

**Detection and evolutionary analysis of functional conserved
non-coding sequences in higher plants by integrating evolutionary
and functional conservation**

Xi Wang

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Dr. J. Durner

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H.-W. Mewes
2. Univ.-Prof. Dr. J. Parsch
(Ludwig Maximilian Universität München)

Die Dissertation wurde am 08.02.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 22.06.2012 angenommen.

Content

1. Introduction

1.1 Transcriptional regulation: transcription factor and promoter architecture

1.2 Cis-regulatory element: representation and detection

1.2.1 Representing cis-regulatory elements

1.2.2 Discovery cis-regulatory elements

1.2.2.1 Experimental approaches

1.2.2.2 Computational approaches

1.2.2.2.1 Detection of known sites

1.2.2.2.2 Detection of unknown sites

1.3 Evolution of cis-regulatory element

1.4 Goals of this work

2. Results

2.1 Detection and evaluation of cis-regulatory elements in higher plants

2.1.1 Cis-regulatory element detection in rice and sorghum by combination of co-expression and evolutionary conservation

2.1.1.1 Preparation of genomic sequence datasets

2.1.1.2 Rice expression data processing

2.1.1.3 Determination of co-expressed groups

2.1.1.4 *PhyloCon* motif discovery and analysis

2.1.1.5 Validation of *PhyloCon* detected motifs

2.1.2 Cis-regulatory element detection in rice and sorghum by “global” evolutionary conservation

2.1.2.1 Motif detection by network-level conservation: implementation

2.1.2.2 Detection of dyadic motifs

2.1.2.3 Validation of detected motifs

2.1.3 Cis-regulatory element detection in *Arabidopsis thaliana* by „mutual“ information among co-expressed gene groups

2.1.3.1 „Mutual“ information and its application for cis-element detection by *FIRE*

2.1.3.2 Cis-regulatory element detection in *Arabidopsis Thaliana* using *FIRE*

2.1.3.3 Validation of detected motifs

2.1.3.4 Evolutionary conservation of functional candidate motifs detected by *FIRE*

2.2 Evolutionary analysis of cis-regulatory elements in plants

2.2.1 Evolution of cis-regulatory elements during duplication and speciation of *A. thaliana* and *A. lyrata*

2.2.1.1 Identification of orthologous and paralogous genes in *A. thaliana* and *A. lyrata*

- 2.2.1.2 Cis-element conservation during speciation and gene duplication
- 2.2.1.3 Evolution of cis-elements in complex paralogue-ortholog gene networks
- 2.2.1.4 Summary
- 2.2.2 Identification of genetic determinants for genome-plastome incompatibility of *Oenothera*
 - 2.2.2.1 Compartmentalized genetic system and co-evolution of intracellular genetic compartments
 - 2.2.2.2 *Oenothera* and its diverse genome-plastome compatibilities
 - 2.2.2.3 Search for putative genetic determinants responsible for PGI in *Oenothera*
 - 2.2.2.4 ClpP-psbB intergenic region is an AB-I incompatibility determinant
 - 2.2.2.5 Summary

3 Discussion

3.1 Discovery of cis-regulatory elements in higher plants

- 3.1.1 Motivation and approaches
- 3.1.2 Detected cis-regulatory motifs and sites
- 3.1.3 Evaluation of detected cis-regulatory elements
- 3.1.4 Summary and future work

3.2 Evolutionary Analysis of cis-regulatory elements in higher plants

- 3.2.1 Evolution of cis-regulatory elements in *A. thaliana* and *A. lyrata*
- 3.2.2 Search for genetic determinants causal to PGIs in *Oenothera*

4 Summary & outlook

5 Material and methods

- 5.1 Array lists included in each data set
- 5.2 Mapping probes of expression data to current annotation in PhloCon analysis
- 5.3 Expression data processing and filtering in PhyloCon analysis
- 5.4 Determination of co-expression gene groups in PhyloCon analysis
- 5.5 PhyloCon motif discovery and analysis
- 5.6 Validation of PhyloCon motifs
- 5.7 Motif detection in rice and sorghum by network-level conservation
- 5.8 Dyadic motif detection by network-level conservation
- 5.9 Motif detection by FIRE analysis
- 5.10 Motif detection between *A. lyrata* and *A. thaliana* by network-level conservation
- 5.11 Determination of tandem and segmental duplication in *A. thaliana* and *A. lyrata*
- 5.12 Determination of complete paralogue-ortholog gene networks
- 5.13 The flowering plant genus *Oenothera* subsection (*Eu*)*oenothera*

