Running Title:
Cis-elements in Arabidopsis

Corresponding Author:
Name   Klaus Mayer
Address MIPS, Institute for Bioinformatics, GSF National Research Center for Environment and Health, Ingolstaedter Landstr. 1, 85758 Neuherberg, Germany
Telephone  ++49-(0)89-31873584
E-Mail   Kmayer@gsf.de
Fax       ++49-(0)89-31873585

Journal Research Area:
Genome Analysis
Bioinformatics
Systems Biology and Gene Regulation
1. Financial Sources
This work was funded in the GABI program of the German Ministry of Education and Research (BMBF)

2. Corresponding Author:
Klaus Mayer, e-mail Kmayer@gsf.de, fax +49 (0)89 3187 3585
Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis thaliana and Brassica oleracea

Georg Haberer¹,², Michael T Mader¹,³, Peter Kosarev², Manuel Spannagl², Li Yang², Klaus FX Mayer²,*

¹ both authors equally contributed to this work
² MIPS, Institute for Bioinformatics, GSF National Research Center for Environment and Health, Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany
³ Institute of Stem Cell Research, GSF National Research Center for Environment and Health, Ingolstaedter Landstr. 1, 85764 Neuherberg, Germany

* to whom correspondence should be addressed:
Phone: +49 (0)89 3187 3584
E-mail.: Kmayer@gsf.de
Abstract

The rapidly increasing amount of plant genomic sequences allows for the detection of cis-elements through comparative methods. In addition large-scale gene expression data for Arabidopsis thaliana have recently become available. Co-expression and evolutionarily conserved sequences are criteria, widely used to identify shared cis-regulatory elements. In our study we employ an integrated approach to combine two sources of information, co-expression and sequence conservation. Best candidate orthologous promoter sequences were identified by a bidirectional best blast hit strategy in genome survey sequences from Brassica oleracea. The analysis of 779 microarrays from 81 different experiments provided detailed expression information for Arabidopsis genes co-expressed in multiple tissues and under various conditions and developmental stages. We discovered candidate transcription factor binding sites in 64% of the Arabidopsis genes analyzed. Among them, we detected experimentally verified binding sites and showed strong enrichment of shared cis-elements within functionally related genes. This study demonstrates the value of partially shotgun sequenced genomes and their combinatorial use with functional genomics data to address complex questions in comparative genomics.
Introduction

*Brassica oleracea* (*Brassica*) enjoys a close evolutionary relationship to *Arabidopsis thaliana* (*Arabidopsis*). The two genera separated approximately 12-24 million years ago (Yang et al., 1999b). While *Arabidopsis* serves as a model for many research topics in plant genomics and plant biology, subspecies of *Brassica oleracea* cover a wide range of commercially important vegetable crops such as broccoli, cauliflower and cabbage. The availability of the whole genome sequence for *Arabidopsis* as well as the large amount of genomic survey sequences (GSSs) for *Brassica* and their close evolutionary relationship make them useful for comparative plant genomic studies (Ayele et al., 2005; Katari et al., 2005).

Comparative genomics has been proven to be a powerful tool for the discovery of a large variety of functional elements by their conservation between related species. The usefulness of *Brassica* GSSs for the improvement of genome and specifically gene structure annotation in *Arabidopsis* as well as for comparative studies of the repeat contents of both genomes has been reported (Ayele et al., 2005; Katari et al., 2005; Zhang and Wessler, 2004).

In particular, it has been shown that comparative genomics approaches are able to detect genetic elements that are often difficult to discover due to their small size and/or limited information content. Examples include genetic elements like micro-RNAs and *cis-regulatory* elements (Jones-Rhoades and Bartel, 2004; Wasserman et al., 2000; Cliften et al., 2003; Kellis et al., 2003).

Comparative genomics detects *cis-elements* by their conservation between two or more evolutionary related sequences from orthologous genes. The assumption is that orthologs exhibit a common regulatory mode that is reflected in the conservation of transcription factor binding sites. Phylogenetic footprinting (Wasserman et al., 2000) and a related approach – phylogenetic shadowing (Boffelli et al., 2003) – have been successfully applied over a wide range of genera ranging from bacteria to yeasts and mammals (McCue et al., 2001; Boffelli et al., 2003; Zhang and Gerstein, 2003). For plants, first pioneering
studies have been undertaken and encouraging results have been reported (Lee et al., 2005; Guo and Moose, 2003; Bao et al., 2004; Inada et al., 2003).

In addition to sequence conservation, a different popular approach uses functional information, mainly co-expression information, within one species to discover *cis*-elements. Powerful technologies to monitor transcriptional states and dynamics on a genome scale are well established and widely applied. The analysis of co-expressed genes under different conditions and states has been shown to be highly valuable for the analysis of shared *cis*-regulatory elements (Harmer et al., 2000). Diverse algorithms like expectation-maximization or Gibbs-sampler have been adapted and applied to detect motifs that are overrepresented within sets of functionally related promoter sequences (Thijs et al., 2002; Bailey and Elkan, 1994; Tompa et al., 2005).

The majority of studies have used either sequence conservation or overrepresentation of motifs in promoters of co-expressed genes to discover *cis*-regulatory elements. However, some studies used a combination of both approaches to evaluate and/or screen detected motifs. For instance, Kellis and coworkers (Kellis et al., 2003) scored the motifs both by their overrepresentation within particular genomic regions and their conservation between four yeast species to enrich for functional *cis*-regulatory elements. In a similar approach, conserved word occurrences have been further evaluated by expression data for a variety of species and kingdoms (Elemento and Tavazoie, 2004). In plants, no large-scale integrative studies for *cis*-element detection have been conducted so far.

Recent developments in the detection of *cis* regulatory elements integrate both phylogenetic information as well as co-expression information (Wang and Stormo, 2003, Sinha et al., 2004; Moses et al., 2004; Siddhartan et al., 2005). In our study we apply PhyloCon to consider both conservation between orthologous promoter regions as well as co-expression of genes within a species. Promoters of co-expressed genes and their respective orthologous counterparts are selected and subjected to a PhyloCon based analysis. Initial profiles are generated from multiple alignments of orthologous promoter sequences. Motifs
shared between different orthologous groups emerge by iteratively combining initial profiles. The high performance of PhyloCon has been demonstrated both by the application on simulated data as well as on a test set of known yeast transcription factor binding sites (Wang and Stormo, 2003).

To integrate co-expression and sequence conservation for motif discovery, adequate information sources – co-expression and orthology information – are required. Recent contributions have provided both large-scale expression data for Arabidopsis as well as sequence data from Brassica. A large and high-quality expression dataset comprising about 800 microarrays has recently been made available (Craigon et al., 2004; Schmid et al., 2005). They monitor transcriptional states of the genome under various environmental conditions, for different organs and tissues as well as during distinct developmental phases. These data allow to identify genes that are co-expressed over a wide variety of different conditions. The high-quality Arabidopsis genome sequence and a large amount of Brassica genomic survey sequences provide a second information component. These sequences enable to search for conserved elements between the promoters of corresponding Arabidopsis and Brassica genes.

