New Phytologist

Identifying gene coexpression networks underlying the dynamic regulation of wood-forming tissues in *Populus* under diverse environmental conditions

Matthew Zinkgraf[1,2], Lijun Liu[1], Andrew Groover[1,3] and Vladimir Filkov[2]

[1]USDA Forest Service, Pacific Southwest Research Station, Davis, CA 95618, USA; [2]Department of Computer Science, University of California, Davis, CA 95618, USA; [3]Department of Plant Biology, University of California, Davis, CA 95618, USA

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

Received: *31 October 2016*
Accepted: *25 January 2017*

New Phytologist (2017) **214:** 1464–1478
doi: 10.1111/nph.14492

Key words: data integration, genomics, *Populus*, transcriptional networks, wood formation.

Summary

- Trees modify wood formation through integration of environmental and developmental signals in complex but poorly defined transcriptional networks, allowing trees to produce woody tissues appropriate to diverse environmental conditions.
- In order to identify relationships among genes expressed during wood formation, we integrated data from new and publically available datasets in *Populus*. These datasets were generated from woody tissue and include transcriptome profiling, transcription factor binding, DNA accessibility and genome-wide association mapping experiments. Coexpression modules were calculated, each of which contains genes showing similar expression patterns across experimental conditions, genotypes and treatments.
- Conserved gene coexpression modules (four modules totaling 8398 genes) were identified that were highly preserved across diverse environmental conditions and genetic backgrounds. Functional annotations as well as correlations with specific experimental treatments associated individual conserved modules with distinct biological processes underlying wood formation, such as cell-wall biosynthesis, meristem development and epigenetic pathways. Module genes were also enriched for DNase I hypersensitivity footprints and binding from four transcription factors associated with wood formation.
- The conserved modules are excellent candidates for modeling core developmental pathways common to wood formation in diverse environments and genotypes, and serve as testbeds for hypothesis generation and testing for future studies.

Introduction

Wood formation in trees is highly plastic and involves the dynamic integration of environmental signals into complex developmental pathways, resulting in gene expression profiles and wood tissues that are adaptive for environmental conditions (Schrader *et al.*, 2003; Guerriero *et al.*, 2014). Wood serves multiple functions including mechanical support, nutrient storage and dissemination, and water conduction, and each of these functions can be modified throughout development to mitigate environmental stress (Battipaglia *et al.*, 2014). For example, in *Populus*, saline stress results in lower cell division in the vascular cambium and the formation of 'pressure wood', which is characterized by an increase in the number of water-conducting vessels with smaller lumens that are more resistant to cavitation and water stress (Janz *et al.*, 2012). In the case of leaning stems, gravitational cues trigger the production of 'tension wood' that is characterized by increased cell division in the cambium, and production of wood containing fewer vessels and specialized tension wood fibers that create force to pull stems upright (Mellerowicz & Gorshkova, 2012; Gerttula *et al.*, 2015; Groover, 2016). In addition, wood anatomy shows population-level variation among genotypes within species, including variation in adaptive traits affecting the ability to grow in specific environments (Porth *et al.*, 2013; McKown *et al.*, 2014).

The innovation of woody growth from a bifacial cambium is believed to have evolved in lineages predating the divergence of angiosperm and gymnosperms. Within angiosperms, woody growth is an ancestral trait but has been highly modified in the various angiosperm lineages (Spicer & Groover, 2010). Currently a comprehensive description is lacking for the core set of regulatory mechanisms underlying wood development, or how they are modified to generate anatomical diversity in wood. A plausible hypothesis is that at least some of the genes and mechanisms regulating wood formation in basal angiosperms have been conserved in derived lineages. Additionally, these conserved mechanisms could be modulated by signaling mechanisms in response to environmental cues to produce anatomical variation. To test these hypotheses, the study of transcriptional regulation is

currently one of the most technically tractable and biologically relevant avenues of research.

Transcriptional regulation is a primary mechanism that ultimately integrates environmental and developmental signals during wood formation (Du & Groover, 2010). A variety of experimental approaches have been used to dissect transcriptional regulation in wood-forming tissues at levels ranging from the study of individual transcription factors to natural genetic variation. For example, individual transcription factors have been functionally characterized via transgenesis in *Populus* and shown to regulate specific aspects of cell division, cell differentiation and tissue patterning (Groover *et al.*, 2006; Yordanov *et al.*, 2010; Du *et al.*, 2011; Robischon *et al.*, 2011; Jiang *et al.*, 2014; Etchells *et al.*, 2015). Chromatin immunoprecipitation sequencing (ChIP-seq) for a limited number of transcription factors involved in wood formation revealed binding to thousands of loci for each transcription factor, underscoring the complexity of transcriptional regulation (Liu *et al.*, 2015a,b). Using mRNA-sequencing or microarrays, differentially expressed genes have been identified through comparisons of experimental treatments affecting wood development, through comparisons of different stages of wood formation (Schrader *et al.*, 2004; Bao *et al.*, 2009; Dharmawardhana *et al.*, 2010), or through comparison of wood to other tissue development (Rodgers-Melnick *et al.*, 2012). For example, comparison of expression profiles across multiple tissues in *Populus trichocarpa* allowed for the identification of genes that display tissue-specific expression, and provide an estimate of the number and function of genes involved in tissue specific pathways (Quesada *et al.*, 2008). Other approaches identified naturally occurring genetic variation for wood formation. For example, large genome-wide association studies (GWAS) in *Populus* have revealed numerous associations between genetic loci, including transcription factor-encoding genes, and wood-related traits (Porth *et al.*, 2013; McKown *et al.*, 2014) but only 40% of these associations were affiliated with genes that have *a priori* involvement in wood formation. Critically, an effective integration of data from these various genomic studies is needed to provide a more comprehensive understanding of regulatory mechanisms, and how they interact to modify gene expression and wood traits.

Coexpression network approaches have the potential to provide a framework for integrating different data types and extracting additional biological meaning through comparisons across experiments (Usadel *et al.*, 2009; Serin *et al.*, 2016). In practice, transcript levels from large numbers of genes are assayed across biological samples from multiple experimental conditions or tissues, and computational analyses are employed to cluster genes that show similar expression (i.e. high correlation) across samples into coexpression modules. Coexpressed genes cluster together often because they are involved in similar biological pathways or subject to similar regulatory pathways (D'haeseleer *et al.*, 2000). Coexpression networks also have features reflecting the biological organization of the underlying biological pathways, including scale-free topology (Carter *et al.*, 2004). Gene coexpression modules can be made biologically meaningful by overlaying them with functional annotations (e.g. gene ontology), transcription factor binding, or correlations with phenotypes. In this way, modules also provide the means for integrating different data types and providing models for dissecting complex developmental processes.

Previous studies have demonstrated the benefit of integrative analyses in resolving pathway information in various organisms, for example the ENCODE project (Kundaje *et al.*, 2015). In plants, integrative studies in *Arabidopsis* have been shown to have higher resolving power than those of the individual datasets (Lee *et al.*, 2010; Bassel *et al.*, 2012; Amrine *et al.*, 2015). Integrated approaches allow for the identification of genes that are highly correlated with the same partners across multiple experiments (conserved modules) and genes that interact with different partners in a context-specific manner (experiment-specific modules) (Rasmussen *et al.*, 2013; Shaik & Ramakrishna, 2013). This phenomenon is not unique to *Arabidopsis*; similar results also have been observed in *Populus*. For example, comparison of differentially expressed genes between pressure wood and tension wood experiments revealed that similar sets of genes were differentially regulated between these wood types (Janz *et al.*, 2012). Tension wood has been characterized by the upregulation of COBRA-like 4, fasciclin-like arabinogalactan and xyloglucan endotransglycolyase genes (Andersson-Gunneras & Mellerowicz, 2006; Gerttula *et al.*, 2015), and similar sets of genes were downregulated in pressure wood (Janz *et al.*, 2012). These results are consistent with the hypothesis that different environmental signals converge onto similar pathways, and that regulatory mechanisms integrate signals and alter expression of a core set of genes to produce context-specific developmental outcomes.

In the present study, we integrated gene expression data from multiple experimental conditions to define modules of coexpressed genes and tested the hypothesis that coexpression networks for wood-forming tissues comprise a combination of conserved and condition-specific modules. Gene modules that exhibit conserved coexpression across a variety of conditions are presented that could encompass core mechanisms of wood formation, as well as experiment-specific modules involved in modifying wood development under specific conditions. Additionally, we show that integration of different genomic data types (e.g. ChIP-seq and GWAS) into a coexpression framework is an effective means of annotating and dissecting the complex genetic regulation of wood formation across experimental and environmental conditions.

