the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **82,** 887–898 doi: 10.1111/tpj.12850

RESOURCE

A resource for characterizing genome-wide binding and putative target genes of transcription factors expressed during secondary growth and wood formation in *Populus*

Lijun Liu[1], Trevor Ramsay[2], Matthew Zinkgraf[1], David Sundell[3], Nathaniel Robert Street[3], Vladimir Filkov[2] and Andrew Groover[1,4,*]

[1]*USDA Forest Service, Pacific Southwest Research Station, Davis, CA, 95618, USA,*

[2]*Department of Computer Science, University of California Davis, Davis, CA, 95618, USA,*

[3]*Umeå Plant Science Centre, Department of Plant Physiology, University of Umeå, SE-901-87 Umeå, Sweden, and*

[4]*Department of Plant Biology, University of California Davis, Davis, CA, 95618, USA*

Received 20 December 2014; revised 1 April 2015; accepted 2 April 2015; published online 22 April 2015.

*For correspondence (email agroover@fs.fed.us).

Accession numbers: NCBI SRA accession numbers SRP053368 and SRP042635.

SUMMARY

Identifying transcription factor target genes is essential for modeling the transcriptional networks underlying developmental processes. Here we report a chromatin immunoprecipitation sequencing (ChIP-seq) resource consisting of genome-wide binding regions and associated putative target genes for four *Populus* homeodomain transcription factors expressed during secondary growth and wood formation. Software code (programs and scripts) for processing the *Populus* ChIP-seq data are provided within a publically available iPlant image, including tools for ChIP-seq data quality control and evaluation adapted from the human Encyclopedia of DNA Elements (ENCODE) project. Basic information for each transcription factor (including members of Class I KNOX, Class III HD ZIP, BEL1-like families) binding are summarized, including the number and location of binding regions, distribution of binding regions relative to gene features, associated putative target genes, and enriched functional categories of putative target genes. These ChIP-seq data have been integrated within the Populus Genome Integrative Explorer (PopGenIE) where they can be analyzed using a variety of web-based tools. We present an example analysis that shows preferential binding of transcription factor ARBORKNOX1 to the nearest neighbor genes in a pre-calculated co-expression network module, and enrichment for meristem-related genes within this module including multiple orthologs of Arabidopsis KNOTTED-like Arabidopsis 2/6.

Keywords: *Populus trichocarpa*, chromatin immunoprecipitation sequencing, transcription factor, cambium, secondary growth, wood formation.

INTRODUCTION

To fully understand the transcriptional regulation of a developmental process, it is necessary to determine the binding of individual transcription factors to their target genes. Transcription factor binding data can be used in modeling of transcriptional regulatory networks, which provide precise specifications of complex interdependencies underpinning these biological systems (Ideker *et al.*, 2001; Long *et al.*, 2008; Van de Poel *et al.*, 2014).

The putative target genes for a given transcription factor can be estimated using a variety of techniques. For example, high-throughput yeast one-hybrid assays have been used to identify transcription factor putative target genes in Arabidopsis *in vitro* (Brady *et al.*, 2011; Gaudinier *et al.*, 2011; Taylor-Teeples *et al.*, 2015). *In vitro* techniques typically have the limitation of not measuring the effects of chromatin states or involvement of interacting proteins that may be important for understanding regulation *in vivo*. Chromatin immunoprecipitation-coupled high-throughput sequencing (ChIP-seq) can identify putative target genes of transcription factors genome-wide during

normal development (Barski *et al.*, 2007; Johnson *et al.*, 2007; Robertson *et al.*, 2007). However, ChIP-seq also presents several technical challenges in the production, evaluation, and interpretation of high quality datasets (Landt *et al.*, 2012; Liu *et al.*, 2014). The consortium of the ENCODE project has developed rigorous standards for producing and evaluating ChIP-seq datasets (ENDODE Project Consortium 2012), including evaluation of antibodies used for ChIP-seq, evaluation of ChIP-seq datasets using cross-correlation analysis and IDR (irreproducible discovery rate) analyses (see Results). These standards have only recently been extended to plants (Liu *et al.* 2015).

Transcriptional regulation acts as a primary mechanism regulating the radial, secondary growth of woody stems in the model tree genus, *Populus.* For example, microarray analyses show good correlation of transcript levels of genes with functions corresponding to specific stages of cambial divisions and cell differentiation in wood development (Hertzberg *et al.*, 2001; Schrader *et al.*, 2004). Several transcription factors have been functionally characterized as critical regulators of *Populus* secondary growth and wood formation. For example, Class I KNOX homeodomain transcription factors ARBORKNOX1 (ARK1) and ARK2 are expressed broadly in the cambial zone and influence cell differentiation (Groover *et al.*, 2006; Du *et al.*, 2009), Class III HD ZIP popREVOLUTA (PRE) regulates vascular cambium initiation and patterning of secondary vascular tissues (Robischon *et al.*, 2011), while Class III HD ZIP popCORONA (PCN) primarily affects cell differentiation (Du *et al.*, 2011).

We recently applied ChIP-seq to identify genome-wide binding characteristics of ARK1 in *Populus* cambium and recent derivatives. Similar to other transcription factors in both plants and animals, ARK1 has thousands of binding loci in the *Populus* genome (Liu *et al.* 2015). The binding loci are found highly enriched around the transcriptional start sites of genes, and are conserved among paralogs resulting from a genome duplication event in the lineage encompassing *Populus* and *Salix.* However, the analysis of a single transcription factor by ChIP-seq is limited. As has been shown in the ENCODE project (ENDODE Project Consortium 2012), models with good predictive power of gene expression must take into account such fundamental features of transcriptional regulation as combinatorial binding, which requires binding data from multiple transcription factors.

In this report, we present a resource for extending studies of transcription factor binding during secondary growth and wood development in *Populus.* We describe a publically available virtual machine image in iPlant (Goff *et al.*, 2011) containing tools for processing and analyzing existing or new *Populus* ChIP-seq datasets. We also present primary analyses of new ChIP-seq datasets for *Populus* ARK2, PCN, PRE, and a BELL-like homeodomain family member popBELLRINGER (BLR) that is expected to heterodimerize with KNOX proteins (Byrne *et al.*, 2003; Smith and Hake, 2003). Results include genome-wide distributions of each transcription factor's binding regions relative to gene features, functional enrichment of putative target genes, and pair-wise comparison of transcription factor binding regions and putative target genes. All the *Populus* ChIP-seq data have been integrated into the Populus Genome Integrative Explorer (PopGenIE) (http://popgenie.org/) (Sjodin *et al.*, 2009), and we present examples of analyses enabled by the integrated analysis and visualization tools.

RESULTS

ChIP-seq workflow and virtual machine image for data analysis

The major steps of our established ChIP-seq pipeline are outlined in Figure 1. First, transcription factor-specific peptides are selected based on antigenicity and specificity, and used to generate antibodies against each native transcription factor. Antibody specificity is experimentally tested with western blotting of the antibody against recom-

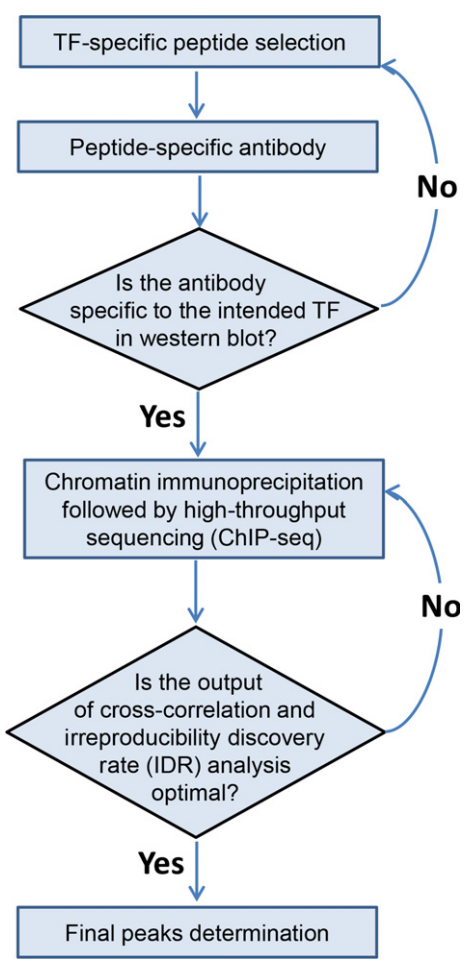


Figure 1. Overview of the ChIP-seq pipeline.

binant proteins. Chromatin immunoprecipitation (ChIP) is then performed using biological replicates of vascular cambium and its recent derivatives with each antibody to precipitate DNA fragments bound by the target transcription factor. The ChIP'd DNA fragments are then used for Illumina sequencing library preparation and subjected to high-throughput sequencing (see Experimental procedures).