In this study we undertook a comprehensive analysis of thousands of Brassica-Arabidopsis orthologous promoter pairs. To generate co-expression information we analyzed a set of 81 microarray experiments from Arabidopsis totaling 779 chips. Promoters from co-expressed genes and their respective Brassica orthologous counterpart were selected. The resulting promoter sets were analyzed and a large number of candidate sites have been discovered. These sites are derived from profiles, which are conserved between orthologous promoters as well as are associated with co-expression. Many of the detected motifs are enriched for specific biological processes and pathways. Evaluation of our analysis with the aid of experimentally validated cis-regulatory elements from Arabidopsis confirms their significance. This study provides the basis for future cis-regulatory module analysis and analysis of regulatory circuits not restricted to Brassicaceae. It further demonstrates the benefits of partial genome sequences to address complex problems in comparative genomics.
Results

The main goal of our study was the genome wide discovery of candidate cis-regulatory regions in *Arabidopsis*. We selected PhyloCon to combine two sources of information for cis-element discovery - co-expression and sequence conservation. PhyloCon has been demonstrated to be very powerful both on biological and on controlled artificial data (Wang and Stormo, 2003). Motif discovery by PhyloCon considers two independent axes of information, conservation of a motif between orthologous promoters and overrepresentation in a set of co-expressed genes. The necessary information was deduced from the identification of best candidates for orthologous upstream sequences between *Arabidopsis* and assemblies of *Brassica* genomic survey sequences (GSSs). The *Brassica oleracea* genome has undergone a recent large-scale duplication event postdating the divergence of the two genera, *Arabidopsis* and *Brassica* (Town et al., 2006). Thus, a substantial number of *Arabidopsis* genes may not have an one-to-one ortholog in *Brassica*. However, previous studies have shown that recently duplicated genes are similar in their expression characteristics and cis-element composition (Blanc and Wolfe, 2004; Haberer et al., 2004). Thus, recently duplicated paralogs in *Brassica* can be expected to share high similarities to the corresponding *Arabidopsis* promoter.

To avoid potential misassignments we applied a stringent bidirectional best BlastN hit (BBH) strategy to detect the best available candidates for orthologous promoter sequences (Material and Methods). In the following we use the term ‘orthologous promoters’ for these candidate pairs. Co-expressed *Arabidopsis* genes were determined using 779 microarray hybridization data (Craigon et al., 2004; Schmid et al., 2005). Figure 1 shows a workflow of the analysis undertaken.

*Brassica* orthologous upstream sequences

We assembled a set of 567,365 *Brassica oleracea* GSSs by applying highly stringent clustering to minimize both redundancy within the GSSs and to prevent
the generation of erroneous hybrid clusters by exclusion of repetitive and/or ambiguous sequences. Assembly and repeat masking/filtering resulted in 142,489 clusters with an average length of 987 bp totaling 140.6 Mb non-redundant sequences. The genome size of *Brassica oleracea* has been estimated to about 600 Mb (Arumuganathan and Earle, 1991). Thus the clustered and repeat-filtered sequences correspond to approximately a quarter of the genome.

Next, we determined orthologous upstream sequences between *Arabidopsis* and *Brassica* by reciprocal BlastN comparisons (E ≤ 10\(^{-10}\)). Starting from 26,535 genes in MAtDB (Schoof et al., 2004) homologous sequences for 7,427 *Arabidopsis* upstream sequences were detected in the *Brassica* GSS assemblies. Applying the reciprocal best Blast hit criterion, 4,007 were retained as putative orthologous sequence pairs.

**Determination of co-expressed *Arabidopsis* genes**

Expression data available from the Nottingham *Arabidopsis* Stock Centre (Craigon et al., 2004) and generated within AtGenExpress (Schmid et al., 2005) were used for the analysis. Affymetrix probes were remapped onto the current annotation of MatDB and ambiguous probes were excluded from the analysis (see Materials and Methods for details). Quality filtering resulted in a total of 21,559 genes for which expression data were obtained. The expression experiments cover a broad variety of biological processes, tissues and stages. Expression data were normalized and Pearson correlation coefficients of all against all genes were computed (Materials and Methods).

The background distribution was derived from the pairwise correlations of all 21,559 genes used in this study (Fig. 2). To define groups of co-expressed genes, the 99%-quantile of this distribution was considered as significant (r = 0.803; Fig. 2). For each *Arabidopsis* gene we assigned all genes exceeding a Pearson correlation r ≥ 0.803 to its co-expression group (CEG). A particular CEG is defined by an anchor gene to which all other CEG members are significantly correlated (Fig. 3A). A gene can be assigned to multiple CEGs and therefore, two
CEGs may share subsets of genes (Fig. 3A). This procedure enables us to differentiate genes and gene groups that participate in several distinct biological processes. From the initial 21,559 genes, 13,254 genes provided anchor points for CEGs while 8,305 genes remained singletons. Approximately one third (4245) of the CEGs consisted of less than 25 members, and 49.4% (6553) of all groups had less than 100 members, indicative of a stringent selection scheme (Fig. S2 in the Supplement). CEGs for each anchor gene are provided via a web accessible database hosting the results from this study (http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp).

**Motif discovery by PhyloCon**

Each PhyloCon analysis group (PAG) is composed of the orthologous gene pairs of an individual CEG. *Arabidopsis* genes with no detectable orthologous upstream sequence in *Brassica* were removed from the analysis set. In the following, we refer to an orthologous pair of a *Brassica* GSS assembly and *Arabidopsis* promoter/upstream sequence as an orthologous promoter group (OPG). Thus, the collection of all OPGs of a particular CEG represents a PhyloCon analysis group (PAG).

Elimination of genes without corresponding *Brassica* orthologous promoter sequence significantly reduced the size and number of co-expressed groups since an OPG has been identified on average for only about one sixth of the *Arabidopsis* genes. In addition, each PAG had to consist of at least two OPGs. This filtering resulted in 4,540 PAGs that were subjected to a PhyloCon analysis. PhyloCon initially creates profiles from pairwise OPG alignments. In subsequent cycles, merging and trimming profiles of preceding cycles generates derived profiles. Thus, profiles of later cycles are derived from alignments of an increasing number of distinct OPGs. An example for profile generation by PhyloCon is given in Figure 3B. For each PAG both the final alignment matrices as well as profiles of previous cycles to which we refer as intermediate matrices were collected. This step reduces the likelihood to miss significant motifs in a noisy data set (see Materials and Methods for details). Analysis of all PAGs
revealed a total of 322,079 preliminary profiles including a large number of redundant intermediate matrices (see Materials and Methods). As (CT)$_n$-repeats (or its respective complement, (GA)$_n$) are very prominent in *Arabidopsis* promoters we pre-filtered consensus sequences of our matrices for the presence of such simple repeats. We analyzed the filtered matrices for overrepresentation within the associated CEG as compared to their frequency in all 21,559 analyzed *Arabidopsis* genes by testing against the cumulative binomial distribution. Within 3,861 PAGs we detected at least one motif model that was significantly overrepresented for the respective CEG ($p \leq 0.01$). Due to the partial coverage of the *Brassica* genome by GSSs, we could not identify putative orthologous promoter sequences for all members of a particular CEG. Hence, PAG sizes are generally smaller than the size of their corresponding CEG. Consequently for only a subset of the CEG members significant profiles are directly supported by conserved alignments. To overcome the limited PAG sizes and to transfer knowledge from the PAG analysis to all members of the CEG, we projected these profiles to all upstream sequences of the CEG. A profile was considered as a candidate motif for those genes of the CEG that contained at least one instance of the respective profile in their promoters.