Materials and Methods

RNA-seq datasets and processing

The transcriptomic data used in the present study came from four independent experiments and are publicly available on the NCBI sequence reads archive (SRA) (Table 1). All RNA-seq experiments used in the analysis sampled recent derivatives of the vascular cambium by lightly scraping the xylem or the phloem (bark) side of a debarked stem. The first experiment sampled developing xylem tissues from hybrid aspen (*Populus alba* × *P. tremula* INRA 717-1B4) including the opposite wood and tension wood of stems that had been gravi-stimulated for

Table 1 Transcription profiling (RNA-seq) datasets used in coexpression analysis

Experiment	Organism	NCBI SRA	Number libraries	Illumina read type
Gravitropism	*Populus alba* × *P. tremula*	SRP058772	56	50 bp SE
Drought vascular tissues	*P. alba* × *P. tremula*	SRS616268–SRS616303	36	50 bp PE
Provenance	*P. trichocarpa*	SRP004333	20	50 bp PE
Woody tissues	*P. trichocarpa*	SRP028935 SRP072680	15	50 bp SE

48 h, and the normal wood of upright-grown control trees (Gerttula *et al.*, 2015). In addition, the gravitropism experiment included a fully factorial sampling of wild-type (WT) plants, two ARBORKNOX2 mutants and a gibberellic acid (GA) hormone treatment. The second experiment sampled xylem and bark vascular tissues from upright-grown trees that were well watered, drought stressed or drought recovered plants (SRS616268–SRS616303; Xue *et al.*, 2016). In addition, the drought experiment sampled both INRA 717-1B4 (WT) and a RNAi knockdown mutant of a sucrose transporter (SUT4; Potri.004G190400). The third experiment sampled developing xylem from 20 *P. trichocarpa* genotypes collected from 20 provenances ranging from 44.0°N to 59.6°N that were grown in a common garden at the University of British Columbia (Bao *et al.*, 2013). The fourth experiment sampled developing xylem and phloem from seven large, actively growing *P. trichocarpa* genotypes from a riparian site in Clatskanie, Oregon (46.1°N) and prepared using methods from Liu *et al.* (2014).

RNA-seq datasets from the earlier studies (Table 1) were downloaded from the NCBI in October of 2015 and all datasets were uniformly reprocessed using the same bioinformatics pipeline. First, adaptor contaminations were removed using Scythe v.0.950 (https://github.com/vsbuffalo/scythe) and reads were trimmed using Sickle v.1.200 in either single-end or paired-end mode with default settings (Joshi & Fass, 2011). Sequenced reads were then mapped to the *Populus* genome v.3.0 (http://www.phytozome.net/poplar.php) using TopHat v.2.0.6 (Trapnell *et al.*, 2009), and uniquely mapped reads were counted for each *Populus* gene model using HTSeq v.0.6.1p1 (Anders *et al.*, 2015) with default settings. Gene expression was calculated using the TMM normalization method in edgeR v.3.10.2 (Robinson *et al.*, 2010) and standardized expression was output as reads per kilobase per million reads (rpkm). All statistical analyses were implemented in R (R Core Team, 2015) unless stated otherwise.

ChIP-seq data and processing

ChIP-seq experiments from five transcription factors and RNA polymerase II (RNA-Pol II) were generated from vascular cambium and recent derivatives from mature *P. trichocarpa* growing in Clatskanie, Oregon, as described previously (Liu *et al.*, 2014, 2015a,b). These data describe genome-wide protein binding locations for two Class I KNOX, two Class III HD-ZIP and one BELL-like homeodomain transcription factors (Table 2). The RNA-Pol ll experiment from Liu *et al.* (2014) was reprocessed using the ENCODE standards and irreproducible discovery rate pipeline (Li *et al.*, 2011) using parameters from Liu *et al.* (2015a).

Table 2 Protein binding (ChIP-seq) datasets from vascular tissues in *Populus trichocarpa* that were used to identify regulatory links between genes

Transcription factor	v3.0 Gene model	No. of peaks	No. of target genes[a]	References
ARK1	Potri.011G011100	14 463	15 182	Liu *et al.* (2015a)
ARK2	Potri.002G113300	2287	2717	Liu *et al.* (2015b)
BLR	Potri.010G197300	5674	3909	Liu *et al.* (2015b)
PCN	Potri.001G188800	3148	4689	Liu *et al.* (2015b)
PRE	Potri.004G211300	658	318	Liu *et al.* (2015b)
RNA-Pol II		4563	1853	Liu *et al.* (2014)

[a]Target genes were designated as having a ChIP peak located within ± 1000 bp of the gene model.

Genomic coordinates of peaks from each ChIP-seq dataset were assigned to target genes based on the location of *Populus* gene models, with peaks assigned to genes if a peak was located within 1000 bp upstream and 1000 bp downstream of a gene. In addition, this algorithm allows peaks to be assigned to multiple genes because some peaks were in close proximity (≤ 1000 bp) to multiple genes and does not assign peaks to the single closest feature. The function for assigning peaks to gene features (Peaks2Genes) is available at https://github.com/mzinkgraf/ConsensusCoExpression.

DNase-seq data and processing

DNase I hypersensitivity sequencing (DNase-seq) was performed on vascular cambium and recent derivatives harvested in June 2013 by lightly scraping the debarked stem from a single mature *P. trichocarpa* located in Clatskanie, Oregon, and flash-freezing the sample in the field. DNase samples were ground to a fine powder in liquid nitrogen and nuclei were isolated using CellLytic™ PN isolation kit (Sigma-Aldrich). The isolated nuclei were resuspended in digestion buffer (10 mM Tris at pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) and digested with DNase I (Zymo Research, Irvine, CA, USA) using concentrations from 0.5 to 5.0 enzyme units for 10 min at 37°C. Digested DNA was extracted with chloroform-isopropanol, and gel size-selected to isolate 200–500-bp fragments. Library construction was performed using the TruSeq DNA Sample Prep kit (Illumina, San Diego, CA, USA) according to the manufacturer's protocol and sequenced using Illumina HiSeq 2500 in 50-bp single-end

sequencing mode. Adaptor contamination was removed from samples using SCYTHE v.0.950 and reads were trimmed using SICKLE v.1.200. Trimmed reads were mapped to v.3.0 of the *Populus* genome using BOWTIE2 v.2.0.2 (Langmead *et al.*, 2009) and only uniquely mapped reads ($q \geq 40$) were kept for further analysis. To identify footprints for each sample, F-SEQ (Boyle *et al.*, 2008) was used with a bandwidth of 300 bp and a signal threshold of two. These modified parameters have been shown to increase performance of F-SEQ (Koohy *et al.*, 2014). The quality of each DNase-seq sample was assessed using descriptive statistics (number and width of footprints) and the similarity of footprint profiles between samples was calculated using Jaccard's similarity from BEDTOOLS v.2.24.0 (Quinlan & Hall, 2010). The final set of reproducible footprints was generated by intersecting three DNase-seq samples (2.0, 3.0, 4.0 units DNase I) that displayed high similarity using DIFFBIND v.1.14.4 (Ross-Innes *et al.*, 2012). Descriptive statistics of footprints were calculated using CHIPPEAKANNO v.3.2.2 (Zhu *et al.*, 2010). To assess the chromatin structure and accessibility of genes to DNase I degradation, footprints were assigned to *Populus* gene models using the PEAKS2GENES function as described in the ChIP-seq section. DNase-seq footprints could be assigned to one of six possible categories: located within 1000 bp upstream of the transcriptional start site (TSS), overlapping the TSS, inside the gene, overlapping the transcriptional end site (TES), within 1000 bp downstream of the TES or no target gene.

Coexpression networks

Coexpression analysis and module identification were conducted for each individual RNA-seq dataset using functions from WGCNA v.1.47 (Langfelder & Horvath, 2008). For each dataset, the soft threshold was determined as that producing a > 80% model fit to scale-free topology and low mean connectivity. Experiment-specific coexpression relationships were calculated using Pearson's correlation coefficients raised to the soft threshold and grouped using hierarchical clustering of dissimilarity among the topological overlap measures (TOM). Coexpressed modules were determined using dynamic tree cutting with parameters from Gerttula *et al.* (2015), and included a minimum module size of 500 and cut height of 0.998. Dynamic tree cutting is a flexible approach to identify modules from complex hierarchical dendrograms, and the minimum module size determines the smallest number of genes on a branch that can be considered a module and the cut height controls the maximum branch height that can be joined into a cluster. Dynamic tree cutting may identify modules that have similar expression profiles (Langfelder *et al.*, 2008) and modules with correlated expression profiles (> 0.75) were collapsed because these modules contain highly coexpressed genes. Furthermore, we selected these parameters because random sampling of gravitropism samples showed that this approach to module identification was robust to outliers and produced stable modules (Gerttula *et al.*, 2015). It is possible to select parameters that generate smaller modules but these modules are not reproducible with different parameter choices or a subset of samples (Langfelder *et al.*, 2011).

Data integration

Integration of the coexpression results from the individual RNA-seq experiments was performed using consensus clustering (Langfelder & Horvath, 2007; Langfelder *et al.*, 2013). Briefly, adjacency matrices from each experiment were scaled using a 0.95 quantile transformation and consensus adjacency was calculated by combining the scaled matrices using parallel quartiles with a probability of 0.25. The final consensus network was defined by calculating the TOM of the consensus adjacency matrix. The identification of gene modules in the consensus network was performed using hierarchical clustering of the consensus TOM matrix and dynamic tree cutting of the hierarchical dendrogram with the following parameters; dendrogram cut height of 0.990, minimum module size of 300 and merge cut height of 0.25. The modules identified in the consensus network represent gene clusters that had conserved coexpression patterns across all RNA-seq experiments. Conserved modules were summarized using modules eigengene (ME) values and represent the first principle component of the standardized expression data for genes in each module (Langfelder & Horvath, 2007). Next we calculated module membership (kME) for each gene to its respective module to assess how tightly connected genes were to the ME.

Two approaches were used to determine how conserved modules relate to coexpression patterns in individual experiments. First, the change in correlation structure of ME values was compared across each RNA-seq experiment. Module eigengenes were calculated as the first principal component of the gene expression matrix for each module. Second, the preservation of conserved modules was assessed in individual datasets by pairwise cross-tabulation between conserved modules and modules from individual coexpression networks (Langfelder *et al.*, 2011). Significance of preservation for each conserved module in experiment-specific modules was obtained using a one-sided fisher exact test on the cross-tabulation between conserved modules and experiment-specific modules.

We performed coexpression analysis and consensus clustering on all possible combinations of the four RNA-seq datasets to assess how integration of multiple experiments influenced the size and resolution of coexpression networks. The effect of increasing the number of datasets on network structure, such as total number of genes in coexpression network, number of modules and average size of modules, were determined using log–linear regression models.