Processing of the data and evaluation of results are achieved using a series of custom scripts and established programs, which have been made available as a virtual machine (VM) through iPlant's Atmosphere (http://www.iplantcollaborative.org/ci/atmosphere) as Populus_ChIPseq_VM_1.0, and also available as a VM image at http://web.cs.ucdavis.edu/~filkov/software/iplant/. Briefly, with the VM, raw reads can be trimmed and filtered, and used to produce sets of high quality mapped reads (see Experimental procedures). ChIP-seq replicates are then evaluated with cross-correlation analysis. This analysis uses the observation that bona-fide ChIP-seq peaks accumulate sequence reads staggering the peak center on the forward and reverse strands, and calculates a related RSC (relative strand correlation) statistic that can be used as an estimate of peak quality (Li *et al.*, 2011). The consistency between ChIP-seq replicates is then evaluated using IDR analysis, which assumes binding signals are highly reproducible (Li *et al.*, 2011). Finally, replicates are pooled for IDR analysis to identify the most reproducible peaks and associated high confidence binding regions.

Identification of genome-wide binding regions for *Populus* transcription factors

Western blots were probed with antibodies raised against *Populus* transcription factors ARBORKNOX2 (ARK2), popCORONA (PCN), popREVOLUTA (PRE) and BELLRINGER (BLR) as shown in Figure S1. In each case, the peptide-specific antibodies exclusively recognized the target but not control transcription factors (Figure S1), suggesting high antibody specificity resulting from genome-enabled antibody design. Each antibody was used to prepare two biological replicates of ChIP-seq libraries from *Populus* cambium and recent derivatives, and submitted for Illumina sequencing (see Experimental procedures). A total of 32–81 million mapped reads for peak calling were obtained for each replicate (Table S1).

Cross-correlation analysis showed the RSC value of each replicate was higher than 0.65 for most samples and thus within ENCODE acceptable standards, except for BLR replicate1 (BLR_r1), and PCN replicate1 and replicate2 (PCN_r1, and PCN_r2) (Figure S2a–d), which was possibly due to inadequate numbers of mapped reads for those three replicates (Table S1). However, the pooled ChIP-seq samples for each transcription factor that were used for final peak detection all displayed acceptable RSC values (Figure S2e–h).

Next, IDR analysis was performed with each transcription factor's ChIP-seq replicates. As shown in Figure 2, the number of significant peaks with high reproducibility identified between the two original replicates (Figure 2a, c, e, g) or pseudoreplicates of the pooled sample (Figure 2b, d, f, h) varied across different transcription factors. In total, 2287, 5674, 3148, and 658 significant peaks (Tables 1 and S2) for ARK2, BLR, PCN, and PRE, respectively, were detected using a conservative threshold (see Experimental procedures). For simplicity, these peaks are referred to as transcription factor binding regions in the following studies.

Transcription factors vary in their binding regions relative to putative target genes

We next assigned the closest gene to each binding region as its putative target gene. Most transcription factors were associated with a single binding region relative to a putative target gene. However a small number of binding regions were localized to small scaffolds which have not been placed within the 19 chromosomes in the current *Populus* genome and do not contain any annotated genes (Table S2). Other putative target genes were targeted by multiple binding regions (Table S2). Together, 2277, 4925, 2857, and 551 unique putative target genes were identified for ARK2, BLR, PCN, and PRE, respectively (Table 1).

Comparison of the distribution of ChIP-seq peaks relative to the transcription start site (TSS) of the putative target gene revealed striking differences in binding profiles among different transcription factors (Figure 3). While around 90% of ARK1 (Liu *et al.* 2015) and ARK2 binding regions were within 1 kb of the TSS, only around 30% of BLR, PCN, and PRE binding regions fell within this range. The majority of binding regions were within 5 kb of the TSS for all transcription factors, however, showing a general preference for binding near genes (Table 2). Notably, the distribution of ARK2 binding regions relative to gene features (Figure 3a) was very close to that of ARK1 (Liu *et al.* 2015), which is consistent with their close phylogenetic relationship and could reflect some level of genetic redundancy similar to their Arabidopsis orthologs, SHOOTMERISTEMLESS and BREVIPEDICELLUS (Byrne *et al.*, 2002). Compared with ARK1 and ARK2, more binding regions of BLR, PCN, and PRE resided in the upstream or downstream regions or overlapped with the 3′-end of the putative target genes (Figure 3b–d). Moreover, BLR uniquely showed a bimodal distribution of binding regions relative to the TSS of genes, with a sharp peak of binding immediately upstream of the TSS, and a smaller peak of binding downstream of genes. Overall, the results showed that binding regions for these transcription factors were generally enriched around the TSS of genes in the *Populus* genome; however, different transcription factors also had distinct binding features, such as proximal versus distal binding to the putative target genes.

