# Comparative transcriptomics reveals divergence in pathogen response gene families amongst 20 forest tree species

[1*]Mengmeng Lu, [2]Min Cao, [2]Jie Yang, [1,3]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] University of Notre Dame Environmental Research Center, 736 Flanner Hall, University of Notre Dame, Notre Dame, IN 46556, USA

[*]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 (FAO 2020; Gamfeldt et al. 2013; Miura et al. 2015). 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 (Allen et al. 2010; Hammond et al. 2022; Kirilenko and Sedjo 2007; Rizvi et al. 2015). 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 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 (Jiao et al. 2012; Tang et al. 2008). Alternatively, pervasive SSD, including tandem duplication (Zhang 2003), transposon-mediated duplication (Jiang et al. 2004), segmental duplication (Bailey et al. 2002), 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 (One Thousand Plant Transcriptomes Initiative 2019). They are sources of plant functional evolution and innovation (Force et al. 1999; Ohno 1970), exhibiting a clear functional basis, which are influenced by the mechanisms of duplication (Hanada et al. 2008; Maere et al. 2005). 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 (Isabel et al. 2020; Neale and Kremer 2011), 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 (Demuth and Hahn 2009; Lynch and Conery 2003). Large segmental duplication and changes in copy number are known to be associated with disease response and environmental adaptation (Bailey et al. 2002; 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 hazelnut), *Dirca palustris* (Eastern leatherwood), *Fraxinus nigra* (black ash), *Ostrya virginiana* (American hophornbeam), *Populus grandidentata* (bigtooth aspen), *Populus tremuloides* (quaking aspen), *Prunus serotina* (black cherry), *Prunus virginiana* (chokecherry), *Quercus rubra* (northern red oak), *Ribes cynosbati* (prickly gooseberry), *Tilia americana* (American basswood), *Ulmus americana* (American elm). They are all angiosperms (flowering plants) and eudicots, belonging to the clades of asterids, saxifragales, 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.

*RNA-sequencing, de novo transcriptome assembly, and annotation*

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 NovaSeq 6000 platform to produce 150 bp paired-end reads. 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 (Phred quality score ≤5) which is over 50% of the total base. An average of 6 Gb 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 (TPM) value ≥1.5 for each gene was retained. Redundancy was removed from the retained transcripts using the software CD-HIT v4.8.1 with parameters of -c 0.95 -n 8 -T 32 -M 1540 (Fu et al. 2012). To remove potential contamination, the retained transcripts were aligned against the non-redundant (nr) protein database (downloaded on January 22, 2022) using the blastx function implemented by the software DIAMOND v2.0.13.151

(Buchfink et al. 2021) with parameter of --outfmt 102 to assign the best known taxon ID to each query transcript. The file that was used to convert accession numbers to taxonomy was downloaded from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/, and the taxon ID file was downloaded from https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/. The query transcripts that could be assigned to green plant (Viridiplantae) taxon IDs were retained. The filtered transcripts were aligned to the nr database that only includes green plant subjects using the blastx function implemented by the software DIAMOND v2.0.13.151 with default parameters. The aligned transcripts were combined with those transcripts assigned with green plant taxon IDs to form the final transcriptomes. The completeness of the final transcriptomes was examined using the 1,614 orthologues in the BUSCO (Benchmarking Universal Single Copy Orthologs) v5.2.2 set of embryophyta_odb10 (Manni et al. 2021).

The transcriptomes were annotated with Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) terms. The transcripts were aligned to the nr database that only includes green plant subjects using the blastx function implemented by the software DIAMOND v2.0.13.151 with parameters of -taxonlist 33090 --very-sensitive -f5. The .xml outputs were used as input for mapping and annotation using the software Blast2GO (Götz et al. 2008). The transcripts were translated to protein sequences using the software TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/wiki) with default parameters to get the longest open reading frame (ORF) per transcript, and then were annotated by the software InterProScan v5.54-87.0 (Jones et al. 2014) with the parameters of -dra -f XML -goterms -pa. The final GO terms were acquired by merging Blast2GO and InterProScan annotation outputs using the software Blast2GO. To get the KEGG terms, the longest ORF per transcript were also annotated using EggNOG-mapper v2.1.7 (Cantalapiedra et al. 2021) with parameters of -m diamond -sensmode more-sensitive -tax_scope 33090 -tax_scope_mode 33090. The assembled transcriptome and annotation were used in the downstream analyses as shown in the flowchart (Supplementary Fig. S1).

*Ks analyses*

The level of synonymous substitutions (Ks) between two homologous sequences increases approximately linearly with time. Thus pairs of paralogous genes within a species can be sorted to estimate the relative ages of gene duplication (Blanc and Wolfe 2004). To identify paralogous gene pairs within a species, an all-versus-all blastp was performed within each species. The peptide sequences (the longest ORF per transcript) of each species were used as input to run all-versus-all 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 blastp output was filtered by removing self blastp results, keeping outputs with pident (percent of identical matches) ≥20 and nident (number of identical matches) ≥50, and removing sequences with more than 10 hits. The filtered paralogous gene pairs were aligned using the software MAFFT v7.490 (Katoh et al. 2002) with default parameters. The alignments were used as input for estimating the number of synonymous substitutions per synonymous site (Ks) and the number of non-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 $\geq$0.01 and $\leq$2 using the R package mixR (Yu 2022). The best model from a candidate of mixture models was selected using the select function based on the information criterion BIC.

*Estimating duplication events and divergence time*

To infer duplication events, the species tree and gene trees were first inferred using STAG and STRIDE algorithms (Emms and Kelly 2017; 2018) implemented by the program OrthoFinder v2.5.4 (Emms and Kelly 2019). 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 duplicate 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 in November, 2022 from http://www.timetree.org/ (Kumar et al. 2017).

*Gene families and their evolutionary rate*

To infer gene families, the peptide sequences (the longest ORF per transcript) of the twenty 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 a 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 (One Thousand Plant Transcriptomes Initiative 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 a mean Ka/Ks >10 were filtered out. The graphs were plotted using the R package ggplot2 and ggpubr (Kassambara 2023; Wickham 2016). 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 2023). 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 v2.0.13.151. 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. 2021) with the command python prep_r8s.py -s 60164 -p 'Fraxinus_nigra,Corylus_cornuta' -c 115. "60164" was the number of sites in the alignment. "-c 115" represents a median of 115 MYA divergence time, it is the calibration points between *Fraxinus nigra* and *Corylus cornuta*. Then the ultrametric tree was estimated using the program r8s v1.81 with default parameters. The ultrametric tree and gene family counts were used as input for the program CAFE5. Since CAFE5 was not recommended to analyze recent polyploid species, only diploid species were used to run 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 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_sac
charum: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 have similar Ks shapes, so do the two *Populus* species, suggestive of common WGDs before the two species diverged from the same ancestors. By examining the orthologous Ks of the two *Prunus* and of the two *Populus* species, respectively, we found a high peak close to zero, which may represent the duplication events after the WGDs and divergence (Supplementary Fig. S3).

