T., Maier, T.R., Jobet, E., Descombin, J., Zhang, P., Paramasivan, V., Baum, T.J., Navarro, L., Dvořáčková, M., Mirouze, M., and Pontvianne, F. 2020. Large tandem duplications affect gene expression, 3D organization, and plant-pathogen response. Genome Res 30:1583-1592.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641-1650.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Reid, N.M., Proestou, D.A., Clark, B.W., Warren, W.C., Colbourne, J.K., Shaw, J.R., Karchner, S.I., Hahn, M.E., Nacci, D., Oleksiak, M.F., Crawford, D.L., and Whitehead, A. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science 354:1305-1308.

Rizvi, A.R., Baig, S., Barrow, E., and Kumar, C. 2015. Synergies between climate mitigation and adaptioon in forest landscape restoration. Gland, Switzerland: IUCN.

Sanderson, M.J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301-302.

Seoighe, C., and Gehring, C. 2004. Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends in Genetics 20:461-464.

Soltis, P.S., and Soltis, D.E. 2016. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol 30:159-165.

Swenson, N.G., and Jones, F.A. 2017. Community transcriptomics, genomics and the problem of species co-occurrence. Journal of Ecology 105:563-568.

Swenson, N.G., Iida, Y., Howe, R., Wolf, A., Umaña, M.N., Petprakob, K., Turner, B.L., and Ma, K. 2017. Tree co-occurrence and transcriptomic response to drought. Nature Communications 8:1996.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. 2008. Synteny and Collinearity in Plant Genomes. Science 320:486-488.

Trouern-Trend, A.J., Falk, T., Zaman, S., Caballero, M., Neale, D.B., Langley, C.H., Dandekar, A.M., Stevens, K.A., and Wegrzyn, J.L. 2020. Comparative genomics of six Juglans species reveals disease-associated gene family contractions. The Plant Journal 102:410-423.

Van de Peer, Y., Maere, S., and Meyer, A. 2009. The evolutionary significance of ancient genome duplications. Nature Reviews Genetics 10:725-732.

Wickham, H. 2011. The Split-Apply-Combine strategy for data analysis. J. Stat. Softw. 40:1-29.

Wickham, H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York, NY, US.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586-1591.

Yang, Z., and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32-43.

Yu, Y. 2022. mixR: An R package for finite mixture modeling for both raw and binned data. Journal of Open Source Software 7:4031.

850 Zeng, L., Zhang, N., Zhang, Q., Endress, P.K., Huang, J., and Ma, H. 2017. Resolution of deep
851 eudicot phylogeny and their temporal diversification using nuclear genes from
852 transcriptomic and genomic datasets. New Phytol 214:1338-1354.
853 Zhang, J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-
854 298.
855 Zhao, J., Davis, L.C., and Verpoorte, R. 2005. Elicitor signal transduction leading to production of
856 plant secondary metabolites. Biotechnol Adv 23:283-333.
857 Zwaenepoel, A., and Van de Peer, Y. 2019. Inference of Ancient Whole-Genome Duplications
858 and the Evolution of Gene Duplication and Loss Rates. Molecular Biology and Evolution
859 36:1384-1404.
860