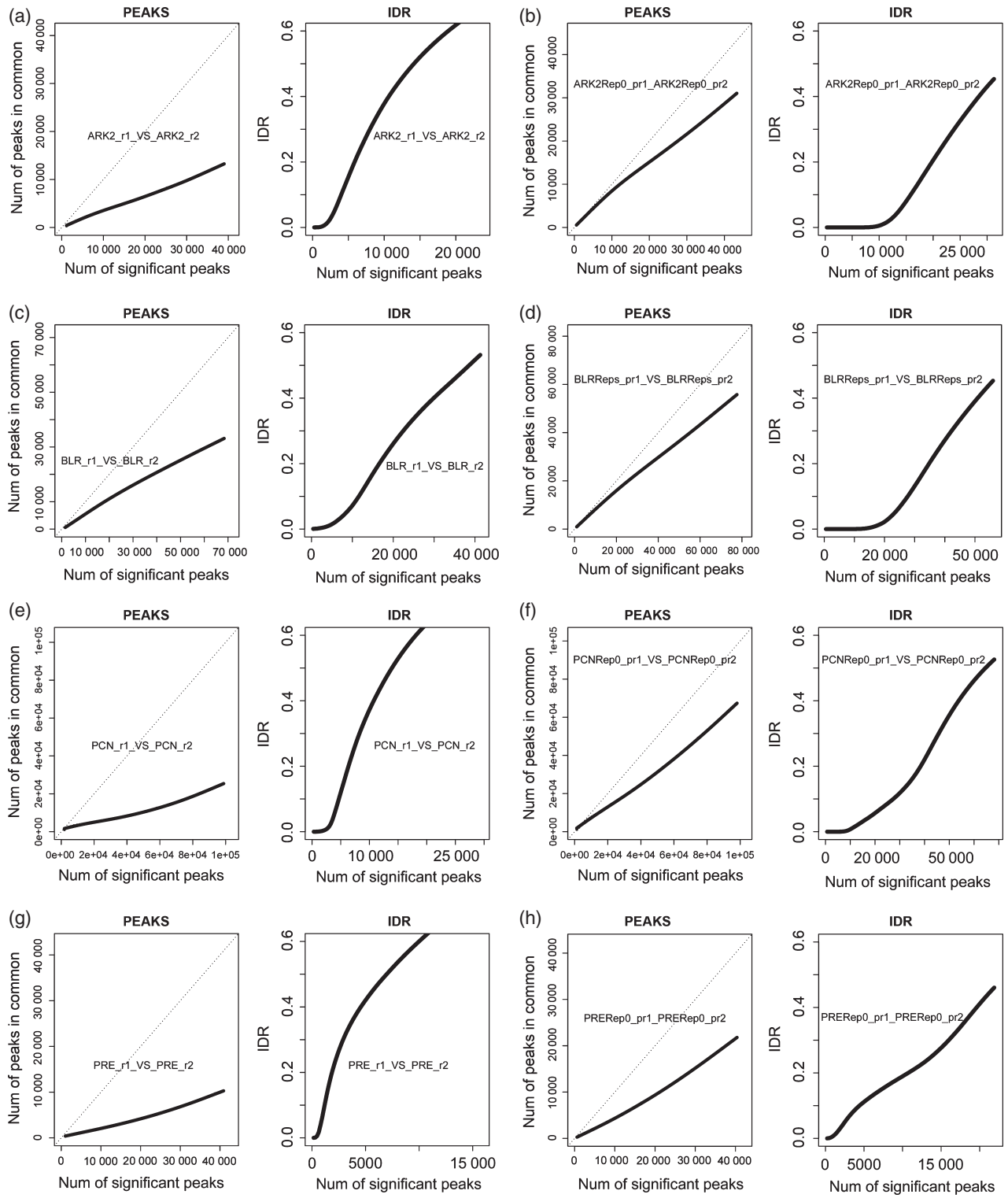


Figure 2. Reproducibility and Irreproducible Discovery Rate (IDR) analysis of ARK2, BLR, PCN, and PRE ChIP-seq replicates.
(a–d) Reproducibility plots of each transcription factor ChIP-seq replicates. Plots on the left compare the number of significant peaks shared by two replicates for increasing numbers of peaks included in the analysis (reproducibility profile). Theoretical perfect congruence between replicates is indicated by the dotted line. Plots on the right compare the IDR of increasing numbers of ChIP peaks for the comparison of two replicates.
(e–h) Reproducibility plots of pooled ChIP-seq sample for each transcription factor. pr1 and pr2 stand for pseudoreplicates of the pooled samples which were derived by randomly splitting the mapped reads into two samples. The plot on left shows the reproducibility profile between all the peaks identified in two replicates. The plot on the right shows the IDR at increasing numbers of peaks selected by IDR criterion.

Table 1 Summary of ChIP-seq peaks and putative target genes for each transcription factor

	#Peaks[a]	Average width (bp)	Average score[b]	#Unassigned peaks[c]	#Targets of multiple peaks[d]	#Targets[e]
ARK1	14 463	1085	740	7	500	13 944
ARK2	2287	1090	855	0	10	2277
BLR	5674	619	517	8	628	4925
PCN	3148	523	884	14	227	2857
PRE	658	568	147	25	54	551

[a]Number of final peaks identified for each transcription factor ChIP-seq.
[b]Average score of all peaks from IDR output.
[c]Number of peaks that have no closest genes assign to in the peak annotation.
[d]Number of genes that are assigned to more than one peak.
[e]Total number of each transcription factor's putative target genes.

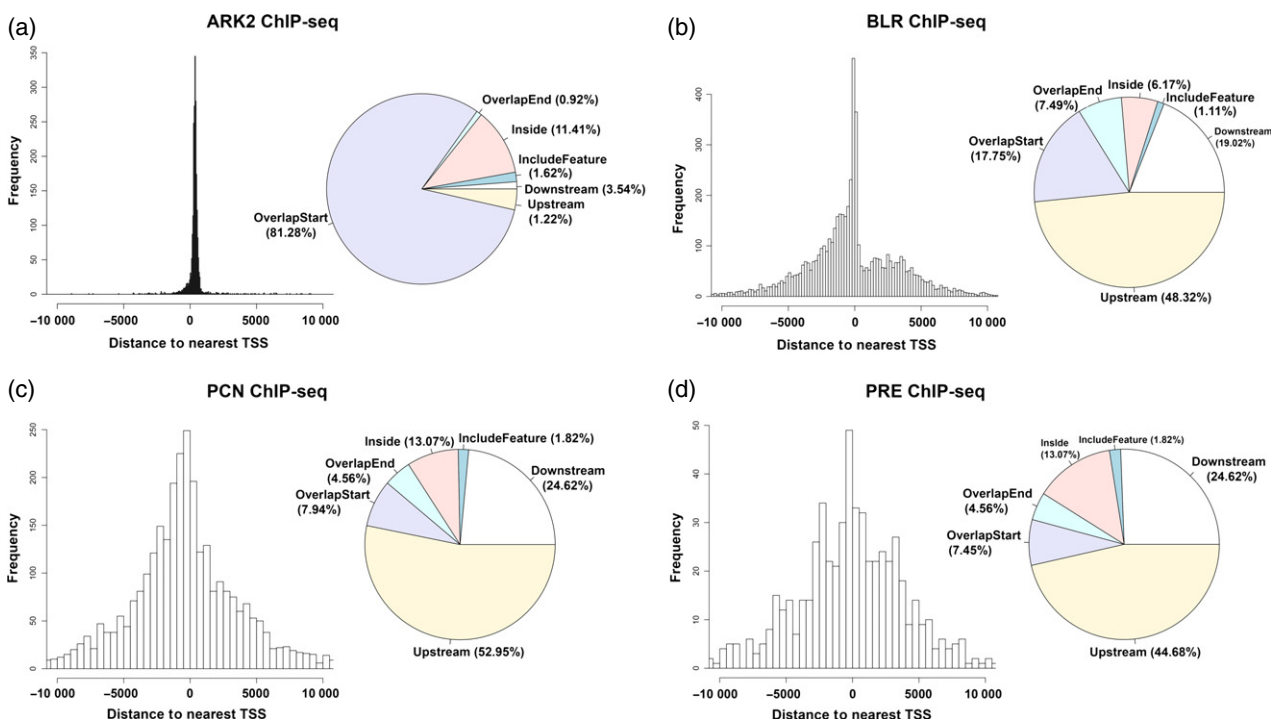


Figure 3. Distribution of ARK2, BLR, PCN, and PRE binding regions relative to putative target genes in *Populus* genome.
The plots on left show the peaks distribution relative to the transcriptional start site (TSS) of putative target genes using the distance calculated from peak center to the TSS. The pie charts on the right show the distribution of binding regions relative to gene features.

Table 2 Distribution of binding regions relative to putative target gene transcription start site (TSS) for each transcription factor. Numbers in the table represent the percentage (%) of corresponding transcription factor's total peaks within each distance range

	< 0.1 kb[a] (%)	0.1–0.5 kb[a] (%)	0.5–1 kb[a] (%)	1–2 kb[a] (%)	2–3 kb[a] (%)	3–5 kb[a] (%)	5 kb[a] < (%)
ARK1	3.67	66.8	17.9	4.58	2.45	2.55	1.99
ARK2	3.98	72.85	16.83	2.40	1.445	1.4	1.09
BLR	8.76	13.91	9.68	17.99	13.41	19.42	16.69
PCN	4.0	10.13	11.02	17.09	14.07	17.50	25.73
PRE	4.2	8.21	9.42	13.22	16.11	17.78	27.20

[a]Distance from peak center to TSS of the putative target gene.

