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, USA 46556

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

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

[*]Correspondance: Email: mlu@nd.edu

Funding: This research was supported by the National Science Foundation (NSF DEB-2124466) and NSFC China-US Dimensions of Biodiversity Grant (DEB-32061123003).

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 have 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, and a signal of recent duplication events specific to particular species. We also found rapid evolution, namely fast expansion or fast contraction of gene families, in plant-pathogen interaction genes amongst the diploid species studied. 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.

Key words: angiosperm; 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 they supply other organisms (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 to our understanding of the abiotic and biotic factors governing the distribution and dynamics of species populations, but also 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), where 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). 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.

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.

Material and methods

Samples, sequencing data, de novo assembly, and annotation

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. (Figure 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, belong to the clades of asterids, saxifragles, and rosids. While these species all co-occur at the UNDERC site, their geographic ranges largely overlap on the continental scale and the species have largely co-occurred in North America since at least the Pleistocene. Upon collection, the samples were flash frozen in liquid nitrogen, then transported and stored in an ultracold freezer at the field station. The samples were sent to Novogene corporation inc. (Chula Vista, CA, USA) for mRNA extraction, library preparation, mRNA sequencing, and quality control. The samples were sequenced on an Illumina NovoSeq 6000 platform to produce 150bp paired-end reads.

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

Raw reads filtering was performed by Novogene to get the clean reads including removing reads containing adapters, removing reads containing N >10%, and removing reads containing low quality base which is over 50% of the total base. An average of 6Gb paired-end clean reads were acquired for each sample and deposited in NCBI Sequence Read Archive (SRA accession number: PRJNA896247). *De novo* transcriptome assembly was performed for each sample using the software TransLiG v1.3 with default parameters (Liu et al., 2019). The clean reads were then mapped to the assembled transcriptome using the software RSEM v1.2.25 with default parameters (Li and Dewey, 2011). The longest transcript (isoform) with a transcript per million

170 (TPM) value ≥ 1.5 for each gene was retained. Redundancy was removed from the retained
171 transcripts using the software CD-HIT v4.8.1 with parameters of -c 0.95 -n 8 -T 32 -M 1540 (Fu
172 et al., 2012). To remove potential contamination, the retained transcripts were aligned against the
173 non-redundant (nr) protein database (downloaded on January 22, 2022) using the blastx function
174 implemented by the software DIAMOND v2.0.13.151 (Buchfink et al., 2021) with parameter of -
175 -outfmt 102 to assign the best known taxon ID to each query transcript. The file that was used to
176 convert accession numbers to taxonomy was downloaded from
177 https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/, and the tax ID file was downloaded
178 from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/. The query transcripts that could
179 be assigned to green plant (Viridiplantae) taxon IDs were retained. The filtered transcripts were
180 aligned to the nr database that only includes green plant subjects using the blastx function
181 implemented by the software DIAMOND v2.0.13.151 with default parameters. The aligned
182 transcripts were also retained and then were combined with the previously retained transcripts to
183 form the final transcriptomes. The completeness of the final transcriptomes was examined using
184 the 1,614 orthologues in the BUSCO (Benchmarking Universal Single Copy Orthologs) v5.2.2
185 set of embryophyta_odb10 (Manni et al., 2021).
186
187 The transcriptomes were annotated with Gene Ontology (GO) and Kyoto Encyclopedia of Genes
188 and Genomes (KEGG) terms. The transcripts were aligned to the nr database that only includes
189 green plant subjects using the blastx function implemented by the software DIAMOND
190 v2.0.13.151 with parameters of -taxonlist 33090 --very-sensitive -f5. The .xml outputs were
191 used as input for mapping and annotation using the software Blast2GO (Götz et al., 2008). The
192 transcripts were translated to protein sequences using the software TransDecoder v5.5.0
193 (https://github.com/TransDecoder/TransDecoder/wiki) with default parameters to get the longest
194 ORF (open reading frame) per transcript, and then were annotated by the software InterProScan
195 v5.54-87.0 (Jones et al., 2014) with the parameters of -dra -f XML -goterms -pa. The final GO
196 terms were acquired by merging Blast2GO and InterProScan annotation outputs using the
197 software Blast2GO. To get the KEGG terms, the longest ORF per transcript were also annotated
198 using EggNOG-mapper v2.1.7 (Cantalapiedra et al., 2021) with parameters of -m diamond -
199 sensmode more-sensitive -tax_scope 33090 -tax_scope_mode 33090.
200
201 *Ks analyses*
202 The level of synonymous substitutions (Ks) between two homologous sequences increases
203 approximately linearly with time. Thus pairs of paralogous genes within a species can be sorted
204 to estimate the relative ages of gene duplication (Blanc and Wolfe, 2004). To identify paralogous
205 gene pairs within a species, an all-versus-all blastp was performed within each species. The
206 peptide sequences (the longest ORF per transcript) of each species were used as input to run all-
207 versus-all blastp using the software DIAMOND v2.0.13.151 with parameters of -p 16 --max-
208 target-seqs 20 --evalue 1e-10 --outfmt 6 qseqid sseqid pident nident evalue. The blastp output
209 was filtered by removing self blastp results, keeping outputs with pident (percent of identical
210 matches) ≥20 and nident (number of identical matches) ≥50, and removing sequences with more
211 than 10 hits. The filtered paralogous gene pairs were aligned using the software MAFFT v7.490
212 (Katoh et al., 2002) with default parameters. The alignments were used as input for estimating
213 the number of synonymous substitutions per synonymous site (Ks) and the number of non-
214 synonymous substitutions per nonsynonymous site (Ka) using the program yn00 in the software

PAML v4.9j (Yang, 2007) with parameters of icode=0 weighting=0 commonf3×4=0. The Ks of each paralogous gene pairs were acquired using the method by Yang and Nielsen (2000).

Ks can also be calculated using the orthologous gene pairs between species to estimate speciation events (Blanc and Wolfe, 2004). To identify orthologous gene pairs between species, reciprocal best hit blastp were run between the compared pairs. The peptide sequences (the longest ORF per transcript) of the paired species were used as input to run blastp using the software DIAMOND v2.0.13.151 with parameters of -p 16 --max-target-seqs 20 --evalue 1e-10 --outfmt 6 qseqid sseqid pident nident evalue. 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 and Ks as abovementioned.

The Ks distribution is proportional to the time lapse. Thus, the Ks distribution within species can be used to infer the timing of WGD, and the Ks distribution between two different species can infer time since speciation (Blanc and Wolfe, 2004). To plot the Ks distribution, the mixture model was fit and plot for the Ks values ≥0.01 and ≤2 using the R package mixR (Yu, 2022). In case there are discernably separate peaks (multiple components), bootstrap likelihood ratio test (LRT) was used to select the models.

