Plant Physiology Preview. Published on April 8, 2009, as DOI:10.1104/pp.109.136028

Running title: Cis-elements and coexpression networks in Arabidopsis

Mailing address: Yves Van de Peer
Department of Plant Systems Biology, VIB2-Universiteit Gent
Technologiepark 927, B-9052 Gent (Belgium)

Tel. 32-9-3313807; fax 32-9-3313809; e-mail yves.vandepeer@psb.ugent.be

Journal research area: Bioinformatics

Keywords: transcriptional regulation, coexpression, cis-regulatory element, Arabidopsis, cell cycle, E2F, OBP1

Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks [1][W]

Klaas Vandepoele, Mauricio Quimbaya, Tine Casneuf, Lieven De Veylder and Yves Van de Peer*

Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Ghent, Belgium. Department of Molecular Genetics, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium.

Footnotes

[1] K.V. and L.D.V. are postdoctoral fellows of the Fund for Scientific Research, Flanders (FWO). This work was supported by the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

[2] T.C. current address: Ortho Biotech Oncology Research & Development, a division of Janssen Pharmaceutica, Turnhoutseweg 30, B-2340 Beerse, Belgium

* Corresponding author; e-mail yves.vandepeer@psb.ugent.be; fax +32 9 33 13809.

[W] The online version of this article contains Web-only data.

Abstract

Analysis of gene expression data generated by high-throughput microarray transcript profiling experiments has demonstrated that genes with an overall similar expression pattern are often enriched for similar functions. This guilt-by-association principle can be applied to define modular gene programs, identify cis-regulatory elements or predict gene functions for unknown genes based on their coexpression neighborhood. We evaluated the potential to use Gene Ontology (GO) enrichment of a gene's coexpression neighborhood as a tool to predicts its function but found overall low sensitivity scores (13-34%). This indicates that for many functional categories coexpression alone performs poorly to infer known biological gene functions. However, integration of cis-regulatory elements shows that 46% of the gene coexpression neighborhoods are enriched for one or more motif, providing a valuable complementary source to functionally annotate genes. Through the integration of coexpression data, GO annotations and a set of known cis-regulatory elements combined with a novel set of evolutionary conserved plant motifs, we could link many genes and motifs to specific biological functions. Application of our coexpression framework extended with cis-regulatory element analysis on transcriptome data from the cell cycle-related transcription factor *OBP1* yielded several coexpressed modules associated with specific cis-regulatory elements. Moreover, our analysis strongly suggests a feed forward regulatory interaction between *OBP1* and the E2F pathway. The ATCOECIS resource (http://bioinformatics.psb.ugent.be/ATCOECIS/) makes it possible to query coexpression data, GO and cis-regulatory elements annotations, submit user-defined gene sets for motif analysis and provides an access point to unravel the regulatory code underlying transcriptional control in Arabidopsis.

Introduction

The rapid accumulation of genome-wide data describing both genome sequences and functional properties of genes facilitates the development of systems biology approaches. Especially the application of microarray experiments for several model organisms now provides us with detailed catalogues of condition-dependent transcriptional activity during development, in different organs, cell types or in response to various endo- or exogenous stimuli (Birnbaum et al., 2003; Schmid et al., 2005). In plants, transcriptional regulation is mediated by a large number (>1500) of transcription factors (TFs) controlling the expression of tens or hundreds of target genes in various, sometimes intertwined, signal transduction cascades (Wellmer and Riechmann, 2005). Whereas the similarity in gene expression patterns can be used to infer modular gene programs (~regulatory networks), the integration of expression and sequence data makes it possible to identify cis-regulatory elements, the functional elements responsible for the timing and location of transcriptional activity (Haberer et al., 2006; Ma et al., 2007). Transcription factor binding sites (or DNA sequence motifs, shortly motifs) are the functional elements that determine the timing and location of transcriptional activity. Complementary, the identification of differentially expressed genes in response to a treatment/stimulus or in a transgenic over-expression/knock-out experiment can identify new target genes and provide insights into the underlying regulatory interactions (Vandepoele et al., 2005; Zhang et al., 2005).

Systematic computational analysis of DNA motifs illustrated the presence of TATA-boxes as well as Y patches characterizing a large fraction of plant core promoters (Yamamoto et al., 2007). Other motifs have been described showing strong position- and/or strand-dependent localization and a subset of these correspond to known cis-regulatory elements (Molina and Grotewold, 2005; Obayashi et al., 2007; Yamamoto et al., 2007). Through the combination of motif mapping data on Arabidopsis promoters with gene expression patterns, Walther and co-workers found a positive correlation between multi-stimuli response genes and cis-element density in upstream regions (Walther et al., 2007). Studies focusing on the combinatorial nature of transcriptional control have identified several examples of cooperative elements (or cis-regulatory modules) driving time-of-day-specific expression patterns or regulating genes involved in processes like photosynthesis or protein biosynthesis (Vandepoele et al., 2006; Michael et al., 2008). Interestingly, evolutionary analysis suggests that these regulatory modules are conserved between species belonging to different plant families (Kim et al., 2006).

The exploitation of the idea that correlated expression implies a biological relevant relationship resulted in the development of several meta-analysis tools that infer Arabidopsis gene functions using a guilt-by-association principle, such as ACT (Jen et al., 2006), ATTED-

II (Obayashi et al., 2007) and CressExpress (Srinivasasainagendra et al., 2008). In general, these methods determine, for a gene of interest, a set of coexpressed genes while significant functional annotations in the gene's coexpression neighborhood are used to draw new biological hypotheses. The Gene Ontology (GO) or AraCyc functional annotation systems in combination with a statistical test are mostly used to determine functional enrichment. While generally coexpression networks cover all correlated expression patterns between genes within an expression compendium, detailed analysis of the topology or node-to-node relationships within the network provides an overview of the organization and complexity of transcriptional regulation. For example, Persson et al. nicely illustrated the existence of several coexpression clusters corresponding to functional modules involved in primary and secondary cell wall formation (Persson et al., 2005). Similarly, Ma et al. identified several highly connected subclusters in an Arabidopsis gene network grouping genes related to biochemical pathways and cold stress (Ma et al., 2007). Apart from studying gene coexpression networks within one organism, the comparison of expression data between different species using orthologous genes makes it possible to identify evolutionary conserved regulatory programs as well as species-specific adaptations in response to changes in lifestyle or environmental conditions (Stuart et al., 2003).

Although these examples demonstrate the potential of coexpression-based meta-analysis, our current understanding of the relationship between regulatory elements and the observed expression states in different developmental stages, tissues or treatments remains limited. The main objectives of this study were i) to analyze the properties and the functional predictive power of coexpression networks in Arabidopsis ii) to extend coexpression frameworks with information about cis-regulatory elements to functionally annotate genes iii) to apply Gene Ontology and motif enrichment analysis to dissect cell cycle regulatory control using publicly available transcriptome data and vi) to study the organization of cis-regulatory elements in Arabidopsis promoters.

Results

The biological significance of expression similarity

Starting from a set of 322 Affymetrix ATH1 microarray slides retrieved from various publicly available sources, data normalization and averaging of replicates resulted in a non-redundant expression data set of 129 experiments (see Material and Methods). Using a custom-made Chip Description File (CDF) grouping only non cross-hybridizing probes in probesets (Casneuf et al., 2007), the expression patterns of 19,937 genes could be monitored. Although it does not cover all annotated protein-coding genes in Arabidopsis, the CDF file has the advantage that it can reliably measure and discriminate between the expression of both copies of duplicated gene pairs with valid probesets (overcoming potential cross-hybridization caused by high sequence similarity).