Putative target genes of different transcription factors show distinct functional enrichment

To explore the functions of putative target genes for the different transcription factors, we performed Gene Ontology (GO) enrichment analysis and identified 605 (Liu *et al.* 2015), 282, 106, 80, and 28 over-represented GO categories for ARK1, ARK2, BLR, PCN, and PRE putative target genes, respectively (Table S3).

Detailed analysis showed that putative target genes of different transcription factors have distinct enrichment of 'cellular component' localization: ARK1 and ARK2 putative target genes mainly localize to vesicle membrane, nucleus, and protein complex; BLR putative target genes only have enrichment in ubiquitin ligase complex and cytoskeleton; PCN putative target genes mainly localize to photosynthetic membrane; and PRE putative target genes do not show any 'cellular component' enrichment (Table S3). The enrichment of 'biological process' and 'molecular function' categories display more complex features between different transcription factors (Figure 4 and Table S3). For example, many enriched GO categories from ARK2 and BLR targets were also detected in ARK1 targets. However, categories including 'vesicle-mediated transport' and 'DNA repair' were specifically shared by ARK1 and ARK2 targets while categories such as 'lipid biosynthetic process' and 'actin cytoskeleton organization' were specifically shared by ARK1 and BLR targets. The majority of the GO overlapping categories between ARK2 and BLR targets were broadly involved in metabolic or biosynthetic processes. PCN targets were specifically enriched in 'photosynthesis' related categories; and PRE putative target genes were specifically enriched in 'apoptotic process' and 'programmed cell death' processes.

Table 3 Overlap between ChIP-seq peaks for different transcription factors. Numbers represent the overlapping peaks between the (row, column) pairs of ChIP-seq datasets

	ARK1	ARK2	BLR	PCN	PRE
ARK1	14 463				
ARK2	2285	2287			
BLR	829	122	5674		
PCN	144	13	80	3148	
PRE	52	17	64	17	658

Comparison of transcription factors' binding regions and putative target genes

The transcription factors included in this study were either genetically demonstrated as critical regulators during vascular cambium maintenance and differentiation (ARK1, ARK2, PCN, PRE) or predicted to interact with each other in *Populus* (BLR with ARK1/ARK2) as is found for their Arabidopsis orthologs (e.g. Byrne *et al.*, 2003). We thus analyzed the frequency of common binding regions and putative target genes between different transcription factors, as a first mean of evaluating combinatorial binding between different transcription factors.

Pair-wise comparison of transcription factor binding regions identified significant overlap between Class I KNOX transcription factor binding regions (Table 3): more

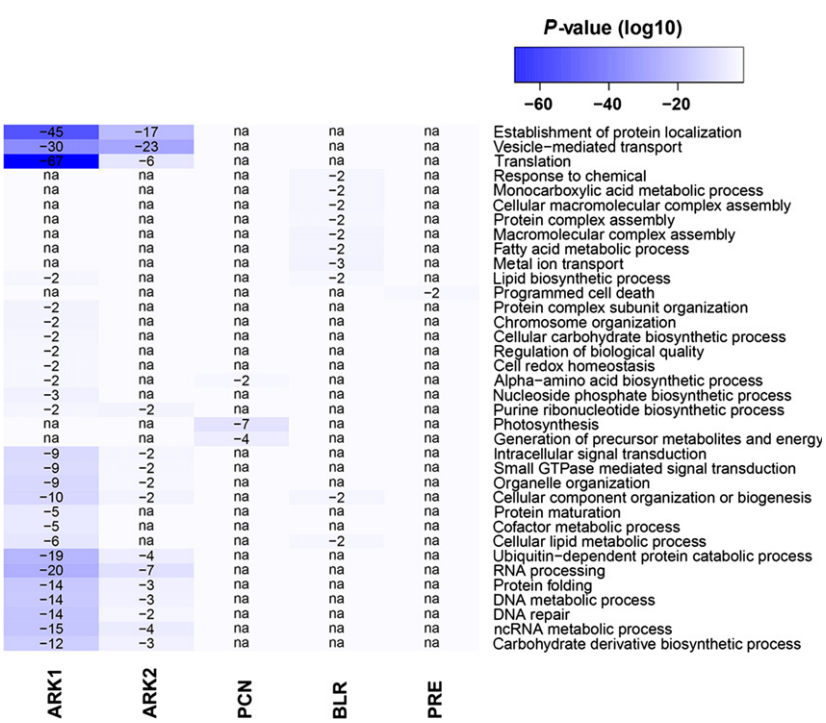


Figure 4. Comparison of enriched Gene Ontology (GO) categories in different transcription factors' putative target genes.

than 99% of ARK2 binding regions overlapped with ARK1 binding regions. This could reflect that these two transcription factors are within the same family and, as in the case of their Arabidopsis orthologs (Byrne *et al.*, 2002), have partially redundant functions. A lower degree of overlap, 829 binding regions, was found between ARK1 and BLR, which accounted for 15% of BLR binding regions. The percentage of overlapping binding regions in other pair-wise comparisons involving Class III HD ZIPs were lower (Table 3), roughly reflecting the expected relationships among these transcription factors based on their phylogenetic relationships and previous functional and genetic characterization in *Populus* and other species.

We also performed pair-wise comparison with transcription factors putative target genes (Table 4). While the trends in relationships among the transcription factors were similar with the binding regions comparison, the number of overlapping putative target genes was not the same as the number of binding regions due to differences in binding site distributions relative to genes. For example, 2285 peaks overlapped between ARK1 and ARK2 while only 2250 putative target genes overlapped between the two, which was mainly caused by cases where multiple binding regions were assigned to one gene. Conversely, 1891 putative target genes overlapped between ARK1 and BLR, accounting for 38% of BLR putative target genes. This percentage is much higher than the percentage of overlapping binding regions, which was the result of a common putative target gene assigned to non-overlapping binding regions of the two transcription factors. These same factors also caused the increase of overlapping putative target genes in other pair-wise comparisons involving PCN and PRE.

The studies of overlapping transcription factors binding regions and putative target genes (Tables 3 and 4) showed that the overlapping putative target genes between two transcription factors can be further divided into two subgroups: genes that are bound by both transcription factors with overlapping binding regions versus genes that are bound by both transcription factors but with non-overlapping binding regions. We used the comparison of ARK1 and BLR as an example to further dissect whether there are major differences between these two subgroups of putative target genes that could, for example, reflect the effects of direct interactions (e.g. heterodimerization) between the transcription factors. Out of the 1891 overlapping putative target genes bound by either, 787 were bound by overlapping binding regions (ARK1/BLR_overlap) while 1104 were bound by non-overlapping binding regions (ARK1/BLR_non-overlap). First, we performed GO analysis to test whether the functional enrichment among these two subgroups of putative target genes were different. Totally, 128 and 101 enriched GO categories were identified for the 787 ARK1/BLR_overlap and 1104 ARK1/BLR_non-overlap putative target genes, respectively (Tables S4 and S5). Surprisingly, only seven categories overlapped between the 128 and 101 enriched GO categories, suggesting those two subgroups of overlapping putative target genes participate in fundamentally different biological processes (Figure 5a and Tables S4 and S5). Next, we looked in more detail whether the ARK1 and BLR binding regions associated with the ARK1/BLR_overlap or ARK1/BLR_non-overlap genes also have significant differences in distribution relative to the putative target genes. We found that the ARK1 binding regions associated with these two subgroups of putative target genes had similar distribution patterns (Figure 5b, c), however, the BLR binding regions associated with these two subgroups of putative target genes were fundamentally different: BLR binding regions associated with the ARK1/BLR_overlap putative target genes were highly enriched around the TSS regions (Figure 5d), while BLR binding regions associated with the ARK1/BLR_non-overlap putative target genes were largely absent from the TSS regions (Figure 5e). Distribution of the binding regions which do not associate with the overlapping putative target genes displayed similar patterns with all binding regions for both ARK1 and BLR, respectively (Figure S3).