*Duplication events and divergence time*

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 WGDs across eudicots 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 Saxifragales was about 100 MYA.

*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. S4, 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 metabolite biosynthesis.

*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. S5, respectively. 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.

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. S6) 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).

## Discussion

Gene duplication events are prevalent in all organisms and provide genetic variation for evolutionary innovation and adaption to fluctuating conditions (Panchy et al. 2016; Van de Peer et al. 2009; Zhang 2003). 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 an 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 (Fig. 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. The duplication events at the root (Fig. 3) of the tree 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. WGD events followed by different evolutionary events might also occur in some clades. For *Acer*, it is likely that *Acer spicatum* experienced a diploidization event to revert to diploid status, while *Acer rubrum* maintained a polyploid status.

Divergence time mostly agrees with the previous reports, which stated rosids and asterids diverged about 125 MYA, malvids and fabids diverged about 110 MYA (Zeng et al. 2017). However, divergence estimate in the present study is rather rough. Further studies need to consider using dense taxon sampling, diverse markers like nuclear internal transcribed spacer or chloroplast DNA fragments, and molecular dating with fossil corrections methods to better resolve the evolutionary relationships.

*Sustained large size gene families*

Though most duplicate genes were eliminated or pseudonized (Panchy et al. 2016), a small portion of gene families can actively maintain duplicate genes as selection favors the fitness related neo-functionalization. The main functions of these sustained large size gene families are mainly involved in plant-pathogen interaction, signaling, and metabolite biosynthesis (Fig. 4b & Supplementary Fig. S4). This is consistent with previous report that genes encoding functions like signal transduction and stress response tend to have paralogs (Panchy et al. 2016), and genes related to the function like transcription regulation and signal transduction had higher possibility to survive in multiple rounds of genome duplication (Seoighe and Gehring 2004). Plants constantly combat with pathogens and pests. The major class of plant disease resistance (R) genes contain a nucleotide-binding site (NBS) and C-terminal leucin-rich repeats (LRRs). Tandem and segmental gene duplication explain much of the diversification among R genes (Andersen et al. 2020; Leister 2004; Meyers et al. 2003), possibly because they have experienced frequent deletion and duplication throughout evolution, and thus are highly variable (Gao et al. 2018; Hanada et al. 2008). Plant secondary metabolites, such as monoterpenoid, flavonoid, benzoxazinoid, play important roles in defense against predation and pathogens (Bergman et al. 2020; de Bruijn et al. 2018; Falcone Ferreyra et al. 2012). The regulation of plant secondary metabolism requires the intimation of signal transduction network that leads to activation of biosynthetic genes (Zhao et al. 2005).

Gene duplication and deletion give rise to gene family expansion and contraction. The high gene gain rate likely reflects the occurrence of multiple WGDs and frequent SSDs (Carretero-Paulet et al. 2015). While for the other species with contraction outnumbered expansion, the pattern is mainly caused by gene loss and retention of specific gene families followed the common WGDs

preceded their common ancestor (Fig. 3). We found a high gene family turnover rate in rosids, but we must caution that the unbalanced taxon sampling in our study species might 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. S6). 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).

*Limitation of the current study*

Our conclusions are drawn from samples collected from co-occurring species from a single forest. The forest type sampled and the species are typical for the entire region including the Ottawa, Nicolet, and Hiawatha National Forests in northern Wisconsin and the Upper Peninsula of Michigan, USA. As individuals within species vary in their adaptive and plastic responses to the environment, our results are, therefore, most applicable to the study region and not the entire geographic range of the species. Furthermore, the assembled transcripts of each species in this study are derived from leaf material and, therefore, cannot represent the whole gene set expressed by the organism. Lastly, we also note that we filtered out genes that had low expression levels to ensure the quality of the assembled transcriptome. Thus, our study does omit some genes, and it is possible that we might end up with different conclusions if we had more genes.

## 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. We appreciate Dr. Sam Yeaman and Dr. James R. Whiting for constructive suggestions. 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**

ML and NGS were supported by the Dimensions of Biodiversity grant (DEB-2124466) from National Science Foundation. JY and MC were supported by the China-US Dimensions of Biodiversity grant (DEB-32061123003) from the National Natural Science Foundation of China.

**Conflicts of interest**

The authors declare no conflict of interest.

**Literature cited**

Alexa A, Rahnenfuhrer J. 2023. topGO: Enrichment analysis for gene ontology. R package version 2.52.0.

Allen CD, Macalady AK, Chenchouni H, Bachelet D, McDowell N, Vennetier M, Kitzberger T, Rigling A, Breshears DD, Hogg EHT, et al. 2010. A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. Forest Ecol Manag. 259(4):660-684. doi:10.1016/j.foreco.2009.09.001.

Amborella Genome Project. 2013. The Amborella genome and the evolution of flowering plants. Science. 342(6165):1241089. doi:10.1126/science.1241089.

Andersen EJ, Nepal MP, Purintun JM, Nelson D, Mermigka G, Sarris PF. 2020. Wheat disease resistance genes and their diversification through integrated domain fusions. Front Genet. 11:898. doi:10.3389/fgene.2020.00898.

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. Science. 297(5583):1003-1007. doi:10.1126/science.1072047.

Bergman ME, Chávez Á, Ferrer A, Phillips MA. 2020. Distinct metabolic pathways drive monoterpenoid biosynthesis in a natural population of *Pelargonium graveolens*. J Exp Bot. 71(1):258-271. doi:10.1093/jxb/erz397.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes[W]. Plant Cell. 16(7):1667-1678. doi:10.1105/tpc.021345.

Brosius J. 1991. Retroposons—seeds of evolution. Science. 251(4995):753. doi:10.1126/science.1990437.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 18(4):366-368. doi:10.1038/s41592-021-01101-x.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 38(12):5825-5829. doi:10.1093/molbev/msab293.

Carretero-Paulet L, Librado P, Chang T-H, Ibarra-Laclette E, Herrera-Estrella L, Rozas J, Albert VA. 2015. High gene family turnover rates and gene space adaptation in the compact genome of the Carnivorous plant *Utricularia gibba*. Mol Biol Evol. 32(5):1284-1295. doi:10.1093/molbev/msv020.

Casola C, Koralewski TE. 2018. Pinaceae show elevated rates of gene turnover that are robust to incomplete gene annotation. Plant J. doi:10.1111/tpj.13994.

Cazzolla Gatti R, Reich PB, Gamarra JGP, Crowther T, Hui C, Morera A, Bastin J-F, de-Miguel S, Nabuurs G-J, Svenning J-C, et al. 2022. The number of tree species on Earth. Proc Natl Acad Sci USA. 119(6):e2115329119. doi:10.1073/pnas.2115329119.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17):i884-i890. doi:10.1093/bioinformatics/bty560.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr Opin Plant Biol. 19:91-98. doi:10.1016/j.pbi.2014.05.008.