The biological meaning of the conserved modules was investigated using two approaches. First, MEs were correlated to experimental treatments to determine how individual conserved modules were associated with experimental perturbations such as gravi-stimulation, drought and woody tissues. Second, functional annotation of conserved modules was performed using gene ontology (GO) enrichment analysis. GO enrichment of conserved modules was performed using *Arabidopsis* best BLAST hits of *Populus* gene models and significant ($P < 0.01$) enrichment of GO terms was calculated using GOSTATS v.2.37.0 (Falcon & Gentleman, 2007). *Arabidopsis* annotations were used for the

GO enrichment because categories associated with meristem functions are not annotated in the *Populus* v.3.0 and the vascular cambium is an important meristem involved in the formation of woody tissues. The *Arabidopsis* GO annotations for TAIR10 were downloaded from agriGO (http://bioinfo.cau.edu.cn/agriGO/).

In order to determine how transcriptional regulation of gene expression and chromatin structure potentially influence conserved modules, genes from the conserved modules were tested for enrichment for binding from five transcription factors and RNA-Pol II, as well as changes in gene accessibility as measured with DNase-seq footprints. A hypergeometric distribution was used to determine the probability that a conserved module was overrepresented for target genes from each of the ChIP-seq and DNase-seq experiments, and the probability was calculated using all genes (34 361) in the consensus network.

In order to understand the effect of evolutionary processes on coexpression relationships, we assessed two levels of natural genetic variation. First, we tested if paralogous genes arising from the Salicoid whole genome duplication event co-occurred in conserved modules. The paralogous relationships of *Populus* genes were obtained from the study by Rodgers-Melnick *et al.* (2012), and co-occurrence between paralogs and conserved modules was calculated using a chi-squared test. Second, we tested if modules were enriched for single nucleotide polymorphisms (SNPs) that have previously been associated with wood chemistry (Porth *et al.*, 2013) and biomass-related (McKown *et al.*, 2014) traits in a natural population of *P. trichocarpa*. SNPs were assigned to the closest gene and a hypergeometric test was used to determine if genes associated with a specific trait were enriched compared with all the possible gene models that were represented on the 34K *Populus* SNP array (Geraldes *et al.*, 2013). The probe locations from the 34K array were obtained from McKown *et al.* (2014).

Computer code and data archiving

All R code used to generate consensus coexpression networks and integration of multiple datasets is publically available on github at https://github.com/mzinkgraf/ConsensusCoExpression. The sequences associated with the *P. trichocarpa* woody tissues RNA-seq experiments has been deposited in NCBI-SRA under SRP072680. The sequences associated with the DNase-seq experiment have been deposited under SRP072559 and the final DNase-seq footprints are available as a genomic track at http://PopGenIE.org/gbrowse.

Results

Identification of gene modules coexpressed during woody growth in diverse conditions

A primary goal of the experiments here was to identify and functionally describe coexpressed gene modules commonly associated with wood formation under diverse environmental conditions, experimental treatments, and genotypes in *Populus*. A gene module in this study is defined as a group of genes all pairs of which have highly correlated expression (i.e. coexpressed) in individual experiments or across all experiments. As shown in Fig. 1, the strategy taken was to first calculate gene coexpression networks for each individual experiment surveyed. The data for each experiment are publically available RNA-seq datasets that survey multiple tissues, genotypes and experimental perturbations involved in woody development (Table 1), and include a gravitropism and reaction wood experiment in hybrid aspen (Gerttula *et al.*, 2015), a drought experiment in hybrid aspen (Xue *et al.*, 2016), a provenance analysis of *P. trichocarpa* (Bao *et al.*, 2013), and a survey of woody tissues from naturally growing *P. trichocarpa* (Liu *et al.*, 2014). Next, consensus clustering was used to identify gene modules conserved in all experiments (i.e. modules with genes showing coexpression across all conditions). These modules were then quantified for module preservation parameters, and overlaid with functional annotations to facilitate quantitative biological interpretation.

Coexpression networks and module assignments of genes from individual experiments were calculated, and revealed large differences in coexpression patterns across experiments. Briefly, soft thresholding of individual experiments produced networks that displayed high model fit to a scale-free topology commonly observed in biological networks (Supporting Information Fig. S1a), and reduced the mean connectivity in the network by decreasing the influence of low correlations between genes (Fig. S1a). Each experiment was associated with multiple coexpressed gene modules, which were assigned arbitrary color labels that are unique to each experiment (Fig. 1). We identified 11, seven, 20 and 13 modules in the respective gravitropsim, drought, provenance and woody tissue experiments. The assignment of *Populus* genes to coexpression modules for individual experiments is summarized in Table S1. Modules in the individual experiments were assigned an arbitrary color by WGCNA as a label (note that module colors are not comparable between the individual experiments).

We next identified conserved modules of genes that were commonly coexpressed across all experiments. Such modules represent candidates for genes and mechanisms associated with core developmental processes (e.g. cell division or meristem function) that may be universally involved in wood formation irrespective of environmental conditions. Integration of the RNA-seq experiments using consensus clustering identified four modules (Fig. 2a) and were named conserved module (CM) one to four. Genes assigned to the nonconserved group (25 963 genes) in the consensus analysis represent genes that could not be clustered across all experiments and show low module membership, a measurement of the correlation of individual genes expression to the average expression of all genes in the module (Fig. S2). By contrast, genes within CM1, CM2, CM3 and CM4 showed high module membership scores (Fig. S2), and contained 2829, 2289, 553 and 2727 genes, respectively. Correlation relationships among the conserved modules were dynamic, and the connections between modules changed across each experiment, as shown in the form of a dendrogram in Fig. 2(b) and a heat map of correlations among all modules in Fig. 2(c). For example, CM3 was positively correlated with CM1 and CM2 modules in the gravitropism experiment, weakly correlated with CM1 and CM2

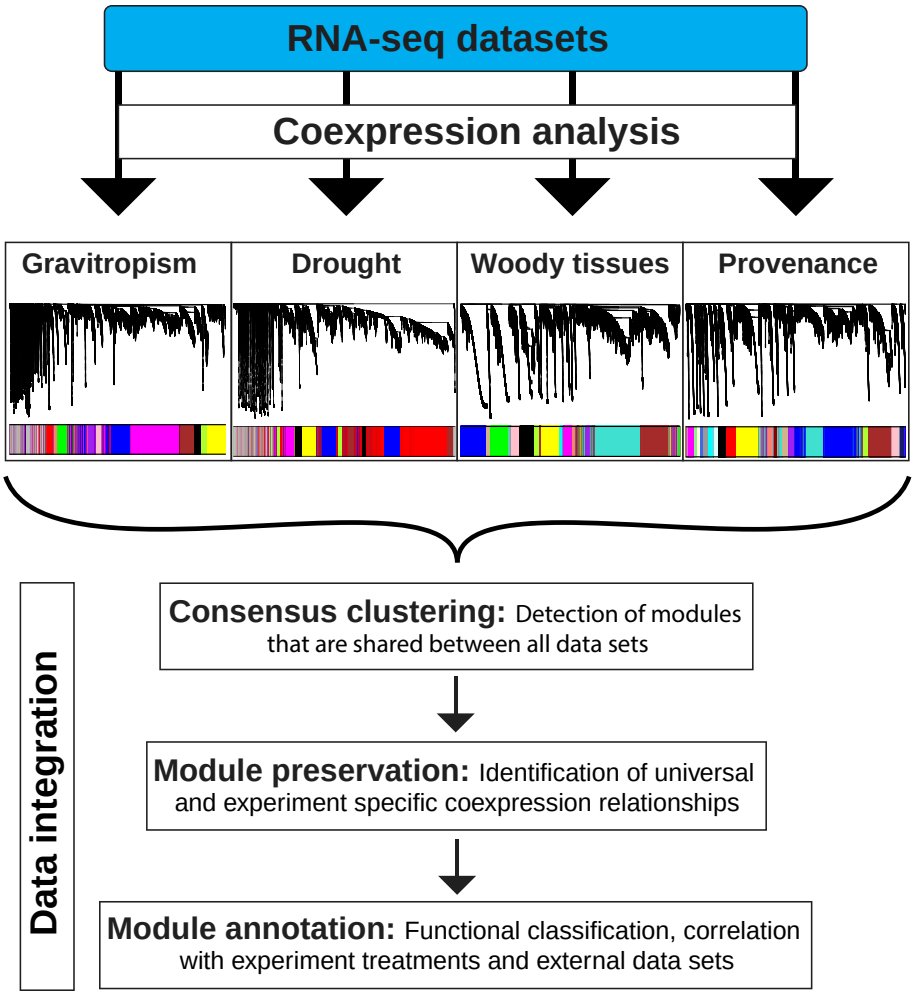

Fig. 1 Flow chart depicting the experimental approach for modeling coexpression networks underlying wood development using data integration and consensus clustering.

modules in the drought experiment and negatively correlated to CM1 and CM2 modules in the *P. trichocarpa* woody tissues experiment. Such changes in relationships among modules across experimental conditions are consistent with modules on which modifying signals in different experiments (e.g. from environmental cues and evolutionary divergence) converge.

