Running head
Shared plant cis-regulatory sequence combinations

Corresponding author
Xiaoman Li
Harris Corporation and Engineering Center 346
4000 Central Florida Blvd Orlando, FL, 32816, USA
Phone: 407-823-4811
Fax: 407-823-5835
Email: xiaoman@mail.ucf.edu

Journal research area
Bioinformatics
Thousands of cis-regulatory sequence combinations are shared by *Arabidopsis* and Poplar

Jun Ding¹, Haiyan Hu¹, Xiaoman Li²

¹Department of Electrical Engineering and Computer Science, ²Burnett School of Biomedical Science, University of Central Florida, Orlando, FL, 32816, USA
Financial source: NIH R01HG004359, NSF CBET 1125676
Corresponding authors: Haiyan Hu (haihu@cs.ucf.edu), Xiaoman Li (xiaoman@mail.ucf.edu)
ABSTRACT
The identification of cis-regulatory modules can greatly advance our understanding of gene regulatory mechanisms. Despite the existence of binding sites of more than three transcription factors in a cis-regulatory module, studies in plants often consider only the co-occurrence of binding sites of one or two transcription factors. In addition, cis-regulatory module studies in plants are limited to combinations of only a few families of transcription factors. It is thus not clear how widespread plant transcription factors work together, which transcription factors work together to regulate plant genes, and how the combinations of these transcription factors are shared by different plants. To fill these gaps, we applied a frequent pattern mining based approach to identify frequently used cis-regulatory sequence combinations in the promoter sequences of two plant species, *Arabidopsis thaliana* and *Populus trichocarpa*. A cis-regulatory sequence here corresponds to a DNA motif bound by a transcription factor. We identified 18638 combinations composed of 2 to 6 cis-regulatory sequences that are shared by the two plant species. In addition, with known cis-regulatory sequence combinations, gene function annotation, gene expression data, and known functional gene sets, we shown that the functionality of at least 96.8% and 65.2% of these shared combinations in *Arabidopsis* are partially supported, under a false discovery rate of 0.1 and 0.05, respectively. Finally, we discovered that 796 of the 18638 combinations might relate to functions that are important in bioenergy research. Our work will facilitate the study of gene transcriptional regulation in plants.
INTRODUCTION

Identifying cis-regulatory modules (CRMs) is important for the understanding of gene transcriptional regulation (Singh, 1998; Yuh et al., 1998; Blanchette et al., 2006; Hu et al., 2008; Cai et al., 2010). CRMs are short DNA regions of a few hundred base pairs (bp) that contain multiple transcription factor binding sites (TFBSs). It is estimated that there are at least five times more CRMs than genes in high eukaryotic species (Davidson, 2006). In addition, in high eukaryotes such as metazoans, CRMs instead of individual TFBSs often determine the spatial and temporal expression patterns of neighboring genes (Singh, 1998; Yuh et al., 1998). Therefore, identification of CRMs is crucial for studying gene transcriptional regulation.

In past decades, many studies, both experimental and computational, have identified CRMs in animals (Yuh et al., 1998; Kel-Margoulis et al., 2000; Loots et al., 2000; Frith et al., 2001; Andrioli et al., 2002; Zhou and Wong, 2004; Gupta and Liu, 2005; Blanchette et al., 2006; Hu et al., 2008; Cai et al., 2010). Two of the best experimentally studied CRM systems so far may be the CRMs in the Eve gene in Drosophila and those in the Endo16 gene in the sea urchin, in which not only the locations of TFBS instances but also the expression patterns controlled by these CRMs are identified (Howard and Davidson, 2004). Arguably, two databases, the Redfly database for CRMs in Drosophila (Gallo et al., 2006) and the VISTA Enhancer Browser for CRMs in human and mouse (Visel et al., 2007), have the largest collections of experimentally verified CRMs. Besides the experimentally verified CRMs, there are many computational methods developed for CRM prediction (Frith et al., 2002; Bailey and Noble, 2003; Frith et al., 2003; Johansson et al., 2003; Sinha et al., 2003; Alkema et al., 2004; Zhou and Wong, 2004; Gupta and Liu, 2005; Hu et al., 2008; Cai et al., 2010). Several computationally predicted CRM resources are also available (Blanchette et al., 2006; Cai et al., 2010). Computationally predicted CRMs are indispensable and useful due to the enormous size of genomes and the time consuming process to verify a CRM experimentally.

Compared with CRM studies in animals, research on plant CRMs is limited. Many experimental CRM studies in plants consider co-occurrence of TFBSs of individual TFs with multiple DNA binding domains or a pair of interacting TFs (Solano et al., 1995; Chen et al., 1996; Singh, 1998; Kagaya et al., 1999; Nagano et al., 2001; Abe et al., 2003; Konishi and Yanagisawa, 2010),
although a plant CRM may consist of TFBSs of 3 or more TFs (Baudry et al., 2004; Kim et al., 2004; Akyildiz et al., 2007; Kawashima et al., 2009; Wang et al., 2010). In addition, these studies often identify only individual CRMs for individual genes, and only consider a limited number of TFs (Steffens et al., 2005). Furthermore, compared with a large number of computational CRM studies on animals mentioned above, only a couple of computational CRM studies in plants exist. These studies often focus on the identification of 2 to 3 TFs that are potentially working together to regulate target genes (Steffens et al., 2005; Vandepoele et al., 2006; Chang et al., 2008; Michael et al., 2008). Therefore, it is not clear how widespread CRMs are in plants, which multiple TFs likely coordinate the transcriptional regulation of their target genes, and how these TF combinations are shared by different plant species.

To fill these gaps in our knowledge about plant CRMs, we developed computational approaches to study plant CRMs in the upstream 1 kilo base (kb) regions of all genes in two plant species, *Arabidopsis thaliana* (*Arabidopsis*) and *Populus trichocarpa* (poplar) by using cis-regulatory sequences from the PLACE database (Higo et al., 1999). These cis-regulatory sequences in the PLACE database are normally called motifs, each of which represents the common pattern of TFBSs bound by a TF. In this paper, we called them cis-regulatory sequences instead of motifs because a PLACE motif often has only a few experimentally verified TFBSs and we often do not know the TF behind a PLACE motif. We found that 18638 cis-regulatory sequence combinations composed of 2 to 6 cis-regulatory sequences are shared by *Arabidopsis* and Poplar. In addition, we discovered that a large number of these shared combinations in *Arabidopsis* are partially supported by various sources of functional evidence. The developed methods and the predicted shared cis-regulatory sequence combinations will facilitate future CRM studies in plants.