In order to verify the guilt-by-association relationship between expression similarity and similarity in gene function, for pre-defined functional sets of genes grouped in GO categories, we first quantified their level of expression similarity using the expression coherence (EC). EC is a measure for the amount of expression similarity within a set of genes, ranging between zero and one and is high for sets of genes that converge into one or a few tight coexpression clusters (Pilpel et al., 2001). As shown in figure 1A, for many GO categories the EC is higher than expected by chance. For Biological Process and Cellular Component approximately 41% and 74% of all categories have EC values higher than expected by chance, respectively, whereas for Molecular Function 36% of the GO categories show elevated coexpression levels. Also for genes annotated in biochemical pathways through AraCyc 33% of all categories show EC values higher than random (Figure 1B). The highest EC values for GO Biological Process cover categories involved in photosynthesis (EC=0.60, 124 genes), porphyrin biosynthesis (EC=0.35, 45 genes), ribosome biogenesis and assembly (EC= 0.44, 114 genes), tetraterpenoid biosynthesis (EC=0.31, 21 genes) and starch metabolism (EC=0.19, 27 genes). For the AraCyc pathways, the categories ‘glucosinolate biosynthesis from tryptophan’, ‘photosynthesis, light reaction’, ‘carotenoid biosynthesis’ and ‘urea cycle’ all have EC values of more than 50%. Nevertheless, since most functional categories have only low EC values (see Supplemental table 1), these results indicate that genes within a functional category do not completely correspond to transcriptional modules and suggest that several coexpression sub-groups might exist for genes annotated in the same functional category. Therefore, an unsupervised approach based on clustering of expression profiles should offer a better strategy to identify transcriptional modules compared to predefined functional categories. Since the AraCyc pathways only group a small number of genes (per pathway) and many pathways are also

present in the GO annotation, we only retained GO categories (containing 25 or more genes) for further analysis.

Construction of Arabidopsis coexpression networks

Before building a gene coexpression network, expression similarities between gene pairs were calculated using the Pearson correlation coefficient (PCC). To determine valid coexpressed genes we applied three similarity thresholds (PCC bigger than 0.63, 0.72 and 0.83) corresponding to the 90^{th}, 95^{th} and 99^{th} percentile of a background PCC distribution containing nearly half a million gene pairs sampled from 1,000 randomly selected genes. Subsequently, all gene-gene coexpression relationships with PCC values above a selected threshold were grouped resulting in three networks (hereafter referred to as ATH90, ATH95 and ATH99) for which the estimated amount of false-positive coexpressed gene pairs is 10%, 5% and 1%, respectively. These networks can be represented as undirected graphs where genes (~nodes) are connected by edges if they are coexpressed. Our approach to initially build multiple networks with different expression similarity constrains is motivated by the fact that it is difficult to *a priori* define an optimal threshold to capture biological knowledge from the network. Therefore, we first performed an evaluation experiment to estimate the biological knowledge captured in the three networks delineated used different similarity thresholds.

We applied a guide gene clustering method to group genes with similar expression patterns followed by gene set enrichment analysis (Wolfe et al., 2005; Aoki et al., 2007). Guide gene clusters group for each query gene all coexpressed genes resulting, on a genome-wide scale, in potentially overlapping clusters. Gene Ontology enrichment analysis was then applied to functionally annotate coexpression clusters (Figure 2) and to assess the predictive power of the three networks to recover known functional annotations. For each gene belonging to a GO category *i*, we determined if the functional GO enrichment in its coexpression guide cluster (~neighborhood) could predict the correct function. Likewise, using sets of randomly selected genes not annotated with GO category *i*, we estimated the number of false positive predictions. The assessment of the prediction power using this approach aims to estimate the optimal size of a gene's coexpression neighborhood to retrieve relevant GO enrichments and to associate unknown genes to specific biological processes. Although GO function predictions for some negative genes might correspond to false negatives (i.e. a correctly predicted functional gene association not yet described in the current GO annotations), application of an iterative random sampling procedure makes it possible to compare the false positive rates between different GO categories and for different similarity thresholds (see Material and Methods). Based on a subset of 50 different GO categories (18 Biological Process, 16 Molecular Function and 16 Cellular Components

categories; covering in total 11,838 genes) we observe that the positive predictive value (PPV or precision rate), referring to the proportion of genes with a functional prediction being correctly predicted, is highest for the ATH90 and the ATH95 network (0.93 and 0.92, respectively; Table1 and Figure 3). Complementary to the PPV, the sensitivity (SN or recall) measures the proportion of actual positives (i.e. known functional annotations) that are correctly identified as such. Although for GO Biological Process the ATH90 and ATH95 networks again have the highest average sensitivity (Figure 3), their actual values (SN=0.39, see Table 1) indicate that many known biological annotations cannot be inferred from a gene's coexpression neighborhood. Assessing the functional predictive power of a recently published Arabidopsis gene network (Ma et al., 2007) based on a graphical Gaussian model (called ATHGGM in Table 1) reveals that the limited number of genes in this network together with the sparse nature of this network – the median number of coexpressed genes is 4 – is responsible for low PPV and sensitivity values. Since the ATHGGM network aims to discover regulatory interactions, the low prediction scores are not surprisingly and suggest that it captures complementary information compared to coexpression networks. Although for some Arabidopsis coexpression platforms like ATTED-II genome-wide data about coexpressed genes is available (Obayashi et al., 2007), the absence of a predefined coexpression neighborhood for each gene makes it practically impossible to systematically evaluate and compare the predictive power of other meta-analysis tools.

Comparing the sensitivity scores over the different ontologies in this benchmark experiment shows that, when requiring that at least half of the known annotations are recovered, approximately 22%, 25% and 56% of the GO Biological Process, Molecular function and Cellular Component categories are retained, respectively (see series ATH95 in Figure 3A-C). Whereas the PPV and SN values both confirm that the ATH90 and ATH95 coexpression networks are better able to infer gene functions than the ATH99 network, we selected the ATH95 expression similarity threshold for further analysis since the functional enrichment folds are higher than those from the ATH90 network (median enrichment fold of 3.97 and 3.24, respectively).

Properties of the Arabidopsis coexpression network

After reconstructing the ATH95 network using all 19,937 measurable genes present on the ATH1 microarray, the final network covers 19,064 genes with a median number of 548 coexpressed genes. When comparing the average connectivity per GO category, we observed that genes involved in processes like rRNA metabolism, histone modification, amino acid activation, photosynthesis or DNA repair all have >1500 coexpression partners (top 5% of connectivity distribution). In contrast, several categories involved in 'response to'

(response to water, reactive oxygen species, brassinosteroid stimulus, extracellular stimulus, high light intensity and hyperosmotic salinity response) show low average connectivity values (lowest 10% of distribution; Supplemental table 2). This finding does not indicate that response-to genes are less coexpressed than the general house-keeping functions described above. Rather, it suggests that the latter form large coexpression modules compared to the stress-related coexpression modules.