We next quantified correlations between conserved modules and experiment-specific modules to test the hypothesis that coexpression networks from individual experiments are a combination of context-specific regulatory interactions and conserved interactions from core developmental pathways. The conserved modules were mapped onto individual networks, and showed high preservation (P-value < 0.0001) in individual coexpression networks (Fig. 2d). Multiple experiment-specific correlations with conserved modules were identified, with each conserved module being correlated with one or two experiment-specific modules in each experiment. Interestingly, a limited number of experiment-specific modules within each experiment were responsible for correlations, ranging from five experiment-specific modules in the gravitropism experiment to only two experiment-specific modules displaying significant correlations with conserved modules. These experiment-specific modules could represent putative mechanisms that integrate specific environmental or experimental cues to modify multiple core developmental processes represented by the conserved modules.

Including more treatments and experimental conditions improves the resolution of coexpression networks

We next examined the effect of increasing number of experimental conditions on the identification and characteristics of consensus networks. The results from such analyses provide practical insights into the utility of performing and integrating additional experiments, and can also provide insights into the biological interpretation of conserved modules. Coexpression analyses and consensus clustering were performed with all possible combinations of the four RNA-seq datasets and showed that increasing the number of datasets in the coexpression analysis increased the *resolution*, that is, the module sizes and specificity for function, of the coexpression networks (Fig. 3). Analysis of individual datasets yielded coexpression networks containing large numbers of genes and 90% (mean = 32 904 genes) of the expressed genes in the

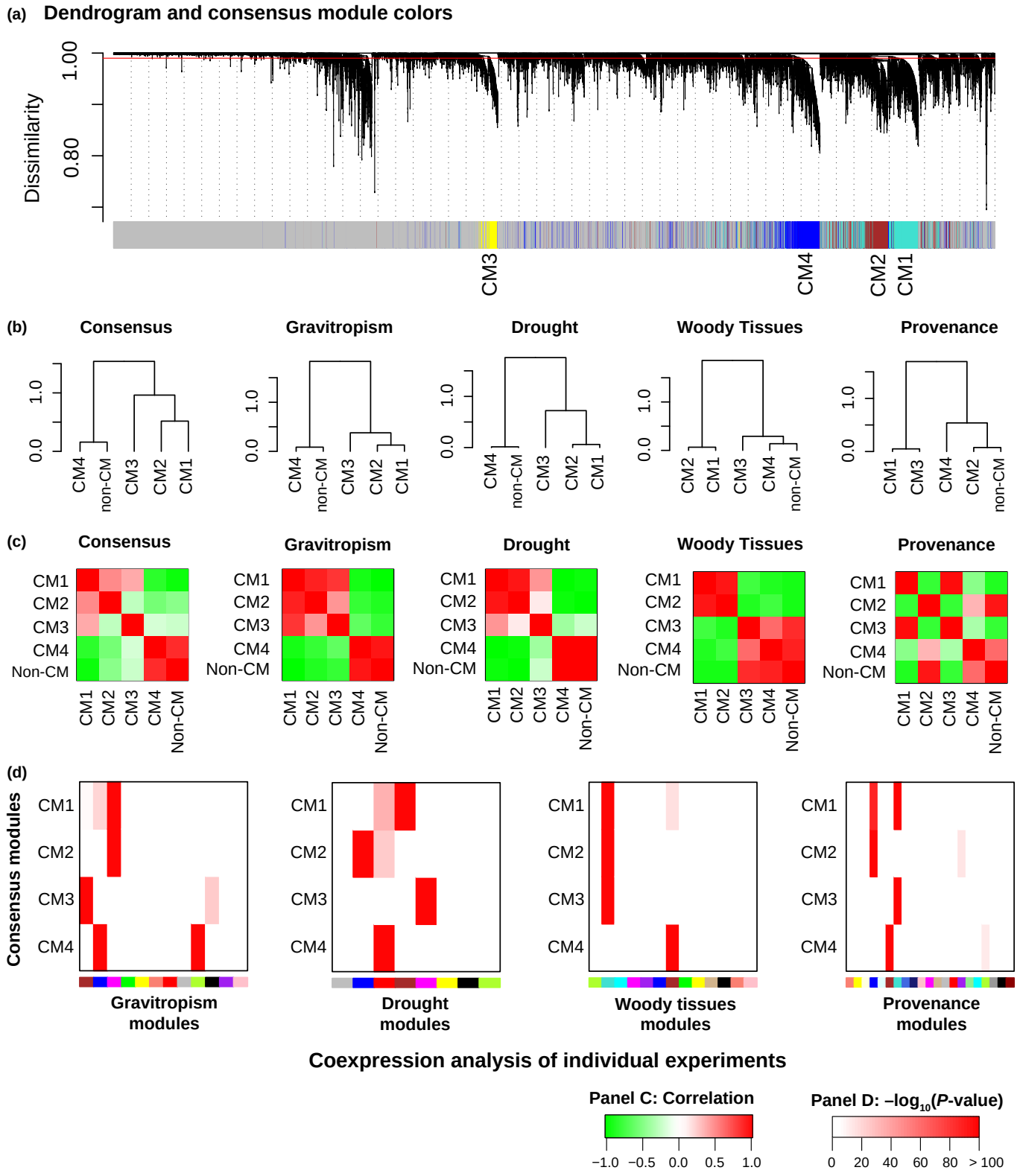

Fig. 2 Consensus network analysis across four RNA-seq datasets in *Populus*. (a) Hierarchical clustering dendrogram of the average network adjacency for the identification of conserved coexpression modules. Red line represents the cut height used in module identification. (b) Dendrogram of relationships among conserved modules for individual RNA-seq datasets. (c) Preservation of conserved modules in individual datasets quantified and presented as a heat map. Color scale is located at bottom of the figure. (d) Quantification of overlap between individual experiment-specific coexpression modules and conserved modules. Color scale for $-\log_{10}$(*P*-value) is located at the bottom of the figure. CM, conserved module.

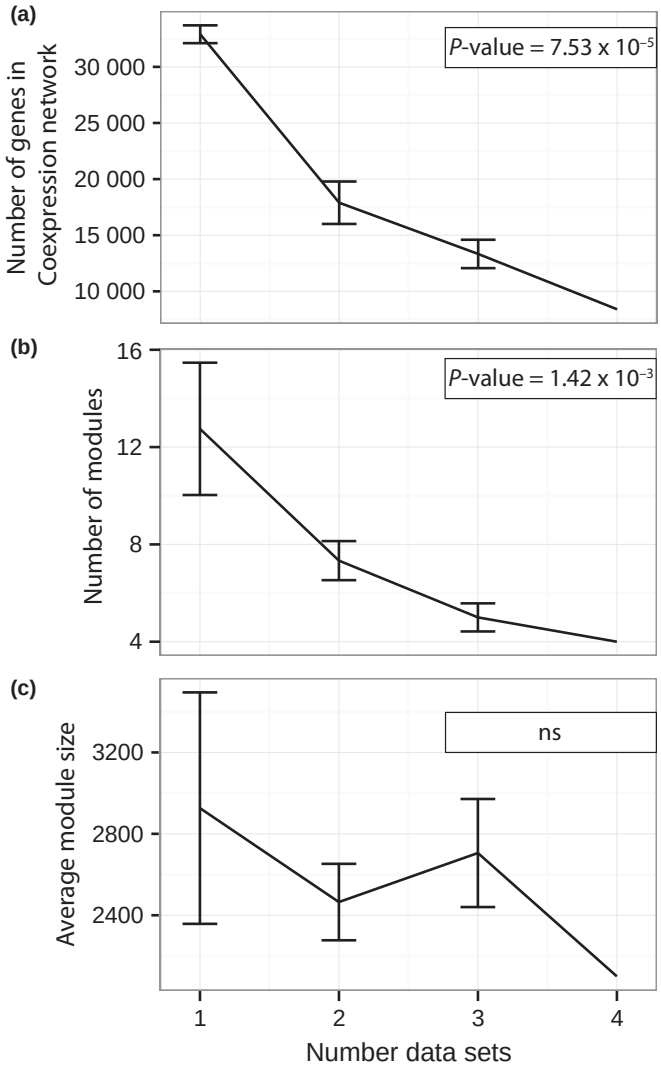

Fig. 3 Effects of increasing amounts of experimental data on the resolution of coexpression networks. Analysis of the effect of increasing data (a) on the number of genes included within coexpression networks, (b) the number of coexpression modules and (c) the average size of coexpression modules. Error bars indicate variation associated with permutations of datasets for inclusion in analyses. Permutations could not be performed in the cases including all four datasets, precluding estimation of variation. Description of analyses is found in the Materials and Methods section. ns, not statically significant (*P*-value > 0.05)

Populus genome that could be assigned to a coexpressed module (Fig. 3a). Such networks contain many genes and have large numbers of modules, and thus have little power to discriminate among genes specifically regulating wood development from other genes that happen to display coexpression under a limited number of conditions. By contrast, integration of data from increasing numbers of experiments using consensus clustering was progressively more stringent, as module membership requires correlated expression across larger numbers of conditions. Integration of data from all expression experiments decreased the number of genes coexpressed across all conditions by 75%. Increasing stringency also resulted in the identification of decreasing numbers of modules as additional experiments were added to the analysis (Fig. 3b). This trend supports the hypothesis that signals from diverse conditions converge on a limited number of core regulatory processes or pathways. Furthermore, increasing the number of datasets did not significantly influence the average size of coexpression modules, suggesting that the retained modules are robust and more function-specific than those identified from the individual datasets (Fig. 3c).

Conserved modules are distinguished by correlations with biological processes

Functional features of the conserved coexpression modules indicate that each module is associated with distinct developmental mechanisms. The behavior of gene expression for each module was modeled using the eigengene value of each module. Formally, the eigengene value is defined as the first principal component of the module expression matrix, and conceptually summarizes expression in terms of the most representative gene within the module. To understand how expression of genes in conserved modules responded to experimental treatments, we calculated the correlation between each module and each experiment including wood traits, experimental treatments and genotypes. As shown in Fig. 4, eigengene expression of each conserved module was unique, with modules showing significant correlations (Fig. S3) with variables from individual experiments in both hybrid aspen and *P. trichocarpa* (Fig. 4). In addition, conserved modules were significantly enriched for differentially expressed genes. Using a hypergeometric test, we show that a subset of differentially expressed genes from the gravitropism experiment (Gerttula *et al.*, 2015) were over-represented in conserved modules (Table S2).