**RESULTS**

**CRMs are widespread in *Arabidopsis* and poplar gene upstream sequences**

We modified our previously developed method (Cai et al., 2010) for the identification of cis-regulatory sequence combinations and applied it to the upstream 1kb region of every gene (relative to gene transcription start sites) in *Arabidopsis* and poplar (Figure 1). In brief, first, starting from all cis-regulatory sequences in the PLACE database (Higo et al., 1999), we scanned the upstream 1kb sequence of each gene in the above two plant species. Here a cis-regulatory
sequence is a motif in the PLACE database (Higo et al., 1999), which is represented as consensus sequence such as RTACGTGGCR. It is also worth pointing out that scanning 1kb long sequences with cis-regulatory sequences will result in many false positive cis-regulatory sequence instances, which we filtered below by considering co-occurrence of these instances in a large number of sequences and by assessing the statistical significance of this co-occurrence. Second, we identified individual cis-regulatory sequence combinations composed of multiple cis-regulatory sequences, whose instances co-occurred in upstream 1kb sequences of at least 100 genes in one species. Third, we assessed the statistical significance of these combinations by the Poisson clumping heuristic approximation (Aldous, 1989). Finally, significant cis-regulatory sequence combinations were reported as the final predicted combinations. The upstream 1kb sequences containing instances of a cis-regulatory sequence combination were considered as CRMs of this combination and the genes containing CRMs of a combination were taken as target genes of this combination. See Material and Method Section for details.

In total, we identified 51176 cis-regulatory sequence combinations in 92.2% of Arabidopsis genes. These combinations in Arabidopsis consist of 2 to 6 cis-regulatory sequences, and 3.88 cis-regulatory sequences on average (Figure 2). Similarly, we identified 58031 cis-regulatory sequence combinations in 98.2% of genes in poplar. These poplar combinations consist of 2 to 6 cis-regulatory sequences, and 3.91 cis-regulatory sequences on average (Figure 2). Among all these cis-regulatory sequence combinations, 18638 combinations are shared by Arabidopsis and poplar. Here a shared combination has exactly the same cis-regulatory sequence combination in the two species, but the majority of its target genes in Arabidopsis may be not orthologous to its target genes in poplar. In fact, for a shared cis-regulatory sequence combination, on average, only 33.3% of its target genes in Arabidopsis have defined orthologous genes in poplar in the Inparanoid database (Ostlund et al., 2009). In addition, for a shared combination, among the target genes with orthology information in the two species, no more than 13.6% of its target genes in Arabidopsis are orthologous to its target genes in poplar. These percentages could be higher, since we have no orthology information for the majority of genes in the two species. Because we identified these combinations in the two species separately, the 18638 shared cis-regulatory sequence combinations are likely biologically meaningful. See the supplemental file S1 for all 18638 combinations and their target genes.
**Shared cis-regulatory sequence combinations are supported by known combinations**

To assess the reliability of the above 18638 shared cis-regulatory sequence combinations, we compared these combinations with known cis-regulatory sequence combinations. Many experimental studies have identified co-occurring TFBSs of a pair of interacting TFs or individual TFs with multiple DNA binding domains in plants (Solano et al., 1995; Zhang et al., 1995; Chen et al., 1996; Davies et al., 1996; Singh, 1998; Chen and Singh, 1999; Kagaya et al., 1999; Yanagisawa and Schmidt, 1999; Nagano et al., 2001; Abe et al., 2003; Baudry et al., 2004; Konishi and Yanagisawa, 2010). In Table 1, we list seven well known cis-regulatory sequence combinations. TF or TFBS combinations were considered as cis-regulatory sequence combinations here since each TF or TFBS corresponds to a cis-regulatory sequence. There are other known TF combinations as well. However, cis-regulatory sequence information for known TF combinations is not always available. In case that there are known cis-regulatory sequences for TFs, it is difficult to match these cis-regulatory sequences with the PLACE cis-regulatory sequences we used due to different degrees of mismatches between the cis-regulatory sequences. Therefore, we only used the seven known cis-regulatory sequence combinations in Table 1.

We found that our approach predicted all seven known combinations. For each known combination, there are two or more predicted cis-regulatory sequence combinations that contain this known combination. For instance, for the known bHLH-MYB (MYC-MYB) combination, 32 predicted cis-regulatory sequence combinations contain two motifs from the bHLH and MYB families, respectively. Multiple predicted cis-regulatory sequence combinations are consistent with a known combination because different TFs in the same family may bind different motifs. In addition, additional cis-regulatory sequences together with the same known combinations form different predicted cis-regulatory sequence combinations, which regulate different target genes. For instance, M9269 (composed of three cis-regulatory sequences AAATTAACCAA (LBOXLERBCS), CTCAAGTGA (SORLIP3AT), and TACGTGTC (ABREMOTIFAOSEOM)) and M9270 (composed of the cis-regulatory sequences AAATTAACCAA (LBOXLERBCS), CTCAAGTGA (SORLIP3AT), and TTCCCTGTT (ANAERO5CONSENSUS)) are two predicted combinations that overlap with the known combination bHLH-MYB (composed of the PLACE motifs LBOXLERBCS and SORLIP3AT). M9269 and M9270 regulate 108 and 110
target genes, respectively. However, only 22 target genes are regulated by both combinations (See Supplemental file S2).

Since the predicted cis-regulatory sequence combinations include additional cis-regulatory sequences, we further investigated whether there is functional evidence to support the addition of other cis-regulatory sequences. We found that this is indeed the case for at least several predicted combinations. For instance, two predicted combinations, M9269 and M9274, contain the known bHLH-MYB combination. Besides the two PLACE motifs SORLIP3AT and LBOXLEROBCS that correspond to a bHLH TF and an MYB TF, M9269 and M9274 contain an extra cis-regulatory sequence TACGTGTC (ABREMOTIFAOSOSEM) and CAATNATTG (ATHB5ATCORE), respectively. The bHLH and MYB function as transcriptional activators in abscisic acid (ABA) signaling (Abe et al., 2003). The extra motif ABREMOTIFAOSOSEM in M9269 is also related to ABA signaling (Hobo et al., 1999). Similarly, the TF that binds the extra motif ATHB5ATCORE in M9274 is a homeodomain-leucine zipper protein, ATHB5 (Johannesson et al., 2001), which relates to the ABA signaling as well (Johannesson et al., 2003). Therefore, the function of two extra cis-regulatory sequences is consistent with that of the known combinations, both of which relate to ABA signaling.