Estimating duplication events and divergence time

To infer duplication events, the species tree and gene trees were first inferred using STAG and STRIDE algorithms implemented by the program OrthoFinder v2.5.4 (Emms and Kelly, 2017, 2018). The inferred species tree was similar to that expected given current knowledge of the Angiosperm phylogeny with the exception of the placement of *Populus tremuloides* and *P. grandidentata*. Thus, for downstream analysis and discussion, we removed these two species for the species tree inference. The inferred species tree and gene trees were used to estimate duplication events and divergence time. The gene duplication events on the species tree node were reported if, at least 50% of the descendant species have retained both copies of the duplicated gene. The divergence time was estimated using the software r8s v1.81 (Sanderson, 2003) with calibration points of 115 million years (MYA) between *Fraxinus nigra* and *Corylus cornuta*, which were obtained from http://www.timetree.org/ (Kumar et al., 2017).

Gene family and their evolutionary rate

To infer gene families, the peptide sequences (the longest ORF per transcript) of the 20 study 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 20 study 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 study 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). We studied the relationship between gene family size and evolutionary rate, which is represented by mean Ka/Ks per gene family, using scatter plot. All statistical analyses were performed in R (R Core Team, 2021). Gene families with a family size ≥5, and Ka/Ks ≥1.5 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 study 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 λ (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 λ, which represents the probability that any gene will be gained or lost. Multiple λ 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 evolution 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 (Figure 5b). 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

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 (Table S1). We retrieved 76 ~ 88 % of the BUSCOs in the assembled transcriptomes (Figure S1).

We found a large Ks peak around 0.71~0.75 shared by all the study species (Figure S2), and discernable separate peaks on the left of the major peak can be observed from Ks plots of *Amelanchier laevis*, *Dirca palustris*, *Fraxinus nigra*, *Populus grandidentata*, *Prunus serotina*, *Populus tremuloides*, *Prunus virginiana*, and *Tilia americana* (Figure 2). Bootstrap LRT indicates that considering these small peaks as a separate component could fit the mixture model better. The shared large peaks by all species may correspond to large-scale duplication events, perhaps the common WGDs across core eudicots, and the small peaks may represent recent duplication events specific to certain species. The two *Prunus* species have the small peaks with similar shapes at a similar position (Ks=0.10 or 0.11) and so do the two *Populus* species (Ks=0.12). This pattern may represent a common WGD before the two species diverged from the same ancestors. By examining the orthologous Ks between the two species, we found a high peak closer to zero, which may represent the duplication events after the WGDs and divergence.

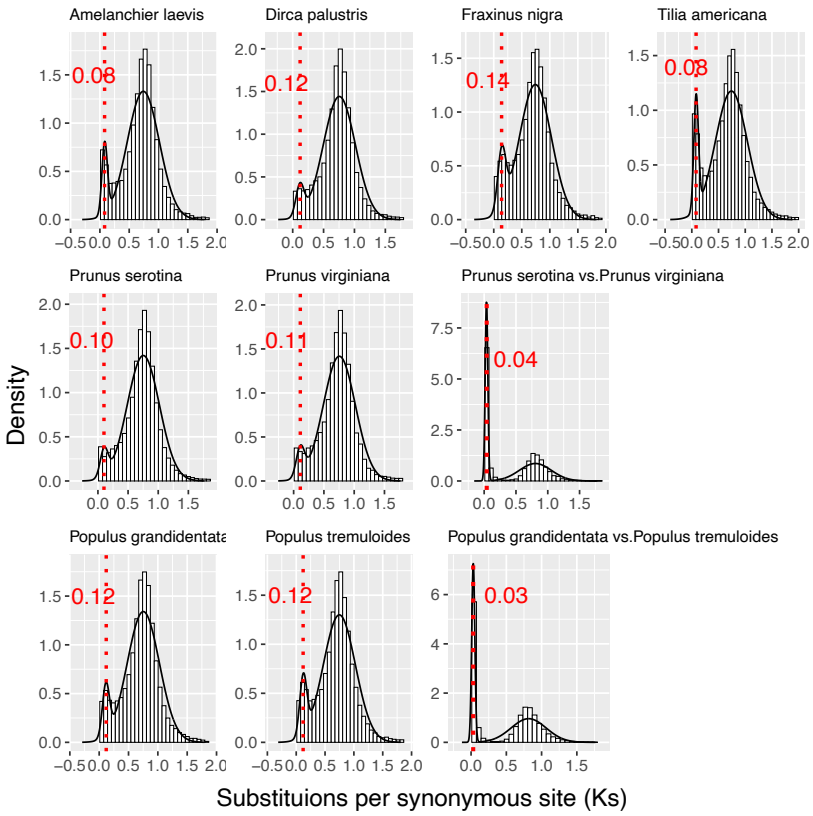

Figure 2. The Ks distributions for the studied species with discernable separate peaks on the left of the major peak.

Using a species tree with estimated branch lengths, we inferred divergence times and duplication events (with at least 50 % support rate, Figure 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 (Figure S3), 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 (Figure 3). The divergence time of the studied malvids and fabids from a common ancestor with Saxifragles was about 100 MYA.

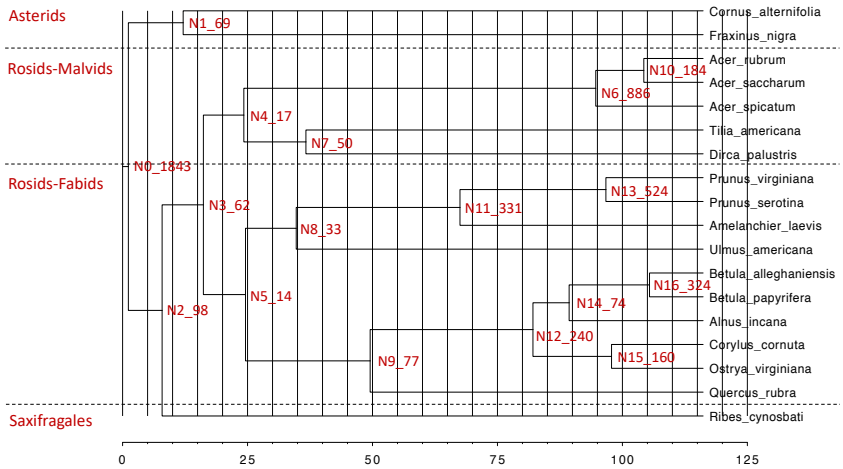

Figure 3. Divergence time in million years and gene duplication events with 50 % support rate. Node numbers were followed by number of gene duplication events.

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. Most gene families have a size <5, only 1.8~5.9 % of the total gene families across all species have a size ≥5 (Figure 4a). 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 ≥1.5 in red (Figure 4a). We tested the KEGG and GO terms that are enriched in these highlighted gene families (Figures 4b &4c & S4, see the whole list of GO enrichment in Table S2). The most highly shared enriched KEGG and GO terms are related to plant-pathogen interactions and secondary metabolite biosynthesis. Across all species, percentage of gene families with a size ≥5 and Ka/Ks ≥1.5 is 4.4 % among plant-pathogen interaction gene families, and is 0.7 % among all gene families.