Whereas the average sensitivity (covering all GO categories) of functional predictive power for the complete coexpression network is low (SN = 0.19 with average PPV of 0.92), several examples of GO categories with good sensitivity scores can be found (Supplemental table 3). These include photosynthesis (0.80), ribosome biogenesis and assembly (0.70), tRNA metabolism (0.64), starch metabolism (0.59) and amino acid activation (0.58). In contrast, very general GO categories receive low PPV scores due to the large number of putative false positive predictions (e.g. PPV Biological Process term metabolism = 0.035, PPV Molecular Function term catalytic activity = 0.16). Although comparing average connectivity with sensitivity per GO category suggests that primarily genes with large coexpression neighborhoods yield good prediction sensitivity, plotting both variables against each other (Supplemental figure 1) reveals that also many small coexpression neighborhoods provide good predictive power. Examples of GO categories with small coexpression neighborhoods but high prediction sensitivity include response to hydrogen peroxide, starch metabolism and response to high light intensity (average connectivity < 600 and SN ≥ 0.38). Examples of GO categories with low sensitivity scores but high connectivity (e.g. protein ubiquitination, meiosis, protein targeting and posttranscriptional gene silencing) suggest that the primary regulation of these genes is not at the transcriptional level, explaining the bad prediction scores.

Identification of cis-regulatory elements and integration with coexpression clusters

Complementary to functional enrichment using Gene Ontology, we also mapped putative cis-regulatory elements on all genes and calculated motif enrichment for the different gene coexpression clusters. Whereas known plant cis-regulatory elements were retrieved from PLACE (Higo et al., 1999) and AGRIS (Palaniswamy et al., 2006), a complementary set of elements was identified using the network-level conservation principle which applies a systems-level constraint (Elemento and Tavazoie, 2005). Briefly, this method exploits the well-established notion that each TF regulates the expression of many genes in the genome and that the conservation of global gene expression between two related species requires that most of these targets maintain their regulation. In practice, this assumption is tested for each candidate motif by determining its presence in the upstream regions of two related

species and by calculating the significance of conservation over orthologous genes (see Materials and methods). Whereas the same principle of evolutionary conservation is also applied in phylogenetic footprinting methods to identify transcription factor binding sites, it is important to note that here the conservation of several targets in the regulatory network is evaluated simultaneously and that aligned non-coding DNA sequences are not required. This is in contrast with standard footprinting approaches, which only use sequence conservation in upstream regions on a gene-by-gene basis to detect functional DNA motifs. Using motif conservation over orthologous genes between Arabidopsis and poplar (*Populus trichocarpa*) 866 non-redundant 8-mer motifs with significant Network-level Conservation Scores (NCS; p-value<0.05) were identified. Comparing these NCS-motifs with the known cis-regulatory elements from PLACE and AGRIS revealed that 63% (544/866) match described elements. Reversely, 24% of the known motifs show significant evolutionary conservation when applying the network-level conservation principle, suggesting that some of these motifs might be too stringently defined to show cross-species conservation or represent species-specific regulatory elements. Plotting the NCS values for the remaining 322 NCS-motifs not matching known motifs indicates that they have similar conservation scores (inter-quantile values 13.91-15.21-17.63) compared to the known motifs (inter-quantile values 13.81-15.33-17.99). This indicates that both sets of motifs (i.e. NCS-motifs matching and not matching known motifs) are equally well conserved between Arabidopsis and poplar at a genome-wide level and that the new motifs can be considered as putative cis-regulatory elements.

Although the network-level conservation method provides an elegant way to uncover candidate cis-regulatory elements, identifying individual biological functional motif instances on promoter sequences remains problematic. Especially the short and sometimes degenerate nature of these 8-mers (or transcription factor binding sites in general) yields a large fraction of false-positive motif matches. Therefore, for NCS-motifs we only considered Arabidopsis instances showing evolutionary conservation in one or more orthologous poplar promoters. This filtering step yields overall higher enrichment values when validating motif instances using GO (Table 2). In contrast, for known experimentally defined plant motifs from PLACE and AGRIS all motif instances on Arabidopsis promoters were retained for further analysis. Although these databases sometimes report highly similar motifs that might be considered as redundant entries, we observed that in several cases motif variants, when performing genome-wide mappings, yielded sets of target genes showing different GO enrichment. For example, when considering the Gbox related motifs CACGTG, *A*CACGTG, CACGTG*TA* and CACGTG*GC*, we observed that the first two show GO enrichment to response to cold, the last motif variant towards photosynthesis and starch metabolism, and that the third motif with TA suffix does not show any significant enrichment to any of these GO terms. Complementary, of these four motifs only the second *A*CACGTG motif shows

enrichment towards response to abscisic acid (ABA) stimulus (p-value < 0.017) although the more degenerate ACGTGKC PLACE motif shows a stronger association with ABA responsive genes (p-value < 1.1e-04). Since these examples confirm the biological relevance of motif variants (Geisler et al., 2006), for all PLACE and AGRIS elements motif variants were maintained.

Performing motif enrichment using the complete ATH95 network reveals that 46% of the genes have one or more significant motif in their coexpression neighborhood. In total 762 of the 866 NCS-motifs (or 88%) and 249 of the known 721 motifs (35%) were found to be enriched. An overview of the ten most frequently enriched NCS-motifs together with their biological role determined using GO enrichment is shown in Table 2. All ten motifs correspond with well-described plant cis-regulatory elements. Examples of frequent motifs include the TELO and UP1 motif driving the expression of ribosomal genes, the Ibox and Gbox present in genes involved in photosynthesis and stress reponse, the ABA responsive element (ABRE), the E2F motif regulating DNA replication genes and the MSA element responsible for M-phase specificity during the cell cycle. For each motif the full set of putative target genes, including GO enrichments, can be found online (http://bioinformatics.psb.ugent.be/ATCOECIS/).

Dissecting the cell cycle regulatory network using E2Fa and OBP1 target genes

To test the applicability of our approach to unravel biological coexpression networks and infer regulatory logic, we used data from a detailed transcription factor overexpression experiment studying cell cycle control in Arabidopsis. Based on transcriptome analysis of *OBP1* overexpression lines, Skirycz and colleagues recently identified that this DOF transcription factor is involved in cell cycle initiation (Skirycz et al., 2008). To identify cis-regulatory elements and predict new regulatory interactions, we combined expression data reporting oscillating transcripts in synchronized Arabidopsis cell suspensions (Menges et al., 2003) with clustering, GO and motif enrichment analysis. For the 632 genes upregulated by *OBP1*, a significant enrichment of the corresponding cis-regulatory element TAAAG is observed (Table 3). Partitioning the genes using phase expression during cell division reveals that for the DOF upregulated genes showing periodic expression 69% peaks at M phase. This expression pattern is clearly reflected in the motif analysis with the MSA element (M-specific activator) being 11-fold enriched (GACCGTTN; p-value<6.64e-30).