Lengths of profiles predominantly range between 6 and 15 bp (Supplement Fig. 1). This is in good agreement with sizes of known individual transcription factor binding sites. To estimate the number of candidate sites per gene, overlapping sites/instances of different motifs were merged (Materials and Methods). Fusion of overlapping sites did not change the size distribution (Supplement Fig. S1). This indicates that our profiles detect well-confined regions within the promoters. We found on average 7.3 non-redundant sites per gene and a total of 61,745 sites in 8,407 *Arabidopsis* genes (out of 13,254 genes contained in all CEGs). Each CEG and each gene can be queried for a list of significant profiles at [http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp](http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp). For the web display, we assorted identical or nearly identical profiles exceeding a Pearson correlation $r \geq 0.98$ into one cluster (Materials and Methods). However, as merged profiles generally result in a higher degeneracy compared to the single
profiles, we only grouped but did not merge any profiles to retain the specificity for each profile for subsequent analysis.

**Detected Profiles Match Known *cis*-Elements**

We investigated whether profiles detected in this study match known *cis*-elements. For this purpose we screened elements from the PLACE and AGRIS databases (Higo et al., 1999; Davuluri et al., 2003). Both databases contain experimentally validated *cis*-regulatory elements specific for plants (Materials and Methods).

In total 537 motifs are contained within the two databases. However, there is a significant degree of redundancy both between the databases as well as within individual databases. In many cases it is difficult to decide whether two motif variants constitute binding sites for two distinct transcription factors or represent two sites for one transcription factor. Therefore, we used all binding sites listed in both databases. For 255 sites (out of 537 sites; 47.5%), we detected a profile similar to the described motif within PLACE or AGRIS. Table 1 lists a selection of detected matches.

Several reasons complicate the evaluation for profile matches to motifs reported in PLACE or AGRIS. Firstly, motifs in PLACE are derived from various plant species. Hence, some motifs might not be present within *Arabidopsis* and *Brassica*. Secondly, many motifs are either reported as consensus sequences or experimental reports are restricted to only one specific site in a particular promoter. Particularly, the latter description is likely too specific as many transcription factor binding sites are degenerated. In addition, some consensus sequences do not describe binding sites for individual transcription factors but instead give the (degenerated) consensus for a family of transcription factors like for example the Myb transcription factors (see Table 1). This problem is especially pronounced for known motifs for which only a short core sequence is present and which are involved in the regulation of numerous pathways. Examples comprise the ACGT-element or the CAAT-Box. Most importantly, the *Brassica* assembly only partially covers the *Brassica* genome and the average
length of the *Brassica* GSSs is about half of the average length of *Arabidopsis* promoters used in this study. Thus, we are missing a considerable number of genes or promoter regions for comparison. Nevertheless, our findings for several known motifs are consistent with experimental findings and functional enrichments described below. For instance, the PALBOX is frequently found in promoters of genes catalyzing steps in the phenylpropanoid biosynthesis. Consistently, we detected a significant enrichment of several profiles highly similar to it (e.g. Table 2) in the flavonoid, phenylpropanoid and lignin biosynthesis and in the category ‘response to UV-C’. A detailed description of the detected sites within the PAL promoter and their matches to known sites within this promoter is given in the last section. We also detected several profiles matching the G-box related ABA responsive element GCCACGTG (ABRE). In agreement with its regulatory function, these profiles were significantly overrepresented in the functional category ‘response to abscisic acid stimulus’ (Table 2).

**Detected Profiles are overrepresented within Specific Functional Categories and Biochemical Pathways**

Numerous studies demonstrated that co-expressed genes have an increased probability to be involved in a common biological process (DeRisi et al., 1997; Hughes et al., 2000; Schmid et al. 2005). Furthermore, correlations between the occurrence of particular motifs and specific functional categories have been demonstrated for instance in yeast (Kellis et al., 2003). To investigate whether we find enrichment for specific functional categories in the detected profiles, we made use of Gene Ontology information (GOSlim catalog) as well as information on biochemical pathways in *Arabidopsis* (KEGG pathways) (Berardini et al., 2004). Both GOSlim annotations as well the KEGG biochemical pathway information for *Arabidopsis* have been obtained from The Arabidopsis Information Resource (TAIR; www.arabidopsis.org). GOSlim and KEGG assignments for the genes contained within our analysis set were selected and only categories that contain at least two genes were considered. Furthermore,
we restricted our analysis for enrichment to GOSlim to categories describing biological processes. The enrichment in a particular KEGG pathway or a biological process defined by a GO term (Materials and Methods) was determined using the binomial coefficient against the genome-wide background distribution. P-values have been Bonferroni-corrected for multiple testing and corrected p-values ≤ 0.05 were considered as significant (Materials and Methods). For 538 out of 923 GOSlim categories and for 99 out of 292 KEGG pathways we found significant enrichments for at least one candidate profile. Table 2 lists several example profiles and their respective functional or pathway category for which we detected a significant enrichment. To compile a list of candidate profiles for each gene, we employed a similar schema as applied for the analysis of profiles overrepresented in co-expression groups. Using profiles located within promoters of genes associated to pathway or functional category assignments we analyzed for profiles which are significantly enriched in the respective category/pathway. This association rule even reports candidate motifs for genes that were not included in the analysis involving PAGs, e.g. singletons. One example are genes involved in the gibberellic acid (GA) biosynthetic pathway. For instance, a profile with a consensus sequence CACGTkTGGT (Table 2) is found in 61.5 % of the genes assigned to the gibberellic acid biosynthesis pathway and hence is more than fourfold enriched compared to a random expectation. Genes containing the motif in their promoters comprise almost all steps in the biosynthetic pathway for which Arabidopsis genes are known: an ent-copalyl diphosphate synthetase, the ent-kaurene oxidase GA3, several GA 20-oxidases and a GA 3ß-hydroxylase forming the last step to biologically active GAs. Comprehensive results can be found and queried at the associated website (http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp). Our findings and correlations will support the analysis of functional relations and aid to identify candidate regulatory sites within the promoter of individual genes of interest.
PhyloCon Sites Correspond to Experimentally Confirmed Sites

We evaluated our results using experimental findings. CRABS CLAWS (CRC), a member of the YABBY gene family, is required for nectary and carpel development in Arabidopsis (Lee et al., 2005). In our analysis CRC was present in various CEGs. Their constituting members can group these into two distinct biological processes. In the first group, CRC co-clusters with AP1, the floral organ identity gene AP3 and the SEPALLATA group (SEP1, SEP2 and SEP3). All these genes play critical roles in flower development and AP3 and SEP1-3 have been suggested to be involved in CRC regulation (Lee et al., 2005). The second cluster group is enriched for genes involved in sporocyte morphogenesis and contains genes like MALE STERILITY 2 (MS2; Aarts et al., 1997), SPOROCYTELESS (SPL; Yang et al., 1999a) and several genes with their expression patterns indicative of a role within the sporocyte. The promoter of CRC has been shown to contain five distinct regions which are conserved between three different species (Arabidopsis thaliana, Lepidium africanum and Brassica oleracea) (Lee et al., 2005). Detailed promoter studies have shown that these five regions are required for correct CRC expression and individual functional binding sites have been identified by site-directed mutagenesis. Within our analysis we identified 22 different conserved sites. 19 sites were located within one or another of the five regions and one site overlapped with an experimentally verified CArG-box (Fig. 4). It is noteworthy that several profiles were specific for one or the other of the biological process cluster groups. For example, a profile with the consensus sequence TGGAGGCA was present within the promoters of CRC and sporocyte-specific genes but missing in floral organ identity genes (Fig. 4). All sites detected in the CRC promoter are located within conserved regions.