Cortés AJ, Restrepo-Montoya M, Bedoya-Canas LE. 2020. Modern strategies to assess and breed forest tree adaptation to changing climate. Front Plant Sci. 11:583323. doi:10.3389/fpls.2020.583323.

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16(6):738-749. doi:10.1101/gr.4825606.

de Bruijn WJC, Gruppen H, Vincken J-P. 2018. Structure and biosynthesis of benzoxazinoids: Plant defence metabolites with potential as antimicrobial scaffolds. Phytochemistry. 155:233-243. doi:10.1016/j.phytochem.2018.07.005.

De Smet R, Adams KL, Vandepoele K, Van Montagu MCE, Maere S, Van de Peer Y. 2013. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA. 110(8):2898-2903. doi:10.1073/pnas.1300127110.

Demuth JP, Hahn MW. 2009. The life and death of gene families. Bioessays. 31(1):29-39. doi:10.1002/bies.080085.

Diss G, Gagnon-Arsenault I, Dion-Coté A-M, Vignaud H, Ascencio DI, Berger CM, Landry CR. 2017. Gene duplication can impart fragility, not robustness, in the yeast protein interaction network. Science. 355(6325):630-634. doi:10.1126/science.aai7685.

Dodds PN, Rathjen JP. 2010. Plant immunity: towards an integrated view of plant–pathogen interactions. Nat Rev Genet. 11(8):539-548. doi:10.1038/nrg2812.

Emms DM, Kelly S. 2017. STRIDE: Species tree root inference from gene duplication events. Mol Biol Evol. 34(12):3267-3278. doi:10.1093/molbev/msx259.

Emms DM, Kelly S. 2018. STAG: Species tree inference from all genes. bioRxiv. doi:10.1101/267914.

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238. doi:10.1186/s13059-019-1832-y.

Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30(7):1575-1584. doi:10.1093/nar/30.7.1575.

Falcone Ferreyra ML, Rius SP, Casati P. 2012. Flavonoids: biosynthesis, biological functions, and biotechnological applications. Front Plant Sci. 3:222. doi:10.3389/fpls.2012.00222.

FAO. 2020. Global Forest Resources Assessment 2020: Main report. Rome. https://doi.org/10.4060/ca9825en.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151(4):1531-1545. doi:10.1093/genetics/151.4.1531.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 28(23):3150-3152. doi:10.1093/bioinformatics/bts565.

Gamfeldt L, Snäll T, Bagchi R, Jonsson M, Gustafsson L, Kjellander P, Ruiz-Jaen MC, Fröberg M, Stendahl J, Philipson CD, et al. 2013. Higher levels of multiple ecosystem services are found in forests with more tree species. Nat Commun. 4(1):1340. doi:10.1038/ncomms2328.

Gao Y, Wang W, Zhang T, Gong Z, Zhao H, Han G-Z. 2018. Out of water: The origin and early diversification of plant R-genes. Plant Physiol. 177(1):82-89. doi:10.1104/pp.18.00185.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36(10):3420-3435. doi:10.1093/nar/gkn176.

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianini N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15(8):1153-1160. doi:10.1101/gr.3567505.

Hammond WM, Williams AP, Abatzoglou JT, Adams HD, Klein T, López R, Sáenz-Romero C, Hartmann H, Breshears DD, Allen CD. 2022. Global field observations of tree die-off reveal hotter-drought fingerprint for Earth's forests. Nat Commun. 13(1):1761. doi:10.1038/s41467-022-29289-2.

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH. 2008. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. 148(2):993-1003. doi:10.1104/pp.108.122457.

Hurley BP, Garnas J, Wingfield MJ, Branco M, Richardson DM, Slippers B. 2016. Increasing numbers and intercontinental spread of invasive insects on eucalypts. Biol Invasions. 18(4):921-933. doi:10.1007/s10530-016-1081-x.

Isabel N, Holliday JA, Aitken SN. 2020. Forest genomics: Advancing climate adaptation, forest health, productivity, and conservation. Evol Appl. 13(1):3-10. doi:10.1111/eva.12902.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 431(7008):569-573. doi:10.1038/nature02953.

Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 13(1):R3. doi:10.1186/gb-2012-13-1-r3.

Johri P, Gout J-F, Doak TG, Lynch M. 2022. A population-genetic lens into the process of gene loss following whole-genome duplication. Mol Biol Evol. 39(6):msac118. doi:10.1093/molbev/msac118.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al. 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 30(9):1236-1240. doi:10.1093/bioinformatics/btu031.

Kassambara A. 2023. ggpubr: 'ggplot2' based publication ready plots. R package version 0.6.0. https://rpkgs.datanovia.com/ggpubr/.

Katoh K, Misawa K, Kuma Ki, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059-3066. doi:10.1093/nar/gkf436.

Kirilenko AP, Sedjo RA. 2007. Climate change impacts on forestry. Proc Natl Acad Sci USA. 104(50):19697-19702. doi:10.1073/pnas.0701424104.

Kolde R. 2019. pheatmap: Pretty Heatmaps. R package version 1.0.12.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A resource for timelines, timetrees, and divergence times. Mol Biol Evol. 34(7):1812-1819. doi:10.1093/molbev/msx116.

Kuzmin E, VanderSluis B, Nguyen Ba AN, Wang W, Koch EN, Usaj M, Khmelinskii A, Usaj MM, van Leeuwen J, Kraus O, et al. 2020. Exploring whole-genome duplicate gene retention with complex genetic interaction analysis. Science. 368(6498):eaaz5667. doi:10.1126/science.aaz5667.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends Genet. 20(3):116-122. doi:10.1016/j.tig.2004.01.007.

Li B, Dewey CN. 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 12(1):323. doi:10.1186/1471-2105-12-323.

Linnakoski R, Forbes KM. 2019. Pathogens—The hidden face of forest invasions by wood-boring insect pests. Front Plant Sci. 10:90. doi:10.3389/fpls.2019.00090.

Liu J, Yu T, Mu Z, Li G. 2019. TransLiG: a de novo transcriptome assembler that uses line graph iteration. Genome Biol. 20(1):81. doi:10.1186/s13059-019-1690-7.

Lu M, Feau N, Obreht Vidakovic D, Ukrainetz N, Wong B, Aitken SN, Hamelin RC, Yeaman S. 2021. Comparative gene expression analysis reveals mechanism of *Pinus contorta*

response to the fungal pathogen *Dothistroma septosporum*. Mol Plant Microbe Interact. 34(4):397-409. doi:10.1094/mpmi-10-20-0282-r.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290(5494):1151-1155. doi:10.1126/science.290.5494.1151.

Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics. 3(1-4):35-44. doi:10.1023/A:1022696612931.

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA. 102(15):5454-5459. doi:10.1073/pnas.0501102102.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO Update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 38(10):4647-4654. doi:10.1093/molbev/msab199.

Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2021. CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics. 36(22-23):5516-5518. doi:10.1093/bioinformatics/btaa1022.

Meng X, Zhang S. 2013. MAPK cascades in plant disease resistance signaling. Annu Rev Phytopathol. 51:245-266. doi:10.1146/annurev-phyto-082712-102314.

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. 2003. Genome-wide analysis of NBS-LRR-encoding genes in Arabidopsis. Plant Cell. 15(4):809-834. doi:10.1105/tpc.009308.

Miura S, Amacher M, Hofer T, San-Miguel-Ayanz J, Ernawati, Thackway R. 2015. Protective functions and ecosystem services of global forests in the past quarter-century. Forest Ecol Manag. 352(7):35-46. doi:10.1016/j.foreco.2015.03.039.

Neale DB, Kremer A. 2011. Forest tree genomics: growing resources and applications. Nat Rev Genet. 12(2):111-122. doi:10.1038/nrg2931.

Ngou BPM, Heal R, Wyler M, Schmid MW, Jones JDG. 2022. Concerted expansion and contraction of immune receptor gene repertoires in plant genomes. Nat Plants. 8(10):1146-1152. doi:10.1038/s41477-022-01260-5.

Ohno S. 1970. Evolution by gene duplication. New York: Springer.

One Thousand Plant Transcriptomes Initiative. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 574(7780):679-685. doi:10.1038/s41586-019-1693-2.

Panchy N, Lehti-Shiu M, Shiu SH. 2016. Evolution of gene duplication in plants. Plant Physiol. 171(4):2294-2316. doi:10.1104/pp.16.00523.

Piasecka A, Jedrzejczak-Rey N, Bednarek P. 2015. Secondary metabolites in plant innate immunity: Conserved function of divergent chemicals. New Phytol. 206(3):948-964. doi:10.1111/nph.13325.

Picart-Picolo A, Grob S, Picault N, Franek M, Llauro C, Halter T, Maier TR, Jobet E, Descombin J, Zhang P, et al. 2020. Large tandem duplications affect gene expression, 3D organization, and plant-pathogen response. Genome Res. 30(11):1583-1592. doi:10.1101/gr.261586.120.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 26(7):1641-1650. doi:10.1093/molbev/msp077.

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 NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, Shaw JR, Karchner SI, Hahn ME, Nacci D, Oleksiak MF, et al. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science. 354(6317):1305-1308. doi:10.1126/science.aah4993.

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

Sanderson MJ. 2003. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics. 19(2):301-302. doi:10.1093/bioinformatics/19.2.301.

Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet. 20(10):461-464. doi:10.1016/j.tig.2004.07.008.

Soltis PS, Soltis DE. 2016. Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol. 30:159-165. doi:10.1016/j.pbi.2016.03.015.

Swenson NG, Iida Y, Howe R, Wolf A, Umaña MN, Petprakob K, Turner BL, Ma K. 2017. Tree co-occurrence and transcriptomic response to drought. Nat Commun. 8(1):1996. doi:10.1038/s41467-017-02034-w.

Swenson NG, Jones FA. 2017. Community transcriptomics, genomics and the problem of species co-occurrence. J Ecol. 105(3):563-568. doi:10.1111/1365-2745.12771.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. Science. 320(5875):486-488. doi:10.1126/science.1153917.

Trouern-Trend AJ, Falk T, Zaman S, Caballero M, Neale DB, Langley CH, Dandekar AM, Stevens KA, Wegrzyn JL. 2020. Comparative genomics of six Juglans species reveals disease-associated gene family contractions. Plant J. 102(2):410-423. doi:10.1111/tpj.14630.

Van de Peer Y, Maere S, Meyer A. 2009. The evolutionary significance of ancient genome duplications. Nat Rev Genet. 10(10):725-732. doi:10.1038/nrg2600.

Wickham H. 2011. The Split-Apply-Combine strategy for data analysis. J Stat Softw. 40(1):1-29. doi:10.18637/jss.v040.i01.

Wickham H. 2016. ggplot2: Elegant graphics for data analysis. New York: Springer.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8):1586-1591. doi:10.1093/molbev/msm088.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 17(1):32-43. doi:10.1093/oxfordjournals.molbev.a026236.

Yu Y. 2022. mixR: An R package for finite mixture modeling for both raw and binned data. Journal of Open Source Software. 7(69):4031. doi:10.21105/joss.04031.

Zeng L, Zhang N, Zhang Q, Endress PK, Huang J, Ma H. 2017. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. New Phytol. 214(3):1338-1354. doi:10.1111/nph.14503.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18(6):292-298. doi:10.1016/S0169-5347(03)00033-8.

Zhao J, Davis LC, Verpoorte R. 2005. Elicitor signal transduction leading to production of plant secondary metabolites. Biotechnol Adv. 23(4):283-333. doi:10.1016/j.biotechadv.2005.01.003.

Zwaenepoel A, Van de Peer Y. 2019. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. Mol Biol Evol. 36(7):1384-1404. doi:10.1093/molbev/msz088.