Eigengene values revealed consistency of module behavior across experiments in three of the four conserved modules. For example, woody tissues of *P. trichocarpa* displayed similar patterns of expression as woody tissues sampled in the hybrid aspen drought experiment. Specifically, in both *P. trichocarpa* and hybrid aspen the CM1 and CM2 modules had high expression in the xylem and low phloem/bark expression (Fig. 4). Conversely, CM4 displayed low xylem and high phloem/bark expression in both experiments. Eigengene values also revealed dynamic regulation of modules within individual experiments. For example, in the hybrid aspen gravitropism experiment, the CM1, CM2 and CM3 modules were strongly associated with increased expression in response to treatment with GA, and showed decreased expression in opposite wood (OW). By contrast, CM4 was associated with decreased expression from the GA treatment and increased expression in OW (Fig. 4). These results suggest that GA has global effects on wood formation, as has been suggested previously based on experimental manipulation of GA and measurement of endogenous GA concentrations across woody tissues (Israelsson *et al.*, 2005; Mauriat & Moritz, 2009). They also suggest that there are major differences between OW and normal wood, as suggested previously (Gerttula *et al.*, 2015).

Modules also showed correlations with natural variation in *P. trichocarpa* provenances. After controlling for sampling year, the eigengene expression of the CM1, CM2 and CM3 modules

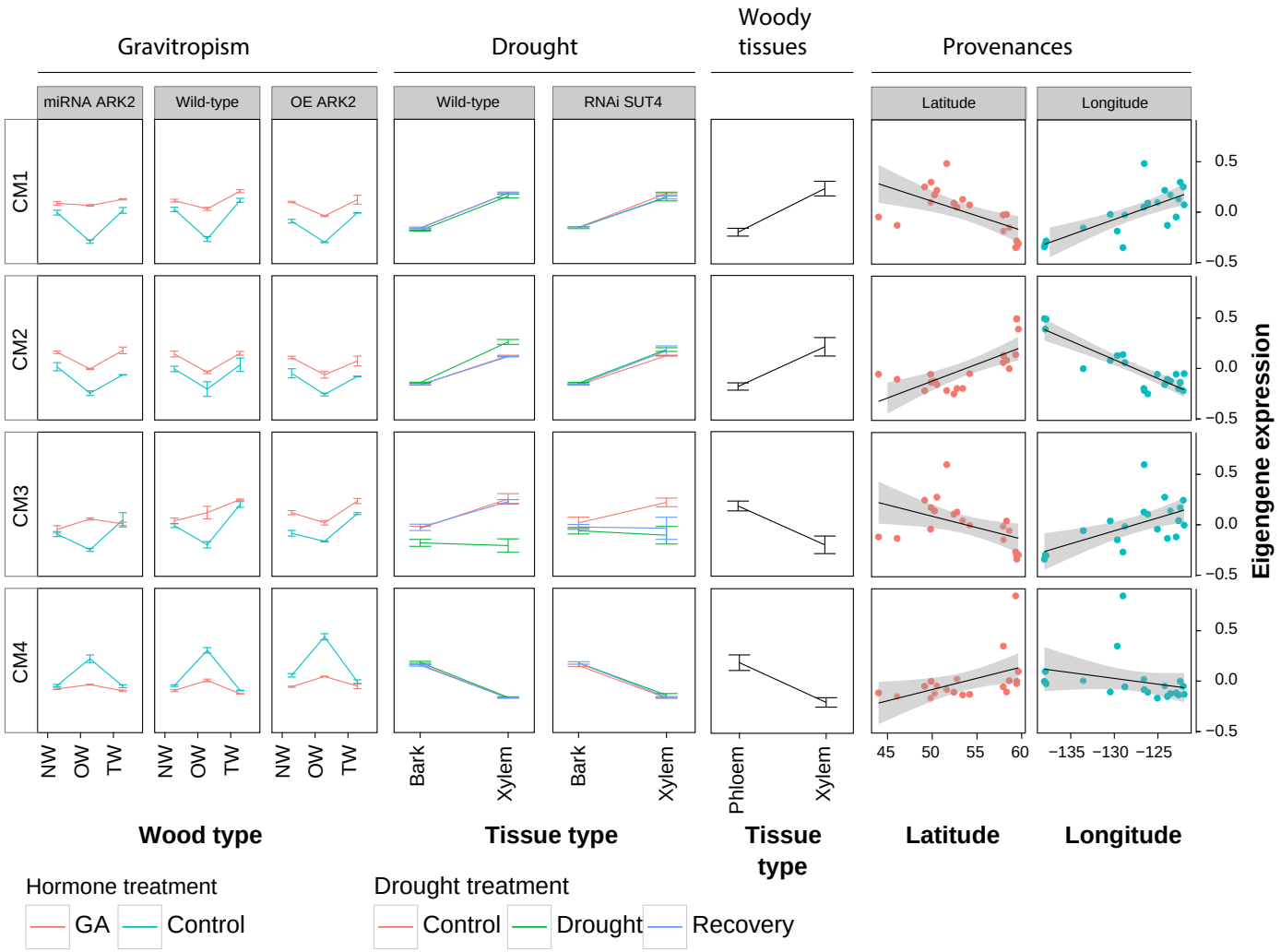

Fig. 4 Eigengene expression for each conserved module within experimental treatments and variables. Plots depict eigengene expression of conserved modules across samples and treatments of each experiment. CM, conserved module; NW, normal wood; OW, opposite wood; TW, tension wood; miRNA ARK2, hybrid aspen genotype expressing a miRNA targeting *ARBORKNOX2* transcripts; OE ARK2, hybrid aspen overexpressing *ARBORKNOX2*; RNAi SUT4, hybrid aspen expressing an interfering RNA targeting Sucrose Transporter 4. Descriptions of individual experiments and variables are shown in Table 1 and the Materials and Methods section. The horizontal bars represent ± 1 SE and gray shading represents 95% confidence intervals.

were significantly (P-value = 7.53×10^{-7}; 1.74×10^{-4}; 4.37×10^{-3}, respectively) associated with changes in longitude of *P. trichocarpa* provenances (Fig. 4). When grown in a common garden, expression of CM2 was highest in provenances originating near the Pacific coast, and interior provenances displayed lower expression of these same genes. Conversely, the CM1 and CM3 modules displayed high expression in interior provenances and lower expression in the coastal provenances.

Functional annotation of conserved modules using GO enrichment analysis showed that modules were enriched for hundreds of GO terms in the molecular function, biological process and cellular component pathways (Table S3). We focused on GO terms from biological processes involved in five categories fundamental to wood formation based on term annotations: hormone (including auxin, GA, brassinosteroid, cytokinin), cell-wall (including cellulose, xylan, xylose and lignin biosynthesis), meristem (including shoot development, xylem/phloem patterning), protein localization and epigenetic modifications (including histone, methylation and chromatin processes). Based on these categories, each conserved module was enriched in genes representing distinct biological pathways (Fig. 5; Table S4). The CM2 module was highly enriched with genes associated with cell-wall biogenesis, and cellulose, lignin and xylan biosynthesis, and to a lesser extent meristem development and maintenance, and protein localization. The CM1 module was broadly associated with regulation of hormone levels, cellulose biosynthesis, meristem development and protein localization. The CM3 module was highly enriched for genes associated with epigenetic modifications and protein localization. The CM4 module was enriched with genes involved in the regulation and response to hormone levels, meristem development, xylem–phloem patterning and protein localization. Furthermore, the nonconserved genes do not represent discrete functional groups of GO terms and displayed nonspecific results similar to random gene sets of the same size (Fig. S4).

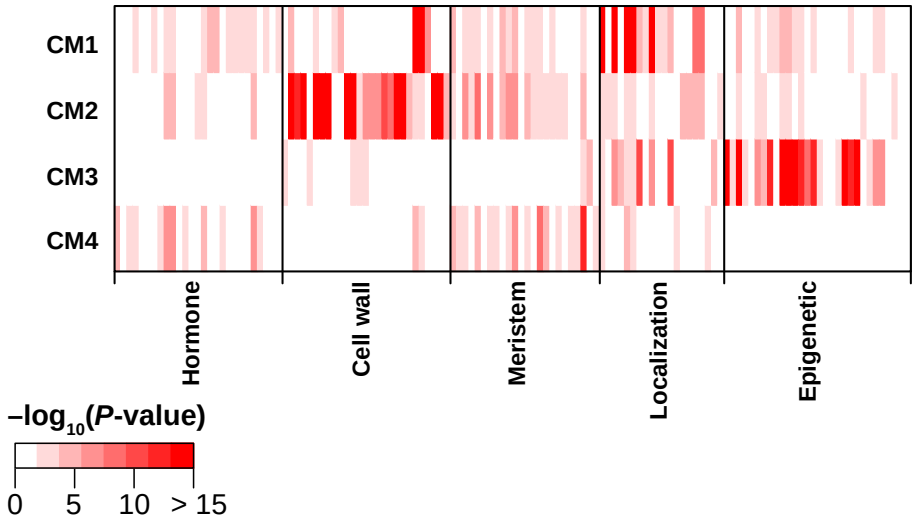

Fig. 5 Heat map summarizing gene ontology (GO) enrichment of conserved modules. Color intensity represents the statistical significance of difference in observed vs expected frequency of genes characterized by the selected GO categories. CM, conserved module.