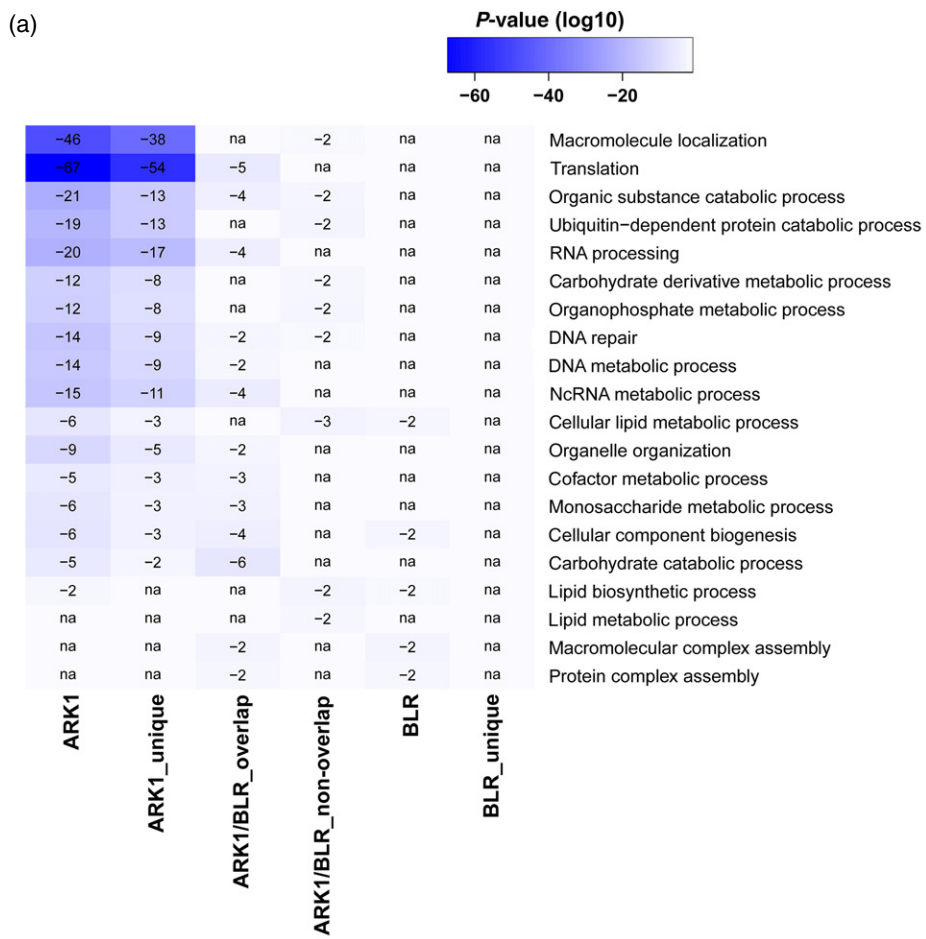
Integration of ChIP-seq data into PopGenIE

We incorporated all the transcription factor ChIP-seq results described here into the web-based resource PopGenIE (http://popgenie.org) to facilitate visualization of transcription factor binding in the *Populus* genome and to integrate these data with other publicly available *Populus* gene expression data. We provide here examples of the

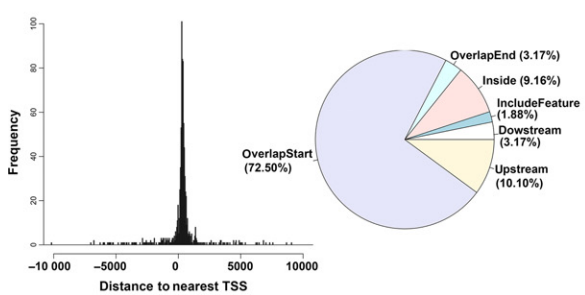
Table 4 Co-binding of putative target genes identified in different ChIP-seq experiments for the indicated transcription factors. Numbers represent number of unique genes targeted by the transcription factors combinations compared in the rows and columns. *P*-values represent the exact hypergeometric probability. Representation Factor (RF) indicates overrepresentation (RF > 1) or underrepresentation (RF < 1) of co-binding

	ARK1	ARK2	BLR	PCN	PRE
ARK1	13 944				
ARK2	2250 ($P < 0.000e^{+00}$, RF: 3.0)	2277			
BLR	1891 ($P < 2.464e^{-15}$, RF: 1.1)	245 ($P < 0.060$, RF: 0.9)	4925		
PCN	848 ($P < 5.920e^{-06}$, RF: 0.9)	105 ($P < 3.105e^{-06}$, RF: 0.7)	473 ($P < 3.773e^{-15}$, RF: 1.4)	2857	
PRE	125 ($P < 1.998e^{-08}$, RF: 0.7)	22 ($P < 0.072$, RF: 0.7)	105 ($P < 4.694e^{-07}$, RF: 1.6)	58 ($P < 7.824^{e-04}$, RF: 1.5)	551

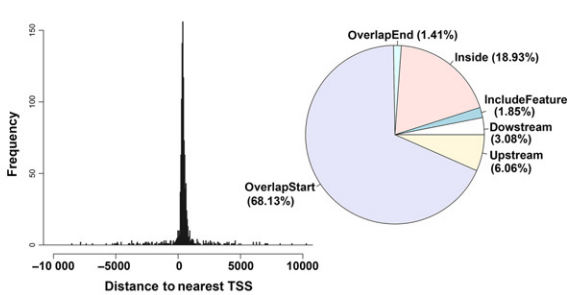
(a)



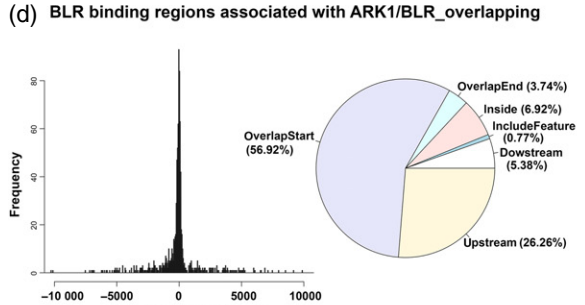
(b) ARK1 binding regions associated with ARK1/BLR_overlapping



(c) ARK1 binding regions associated with ARK1/BLR_non-overlapping



(d) BLR binding regions associated with ARK1/BLR_overlapping



(e) BLR binding regions associated with ARK1/BLR_non-overlapping

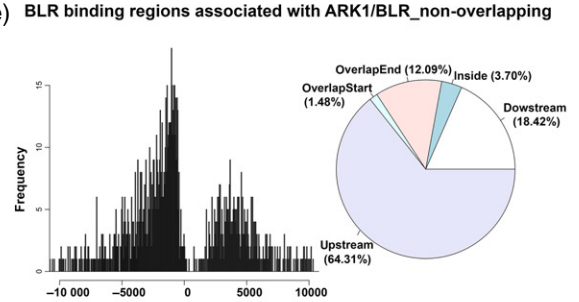


Figure 5. Overlapping study of ARK1 and BLR putative target genes.
(a) The ARK1/BLR_overlap and ARK1/BLR_non-overlap subgroups of ARK1 and BLR overlapping putative target genes have distinct functional enrichments. Categories: ARK1, all ARK1 putative target genes; ARK1_unique, ARK1 putative target genes that were not bound by BLR; ARK1/BLR_overlap, genes which were bound by overlapping ARK1 and BLR binding regions; ARK1/BLR_non-overlap, genes which were bound by ARK1 and BLR but with non-overlapping binding regions; BLR, all BLR putative target genes; BLR_unique, BLR putative target genes that were not bound by ARK1.
(b, c) Genome-wide distribution of ARK1 binding regions relative to ARK1/BLR_overlap and ARK1/BLR_non-overlap.
(d, e) Genome-wide distribution of BLR binding regions relative to ARK1/BLR_overlap and ARK1/BLR_non-overlap. (b–e) The plots on the left show the distribution of binding regions relative to transcriptional start site (TSS) of genes and the pie charts on the right show the distribution of binding regions relative to gene features.

