New Phytologist

Research

Evolutionary network genomics of wood formation in a phylogenetic survey of angiosperm forest trees

Matthew Zinkgraf[1,2]*, Shu-Tang Zhao[3]*, Courtney Canning[1], Suzanne Gerttula[1], Meng-Zhu Lu[3], Vladimir Filkov[4] and Andrew Groover[1,5]

[1]USDA Forest Service, Pacific Southwest Research Station, Davis, CA 95618, USA; [2]College of Science and Engineering, Western Washington University, Bellingham, WA 98225-9063, USA; [3]State Key Laboratory of Tree Genetics and Breeding, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China; [4]Computer Science, University of California Davis, Davis, CA 95618, USA; [5]Department of Plant Biology, University of California Davis, Davis, CA 95616, USA

Authors for correspondence:
Andrew Groover
Tel: +1 530 759 1738
Email: agroover@fs.fed.us

Vladimir Filkov
Tel: +1 530 752 8393
Email: filkov@cs.ucdavis.edu

Meng-Zhu Lu
Tel: +86 571 6383 9132
Email: lumz@caf.ac.cn

Received: *20 May 2020*
Accepted: *6 July 2020*

New Phytologist (2020)
doi: 10.1111/nph.16819

Key words: computational biology, evolutionary genomics, forest trees, gene coexpression networks, plants, wood formation.

Summary

- Wood formation was present in early angiosperms, but has been highly modified through evolution to generate the anatomical diversity seen in extant angiosperm lineages. In this project, we modeled changes in gene coexpression relationships associated with the evolution of wood formation in a phylogenetic survey of 13 angiosperm tree species.
- Gravitropic stimulation was used as an experimental treatment to alter wood formation and also perturb gene expression. Gene transcript abundances were determined using RNA sequencing of developing wood tissues from upright trees, and from the top (tension wood) and bottom (opposite wood) tissues of gravistimulated trees.
- A network-based approach was employed to align gene coexpression networks across species based on orthologous relationships. A large-scale, multilayer network was modeled that identified both lineage-specific gene coexpression modules and modules conserved across multiple species. Functional annotation and analysis of modules identified specific regulatory processes associated with conserved modules, including regulation of hormones, protein phosphorylation, meristem development and epigenetic processes.
- Our results provide novel insights into the evolution and development of wood formation, and demonstrate the ability to identify biological processes and genes important for the evolution of a foundational trait in nonmodel, undomesticated forest trees.

Introduction

A fundamental goal of biology is to describe the evolution of genetic mechanisms regulating phenotypic traits, including how those mechanisms are modified to produce phenotypic diversity and novelty. In plants, the evolution and development of simple traits such as flower morphology have been modeled based on diversification of a small number of well-studied homeotic genes (Chanderbali *et al.*, 2016). But the majority of traits that are of ecological or economic importance for plants are complex, quantitative traits, influenced by large numbers of genes (Holland, 2007; Ingvarsson & Street, 2011). For example, with forest trees, economic traits including yield (Bastiaanse *et al.*, 2019), wood biochemistry (Wierzbicki *et al.*, 2019) and architecture (Wu & Stettler, 1994) are all quantitative. Likewise ecological traits of trees, including disease and insect interactions (Newcombe & Bradshaw Jr, 1996), phenology (Bradshaw-Jr & Stettler, 1995) and response to drought (Street *et al.*, 2006; Rodriguez-Zaccaro & Groover, 2019), are all quantitative. A further challenge is posed by nonmodel plants, including undomesticated trees, for which candidate genes and tractable research approaches are often lacking (Abzhanov *et al.*, 2008). Evolutionary genomic approaches that exploit recent advances in DNA sequencing technologies and computational biology can now be used to investigate these previously intractable problems (Groover & Cronk, 2013).

Conceptually, comparative and evolutionary genomic approaches can be used to uncover fundamental properties of phenotypic traits (Wray, 2013). A general approach is to seek correlations between phenotypic traits and changes in DNA or protein sequence within a phylogenetic context. For example, phylostratigraphic analysis seeks to identify the last common ancestor that contains a gene or gene family of interest (Domazet-Lošo *et al.*, 2007). This information can then be used for hypothesis generation regarding co-occurrence of a gene and a new trait. For example the comparative genomics project, PLAZA 4.0 (Van Bel *et al.*, 2017), integrates phylogenetic relationships with protein similarity and sequence-derived functional

*These authors contributed equally to this work.

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

annotations to analyze the evolution of groups of genes in plant species. However, additional approaches are needed to fully understand the evolution of traits, as evolutionary innovations arise not only by the appearance of new genes, but also through the co-option and modification (e.g. change in expression pattern) of existing genes and genetic mechanisms (Abzhanov *et al.*, 2008). Additionally, phenotypic traits are not conditioned by individual genes, but rather by the collective interaction of many genes. It is thus ultimately desirable to model the interaction of genes underlying quantitative trait variation, including how those interactions are modified to create phenotypic diversity during evolution.

Gene coexpression analysis within a phylogenetic context can potentially reveal putative ancestral mechanisms underlying a phenotypic trait, by identifying orthologous groups of genes whose coexpression relationships are maintained across species (Ruprecht *et al.*, 2017b). Analysis of such gene modules can ultimately describe general features of conserved mechanism, and how those mechanisms have been modified through evolution and speciation to create new traits and trait diversity (Masalia *et al.*, 2017; Ruprecht *et al.*, 2017a,b). DNA sequencing technologies enable comprehensive cataloging and quantification of gene expression in species and tissues of interest using RNA sequencing. The challenge now is how to use these extensive datasets to model the properties of individual genes and functional groups of genes associated with traits within a phylogenetic context. Coexpression analyses cluster genes into modules based on correlated expression levels across samples (Zhang & Horvath, 2005). Experimentally, gene expression can be perturbed for coexpression approaches through experimental treatments or genetic modifications, or else through examination of different cell or tissue types relevant to the trait being studied (Langfelder & Horvath, 2008). Gene modules have the advantage of enabling functional analysis, for example through determination of enrichment of genes within modules associated with specific cellular processes, localization, or biochemical function (e.g. gene ontology (GO) analysis; Langfelder & Horvath, 2008). Recently, coexpression analysis has been applied to wood formation in poplar trees, and used to identify gene modules associated with specific biological functions and correlated to complex traits of ecological and economic importance (Gerttula *et al.*, 2015; Sundell *et al.*, 2016; Zinkgraf *et al.*, 2017, 2018a).

As a group, the angiosperms (flowering plants) encompass large numbers of species characterized by surprising degrees of variation and diverse phenotypic traits (Wang *et al.*, 2009). Indeed, the rapid radiation and diversification of the angiosperms posed an 'abominable mystery' to Darwin (Friedman, 2009). Advances in comparative genomics and phylogenetics have provided important clues as to the nature of angiosperm evolution (Chase *et al.*, 2016). In general, the evolutionary history of angiosperm genomes highlights a surprisingly plastic nature, with many lineages' genomes being shaped by polyploidization and gene duplications, subsequent selective retention and fractionation of gene content, tandem duplication, and translocations (Soltis *et al.*, 2009; Jiao *et al.*, 2011; AmborellaGenomeProject, 2013). As a result, chromosomal number and syntenic relationships are highly variable across angiosperm genomes. How these genomes have undergone such radical changes without catastrophic consequences or more dramatic effects on phenotypes is in some regards a new abominable mystery, and presents the challenge of comprehensively describing the genetic changes and evolutionary mechanisms that generated the observed diversity in extinct and extant angiosperms.

Wood formation is an ecologically and economically important process that was present in early angiosperms (Sinnott & Bailey, 1915; Spicer & Groover, 2010). Wood formation as produced by a bifacial cambial meristem has arisen independently in the evolution of land plants, but it appears likely that wood formation in angiosperms ultimately derives from progymnosperm progenitors to the angiosperms (Spicer & Groover, 2010; Tomescu & Groover, 2019). Among extant angiosperms, *Amborella trichopoda* represents the most basal lineage (AmborellaGenomeProject, 2013). *Amborella* grows as a shrub or small tree, and possesses a bifacial cambium that produces wood with 'primitive' features, including lacking the water-conducting vessel elements found in more derived angiosperms (Carlquist & Schneider, 2001). Notably, wood formation has undergone extensive modification during angiosperm evolution (Spicer & Groover, 2010). The mechanical properties of wood have been modified to enable an array of growth forms ranging from massive forest trees to lianas (Chery *et al.*, 2020; Groover, 2020). At the same time, wood is the water-conducting tissue of woody stems, and wood anatomy has been modified in various lineages to exploit habitats with extremes in water stress (Rodriguez-Zaccaro & Groover, 2019). Biochemical variation in angiosperm wood is also abundant, affecting both ecological traits as well as lumber, bioenergy and pulp traits critical to forest industries (Higuchi, 2012).