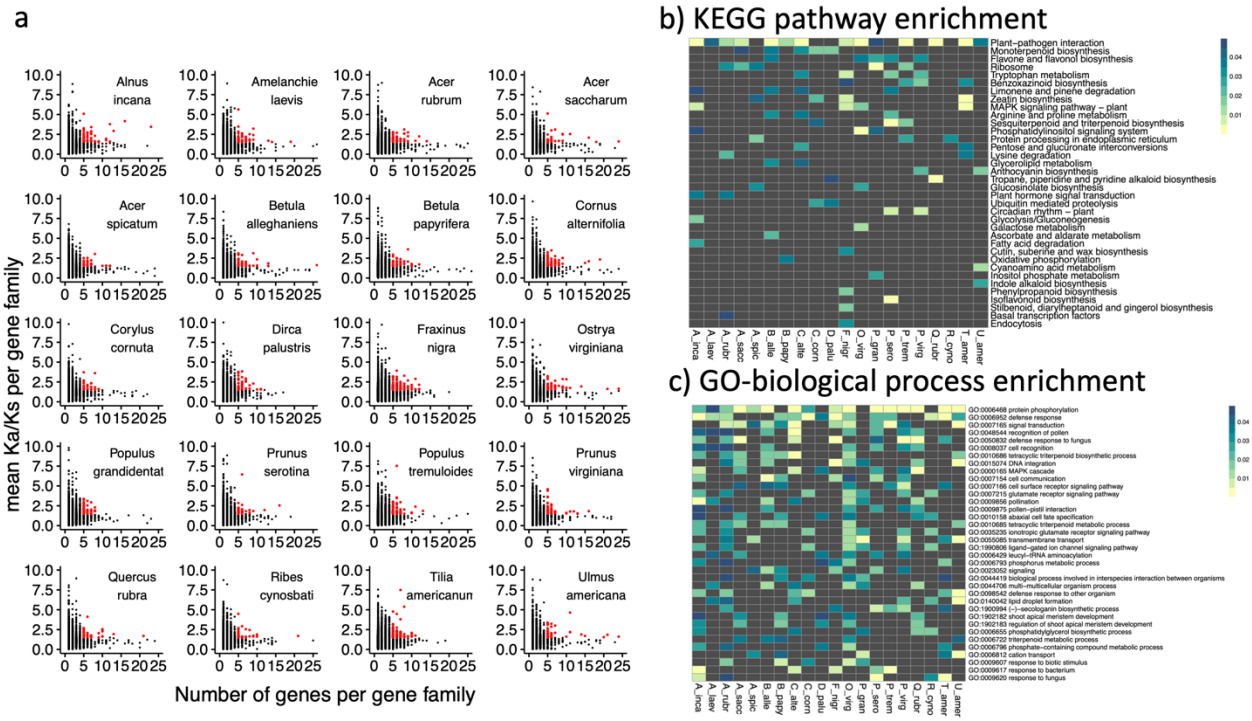

Figure 4. Scatter plot between gene size and Ka/Ks per gene family (a), KEGG pathway (b) and GO terms enrichment (c) in gene families with a size ≥5 and Ka/Ks ≥1.5.

We inferred the birth-death parameter, λ (number of gene gains and losses per gene per million years), which is specified in different clades, using the program CAFE. $\lambda 1$ (saxifragales) =0.0034, $\lambda 2$ (rosids) =0.0042, $\lambda 3$ (asterids) =0.0027, which means that the lineage leading to rosids have the highest gene family turnover rates, followed by saxifragales, and then by asterids. The four species with recent duplication events (Figure 2) were *Fraxinus nigra*, *Dirca palustris*, *Tilia americana*, and *Amelanchier laevis*. Specifically, these species have more expanding than contracting gene families (Figure 5a, S5). All of the study species have more rapidly expanding than rapidly contracting gene families (family-wide p-value <0.01, Figure 5b), which may contribute to their lineage-specific novelty.

We tested the enriched GO and KEGG terms in the fast evolving (expanding or contracting) gene families. All of the study species shared the enriched KEGG pathway term plant-pathogen interaction (Figure 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 (Figure S6) related to disease resistance protein, leucine-rich repeat receptor like protein kinase and others. Some species shared the enriched KEGG pathway term Tryptophan metabolism and MAPK signaling pathway (Figure 6).

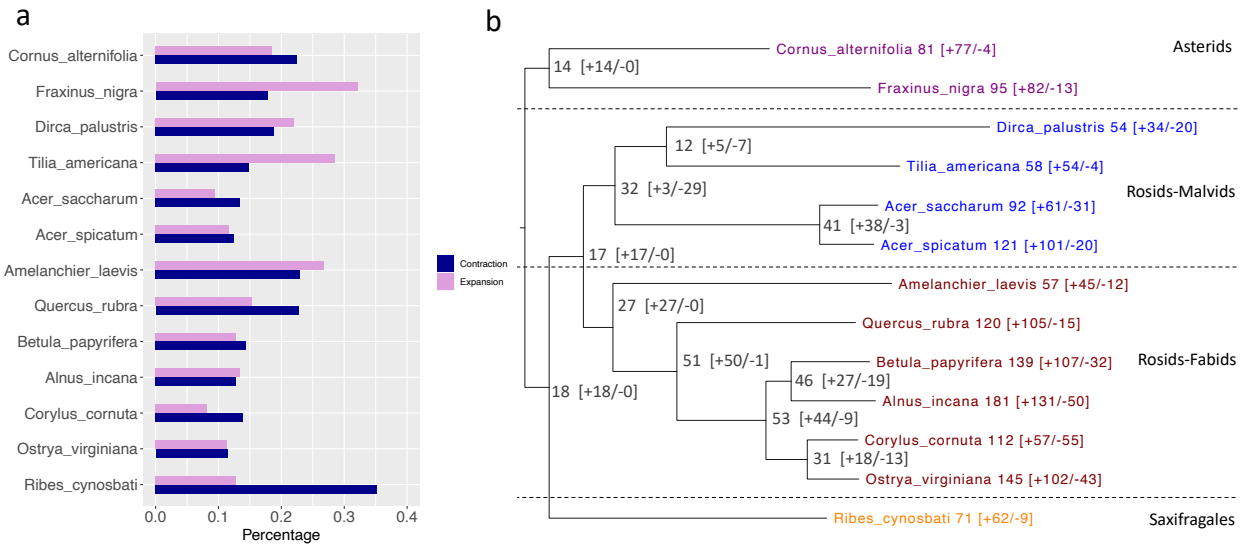

Figure 5. The percentage of gene family expansion and contraction (a) and the number of fast evolving gene families (b), with + representing fast expanding gene family and – representing fast contracting gene family.