## Figure legends

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

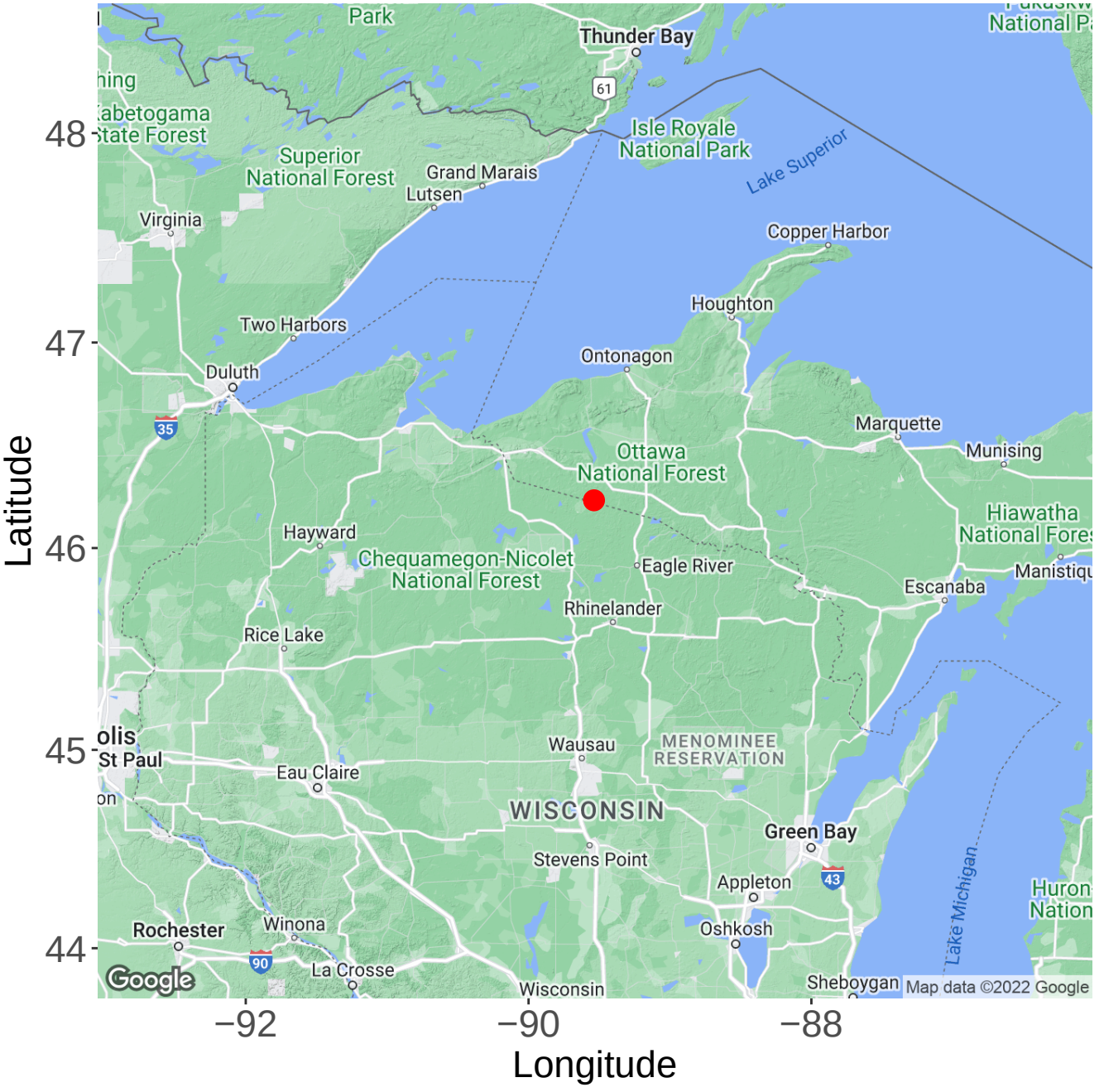
**Fig. 2. Distribution of Ks (the level of synonymous substitutions). A mixture model was fit and plot for the Ks values ≥0.01 and ≤2 of each species using the R package mixR.**

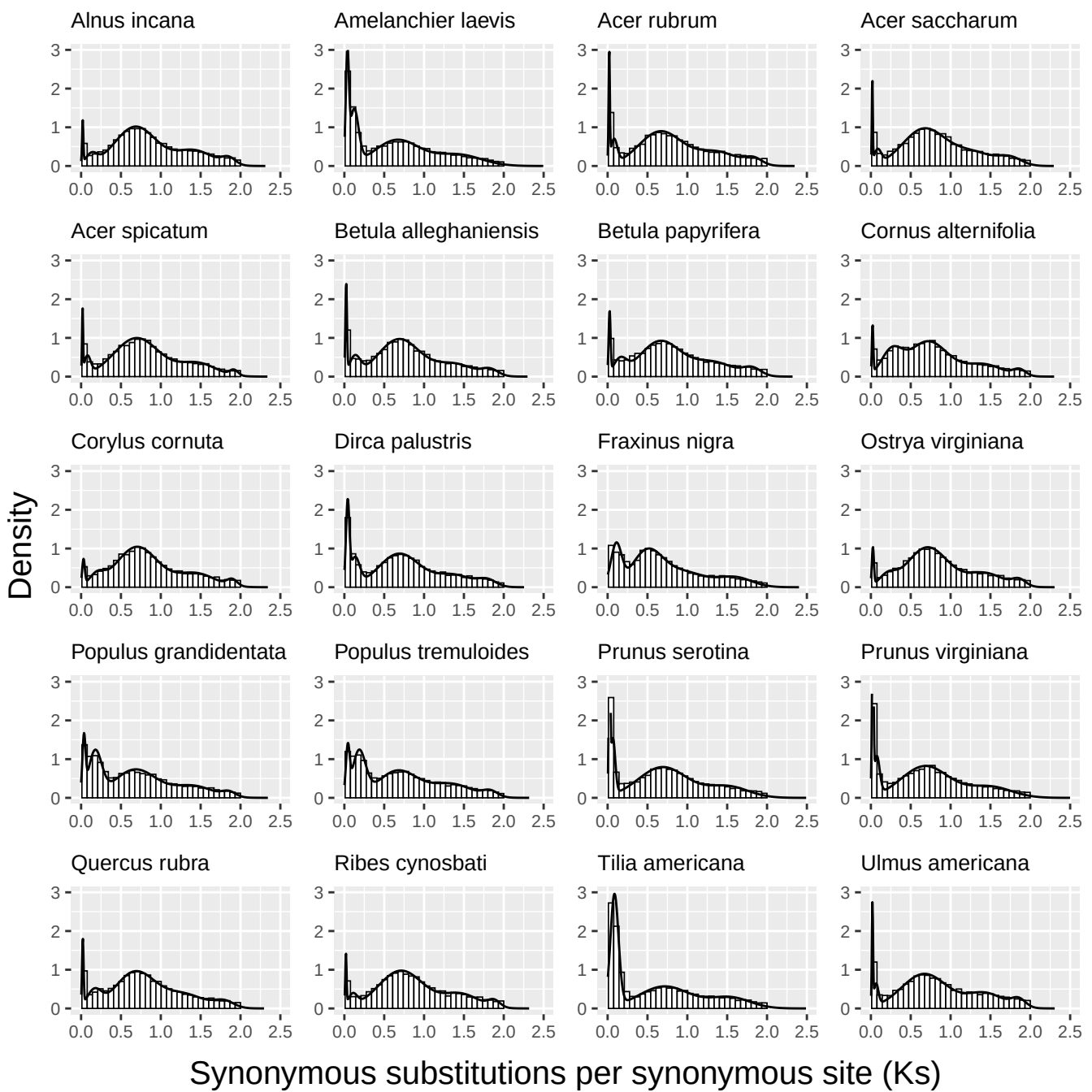
**Fig. 3. Divergence time in million years and gene duplication events with 50% support rate. Node numbers were followed by number of gene duplication events.**