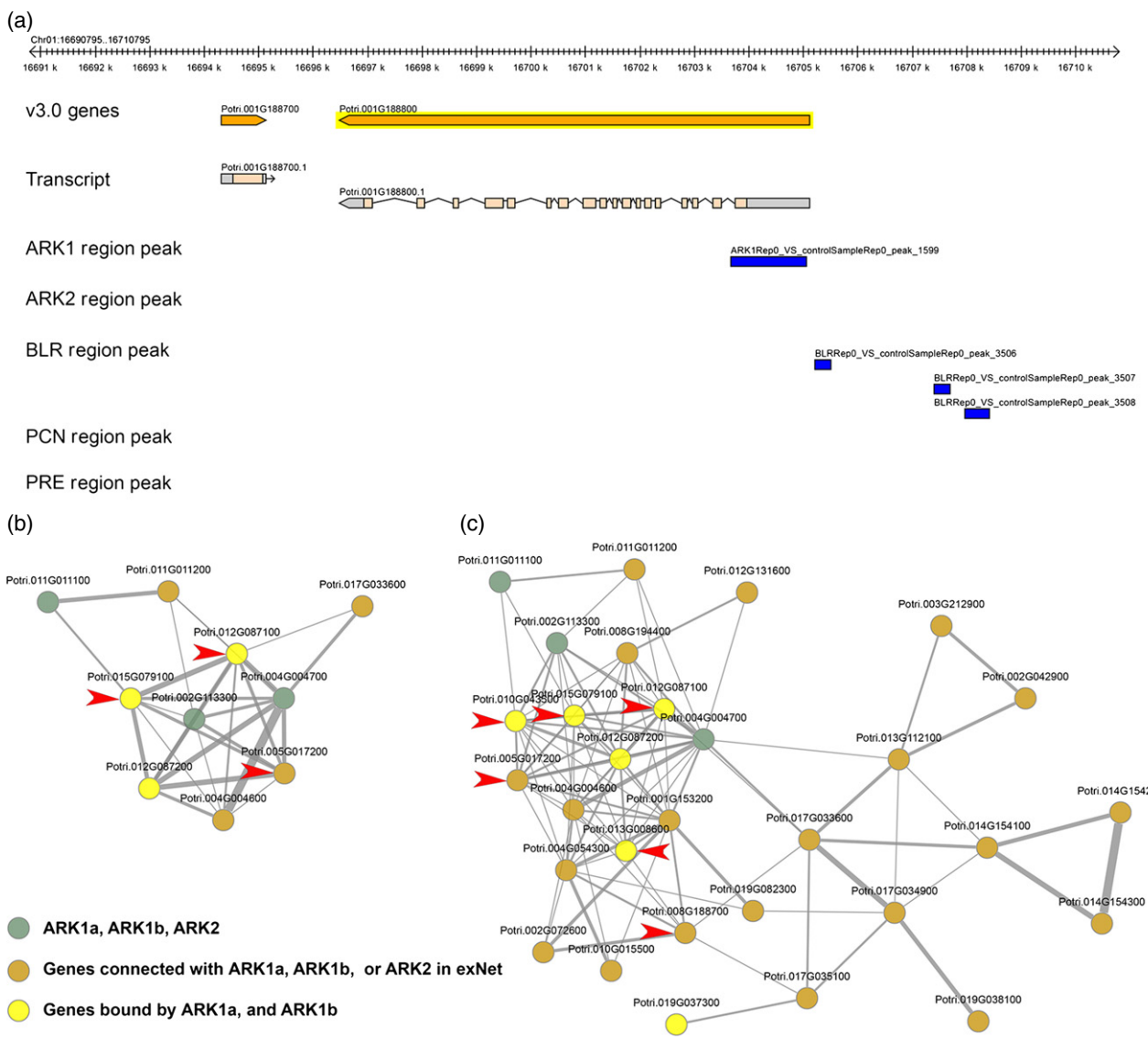


Figure 6. Illustrations of the transcription factors ChIP-seq data with PopGenIE.
(a) The output from the Genome Browser (GBrowse) showing ARK1 and BLR binding to *PCN* (Potri.001G188800).
(b, c) Visualization of the distribution of ARK1 targets among its neighbor genes in the pre-calculated co-expression network of exNet. We seeded exNet with *ARK1a, ARK1b,* and *ARK2,* and then expanded the network to 10 (b) or 30 genes (c) and colored the ARK1 bound genes after each expansion. Orthologs of Arabidopsis KNAT2/6 are indicated with arrowheads.

utility of PopGenIE using the ChIP-seq data. Visualization of binding regions in the genome is facilitated with the 'Genome Browser (GBrowse)' tool, and is illustrated by ARK1 and BLR binding around the *PCN* transcription factor gene (Potri.001G188800) (Figure 6a), which was previously reported as an ARK1 target gene (Liu *et al.* 2015). The ChIP-seq data can also be explored with the PopGenIE co-expression network visualization tool 'exNet'. This tool utilizes precomputed co-expression networks that can be overlaid with ChIP-seq binding status of specific transcrip-

tion factors to individual putative target genes. As an example, an exNet analysis was seeded with the genes *ARK1a*, *ARK1b*, and *ARK2*, and then the network was expanded to include co-expression neighbors using the default 'expand' and 'display' threshold values, as shown in Figure 6(b, c). The co-expression neighbors that are also putative target genes of ARK1 were then colored. Interestingly, the results showed that within this small-scale network, ARK1 preferentially binds to genes tightly co-expressed with it (neighbors with path lengths on one) than to genes with longer path lengths, consistent with a role of an upstream, fan-out, signal mediator and stabilizer. Genes within the module were evaluated for GO enrichment using both the most recent annotations for *Populus* (Version 3.0) and annotations for Arabidopsis orthologs (http://www.phytozome.net/poplar.php). Interestingly, of the 22 genes in the module with annotated Arabidopsis orthologs, nine are assigned GO categories involving meristem functions (GO:0048507 meristem development, GO:0010014 meristem initiation, GO:0010073 meristem maintenance, GO:0009934 regulation of meristem structural organization: Table S6). These GO terms are only found in Arabidopsis, highlighting a limitation of the current *Populus* annotations. Six of the genes in the cluster encode Class I KNOX transcription factors are annotated as similar to KNOTTED-like Arabidopsis 6 (KNAT6) (arrowheads, Figure 6c). It should be noted that these annotations are an oversimplification, and that these genes are actually co-orthologs of the Arabidopsis KNAT2/6 subclass described previously (Mukherjee *et al.*, 2009). Mirroring this result in Arabidopsis, KNAT6 has previously been shown act redundantly with ARK1 ortholog SHOOTMERISTEMLESS (STM) (Belles-Boix *et al.*, 2006) and to physically interact with STM interactor BELLRINGER (BLR) (Ragni *et al.*, 2008). Taken together, these brief examples show that integrating *Populus* transcription factor ChIP-seq data with co-expression analysis can be valuable for discovery and hypothesis generation.