Wood formation in angiosperms is highly modified in response to gravistimulation (Groover, 2016). For example when an angiosperm tree is displaced from the vertical (e.g. by wind, avalanche, or erosion), the tree will produce a highly modified 'tension wood' on the upper facing side of the stem (Timell, 1986). Tension wood is capable of creating tremendous contractile force, pulling the tree upright against the force of gravity (Mellerowicz *et al.*, 2008). Gravistimulation and tension wood formation are accompanied by complex changes in gene expression and can be used as a controlled experimental treatment to perturb gene expression during wood formation. For example, gravistimulation has been used to perturb wood gene formation, allowing identification of biologically functional gene coexpression modules associated with specific biological processes underlying wood formation in the model *Populus* (Gerttula *et al.*, 2015; Zinkgraf *et al.*, 2018a).

Angiosperm trees pose challenges for evolutionary genomic studies, as they are typically highly outcrossing, heterozygous, undomesticated, and have complex genomes shaped by varied evolutionary histories (White *et al.*, 2007). In this report, we took an evolutionary genomics approach based on gene coexpression relationships in a phylogenetic survey of angiosperm trees from North America and Asia. We report the identification and characterization of conserved coexpression gene modules underlying

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

wood formation, including conserved modules highly enriched for genes of specific functions. Lineage-specific modules were also identified that point to potential mechanisms involved in diversification of wood development in different lineages. Together these results demonstrate that dissection of the evolution and development of a complex trait in trees is now tractable, and they provide specific new insights into the evolution of wood formation in trees.

Materials and Methods

Plant material and sample collection

The wood formation experiments were conducted at the Institute of Forest Genetics in Placerville, CA, USA, and at the Chinese Academy of Forestry in Beijing, China, during July 2015. A total of 13 species were included in the experiments and are listed in Table 1. Experiments for each species were started with upright grown trees that were 2–3 years old with two to three replicates per species that were randomly assigned to control (upright grown) and gravistimulation groups. Gravistimulation was conducted by placing potted plants horizontally on a glasshouse bench for 48 h as previously described (Gerttula *et al.*, 2015).

Woody tissues were collected from the middle third of the stem and harvested by lightly scraping the xylem side of a debarked stem using a double-sided razor blade. For gravistimulation experiments, developing xylem was collected from the upper surface (tension wood) and lower surface (opposite wood) of the leaning stem. For upright-grown control plants, two samples of developing xylem were sampled per stem by vertically dividing the stem in half. All tissue samples were immediately flash-frozen in liquid nitrogen and stored at −80°C.

Table 1 List of woody species included in study.

Species	Order	Native range	Experimental location[a,b]
Liriodendron chinense	Magnoliales	Asia	Bejing, China
Liriodendron tulipifera	Magnoliales	North America	CA, USA
Liquidambar formosana	Saxifragales	Asia	Bejing, China
Liquidambar styraciflua	Saxifragales	North America	CA, USA
Eucalyptus grandis	Myrtales	Australia	CA, USA
Salix aegyptiaca	Malpighiales	Southwest Asia	CA, USA
Salix babylonica	Malpighiales	Northern China	CA, USA
Salix chaenomeloides	Malpighiales	Asia	CA, USA
Salix purpuria	Malpighiales	North America	CA, USA
Salix suchowensis	Malpighiales	Asia	Bejing, China
Populus tomentosa	Malpighiales	Asia	Bejing, China
Populus tremuloidies	Malpighiales	North America	CA, USA
Populus trichocarpa × deltoides	Malpighiales	North America	CA, USA

[a]Institute of Forest Genetics, Pacific Southwest Research Station, USDA Forest Service, Placerville, CA 95667, USA.
[b]State Key Laboratory of Tree Genetics and Breeding, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China.

RNA library construction and sequencing

Frozen tissue was ground in liquid nitrogen to a fine powder and stored at −80°C. Total RNA was extracted from the frozen tissue using TRIzol reagent (Invitrogen) and cleaned using a Qiagen RNeasy on-column DNase treatment (Qiagen) following the manufacturer protocols. The quantity of the RNA was measured using Qubit V2 (Invitrogen) and quality assessed using Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA).

North American samples were sequenced using two approaches. First, multiplexed libraries generated from individual samples were sequenced using 50 bp single-end runs and utilized for expression analysis. Second, single pooled libraries for each species containing equal proportions of library product from each sample (normal, tension and opposite wood) were sequences using 150 bp paired-end runs and used to assemble species-specific transcriptomes. All library preparation for North American samples were generated using KAPA mRNA Hyper Prep Kit (KAPA BioSystems, Indianapolis, IN, USA) following the manufacturer's instructions and multiplexed using 12 unique adapter sequences (KAPA BioSystems). Individual and pooled libraries were sequenced on an Illumina HiSeq 4000 system at the QB3 Vincent J. Coates Genomics Sequencing Laboratory in October 2016.

Library preparations for all Chinese samples were generated using NEBNext® Ultra™ RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) and individual samples were multiplexed using 12 unique adapter sequences. Multiplexed libraries were sequenced using 150 bp paired-end sequencing on an Illumina HiSeq 4000 system at Novogene in Beijing, China, in April 2016. The frequency of sequencing read counts for libraries used in the analyses presented here are shown in Fig. S1.

Transcriptome assembly and annotation

To construct transcriptome assemblies for the eight species that were lacking genomic-level resources (Table S1), we uniformly processed the 150 bp paired-end reads for each species using the following steps. First, adaptor contaminations were removed using Scythe v.0.991 (https://github.com/vsbuffalo/scythe) and reads were trimmed using Sickle v.1.33 using an average phred quality of 30 and minimum length of 100 bp (https://github.com/najoshi/sickle). Second, cleaned reads were assembled using Trinity v.2.2.0 with a minimum contig length of 350 bp and default parameters (Grabherr *et al.*, 2011). *De novo* assemblies were generated for species from the genera *Liquidambar* and *Liriodendron*. Genome-guided assemblies were generated for the other species, using *Populus trichocarpa* v.3.0 to guide the *Populus tomentosa* assembly and *Salix purpurea* v.1.0 to guide the *Salix* assemblies. Third, we applied the tr2aacds analysis pipeline with default setting from EvidentialGene (http://arthropods.eugenes.org/EvidentialGene/evigene/) to identify an optimal set of transcripts, and to predict the coding and amino acid sequences. The tr2aacds pipeline reduces the complexity of the transcriptome assembly by filtering out: partial or fragmented transcripts; transcripts of high sequence similarity; and transcripts with low coding potential. Fourth, the predicted

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

protein sequences for each transcriptome assembly were annotated using best *Arabidopsis* BLAST hits and determined using TAIR10 peptide sequences (https://www.arabidopsis.org/) and BLASTP v.2.2.25 (Altschul *et al.*, 1997) with an *e*-value cutoff of less than 1e–05. The quality of each assembly was assessed using N50 statistics, the percentage of reads that were represented by each assembly and single-copy orthologs from BUSCO. To do this, cleaned reads were mapped back onto an assembly using BOWTIE2 v.2.2.9 (Langmead & Salzberg, 2012) and the percentage was quantified using a Trinity utility (SAM_nameSorted_to_uniq_count_stats.pl). The completeness of the new transcriptome assemblies were assessed using BUSCO (v.3.1.0) (Simão *et al.*, 2015) and single-copy orthology from the embryophyta (version odb10) (Kriventseva *et al.*, 2019) dataset.

Orthologous relationships of protein sequences between all species were calculated using INPARANOID v.4.1 (Remm *et al.*, 2001) with the BLOSUM62 substitution matrix and default settings. Clusters of orthologous groups across all species were identified using the Markov clustering algorithm (MCL) on the orthology and paralogy relationships generated from INPARANOID. MCL clustering was performed using an inflation value of 1.5 and overlap of orthologous groups was visualized using OrthoVenn (Wang *et al.*, 2015).

Gene expression and comparative network analysis

Gene expression for each tissue sample (normal, tension and opposite wood) across all 13 species were quantified using 50 bp single-end reads for North American experiments and 150 bp for Chinese experiments (Table 1). RNA-seq libraries were demultiplexed and uniformly processed using the following steps. First, adaptor contaminations were removed using SCYTHE v.0.991 and reads were trimmed using SICKLE v.1.33 using an average phred quality of 20 and minimum length of 30 bp. Second, cleaned sequencing reads were mapped using HISAT2 v.2.1.0 (Anders *et al.*, 2014; Kim *et al.*, 2015) to each species assembly that had organellar contamination removed. Potential chloroplast and mitochondria gene models in each transcriptome and genome assembly were identified using TAIR10 peptide sequences and BLASTP with an e-value cutoff of 1e–05. Third, uniquely mapped reads were counted for each gene model using HTSEQ v.0.9.1 (Anders *et al.*, 2014) for nonstranded reads with default settings. Fourth, normalized expression for each gene model was calculated using the TMM normalization method in EDGER v.3.22.3 (Robinson *et al.*, 2010) and output as reads per kilobase per million reads (RPKM).