The genes repressed by *OBP1* show GO enrichment for cell wall modification and response to biotic stimulus. To study the underlying regulatory control, we applied the CAST clustering algorithm (Ben-Dor et al., 1999) on our full expression matrix and analyzed these coexpression clusters containing 5 or more DOF downregulated genes. Advantages of CAST

clustering over more classical algorithms such as hierarchical or K-means clustering are that only two parameters have to be specified (the affinity measure, here defined as PCC ≥ 0.72 and the minimal number of genes within a cluster, here set to 5) and that it independently determines the total number of clusters and whether a gene belongs to a cluster. In addition, only genes are grouped in a cluster if they all show a minimal expression similarity with all other genes present in that cluster, yielding global non-overlapping gene clusters with homogeneous expression patterns. The largest cluster covers 164 of the 842 downregulated genes and is strongly enriched for photosynthesis and the Ibox (CTTATCCN). Additionally, five smaller clusters were found all showing stress or defense response, of which two also showed motif enrichment. The first cluster contains 25 genes showing strong shoot osmotic-stress response in the expression data and is enriched for ANCATGTG (MYCATRD22), a dehydration responsive element. The second cluster contains 11 genes mainly showing expression in leaf, is enriched for GO category 'systemic acquired resistance' and shows motif enrichment for the ACGTCATAGA motif (LS7ATPR1), a salicylic acid-inducible element involved in systemic inducible plant defense responses (Despres et al., 2000). Whereas the downregulation of several stress-responsive regulons coincides with the negative link between stress and cell proliferation, the downregulation of the photosynthetic machinery is in agreement with the lack of Rubisco expression in meristems (Fleming et al., 1996).

The observation that 38 DOF upregulated genes peak during S phase are enriched for the E2F motif (5-fold for GCGGGAAN; p-value<9.97e-06) suggested a link between *OBP1* and *E2F*, a well-studied regulator controlling the activation of genes required for cell cycle progression and DNA replication (Vandepoele et al., 2005). Therefore, we compared these DOF target genes and a set of putative E2F target genes that were also identified through microarray analysis on *E2Fa/DPa* overexpressing plants (Vandepoele et al., 2005). Comparing the upregulated genes from the E2Fa and *OBP1* experiments revealed that a significant number of 65 genes are shared between both overexpression lines (Table 3, data set DOF/E2F_UP). Although this set of genes does not show enrichment for the TAAAG DOF motif, 74% of these genes have a WTTSSCSS E2F binding site in their promoter. Together with the observation that the transcription factor E2Fa is upregulated by *OBP1*, these results suggest a feed-forward mechanism between both regulators. Our hypothesis that the *E2Fa* transcription factor is a downstream OBP1 target is in agreement with the observation that *OBP1* is involved in cell cycle initiation in response to developmental and environmental signals (Skirycz et al., 2008). Similarly, the strong enrichment of the MSA element in the DOF target genes showing a strong M phase peak expression suggests that other factors are involved in the signaling between *OBP1* and the activation of these mitotic cell cycle genes.

The organization of cis-regulatory elements in Arabidopsis promoters

Complementary to the enrichment analysis of gene coexpression neighborhoods to gain novel insights about gene functions, summarizing all motif instances over all target genes provides a global view on motif organization in Arabidopsis. Enrichment analysis of cis-regulatory elements over all GO categories yielded several examples of strong associations between motifs and biological processes (Figure 4; for complete lists see Supplemental figures 2 and 3 or the ATCOECIS website). Similar to the motif enrichment analysis using gene coexpression neighborhoods, we found more NCS-motifs enriched over one or more GO categories compared to known plant motifs (50% and 10%, respectively). Moreover, combining genes from different GO categories with conserved motif instances reveals the existence of specific and global cis-regulatory elements. Whereas more than three quarter (327/430) of all NCS-motifs are only enriched in less than five GO categories, the remaining 103 motifs are enriched in multiple (between 5 and 45) GO categories. Examples of global cis-regulatory elements enriched in 15 or more categories are the TELO motif, the Ibox, the E2F motif and the AGATCTNN motif (Supplemental figure 2). Complementary, we found two other (sets of) motifs, CTATATAN and CT-dinucleotide motifs, showing strong position and strand specificity (i.e. close to the startcodon of the gene and on the same strand of the transcribed gene) resembling TATA and Y patch core promoter motifs, respectively. In agreement with Yamamoto and co-workers (Yamamoto et al., 2007), the Y path (e.g. ACAGAGNG or CNTCTCTC) is preferentially located closer to the transcription start site than the TATA motif (Supplemental table 4). Examples of specific motifs consist of the Heat Shock Element GAANNTTC found to be enriched in 'response to heat' genes, a DRE-like motif GNNGACCA enriched in red light signaling genes and ANGAAAGA enriched in cytokinin mediated signaling genes. When comparing motif position biases (Supplemental table 4), we found that 40% of the global cis-regulatory elements show a preferential promoter location compared to 12% for the specific elements. This tendency for the former to be preferentially located close to the transcription start site confirms their role as core promoter elements. Although this strong positional bias of CT-dinucleotide motifs confirms their putative function as core elements, several GO categories were found enriched for the presence of conserved CT-dinucleotide motifs, suggesting a biological role for these low complexity motifs. Examples include kinase regulatory activity (70% of genes have CNTCTCTC), microtubule motor activity (46% of genes have CTCTNCNC), cell wall biosynthesis (45% of genes have CNTCTCTC) and Golgi membrane localization (52% of genes have GNCTCTCN). Apart from positional biases for individual motifs, for a set of ribosome biogenesis genes we found a clear and strict promoter organization when

comparing TELO and UP1 motif instances. On a genome-wide scale both motifs are significantly enriched in ribosomal genes and in 92% of the genes containing both motifs the TELO motif is located more upstream compared to the UP1 motif (Supplemental figure 4). This observation confirms the existence of cis-regulatory motifs in plants showing clear organizational constrains (Kim et al., 2006).

Discussion

The aim of our study was to investigate the applicability of coexpression networks to infer functional information for Arabidopsis genes. For a large fraction of genes with similar functional annotation, either using GO categories or AraCyc pathways, elevated coexpression levels were found using the expression coherence measure (Wei et al., 2006). Although many of these functional categories only partially correspond to transcriptional modules, the clustering of expression profiles using a gene-centric approach provides a practical starting point to study the coexpression neighborhood of a certain gene. Whereas enrichment analysis using GO confirms the 'guilt-by-association' principle for many genes (Horan et al., 2008), our benchmark experiment quantifying the predictive power of coexpression networks to infer known functional annotations reveals that for a majority of biological processes many known GO associations cannot be deduced from the coexpression network. Although more advanced computational classification systems trained for a specific biological process can partially solve this problem (Li et al., 2006), the integration of information about cis-regulatory elements provides an alternative approach to further characterize gene functions.