Figure 5a depicts a PAG enriched for genes involved in phenylpropanoid biosynthesis. Within this group, an enzymatic chain involving Phenylalanine-ammonia-lyase (PAL1, the entry point of the biosynthetic pathway), trans-cinnamate 4-monooxygenase (C4H/CYP73A5), p-coumarate 3-hydroxylase (CYP98A3) and a caffeoyl-CoA O-methyltransferase like protein is present. We
found these genes to frequently co-cluster within our analysis. In addition, many of these PAG clusters contained two isoforms of the coumarate CoA-ligase (4CL1, 4CL2) and a second variant of the Phenylalanine-ammonia lyase (PAL2). Tight co-expression of PAL1, C4H and 4CLs has been reported not only for Arabidopsis (Mizutani et al., 1997) but has also been found in poplar (Hertzberg et al., 2001). The PAL-enzymes, C4H and 4CL1/2 make up the core of the phenylpropanoid pathway while CYP98A3 and the methyl-transferase catalyze steps of a specific branch within the pathway, the synthesis of precursors of the lignin biosynthesis (Figure 5a). For C4H several motifs characteristic of enzymes of the phenylpropanoid pathway have been reported, 4 P-boxes, 3 A-boxes and 2 L-boxes. As shown in Figure 5b, almost all box instances are either fully or partially covered by conserved sites detected from the analysis of several C4H containing PAGs.

In summary, the examples illustrate that conserved sites detected in this study correlate very well with known transcription factor binding sites.
Discussion

The detection of cis-regulatory elements in higher eukaryotes is a major challenge in functional genomics. Bioinformatic sequence analysis of transcription factor binding sites is notoriously difficult as cis elements are difficult to distinguish from background. Two major approaches are commonly undertaken. In the first strategy, co-expressed genes are selected and analyzed for shared cis sequence elements. A second approach, phylogenetic footprinting, aims to detect candidate transcription factor binding sites from conserved regions in alignments of orthologous promoters. Both strategies have been demonstrated to be powerful (Zhang and Gerstein, 2003). Wang and Stormo (2003) proposed to enhance cis-element discovery by a combination of two sources of information, co-expression and sequence conservation.

In our analysis, we applied PhyloCon to discover cis-elements in Arabidopsis upstream sequences. Co-expression information for Arabidopsis genes was derived from a large set of microarray experiments (Craigon et al., 2004). For the orthologous reference genome Brassica oleracea, about 560 Mb redundant genomic survey sequences (GSS) were available (Ayele et al., 2005; Katari et al., 2005). After repeat masking and clustering, 142 Mb of non-redundant GSS were analyzed for orthologous Arabidopsis equivalents through bidirectional best hits. We identified 4007 Arabidopsis genes with a corresponding best candidate orthologous promoter in Brassica. We detected a large number of candidate transcription factor binding profiles supported by sequence conservation and statistically significant overrepresentation within co-expressed genes. The mapping of these profiles on the 13,254 CEGs, i.e. including genes with no orthology information, revealed more than 60,000 sites in 8,407 Arabidopsis promoters (63.4 % of all genes found in one or more CEGs).

One limitation in our approach is the incomplete Brassica genome sequence that covers approximately one fourth of the genome. OPGs identified in this study represent the best available candidates for orthologous promoter pairs. Albeit PhyloCon uses sequence conservation for motif discovery strict orthologous relationship of sequence pairs is not compulsory. For the identification of cis-
regulatory elements paralogous promoters that contain conserved regulatory regions are useful as well (Haberer et al., 2004). Further limitations also include genomic rearrangements (e.g. insertions, deletions, translocations or duplications) that in the absence of positional information may affect the correct promoter identification for an OPG. However, noise potentially caused by these effects is limited by two provisions. Firstly PhyloCon generates new motifs by stepwise additions and combination of motifs found in individual OPG comparisons. Profiles from falsely assigned OPGs represent random matches unrelated to functional elements present in other OPGs of the respective PAG. Thus, they are unlikely to be added to the list of motifs reported. In addition we statistically evaluated overrepresentation of both intermediate and final motifs in the respective group of co-expressed genes.

A number of complete plant genomes will be available in the near future and will help to circumvent some of the limitations encountered with partial genomes. Map information for these genomes will enable to detect syntenic relationships and thus support the detection of true corresponding orthologous promoters. Map derived synteny relations, however, are impaired by the highly dynamic nature of plant genomes. Genome duplications, segmental and tandem duplications are prevalent and in plant genomes gene families are often highly expanded (Vision et al., 2000; Arabidopsis Genome Initiative, 2000; Simillion et al., 2002; Bowers et al., 2003). For *Brassica oleracea*, a report describes a recent partial duplication of a 2 MB contig (Town et al., 2006). This complicates the detection and definition of orthologous gene and promoter relationships. Specifically, one *Arabidopsis* gene may have none, one or more ortholog(s) in *Brassica* and a one-to-one mapping does not necessarily reflect the correct orthologous relationship.

Detection of cis-regulatory elements is known to be an error-prone process. Nevertheless, several observations indicate a successful enrichment for functional transcription factor binding sites in our study: sequence conservation of motifs between *Arabidopsis* and *Brassica*, enrichment of motifs in functional categories and detection of known sites. Sequence conservation between
evolutionary related species is generally considered as an indicator for either short divergence times or the functional importance of the respective elements. Insufficient sequence divergence imposes a severe problem for classic phylogenetic footprinting analysis based on sequence alignments as non-functional elements cannot be delimited from functional elements. From a large set of lambda clones Windsor et al. (2006) identified orthologous upstream sequences between *Arabidopsis* and its close relative *Boechera stricta*. They reported significant sequence conservation within these sequences. Albeit *Brassica oleracea* is more distantly related to *Arabidopsis* than *Boechera stricta*, sequence conservation may still be caused by insufficient divergence times. Support for this assumption comes from a previous analysis of one of our examples, *CRABS CLAW (CRC)*. A phylogenetic footprinting analysis by promoter alignments identified five conserved regions between *Arabidopsis* and *Brassica oleracea*, each comprising up to several hundred basepairs (Lee et al., 2005). In contrast, in our study motifs discovered in the CRC promoter are more refined (Figure 4). The analysis of recently diverged sequences in our analysis is far less likely to result in extended motifs compared to phylogenetic footprinting. Promoters between different OPGs of one PAG are functionally related by co-expression but not by evolutionary relationship. Thus for evolutionary unrelated OPGs from one PAG sequence similarity beyond shared cis-elements is expected to be equal to background promoter similarity. Delimited motifs consequently emerge from extended alignments after comparing profiles from different OPGs. This powerful feature of PhyloCon has already been demonstrated in a study of four closely related yeast species (Stormo and Wang, 2003). In our study, mean and median sizes of motifs detected are in good agreement with sizes of known transcription factor binding sites and indicate the detection of well delineated elements.