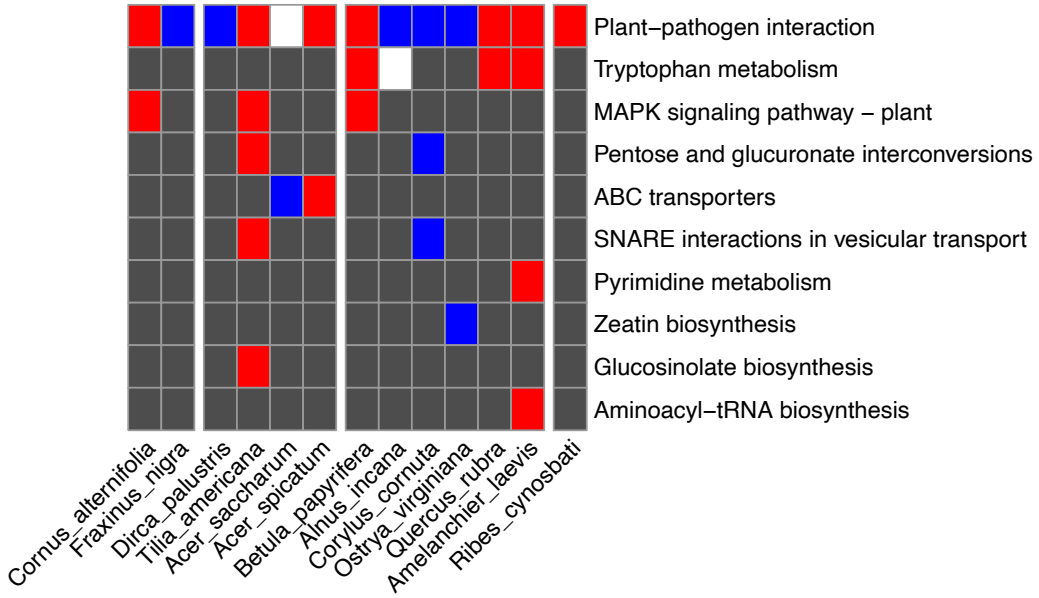

Figure 6. Enriched KEGG pathway terms in the study diploid species. Red represents fast expanding gene families; blue represents fast contracting gene families; white represents both fast expanding and contracting gene families.

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 and metabolite biosynthesis actively remain extensive variation, and fast expansion or fast contraction in plant-pathogen interaction genes amongst the study 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). All the study species shared a duplication event as seen by the large Ks peaks around 0.71~0.75 (Figure S2), which probably correspond to the common WGDs across core eudicots. Besides the large peak, some species had a discernable recent duplication event as seen by the small peaks on the left of the major peak (Figure 2). These abundant duplication events provide resources for genetic variation, although most duplicate genes were lost by the process of deletion of the entire duplicate sequence and/or pseudogenization through loss-of-function mutations, the retained duplicate genes lead to lineage-specific innovation (Panchy et al., 2016). Earlier studies placed at least three WGD events in angiosperm radiation, the oldest WGD occurred before the monocot-eudicot divergence, a second WGD occurred specifically to eudicots, a more recent WGD, gamma (γ) triplication event, occurred before the separation of rosids and asterids (Tang et al., 2008; Jiao et al., 2012). Besides, the lineage-specific WGDs and numerous SSDs continuously lead to new and complex gene contents. 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).

The predicted species tree (Figure 3) generally agrees with current phylogenetic hypotheses for the angiosperms with the exception of the two *Populus* species, which is probably caused by the incomplete gene sets obtained from one leaf RNA-seq sample and/or interspecific gene flow. We identified the most (1,843) duplication events at the root (Figure 3) of the tree, which is likely to be related to the gamma WGD before the divergence of rosids and asterids. Such large-scale duplication may contribute to genome structure variation, organismal complexity, and rapid speciation in angiosperm species. The parental node for *Acer*, *Prunus*, *Populus* species (Figure 3 & S3) also have a large number of duplication events. The most-likely scenario is WGD event occurred in the common ancestor, followed by different evolutionary events. For *Acer*, *Acer spicatum* experienced a diploidization event to revert to diploid status, while *Acer rubrum* maintained a polyploid status. Similarly, the recent Ks peaks on the similar location in two *Prunus* species and two *Populus* species are in line with their large number of duplication events in parental nodes.

457 Divergence time mostly agrees with the previous reports, which stated rosids and asterids
458 diverged about 125 MYA, malvids and fabids diverged about 110 MYA (Zeng et al., 2017).
459 However, divergence estimate in the present study is rather rough. Further studies need to
460 consider using dense taxon sampling, diverse markers like nuclear internal transcribed spacer
461 (ITS) or chloroplast DNA fragments, and molecular dating with fossil corrections methods to
462 better resolve the evolutionary relationships.
463
464 *Sustained large size gene families*
465 A small portion of gene families can actively maintain duplicate genes. Gene family evolution is
466 dynamic process resulting from gene or genome duplication events where most of these
467 duplicate genes are eventually eliminated or pseudonized (Panchy et al., 2016). The extreme
468 example is diploidization of paleopolyploidy species, which revert to one copy of genes after
469 WGDs. Nonetheless, some duplicate genes can be retained, perhaps due to the gene balance and
470 dosage effect (Birchler and Veitia, 2012). Alternatively, selection favors the neofunctionalization
471 in the duplicate genes, as these duplicate genes help organisms adapt to variable environments
472 (Cortesi et al., 2015). Recent studies explained retention of duplicate genes as some paralogous
473 genes depend on or compensate for their cognate copy (Diss et al., 2017), while some paralogous
474 gene pairs are stuck in a functional and structural entanglement status over evolutionary time
475 scale (Kuzmin et al., 2020).
476
477 What are the main functions of these sustained large size gene families? They are mainly
478 involved in plant-pathogen interaction and metabolite biosynthesis (Figure 4b & 4c). This is
479 consistent with previous report that genes encoding functions like signal transduction and stress
480 response tend to have paralogs (Panchy et al., 2016), and genes related to the function like
481 transcription regulation and signal transduction had higher possibility to survive in multiple
482 rounds of genome duplication (Seoighe and Gehring, 2004). Plants constantly combat with
483 pathogens and pests. The major class of plant disease resistance (R) genes contain a nucleotide-
484 binding site (NBS) and C-terminal leucin-rich repeats (LRRs). Tandem and segmental gene
485 duplication explain much of the diversification among R genes (Meyers et al., 2003; Leister,
486 2004; Andersen et al., 2020), possibly because they have experienced frequent deletion and
487 duplication throughout evolution, and thus are highly variable (Hanada et al., 2008; Gao et al.,
488 2018). Plant secondary metabolites, such as monoterpenoid, flavonoid, benzoxazinoid, play
489 important roles in defense against predation and pathogens (Falcone Ferreyra et al., 2012; de
490 Bruijn et al., 2018; Bergman et al., 2020). The regulation of plant secondary metabolism requires
491 the intimation of signal transduction network that leads to activation of biosynthetic genes (Zhao
492 et al., 2005).
493
494 Gene duplication and deletion give rise to gene family expansion and contraction. The greater
495 number in gene family expansions than contractions (Figure 5a) in *Fraxinus nigra*, *Dirca*
496 *palustris*, *Tilia americana*, and *Amelanchier laevis* is associated with their recent duplication
497 events as shown in the Ks distribution (Figure 2). A previous study showed that estimates of
498 birth and death rate was consistent with the estimates based on number of recent gene duplicates
499 (Demuth and Hahn, 2009), thus it is reasonable to see the expansion numbers parallel the Ks
500 distributions. The high gene gain rate likely reflects the occurrence of multiple WGDs and
501 frequent SSDs (Carretero-Paulet et al., 2015). While for the other species with contraction
502 outnumbered expansion, the pattern is mainly caused by gene loss and retention of specific gene