Besides predicting known cis-regulatory sequence combinations, we also predicted known target genes as target genes of the corresponding predicted combinations. For 6 out of 7 known combinations in Table 1, we predicted at least one known target gene as a target gene of multiple corresponding predicted combinations. Note that not all known target genes are included in our prediction. For instance, AT1G61720 is a target gene of the known combination bHLH and MYB (Debeaujon et al., 2003), and was not predicted as a target gene of any corresponding predicted combination. The lack of known target genes could be due to the limited number of known motifs in the PLACE database (only 469 motifs) and upstream sequences (1kb) used. It could also be a result of the choice of parameters used in the MOPAT software (Hu et al., 2008). See Table 1 and the supplemental file S2 for the target genes and the details of comparisons of predicted cis-regulatory sequence combinations with known combinations.

**Gene ontology (GO) enrichment analysis supports shared cis-regulatory sequence**
To investigate whether the 18638 shared cis-regulatory sequence combinations make sense, we also performed GO enrichment analysis (Boyle et al., 2004) on the target genes of each of the 18638 cis-regulatory sequence combinations in *Arabidopsis*. The GO enrichment analysis is to assess whether a group of genes significantly shares any function, where a function is described by a standard GO term. That is, we want to check whether the target genes of a cis-regulatory sequence combination significantly share any function, represented by GO terms here, when compared with all annotated genes in *Arabidopsis*. We did GO enrichment analysis for each cis-regulatory combination only in *Arabidopsis* because the *Arabidopsis* genome is much better annotated than the poplar genome.

By performing GO enrichment analysis with the tool GOTermFinder (Boyle et al., 2004), we found that target genes of 58.1% and 77.5% of cis-regulatory sequence combinations in *Arabidopsis* significantly share GO terms, with a false discovery rate (FDR) of 0.05 and 0.1, respectively. Below, we provide two examples of the predicted combinations with literature support of their functions. Researchers with interest in specific genes will find more supported cis-regulatory sequence combinations in the supplemental file S1.

**Example 1.** The cis-regulatory sequence combination M3121 is composed of three PLACE cis-regulatory sequences, TTGCATGACT (SORLREP5AT), ATAAACGT (SORLREP2AT), and TAAAAGTTAAAAAC (BOX1PVCHS15). SORLREP5AT and SORLREP2AT are known to be phyA-repressed motifs (Hudson and Quail, 2003) and phyA is critical for red or far-red light photoreceptor activities (Dehesh et al., 1993). BOX1PVCHS15 is known to be bound by GT-1 (Villain et al., 1996) and GT-1 is also related to red light and far-red light response (Gilmartin and Chua, 1990; Escobar et al., 2004). Thus, this cis-regulatory sequence combination M3121 is important for the red or far-red light photoreceptor activity and all three cis-regulatory sequences in this combination may interact to regulate their 187 target genes. We found the following GO terms significantly shared by the target genes of this predicted combination: GO:0009883 (red or far-red light photoreceptor activity, p-value 8.47E-05) and GO:0008020 (G-protein coupled photoreceptor activity, p-value 1.68E-04). These GO terms of target genes are consistent with the
function of the three cis-regulatory sequences in this combination, which supports the functionality of this cis-regulatory sequence combination.

**Example 2.** The cis-regulatory sequence combination M64 is composed of two cis-regulatory sequences, TGACACGTGGCA (HY5AT) and CCNNNNNNNNNNNNCCACG (UPRMOTIFIIAT). The *Arabidopsis* bZIP TF HY5 binds to the HY5AT motif (Chattopadhyay et al., 1998). HY5 relates to photomorphogenesis, which means light-mediated development (Hardtke et al., 2000). UPRMOTIFIIAT is a cis-acting element of gene CNX1 (Oh et al., 2003). It has been shown that both CNX1 and HY5 mediate the development of plants in the light (Mangeon et al., 2011). Thus, the function of the two cis-regulatory sequences in this combination is consistent and the two cis-regulatory sequences in this combination may interact to regulate their 245 target genes. We found that the following GO terms are significantly shared by the target genes of this combination: GO:0019684 (photosynthesis, light reaction, p-value 2.31E-06), GO:0015979 (photosynthesis, p-value 1.41E-08), and GO:0010207 (photosystem II assembly, p-value 6.36E-05). These functions of target genes are consistent with that of cis-regulatory sequences in this combination, which supports the functionality of this combination.

**Co-expressed gene clusters and known functional gene sets in literature supports predicted cis-regulatory sequence combinations shared between *Arabidopsis* and poplar**

In addition to GO enrichment analysis, we also compared target genes of the 18638 shared cis-regulatory sequence combinations with co-expressed gene clusters in *Arabidopsis*. Since co-expressed target genes are often co-regulated (Allocco et al., 2004), the significant overlap of target genes of a cis-regulatory sequence combination with co-expressed gene clusters supports the functionality of the predicted combinations as well. We found that for 954 of the 18638 shared cis-regulatory sequence combinations, under a FDR of 0.05, their target genes significantly overlap with at least one co-expressed gene cluster from 8 microarray datasets. When we changed the FDR to 0.10, we found that target genes of 14910 combinations significantly overlap with at least one co-expressed gene cluster. The co-expressed gene clusters were listed in the supplemental file S5. These 954 and 14910 cis-regulatory sequence combinations could be obtained by sorting corresponding columns in the supplemental file S1.
In addition, we compared target genes of the 18636 shared cis-regulatory sequence combinations with all gene sets mentioned in the poplar genome paper (Tuskan et al., 2006) and are available in the literature. In total, we obtained six and five gene sets in Arabidopsis and in poplar, respectively (Table 2, row 5 to 15). Each of these eleven gene sets contains genes of a specific function that are important for the production of bioenergy. We found that each of the eleven gene sets enriched the target genes of several cis-regulatory sequence combinations. In total, we obtained 796 shared cis-regulatory sequence combinations that are likely to regulate genes in these functional gene sets, which are important for bioenergy research. For instance, target genes of the cis-regulatory sequence combination M164 significantly overlap with the cell wall formation gene set in Arabidopsis. This combination is composed of cis-regulatory sequences CCCACCTACC (ACIPVPAL2) and TCCATGCA (SPHCOREZMC1). The motif ACIPVPAL2 interacts directly with the PAL2 gene (Phenylalanine ammonia-lyase gene family). The PAL2 gene is involved with phenylpropanoid biosynthesis pathway (Hatton et al., 1995). C1, one of the target genes of the motif SPHCOREZMC1, relates to the cell wall formation process (Suzuki et al., 1997; Yamaguchi et al., 2000). Therefore, the two cis-regulatory sequences corresponding to the two motifs both relate to the cell wall formation process, which conforms to the function of genes in this gene set. The supplemental file S3 provides the details of shared cis-regulatory sequence combinations with target genes significantly overlapping with these eleven gene sets.