We investigated for enrichments within functional categories by making use of Gene Ontology (GOSlim) and the KEGG biochemical pathways annotations for the respective *Arabidopsis* genes. Many profiles detected are enriched in a wide range of biological functional categories involving metabolism (e.g. gluconeogenesis), development (flower development), signaling (abscisic and
giberellic acid signaling) as well as cell maintenance tasks like ribosome biogenesis. Applying the "guilt-by-association" rule, the occurrence of particular profiles or the functional enrichment within particular CEGs may assist to transfer knowledge to genes of yet unknown functions. To assist in this task, we implemented a database and a web portal providing structured access to all results of this study (http://mips.gsf.de/proj/plant/webapp/expressionDB/index.jsp). In addition, listings of all CEGs and candidate sites for all genes are provided as supplementary material.

We analyzed to which extent known Arabidopsis and plant cis-elements present within the AGRIS and PLACE databases overlap to the motifs detected within our analysis. We compared all cis-element entries present within the two databases with the profiles resulting from our analysis. We successfully detected 255 out of 537 elements present within the databases. Limiting factors in this analysis are the incomplete Brassica genome and the partial coverage of many Arabidopsis promoters by corresponding orthologous Brassica GSS contigs. An additional limitation is the partial coverage of the Arabidopsis transcriptome by Affymetrix GeneChips (21559 out of 26535 genes in MATDB). Given these restrictions, the successful detection of 47.5% of described cis-elements from PLACE and AGRIS can be viewed as highly satisfactory. However, due to the incomplete data set as well as some limitations of motifs within the databases (e.g. single site reports, consensus sequences of transcription factor families, see Results), an exact global assessment of specificity and sensitivity for our results is not feasible.

Two examples that we have studied in detail illustrate the correlation of sites detected in our analysis with regulatory elements involved in transcriptional regulation. CRABS CLAW (CRC) co-clustered with several floral development genes that have been shown to interact with CRC (Lee et al., 2005). Five enhancers required for correct CRC expression have been reported (Lee et al., 2005). As shown in Figure 4, detected elements are preferentially located within these regions and a CArG box confirmed by site-directed mutagenesis is covered by one of the cis-elements detected in our analysis. The trans-cinnamate 4-
monooxygenase (C4H) gene is involved in the phenylpropanoid pathway. Tight co-expression of PAL1, C4H and 4CLs has been observed for both *Arabidopsis* and poplar (Mizutani et al., 1997; Hertzberg et al., 2001). Promoter elements for C4H have been characterized and it has been shown that C4H contains 4 P-boxes, 3 A-boxes and 2 L-boxes (Figure 5b) (Mizutani et al., 1997). Eight out of nine sites for the gene were identified. It is noteworthy that even the borders of several elements, a well-known problem in *cis*-element detection via phylogenetic alignments, were very well approximated.

In our work we demonstrate that complex questions in comparative genomics can be addressed by using fragmented genome information and an integrative analytical approach i.e. the combination of expression data with comparative sequence analysis and phylogenetic footprinting. Our analysis uses the comparison of a full and a partial genome sequence. The approach can be extended to additional partial or complete genomes to enhance the support for and the refinement of discovered motifs. The simultaneous analysis of several partial genomes, however, would decrease the number of OPGs available for the analysis, as best candidate orthologs have to be detected in multiple partial sequence sets. For instance, for the analysis of two genomes with a coverage of one quarter each, on average one would expect 1/16\(^{th}\) of candidates to be present in both sets. Instead of a simultaneous analysis partial genomes may be sequentially subjected to a comparison against a complete genome. Subsequent processing and merging of results derived from pairwise comparisons would lead to a more comprehensive *cis* element catalogue.

Full genome sequences are labor- and cost-intensive and high quality genome projects are expected in the near future for only a few model organisms and economically important species. Large-scale expression data will underlie similar limitations. Genome scale comparative genomics would thus have to rely on a few species that may be separated by large evolutionary distances restricting the scope of comparative analyses. In plants, this problem is particularly accentuated as up to now only two genomes, rice and *Arabidopsis*, have been analyzed.
extensively (Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005). On the other hand, shotgun GSSs can be produced at a fraction of the costs and provide a rapid and promising alternative for comparative genomics of closely related species. Besides being useful for gene discovery, these resources will be of importance for the detection of conserved genomic features beyond the genes and will allow the elucidation of additional functional elements like promoter elements and architectures.
Materials and Methods

Brassica oleracea dataset

Sequences were retrieved from NCBI selecting for the keyword ‘Brassica oleracea’ in the field ‘Organism’. The vast majority of the sequences represent genomic survey sequences (GSSs) of Brassica oleracea deposited by a sequencing consortium of The Institute of Genomic Research (TIGR), the Cold Spring Harbor Laboratories and Washington University. A total of 567,985 sequences (567,365 GSS) were obtained. The 567,365 GSS sequence reads represented approximately 384 Mb of sequence. A rigid clustering regime using the Harvester assembly pipeline (BIOMAX Informatics, AG, Martinsried, Germany) was applied. The assembly method of Harvester is based on the CAP3 program (Huang and Madan, 1999). Default settings have been applied. This step also removed a large number of clones containing repetitive sequences. The remaining 190,513 GSSs defined 142,489 assemblies with an average size of 987 bp, totaling 140.6 Mb of non-redundant genomic sequences of Brassica oleracea.

Determination of candidate orthologous upstream sequences

Individual Arabidopsis thaliana upstream sequences were selected from the genomic sequence. Sequences were delimited either by the 5' neighboring gene or a maximum size of 3kb (excluding the 5'-untranslated region). Since 5'-untranslated regions (5'-UTRs) may harbor motifs or signals relevant to the transcriptional activity of a gene, 5'-UTR sequences were included in the analysis. To identify best available candidates for orthologous promoter regions between partial genome information of Brassica oleracea and the complete Arabidopsis thaliana genome sequence, a bidirectional best Blast hit (bbh) analysis strategy was applied. Upstream Arabidopsis sequences were compared against the Brassica GSS assemblies by BlastN (E ≤ 1^{-10}) while the Brassica GSSs were compared to the whole genome sequence of Arabidopsis. The
genomic position of the highest scoring BlastN hit of a *Brassica* GSS had to correspond to the location of the original upstream sequence. This modified reciprocal best hit criterion was used to group the candidate orthologous *Arabidopsis* and *Brassica* sequences.

**Microarray transcriptomics datasets**

*Arabidopsis thaliana* genome scale expression data have become available from a variety of microarray platforms. Among them are several cDNA arrays, both commercial and custom-made, as well as two Affymetrix oligonucleotide GeneChips (http://www.affymetrix.com/products/arrays/index.affx?Arabidopsis). However, it is well known that comparisons among different platforms are problematic (The Toxicogenomics Research Consortium, 2005). Hence we made use of measurements from a single platform only, the Affymetrix ATH1 GeneChip.

Experiments available from NASC (http://nasc.nott.ac.uk, CD-ROM release as of November 2004, (Craigon et al., 2004)) and AtGenExpress (Schmid et al., 2005; kindly provided by Markus Schmid and Detlev Weigel) have been used in this study.