503 families followed the common WGDs preceded their common ancestor (Figure S2). We found a
504 high gene family turnover rate in rosids, but we must caution that the unbalanced taxon sampling
505 in our study species might cause this high turnover value. Moreover, future studies should
506 consider distinguishing the branches with or without recent WGDs, as the branches with recent
507 WGDs tend to have a high gene turnover rate (Casola and Koralewski, 2018).
508
509 *Rapid evolution in plant-pathogen interaction genes*
510 We found rapid evolution of genes involved in plant-pathogen interaction in the adaptive
511 expansion and adaptive contraction gene families across the study species (Figure 6). The
512 specific expanding or contracting gene families involved in plant-pathogen interaction in each
513 species differed from each other, implying lineage-specificity in changed gene families (Figure
514 S6). Other rapidly evolving gene families are related with functions such as tryptophan
515 metabolism and MAPK signaling pathway (Figure 6). This result is consistent with previous
516 studies, which found that fast evolving gene families were mostly groups with functions related
517 with immune defense/stress response, metabolism, cell signaling, chemoreception, and
518 reproduction (Demuth and Hahn, 2009). Additionally, a study on *Juglans* (walnuts) found
519 notable contractions in gene families that are involved in disease resistance and other abiotic
520 stress response (Trouern-Trend et al., 2020). Both metabolism and MAPK signaling pathway are
521 associated with plant defense. MAPK cascades sense pathogen/microbe-associated molecular
522 patters and pathogen effectors, and thus play pivotal roles in signaling plant defense against
523 pathogen attack (Meng and Zhang, 2013). Plant secondary metabolites contribute to plant
524 immunity as antibiotics, or to control callose deposition and programmed cell death (Piasecka et
525 al., 2015). The biotic stress exerted by pathogen or pests is a major source of selection pressure
526 on the study trees, which drive the rapid radiation of diverse and large plant immune receptor
527 gene repertoires. Expansion and/or contraction of cell-surface receptor genes correlate with the
528 expansion and/or contraction of intracellular receptor genes (Ngou et al., 2022).
529
530 *Conclusions*
531 In summary, we found the common WGDs shared by all the study species with some species
532 having recent duplication events, which contribute to the gene family expansions. Gene families
533 involved in plant-pathogen interaction and metabolite biosynthesis actively remain extensive
534 variation and are fast expanding or contracting amongst the study species. This knowledge calls
535 for an emphasis of gene family size variation, i.e. copy number variation, when studying forest
536 genetic diversity, which will inform forest adaptive strategy. Future studies may focus on
537 classification of duplications in forest tree species, and investigate how the duplications give rise
538 to species-specific innovation and adaption.
539
540
541
542 **Data availability**
543 The raw RNA-seq data were deposited in NCBI SRA (accession number: PRJNA896247;
544 http://www.ncbi.nlm.nih.gov/sra). The assembled transcriptomes and their annotation were
545 deposited in Dryad. The scripts used for analyses were deposited in
546 https://github.com/Mengmeng-Lu.
547
548

549 **References**

550 Alexa, A., and Rahnenfuhrer, J. 2018. topGO: Enrichment Analysis for Gene Ontology. R
551 package version 2.34.0.
552 Allen, C.D., Macalady, A.K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M.,
553 Kitzberger, T., Rigling, A., Breshears, D.D., Hogg, E.H.T., Gonzalez, P., Fensham, R.,
554 Zhang, Z., Castro, J., Demidova, N., Lim, J.-H., Allard, G., Running, S.W., Semerci, A.,
555 and Cobb, N. 2010. A global overview of drought and heat-induced tree mortality reveals
556 emerging climate change risks for forests. For Ecol Manage 259:660-684.
557 AMBORELLA GENOME PROJECT, Albert, V.A., Barbazuk, W.B., dePamphilis, C.W., Der,
558 J.P., Leebens-Mack, J., Ma, H., Palmer, J.D., Rounsley, S., Sankoff, D., Schuster, S.C.,
559 Soltis, D.E., Soltis, P.S., Wessler, S.R., Wing, R.A., Albert, V.A., Ammiraju, J.S.S.,
560 Barbazuk, W.B., Chamala, S., Chanderbali, A.S., dePamphilis, C.W., Der, J.P.,
561 Determann, R., Leebens-Mack, J., Ma, H., Ralph, P., Rounsley, S., Schuster, S.C., Soltis,
562 D.E., Soltis, P.S., Talag, J., Tomsho, L., Walts, B., Wanke, S., Wing, R.A., Albert, V.A.,
563 Barbazuk, W.B., Chamala, S., Chanderbali, A.S., Chang, T.-H., Determann, R., Lan, T.,
564 Soltis, D.E., Soltis, P.S., Arikit, S., Axtell, M.J., Ayyampalayam, S., Barbazuk, W.B.,
565 Burnette, J.M., Chamala, S., De Paoli, E., dePamphilis, C.W., Der, J.P., Estill, J.C.,
566 Farrell, N.P., Harkess, A., Jiao, Y., Leebens-Mack, J., Liu, K., Mei, W., Meyers, B.C.,
567 Shahid, S., Wafula, E., Walts, B., Wessler, S.R., Zhai, J., Zhang, X., Albert, V.A.,
568 Carretero-Paulet, L., dePamphilis, C.W., Der, J.P., Jiao, Y., Leebens-Mack, J., Lyons, E.,
569 Sankoff, D., Tang, H., Wafula, E., Zheng, C., Albert, V.A., Altman, N.S., Barbazuk,
570 W.B., Carretero-Paulet, L., dePamphilis, C.W., Der, J.P., Estill, J.C., Jiao, Y., Leebens-
571 Mack, J., Liu, K., Mei, W., Wafula, E., Altman, N.S., Arikit, S., Axtell, M.J., Chamala,
572 S., Chanderbali, A.S., Chen, F., Chen, J.-Q., Chiang, V., De Paoli, E., dePamphilis, C.W.,
573 Der, J.P., Determann, R., Fogliani, B., Guo, C., Harholt, J., Harkess, A., Job, C., Job, D.,
574 Kim, S., Kong, H., Leebens-Mack, J., Li, G., Li, L., Liu, J., Ma, H., Meyers, B.C., Park,
575 J., Qi, X., Rajjou, L., Burtet-Sarramegna, V., Sederoff, R., Shahid, S., Soltis, D.E., Soltis,
576 P.S., Sun, Y.-H., Ulvskov, P., Villegente, M., Xue, J.-Y., Yeh, T.-F., Yu, X., Zhai, J.,
577 Acosta, J.J., Albert, V.A., Barbazuk, W.B., Bruenn, R.A., Chamala, S., de Kochko, A.,
578 dePamphilis, C.W., Der, J.P., Herrera-Estrella, L.R., Ibarra-Laclette, E., Kirst, M.,
579 Leebens-Mack, J., Pissis, S.P., Poncet, V., Schuster, S.C., Soltis, D.E., Soltis, P.S., and
580 Tomsho, L. 2013. The Amborella Genome and the Evolution of Flowering Plants.
581 Science 342:1241089.
582 Andersen, E.J., Nepal, M.P., Purintun, J.M., Nelson, D., Mermigka, G., and Sarris, P.F. 2020.
583 Wheat Disease Resistance Genes and Their Diversification Through Integrated Domain
584 Fusions. Front Genet 11:898.
585 Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers,
586 E.W., Li, P.W., and Eichler, E.E. 2002a. Recent segmental duplications in the human
587 genome. Science 297:1003-1007.
588 Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers,
589 E.W., Li, P.W., and Eichler, E.E. 2002b. Recent Segmental Duplications in the Human
590 Genome. Science 297:1003-1007.
591 Bergman, M.E., Chávez, Á., Ferrer, A., and Phillips, M.A. 2020. Distinct metabolic pathways
592 drive monoterpenoid biosynthesis in a natural population of Pelargonium graveolens.
593 Journal of Experimental Botany 71:258-271.