To generate coexpression networks from our 13 species experiments, we applied FASTOC (Zinkgraf *et al.*, 2018b), a multilayer network approach developed by expanding on an existing tool, ORTHOCLUST (Yan *et al.*, 2014). FASTOC identifies gene modules conserved across species, and unique to phylogenetic lineages and individual species. It comprises the following steps. First, within-species coexpression relationships were identified by finding and connecting to a gene the 10 genes with the most highly correlated, three-tissue sample expression profiles. The correlations were determined using normalized RNA-seq expression (RPKM) and Pearson's correlation coefficient analysis. We only used the top 10 most correlated profiles to make the computation tractable. Second, the Louvain community detection method was applied to the above network to find compact clusters (i.e. modules) of genes. The Louvain method is a heuristic that builds communities based on random starting points in large networks and assigns genes (nodes) to communities to optimize community modularity. Using many Louvain runs ($n = 850$), we calculated how often genes coappear in the same Louvain communities. Groups of genes (modules) that display high coappearance in Louvain communities were identified using hierarchical clustering and dynamic tree cutting (Langfelder & Horvath, 2008). Third, weighted orthologous gene relationships were used to align the within-species modules across all species. Between-species links in the network were determined using all pairwise orthologous relationships and weighted based on the normalized number of orthologs per gene.

Functional enrichment of modules

To ascertain the functional role of each conserved module, two approaches were utilized. First, GO enrichment analysis of conserved modules was performed using *Arabidopsis* best BLAST hit of gene models, and significant ($P < 0.01$) enrichment of GO terms was calculated using GOSTATS v.2.50.0 (Falcon & Gentleman, 2007). *Arabidopsis* annotations for TAIR10 were used for the GO enrichment and downloaded from AGRIGO (http://bioinfo.cau.edu.cn/agriGO/). Visualization and parsing of the GO terms were conducted using the package TREEGO (https://github.com/mzinkgraf/treeGO). GO terms from the Biological Processes (BP) were parsed for functions associated with hormones, peroxisome, protein localization, saccharides, cell walls, meristem/cambium, phosphorylation, epigenetics, development and gravitropism. Second, we compiled a list of genes and orthologs (4345 genes) that have previously been implicated in vascular cambium development, xylem formation and secondary growth (Nieminen *et al.*, 2015; Ye & Zhong, 2015; Roodt *et al.*, 2017). This list of genes was based on direct molecular experimentation, gene expression profiling in vascular tissues and bioinformatic approaches across 30 + angiosperm species. As our understanding of genes associated with wood formation and functional annotations has been highly influenced by the model plant *Arabidopsis* and *A. thaliana* gene IDs are a common output of gene annotation pipelines, we used *A. thaliana* gene IDs to conduct this analysis. Using these genes, we tested if conserved modules were overrepresented with wood-related genes using a hypergeometric test and adjusted *P*-value significance to control for multiple testing using the false discovery rate (FDR; Benjamini & Hochberg, 1995) of FDR < 0.05.

Phylogenetic analysis

The evolutionary relationship between the 13 woody species (Table 1) and *Amborella trichopoda* was generated using 948 single-copy gene trees that could be identified in all species using MCL clustering on the orthology and paralogy relationships generated from INPARANOID. For each gene, DNA sequence alignments were generated using MUSCLE and maximum likelihood

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

trees were generated using the GTRGAMMA model in RAxML (v.8.2.12) (Stamatakis, 2014) with 200 bootstraps. A species tree was constructed using the coalescence approach ASTRAL-MP (v.5.14.3) (Yin *et al.*, 2019), with the 948 best maximum likelihood trees for the 948 gene trees, and significance was obtained based on 200 bootstraps.

Results

Experimental design and tree species analyzed

As summarized in Table 1, a total of 13 forest tree species were selected from five diverse angiosperm genera and included in the experiments here. As shown in an angiosperm phylogeny in Fig. 1, genera sampled include the *Liriodendron* (order Magnoliales in the Magnoliids), *Liquidambar* (order Saxifragales in the core eudicots), *Eucalyptus* (order Myratales in the Rosids), *Salix* (Family Salicaceae, order Malphighiales, in the Rosid 1 clade), and *Populus* (Family Salicaceae, order Malphighiales, in the Rosid 1 clade). Additionally, with the exception of *Eucalyptus*, all other genera included sampling of species from both North America and Asia. Although not included in the analyses presented here, these species pairs allow potential comparisons within genera as well as speciation events associated with the eastern Asian–eastern North American disjunct (Wen, 1999) (e.g. *Liriodendron tulipifera* from North America and *Liriodendron chinense* from Asia). Five *Salix* and three *Populus* species were included, allowing for comparisons at the family level (*Salix* and *Populus* are within the Salicaceae).

For each species, gravistimulation was used as an experimental treatment to perturb gene expression, by assigning replicates of each species to be placed either upright or placed horizontally. After 48 h of treatment, wood-forming tissues were harvested for RNA sequencing from 'normal wood' of upright trees, or downward-facing 'opposite wood' and upward facing 'tension wood' from horizontal trees (see the Materials and Methods section).

Transcriptome analysis

For the eight species in the experiment lacking reference genomes, reference transcriptomes were assembled (see the Materials and Methods section) and are summarized in Tables S1 and S2. Quality of transcriptome completeness was assessed using BUSCO, with 85.2 ± 7.5% (± SD) identification of complete orthologs (single and duplicated), *c.* 10% lower than the five whole genomes included in the study (Fig. S2) and consistent with high-quality assemblies.

To account for complex orthologous relationships that arise because of shared and lineage-specific gene duplications across plant lineages, clusters of orthologous groups (COGs) among all 13 species were calculated using protein sequences from reference genomes and transcriptome gene models (see the Materials and Methods section). Each orthologous group contains clusters of individual proteins, paralogs or groups of co-orthologs found in at least one species based on sequence similarity. The number of orthologous groups was similar across species and did not scale with genome size or phylogenetic distance (Fig. 2a). Comparison of COGs among representative species from each genus (Fig. 2b) show that many of the COGs were present across all species (9258) and between more closely related genera, such as *Liquidambar–Liriodendron* (2582), *Populus–Salix–Eucalyptus–Liquidambar* (2270) and *Populus–Salix* (2175). Modest numbers of COGs were specific to individual species (Fig. 2b). The total number of genes across the eight species was within a two-fold range, with the exception of *L. chinense*, which showed an unexpectedly *c.* two-fold larger total number of genes than other species. It is not clear if this is a reflection of a biological feature or of the transcriptome assembly, but recent sequencing of the *L. chinense* genome and resequencing of *L. tulipifera* (Chen *et al.*, 2019) revealed 10-fold higher nucleotide diversity in *L. chinense* than in *L. tulipifera*. High nucleotide diversity has been shown to increase the likelihood of assembly of haplotype-specific gene models (Pryszcz & Gabaldón, 2016).

To further validate our *Liriodendron* transcriptome assemblies, we compared the transcriptome gene features against unique protein-coding sequences from the recent *L. chinense* NJFU_Lchi_2.0 genome (Table S3). Based on BLASTP results for each of the transcriptome assemblies and reference protein sequences, > 90% of the reference proteins ($n = 35\,269$ proteins) had a significant (*e*-value < 1e-5) match to at least one of the transcriptome features. For each of the recovered proteins, the mean percentage alignment per reference protein was high. For example, when comparing the *L. chinense* transcriptome to the reference genome, the single longest alignment recovered 76.28 ± 23.8% (mean ± SD) and all significant alignments recovered 91.98 ± 14.3% of the amino acid sequence for reference proteins. Overall, these results are constant with our transcriptome assemblies being of high quality and robust for the analyses described in the following sections.

Comparative gene network analyses reveal conserved gene modules

A comparative network analysis was used to compare and align gene coexpression modules across species. As described in the Materials and Methods section, gene coexpression modules were identified within each species, and then genes within modules were aligned across species based on weighted orthologous relationships (Zinkgraf *et al.*, 2018b). The approach accommodated and applied weighted values of orthology not only to the one-to-one but also to the more complex paralogous relationships (e.g. one-to-many paralogs). The approach and experimental design enabled identification of gene coexpression modules that are unique to a species, unique to a taxonomic unit (e.g. a genus), or modules that are shared across all species (our primary goal). The coexpression relationships for modules shared across species have been conserved across the entire phylogenetic spectrum surveyed, and thus represent excellent candidates for genes and mechanisms that were involved in wood formation in the common ancestor of the sampled angiosperm trees.