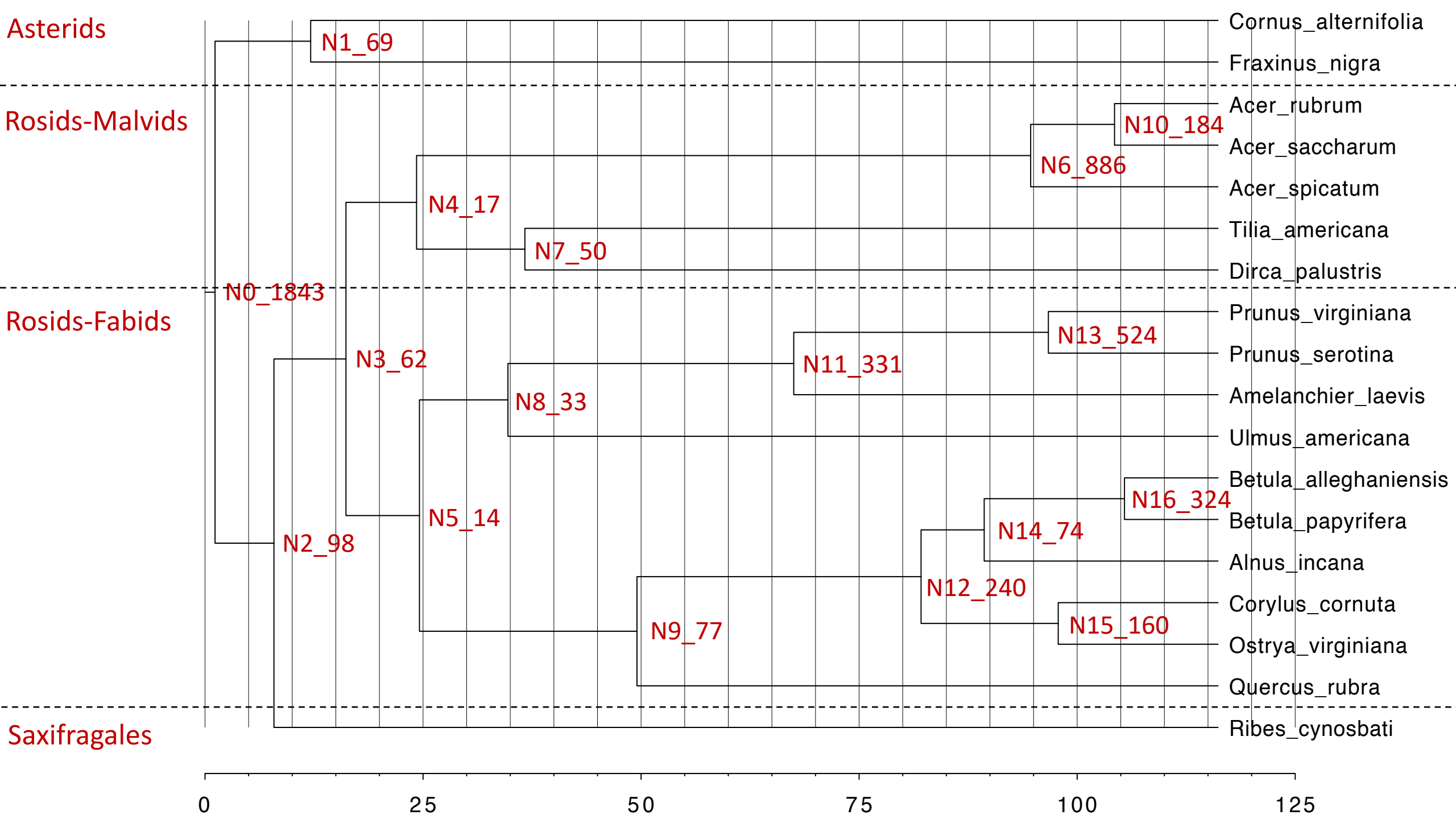
**Fig. 4. Scatter plot between gene size and Ka/Ks per gene family (a) and KEGG pathway terms enrichment (b) in gene families with a size ≥5 and Ka/Ks >1. The bar in b represents the adjusted p-values of KEGG enrichment test.**

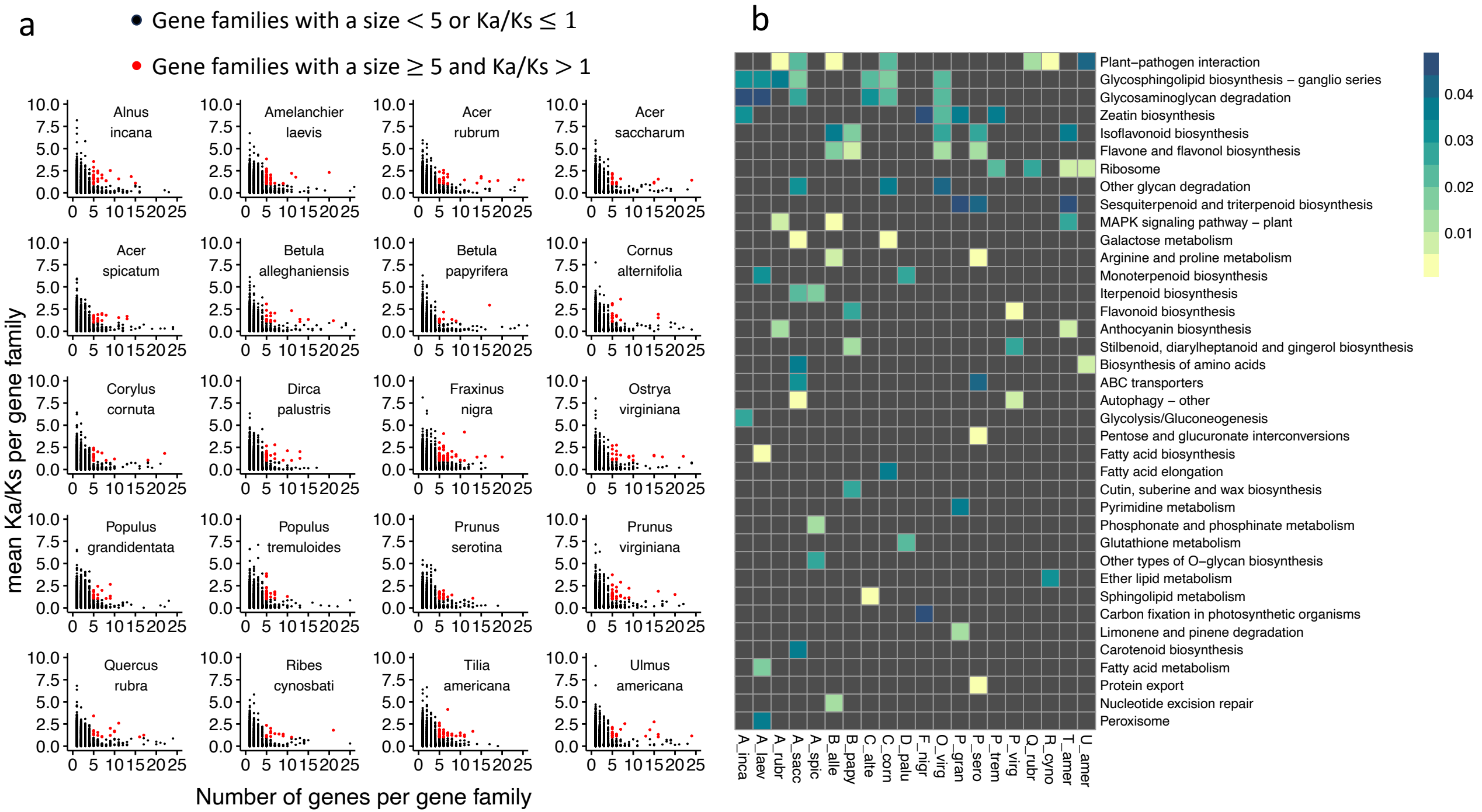
**Fig. 5. The percentage of gene family expansion and contraction (a) and the number of fast evolving gene families (b), with + representing fast expanding gene family and – representing fast contracting gene family.**