Birchler, J.A., and Veitia, R.A. 2012. Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. Proceedings of the National Academy of Sciences 109:14746-14753.

Blanc, G., and Wolfe, K.H. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes[W]. The Plant Cell 16:1667-1678.

Brosius, J. 1991. Retroposons—Seeds of Evolution. Science 251:753-753.

Buchfink, B., Reuter, K., and Drost, H.-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods 18:366-368.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825-5829.

Carretero-Paulet, L., Librado, P., Chang, T.-H., Ibarra-Laclette, E., Herrera-Estrella, L., Rozas, J., and Albert, V.A. 2015. High Gene Family Turnover Rates and Gene Space Adaptation in the Compact Genome of the Carnivorous Plant Utricularia gibba. Molecular Biology and Evolution 32:1284-1295.

Casola, C., and Koralewski, T.E. 2018. Pinaceae show elevated rates of gene turnover that are robust to incomplete gene annotation. Plant J.

Cazzolla Gatti, R., Reich, P.B., Gamarra, J.G.P., Crowther, T., Hui, C., Morera, A., Bastin, J.-F., de-Miguel, S., Nabuurs, G.-J., Svenning, J.-C., Serra-Diaz, J.M., Merow, C., Enquist, B., Kamenetsky, M., Lee, J., Zhu, J., Fang, J., Jacobs, D.F., Pijanowski, B., Banerjee, A., Giaquinto, R.A., Alberti, G., Almeyda Zambrano, A.M., Alvarez-Davila, E., Araujo-Murakami, A., Avitabile, V., Aymard, G.A., Balazy, R., Baraloto, C., Barroso, J.G., Bastian, M.L., Birnbaum, P., Bitariho, R., Bogaert, J., Bongers, F., Bouriaud, O., Brancalion, P.H.S., Brearley, F.Q., Broadbent, E.N., Bussotti, F., Castro da Silva, W., César, R.G., Češljar, G., Chama Moscoso, V., Chen, H.Y.H., Cienciala, E., Clark, C.J., Coomes, D.A., Dayanandan, S., Decuyper, M., Dee, L.E., Del Aguila Pasquel, J., Derroire, G., Djuikouo, M.N.K., Van Do, T., Dolezal, J., Đorđević, I.Đ., Engel, J., Fayle, T.M., Feldpausch, T.R., Fridman, J.K., Harris, D.J., Hemp, A., Hengeveld, G., Herault, B., Herold, M., Ibanez, T., Jagodzinski, A.M., Jaroszewicz, B., Jeffery, K.J., Johannsen, V.K., Jucker, T., Kangur, A., Karminov, V.N., Kartawinata, K., Kennard, D.K., Kepfer-Rojas, S., Keppel, G., Khan, M.L., Khare, P.K., Kileen, T.J., Kim, H.S., Korjus, H., Kumar, A., Kumar, A., Laarmann, D., Labrière, N., Lang, M., Lewis, S.L., Lukina, N., Maitner, B.S., Malhi, Y., Marshall, A.R., Martynenko, O.V., Monteagudo Mendoza, A.L., Ontikov, P.V., Ortiz-Malavasi, E., Pallqui Camacho, N.C., Paquette, A., Park, M., Parthasarathy, N., Peri, P.L., Petronelli, P., Pfautsch, S., Phillips, O.L., Picard, N., Piotto, D., Poorter, L., Poulsen, J.R., Pretzsch, H., Ramírez-Angulo, H., Restrepo Correa, Z., Rodeghiero, M., Rojas Gonzáles, R.D.P., Rolim, S.G., Rovero, F., Rutishauser, E., Saikia, P., Salas-Eljatib, C., Schepaschenko, D., Scherer-Lorenzen, M., Šebeň, V., Silveira, M., Slik, F., Sonké, B., Souza, A.F., Stereńczak, K.J., Svoboda, M., Taedoumg, H., Tchebakova, N., Terborgh, J., Tikhonova, E., Torres-Lezama, A., van der Plas, F., Vásquez, R., Viana, H., Vibrans, A.C., Vilanova, E., Vos, V.A., Wang, H.-F., Westerlund, B., White, L.J.T., Wiser, S.K., Zawiła-Niedźwiecki, T., Zemagho, L., Zhu, Z.-X., Zo-Bi, I.C., and Liang, J. 2022. The number of tree species on Earth. Proceedings of the National Academy of Sciences 119:e2115329119.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884-i890.