- 6 Reference**
- 7 Appendixes**
- 8 Publications**
- 9 Acknowledgments**

1. Introduction

1.1 Transcriptional regulation: transcription factor and promoter architecture

In higher organisms, large portion of cell functional attributes contributing establishment of a master body plan, cell differentiation and proper response to a changing environment is determined by tight regulation of gene and protein activities (Maeda et al 2011). Such regulation can be achieved at any steps from a gene to production of its active protein. Mechanisms include epigenetic changes by e.g. DNA methylation and chromatin remodeling, post-transcriptional controls like mRNA degradation and stability via miRNA, or post-transcriptional modifications of protein by e.g. membrane anchoring or phosphorylation (Gräff et al, 2010). However, one of the first and essential regulations is control of transcription, a process from a gene to its mRNA product. Such transcriptional regulation is realized by binding of proteins, the transcription factors, to particular genomic regions, the promoters.

Transcription factors bind to either enhancer or promoter regions of DNA adjacent to the genes they regulate. In the human genome, approximately 10% of genes in the genome code for transcription factors which makes this family the largest component in the human proteome (Babu et al, 2004). In higher plants like *Arabidopsis thaliana*, more than 5% of the genes encode for transcription factors (Riechmann et al, 2000). The large number of transcription factors in living organisms underpins the relevance of tight and transcriptional regulation at various levels. Transcription factors are modular in structure and eventually contain three different domains: 1) Trans-activating domain which contains binding sites for other proteins such as transcription co-regulators, 2) DNA-binding domain which attach to specific sequences (often referred as cis-regulatory elements or response elements) of promoter or enhancer adjacent to regulated genes, and 3) an optional signal sensing domain which senses external signals and transmit them to the rest of the transcription components. Transcription factors typically attach to the gene loci that don't encode them and are thus acting in *trans* (Latchman, 1997). Nevertheless auto-regulatory loops, that is regulation of transcription by the transcription factor protein encoded by the locus is frequently observed (Salgado et al, 2001).

Promoters, on the other hand, are regions bound by *trans*-acting transcription factors and usually close to the transcription initiation site of the regulated genes. Hence, they are often called to be *cis*-active. Promoters contain specific DNA sequences and response elements that accomplish attachment of transcription factors in a sequence-specific manner. Such binding elements are mainly interspersed in three hierarchical-structured promoter parts: core promoter, proximal and optional distal promoter:

- The core promoter surrounds transcription start sites and has a small size of approximately 100 bp or less (Smale et al, 2003). Assembly of the pre-initiation complex at the core promoter, including general transcription factors, co-factors

and a DNA-dependent RNA polymerase is the first step of transcription. Most knowledge how core promoters interact with the general transcriptional machinery is derived from studies in yeast, fly and human (Thomas et al, 2006). In bacteria, the core promoter consists of two short sequences at -10 and -35 positions upstream from transcription start site which are recognized by RNA polymerase and an associated sigma factor. In animals, at least seven different sub-sequences including transcription start site, TATA box, downstream core element have been identified in the core promoters and are necessary for general transcription machinery (Thomas et al, 2006). Higher plants accomplish the same general transcription machinery as animals (Thomas et al, 2006). However, important differences in core promoter structures have been revealed thanks to several large-scale studies. For the majority of *Arabidopsis thaliana* genes, for example, alternative initiation sites have been discussed (Seki 2002; Yamamoto 2009; Tanaka 2009). Furthermore, although sequences as well as distance and orientation of the TATA-box to the transcription start site seems to be well conserved between both kingdoms, higher plants apparently lack most other functional sub-elements including downstream promoter and downstream core elements (Molina et al, 2005; Yamamoto et al, 2007). Instead of these elements, two additional plant-specific core promoter elements, Y-patch and GA-element, have been identified by the local distribution of short sequences (LDSS) profiles investigated in a recent study (Yamamoto et al, 2007; 2009).