**DISCUSSION**

In this study, we identified 18638 cis-regulatory sequence combinations shared by Arabidopsis and poplar. By comparing these shared combinations with known cis-regulatory sequence combinations, analyzing functional similarity of target genes, comparing target genes with co-expressed gene clusters, and overlapping target genes with known functional gene sets, we have shown that various sources of evidence support the functionality of more than 96.8% and 65.2% of the 18638 shared cis-regulatory sequence combinations in Arabidopsis, under a FDR of 0.05 and 0.1, respectively. All 18638 combinations, the sources of functional evidence supporting their functions, and their target genes are in the supplemental files S1, S2, and S3.

We compared the 18638 shared cis-regulatory sequence combinations with seven categories of known combinations. Although all seven categories of known combinations were included in our
predictions, not all known specific TF-TF interactions of each type were predicted. For instance, the TF combination OBP1-OBF (Dof-bZIP) has been reported by several papers (Zhang et al., 1995; Chen et al., 1996; Yanagisawa and Schmidt, 1999). However, we did not find this specific combination in our predicted combinations that contain cis-regulatory sequences corresponding to bZIP and Dof. The missed specific TF-TF combinations could be due to the reason that we removed many cis-regulatory sequence instances and putative cis-regulatory sequence combinations as well to control false positives.

For the 18638 cis-regulatory sequence combinations, on average, about two-thirds of their target genes in the two plant species have no orthologous gene information. The incompleteness of orthologous gene information, together with the limited number of known cis-regulatory sequences and the short non-coding sequences around each gene considered, results in fewer than 13.8% of Arabidopsis target genes with defined poplar orthologs orthologous to the corresponding poplar target genes, for each of the 18638 cis-regulatory sequence combinations. However, we doubt that the percentage of orthologous target genes for most of the shared combinations will be increased significantly with more information, given the long divergence time of the two species. In fact, previous studies have confirmed that substantial changes occur between the target genes of the same TFs (TF combinations) between even closely related species (Borneman et al., 2007; Tuch et al., 2008; Perez and Groisman, 2009).

In each of the 18638 predicted cis-regulatory sequence combinations, there are two or six cis-regulatory sequences. Note that the same sequences may be included in a cis-regulatory sequence combination multiple times, if there are multiple cis-regulatory sequences in the PLACE database corresponding to the same TFs. For instance, two cis-regulatory sequences, GCCACGTGGc (ACGTROOT1) and GCCACGTGGg (ABREAZMRAB28), are likely bound by the same TF. We found instances of these two cis-regulatory sequences co-occur in 221 1kb long upstream sequences. Thus, we predicted a cis-regulatory sequence combination, M103, which consists of the two cis-regulatory sequences. It is worth mentioning that similar cis-regulatory sequences occurring in the same combination is due to their multiple occurrences in 1kb long sequences instead of the overlap of their instances, since overlapping instances were removed from the beginning (See Materials and Methods). Because of filtering overlapping
instances in our method, it is also evident that there could be more TFs in a plant TF combination.

Among the 18638 shared cis-regulatory sequence combinations, 796 combinations are likely to regulate genes that are important in bioenergy research (Table 2). These genes are known to relate to cell wall formation, secondary metabolism, phytohormones, membrane transporter, and disease resistance (Tuskan et al., 2006). We found enrichment with genes related to bioenergy production among the target genes of 796 shared cis-regulatory sequence combinations. We have also shown that cis-regulatory sequences in some combinations are consistent with these functions. These 796 predicted cis-regulatory sequence combinations will be useful for generating hypotheses for future bioenergy research.

Our method is useful for genome-wide studies of gene transcriptional regulation in plants. To our knowledge, previous computational methods for plant CRM identification are often restricted to a small group of genes (Vandepoele et al., 2006) or require TFBS co-occurrence in short regions (≤200 bp) (Steffens et al., 2005; Chang et al., 2008). In addition, these methods often consider combinations of only two cis-regulatory sequences at a time. Thus, from previous studies, it is not clear how prevalent CRMs are in the plant genomes. Compared with previous methods, our method can identify new cis-regulatory sequence combinations from a large number of motifs (such as all motifs in plants), can consider cis-regulatory sequence combinations composed of any number of motifs, and can identify frequently used cis-regulatory sequence combinations from upstream 1kb regions of all genes in one or multiple plant species. Our method is thus a useful addition to tools for plant gene regulation studies.

There are several limitations in our current study. We only considered instances of cis-regulatory sequences in the upstream 1kb of all genes in a species here. It has been shown, however, that TFBSs can be anywhere in the non-coding sequences (Blanchette et al., 2006). In future studies, we should extend our method to include the entire non-coding regions of a plant genome (Cai et al., 2010). In addition, we could consider plant motifs in other databases (Wingender et al., 1996; Lescot et al., 2002; Sandelin et al., 2004) in addition to the PLACE database. This is feasible because the main component of our method, the frequent pattern mining technique, has been used to effectively deal with a much larger number of different items (cis-regulatory sequences).
in transactions (upstream sequences) in the research field of databases (Han et al., 2000; Grahne and Zhu, 2005). Finally, we only examined the shared combinations in two plant species here. With more genomic data such as gene expression data and epigenetic data in Arabidopsis and poplar, we could consider the function of other combinations in individual species as well.

Our study provides a useful method for the study of gene transcriptional regulation in plants. A software package based on the developed method will be freely available at http://www.cs.ucf.edu/~xiaoman/arabpoplar/webserver. This software package can be readily applied to a genome-wide study of any sequenced plant species. Our study also supplies a valuable resource of predicted CRMs and cis-regulatory sequence combinations in Arabidopsis and poplar. Besides the prediction provided in the supplementary files, we also provided a web server at http://www.cs.ucf.edu/~xiaoman/arabpoplar/webserver for researchers to extract TF combinations they are interested in. The predicted CRMs and shared cis-regulatory sequence combinations in two species, despite the support of many sources of evidence, needs to be experimentally validated. We expect that the application of next generation sequencing techniques to plant genomes will greatly accelerate the validation process and facilitate their use in plant and bioenergy research.