**Mapping of the Affymetrix probe sets onto the *Arabidopsis* genome**

Due to annotation updates and enhanced gene modeling, GeneChip oligonucleotide mapping is frequently erroneous and outdated. Therefore, probe sets were re-calculated using an enhanced oligonucleotide mapping against the *Arabidopsis* genome template.

All oligonucleotides present on the ATH1 GeneChip of Affymetrix (sequences downloaded from www.affymetrix.com as of October 2004) were realigned against the coding sequence and, for genes with associated full length cDNA information, against the UTR sequences (MAtDB release from 24th of September, 2004, Schoof et al., 2004).
Oligonucleotides aligning to more than one gene and probes without perfect matches were excluded. For subsequent calculations, only probe sets with at least five probe pairs were considered. Most of the probe sets still consist out of 9 to 11 probe pairs. Four percent of the probes matched perfectly to at least two genes and led to partial unspecific estimates for 10% of the original probe sets indicating the need for the re-alignments. We excluded those probe sets from our refined sets. In summary expression measurements from 21,559 genes met the quality criteria and were used for subsequent analyses.

Statistical Processing

The statistical analysis of the expression data was carried out in R (R Development Core Team, 2004) using the FunDaMiner system (http://mips.gsf.de/proj/express). We calculated MAS 5.0, dChip (Li and Wong, 2001) and RMA (Bolstad et al., 2003) probe set summaries according to the redefined probe sets for every experiment. All probe set summary data were transformed to log-scale (basis 2). The complete dataset was normalized by applying the LMPN method. LMPN is based on the local polynomial regression fitting method loess (Cleveland, 1979; Cleveland et al., 1992) operating on MA-scale (Dudoit et al., 2000). Non-linear normalizations like the polynomial regression method are required if datasets consist of experiments from various researchers employing e.g. different RNA extraction protocols like in our study. Ignoring this would introduce additional, technical rather than biological, correlative components. Replicates (usually around three) were summarized by the mean leading to more robust estimates of the real expression level. The correlation coefficients were computed for mismatch-corrected MAS 5.0 summaries.

Correlation Matrices and Distribution of Correlation Coefficients

For 779 measurements, i.e. microarray experiments, we computed the correlation matrix of all-against all probe sets. The full matrix consists of about
4.65x10^8 (21,559^2) correlation pairs. Correlations were determined as metric (Pearson) correlation coefficients. The full correlation matrix (except self-correlations) served as background and the 99%-quantile has been derived from this distribution (r=0.803). Correlations with a correlation coefficient higher than the 99%-quantile of the background distribution, analogous to a one sided 1% significance level, were considered as relevant.

**Extraction of Co-expressed Groups**

For each of the 21,559 genes, its co-expression group (CEG) was defined as those genes showing a Pearson correlation r ≥ 0.803. By this definition, one gene may be associated with multiple groups. By applying the above cut-off for correlations, we obtained 13,254 CEGs while 8,305 genes remained as singletons.

**PhyloCon Analysis**

PhyloCon was downloaded from http://ural.wustl.edu/~twang/Phylocon/ (Wang and Stormo, 2003).

A common problem in motif discovery is the degree of noise in the selected set of genes. One reason is that co-expression does not necessarily result from co-regulation as co-expression of two genes can be attributed to secondary effects like e.g. transcription factor cascades. Measurement errors, cross hybridization, biological variation and erroneous annotations are additional sources of noise. In this study, the incomplete and fragmented genome of *B. oleracea* represents an additional difficulty. With an average length of about 1 kb for the GSS assemblies, and about 1.8 kb for the *Arabidopsis* upstream sequences, orthologous information is available on average only for approximately 55% of the promoter region. Thus, even if all studied promoters of one specific co-expression group contain a conserved binding site, approximately half of the phylogenetic comparisons will on average fail to detect it as alignments cannot cover the conserved site. However, PhyloCon computes candidate matrices in a
step-wise manner. Starting from pairwise orthologous promoter group (OPG) alignments it sequentially adds OPGs to the alignments from which new matrices are built. Importantly, PhyloCon allows for a report of intermediate matrices, i.e. matrices derived from preceding cycles.

To overcome the limitations of missing sites or noisy expression groups, we retrieved a maximum of 10 intermediate matrices per cycle. As additional parameters we allowed for 200 temporary (or test) matrices per cycle and the number of standard deviations was set to 0.5 (for details, see Wang and Stormo, 2003). Both strands were analyzed.

**Analysis of Profiles and Conserved Sites**

Primary profiles were filtered for (CT)$_n$- and (GA)$_n$-repeats, as these repeats are very prominent on *Arabidopsis* promoters. Alignment matrices reported by PhyloCon were transformed into position weight matrices (PWMs) to generate a scoring function for sequence instances. For an alignment matrix of length $m$ we determined the number of occurrences $n_{ij}$ of the four possible nucleotides $i \in \{A,C,G,T\}$ in each column $j = 1,2,\ldots,m$. The following formula has been applied to transform this $(4 \times m)$-count/frequency matrix into a $(4 \times m)$-PWM (Hertz and Stormo, 1999):

$$a_{ij} = \log \left( \frac{n_{ij} + p_i}{(N + 1)} \frac{N}{p_i} \right),$$

where $N$ is the number of instances in the alignment, $p_i$ the *Arabidopsis* background probability of nucleotide $i$ and $n_{ij}$ the counts of nucleotide $i$ at position $j$ in the alignment of the instances found by PhyloCon. A single cell $a_{ij}$ of a $(4 \times m)$-PWM $\mathbf{A}$ is the respective score for nucleotide $i$ at position $j$. Let $S=s_1,s_2,\ldots,s_m$ be a sequence of length $m$ with each $s_j=1,\ldots,m$ representing a (particular) letter in the alphabet. A PWM $\mathbf{A}$ assigns a score to sequence $S$ by summing up all $j=1,2,\ldots,m$ single cell scores $a_{ij}$ where $i$ corresponds to letter $s_j$. A sequence $S$ is considered an instance of a PWM if its score exceeds a threshold score. This
cut-off score is specific for each PWM. It is derived from the instance in the original PhyloCon alignment matrix with the lowest score. For each motif we determined all occurrences in the *Arabidopsis* upstream sequences. Based on this mapping we derived expected and observed frequencies of motif occurrences.

For each PWM we tested the statistical significance of its overrepresentation within the respective co-expression group (CEG) in comparison to all 21,559 *Arabidopsis* upstream sequences (see Materials and Methods). P-values were obtained from the cumulative binomial distribution and a PhyloCon PWM was considered to be significantly overrepresented for $p \leq 0.01$.

To identify identical and almost identical profiles, profiles with a Pearson correlation coefficient $r \geq 0.98$ were grouped into clusters. Overlapping clusters were recursively merged. To determine the similarity between profiles, we used the Pearson correlation coefficient (PCC) between two columns of a profile as described in Schones et al., 2005. To compare multiple columns of two profiles, the scores of each column comparison were summed up and normalized against the number of compared columns. For motifs of different lengths, we compared the shorter profile against longer profiles applying sliding windows. For these cases, the window with the highest PCC has been considered for the comparison.