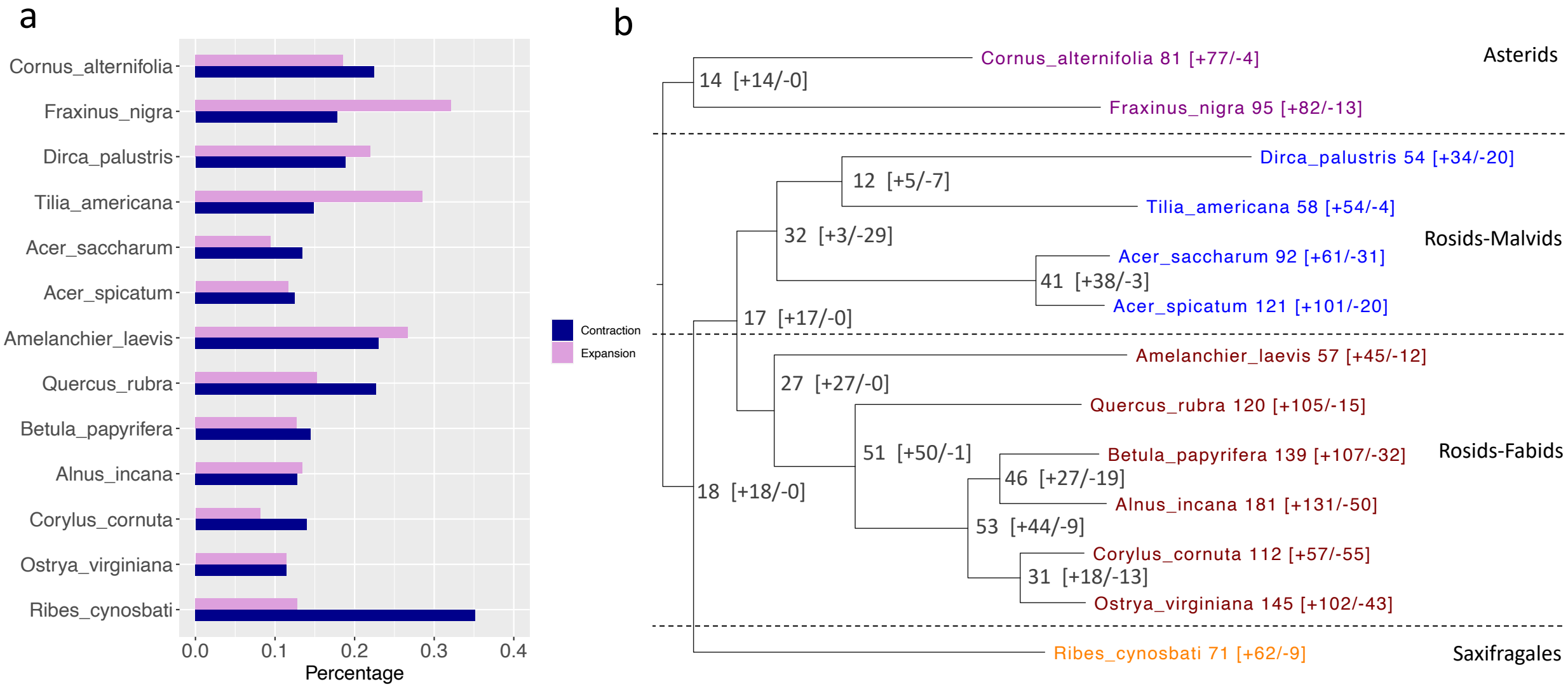
**Fig. 6. Enriched KEGG pathway terms in the studied diploid species. Red represents fast expanding gene family; blue represents fast contracting gene family; white represents both fast expanding and contracting gene family. Dark grey represents the KEGG terms were not enriched in fast expanding or contracting gene family.**