- Proximal promoters locate 5'-adjacent to the core promoter. In contrast, the position of distal promoters is less defined. They can locate up to several thousand base pairs upstream of transcription start site, and can even be located in introns and downstream regions of the respective gene (Banerji et al, 1981; Carter et al, 2002; Vokes et al, 2007). Compared to core promoters which assemble the pre-initiation complex, proximal and distal promoters contain regulatory elements attached by sequence-specific transcription factors. Though required, the interaction of pre-initiation complex with core promoter in general realizes a low-level basal transcription rate. Only the cooperation of transcription factors attaching to proximal and distal promoters achieves a tight modulation of the basal transcription rate and specific spatiotemporal expression patterns (Seipel et al 1992).

The mechanism of transcription regulation can be regarded as interpretation of genetic “blueprint” of promoters by diverse recruited transcription factors and co-factors. On one hand, diverse cis-acting element subsets of a promoter can associate with different transcription factor complexes to activate the promoter. On the other hand, a single transcription factor can trigger different cellular responses in diverse promoter architectures dependent on the type of additionally recruited Transcription factors or co-factors. The combinatorial binding between diverse subsets of cis-regulatory elements and different recruited transcription factor complexes enables manifold cellular responses by limited transcription factors and promoter architectures.

1.2 *Cis*-regulatory element: representation and detection

Since the discovery of the lac operon and the realization that its expression was regulated by a protein factor (Jacob et al, 1961), a major objective in molecular biology has been to understand sequence-specific binding of transcription factors. *Cis*-regulatory elements, also known as response elements, are functional genomic sequences binding to transcription factors. They generally are assumed to be scarcely distributed in the proximal and distal promoter regions. As a complementary of core promoters, such regulatory sites play a crucial role to modulate levels and specify patterns of transcriptional gene activity by recruitment of specific transcription factors to their corresponding promoter parts (Seipel et al, 1992). Therefore, detection and characterization of *cis*-regulatory elements in various organisms have become one of the major tasks for biologists and bioinformaticians to understand the regulation mechanism of gene transcription.

1.2.1 Representation of *cis*-regulatory elements

Cis-regulatory elements are short DNA motifs of a size of 5-20 that are bound by transcription factors (TF) in a sequence-specific manner. Multiple *cis*-elements are usually clustered into *cis*-regulatory modules to achieve the transcription regulation under the cooperation of diverse sets of transcription factors and co-factors (Ben-Tabou et al, 2007). Two notable features of *cis*-regulatory element, variability and specificity, have been characterized. On one hand, binding of transcription factors tolerates in most cases a certain degree of sequence variability of binding sites. Transcription factors deduce different binding affinities by interacting diverse sequences. Regulatory systems can take advantage of the variability in the sites to better control transcription and accomplish the variability of gene expression. On the other hand, the majority of the genome comprises non-specific but weak affinity binding sites for particular TFs. Therefore, functional *cis*-regulatory elements that are bound by TFs with high affinity are a family of similar sequences. They must display a much higher binding affinity with corresponding proteins than non-specific binding sites. Such specificity of regulatory sites is essential for the regulatory system to work properly.

Several representations of *cis*-regulatory element have been developed to optimally estimate variability and specificity. More importantly, Representations of a given collection for known binding sites are employed to search new sequences and reliably predict additional DNA elements (see below). Among these models, exact words – a simple DNA sequence - apparently are the most simple motif representation (Stormo, 2000; fig. 1A). However, they can not capture any motif variability. Biologists, thus, typically use two other representations of *cis*-regulatory elements: the concept of consensus sequence and position-specific weight matrix. Given an alignment of several known sites, consensus sequence displays nucleotide variability within a motif in a qualitative manner using the full IUPAC alphabet (fig. 1B). In general it refers to

a sequence that matches all of the example sites closely, but not necessarily exactly. There is a trade-off between the number of mismatches allowed and the sensitivity as well as precision of the representation. Day and McMorris (1992) compared several methods for generating consensus sequences and outlined their strengths and weaknesses. Position-specific weight matrix, on the other hand, consists of probability, or “weight”, for each nucleotide alphabet at each motif position from a given alignment. The weights can be determined by information content or relative entropy, if occurrence of each base is adjusted by its background frequency in the genome (Schneider et al, 1986; Stormo, 1988; 1990; fig. 1C-D). Hence, in contrast to consensus sequence, position-specific weight matrix provides a statistical model with free parameters for the representation of cis-regulatory elements.