By combining known plant motifs and a new set of evolutionary conserved motifs we could annotate 9,117 coexpression neighborhoods with one or more motif. Compared to the Pathway-Level Co-expression method implemented in CressExpress (Srinivasasainagendra et al., 2008), which selects relevant genes if they are coexpressed with multiple query genes, the application of a statistical test for enrichment of functional annotation provides a robust and complementary method to identify new genes involved in different biological processes. Similar GO enrichment tools are also available in Arabidopsis coexpression tools like ACT (Jen et al., 2006), ATTED-II (Obayashi et al., 2007) and Plant Gene Expression Database (Horan et al., 2008). Clearly, the annotation of enriched cis-regulatory elements in guide gene clusters provides additional information compared to existing coexpression tools for Arabidopsis like ACT (Jen et al., 2006), CressExpress (Srinivasasainagendra et al., 2008) and the Plant Gene Expression Database (Horan et al., 2008). For a set of 866 putative cis-regulatory elements identified using the network-level evolutionary conservation principle, we

found that 88% of them are significantly enriched in one or more coexpression neighborhoods and that half of these NCS-motifs are enriched in one or more GO category. Since 37% of these motifs do not match any known plant cis-regulatory element, the detailed information about conserved motif instances provides a valuable resource to further enlarge our knowledge about transcriptional control in plants. Whereas the ATTED-II coexpression database also provides information about cis-regulatory elements, only 7-bp words are used to predict functional elements using the CEG method (i.e. correlation between gene expression and a defined gene group (Obayashi et al., 2007)). The presence of both known and NCS-based cis-regulatory elements in ATCOECIS offers a complementary set of tools to analyze coexpression gene sets. It contains a diverse set of simple and intuitive search functions that makes it possible to retrieve information about GO and motif enrichment for the gene coexpression neighborhoods described in this study. In addition, user-defined gene sets generated using clustering of dedicated expression data or chromatin immunoprecipitation experiments can be processed to identify motifs overrepresented in the target genes. Although some tools (e.g. ACT) provide clique finders to extract sets of genes showing consistent coexpression, so far we were unable to obtain better results when systematically comparing cliques with other clustering algorithms using GO and motif enrichment (unpublished results).

To demonstrate the utility of our framework to detect new regulatory interactions, we used publicly available transcriptome data of *OBP1* overexpression lines. This DOF transcription factor was recently identified as a regulator integrating developmental signals and cell cycle initiation (Skirycz et al., 2008). Starting from differentially expressed genes, the clustering of expression data and subsequent motif analysis identified 5 different transcription factor binding sites that could be linked with different modes of DOF regulation. Whereas the TAAAG DOF motif and the MSA element were found to be enriched in many upregulated genes (Table 3), clustering of the downregulated genes yielded 3 coexpression clusters with motif enrichment. The largest cluster mainly contained photosynthesis genes having an Ibox in their promoter and the two smaller clusters, showing stress and defense response, were enriched for the MYCATRD22 dehydration responsive element and LS7ATPR1, an element involved in systemic inducible plant defense responses, respectively. The presence of several DOF upregulated genes involved in DNA replication with S phase peak expression in synchronized Arabidopsis cell suspensions (Menges et al., 2003) suggests a link between *OBP1* and the E2F pathway. Indeed, comparison of E2F target genes with these DOF targets showed a significant overlap of 65 genes of which 74% have a WTTSSCSS E2F binding site in their promoter (Table 3). Our hypothesis that a regulatory interaction exists between the *OBP1* and the *E2Fa* transcription factor is supported by the fact that the latter is also upregulated in the *OBP1* overexpression line. We speculate that *OBP1*, linking

developmental and environmental signals with cell cycle initiation, might regulate several transcription factors controlling the progression through the different phases of the cell cycle.

As reported in this analysis, the sensitivity of coexpression functional prediction systems varies largely for genes involved in different biological processes. Also the selected set of microarray experiments, together with the applied distance measures and similarity thresholds, will have a great influence on the biological relevance of predicted gene functions (Yeung et al., 2004). Although it has been shown that coexpression is relatively stable when using >100 arrays (Vandepoele et al., 2006; Aoki et al., 2007), the availability of relevant microarray experiments to infer regulatory networks for a biological process of interest undoubtedly will increase the resolution and prediction accuracy of meta-analysis platforms. Whereas the association of different global and specific regulatory elements with different GO categories provides a fist glimpse on the regulatory logic embedded in plant promoters, the application of biclustering methods on a genome-wide scale can provide more detailed insights about the combinatorial nature of transcriptional control in plants. Similarly, the application of coexpression neighborhood analysis in a multi-species phylogenomic framework using orthologous gene relationships will make it possible to maximally exploit evolutionary conservation and enrichment analysis for gene function inference.

Material and methods

Expression data

A total of 322 (48x3 AtGenExpress Development and Tissue slides + 68x2 AtGenExpress Stress slides + 42 Birnbaum Root slides) Affymetrix ATH1 microarray slides monitoring the transcriptional activity of ~ 20,000 Arabdidopsis genes in different tissues and under different experimental conditions were retrieved from the Nottingham Arabidopsis Stock Centre (http://arabidopsis.info/. Raw data was background corrected and normalized using RMA (Irizarry et al., 2003) and a custom-made Chip Description File (CDF). This high-quality CDF file was build using selected reporter probes that have perfect sequence identity with a single target gene's transcript. Reporters that hybridize with one mismatch to another gene's transcript are filtered out. We also filtered out reverse complementary matching reporters and reporters that hybridize multiple times on the genomic sequence. The latter was done in order to remove reporters that match unannotated sequences. We included probe sets in this study only if they consisted of at least eight reporters which resulted in 19,937 unique probe sets (Casneuf et al., 2007). Note that these stringent criteria used to construct the CDF file make it possible to reliably measure expression values for duplicated

genes (i.e. free from cross-hybridization between paralogs showing high sequence similarity).The mean intensity value was calculated for the replicated slides. As a result, 129 experiments measuring the expression for 19,937 genes were retained for further analysis yielding an expression matrix with approximately 2.5 million data points (Supplemental table 5).

Expression coherence and clustering

The expression coherence (EC), which is a measure for the amount of expression similarity within a set of genes, was calculated as described by Pilpel and co-workers (Pilpel et al., 2001). EC reports the fraction of gene pairs per GO category that show elevated coexpression. Here, the Pearson Correlation Coefficient (PCC) was used as a measure for similarity between expression profiles. Based on the similarity between expression profiles for 1,000 random genes (~1,000*999*0.5 gene pairs), a PCC threshold of 0.72 corresponding with the 95th percentile of this random distribution was used to detect significantly co-expressed genes. To calculate the random EC for Gene Ontology (GO) categories, random gene sets were sampled with the same size as the category under investigation.

To create guide gene clusters, we selected for each gene all coexpression partners showing a PCC bigger then or equal to a defined threshold. Only guide gene clusters containing ten or more genes were retained. Three PCC thresholds were evaluated corresponding with the 90th, 95th and 99th percentile of the random background distribution. Note that guide gene clusters can overlap with each other because each gene present on the ATH1 microarray is initially selected as a guide gene. CAST clusters were identified as described in (Vandepoele et al., 2006).

Gene Ontology functional annotation

Gene Ontology associations for *Arabidopsis* proteins were retrieved from TAIR (www.arabidopsis.org; (Swarbreck et al., 2008)). The assignments of genes to the original GO categories were extended to include parental terms (i.e. a gene assigned to a given category was automatically assigned to all the parent categories as well) using the Perl GO::Parser and GO::Node modules. All GO categories containing fewer than 25 genes were discarded from further analyses. Enrichment values were calculated as the ratio of the relative occurrence in a set of genes to the relative occurrence in the genome. The statistical significance of the functional enrichment within gene sets was evaluated using the hypergeometric distribution adjusted by the Bonferroni correction for multiple hypotheses testing. Corrected p-values<0.05 were considered as significant. GO-motif networks were drawn using Cytoscape (Cline et al., 2007).