Note that we did not merge any redundant profiles/PWMs, i.e. recompute a new alignment matrix derived from the merged profiles. Although this approach results in a significant redundancy we thereby avoid any flaws by for example low-quality matrices that potentially strongly alter the specificity of a merged profile. As a consequence, one would need to re-assess findings for the merged matrices, e.g. enrichments of particular profiles in CEGs and functional categories, which have been obtained from the more specific individual profiles.

To derive an estimate about the number of sites detected in the PAG analysis, we merged detected instances/sites (not profiles!) in each gene if in the promoter two instances overlap by more than 90%.
Known Motif Matches

To compare detected profiles with reported sites, all motifs listed in the AGRIS (http://arabidopsis.med.otheo-state.edu/AtcisDB/bindingSiteContent.jsp) and PLACE (http://www.dna.affrc.go.jp/PLACE) databases were downloaded (Higo et al., 1999; Davuluri et al., 2003). PLACE and AGRIS derived motifs are given as consensus sequences whereas profiles obtained during the course of our analysis are present as position weight matrices. Thus a direct comparison was not feasible. Instead, we decomposed consensus sequences of known motifs into all exact sequences by replacing degenerated IUPAC letters into their respective bases. These sequences can then be tested for their probability to constitute an instance of a particular profile (see above). In analogy to the approach used for the determination of profile similarity, we applied sliding windows as described above.

Functional Categories

GOSlim annotation for Arabidopsis and the KEGG pathway map has been obtained from TAIR (www.arabidopsis.org). All functional categories containing only one member have been excluded from subsequent analysis. Gene lists of categories were matched with the 21,559 genes used in this study. For each profile, we selected the genes containing the respective profile within their upstream sequences. Overrepresentation of the profile in a functional category was consequently checked for each GO annotation associated with the selected genes. P-values for each test were obtained by cumulative binomial probability.

\[
P(k \geq x) = \sum_{k=x}^{n} \binom{n}{k} p^k (1 - p)^{n-k},
\]

where \( n \) is the number of all studied genes associated with a specific GO annotation, \( x \) is the number of observed genes associated with this GO annotation and containing the profile, and \( p \) is the expected frequency of the
profile, i.e. the number of promoters containing the profile divided by the number of all studied genes. Profiles present in only one GO annotation were not considered (x>1) as no reliable statistics can be computed for only one occurrence. Multiple testing corrections were performed by multiplication of the P-value with the total number of assayed GO annotations for each profile. For the KEGG pathways, we employed a similar binomial testing scheme. P-values were corrected for multiple testing by the number of different KEGG pathways.

Acknowledgment

We thank Markus Schmid and Detlev Weigel for providing us microarray data from the AtGenExpress, Chris D. Town from TIGR for making the Brassica oleracea GSS dataset available to us prior to publication. The authors also wish to thank Louise Gregory for helpful discussions. This work was funded by the GABI program of the German Ministry of Education and Research (BMBF).
Literature


Katari, MS, Balija, V, Wilson, RK, Martienssen, RA, and McCombie, WR (2005) Comparing low coverage random shotgun sequence data from Brassica oleracea and Oryza sativa genome sequence for their ability to add to the annotation of Arabidopsis thaliana. Genome Res 15: 496-504
Figure Legends

Figure 1. Analysis Schema
Two sources of evidences, co-expression and conservation, were retrieved for a motif search in thousands of *Arabidopsis* upstream sequences. Expression data from 779 Affymetrix microarrays were normalized and an all-against all correlation matrix was computed. For each gene, correlated genes form a CEG (co-expression group). *Brassica* genomic survey sequences (GSSs) were assembled and filtered for repeats. *Brassica* GSSs are depicted as green lines and *Arabidopsis* upstream sequences as blue lines. Orthologous sequences were derived by a bidirectional best blast hit strategy. All sequences of one CEG with associated orthology information represent a PhyloCon analysis group (PAG). A PAG therefore consists of at least two orthologous promoter pairs (OPG) constituted by one *Arabidopsis* and one *Brassica* sequence. Each PAG was subjected to a motif search and all profiles, including intermediate profiles, were retrieved. Stepwise profile generation by PhyloCon was used to overcome potential shortcomings of the motif discovery procedure, in particular noisy data and partial coverage (see Figure 3B). In an (hypothetical) example PAG, a functional motif is shown as red box and it is present in three of four OPGs while OPG 4 lacks it despite being co-expressed. A combination of OPG4 with any other of the three OPGs containing the motif will result in low quality profiles as marked by a question mark within the matrix. The motif in OPG3 is not covered by a (partial) GSS and any OPG combination may fail to detect the motif. However, the motif will be detected within the analysis of OPG1 and 2. As intermediate profiles were kept, the successive statistical evaluation identifies the profile as significantly overrepresented within the respective CEG while low quality profiles are excluded. In addition, its occurrence in OPG3 is uncovered by a genome wide search for profile instances. Finally, candidate transcription factor binding sites are listed for each gene.
Figure 2. Distribution of Pearson Correlation Coefficients
A correlation matrix of all-against all probe sets was computed and correlations – excluding self-correlations – were binned into 100 bins of width 0.02. Bins are shown on the x-axis and their relative frequency on the y-axis. The dashed vertical line indicates the 99%-quantile of the resulting distribution (r=0.803).

Figure 3. Details of the analysis
(A) gives an example how co-expression groups (CEGs) are defined for this analysis. Two (hypothetical) CEGs and their relationship are shown. Letters in circles stand for individual genes, lines indicate a significant correlation r ≥ 0.803, the 99% quantile. Anchor genes of the two CEGs are gene A and B, respectively. Note that the two CEGs share gene A, B and C; however, there are genes (X and Y) that are significantly correlated to only one of the expression profiles of an anchor gene. This results in overlapping but not identical CEGs.
(B) illustrates the step-wise generation of matrices in a PhyloCon analysis of a particular PAG. Profiles are shown as multiple alignments of several OPGs (orthologous promoter groups) and each OPG consists of one Arabidopsis-Brassica promoter pair. Arabidopsis sequences are indicated by ‘At’, Brassica sequences by ‘Bo’. Alignments are transformed to position weight matrices (see Materials and Methods) that are depicted here as sequence logos. Starting from aligned blocks of all pairwise OPG alignments, i.e. four promoter sequences (not shown), new matrices are generated by stepwise combinations of aligned blocks thereby refining borders and composition of the profiles. Intermediate matrices can be retrieved at each combining step.

Figure 4. Sites within the promoter of CRABS CLAW
At the top of the figure, a scheme of the CRABS CLAW (CRC) promoter is shown. Five previously reported enhancer regions (gray boxes A-E according to Lee et al., 2005) are depicted. Motifs detected in our analysis are displayed as colored boxes. Green, red and cyan boxes designate motifs found in flower developmental genes. The respective genes are shown at the bottom left (AP1, AP3, SEP1 and SEP2) including At1g72290 that encodes an endopeptidase
inhibitor co-occurring with these genes. Blue boxes are specific for genes involved in sporocyte formation (MS2, SPL) or expression patterns implying a possible role in this biological process (ATA1). Black boxes in the CRC promoter are profiles detected by PhyloCon but are either only sporadically present or completely absent within the shown co-expressed promoters.