Coexpression modules were recovered that were specific to all species, specific to families, and specific to individual species. As presented in Table 2, a total of 19 coexpression modules were identified

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

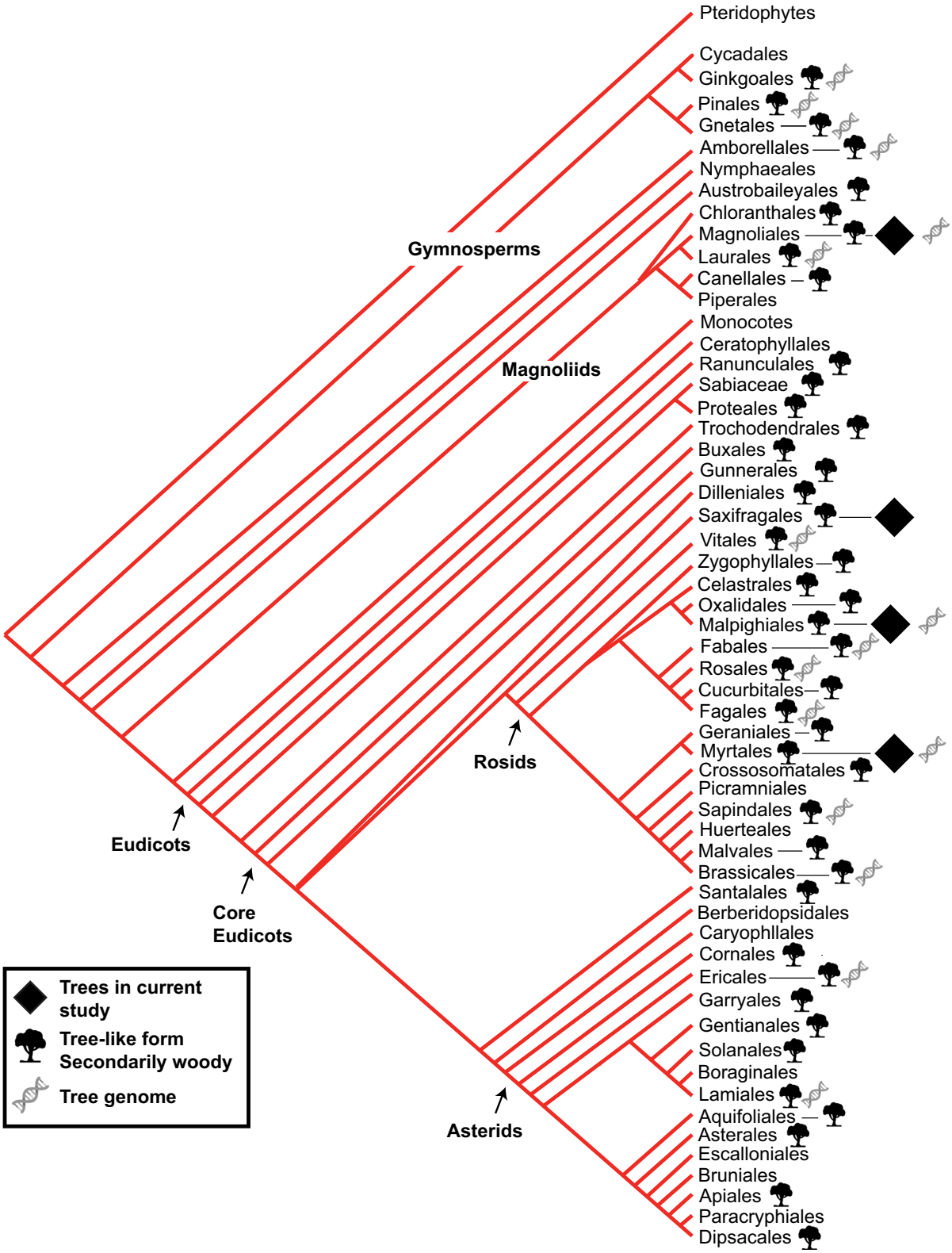

Fig. 1 Angiosperm phylogeny and genera of trees sampled in current study.

that were common to all species. Between one and four modules were specific to individual lineages, with *Liriodendron* spp. and *Liquidambar* spp. having larger numbers of shared modules (Table 2). The number of species-specific modules ranged from zero to 31, with a weak positive relationship between earlier diverging lineages and larger numbers of modules.

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

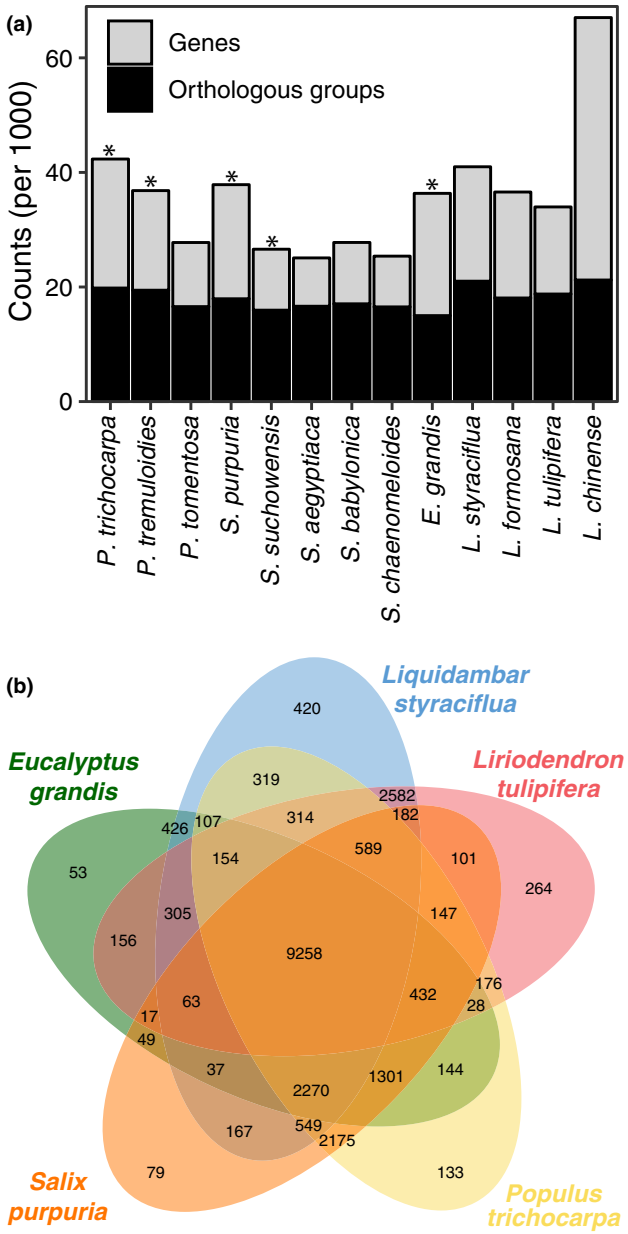

Fig. 2 (a) The number of genes and clusters of orthologous groups identified in the reference assemblies across each species. (b) Overlap of orthologous groups based on protein sequence similarity using representative species from each genera. Asterisks designate a species with whole-genome sequence.

Fig. 3 shows a graphical representation of the coexpression modules in a phylogenetic framework, illustrating their presence within and among species and visualizing spatial characteristics of modules. For each gene combination, co-occurrence in a network module was calculated using weighted scores of both coexpression and orthologous relationships (see the Materials and Methods section), with colors within the heatmap reflecting the significance of co-occurrence. The half-matrix shown is a module-to-module heatmap, with diamonds along the left side of the graph circumscribing modules of coexpressed genes in each species. The rectangles off the main diagonal are pairwise comparisons between species, with comparison of any two species represented by an intersection, as illustrated by the figure inset. Comparisons across species reveal whether modules are conserved only in individual species, or if a module is found in additional species (e.g. if a module co-occurs in one or more off-diagonal rectangles). Thus, within the graph, conserved modules can be visually inspected across species within a phylogenetic framework. For example, the inset panel in Fig. 3 highlights two (c4, c5) of the 19 conserved module between *P. tremuloidies* and *P. trichocarpa* × *deltoides*.