MATERIALS AND METHODS

Plant motifs, genes, upstream sequences, and orthology information
A set of 469 plant motifs were obtained from the PLACE database (Higo et al., 1999) (Supplemental file S4). Detailed description about these PLACE motifs can be retrieved at (http://www.dna.affrc.go.jp/PLACE/info.html). These motifs are represented as consensus sequences, in which the most frequent nucleotide(s) at each position of the sequences is given. For instance, the motif for the sequence bound by CBF2 is CCACGTGG. We called these PLACE motifs “cis-regulatory sequences” in this paper.

We obtained all genes and their upstream 1kb sequences relative to the gene transcriptional start sites in Arabidopsis and poplar from the ENSEMBL Plant (http://plants.ensembl.org/index.html), where the TAIR10 release and the JGI2.0 were used for these two species. In total, we obtained
33518 and 41365 genes and their upstream sequences in *Arabidopsis* and poplar, respectively. Repeats in the upstream sequences were then masked by using RepeatMasker (http://www.repeatmasker.org/). We obtained all pairs of orthologous genes in *Arabidopsis* and poplar from the Inparanoid database (Ostlund et al., 2009), from which 10499 pairs of orthologous genes are defined in the two species. These pairs of orthologous genes are composed of 9707 *Arabidopsis* genes and 9941 poplar genes.

**Gene expression data and co-expressed clusters**

We obtained gene expression data on the GPL198 platform (Affymetrix Arabidopsis ATH1 Genome Array) from the GEO database (Edgar et al., 2002). Data from the GPL198 platform was used because this microarray platform has the largest number of samples in *Arabidopsis*. We further filtered data series that have fewer than 10 samples or have not been normalized by the RMA algorithm (Irizarry et al., 2003). We required at least 10 samples in each dataset to ensure more reliable co-expressed gene clusters. The requirement of normalization by RMA will be helpful for others to repeat this study. In this way, we obtained 15 datasets and 292 samples. Because of our interest in development and bioenergy, we further chose time series data and stress resistance related data from the 15 datasets. In total, we obtained 8 datasets and 188 samples. The series number of these datasets and the number of samples is listed in supplemental file S5. For each dataset, we calculated the pair wise Pearson’s correlation coefficient for each gene pair and performed hierarchical clustering with complete linkage to obtain co-expressed clusters, pairs of genes in each of which have a Pearson’s correlation coefficient larger than 0.6 or smaller than -0.6. In total, we obtained 2461 clusters from the eight GEO datasets (Supplemental file S5). In addition to using 0.6 as the correlation coefficient cutoff, we also obtained clusters based on the cutoffs 0.7 and 0.8. The results from different expression correlation coefficient cutoffs were similar. We reported our results based on the cutoff 0.6 in this paper without specification.

**Identification of cis-regulatory sequence combinations in 1kb long upstream sequences**

Previously, we developed an efficient method (Cai et al., 2010) to identify motif combinations in a large number of sequences of 1kb length, by using 522 known vertebrate motifs from the TRANSFAC database (Wingender et al., 1996). We applied this method here to identify cis-
regulatory sequence combinations in upstream 1kb sequences of all genes in Arabidopsis and poplar with slight modifications. The reason to modify the previous method is that the quality of plant motifs in the PLACE database (Higo et al., 1999) is lower than that in the TRANSFAC database, with a much fewer number of experimentally verified TFBSs for each motif and a consensus representation of motifs in PLACE. The consequence of scanning upstream sequences with low quality motifs is that many motifs occur in almost every upstream sequence even requiring exact match of consensus sequences. We thus called these plant motif consensuses “cis-regulatory sequences”. The details of the method follow.

First, we defined a score cutoff for each plant cis-regulatory sequence. To define the score cutoff, we permutated all upstream sequences and then scanned these random sequences in each species separately, with each PLACE cis-regulatory sequence. For each PLACE cis-regulatory sequence, we calculated the score of each segment of length k in the random sequences, where k is the length of the PLACE cis-regulatory sequence. The score is calculated by the following formula:

$$\text{score} = \sum_{i=1}^{k} \log \frac{p(b,i)}{\hat{f}(b)}.$$  

In this formula, $p(b,i)$ is defined as

$$\frac{1.04}{n^{1.0+0.01}} \cdot Z(b,i) + 0.01 \cdot [1 - Z(b,i)]$$

where $n$ is the number of different nucleotides allowed at the $i$-th position of the PLACE cis-regulatory sequence, $Z(b,i)$ is 1 if $b$ is allowed at the $i$-th position and 0 if $b$ is not allowed, and 0.01 is used as pseudo count to regularize the motif (Claverie and Audic, 1996); $\hat{f}(b)$ is the frequency of the nucleotide $b$ in upstream sequences. With the score of each segment, we obtained the score distribution of random segments in each species. We then chose the 85% quartile of this distribution as the score cutoff of the corresponding PLACE cis-regulatory sequence in the corresponding species. That is, if a segment with the score larger than this cutoff, this segment is a potential cis-regulatory sequence instance of this PLACE cis-regulatory sequence. This 85% cutoff corresponds to at most one mismatch for 6-8 bp long PLACE cis-regulatory sequences.

Second, we defined cis-regulatory sequence instances of each PLACE cis-regulatory sequence in each upstream sequence. With the PLACE cis-regulatory sequences and their score cutoffs, we scanned each upstream sequence and obtained all segments with scores larger than the corresponding score cutoffs. These segments are called cis-regulatory sequence instances. Note that some cis-regulatory sequence instances overlap with each other. For those overlapping cis-
regulatory sequences instances with starting positions shorter than 4 bp away from each other, we only kept the one with the largest score, as in the previous study (Hu et al., 2008). We also removed instances of cis-regulatory sequences whose instances occur in more than 50% of upstream sequences, as these cis-regulatory sequences are too degenerate to be identified in statistically meaningful cis-regulatory sequence combinations. The following cis-regulatory sequence combination analysis is based on the remaining cis-regulatory sequence instances.