Evaluation of functional predictive power

The functional predictive power of a gene's coexpression neighborhood was determined by calculating the sensitivity (SN = TP / (TP + FN)) and the positive predictive value or precision rate (PPV = TP / (TP + FP)). For each guide gene *i* all significant GO enrichments found in the set of coexpressed genes were considered as GO predictions for gene *i*. True positives (TP) are actual positive examples predicted as positives, false negatives (FN) are actual positive examples predicted as negatives and false positives (FP) are actual negative examples predicted as positives. As GO annotations are far from complete, we estimated the number of false predictions using a random sampling approach. Starting from all *j* positive genes annotated with a particular GO term, FP was estimated by randomly selecting *j* negative genes and counting how much positive predictions were made. This procedure was repeated 100 times and FP for a given GO term was calculated as the average fraction of positive predictions.

Cis-regulatory element analysis

Starting from all possible 8-mers (generated using the 5-letter alphabet A,C,G,T,N) we applied the Network-level Conservation Score to determine evolutionary conserved motifs present in the upstream sequences of Arabidopsis genes. This evolutionary filter is used to discriminate between potentially functional and false motifs and applies a systems-level constraint to identify putative cis-regulatory elements (Elemento and Tavazoie, 2005; Vandepoele et al., 2006). The method exploits the well-established notion that each TF regulates the expression of many genes in the genome and that the conservation of global gene expression between two related species requires that most of these targets maintain their regulation. In practice, this assumption is tested for each candidate motif by determining its presence in the upstream regions of two related species (here Arabidopsis and poplar) and by calculating the significance of conservation over orthologous genes. Orthologous groups were identified through protein clustering using OrthoMCL (Li et al., 2003). Starting from an all-against-all BLASTP sequence similarity search using the full proteomes of *A. thaliana* (26,541 proteins) and *Populus trichocarpa* (45,554 proteins), 11,707 orthologous clusters were defined, covering 18,088 *Arabidopsis* and 22,760 poplar genes. These orthologous groups contain inparalogous genes (i.e., genes duplicated after the divergence between Arabidopsis and *Populus*), and thus offer a more realistic representation of orthology compared to, for example, reciprocal best hit approaches. Motif mapping was done using dna-pattern (RSA tools; (van Helden et al., 2000)) and was restricted to the first 1,000bp upstream from the translation start site or to a shorter region if the adjacent upstream gene is located within a distance smaller than 1000 bp. Starting from all 193,584 8-mers the top 5% motifs with the highest NCS values (NCS score > 12.48) were selected and similar motifs

were grouped. We measured the similarity between two motifs as the PCC of their corresponding position weight matrix. Note that all NCS-motifs are represented by consensus sequences and that the transformation to PWMs was only done for internal motif processing. Each motif of length w was represented using a single vector, by concatenating the rows of its matrix (obtaining a vector of length 4*w). Subsequently, the PCC between every alignment of two motifs was calculated, as they are scanned past each other, in both strands (Kreiman, 2004; Xie et al., 2005). Then, all motifs with a PCC > 0.75 were considered as similar and only the motif with the highest NCS value was retained using it's consensus sequence. This resulted in a set of 866 non-redundant motifs that were used for further analysis.

To calculate motif enrichment for clusters, only Arabidopsis NCS-motif matches conserved in one or more orthologous poplar gene were retained. Significance levels were calculated using the hypergeometric distribution adjusted by the Bonferroni correction for multiple hypotheses testing (using the number of evaluated motifs). Corrected p-values < 0.05 were considered as significant.

Acknowledgements

We want to thank Mattias de Hollander for helpful discussions and technical assistance with the analysis of cis-regulatory elements.

References

Aoki K, Ogata Y, Shibata D (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. Plant and Cell Physiology **48:** 381-390

Ben-Dor A, Shamir R, Yakhini Z (1999) Clustering gene expression patterns. J Comput Biol **6:** 281-297

Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, Benfey PN (2003) A gene expression map of the Arabidopsis root. Science **302:** 1956-1960

Casneuf T, Van de Peer Y, Huber W (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. BMC Bioinformatics **8:** 461

Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD (2007) Integration of biological networks and gene expression data using Cytoscape. Nat Protoc **2:** 2366-2382

Despres C, DeLong C, Glaze S, Liu E, Fobert PR (2000) The Arabidopsis NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. Plant Cell **12:** 279-290

Elemento O, Tavazoie S (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol **6:** R18

Fleming AJ, Manzara T, Gruissem W, Kuhlemeier C (1996) Fluorescent imaging of GUS activity and RT-PCR analysis of gene expression in the shoot apical meristem. Plant J **10:** 745-754

Geisler M, Kleczkowski LA, Karpinski S (2006) A universal algorithm for genome-wide in silicio identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in Arabidopsis. Plant J **45:** 384-398

Haberer G, Mader MT, Kosarev P, Spannagl M, Yang L, Mayer KF (2006) Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea. Plant Physiol **142:** 1589-1602

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

Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, Zhu JK, Cushman JC, Gollery M, Girke T (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol **147:** 41-57

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

Jen CH, Manfield IW, Michalopoulos I, Pinney JW, Willats WG, Gilmartin PM, Westhead DR (2006) The Arabidopsis co-expression tool (ACT): a WWW-based tool and database for microarray-based gene expression analysis. Plant J **46:** 336-348

Kim DW, Lee SH, Choi SB, Won SK, Heo YK, Cho M, Park YI, Cho HT (2006) Functional conservation of a root hair cell-specific cis-element in angiosperms with different root hair distribution patterns. Plant Cell **18:** 2958-2970

Kreiman G (2004) Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. Nucleic Acids Res **32:** 2889-2900

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:** 2178-2189

Li Y, Lee KK, Walsh S, Smith C, Hadingham S, Sorefan K, Cawley G, Bevan MW (2006) Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine. Genome Res **16:** 414-427

Ma S, Gong Q, Bohnert HJ (2007) An Arabidopsis gene network based on the graphical Gaussian model. Genome Res **17:** 1614-1625

Menges M, Hennig L, Gruissem W, Murray JA (2003) Genome-wide gene expression in an Arabidopsis cell suspension. Plant Mol Biol **53:** 423-442

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

Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. BMC Genomics **6:** 25

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H (2007) ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Res **35:** D863-869

Palaniswamy SK, James S, Sun H, Lamb RS, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol **140:** 818-829

Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci U S A **102:** 8633-8638

Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet **29:** 153-159

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet **37:** 501-506

Skirycz A, Radziejwoski A, Busch W, Hannah MA, Czeszejko J, Kwasniewski M, Zanor MI, Lohmann JU, De Veylder L, Witt I, Mueller-Roeber B (2008) The DOF transcription factor OBP1 is involved in cell cycle regulation in Arabidopsis thaliana. Plant J **56:** 779-792

Srinivasasainagendra V, Page GP, Mehta T, Coulibaly I, Loraine AE (2008) CressExpress: a tool for large-scale mining of expression data from Arabidopsis. Plant Physiol **147:** 1004-1016

Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science **302:** 249-255

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res **36:** D1009-1014

van Helden J, Andre B, Collado-Vides J (2000) A web site for the computational analysis of yeast regulatory sequences. Yeast **16:** 177-187

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

Vandepoele K, Vlieghe K, Florquin K, Hennig L, Beemster GT, Gruissem W, Van de Peer Y, Inze D, De Veylder L (2005) Genome-wide identification of potential plant E2F target genes. Plant Physiol **139:** 316-328

Walther D, Brunnemann R, Selbig J (2007) The regulatory code for transcriptional response diversity and its relation to genome structural properties in A. thaliana. PLoS Genet **3:** e11

Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A (2006) Transcriptional coordination of the metabolic network in Arabidopsis. Plant Physiol **142:** 762-774

Wellmer F, Riechmann JL (2005) Gene network analysis in plant development by genomic technologies. Int J Dev Biol **49:** 745-759

Wolfe CJ, Kohane IS, Butte AJ (2005) Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics **6:** 227

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature **434:** 338-345

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics **8:** 67

Yeung KY, Medvedovic M, Bumgarner RE (2004) From co-expression to co-regulation: how many microarray experiments do we need? Genome Biol **5:** R48

Zhang W, Ruan J, Ho TH, You Y, Yu T, Quatrano RS (2005) Cis-regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in Arabidopsis thaliana. Bioinformatics **21:** 3074-3081

Figure Legends

Figure 1. Expression Coherence scores for genes functionally annotated using Gene Ontology (A) and AraCyc (B). BP, MF and CC refer to Biological Process, Molecular Function and Cellular Component categories and the number of categories is indicated in parenthesis.