As summarized in Table S4, the number of genes varied both among modules within a species and for individual modules across species. The largest modules were found in *Liquidambar styraciflua* and contained 2826 genes (c11), 1560 genes (c2) and 1344 genes (c12). These modules were outliers, containing an order of magnitude more genes than the same modules in other species. Overall, the number of genes within conserved modules across all species was small, with a median of 181 genes per module and a mean (± SD) of 248.7 ± 187.1 genes per module. These smaller numbers of genes within the conserved modules reflect the strict filtering resulting from the requirement that coexpression relationships are maintained across multiple species.

Conserved modules have distinct enrichments of genes of specific function

Putative functions of modules were assigned using GO to identify enrichment of genes within specific GO categories associated with wood formation. Fig. 4 shows a heatmap of overrepresentation of genes within individual GO terms and species for the 19 conserved modules. In general, conserved modules display distinct patterns of enrichment for specific GO categories, and reject the hypothesis that genes within conserved modules are random assemblages.

Interestingly, the strongest enrichment within individual modules was for GO categories primarily corresponding to regulatory mechanisms, whereas enrichment for categories associated with the biosynthesis of secondary cell walls (which comprises the

Table 2 Number of modules identified in each species.

Species	Conserved	Lineage-specific	Species-specific	Total
Liriodendron chinense	19	4	31	54
Liriodendron tulipifera	19	4	19	42
Liquidambar formosana	19	3	8	30
Liquidambar styraciflua	19	3	15	37
Eucalyptus grandis	19	1	14	34
Salix aegyptiaca	19	1	0	20
Salix babylonica	19	1	1	21
Salix chaenomeloides	19	1	0	20
Salix purpuria	19	1	15	35
Salix suchowensis	19	1	4	24
Populus tomentosa	19	1	0	20
Populus tremuloidies	19	1	8	28
Populus trichocarpa × *deltoides*	19	1	20	40

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

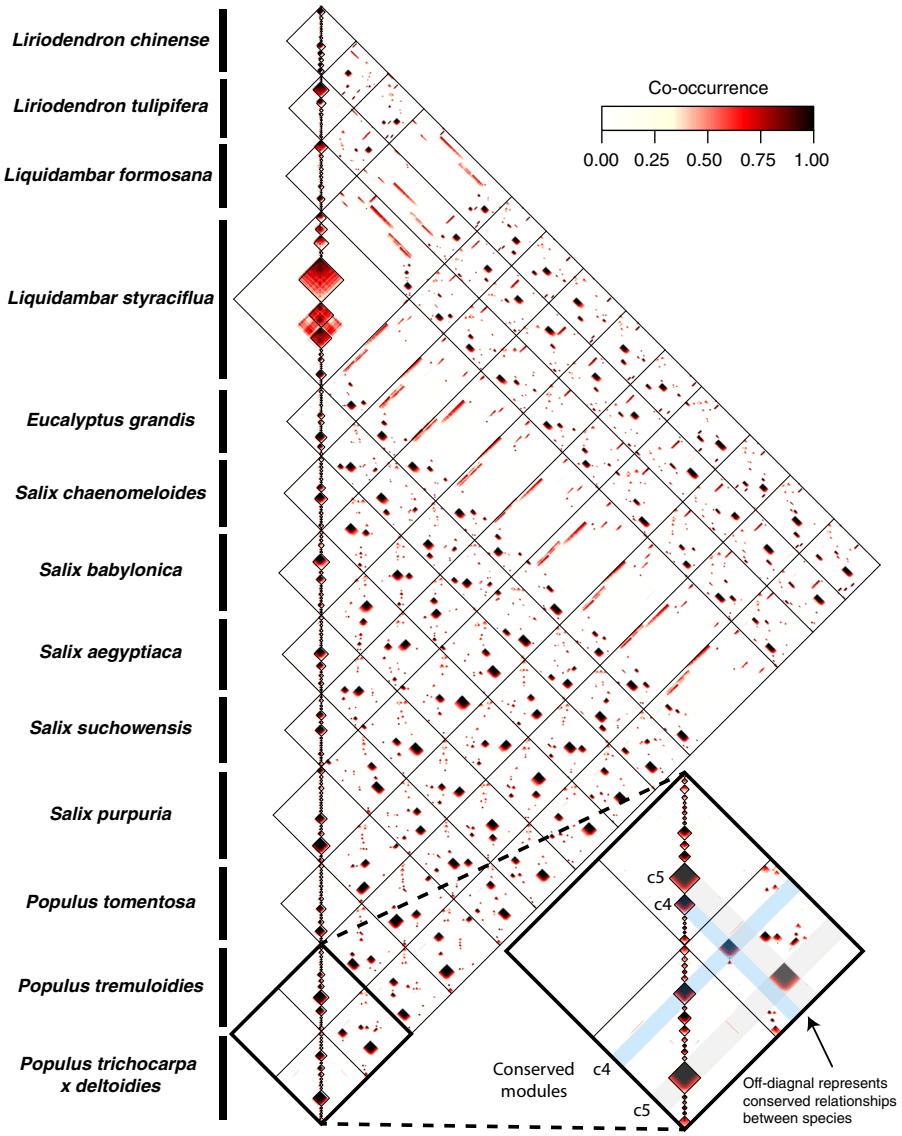

Fig. 3 Comparative network analysis between 13 woody species. Blocks along the diagonal represent genes that occur in coexpression modules within individual species. Blocks along the off-diagonal with strong co-occurrence represent conserved module relationships between species. For each gene combination, co-occurrence in a network module was calculated using weighted scores of both coexpression and orthologous relationships (see the Materials and Methods section), with colors within the heatmap reflecting the significance of co-occurrence. The inset figure shows an expanded view of *Populus tremuloidies* and *Populus trichocarpa* × *deltoides*, and highlights the conserved relationships between two (c4, c5) of the 19 conserved modules.

primary mass of wood) were distributed among many modules (Fig. 4). For example, hormones are known to be fundamental to the regulation of wood formation, and conserved module c1 shows strong enrichment for multiple hormone-related GO categories across all species. Protein localization is another fundamental process regulating wood formation (Tomescu & Groover, 2019), characterized by highly asymmetric polar growth of cell types (e.g. cambial initials), asymmetric deposition of cell walls, and polarized auxin transport (Schrader *et al.*, 2003; Dan *et al.*, 2018). Genes within GO categories associated with these processes are highly enriched in conserved module c5. Genes involved in gravitropism response, the stimulus used in this experiment to perturb gene expression, are highly enriched in conserved modules c15 and c16. Interestingly, little is known about the role of epigenetics in wood formation, but two conserved modules (c4, c5) show strong enrichment for epigenetic GO categories, related to chromatin remodeling, DNA methylation and histone modification. A potentially interesting co-occurrence of functions within modules can also be inferred; for example, conserved module c4 is enriched for meristem, phosphorylation, epigenetic and development GO categories.

Next, we tested if conserved modules were enriched for wood formation genes identified in the literature that have previously been implicated in vascular cambium development, xylem formation, and secondary growth (Nieminen *et al.*, 2015; Ye & Zhong, 2015; Roodt *et al.*, 2017). Of the 4345 *Arabidopsis* orthologs associated with wood formation, 4276 were expressed in the network and a subset of 2793 genes could be assigned to conserved modules using the best hit results from BLASTP (e-value < 1e–5). The number of wood formation genes in each conserved module ranged from 82 to 887 genes and, based on a hypergeometric test, 17 of the 19 conserved modules were significantly enriched for wood-related genes (Table S5). A full cross-tabulated list of wood formation genes (*A. thaliana* IDs) and the number of occurrences in each conserved module can be found in Table S6. Overall, the overlap between genes found in conserved modules and the curated literature provides additional support to the biological relevance of the conserved modules, and also suggests