Third, we identified cis-regulatory sequence combinations that occur in at least 100 upstream sequences. With cis-regulatory sequence instances in each upstream sequence, we applied frequent pattern mining techniques (Han et al., 2000; Grahne and Zhu, 2005) to identify cis-regulatory sequence combinations whose cis-regulatory sequence instances co-occur in at least 100 upstream sequences. The frequent pattern mining techniques were originally developed as a database mining tool to identify frequent combination of patterns in databases (Han et al., 2000). Consider a database containing customer purchase information. With what frequency will a customer who buys item A also buys item B, C, …? The frequent pattern mining techniques can efficiently solve this type of problems. In our case, the 469 PLACE cis-regulatory sequences correspond to items and upstream sequences correspond to customers. The cis-regulatory sequence instances in an upstream sequence are the exact items this sequence contains.

Fourth, we identified significant cis-regulatory sequence combinations by using the Poisson clumping heuristic (Aldous, 1989). For a frequent cis-regulatory sequence combination identified at the third step, say $m_1, m_2, \ldots, m_n$, assume it occurs in $g$ out of $N1$ sequences and $\lambda_i$ is the rate parameter of the Poisson process for the motif $i$. Here $N1$ is the number of all upstream sequences in one species under consideration. Note that we have considered a sequence and its reverse complement sequence when calculating $\lambda_i$. The probability that instances of all cis-regulatory sequences in this combination occur at least one time in one 1kb long upstream sequence is $p = \prod_{i=1}^{n}(1 - e^{-1000\lambda_i})$. Therefore, the p-value that this combination occurs in $g$ sequences is $\sum_{i=g}^{N1} C(N1,i)p^i(1 - p)^{N1-i}$, where $C(N1,i)$ is defined as $\frac{N1(N1-1)\cdots(N1-(i-1)+1)}{i(i-1)\cdots1}$. Since we generated a large number of combinations at the third step, to control the false positive rate by Bonferroni correction, we only selected cis-regulatory sequence combinations with the corresponding p-value smaller than $0.05/C(469,m)$, where $m$ is the largest number of motifs in
the frequent cis-regulatory sequence combinations output at the third step. For each selected cis-regulatory sequence combination, the upstream sequences containing instances of all cis-regulatory sequences in this combination are defined as CRMs of this combination. Similarly, genes corresponding to these upstream sequences are defined as the target genes of this cis-regulatory sequence combination.

**P-value calculation for significant overlap between target genes of a cis-regulatory sequence combination and a gene set of special interest**

For every cis-regulatory sequence combination, we calculated the significance of overlap between its target genes and each co-expressed gene cluster obtained from the 8 microarray expression datasets, by using the hypergeometric test. Assume the set of all genes that are included in the GPL198 platform is S and |S|, the number of genes in S, is equal to N. Given a target gene set S₁ of a cis-regulatory sequence combination and a co-expressed gene cluster S₂, assume the number of genes in the intersection of the three sets S, S₁, and S₂ is |S₁ ∩ S| = n₁ |S₂ ∩ S| = M, and |S₁ ∩ S₂| = m, respectively. Then the significance of the overlap of the target gene set S₁ and the co-expressed gene cluster S₂ is measured by the following p-value based on the hypergeometric test: 

\[ p_{value} = \sum_{i=m}^{\min(n,M)} \frac{C(M,i)C(N-M,n-i)}{C(N,n)} \]

where \( C(x,y) \) is defined above. Likewise, we calculate the p-value of any overlap between target genes of a cis-regulatory sequence combination and a gene set of known function from the literature. With these p-values, we then applied the Q-Value software (Storey and Tibshirani, 2003) to all obtained p-values and found the p-value cutoff, p₁, to control the FDR at the level of 0.05 and 0.1, respectively. We then reported all cis-regulatory sequence combinations with the overlapping p-values smaller than p₁ when compared with a co-expressed cluster or a gene set as the significantly overlapped combinations.

**ACKNOWLEDGEMENTS**

We want to thank the anonymous reviewers for their insightful comments and suggestions, which enable a significant improvement of the paper. This project is supported by a National Human Genome Research Institute grant R01HG004359 (to XL) and a National Science Foundation, Chemical, Bioengineering, Environmental, and Transport Systems grant 1125679 (to HH).
REFERENCES

Abe H, Urao T, Ito T, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. Plant Cell 15: 63-78

Akyildiz M, Gowik U, Engelmann S, Koczor M, Streubel M, Westhoff P (2007) Evolution and function of a cis-regulatory module for mesophyll-specific gene expression in the C4 dicot Flaveria trinervia. Plant Cell 19: 3391-3402

Aldous D (1989) Probability Approximations via the Poisson Clumping Heuristic. Springer-Verlag, New York.

Alkema WB, Johansson O, Lagergren J, Wasserman WW (2004) MSCAN: identification of functional clusters of transcription factor binding sites. Nucleic Acids Res 32: W195-198

Allocco DJ, Kohane IS, Butte AJ (2004) Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics 5: 18

Andrioli LP, Vasisht V, Theodosopoulou E, Oberstein A, Small S (2002) Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms. Development 129: 4931-4940

Bailey TL, Noble WS (1999) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710-3715

Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L (2004) TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in Arabidopsis thaliana. Plant J 39: 366-380

Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, Lefebvre C, Deblois G, Giguere V, Ferretti V, Bergeron D, Coulombe B, Robert F (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16: 656-668

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M (2007) Divergence of transcription factor binding sites across related yeast species. Science 317: 815-819

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710-3715

Cai X, Hou L, Su N, Hu H, Deng M, Li X (2010) Systematic identification of conserved motif modules in the human genome. BMC Genomics 11: 567

Chang WC, Lee TY, Huang HD, Huang HY, Pan RL (2008) PlantPAN: Plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. BMC Genomics 9: 561

Chattopadhyay S, Ang LH, Puente P, Deng XW, Wei N (1998) Arabidopsis bZIP protein HY5 directly interacts with light-responsive promoters in mediating light control of gene expression. Plant Cell 10: 673-683

Chen W, Chao G, Singh KB (1996) The promoter of a H2O2-inducible, Arabidopsis glutathione S-transferase gene contains closely linked OBF- and OBP1-binding sites. Plant J 10: 955-966

Claverie JM, Audic S (1996) The statistical significance of nucleotide position-weight matrix matches. Comput Appl Biosci 12: 431-439

Davidson E (2006) The Regulatory Genome: Gene Regulatory Networks in Development and Evolution., Ed 1st edition. Academic Press., Burlington, MA