Figure 2. Functional enrichment of Gene Ontology and cis-regulatory element annotation for guide gene cluster AT5G59220. Lines indicate coexpression relationships and colored circles show the functional annotation for the individual genes. Enrichment analysis is performed using the hypergeometric distribution and Bonferroni correction for multiple hypotheses testing. Results are shown for the ATH95 benchmark coexpression network.

Figure 3. Functional predictive power for three benchmark coexpression networks build using different expression similarity thresholds (ATH90, ATH95 and ATH99). Panels A-C show cumulative sensitivity scores for a subset of GO categories and panel D shows overall cumulative sensitivity scores. PPV refers to Positive Predictive Values.

Figure 4. Examples of cis-regulatory motifs showing significant associations with one or more Gene Ontology categories. GO-motif networks reveal for different GO categories the fraction of genes having the motif in their promoter (p-value<0.05 using the hypergeometric distribution). The line thickness reflects the motif coverage per GO category and varies from 6% to 70%. Known motifs from AGRIS or PLACE are indicated in italic.

Tables

Table 1. Properties of the different coexpression networks.

Network name (1)	# genes	#edges	median number of edges per gene (lower and upper quartile)	# (%) genes with GO BP enrichment	Average PPV for GO categories (2)	Average SN for GO categories (3)
ATH90 *	19,716	13,580,283	1,625 (341, 2988)	18,668 (94%)	93%	39%
ATH95 *	18,861	6,765,135	650 (88, 1590)	16,663 (84%)	92%	39%
ATH99 *	14,187	1,504,781	91 (11, 449)	9,566 (50%)	88%	33%
ATHGGM	6,653	25,106	4 (2, 8)	1,232 (19%)	85%	9%

(1) * based on 50 GO categories covering 11,838 guide genes (2) PPV = Positive Predictive Value (3) SN = Sensitivity

Table 2. GO enrichment for the 10 most frequent cis-elements enriched in ATH95 gene coexpression neighborhoods (1).

Motif	#genes	Known motif	GO category		Enrichment	
			GO label	GO description	p-value	fold (2)
AAACCCTA	2524	TELO	GO:0042254	ribosome biogenesis and assembly	5.84E-53	7.67 (3.85)
CTTATCCN	1794	Ibox	GO:0015979	photosynthesis	2.18E-87	11.88 (3.15)
GGCCCANN	1601	UP1	GO:0042254	ribosome biogenesis and assembly	6.08E-68	11.93 (2.68)
GCCACGTN	1475	Gbox	GO:0015979	photosynthesis	1.95E-86	13.55 (n.e.)
GCGGGAAN	1303	E2F	GO:0006260	DNA replication	8.89E-26	8.80 (9.40)
GACCGTTN	930	MSA	GO:0007018	microtubule-based movement	9.83E-12	9.82 (n.e.)
AANGTCAA	389	Wbox	GO:0050832	defense response to fungi	6.22E-08	13.00 (1.50)
CNGATCNA	382	AGMOTIFNTMYB2	GO:0048193	Golgi vesicle transport	1.22E-08	17.60 (n.e.)
NCGTGTCN	328	ABRE	GO:0009737	response to abscisic acid stimulus	1.04E-07	6.84 (2.00)
CATGCANN	284	RYREPEATBNNAPA	GO:0048316	lipid transport	9.99E-04	8.10 (1.84)

(1) Similar motifs with the same GO enrichment trends are not shown.

(2) Enrichment values considering all motif matches on Arabidopsis promoters (i.e. both conserved and non-conserved) are shown in parenthesis; n.e. no enrichment.

Table 3. Regulatory analysis of E2Fa/OBP1 target genes.

Features / Data set	OBP1_UP	OBP1_DOWN	E2Fa_UP	E2Fa_DOWN	OBP1/E2Fa_UP
# Genes	632	842	412	220	65*
Cell Cycle Modulated expression					
G1	9	13	2	3	2
S	38	52	70	13	12
G2	4	19	0	1	0
M	114	3	3	0	3
Enrichment Gene Ontology (1)					
GO:0006260 DNA replication	6.6x (2.83E-06)		24.2x (8.59E-36)		55.9x (4.90E-19)
GO:0007017 microtubule-based process	9.9x (2.61E-13)				
GO:0042545 cell wall modification		4.3x (3.52E-03)			
GO:0009607 response to biotic stimulus		2.5x (5.87E-09)			
Enrichment cis-regulatory elements (2)					
GCGGGAAN (E2F)	3% (9.97E-06)		15% (4.19E-65)		26% (1.11E-20)
WTTSSCSS (E2F PLACE)			55% (3.16E-37)		74% (4.72E-14)
TAAAG (DOF PLACE)	94% (4.45E-04)				
GACCGTTN (MSA)	7% (6.64E-30)				
CTTATCCN (Ibox)		4% (4.56E-05)			
CCATGTGN (MYCATERD1)		3% (1.95E-04)			
ANCACATG (MYCATRD22)		6% (2.02E-07)			

* overlap significantly larger than expected by chance (p-value<3.50E-27)

(1) enrichment fold (p-value)

(2) fraction of genes with motif (p-value)

Supplemental data

Supplemental table 1. EC values for different GO and AraCyc categories.

Supplemental table 2. Coexpression network properties per GO category.

Supplemental table 3. Predictive power scores for different GO categories based on the full ATH95 network.

Supplemental table 4. Position and strand biases of conserved NCS-motif instances.

Supplemental table 5. Microarray experiments in expression compendium.

Supplemental figure 1. Correlation between average connectivity and prediction sensitivity per GO category.

Supplemental figure 2. Enrichment of a subset of NCS-motifs over GO categories. For a complete overview, please use http://bioinformatics.psb.ugent.be/ATCOECIS/

Supplemental figure 3. Enrichment of known AGRIS and PLACE motifs over GO categories.

Supplemental figure 4. Genome-wide positional biases of TELO and UP1 motifs in genes enriched for ribosome biogenesis. For a set of 226 genes containing a conserved TELO and UP1 motif the positions upstream from the translation start site are shown in a histogram. The numbers in parenthesis indicate the number of motif instances.