There are advantages and shortcomings for each of the both presentations. Consensus sequence can easily be captured from a collection of sites, and since the presentation has a well-defined search space, algorithms have been developed to efficiently and precisely search for new sites (Stormo, 2000). Moreover, it confers an information loss from the original data, as binding bias towards one of the possible nucleotides is

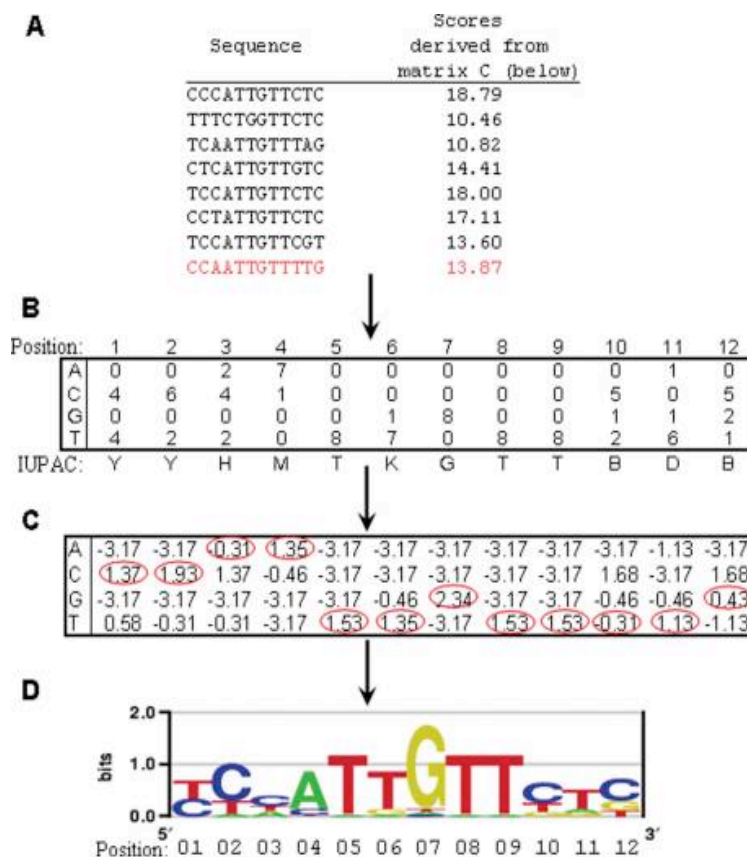


Figure 1: Representations of cis-regulatory element (Guhathakurta 2006)

A) Representation using exact words (left panel). B) Consensus using IUPAC alphabet. C) Position-specific weight matrix. D) Sequence logo representing position-specific weight matrix in C).

not reflected in the model. However, the disadvantages of such model are also obvious. Firstly, though variable positions allowed, such model possesses no statistical power and may fail to report functional variants of a motif. Secondly, the trade-off between the number of allowed variable positions and sensitivities/specificity to express a collection of known sites is hard to determine. Thirdly, the model is hardly able to quantitatively estimate variability and binding specificity of a binding motif. The representation of position-specific weight matrix, on the other hand, provides a quantitative description and can measure the degree of motif variability directly using the probabilities comprised in matrix. Furthermore, based on the experimental observation that the binding energy contributions are additive across the positions, binding specificity of a motif can be summarized from the information contents of all the weight matrix positions (Stormo, 2000) and is directly proportional to transcription factor binding affinity. Position-specific weight matrix representation, thus, can be viewed both as a statistical and as an energy-based model. Based on weight matrix, several methods have also been developed to efficiently calculate the distribution of scores which can be used to determine statistically significant DNA sequence matches (Staden, 1989; Claverie et al, 1996). However, the large number of free parameters for such statistical model can impede motif discovery. Moreover, in the frequent case of only a few known binding sites, a limited amount of available training sets for matrix calculation may lead to biased base frequency. Nevertheless, for representing known sites and discovering new sequences, the matrix model is in general both more sensitive and more precise than the consensus method (Stormo et al, 1982).