Davies B, Egea-Cortines M, de Andrade Silva E, Saedler H, Sommer H (1996) Multiple interactions amongst floral homeotic MADS box proteins. EMBO J 15: 4330-4343

Debeaujon I, Nesi N, Perez P, Devic M, Grandjean O, Caboche M, Lepiniec L (2003) Proanthocyanidin-accumulating cells in Arabidopsis testa: regulation of differentiation and role in seed development. Plant Cell 15: 2514-2531

Dehesh K, Franci C, Parks BM, Seeley KA, Short TW, Tepperman JM, Quail PH (1993) Arabidopsis HYH locus encodes phytochrome A. Plant Cell 5: 1081-1088

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210

Escobar MA, Franklin KA, Svensson AS, Saltar MG, Whiteam GC, Rasmusson AG (2004) Light regulation of the Arabidopsis respiratory chain. Multiple discrete photoreceptor responses contribute to induction of type II NAD(P)H dehydrogenase genes. Plant Physiol 136: 2710-2721
Frith MC, Hansen U, Weng Z (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics 17: 878-889

Frith MC, Li MC, Weng Z (2003) Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res 31: 3666-3668

Frith MC, Spouge JL, Hansen U, Weng Z (2002) Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. Nucleic Acids Res 30: 3214-3224

Gallo SM, Li L, Hu Z, Halfon MS (2006) REDfly: a Regulatory Element Database for Drosophila. Bioinformatics 22: 381-383

Gilmartin PM, Chua NH (1990) Localization of a phytochrome-responsive element within the upstream region of pea rbcS-3A. Mol Cell Biol 10: 5565-5568

Grahne G, Zhu J (2005) Fast algorithms for frequent itemset mining using FP-trees. IEEE transactions on knowledge and data engineering 17: 1347-1362

Gupta M, Liu JS (2005) De novo cis-regulatory module elicitation for eukaryotic genomes. Proc Natl Acad Sci U S A 102: 7079-7084

Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. In ACM SIGMOD International Conference on Management of Data, Dallas, USA, ACM SIGMOD

Hardtke CS, Gohda K, Osterlund MT, Oyama T, Okada K, Deng XW (2000) HY5 stability and activity in arabidopsis is regulated by phosphorylation in its COP1 binding domain. EMBO J 19: 4997-5006

Hatton D, Sablowski R, Yung MH, Smith C, Schuch W, Bevan M (1995) Two classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. Plant J 7: 859-876

Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27: 297-300

Hobo T, Asada M, Kowyma Y, Hattori T (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. Plant J 19: 679-689

Howard ML, Davidson EH (2004) cis-Regulatory control circuits in development. Dev Biol 271: 109-118

Hu J, Hu H, Li X (2008) MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs. Nucleic Acids Res 36: 4488-4497

Hudson ME, Quail PH (2003) Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. Plant Physiol 133: 1605-1616

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264

Johannesson H, Wang Y, Engstrom P (2001) DNA-binding and dimerization preferences of Arabidopsis homeodomain-leucine zipper transcription factors in vitro. Plant Mol Biol 45: 63-73

Johannesson H, Wang Y, Hanson J, Engstrom P (2003) The Arabidopsis thaliana homeobox gene ATHB5 is a potential regulator of abscisic acid responsiveness in developing seedlings. Plant Mol Biol 51: 719-729

Johansson O, Alkema W, Wasserman WW, Lagergren J (2003) Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. Bioinformatics 19 Suppl 1: i169-176

Kagaya Y, Ohmiya K, Hattori T (1999) RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. Nucleic Acids Res 27: 470-478

Kawashima T, Wang X, Henry KF, Bi Y, Weterings K, Goldberg RB (2009) Identification of cis-regulatory sequences that activate transcription in the suspensor of plant embryos. Proc Natl Acad Sci U S A 106: 3627-3632

Kel-Margoulis OV, Romashchenko AG, Kolchanov NA, Wingender E, Kel AE (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. Nucleic Acids Res Suppl 1: 1169-176

Kim YJ, Lee SH, Park KY (2004) A leader intron and 115-bp promoter region necessary for expression of the carnation S-adenosylmethionine decarboxylase gene in the pollen of transgenic tobacco. FEBS Lett 578: 229-235

Konishi M, Yanagisawa S (2010) Identification of a nitrate-responsive cis-element in the Arabidopsis NIR1 promoter defines the presence of multiple cis-regulatory elements for nitrogen response. Plant J 63: 269-282

Lescot M, Dehais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouze P, Rombaux S (2002) PlantCARE, a
Villain P, Mache R, Zhou DX (1996) The mechanism of GT element-mediated cell type-specific transcriptional control. J Biol Chem 271: 32593-32598

Loots GG, Locksley RM, Blankespoor CM, Wang ZE, Miller W, Rubin EM, Frazer KA (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science 288: 136-140

Mangeon A, Bell EM, Lin WC, Jablonska B, Springer PS (2011) Misregulation of the LOB domain gene DDA1 suggests possible functions in auxin signalling and photomorphogenesis. J Exp Bot 62: 221-233

Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM, Givan SA, Yanovsky M, Hong F, Kay SA, Chory J (2008) Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet 4: e14

Nagano Y, Inaba T, Furuhashi H, Sasaki Y (2003) Conservation between animals and plants of the cis-acting element involved in the unfolded protein response. Biochem Biophys Res Commun 301: 225-230

Ostlund G, Schmitt T, Forslund K, Messina DN, Roopra S, Frings O, Sonnhammer EL (2009) Evolution of transcriptional regulatory circuits in bacteria. Cell 138: 233-244

Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32: D91-94

Singh KB (1998) Transcriptional regulation in plants: the importance of combinatorial control. Plant Physiol 118: 1111-1120

Sinha S, van Nimwegen E, Siggia ED (2003) A probabilistic method to detect regulatory modules. Bioinformatics 19 Suppl 1: i292-301

Solano R, Nieto C, Avila J, Canas L, Diaz I, Paz-Ares J (1995) Dual DNA binding specificity of a petal-epidermis-specific MYB transcription factor (MYB.Ph3) from Petunia hybrida. EMBO J 14: 1773-1784

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-9445

Suzuki M, Kao CY, McCarty DR (1997) The conserved B3 domain of VIVIPAROUS1 has a cooperative DNA binding activity. Plant Cell 9: 799-807

Tuch BB, Li H, Johnson AD (2008) Evolution of eukaryotic transcription circuits. Science 319: 1797-1799