DISCUSSION

Recently, network modeling with large-scale datasets conducted in different species has revealed a high complexity of regulation at the transcriptional level. Modeling and fine scale resolution of the intricacies of gene regulation have been shown to require the integration of multiple types of data such as transcription factor binding, gene expression, histone modification, DNA methylation, and protein–protein interaction in both animals (ENDODE Project Consortium 2012) and plants (Heyndrickx *et al.*, 2014). In this report, we described resources for identifying and evaluating genome-wide binding regions for transcription factors in the model tree genus, *Populus,* including genome-wide binding data for four transcription factors (Class I KNOX ARK2; BEL-like homeodomain family member BLR; and Class III HD ZIP PCN and PRE), and publically available tools for the analysis of *Populus* ChIP-seq data. These data and analysis tools are an important step towards modeling of transcriptional networks underlying growth and development of undomesticated tree species, and can now be expanded to include additional transcription factors and chromatin marks, or integrated with gene expression data.

We performed ChIP-seq for the cambium and its recent derivatives from mature *Populus trichocarpa* trees growing in their natural environment, using peptide-specific antibodies for each transcription factor. This approach takes advantage of the radial symmetry of the *Populus* stem to harvest gram amounts of cambium and recent derivatives after removal of the bark, and allows identification of endogenous transcription factor binding regions during natural growth conditions. Using the IDR pipeline developed by the ENCODE project (Li *et al.*, 2011), we identified 2287, 5674, 3148, and 658 highly reproducibly binding regions which were assigned to 2277, 4925, 2857, and 551 unique putative target genes for ARK2, BLR, PCN, and PRE, respectively. We evaluated the specificity of the ChIP-seq data from different perspectives. First, genome-wide distribution analysis found that these transcription factors had distinct binding patterns, such as preferential promoter-proximal binding (ARK2) versus bimodal binding (BLR) relative to putative target gene bodies. Second, GO analysis showed different transcription factors bind to different functional categories of genes. Third, analysis of overlapping transcription factor binding regions and putative target genes found most commonality between the pairs ARK1:ARK2 and ARK1:BLR than all other transcription factor pairs, which parallels the homologous origins of the Class I KNOX transcription factors ARK1 and ARK2, and the well known dimerization within or between Class I KNOX and BEL-like proteins (BLR), e.g. (Bellaoui *et al.*, 2001).

The range for the number of binding sites among transcription factors found here is in line with that seen in similar studies in Arabidopsis (Heyndrickx *et al.*, 2014). While some variation is likely attributable to technical issues such as variation in performance among antibodies used for ChIP, this variation also likely reflects fundamental differences in function of the transcription factors under study. Interestingly, ARK1 (Groover *et al.*, 2006) and ARK2 (Du *et al.*, 2009) are expressed in both the cambial zone as well as the shoot apical meristem. Both transcription factors also have large numbers of binding sites (14 463 sites for ARK1 and 2287 for ARK2), and raises the point that binary responses (transcription on vs transcription off) of target genes by binding of these transcription factors would seem unlikely in these two distinct meristems. Indeed, we previously found that ARK1 binding data was only modestly predictive of differential gene expression in an ARK1 overexpression mutant (Liu *et al.*, 2015), suggesting that ARK1 is not by itself capable of initiating transcription.

We are currently testing the hypothesis that ARK1 and ARK2 are more general factors, whose target genes' expression can be modified by meristem-specific combinatorial binding with other transcription factors. Attractive candidates are the BEL-like transcription factors in *Populus* such as BLR, which are known to heterodimerize with KNOX transcription factors in other plant species (e.g. Byrne *et al.*, 2003; Smith and Hake, 2003).

Indeed, in most cases in both animals (Cheng *et al.*, 2012) and plants (Heyndrickx *et al.*, 2014) binding of a single transcription factor is not sufficient to predict gene expression levels, and instead transcription factors tend to co-associate in context-specific patterns to integrate developmental cues into determining gene expression outcomes (Karczewski *et al.*, 2014; Teng *et al.*, 2014). Our analysis of overlapping binding regions and putative target genes between different transcription factors examined two subgroups of putative target genes: putative target genes that were assigned to overlapping versus putative target genes assigned to non-overlapping binding regions of two transcription factors. We used the overlapping putative target genes of Class I KNOX ARK1 and BEL-like BLR in an illustrative analysis, and found that the functional enrichment and distribution pattern of BLR binding regions associated with these two subgroups of putative target genes were fundamentally different. These results suggest that future studies examining different modes of combinatorial binding of transcription factors in *Populus* could further dissect how the distance, clustering and ordering patterns among different transcription factors and their putative target genes contribute to gene regulation and transcriptional regulatory network properties.

To assist the research community in furthering transcriptional network modeling in trees, we established a VM resource through iPlant (Goff *et al.*, 2011) that can be used to process and evaluate *Populus* ChIP-seq data. We also incorporated our transcription factor ChIP-seq data into the web-based resource PopGenIE, where users can explore the data using a variety of user-friendly, integrated tools. As an example, we used the integration of ChIP-seq and co-expression data in PopGenIE to define a module of genes co-expressed with ARK1 and ARK2. Interestingly, this module revealed co-expression and putative ARK1 binding to multiple KNOTTED-like Arabidopsis 2/6 (KNAT2/6)-like genes. These results clearly point to the *Populus* KNAT2/6 orthologs as genes of interest regarding KNOX regulation of secondary growth and wood formation, and prompt new hypotheses including *Populus* KNAT2/6 heterodimerization and/or co-binding of target genes with ARK1. Likewise, users can deploy the tools within PopGenIE to develop and test new research questions regarding the transcriptional regulation of secondary growth, evolution of transcription factor binding, or regulatory relationships among individual genes or gene modules in *Populus.*

EXPERIMENTAL PROCEDURES

ChIP-seq procedures

All ChIP-seq experiments were performed with vascular cambium and the recent derivatives from mature *Populus trichocarpa* trees. Tissue collection and fixation, chromatin immunoprecipitation, and library preparations were as previously described (Liu *et al.* 2015). Unique peptides with high predicted antigenicity were identified in, ARK2 (Potri.002G113300.1, CHGPLRIFNSDDKSEG), BLR (Potri.010G197300.1, VTKEKSPRYGKTERG), PCN (Potri.001G188800.1, LKSSSEGSESI), and PRE (Potri.004G211300.1, LDKIFNESGRQALYTEF) for antibody production as described (Liu *et al.* 2015). Recombinant ARK1a, ARK1b, and ARK2 proteins were prepared as previously described (Liu *et al.* 2015). *BLR, PCN,* and *PRE* CDS were amplified using primers adding *Bam*HI and *Not*I sites (Table S7), and cloned into expression vector *pET23a.* All constructs were sequenced and transformed into *Escherichia coli* BL21 cells (Invitrogen, Grand Island, NY USA) for protein induction. The protein preparation and western blot were performed as before (Liu *et al.* 2015).

ChIP-seq data analysis

Details of ChIP-seq data analysis are available in (Liu *et al.* 2015). Briefly, processed sequence reads were mapped to the V3 *P. trichocarpa* genome (http://www.phytozome.net/poplar.php). Chip-seq peaks were called and evaluated using the ENCODE IDR pipeline (Li *et al.*, 2011) with MACS2 (Zhang *et al.*, 2008) using 'input' library controls to estimate local sequence read mapping bias. Putative target genes and genome-wide distributions of ChIP-seq peaks were calculated using ChIPpeakAnno (Zhu *et al.*, 2010). Peak intervals identified with the IDR analysis were assigned to the nearest gene based on proximity to peak interval, which were then treated as putative target genes. Overlapping of ChIP-seq peak intervals were evaluated with intersectBed (https://code.google.com/p/bedtools/wiki/Usage).

Gene ontology analysis

The GO enrichment analysis was conducted using the Bioconductor package GOstats (Falcon and Gentleman, 2007) as previously described (Liu *et al.* 2015). For the plots in Figure 4, all enriched goBP terms of five transcription factors were pooled, sorted by the term size, and filtered for enriched GO terms with sizes ranging from 100 to 1000, resulting in 120 terms total. Redundant terms were manually removed resulting in 36 GO categories for the heat map plot. For the plots in Figure 5(a), the enriched GO terms of different subgroups of ARK1 and BLR targets were processed as above yielding 20 GO categories for the heat map plots. The heatmap.2 function from the gplots package was used to plot the enriched GO categories based on *P*-value (log10).