**Figure 5. The Cinnamate 4-Hydroxylase (C4H) Promoter Analysis Group**

Figure 5A shows a (simplified) version of the core phenylpropanoid biosynthesis pathway leading to the synthesis of lignin monomers and anthocyanin pigments (not shown). Enzymes present in one PAG are written next to their reactions: PAL (phenylalanine ammonia-lyase), 4CL (4-coumarate:CoA ligase), C4H (cinnamate 4-hydroxylase), CCoAOMT (caffeoyl-CoA-0-methyltransferase). Figure 5B shows the upstream region of cinnamate 4-hydroxylase (C4H), known motifs are displayed in bold text with the motif name under the site. Asterisks denote detected sites by PhyloCon. Note that P- and L-boxes partially overlap, that both boxes display a high degeneracy (P-box consensus: YTYYMMCMMAMCMMC; L-box consensus: YCYYACCWACC) and are thus particularly difficult to detect.

**Table 1. Known motifs in PLACE and AGRIS match discovered profiles**

Left column shows sequence logos of profile discovered in this study, middle columns lists name/identifier and right column consensus sequence in PLACE/AGRIS for the motif matching the detected profile. This table represents a subset of motifs that match PLACE and AGRIS motifs.

**Table 2. Detected profiles enrich for GOSlim functional categories and KEGG biochemical pathways**

Second column lists sequence logo representation of profiles discovered in this study that are statistically overrepresented in a functional category/KEGG pathway depicted in first column. Following columns show (i) the number of genes within the respective category which contain the motif in their promoters, (ii) the number of genes in the genome annotated for the category and (iii) the
Bonferroni-corrected probability that the enrichments is obtained by chance. Example results are shown for GOSlim annotations in the upper part, for KEGG pathways in the part below the dashed line. The enriched motif for category ‘response to abscisic acid stimulus’ has strong similarity to the known ABA response element (ABRE), the motif for category ‘phenylpropanoid biosynthesis’ is similar to an overlapping P- and L-Box present in phenylpropanoid biosynthesis genes (Fig. 5) and the motif for ‘trehalose biosynthesis’ contains the core G-box.

**Figure S1. Size Distribution of CEG derived Profiles and Sites**
Relative frequencies and size distribution [bp] of profiles, non-fused (‘simple’) and fused (‘nonred’) sites are shown. For details about site fusions, see Materials & Methods.

**Figure S2. Size Distribution of Co-expression Groups (CEGs)**
Large figure shows binned distribution of CEG sizes. CEGs were grouped into bins of size 25. Number of singletons is indicated (‘1’ on the x-axis). Figure shows only the first 40 bins, distribution asymptotically decreases to zero. Inlet figure illustrates number CEG sizes ranging from 2 to 39 members.
<table>
<thead>
<tr>
<th>Detected Motif</th>
<th>Motif identifier in PLACE/Agris database</th>
<th>Consensus sequence in known motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCACGTGG</td>
<td>GBF1/2/3 BS in ADH1</td>
<td>CCTACGTGGCGGG</td>
</tr>
<tr>
<td>ccGACA</td>
<td>LTREC promoter element</td>
<td>ACCGACA</td>
</tr>
<tr>
<td>TAAATGCACC</td>
<td>L1BOXATPDF1</td>
<td>TAAATGYA</td>
</tr>
<tr>
<td>TTGGCC</td>
<td>E2FCONSENSUS</td>
<td>WTSSCSS</td>
</tr>
<tr>
<td>ccGACA</td>
<td>DRECRTCOREAT</td>
<td>RCGAC</td>
</tr>
<tr>
<td>TGGGCC</td>
<td>SITEIATCYTC</td>
<td>TGGGCY</td>
</tr>
<tr>
<td>TCCCCA</td>
<td>PALBOXPPC</td>
<td>YTTYMMCMAMCMMMC</td>
</tr>
<tr>
<td>TACCCAC</td>
<td>MYBPLANT</td>
<td>MACCWAMC</td>
</tr>
<tr>
<td>TCT TCT TCT</td>
<td>TCA1MOTIF</td>
<td>TCATCTTCTTT</td>
</tr>
<tr>
<td>T TCT TCTTC</td>
<td>TL1ATSAR</td>
<td>GGTTCAG</td>
</tr>
<tr>
<td>AAAAAACGTG</td>
<td>T/GBOXATPIN2</td>
<td>AACGTG</td>
</tr>
<tr>
<td>GO/KEGG description</td>
<td>Detected Motif</td>
<td>Genes in category</td>
</tr>
<tr>
<td>--------------------------------------------------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>ribosome biogenesis</td>
<td>TTTTTTTTTTTTTT</td>
<td>94</td>
</tr>
<tr>
<td>polarity specification of adaxial/abaxial axis</td>
<td>TCTTCTCCATT</td>
<td>6</td>
</tr>
<tr>
<td>dolichol biosynthesis</td>
<td>CAGAGACAATTTA</td>
<td>5</td>
</tr>
<tr>
<td>response to osmotic stress</td>
<td>CTTCTCTACCCTC</td>
<td>12</td>
</tr>
<tr>
<td>regulation of transcription</td>
<td>GGTCCAAC</td>
<td>483</td>
</tr>
<tr>
<td>response to auxin stimulus</td>
<td>CATACTGTCG</td>
<td>137</td>
</tr>
<tr>
<td>response to red or far red light</td>
<td>GAGGAGAGAGAGAGAA</td>
<td>19</td>
</tr>
<tr>
<td>response to abscisic acid stimulus</td>
<td>GCCTGAGG</td>
<td>49</td>
</tr>
<tr>
<td>gravitropism</td>
<td>CGTCAGGGGG</td>
<td>6</td>
</tr>
<tr>
<td>sugar mediated signaling</td>
<td>GAGAGAGAGAGAGA</td>
<td>14</td>
</tr>
<tr>
<td>cytidine metabolism</td>
<td>CTCTTCTCAAC</td>
<td>9</td>
</tr>
<tr>
<td>mitochondrial inner membrane transport</td>
<td>TAGGGGGGGGGGG</td>
<td>11</td>
</tr>
<tr>
<td>giberellic acid mediated signaling</td>
<td>CTCAAAGG</td>
<td>18</td>
</tr>
<tr>
<td>carpel development</td>
<td>AAAAGAAGGGG</td>
<td>5</td>
</tr>
<tr>
<td>flower development</td>
<td>TGGCCACCA</td>
<td>24</td>
</tr>
<tr>
<td>microtubule-based movement</td>
<td>TCTCTCTCTCTCTCTCT</td>
<td>66</td>
</tr>
<tr>
<td>phenylpropanoid biosynthesis</td>
<td>ACCTACCA</td>
<td>7</td>
</tr>
<tr>
<td>trehalose biosynthesis</td>
<td>CCACTGTCGC</td>
<td>22</td>
</tr>
<tr>
<td>cellulose biosynthesis</td>
<td>CACGCAATG</td>
<td>42</td>
</tr>
<tr>
<td>gluconeogenesis</td>
<td>GGGCCA</td>
<td>54</td>
</tr>
<tr>
<td>giberellin biosynthesis</td>
<td>ACGTGIGGT</td>
<td>13</td>
</tr>
</tbody>
</table>