Vandepoele K, Casneuf T, Van de Peer Y (2006) Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. Genome Biol 7: R103

Villain P, Mache R, Zhou DX (1996) The mechanism of GT element-mediated cell type-specific transcriptional control. J Biol Chem 271: 32593-32598

Visel A, Minovitsky S, Dubchak I, Pennacchio LA (2007) VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res 35: D88-92

Wang R, Guan P, Chen M, Xing X, Zhang Y, Crawford NM (2010) Multiple regulatory elements in the Arabidopsis NIA1 promoter act synergistically to form a nitrate enhancer. Plant Physiol 154: 423-432
Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24: 238-241
Yamaguchi Y, Nakamura T, Kusano T, Sano H (2000) Three Arabidopsis genes encoding proteins with differential activities for cysteine synthase and beta-cyanoalanine synthase. Plant Cell Physiol 41: 465-476
Yanagisawa S, Schmidt RJ (1999) Diversity and similarity among recognition sequences of Dof transcription factors. Plant J 17: 209-214
Yuh CH, Bolouri H, Davidson EH (1998) Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279: 1896-1902
Zhang B, Chen W, Foley RC, Buttner M, Singh KB (1995) Interactions between distinct types of DNA binding proteins enhance binding to ocs element promoter sequences. Plant Cell 7: 2241-2252
Zhou Q, Wong WH (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A 101: 12114-12119
FIGURE LEGENDS

Figure 1. The procedure of predicting shared cis-regulatory sequence combinations by *Arabidopsis* and poplar.

Figure 2. The number of motifs and target genes for a cis-regulatory sequence combination. (A). The distribution of the number of motifs in a TF combination. Each group of three bars above a number from left to right is for shared cis-regulatory sequence combinations, combinations in *Arabidopsis*, and combinations in poplar, respectively. (B). The distribution of the number of target genes in a cis-regulatory sequence combination. Each group of three bars above a interval from left to right is for shared cis-regulatory sequence combinations, combinations in *Arabidopsis*, and combinations in poplar, respectively. (C) Shared cis-regulatory sequence combinations between *Arabidopsis* and poplar.
Table 1. Comparison with seven well known combinations.

| Known TF combinations | Recognition sequences | # predicted combinations | Inclusion of known target genes |
|------------------------|------------------------|--------------------------|---------------------------------|
| bHLH+MYB               | bHLH-CANNTG, MYB-A/TAACCA or C/TAACG/TG | 32 | AT5G25610 is found to be the target gene of 26 predicted combinations |
| bZIP+Dof               | bZIP-ACGT,Dof-AAAG      | 656 | AT1G78380 is found to be the target gene of 38 predicted combinations |
| MADS+MADS              | MADS-CArG box           | 2  | AT3G54340 is found to be the target gene of 2 predicted combinations |
| bZIP+bZIP              | bZIP-ACGT,bZIP-ACGT     | 429 | AT2G47730 is found to be the target gene of 38 predicted combinations |
| bZIP+MYB               | Bzip-ACGT, MYB-CANNTG   | 707 | AT5G13930 is found to be the target gene of 22 predicted combinations |
| RAV1                   | RAV1-CAACA,RAV1-CACCTG  | 2  | AT2G26710 is found to be the target gene of 2 predicted combinations |
| ZmHOX2a                | ZmHox2a HD1-TCCT,ZmHox2a HD2-GATC | 5  | No known target gene information available in *Arabidopsis* |

The predicted combinations mentioned in the third column contain the known combinations mentioned in the first two columns.
### Table 2. Functional evidence for the 18638 shared cis-regulatory sequence combinations.

| Functional Evidence source(ID) | #Significant combinations |
|-------------------------------|---------------------------|
|                               | FDR = 0.05          | FDR=0.1         |
| Known TF combinations [0]     | 1603 (8.6%)          |                 |
| Gene ontology(1)              | 10840 (58.1%)        | 14455 (77.6%)  |
| Gene expression data from the GEO database(2) | 954 (5.12%) | 14910 (79.5%) |
| Cellulose synthesis Gene Set of Arabidopsis (Wall formation)(3) | 73 (0.392%) | 487 (2.62%) |
| Cellulose synthesis Gene Set of Populus (Wall formation)(4) | 35 (0.188%) | 321 (1.72%) |
| Phenylpropanoid and lignin synthesis Gene Set of Arabidopsis (Wall formation)(5) | 128 (0.687%) | 622 (3.34%) |
| Phenylpropanoid and lignin synthesis Gene Set of Populus (Wall formation)(6) | 63 (0.338%) | 102 (0.547%) |
| Terpenoid synthesis gene set of Arabidopsis (Secondary metabolism)(7) | 35 (0.188%) | 89 (0.478%) |
| Terpenoid synthesis gene set of Populus (Secondary metabolism)(8) | 13 (0.0697%) | 26 (0.139%) |
| Aux/IAA gene set of Arabidopsis (Phytohormones)(9) | 135 (0.724%) | 622 (3.34%) |
| Aux/IAA gene set of Populus (Phytohormones)(10) | 53 (0.284%) | 351 (1.88%) |
| Voltage gated ion channel gene set of Arabidopsis (Membrane transporter)(11) | 172 (0.922%) | 518 (2.78%) |
| TNL gene set of Arabidopsis (Disease resistance)(12) | 81 (0.435%) | 402 (2.15%) |
| TNL gene set of Populus (Disease resistance)(13) | 26 (0.139%) | 129 (0.692%) |
| **Total**                     | **12149 (65.2%)**    | **18047 (96.8%)** |
Define a score cutoff for each plant cis-regulatory sequence in each species

Use these cis-regulatory sequences to scan each 1kb upstream sequence

Segments with scores larger than corresponding cis-regulatory sequence score cutoff

Yes

Cis-regulatory sequence instances

Instances for cis-regulatory sequence occur in more than 50% of upstream sequences

Yes

Cis-regulatory sequence was discarded

NO

NO

Significant cis-regulatory sequence combinations

Use Poisson clumping heuristic to identify significant cis-regulatory sequence combinations

Cis-regulatory sequence combinations

Use frequent pattern mining techniques to identify cis-regulatory sequence combinations whose instances occur together in at least 100 upstream sequences

Remaining cis-regulatory sequence instances

Instances for cis-regulatory sequence with starting positions shorter than 4bp away from each other

NO

Yes

Cis-regulatory sequence instances with smaller score were discarded