Comparing GO enrichment and eigengene expression revealed additional functional features of each module. For example, CM1 and CM2 modules showed higher expression in xylem in both the hybrid aspen drought and *P. trichocarpa* tissue type experiments, and were also upregulated in tension wood in the gravitropism experiment (Fig. 4). These same modules also showed dramatic enrichment for multiple cell wall and meristem-related GO categories. In addition, the CM3 showed the strongest response to drought treatments, and also had enrichment for numerous GO categories associated with epigenetic modifications that have been previously implicated in drought response (Gourcilleau *et al.*, 2010; Liang *et al.*, 2014).

Integration of transcription factor binding and DNA accessibility data identify features of transcriptional regulation of coexpression gene modules

We next assayed genome-wide chromatin accessibility, to enable identification of cis-elements and trans-regulatory factors associated with regulation of gene expression within modules. Genome-wide footprints of DNase-seq representing regions of DNA accessible for protein binding were identified in samples of vascular tissue from *P. trichocarpa* using DNase-seq. The number of DNase-seq footprints per sample ranged from 300 123 to 371 692 footprints (Fig. S5a) and the mean footprint width per sample ranged from 134 to 156 bp (Fig. S5b). The global characteristics of DNase-seq footprints were similar among samples and the highest Jaccard's similarity occurred between samples of similar DNase I concentration (Fig. S6). The intersection of three highly similar DNase-seq samples (2.0, 3.0, 4.0 units DNase I) identified 125 415 reproducible footprints that were used in additional analyses. The most highly reproducible footprints occurred within close proximity to the TSS of gene features (Fig. S7a) and 82.6% of the footprints could be assigned to target genes (Fig. S7b). Approximately 36.1% of the footprints occurred inside gene features, 13.4% overlapped the TSS, 11.2% overlapped the TES, 11.8% occurred within 1 kb downstream of genes, and the remaining 10% either occurred within 1 kb upstream of genes or overlapped the entire gene feature.

Data from DNase-seq, and from ChIP-seq experiments describing the binding of individual transcription factors within the *Populus* genome (Liu *et al.*, 2014, 2015a,b) were next integrated with the coexpression modules. Enrichment of DNase footprints and transcription factor binding sites within modules showed that DNA accessibility and specific regulatory relationships may be important in defining conserved modules (Fig. 6). Reproducible footprints from the DNase-seq experiment were enriched in and around genes belonging to the four conserved modules (Fig. 6a). Additionally, three modules (CM1, CM2, CM3) were significantly enriched binding from ARK1, ARK2, BLR and PCN transcription factors (Fig. 6b). For example, ARK1 bound to 69.2% (P-value = 1.57×10^{-195}) of the genes in CM1, 55.4% (P-value = 3.96×10^{-37}) of the genes in CM2, and 73.2% (P-value = 1.65×10^{-49}) of the genes in CM3. Furthermore, the same three modules showed enriched binding (P-value < 0.05) from RNA-Pol II.

Conserved coexpression modules associate with paralogous genes and population-level adaptive traits

In order to understand how genetic variation and evolutionary processes influence the coexpression relationships involved in wood development, we assessed two levels of natural genetic variation. First, we found that duplicated genes arising from the Salicoid whole genome duplication (Tuskan *et al.*, 2006) were more likely to occur in conserved modules (P-value = 1.12×10^{-134}) than randomly selected genes. Further analysis of paralogs show that gene pairs displayed both conserved and divergent coexpression relationships (Fig. 7; $\chi^2 = 3049$; df = 16; P-value < 0.0001). The majority of paralogous genes showed that gene pairs were more likely to co-occur in the same coexpression module. However, a subset of paralogous genes displayed divergent coexpression relationships and these paralogs were assigned to either CM1 or CM2 modules (Fig. 7). Second, we assessed the

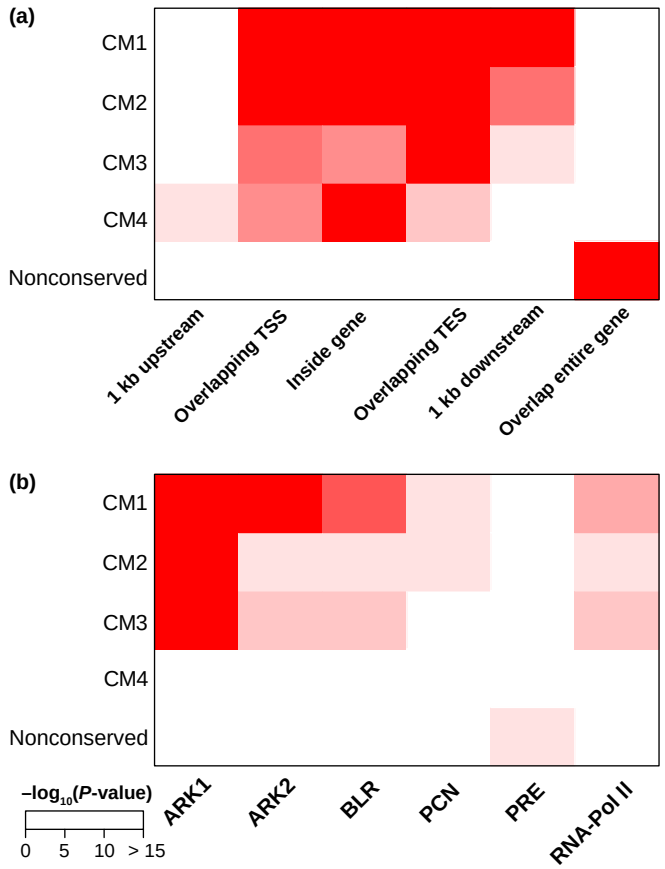

Fig. 6 Enrichment of DNase-seq footprints and binding of transcription factors within conserved modules. (a) Enrichment of reproducible DNase I hypersensitivity footprints in and around genes within each conserved module. (b) Enrichment of transcription factor and RNA-Pol II binding to genes found in conserved modules. CM, conserved module; TSS, transcriptional start site; TES, transcriptional end site.

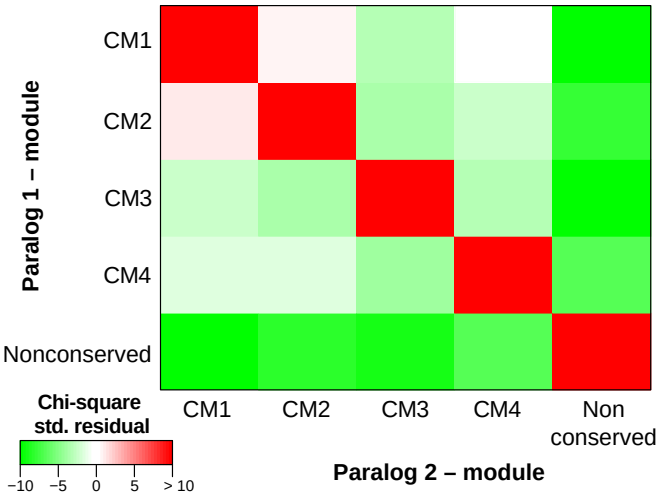

Fig. 7 Frequency of co-occurrence of paralogous genes within the same vs different gene coexpression modules. The observed vs expected frequency of co-occurrence of paralogous genes arising from the Salicoid whole genome duplication event was calculated and summarized here in a heat map. Intensity of red color indicates that the observed frequency of module assignments for paralogous gene pairs is greater than the expected frequency. Green indicates that the observed frequency is less than expected. Paralogous gene pairs primarily co-occur in the same conserved module (red diagonal), with the exception of gene pairs that occurred in the CM1 and CM2 modules. All other module combinations show fewer paralogs than would be expected by chance. CM, conserved module.

enrichment of conserved modules with GWAS for genetic variation associated with wood chemistry (Porth *et al.*, 2013) and biomass-related (McKown *et al.*, 2014) traits from a population genetic survey. Overall, 17 of the 36 traits were enriched in at least one of the four conserved modules (Fig. 8). The SNPs associated with three wood chemistry and two biomass traits were enriched in CM1, four wood chemistry and five biomass traits were enriched in CM2, and the CM4 module was enriched with one wood chemistry and three biomass traits. In addition, integration of GWAS shows that genetic variation in conserved module genes can influence wood formation processes such as the production of lignin, microfibril angle, cellulose crystallinity and xylose (Fig. 8). Together, these results indicate that the conserved modules are biologically significant and explain variation in gene expression at various scales ranging from individual experiments, population-level variation and across species, such as *P. trichocarpa* and hybrid aspen.

Discussion

A primary goal of the study reported here was to identify refined modules of coexpressed genes in different genotypes and under various environmental conditions that influence wood formation. We successfully identified and characterized such coexpressed gene modules, which we refer to as conserved modules, whose coexpression relationships were significant across the diverse experiments sampled here. We hypothesize that these conserved modules represent core biological mechanisms that are universally involved in wood development, which are modified to affect the various developmental outcomes associated with environmental, experimental or genetic perturbations.

We identified conserved gene modules using data from four diverse *Populus* experiments, and the coexpression relationships within modules were highly preserved in each experiment. These properties are consistent with core mechanisms (e.g. meristem function or cell wall biosynthesis) that would be common to wood formation but that are modified in response to environmental or experimental perturbations. The interactions between conserved modules changed across experiments (Fig. 2b,c) and suggest that experiment-specific perturbations such as gravistimulation, drought, tissue types and genetic variation across provenances may converge on core mechanisms to produce context-specific wood phenotypes.