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

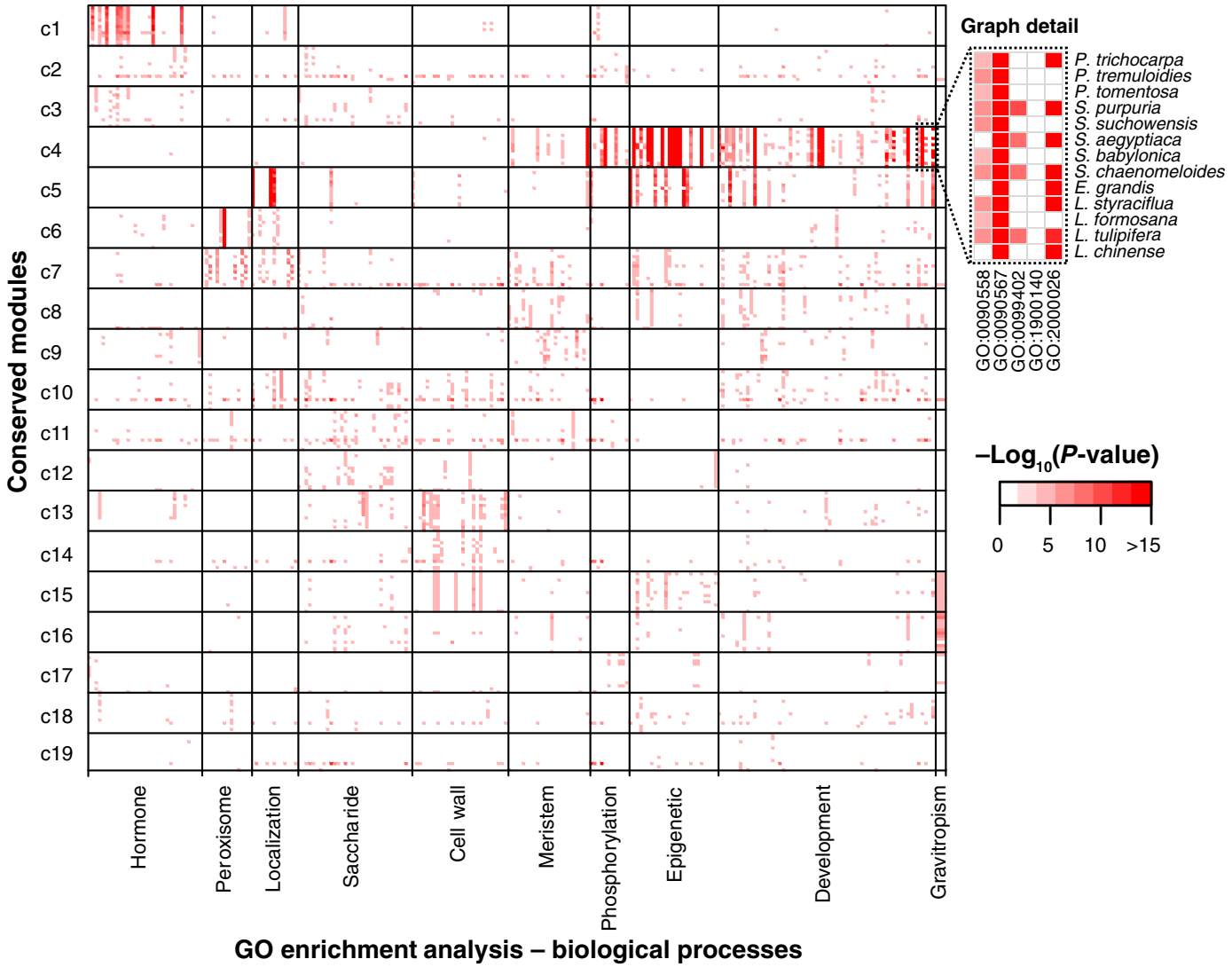

Fig. 4 Functional annotation of 19 conserved modules using gene ontology (GO) enrichment analysis with annotations from biological processes. Individual colored cells represent significance of enrichment ($-\log_{10}$(*P*-value)) of a unique GO term in a conserved module gene set from each species (see graph detail).

what genes identified from previous studies in *Arabidopsis* might be most relevant to wood formation in other angiosperm species.

Discussion

In this study, we examined the evolutionary processes underlying wood formation in angiosperm tree species through the perspective of gene coexpression relationships in a phylogenetic framework. The approach was successful in identifying gene modules comprising orthologous genes whose coexpression relationships were maintained across species, and thus through evolutionary history. Thus, these results move beyond the observations made in studies of a limited number of model species to give new insights into what genes and mechanisms may be fundamental to wood formation in angiosperms.

Our ability to detect signal, in terms of conserved coexpression modules, suggests that the plasticity of angiosperm genomes is enabled by maintenance of key gene regulatory relationships during genome rearrangements, irrespective of physical locations of genes on chromosomes. Indeed, despite extensive genome duplication, rearrangements, fractionation and other differences among species included in the study, some coexpression relationships have been maintained over the 140 Myr separating the divergence of the lineages included in the study (Magallón & Sanderson, 2005). Thus, our analyses stress the importance of the interactions of genes and gene products, irrespective of their physical location within the genome, as being a central feature of angiosperm genome evolution. The variation we observed among conserved modules in terms of size (number of genes) and gene connectivity probably reflect in part the whole-genome duplications and subsequent gene fractionation common to many angiosperm lineages (Panchy *et al.*, 2016), including the Salicoid duplication (*Populus* and *Salix* (Tuskan *et al.*, 2006); *Liriodendron*-specific duplication (Chen *et al.*, 2019)) and the

gamma whole-genome triplication shared with all eudicots (Jaillon *et al.*, 2007). Gene duplications have been shown to have a significant impact on the evolution of phenotypic variation (Panchy *et al.*, 2016) and genetic networks (Presser *et al.*, 2008; De Smet & Van de Peer, 2012), for example, when duplicated gene pairs have the potential to influence gene interactions and network topology by adding redundancy to existing pathways, dosage-based responses or divergent interactions that can lead to expression variation (Presser *et al.*, 2008). Relaxed selection that results from gene duplications allows for rewiring of the network, and suggested for *Populus* where genes with low connectivity display increased sequence diversity and natural variation in expression (Mähler *et al.*, 2017).

Our results show that it is possible to take an evolutionary coexpression approach to understand fundamental features of trait evolution, a central challenge for nonmodel species lacking genomic resources and for poorly understood traits (Abzhanov *et al.*, 2008). Technically, a major challenge of our approach initially was determining orthologous relationships across multiple species for all expressed genes, and aligning coexpression networks across species. Additionally, some of the species included in this study did not have fully sequenced genomes available, thus requiring *de novo* assembly of transcriptomes. This situation excluded any analysis of orthology based on syntenic relationships but does capture paralog and co-ortholog relationships that result from gene duplication events, such as whole-genome duplication. Our approach was enabled by expanding on an existing tool, OrthoClust (Yan *et al.*, 2014), to make it more efficient and scalable across multiple species using only gene expression data (Zinkgraf *et al.*, 2018b). Importantly, the size of the conserved modules recovered using this approach were modest. The smaller module sizes arise from additional criteria of phylogenetic conservation in defining modules, and may be an effective way of fractionating genes from analyses by removing genes that represent random associations or species-level variation. In a practical sense, this stringent selection also reduces data dimensionality for downstream analyses or candidate gene identification.

Not all genes are equally tractable to the approach here, however. In general, genes that are best captured within conserved coexpressed modules should have features including strong changes in expression in response to the experimental stimulus (in this study, gravistimulation), consistent changes in expression across species, and conserved sequence across species. Indeed, genes captured within conserved coexpression modules include orthologs of *FASCICLIN-Like Arabinogalactan 11* (*FLA11*: conserved module c13) and *ENDOTRANSGLUCOSYLASE/HYDROLASE*-like genes (*XTH15* and *XTH16-like*: conserved modules c1), that were initially identified as among the most strongly upregulated genes expressed in tension wood (Lafarguette *et al.*, 2004; Andersson-Gunneras *et al.*, 2006). Additional well-characterized genes regulating wood formation captured within conserved modules include orthologs of *VND7* (module c2), *WOX4* (module c4), *ATHB8* and *ATHB15* (module c17), *KAN1* and *KAN2* (module c9), and *KNAT7* (module c13) (Nieminen *et al.*, 2015; Ye & Zhong, 2015; Roodt *et al.*, 2017). On the other hand, genes that do not change in expression in response to the experimental stimulus, that show sporadic changes in expression across species, or that have poor sequence conservation across species are less likely to be captured within conserved coexpression modules. Examples of known genes important for wood formation that are in the network model but not captured within conserved modules include orthologs of *SHOOT MERISTEMLESS* (Groover *et al.*, 2006), *BREVI PEDILELLUS* (Du *et al.*, 2009), *MYB83* and *MYB55* (Zhong & Ye, 2015), *REVOLUTA* (Robischon *et al.*, 2011), *PHLOEM INTERCALATED WITH XYLEM* (Etchells *et al.*, 2015) and *MONOPTEROS/AUXIN RESPONSE FACTOR 5* (Brackmann *et al.*, 2018).