1.2.2 Discovering cis-regulatory elements

A large number of experimental and computational approaches have been designed for identification of cis-regulatory elements (Narlikar et al, 2009; Guhathakurta 2006). Nowadays, experimentalists frequently start a project by *in silico* analysis and by expanding an experimental observation into genome-wide predictions. Despite high accuracy of transcription factor binding site detection, these experiments are technically challenging and time consuming. Bioinformaticians, on the other hand, are well trained in developing computational tools for efficient large-scale identification of cis-elements. However, frequently observed high false-positive rate (see below) requires verification of binding site predictions and improvement of pipelines using experimental data. Therefore, the mixture of experimental and computational methods might be required and is powerful for an iterative refinement process.

1.2.2.1 Experimental approaches

Several techniques for experimental identification of transcription factor binding sites have been developed. They range in scope from localized, site-specific analyses to high-throughput assays that generate broad conclusions about binding site preferences and regulation of gene expression. Depending on the respective biological questions,

prior knowledge of transcription factors, promoter sequences and evidences of transcription factor/DNA interaction, experimental approaches for identification of functional elements can be summarized into two categories:

- 1) The methods lead to detect regulatory elements without direct measurement of transcription factor/DNA interactions. Such techniques include analysis of chromatin structure alterations and experimental manipulation of defined DNA segments. A typical example of the former approach is DNaseI hypersensitivity, which refers to genomic regions unbound by proteins and showing extreme sensitivity to the cleavage effects of the enzyme (Gross et al, 1998). Insensitivity, therefore, serves as a marker for functional regions located in non-coding sequences including promoters, enhancers and silencers (Cereghini et al, 1984; Gross et al, 1998). Notably, the impermanent nature of the nuclease hypersensitivity provides insight into the temporal and tissue-specific stages of activity in the underlying elements when assayed using representative biological samples.

The technology of experimental manipulation of defined DNA segments utilizes gene expression assays to measure changes in the production of a reporter protein in response to cis-acting regulatory signals. Promoter sequences placed upstream of genes of interest can be introduced into a sample of cultured cells and subsequently assayed. The introduction of functional elements creates “gain-of-function” result, whereas “loss-of-function” assays are derived from mutations of functional nucleotides in the target regions. Both techniques are advantageous in helping to capture a functional element when the exact regulatory proteins involved are unknown.

- 2) The techniques directly measure protein/DNA interactions and provide more precise transcription factor binding information. Typical methods include the electrophoretic mobility shift assay (EMSA) (Fried et al, 1981; Garner et al, 1981), DNaseI footprints (Galas et al, 1978), ChIP assay (Kuo et al, 1999) and its high-throughput variant ChIP-chip technique (Ren et al, 2000). EMSA, also known as gel-shift assay, utilizes the sieving power of non-denaturing polyacrylamide gels to separate a protein-bound DNA molecule from one that is unbound. DNaseI footprint combines the binding reaction of an EMSA with the cleavage reaction of DNaseI. Notably, both techniques do not require knowledge of protein identity. The *in vivo* technique of chromatin immunoprecipitation (ChIP), in contrast, is especially useful when the protein of interest is known. It captures *in vivo* protein-DNA interactions by cross-linking proteins to their DNA recognition sites. Before precipitation by a transcription factor-specific antibody, the explored DNA is fragmented into small pieces. After precipitation, reversal of the cross-linking reaction releases the DNA for subsequence detection by PCR amplification and sequencing. ChIP-chip, the high-throughput variation of the ChIP approach, combines the techniques of ChIP assay and cDNA microarray to genome-wide identify putative binding sites. Today, co-immunoprecipitation of bound DNA fragments and their computational analysis with alignment tools or motif finder allocate a comprehensive binding site collection for a particular

transcription factor of interest (Narlikar et al, 2009).

Many hundreds of experimentally characterized and verified cis-regulatory elements, from prokaryotes to mammals, have been identified by various studies. Several public and commercial database resources have collected them. JASPAR and TRANSFAC, for instance, are two high-quality transcription factor binding profile databases (Bryne et al, 2007; Matys et al, 2003). Both resources provide information of cis-regulatory elements with an advanced representation like position-specific weight matrix (see above) and from a wide range of species including vertebrate, insect, fungi and plant (Vlieghe et al, 2006; Wingender et al 2000). Moreover, a few databases exclusively focus on collecting regulatory sites of one or several evolutionary closely related taxa. REDfly, for instance, is a collection of known *Drosophila* transcription factor binding sites and transcriptional cis-regulatory modules (Gallo et al, 2006). AtcisDB of *Arabidopsis* gene regulatory information server (AGRIS) and AtProbe are two promoter binding element databases of *Arabidopsis thaliana* (Davuluri et al, 2003; AtProbe: <http://exon.cshl.org/cgi-bin/atprobe/atprobe.pl>). PLACE and PLANTCARE are resources of verified regulatory sites found in plants, both mono- and dicots (Higo et al, 1999; Lescot et al, 2002). Notably, the collected binding elements are redundant between resources and even within a database due to the tolerance of cis-regulatory motif variability. Nevertheless, the databases provide high valuable resources for experimentalists and computational biologists with in silico analyses or to expand an experimental observation into a genome-wide predictive analysis (see below).