640 Conant, G.C., Birchler, J.A., and Pires, J.C. 2014. Dosage, duplication, and diploidization:
641 clarifying the interplay of multiple models for duplicate gene evolution over time.
642 Current Opinion in Plant Biology 19:91-98.
643 Cortés, A.J., Restrepo-Montoya, M., and Bedoya-Canas, L.E. 2020. Modern Strategies to Assess
644 and Breed Forest Tree Adaptation to Changing Climate. Front Plant Sci 11:583323.
645 Cortesi, F., Musilová, Z., Stieb, S.M., Hart, N.S., Siebeck, U.E., Malmstrøm, M., Tørresen, O.K.,
646 Jentoft, S., Cheney, K.L., Marshall, N.J., Carleton, K.L., and Salzburger, W. 2015.
647 Ancestral duplications and highly dynamic opsin gene evolution in percomorph fishes.
648 Proceedings of the National Academy of Sciences 112:1493-1498.
649 Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S.,
650 Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis,
651 C.W. 2006. Widespread genome duplications throughout the history of flowering plants.
652 Genome Res 16:738-749.
653 de Bruijn, W.J.C., Gruppen, H., and Vincken, J.-P. 2018. Structure and biosynthesis of
654 benzoxazinoids: Plant defence metabolites with potential as antimicrobial scaffolds.
655 Phytochemistry 155:233-243.
656 De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van de Peer,
657 Y. 2013. Convergent gene loss following gene and genome duplications creates single-
658 copy families in flowering plants. Proceedings of the National Academy of Sciences
659 110:2898-2903.
660 Demuth, J.P., and Hahn, M.W. 2009. The life and death of gene families. Bioessays 31:29-39.
661 Diss, G., Gagnon-Arsenault, I., Dion-Coté, A.-M., Vignaud, H., Ascencio, D.I., Berger, C.M.,
662 and Landry, C.R. 2017. Gene duplication can impart fragility, not robustness, in the yeast
663 protein interaction network. Science 355:630-634.
664 Dodds, P.N., and Rathjen, J.P. 2010. Plant immunity: towards an integrated view of plant–
665 pathogen interactions. Nature Reviews Genetics 11:539-548.
666 Emms, D.M., and Kelly, S. 2017. STRIDE: Species Tree Root Inference from Gene Duplication
667 Events. Mol Biol Evol 34:3267-3278.
668 Emms, D.M., and Kelly, S. 2018. STAG: Species Tree Inference from All Genes (bioRxiv).
669 Enright, A.J., Van Dongen, S., and Ouzounis, C.A. 2002. An efficient algorithm for large-scale
670 detection of protein families. Nucleic Acids Research 30:1575-1584.
671 Falcone Ferreyra, M.L., Rius, S., and Casati, P. 2012. Flavonoids: biosynthesis, biological
672 functions, and biotechnological applications. Frontiers in Plant Science 3.
673 FAO. 2020. Global Forest Resources Assessment 2020: Main report. Rome.
674 Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999.
675 Preservation of duplicate genes by complementary, degenerative mutations. Genetics
676 151:1531-1545.
677 Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-
678 generation sequencing data. Bioinformatics 28:3150-3152.
679 Gamfeldt, L., Snäll, T., Bagchi, R., Jonsson, M., Gustafsson, L., Kjellander, P., Ruiz-Jaen, M.C.,
680 Fröberg, M., Stendahl, J., Philipson, C.D., Mikusiński, G., Andersson, E., Westerlund, B.,
681 Andrén, H., Moberg, F., Moen, J., and Bengtsson, J. 2013. Higher levels of multiple
682 ecosystem services are found in forests with more tree species. Nature Communications
683 4:1340.
684 Gao, Y., Wang, W., Zhang, T., Gong, Z., Zhao, H., and Han, G.-Z. 2018. Out of Water: The
685 Origin and Early Diversification of Plant R-Genes. Plant Physiology 177:82-89.

686 Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M.,
687 Talón, M., Dopazo, J., and Conesa, A. 2008. High-throughput functional annotation and
688 data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420-3435.
689 Hahn, M.W., De Bie, T., Stajich, J.E., Nguyen, C., and Cristianini, N. 2005. Estimating the
690 tempo and mode of gene family evolution from comparative genomic data. Genome Res
691 15:1153-1160.
692 Hammond, W.M., Williams, A.P., Abatzoglou, J.T., Adams, H.D., Klein, T., López, R., Sáenz-
693 Romero, C., Hartmann, H., Breshears, D.D., and Allen, C.D. 2022. Global field
694 observations of tree die-off reveal hotter-drought fingerprint for Earth's forests. Nature
695 Communications 13:1761.
696 Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S.H. 2008. Importance of
697 lineage-specific expansion of plant tandem duplicates in the adaptive response to
698 environmental stimuli. Plant Physiol 148:993-1003.
699 Hurley, B.P., Garnas, J., Wingfield, M.J., Branco, M., Richardson, D.M., and Slippers, B. 2016.
700 Increasing numbers and intercontinental spread of invasive insects on eucalypts.
701 Biological Invasions 18:921-933.
702 Isabel, N., Holliday, J.A., and Aitken, S.N. 2020. Forest genomics: Advancing climate
703 adaptation, forest health, productivity, and conservation. Evolutionary Applications 13:3-
704 10.
705 Jiang, N., Bao, Z., Zhang, X., Eddy, S.R., and Wessler, S.R. 2004. Pack-MULE transposable
706 elements mediate gene evolution in plants. Nature 431:569-573.
707 Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf,
708 M., Ruzicka, D.R., Wafula, E., Wickett, N.J., Wu, X., Zhang, Y., Wang, J., Zhang, Y.,
709 Carpenter, E.J., Deyholos, M.K., Kutchan, T.M., Chanderbali, A.S., Soltis, P.S.,
710 Stevenson, D.W., McCombie, R., Pires, J.C., Wong, G.K., Soltis, D.E., and Depamphilis,
711 C.W. 2012. A genome triplication associated with early diversification of the core
712 eudicots. Genome Biol 13:R3.
713 Johri, P., Gout, J.-F., Doak, T.G., and Lynch, M. 2022. A Population-Genetic Lens into the
714 Process of Gene Loss Following Whole-Genome Duplication. Molecular Biology and
715 Evolution 39:msac118.
716 Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J.,
717 Mitchell, A., Nuka, G., Pesseat, S., Quinn, A.F., Sangrador-Vegas, A., Scheremetjew, M.,
718 Yong, S.Y., Lopez, R., and Hunter, S. 2014. InterProScan 5: Genome-scale protein
719 function classification. Bioinformatics 30:1236-1240.
720 Kassambara, A. 2020. ggpubr: 'ggplot2' Based Publication Ready Plots. https://CRAN.R-
721 project.org/package=ggpubr.
722 Katoh, K., Misawa, K., Kuma, K.i., and Miyata, T. 2002. MAFFT: a novel method for rapid
723 multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research
724 30:3059-3066.
725 Kirilenko, A.P., and Sedjo, R.A. 2007. Climate change impacts on forestry. Proceedings of the
726 National Academy of Sciences 104:19697-19702.
727 Kolde, R. 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12, https://CRAN.R-
728 project.org/package=pheatmap.
729 Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. 2017. TimeTree: A Resource for
730 Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution 34:1812-
731 1819.

732 Kuzmin, E., VanderSluis, B., Nguyen Ba, A.N., Wang, W., Koch, E.N., Usaj, M., Khmelinskii,
733 A., Usaj, M.M., van Leeuwen, J., Kraus, O., Tresenrider, A., Pryszlak, M., Hu, M.-C.,
734 Varriano, B., Costanzo, M., Knop, M., Moses, A., Myers, C.L., Andrews, B.J., and
735 Boone, C. 2020. Exploring whole-genome duplicate gene retention with complex genetic
736 interaction analysis. Science 368:eaaz5667.
737 Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A.,
738 Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M.,
739 Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J.,
740 DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B.,
741 Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M.,
742 Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington,
743 S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L.,
744 Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D.,
745 Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P.,
746 Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J.,
747 Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden,
748 G.T., Goh, F., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K.,
749 Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B.,
750 Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K.,
751 Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Lowe, A.J.,
752 Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T.,
753 McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W.,
754 Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.-M.,
755 Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C.,
756 Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A.,
757 Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S.G., Wenzel, M.,
758 Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D.,
759 Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen,
760 L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z.,
761 Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y.,
762 Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N.-p., Matasci, N., Nelson, D.R.,
763 Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E.,
764 Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H., Lu-Irving, P., Marx, H.E., Qi, X.,
765 Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow,
766 L., Theißen, G., Wong, G.K.-S., and One Thousand Plant Transcriptomes, I. 2019. One
767 thousand plant transcriptomes and the phylogenomics of green plants. Nature 574:679-
768 685.
769 Leister, D. 2004. Tandem and segmental gene duplication and recombination in the evolution of
770 plant disease resistance gene. Trends Genet 20:116-122.
771 Li, B., and Dewey, C.N. 2011. RSEM: accurate transcript quantification from RNA-Seq data
772 with or without a reference genome. BMC Bioinformatics 12:323.
773 Linnakoski, R., and Forbes, K.M. 2019. Pathogens—The Hidden Face of Forest Invasions by
774 Wood-Boring Insect Pests. Frontiers in Plant Science 10.
775 Liu, J., Yu, T., Mu, Z., and Li, G. 2019. TransLiG: a de novo transcriptome assembler that uses
776 line graph iteration. Genome Biology 20:81.