Mapping of conserved modules back onto coexpression networks from individual experiments (Fig. 2d) support previous cross-species analyses (Street *et al.*, 2008), and show that individual networks are a combination of gene interactions that arise from experiment-specific perturbations and interactions that are conserved across all experiments. For example, conserved modules map to five of the 11 modules identified in the coexpression

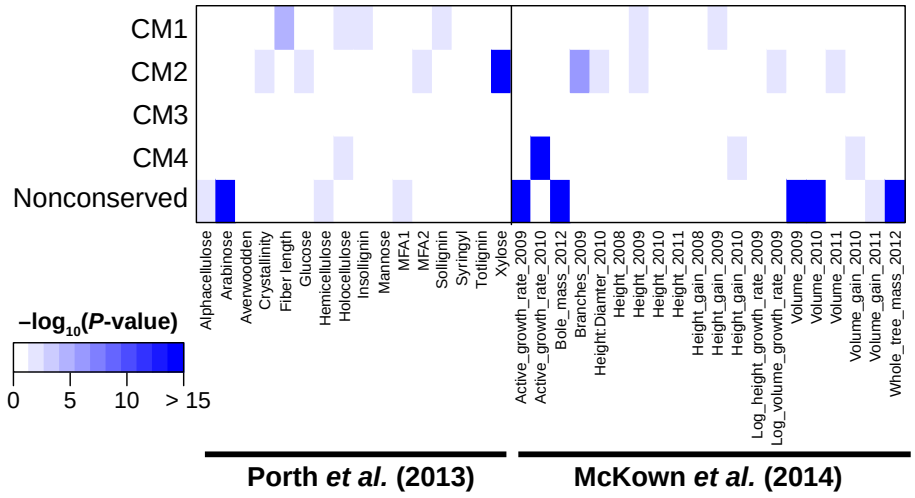

Fig. 8 Conserved module correlations with population association mapping traits in *Populus trichocarpa*. Heat map showing the enrichment of conserved modules for single nucleotide polymorphisms (SNPs) associated with wood chemistry (Porth *et al.*, 2013) and biomass-related traits (McKown *et al.*, 2014) from *P. trichocarpa*. $-\log_{10}$(*P*-value) scale is shown for quantification. MFA, microfibril angle; CM, conserved module.

analysis of the gravitropism study, and four of the seven modules in the drought experiment. Such modules are excellent candidates for defining the mechanisms that respond to the experiment-specific variables (e.g. gravitropism treatment), and modify or interact with the conserved module genes to alter development. In addition, integration of experiments across larger taxonomic scales will aid in the understanding of the ancestral pathways that have led to the diversity of wood formation in angiosperms (Spicer & Groover, 2010).

The coexpression approach here also facilitated the integration of diverse genomic data types from a variety of different experiments. Using computational analyses based on the consensus framework of coexpressed genes, we integrated data types including transcriptome profiling (RNA-seq), protein binding (ChIP-seq), DNA accessibility, and phenotypic data from experiments ranging from genome-wide association studies (GWAS) at population levels to characterizations of individual transcription factors. In general, orthogonal datasets yielded similar results as the consensus coexpression analysis, and led to additional biological insights through correlations of individual conserved modules that have specific responses to experimental treatments and environmental stresses.

Dissection of the conserved modules led to four major findings. First, functional annotation using gene ontology (GO) enrichment analysis suggests that each of the conserved modules represents specific biological pathways involved in cell wall biogenesis, meristem function, epigenetic processes, protein localization or hormones. Second, genes from conserved modules were more accessible to DNase I degradation in differentiating xylem, which suggests that the chromatin structure of co-regulated genes involved in core wood formation pathways is more accessible than noncoexpressed genes. Third, conserved modules were enriched for binding from four key transcription factors (ARBORKNOX1, ARBORKNOX2, BELLRINGER, popCORONA) that play fundamental roles in wood development (Groover *et al.*, 2006; Du *et al.*, 2009; Du & Groover, 2010). Fourth, significant correlations were found between conserved modules, and specific wood types, stress treatments (Fig. 4) and genes implicated in wood biochemistry (Fig. 8).

The conserved modules included gene families previously implicated in wood development (Zhong *et al.*, 2010; Hussey *et al.*, 2013; Nakano *et al.*, 2015; Ye & Zhong, 2015). Our coexpression analyses place these genes and pathways into a larger context, and associate them with unknown genes participating in wood formation. For example, one conserved module, CM2, is highly enriched for genes associated with cell-wall related GO terms, and included first-layer master regulatory transcription factors (NST1, VND1), regulators of first-layer switches (ANAC075, GATA12, SND2, WRKY12), second-layer switches (MYB46, MYB83) and a suite of downstream transcription factors (C3H14, KNAT7, MYB4, MYB42, MYB52, MYB69, MYB103) involved in cell wall formation. In addition, CM2 contains structural genes involved in secondary cell wall biosynthesis, and the production of cellulose (CESA3, CESA4, CESA7, CESA8, COBL4), hemicellulose (GUX1, GUX2, GXM, IRX8, IRX9, IRX10, IRX14-L, PARVUS) and lignin (C4H, CAD5, CCoAOMT1, CCR1, COMT2, F5H1, HCT, LAC4, LAC12, LAC17, PAL1).

Our results show that increasing the number of RNA-seq experiments and perturbations increases the resolution of coexpression networks by identifying smaller numbers of coexpressed genes and fewer modules of genes that underlie wood development in increasingly diverse conditions. We empirically addressed practical issues surrounding the use of coexpression- and computational-based approaches to more precisely narrow the number of genes associated with wood phenotypes. Extrapolating from Fig. 3(a), we estimate that, to identify consensus coexpression networks that contain at most hundreds of genes, approximately eight or more total datasets describing gene expression during wood development in contrastingly diverse conditions as those described here would be required. Adding novel experiments would be the most informative datasets because previously uninvestigated factors would perturb network connections in new ways and refine coexpression modules. Additionally, including diverse tissue types (e.g. tissues other than

wood) is essential for identifying tissue-specific patterns of expression (Quesada *et al.*, 2008).

Much of the previous work on the regulation of wood development has focused on a limited number of genes or specific regulatory interactions in *Arabidopsis* and a handful of woody species (Demura & Fukuda, 2007; Ye & Zhong, 2015). These approaches have been beneficial in providing a starting point to understand wood development in tree species, but lack the power to comprehensively describe the interactions among complex pathways underlying wood formation, which involve thousands of genes. We found that integration of diverse genome-wide datasets directly in a tree species can be used to identify and describe modules of genes that have functional relevance to wood formation, such as overlaying wood chemistry single-nucleotide polymorphisms and consensus coexpression networks. Indeed, although still involving relatively large numbers of genes, the modules described in the experiments here are excellent test beds for further study using genome-scale functional genomic approaches (e.g. Henry *et al.*, 2015). Additionally, although the studies here are restricted to the genus *Populus*, the preservation of coexpression relationships among these modules across diverse conditions and genotypes make them excellent candidates for providing a first glimpse of the ancestral genes required for wood formation in angiosperms. This hypothesis will require additional, comparative studies in additional woody species, but could ultimately describe the ancestral mechanisms that evolved to regulate wood formation, as well as the species- and lineage-specific genes and mechanisms responsible for the amazing diversity in wood development displayed among angiosperms.

Acknowledgements

This work was supported by grants 2011-67013-30062 and 2015-67013-22891 USDA AFRI to A.G. and V.F. M.Z. is supported by NSF PGRP Fellowship grant IOS-1402064. This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303.

Author contributions

A.G., M.Z. and V.F. planned and designed the research project; L.L. and M.Z. performed experiments and analyzed data; and A.G., M.Z. and V.F. wrote the manuscript.

References

Amrine KCH, Blanco-Ulate B, Cantu D. 2015. Discovery of core biotic stress responsive genes in *Arabidopsis* by weighted gene co-expression network analysis. *PLoS ONE* **10**: e0118731.

Anders S, Pyl PT, Huber W. 2015. HTSeq – a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.

Andersson-Gunneras S, Mellerowicz EJ. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant Journal* **45**: 144–165.

Bao H, Mansfield SD, Cronk QCB, El-Kassaby YA, Douglas CJ. 2013. The developing xylem transcriptome and genome-wide analysis of alternative splicing in *Populus trichocarpa* (black cottonwood) populations. *BMC Genomics* **14**: 359.

Bao Y, Dharmawardhana P, Mockler T, Strauss SH. 2009. Genome scale transcriptome analysis of shoot organogenesis in *Populus*. *BMC Plant Biology* **9**: 132.

Bassel GW, Gaudinier A, Brady SM, Hennig L, Rhee SY, De Smet I. 2012. Systems analysis of plant functional, transcriptional, physical interaction, and metabolic networks. *The Plant Cell* **24**: 3859–3875.

Battipaglia G, De Micco V, Sass-Klaassen U, Tognetti R, Mäkelä A. 2014. Special issue: WSE symposium: wood growth under environmental changes: the need for a multidisciplinary approach. *Tree Physiology* **34**: 787–791.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537–2538.

Carter SL, Brechbühler CM, Griffin M, Bond AT. 2004. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **20**: 2242–2250.

Demura T, Fukuda H. 2007. Transcriptional regulation in wood formation. *Trends in Plant Science* **12**: 64–70.

D'haeseleer P, Liang S, Somogyi R. 2000. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16**: 707–726.

Dharmawardhana P, Brunner AM, Strauss SH. 2010. Genome-wide transcriptome analysis of the transition from primary to secondary stem development in *Populus trichocarpa*. *BMC Genomics* **11**: 150.

Du J, Groover A. 2010. Transcriptional regulation of secondary growth and wood formation. *Journal of Integrative Plant Biology* **52**: 17–27.