Studies focusing on single species (often model organisms) have contributed to our understanding of plant development but in a practical sense these studies are also limited. Typically it is not known which findings in a model species can be extrapolated to other species, as there is no measure of the evolutionary history leading to the trait under study. For example, Gerttula *et al.* (2015) conducted an analysis of gene coexpression in hybrid aspen (*Populus alba* × *P. tremula*) wood formation using multiple experimental and genetic treatments to perturb gene expression. This study identified 11 robust coexpression modules, six of which were significantly enriched in the conserved modules (Fig S3), showing that conserved genes and mechanisms can be extracted from previous research using the results reported here. Additionally, our analysis identified which of the previously known genes curated from the wood formation literature were contained within conserved modules, providing additional interpretation of previous studies. Additionally, we identified factors that were not described in previous, single-species, studies. Interestingly, genes associated with epigenetic processes were highly enriched in two conserved modules (c4 and c5; Fig. 4). Although poorly studied in trees, epigenetic regulation could be a very important point of control for developmental traits such as wood formation in long-lived trees exposed to diverse environmental changes (Bräutigam *et al.*, 2013). Together, this comparative approach for forest trees is valuable, as most trees do not have genomic or experimental resources, and thus the informed ability to extend (or not) findings from model trees to other species at different phylogenetic distance is of practical importance for forest conservation and management.

wInterestingly, individual conserved modules were identified that were highly enriched for genes with specific regulatory functions, while genes encoding proteins responsible for the massive biosynthesis of the secondary cell walls comprising the bulk of wood were more distributed across the network of modules. This is consistent with previous findings that gene function gets more specific in transcriptional regulatory networks towards the periphery of the networks, and less specific (i.e., more diffuse) within the network (Filkov & Shah, 2008). It also tracks with findings that highly specific function occurring in well-separated modules can decrease cross-talk between functional modules and increase the overall network robustness through localization of the effects of deleterious perturbations (Maslov & Sneppen,

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

2002). Biologically, these observations are consistent with regulatory genes being more localized and constrained within the overall network, while genes associated with the biosynthesis of cell walls are distributed across the network. This overall topology could enable emergent network properties that underlie the extreme flexibility required in modifying cell wall biosynthesis in response to diverse environmental conditions, and reflective of the variation in anatomy and biochemical features of wood among angiosperm species.

Future studies can further refine our broad findings here. Importantly, the validation of the general approach can now be extended, by adding additional species and treatments, and by integrating phenotypic data to explore correlations between gene modules and traits. Indeed, a significant limitation of our experimental design here was the inability to link findings to phenotypic variation. Predictions from coexpression models could be tested experimentally, using independent datasets or through the use of CRISPR of key genes from conserved and lineage-specific modules in a transformable model (e.g. *Populus*). While challenging, the existing data could be further explored for signal in coexpression relationships related to the Asian–North American floristic disjunction, and within and among genera. Integration of other data types in future experiments may allow for more direct linkages between the evolution of gene networks and phenotypes, and provide new mechanistic insights into Darwin's abominable mystery.

Acknowledgments

The authors thank Reginald King for assistance in plant propagation; Ajiinkkya Bhalerao and Tyson Lathem for assistance with bioinformatics analyses; Larry Smart for supplying cutting of *Salix purpurea*; Steve Strauss for *Eucalyptus grandis*; and the Arnold Arboretum for access to *S. aegyptiaca*, *S. babylonica* and *S. chaenomeloides*. All raw Illumina sequene data for this project have been submitted to the National Center for Biotechnology Information (NCBI) and associated with BioProject PRJNA556244. This study was supported by an Arnold Arboretum Sargent Fellowship to AG, USDA AFRI 2015-67013-22891 to AG and VF, a National Nonprofit Institute Research Grant of CAF (CAFYBB2012039) to SZ, and NSF PGRP Fellowship IOS-1402064 to MZ. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a US Department of Energy Office of Science User Facility operated under contract no. DE-AC02-05CH11231.

Author contributions

AG conceived the project. MZ, S-TZ, M-ZL, VF and AG developed and managed the project. MZ, S-TZ, SG and CC sourced and cultivated germplasm. MZ and S-TZ performed laboratory work. MZ performed bioinformatics analysis with guidance from VF. MZ, VF and AG wrote the manuscript. All authors read the manuscript and contributed to its final form. MZ and S-TZ contributed equally to this work.

ORCID

Andrew Groover https://orcid.org/0000-0002-6686-5774
Meng-Zhu Lu https://orcid.org/0000-0001-9245-1842
Matthew Zinkgraf https://orcid.org/0000-0002-7229-8016

References

Abzhanov A, Extavour CG, Groover A, Hodges SA, Hoekstra HE, Kramer EM, Monteiro A. 2008. Are we there yet? Tracking the development of new model systems. *Trends in Genetics* **24**: 353–360.

AmborellaGenomeProject. 2013. The Amborella genome and the evolution of flowering plants. *Science* **342**: 1241089.

Anders S, Pyl PT, Huber W. 2014. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Andersson-Gunneras S, Mellerowicz EJ, Love J, Segerman B, Ohmiya Y, Coutinho PM, Nilsson P, Henrissat B, Moritz T, Sundberg B. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *The Plant Journal* **45**: 144–165.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Bastiaanse H, Zinkgraf M, Canning C, Tsai H, Lieberman M, Comai L, Henry I, Groover A. 2019. A comprehensive genomic scan reveals gene dosage balance impacts on quantitative traits in *Populus* trees. *Proceedings of the National Academy of Sciences, USA* **116**: 13690–13699.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**: 289–300.

Brackmann K, Qi J, Gebert M, Jouannet V, Schlamp T, Grünwald K, Wallner E-S, Novikova DD, Levitsky VG, Agustí J *et al.* 2018. Spatial specificity of auxin responses coordinates wood formation. *Nature Communications* **9**: 875.

Bradshaw-Jr HD, Stettler RF. 1995. Molecular genetics of growth and development in *Populus*. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* **139**: 963–973.

Bräutigam K, Vining KJ, Lafon-Placette C, Fossdal CG, Mirouze M, Marcos JG, Fluch S, Fraga MF, Guevara MÁ, Abarca D *et al.* 2013. Epigenetic regulation of adaptive responses of forest tree species to the environment. *Ecology and Evolution* **3**: 399–415.

Carlquist S, Schneider KL. 2001. Vegetative anatomy of the New Caledonian endemic *Amborella trichopoda*: Relationships with the *Illiciales* and implications for vessel origin. *Pacific Science* **55**: 305–312.

Chanderbali AS, Berger BA, Howarth DG, Soltis PS, Soltis DE. 2016. Evolving ideas on the origin and evolution of flowers: New perspectives in the genomic era. *Genetics* **202**: 1255–1265.

Chase MW, Christenhusz MJM, Fay MF, Byng JW, Judd WS, Soltis DE, Mabberley DJ, Sennikov AN, Soltis PS, Stevens PF. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Botanical Journal of the Linnean Society* **181**: 1–20.

Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Q, Yang L, Sheng Y, Zhou Y. 2019. Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. *Nature plants* **5**: 18–25.

Chery JG, Pace MR, Acevedo-Rodríguez P, Specht CD, Rothfels CJ. 2020. Modifications during early plant development promote the evolution of nature's most complex woods. *Current Biology* **30**: 237–244.e232.

Dan J, Phoebe E, Noah A, Hilary N, Celia M, Rachel S. 2018. Polar auxin transport is implicated in vessel differentiation and spatial patterning during secondary growth in *Populus*. *American Journal of Botany* **105**: 186–196.

De Smet R, Van de Peer Y. 2012. Redundancy and rewiring of genetic networks following genome-wide duplication events. *Current Opinion in Plant Biology* **15**: 168–176.

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics* **23**: 533–539.

Du J, Mansfield SD, Groover AT. 2009. The *Populus* homeobox gene *ARBORKNOX2* regulates cell differentiation during secondary growth. *The Plant Journal* **60**: 1000–1014.

Etchells JP, Mishra Laxmi S, Kumar M, Campbell L, Turner SR. 2015. Wood formation in trees is increased by manipulating PXY-regulated cell division. *Current Biology* **25**: 1050–1055.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**: 257–258.

Filkov V, Shah N. 2008. A simple model of the modular structure of transcriptional regulation in yeast. *Journal of Computational Biology* **15**: 393–405.

Friedman WE. 2009. The meaning of Darwin's "abominable mystery". *American Journal of Botany* **96**: 5–21.

Gerttula S, Zinkgraf M, Muday G, Lewis D, Ibatullin FM, Brumer H, Hart F, Mansfield SD, Filkov V, Groover A. 2015. Transcriptional and hormonal regulation of gravitropism of woody stems in *Populus*. *The Plant Cell* **27**: 2800–2813.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Groover A. 2016. Gravitropisms and reaction woods of forest trees – evolution, functions and mechanisms. *New Phytologist* **211**: 790–802.

Groover A. 2020. Woody plant evolution: exceptional lianas reveal rules of woody growth. *Current Biology* **30**: R76–R78.