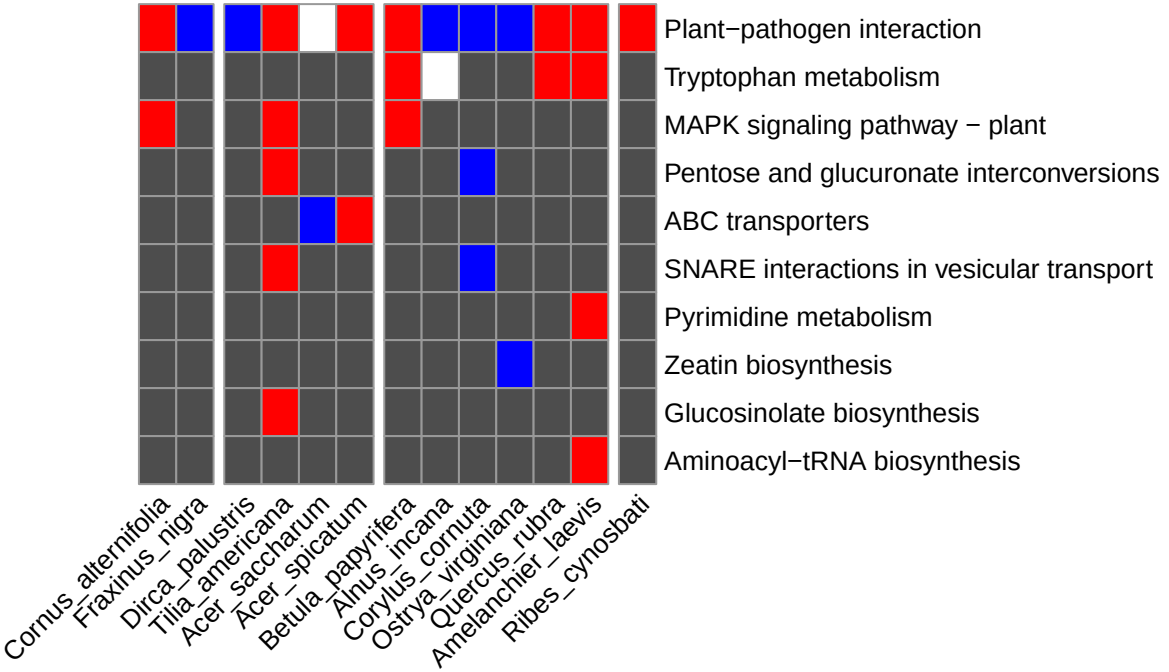

Park
Thunder Bay
61
Isle Royale
National Park
Lake Superior
Superior
National Forest
Grand Marais
Lutsen
Virginia
Copper Harbor
Houghton
Two Harbors
Ontonagon
Duluth
35
Marquette
Munising
Ottawa
National Forest
Hiawatha
Hayward
Chequamegon-Nicolet
National Forest
Eagle River
Escanaba
Rhinelander
Rice Lake
Wausau
MENOMINEE
RESERVATION
Eau Claire
WISCONSIN
Green Bay
Stevens Point
Appleton
43
Lake Michigan
Rochester
Winona
Oshkosh
90
La Crosse
Wisconsin
Sheboygan
Map data ©2022 Google
Google
48
47
46
45
44
Latitude
−92
−90
−88
Longitude


Alnus incana
Amelanchier laevis
Acer rubrum
Acer saccharum
Acer spicatum
Betula alleghaniensis
Betula papyrifera
Cornus alternifolia
Corylus cornuta
Dirca palustris
Fraxinus nigra
Ostrya virginiana
Populus grandidentata
Populus tremuloides
Prunus serotina
Prunus virginiana
Quercus rubra
Ribes cynosbati
Tilia americana
Ulmus americana
Density
0
1
2
3
0.0
0.5
1.0
1.5
2.0
2.5
Synonymous substitutions per synonymous site (Ks)


Asterids
Rosids-Malvids
Rosids-Fabids
Saxifragales
N1_69
N0_1843
N4_17
N7_50
N3_62
N2_98
N5_14
N8_33
N9_77
N11_331
N13_524
N12_240
N14_74
N16_324
N15_160
N6_886
N10_184
Cornus_alternifolia
Fraxinus_nigra
Acer_rubrum
Acer_saccharum
Acer_spicatum
Tilia_americana
Dirca_palustris
Prunus_virginiana
Prunus_serotina
Amelanchier_laevis
Ulmus_americana
Betula_alleghaniensis
Betula_papyrifera
Alnus_incana
Corylus_cornuta
Ostrya_virginiana
Quercus_rubra
Ribes_cynosbati
0
25
50
75
100
125


a
Gene families with a size < 5 or Ka/Ks ≤ 1
Gene families with a size ≥ 5 and Ka/Ks > 1
Alnus incana
Amelanchier laevis
Acer rubrum
Acer saccharum
Acer spicatum
Betula alleghaniensis
Betula papyrifera
Cornus alternifolia
Corylus cornuta
Dirca palustris
Fraxinus nigra
Ostrya virginiana
Populus grandidentata
Populus tremuloides
Prunus serotina
Prunus virginiana
Quercus rubra
Ribes cynosbati
Tilia americana
Ulmus americana
mean Ka/Ks per gene family
Number of genes per gene family
b
Plant–pathogen interaction
Glycosphingolipid biosynthesis – ganglio series
Glycosaminoglycan degradation
Zeatin biosynthesis
Isoflavonoid biosynthesis
Flavone and flavonol biosynthesis
Ribosome
Other glycan degradation
Sesquiterpenoid and triterpenoid biosynthesis
MAPK signaling pathway – plant
Galactose metabolism
Arginine and proline metabolism
Monoterpenoid biosynthesis
Iterpenoid biosynthesis
Flavonoid biosynthesis
Anthocyanin biosynthesis
Stilbenoid, diarylheptanoid and gingerol biosynthesis
Biosynthesis of amino acids
ABC transporters
Autophagy – other
Glycolysis/Gluconeogenesis
Pentose and glucuronate interconversions
Fatty acid biosynthesis
Fatty acid elongation
Cutin, suberine and wax biosynthesis
Pyrimidine metabolism
Phosphonate and phosphinate metabolism
Glutathione metabolism
Other types of O–glycan biosynthesis
Ether lipid metabolism
Sphingolipid metabolism
Carbon fixation in photosynthetic organisms
Limonene and pinene degradation
Carotenoid biosynthesis
Fatty acid metabolism
Protein export
Nucleotide excision repair
Peroxisome
A_inca
A_laev
A_rubr
A_sacc
A_spic
B_alle
B_papy
C_alte
C_corn
D_palu
F_nigr
O_virg
P_gran
P_sero
P_trem
P_virg
Q_rubr
R_cyno
T_amer
U_amer
0.04
0.03
0.02
0.01


a
Cornus_alternifolia
Fraxinus_nigra
Dirca_palustris
Tilia_americana
Acer_saccharum
Acer_spicatum
Amelanchier_laevis
Quercus_rubra
Betula_papyrifera
Alnus_incana
Corylus_cornuta
Ostrya_virginiana
Ribes_cynosbati
0.0
0.1
0.2
0.3
0.4
Percentage
Contraction
Expansion
b
Asterids
Rosids-Malvids
Rosids-Fabids
Saxifragales
Cornus_alternifolia 81 [+77/-4]
Fraxinus_nigra 95 [+82/-13]
Dirca_palustris 54 [+34/-20]
Tilia_americana 58 [+54/-4]
Acer_saccharum 92 [+61/-31]
Acer_spicatum 121 [+101/-20]
Amelanchier_laevis 57 [+45/-12]
Quercus_rubra 120 [+105/-15]
Betula_papyrifera 139 [+107/-32]
Alnus_incana 181 [+131/-50]
Corylus_cornuta 112 [+57/-55]
Ostrya_virginiana 145 [+102/-43]
Ribes_cynosbati 71 [+62/-9]
14 [+14/-0]
12 [+5/-7]
32 [+3/-29]
41 [+38/-3]
17 [+17/-0]
27 [+27/-0]
51 [+50/-1]
46 [+27/-19]
53 [+44/-9]
31 [+18/-13]
18 [+18/-0]


Plant−pathogen interaction
Tryptophan metabolism
MAPK signaling pathway − plant
Pentose and glucuronate interconversions
ABC transporters
SNARE interactions in vesicular transport
Pyrimidine metabolism
Zeatin biosynthesis
Glucosinolate biosynthesis
Aminoacyl−tRNA biosynthesis
Cornus_alternifolia
Fraxinus_nigra
Dirca_palustris
Tilia_americana
Acer_saccharum
Acer_spicatum
Betula_papyrifera
Alnus_incana
Corylus_cornuta
Ostrya_virginiana
Quercus_rubra
Amelanchier_laevis
Ribes_cynosbati