Data archiving

All ChIP-seq sequences have been archived in the NCBI SRA with accession numbers SRP053368 for ARK2, BLR, PCN, PRE ChIP-seq data, and SRP042635 for ARK1 ChIP-seq and input libraries.

ACKNOWLEDGEMENTS

We would like to thank Brian Stanton, Kathy Haiby, and Rich Shuren of Greenwood Resources for facilitating collection of cambium and developing wood/bark samples. Illumina sequencing was performed by the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. This work was supported by the Agriculture and

Food Research initiative competitive grant 2011-67013-30062 of the USDA National Institute of Food and Agriculture. MZ is supported by NSF Postdoctoral Research Fellowship in Biology Grant IOS-1402064.

CONFLICT OF INTEREST

The authors declare no competing interests.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Peptide-specific antibodies used for transcription factors ChIP-seq.

Figure S2. Cross-correlation analysis for evaluating the quality of ChIP-seq replicates.

Figure S3. Genome-wide distribution of ARK1 and BLR binding regions relative to the non-overlapping putative target genes.

Table S1. Summary of transcription factor ChIP-seq replicates.

Table S2. List of ChIP-seq peaks and putative target genes for all transcription factors.

Table S3. GO enrichment of ARK2, BLR, PCN, and PRE putative target genes.

Table S4. GO enrichment of ARK1/BLR_overlap.

Table S5. GO enrichment of ARK1/BLR_non-overlap.

Table S6. Annotations from *Populus* Phytozome V3.0 and Arabidopsis orthologs for genes included in co-expression network presented in Figure 6.

Table S7. Primers used for cloning and PCR verification.

REFERENCES

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Bellaoui, M., Pidkowich, M.S., Samach, A., Kushalappa, K., Kohalmi, S.E., Modrusan, Z., Crosby, W.L. and Haughn, G.W. (2001) The Arabidopsis BELL1 and KNOX TALE homeodomain proteins interact through a domain conserved between plants and animals. *Plant Cell*, **13**, 2455–2470.

Belles-Boix, E., Hamant, O., Witiak, S.M., Morin, H., Traas, J. and Pautot, V. (2006) KNAT6: an Arabidopsis homeobox gene involved in meristem activity and organ separation. *Plant Cell Online*, **18**, 1900–1907.

Brady, S.M., Zhang, L., Megraw, M. *et al.* (2011) A stele-enriched gene regulatory network in the Arabidopsis root. *Mol. Syst. Biol.* **7**, 459. doi: 10.1038/msb.2010.114

Byrne, M.E., Groover, A.T., Fontana, J.R. and Martienssen, R.A. (2003) Phyllotactic pattern and stem cell fate are determined by the Arabidopsis homeobox gene BELLRINGER. *Development*, **130**, 3941–3950.

Byrne, M.E., Simorowski, J. and Martienssen, R.A. (2002) ASYMMETRIC LEAVES1 reveals knox gene redundancy in Arabidopsis. *Development*, **129**, 1957–1965.

Cheng, C., Alexander, R., Min, R. *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.* **22**, 1658–1667.

Du, J., Mansfield, S.D. and Groover, A.T. (2009) The *Populus* homeobox gene ARBORKNOX2 regulates cell differentiation during secondary growth. *Plant J.* **60**, 1000–1014.

Du, J., Miura, E., Robischon, M., Martinez, C. and 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.

ENDODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

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

Gaudinier, A., Zhang, L., Reece-Hoyes, J.S. *et al.* (2011) Enhanced Y1H assays for Arabidopsis. *Nat. Methods*, **8**, 1053–1055.

Goff, S.A., Vaughn, M., McKay, S. *et al.* (2011) The iPlant collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* **2**, doi: 10.3389/fpls.2011.00034.

Groover, A.T., Mansfield, S.D., DiFazio, S.P., Dupper, G., Fontana, J.R., Millar, R. and Wang, Y. (2006) The *Populus* homeobox gene ARBORKNOX1 reveals overlapping mechanisms regulating the shoot apical meristem and the vascular cambium. *Plant Mol. Biol.* **61**, 917–932.

Hertzberg, M., Aspeborg, H., Schrader, J. *et al.* (2001) A transcriptional roadmap to wood formation. *Proc.Natl Acad. Sci. USA*, **98**, 14732–14737.

Heyndrickx, K.S., de Velde, J.V., Wang, C., Weigel, D. and Vandepoele, K. (2014) A functional and evolutionary perspective on transcription factor binding in *Arabidopsis thaliana*. *Plant Cell Online*, **26**, 3894–3910.

Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372.

Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.

Karczewski, K.J., Snyder, M., Altman, R.B. and Tatonetti, N.P. (2014) Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. *PLoS Genet.* **10**, e1004122.

Landt, S.G., Marinov, G.K., Kundaje, A. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831.

Li, Q., Brown, J.B., Huang, H. and Bickel, P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.

Liu, L., Missirian, V., Zinkgraf, M., Groover, A. and 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 Genom.* **15** (Suppl 5), S3.

Liu, L., Zinkgraf, M., Patzold, E., Beers, E., Filkov, V. and Groover, A. (2015) The Populus ARBORKNOX1 homeodomain transcription factor regulates woody growth through binding to evolutionarily conserved target genes of diverse function. *New Phytol.*, **205**, 1469–8137.

Long, T., Brady, S. and Benfey, P. (2008) Systems approaches to identifying gene regulatory networks in plants. *Annu. Rev. Cell Dev. Biol.* **24**, 81–103.

Mukherjee, K., Brocchieri, L. and Bürglin, T.R. (2009) A comprehensive classification and evolutionary analysis of plant Homeobox genes. *Mol. Biol. Evol.* **26**, 2775–2794.

Ragni, L., Belles-Boix, E., Günl, M. and Pautot, V. (2008) Interaction of KNAT6 and KNAT2 with BREVIPEDICELLUS and PENNYWISE in Arabidopsis inflorescences. *Plant Cell*, **20**, 888–900.

Robertson, G., Hirst, M., Bainbridge, M. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657.

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

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

Sjodin, A., Street, N.R., Sandberg, G., Gustafsson, P. and Jansson, S. (2009) The *Populus* Genome Integrative Explorer (PopGenIE): a new resource for exploring the *Populus* genome. *New Phytol.* **182**, 1013–1025.

Smith, H.M.S. and Hake, S. (2003) The interaction of two homeobox genes, BREVIPEDICELLUS and PENNYWISE, regulates internode patterning in the Arabidopsis inflorescence. *Plant Cell*, **15**, 1717–1727.

Taylor-Teeples, M., Lin, L., de Lucas, M. *et al.* (2015) An Arabidopsis gene regulatory network for secondary cell wall synthesis. *Nature*, **517**, 571–575.

Teng, L., He, B., Gao, P., Gao, L. and Tan, K. (2014) Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets. *Nucleic Acids Res.* **42**, e24.

Van de Poel, B., Bulens, I., Hertog, M.L., Nicolai, B.M. and Geeraerd, A.H. (2014) A transcriptomics-based kinetic model for ethylene biosynthesis in tomato (*Solanum lycopersicum*) fruit: development, validation and exploration of novel regulatory mechanisms. *New Phytol.* **202**, 952–963.

Zhang, Y., Liu, T., Meyer, C. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137.

Zhu, L.J., Gazin, C., Lawson, N.D., Pages, H., Lin, S.M., Lapointe, D.S. and Green, M.R. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, **11**, 237.