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

Du J, Miura E, Robischon M, Martinez C, Groover A. 2011. The *Populus* Class III HD ZIP transcription factor *POPCORONA* affects cell differentiation during secondary growth of woody stems. *PLoS ONE* **6**: e17458.

Etchells JP, Mishra Laxmi S, Kumar M, Campbell L, Turner Simon R. 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.

Geraldes A, DiFazio SP, Slavov GT, Ranjan P, Muchero W, Hannemann J, Gunter LE, Wymore AM, Grassa CJ, Farzaneh N *et al.* 2013. A 34K SNP genotyping array for *Populus trichocarpa*: design, application to the study of natural populations and transferability to other *Populus* species. *Molecular Ecology Resources* **13**: 306–323.

Gerttula S, Zinkgraf M, Muday GK, Lewis DR, 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.

Gourcilleau D, Bogeat-Triboulot M-B, Thiec D, Lafon-Placette C, Delaunay A, El-Soud WA, Brignolas F, Maury S. 2010. DNA methylation and histone acetylation: genotypic variations in hybrid poplars, impact of water deficit and relationships with productivity. *Annals of Forest Science* **67**: 208.

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

Groover AT, Mansfield SD, DiFazio SP, Dupper G, Fontana JR, 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.

Guerriero G, Sergeant K, Hausman J-F. 2014. Wood biosynthesis and typologies: a molecular rhapsody. *Tree Physiology* **34**: 839–855.

Henry IM, Zinkgraf MS, Groover AT, Comai L. 2015. A system for dosage-based functional genomics in poplar. *The Plant Cell* **27**: 2370–2383.

Hussey SG, Mizrachi E, Creux NM, Myburg AA. 2013. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 325.

Israelsson M, Sundberg B, Moritz T. 2005. Tissue-specific localization of gibberellins and expression of gibberellin-biosynthetic and signaling genes in wood-forming tissues in aspen. *Plant Journal* **44**: 494–504.

Janz D, Lautner S, Wildhagen H, Behnke K, Schnitzler J-P, Rennenberg H, Fromm J, Polle A. 2012. Salt stress induces the formation of a novel type of 'pressure wood' in two *Populus* species. *New Phytologist* **194**: 129–141.

Jiang Y, Duan Y, Yin J, Ye S, Zhu J, Zhang F, Lu W, Fan D, Luo K. 2014. Genome-wide identification and characterization of the *Populus* WRKY transcription factor family and analysis of their expression in response to biotic and abiotic stresses. *Journal of Experimental Botany* **65**: 6629–6644.

Joshi NA, Fass JN. 2011. *Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. Version 1.33.* [WWW document] URL https://github.com/najoshi/sickle [accessed 22 April 2015].

Koohy H, Down TA, Spivakov M, Hubbard T. 2014. A comparison of peak callers used for DNase-seq data. *PLoS ONE* **9**: e96303.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ *et al.* 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Langfelder P, Horvath S. 2007. Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* **1**: 1–17.

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

Langfelder P, Luo R, Oldham MC, Horvath S. 2011. Is my network module preserved and reproducible? *PLoS Computational Biology* **7**: e1001057.

Langfelder P, Mischel PS, Horvath S. 2013. When is hub gene selection better than standard meta-analysis? *PLoS ONE* **8**: e61505.

Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**: 719–720.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**: 1–10.

Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. 2010. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nature Biotechnology* **28**: 149–156.

Li Q, Brown JB, Huang H, Bickel PJ. 2011. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5**: 1752–1779.

Liang D, Zhang Z, Wu H, Huang C, Shuai P, Ye C-Y, Tang S, Wang Y, Yang L, Wang J *et al.* 2014. Single-base-resolution methylomes of *Populus trichocarpa* reveal the association between DNA methylation and drought stress. *BMC Genetics* **15**(Suppl 1): S9.

Liu L, Missirian V, Zinkgraf M, Groover A, Filkov V. 2014. Evaluation of experimental design and computational parameter choices affecting analyses of ChIP-seq and RNA-seq data in undomesticated poplar trees. *BMC Genomics* **15**: S3.

Liu L, Ramsay T, Zinkgraf M, Sundell D, Street NR, Filkov V, Groover A. 2015b. A resource for characterizing genome-wide binding and putative target genes of transcription factors expressed during secondary growth and wood formation in *Populus*. *Plant Journal* **82**: 887–898.

Liu L, Zinkgraf M, Petzold HE, Beers EP, Filkov V, Groover A. 2015a. The *Populus ARBORKNOX1* homeodomain transcription factor regulates woody growth through binding to evolutionarily conserved target genes of diverse function. *New Phytologist* **205**: 682–694.

Mauriat M, Moritz T. 2009. Analyses of *GA20ox*- and *GID1*-over-expressing aspen suggest that gibberellins play two distinct roles in wood formation. *Plant Journal* **58**: 989–1003.

McKown AD, Klapste J, Guy RD, Geraldes A, Porth I, Hannemann J, Friedmann M, Muchero W, Tuskan GA, Ehlting J *et al.* 2014. Genome-wide association implicates numerous genes underlying ecological trait variation in natural populations of *Populus trichocarpa*. *New Phytologist* **203**: 535–553.

Mellerowicz EJ, Gorshkova TA. 2012. Tensional stress generation in gelatinous fibres: a review and possible mechanism based on cell-wall structure and composition. *Journal of Experimental Botany* **63**: 551–565.

Nakano Y, Yamaguchi M, Endo H, Rejab NA, Ohtani M. 2015. NAC-MYB-based transcriptional regulation of secondary cell wall biosynthesis in land plants. *Frontiers in Plant Science* **6**: 288.

Porth I, Klapste J, Skyba O, Hannemann J, McKown AD, Guy RD, Difazio SP, Muchero W, Ranjan P, Tuskan Ga *et al.* 2013. Genome-wide association mapping for wood characteristics in *Populus* identifies an array of candidate single nucleotide polymorphisms. *New Phytologist* **200**: 710–726.

Quesada T, Li Z, Dervinis C, Li Y, Bocock PN, Tuskan GA, Casella G, Davis JM, Kirst M. 2008. Comparative analysis of the transcriptomes of *Populus trichocarpa* and *Arabidopsis thaliana* suggests extensive evolution of gene expression regulation in angiosperms. *New Phytologist* **180**: 408–420.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2015. *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. [WWW document] URL https://www.R-project.org/ [accessed 1 September 2015]

Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, Bones AM, Nielsen HB, Mundy J. 2013. Transcriptome responses to combinations of stresses in *Arabidopsis*. *Plant Physiology* **161**: 1783–1794.

Robinson MD, McCarthy DJ, Smyth GK. 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.

Rodgers-Melnick E, Mane SP, Dharmawardhana P, Slavov GT, Crasta OR, Strauss SH, Brunner AM, Difazio SP. 2012. Contrasting patterns of evolution following whole genome versus tandem duplication events in *Populus*. *Genome Research* **22**: 95–105.

Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR *et al.* 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**: 389–393.

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**: 10 096–10 101.

Schrader J, Nilsson J, Mellerowicz E, Berglund A, Nilsson P, Hertzberg M, Sandberg G. 2004. A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *The Plant Cell* **16**: 2278–2292.

Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W. 2016. Learning from co-expression networks: possibilities and challenges. *Frontiers in Plant Science* **7**: 1–18.

Shaik R, Ramakrishna W. 2013. Genes and co-expression modules common to drought and bacterial stress responses in *Arabidopsis* and rice. *PLoS ONE* **8**: e77261.

Spicer R, Groover A. 2010. Evolution of development of vascular cambia and secondary growth. *New Phytologist* **186**: 577–592.

Street NR, Sjödin A, Bylesjö M, Gustafsson P, Trygg J, Jansson S. 2008. A cross-species transcriptomics approach to identify genes involved in leaf development. *BMC Genomics* **9**: 589.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

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.

Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ. 2009. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment* **32**: 1633–1651.

Xue L-J, Frost CJ, Tsai C-J, Harding SA. 2016. Drought response transcriptomes are altered in poplar with reduced tonoplast sucrose transporter expression. *Scientific Reports* **6**: 33 655.

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

Yordanov YS, Regan S, Busov V. 2010. Members of the LATERAL ORGAN BOUNDARIES DOMAIN transcription factor family are involved in the regulation of secondary growth in *Populus*. *The Plant Cell* **22**: 3662–3677.

Zhong R, Lee C, Ye Z-H. 2010. Evolutionary conservation of the transcriptional network regulating secondary cell wall biosynthesis. *Trends in Plant Science* **15**: 625–632.

Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, Green MR. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11**: 237.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

Fig. S1 Determining the soft threshold for each individual RNA-seq dataset.

Fig. S2 Module membership of genes assigned to the four conserved modules and the 'nonconserved' group of unclustered genes.

Fig. S3 Significance of module–trait correlations between conserved module eigengenes and treatments for each of the four experiments.

Fig. S4 Comparison of gene ontology enrichment between nonconserved genes and 10 random gene sets of equal size.

Fig. S5 Summary of DNase footprints for increasing concentrations of DNase I enzyme.

Fig. S6 Overlap between DNase-seq footprints found between each of the DNase I samples.

Fig. S7 Genome-wide distribution of the 125 415 reproducible DNase-seq footprints.

Table S1 Module assignments from individual and consensus coexpression analyses, and functional annotations for *Populus* gene models that were expressed across all datasets

Table S2 Conserved modules were enriched with differentially expressed genes

Table S3 Results from gene ontology (GO) enrichment analysis for each conserved module

Table S4 Complete list of results used to generate the gene ontology (GO) enrichment summary of conserved modules (Fig. 5)

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.