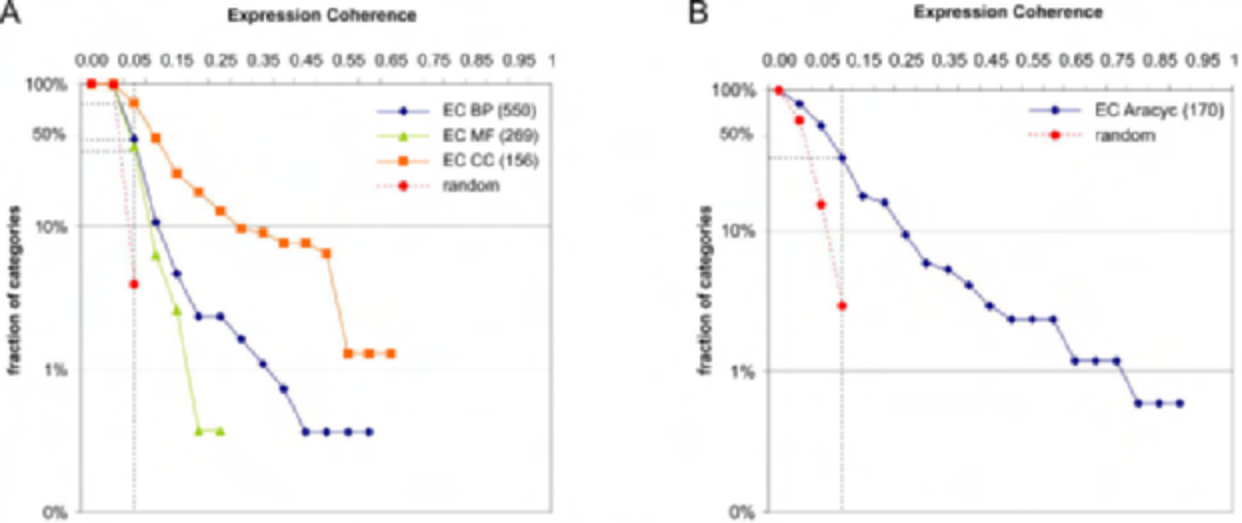

Figure 1. Expression Coherence scores for genes functionally annotated using Gene Ontology (A) and AraCyc (B). BP, MF and CC refer to Biological Process, Molecular Function and Cellular Component categories and the number of categories is indicated in parenthesis.

Enrichment in coexpression neighborhood of AT5G59220 (protein phosphatase 2C):
- GO:0009737 response to abscisic acid stimulus (p-value=0.0057, 28.0 fold enrichment)
- Motif NCACGTGN (p-value=3e-06, 13.6 fold enrichment)

Figure 2. Functional enrichment of Gene Ontology and cis-regulatory element annotation for guide gene cluster AT5G59220. Lines indicate coexpression relationships and colored circles show the functional annotation for the individual genes. Enrichment analysis is performed using the hypergeometric distribution and Bonferroni correction for multiple hypotheses testing. Results are shown for the ATH95 benchmark coexpression network.

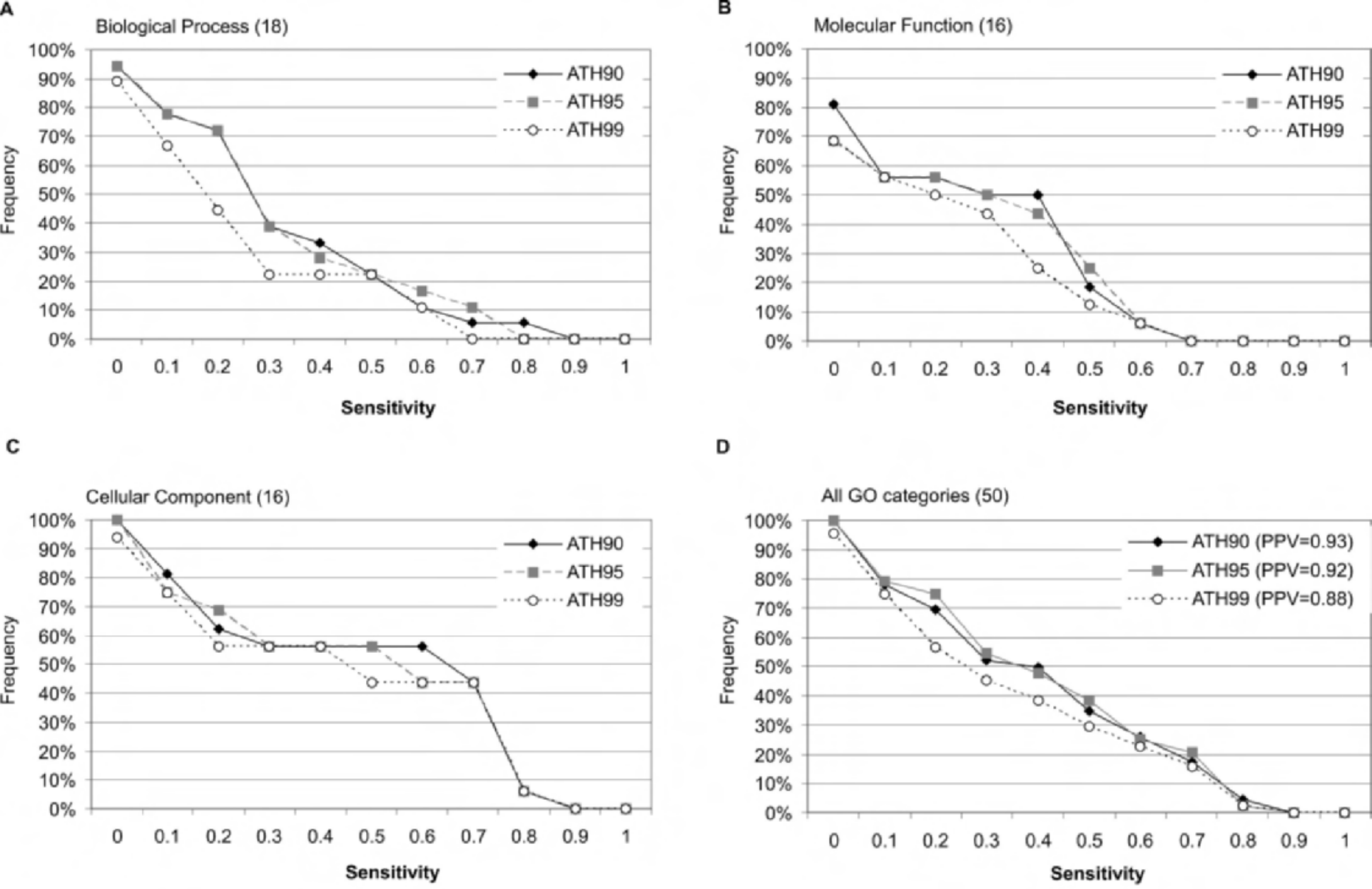

Figure 3. Functional predictive power for three benchmark coexpression networks build using different expression similarity thresholds (ATH90, ATH95 and ATH99). Panels A-C show cumulative sensitivity scores for a subset of GO categories and panel D shows overall cumulative sensitivity scores. PPV refers to Positive Predictive Values.

Figure 4. Examples of cis-regulatory motifs showing significant associations with one or more Gene Ontology categories. GO-motif networks reveal for different GO categories the fraction of genes having the motif in their promoter (p-value<0.05 using the hypergeometric distribution). The line thickness reflects the motif coverage per GO category and varies from 6% to 70%. Known motifs from AGRIS or PLACE are indicated in italic.