Groover A, Cronk Q. 2013. From Nehemiah Grew to genomics: the emerging field of evo-devo research for woody plants. *International Journal of Plant Sciences* **174**: 959–963.

Groover A, Mansfield S, DiFazio S, Dupper G, Fontana J, Millar R, Wang Y. 2006. The *Populus* homeobox gene *ARBORKNOX1* reveals overlapping mechanisms regulating the shoot apical meristem and the vascular cambium. *Plant Molecular Biology* **61**: 917–932.

Higuchi T. 2012. *Biochemistry and molecular biology of wood*. Berlin, Germany: Springer Science & Business Media.

Holland JB. 2007. Genetic architecture of complex traits in plants. *Current Opinion in Plant Biology* **10**: 156–161.

Ingvarsson PK, Street NR. 2011. Association genetics of complex traits in plants. *New Phytologist* **189**: 909–922.

Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97–100.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357–360.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM. 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47**(D1): D807–D811.

Lafarguette F, Leplé J-C, Déjardin A, Laurans F, Costa G, Lesage-Descauses M-C, Pilate G. 2004. Poplar genes encoding fasciclin-like arabinogalactan proteins are highly expressed in tension wood. *New Phytologist* **164**: 107–121.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359.

Magallón S, Sanderson M. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. *Evolution* **59**: 1653–1670.

Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLoS Genetics* **13**: e1006402.

Masalia RR, Bewick AJ, Burke JM. 2017. Connectivity in gene coexpression networks negatively correlates with rates of molecular evolution in flowering plants. *PLoS ONE* **12**: e0182289.

Maslov S, Sneppen K. 2002. Specificity and stability in topology of protein networks. *Science* **296**: 910–913.

Mellerowicz EJ, Immerzeel P, Hayashi T. 2008. Xyloglucan: the molecular muscle of trees. *Annals of Botany* **102**: 659–665.

Newcombe G, Bradshaw H Jr. 1996. Quantitative trait loci conferring resistance in hybrid poplar to *Septoria populicola*, the cause of leaf spot. *Canadian Journal of Forest Research* **26**: 1943–1950.

Nieminen K, Blomster T, Helariutta Y, Mähönen AP. 2015. Vascular Cambium Development. *The Arabidopsis Book* **13**: e0177.

Panchy N, Lehti-Shiu M, Shiu S-H. 2016. Evolution of gene duplication in plants. *Plant Physiology* **171**: 2294–2316.

Presser A, Elowitz MB, Kellis M, Kishony R. 2008. The evolutionary dynamics of the Saccharomyces cerevisiae protein interaction network after duplication. *Proceedings of the National Academy of Sciences, USA* **105**: 950–954.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research* **44**: e113.

Remm M, Storm CE, Sonnhammer EL. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* **314**: 1041–1052.

Robinson M, McCarthy DJ, Smyth G. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.

Robischon M, Du J, Miura E, Groover A. 2011. The *Populus* Class III HD ZIP, *popREVOLUTA*, influences cambium initiation and patterning of woody stems. *Plant Physiology* **155**: 1214–1225.

Rodriguez-Zaccaro FD, Groover A. 2019. Wood and water: How trees modify wood development to cope with drought. *Plants, People, Planet* **1**: 346–355.

Roodt D, Li Z, Van de Peer Y, Mizrachi E. 2017. The evolution of secondary xylem in angiosperms and its loss in monocots. *South African Journal of Botany* **109**: 367.

Ruprecht C, Proost S, Hernandez-Coronado M, Ortiz-Ramirez C, Lang D, Rensing SA, Becker JD, Vandepoele K, Mutwil M. 2017a. Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules. *The Plant Journal* **90**: 447–465.

Ruprecht C, Vaid N, Proost S, Persson S, Mutwil M. 2017b. Beyond genomics: studying evolution with gene coexpression networks. *Trends in Plant Science* **22**: 298–307.

Schrader J, Baba K, May ST, Palme K, Bennett M, Bhalerao RP, Sandberg G. 2003. Polar auxin transport in the wood-forming tissues of hybrid aspen is under simultaneous control of developmental and environmental signals. *Proceedings of the National Academy of Sciences, USA* **100**: 10096–10101.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**: 3210–3212.

Sinnott EW, Bailey IW. 1915. The evolution of herbaceous plants and its bearing on certain problems of geology and climatology. *Journal of Geology* **23**: 289–306.

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS. 2009. Polyploidy and angiosperm diversification. *American Journal of Botany* **96**: 336–348.

Spicer R, Groover A. 2010. The evolution of development of the vascular cambium and secondary growth. *New Phytologist* **186**: 577–592.

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

Street NR, Skogström O, Sjödin A, Tucker J, Rodríguez-Acosta M, Nilsson P, Jansson S, Taylor G. 2006. The genetics and genomics of the drought response in *Populus*. *The Plant Journal* **48**: 321–341.

Sundell D, Street NR, Kumar M, Mellerowicz EJ, Kucukoglu M, Johnsson C, Kumar V, Mannapperuma C, Nilsson O, Tuominen H *et al.* 2016. AspWood: High-spatial-resolution transcriptome profiles reveal uncharacterized modularity of wood formation in *Populus tremula*. *The Plant Cell* **29**: 1585–1604.

Timell TE. 1986. *Compression wood in gymnosperms*. Heidelberg, Germany: Springer-Verlag.

Tomescu AMF, Groover AT. 2019. Mosaic modularity: an updated perspective and research agenda for the evolution of vascular cambial growth. *New Phytologist* **222**: 1719–1735.

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K. 2017. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research* **46**(D1): D1190–D1196.

Wang Y, Coleman-Derr D, Chen G, Gu YQ. 2015. OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Research* **43**: W78–W84.

Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. *Proceedings of the National Academy of Sciences, USA* **106**: 3853–3858.

Wen J. 1999. Evolution of Eastern Asian and Eastern North American disjunct distributions in flowering plants. *Annual Review of Ecology and Systematics* **30**: 421–455.

White TL, Adams WT, Neale DB. 2007. *Forest genetics.* Cambridge, MA, USA: Cabi.

Wierzbicki MP, Christie N, Pinard D, Mansfield SD, Mizrachi E, Myburg AA. 2019. A systems genetics analysis in Eucalyptus reveals coordination of metabolic pathways associated with xylan modification in wood-forming tissues. *New Phytologist* **223**: 1952–1972.

Wray GA. 2013. Genomics and the evolution of phenotypic traits. *Annual Review of Ecology, Evolution, and Systematics* **44**: 51–72.

Wu R, Stettler RF. 1994. Quantitative genetics of growth and development in *Populus.* I. A three-generation comparison of tree architecture during the first 2 years of growth. *Theoretical and Applied Genetics* **89**: 1046–1054.

Yan K-K, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M. 2014. OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biology* **15**: R100.

Ye Z-H, Zhong R. 2015. Molecular control of wood formation in trees. *Journal of Experimental Botany* **66**: 4119–4131.

Yin J, Zhang C, Mirarab S. 2019. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics* **35**: 3961–3969.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network snalysis. *Statistical Applications in Genetics and Molecular Biology* **4**, Article 17. doi: 10.2202/1544-6115.1128

Zhong R, Ye Z-H. 2015. Secondary cell walls: Biosynthesis, patterned deposition and transcriptional regulation. *Plant and Cell Physiology* **56**: 195–214.

Zinkgraf M, Gerttula S, Zhao S, Filkov V, Groover A. 2018a. Transcriptional and temporal response of *Populus* stems to gravi-stimulation. *Journal of Integrative Plant Biology* **60**: 578–590.

Zinkgraf M, Groover A, Filkov V. 2018b. Reconstructing gene networks of forest trees from gene expression data: toward higher-resolution approaches. In Kalajdziski S, Ackovska N. ICT Innovations 2018. Engineering and Life Sciences. Cham, Switzerland: Springer International Publishing. 3–12.

Zinkgraf M, Liu L, Groover A, Filkov V. 2017. Identifying gene coexpression networks underlying the dynamic regulation of wood-forming tissues in Populus under diverse environmental conditions. *New Phytologist* **214**: 1464–1478.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Histogram of RNA-seq library sizes.

Fig. S2 BUSCO assessment of assembly completeness.

Fig. S3 Preservation of conserved network modules with *Populus* gravitropism network.

Table S1 Quality statistics for Illumina sequencing data.

Table S2 Quality statistics for transcriptome assemblies.

Table S3 *Liriodendron* transcriptome assembly comparisons.

Table S4 The number of genes per network module across all species.

Table S5 Enrichment of wood formation genes across each network module.

Table S6 List of wood formation genes and occurrence in each network module.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

No claim to US Government works
New Phytologist © 2020 New Phytologist Trust