Lu, M., Feau, N., Vidakovic, D.O., Ukrainetz, N., Wong, B., Aitken, S.N., Hamelin, R.C., and Yeaman, S. 2021. Comparative Gene Expression Analysis Reveals Mechanism of Pinus contorta Response to the Fungal Pathogen Dothistroma septosporum. Mol Plant Microbe Interact:Mpmi10200282r.

Lynch, M., and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Lynch, M., and Conery, J.S. 2003. The evolutionary demography of duplicate genes. Journal of Structural and Functional Genomics 3:35-44.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. 2005. Modeling gene and genome duplications in eukaryotes. Proceedings of the National Academy of Sciences 102:5454-5459.

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A., and Zdobnov, E.M. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution 38:4647-4654.

Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. 2020. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 36:5516-5518.

Meng, X., and Zhang, S. 2013. MAPK cascades in plant disease resistance signaling. Annu Rev Phytopathol 51:245-266.

Meyers, B.C., Kozik, A., Griego, A., Kuang, H., and Michelmore, R.W. 2003. Genome-Wide Analysis of NBS-LRR–Encoding Genes in Arabidopsis[W]. The Plant Cell 15:809-834.

Miura, S., Amacher, M., Hofer, T., San-Miguel-Ayanz, J., Ernawati, and Thackway, R. 2015. Protective functions and ecosystem services of global forests in the past quarter-century. Forest Ecology and Management 352:35-46.

Neale, D.B., and Kremer, A. 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111-122.

Ngou, B.P.M., Heal, R., Wyler, M., Schmid, M.W., and Jones, J.D.G. 2022. Concerted expansion and contraction of immune receptor gene repertoires in plant genomes. Nature Plants 8:1146-1152.

Ohno, S. 1970. Evolution by gene duplication (Springer-Verlag).

Panchy, N., Lehti-Shiu, M., and Shiu, S.H. 2016. Evolution of Gene Duplication in Plants. Plant 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.

Prunier, J., Caron, S., Lamothe, M., Blais, S., Bousquet, J., Isabel, N., and MacKay, J. 2017. Gene copy number variations in adaptive evolution: The genomic distribution of gene copy number variations revealed by genetic mapping and their adaptive role in an undomesticated species, white spruce (*Picea glauca*). Mol Ecol 26:5989-6001.

821 Puttick, M.N., Clark, J., and Donoghue, P.C.J. 2015. Size is not everything: rates of genome size
822 evolution, not C-value, correlate with speciation in angiosperms. Proceedings of the
823 Royal Society B: Biological Sciences 282:20152289.
824 R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for
825 Statistical Computing, Vienna, Austria. https://www.R-project.org/.
826 Reid, N.M., Proestou, D.A., Clark, B.W., Warren, W.C., Colbourne, J.K., Shaw, J.R., Karchner,
827 S.I., Hahn, M.E., Nacci, D., Oleksiak, M.F., Crawford, D.L., and Whitehead, A. 2016.
828 The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in
829 wild fish. Science 354:1305-1308.
830 Rizvi, A.R., Baig, S., Barrow, E., and Kumar, C. 2015. Synergies between climate mitigation
831 and adaptioon in forest landscape restoration. Gland, Switzerland: IUCN.
832 Rodgers-Melnick, E., Mane, S.P., Dharmawardhana, P., Slavov, G.T., Crasta, O.R., Strauss,
833 S.H., Brunner, A.M., and Difazio, S.P. 2012. Contrasting patterns of evolution following
834 whole genome versus tandem duplication events in Populus. Genome Res 22:95-105.
835 Sanderson, M.J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times
836 in the absence of a molecular clock. Bioinformatics 19:301-302.
837 Seoighe, C., and Gehring, C. 2004. Genome duplication led to highly selective expansion of the
838 Arabidopsis thaliana proteome. Trends in Genetics 20:461-464.
839 Soltis, P.S., and Soltis, D.E. 2016. Ancient WGD events as drivers of key innovations in
840 angiosperms. Curr Opin Plant Biol 30:159-165.
841 Swenson, N.G., and Jones, F.A. 2017. Community transcriptomics, genomics and the problem of
842 species co-occurrence. Journal of Ecology 105:563-568.
843 Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. 2008. Synteny and
844 Collinearity in Plant Genomes. Science 320:486-488.
845 Trouern-Trend, A.J., Falk, T., Zaman, S., Caballero, M., Neale, D.B., Langley, C.H., Dandekar,
846 A.M., Stevens, K.A., and Wegrzyn, J.L. 2020. Comparative genomics of six Juglans
847 species reveals disease-associated gene family contractions. The Plant Journal 102:410-
848 423.
849 Van de Peer, Y., Maere, S., and Meyer, A. 2009. The evolutionary significance of ancient
850 genome duplications. Nature Reviews Genetics 10:725-732.
851 Wickham, H. 2011. The Split-Apply-Combine strategy for data analysis. J. Stat. Softw. 40:1-29.
852 Wickham, H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New York, NY,
853 US.
854 Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.
855 24:1586-1591.
856 Yang, Z., and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates
857 under realistic evolutionary models. Mol Biol Evol 17:32-43.
858 Yu, Y. 2022. mixR: An R package for finite mixture modeling for both raw and binned data.
859 Journal of Open Source Software 7:4031.
860 Zeng, L., Zhang, N., Zhang, Q., Endress, P.K., Huang, J., and Ma, H. 2017. Resolution of deep
861 eudicot phylogeny and their temporal diversification using nuclear genes from
862 transcriptomic and genomic datasets. New Phytol 214:1338-1354.
863 Zhang, J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution
864 18:292-298.
865 Zhao, J., Davis, L.C., and Verpoorte, R. 2005. Elicitor signal transduction leading to production
866 of plant secondary metabolites. Biotechnol Adv 23:283-333.

867 Zwaenepoel, A., and Van de Peer, Y. 2019. Inference of Ancient Whole-Genome Duplications
868 and the Evolution of Gene Duplication and Loss Rates. Molecular Biology and Evolution
869 36:1384-1404.
870