1.2.2.2 Computational approaches

The progress of experimental techniques significantly increased the resolution and accuracy of genome-wide identification of regulatory elements. Their cost, complexity and inefficiency, however, limit large-scale element characterization. To overcome these limitations, an overwhelming number of computational approaches have been developed over the past several years and are widely applied to uncover cis-regulatory element on a large scale.

1.2.2.2.1 Detection of known sites

The development and application of computer algorithms for the analysis and prediction of cis elements can be divided into two sub-problems. The first is, given a collection of known binding sites, develop a representation of those sites that can be used as model to search new binding sequences or identify the position of known sites. Both motif representations introduced above, consensus sequence and position-specific weight matrix, can be utilized to address this problem by straightforward pattern match and sequence scoring criteria, respectively (Staden, 1989; Claverie et al, 1996). However, due to small size of sequence, sequence variation and deficiency of evidence sites, prediction of cis-elements based on both models frequently results in a higher sensitivity but a low specificity. Hence,

additional properties specifying the function of regulatory sequences must be integrated into bioinformatics algorithms to increase the prediction accuracy.

Several complementary observations of the characteristics of regulatory sequences have motivated substantial improvements in the prediction of functional binding sites. For instance, gene regulation is frequently mediated by cooperative interactions between transcription factors that bind to clusters of sites within cis-regulatory modules. Such feature reflects more directly the biochemical mechanisms that regulate gene transcription and can be captured in computational approaches to improve performance of cis-element detection. The algorithms include genome-wide linker scanning of known combinatorial sites and machine-learning techniques to identify characteristics of known regulatory modules that can be used to accurately detect sequences with similar properties. One study based on the former approach is the analysis of two distinct cis-regulatory elements in *Arabidopsis thaliana*, the phytohormone abscisic acid (ABA) response element (ABRE) with an ACGT core and the CE3 coupling element (CE) with a GMCGCGTGKC consensus, which have been experimentally verified as combinatorial regulatory sites for many responses to ABA (Hobo, et al, 1999). A genome-wide linker scan of ABRE-CE module provided substantially better specificity than analysis of isolated sites (Zhang et al, 2005). As an application of the latter algorithm, a study of complex human muscle-specific cis-regulatory module used logistic regression analysis with a vector of five matrix-generated scores that were profiles for the five transcription factors associated with skeletal muscle expression (Wasserman et al, 1998). Compared with the rate of predictions of individual transcription factor binding sites, the focus on combinatorial modules eliminated ~99% of false predictions while retaining 60% of functional regions.

Although the analysis of cis-regulatory modules based on combinatorial interactions between transcription factors can facilitate specificity of prediction, the limited knowledge of binding sites for many transcription factors and reference collections of known cis-regulatory modules precludes the wide application of cluster analysis. Thanks to improvement of sequencing and microarray technologies within the last decade, other relevant characters of regulatory elements have been captured by computational approaches like phylogenetic footprinting/shadowing and overrepresentation in co-expressed gene groups to further improve and enrich the detection of known sites (principles see below). The rVista service (Loots et al, 2004), as an example of internet-based software tool, combines the search of TRANSFAC database (Wingender et al, 2000) with comparative sequence analysis of orthologous sequences using the AVID alignment program (Bray et al, 2003) to reduce the number of false positive predictions by ~95% while maintaining a high sensitivity of the search. The features of unaddressed regulatory system in the cell nucleus like the chromatin structure might also be considered as the next breakthrough to improve the detection of known sites (Felsenfeld, 2003; O'Brien et al, 2003).

1.2.2.2.2 Detection of unknown sites

In mammals and plants, the fact that little is known about most transcription factors and their target binding sites makes *de novo* discovery of cis-regulatory elements an important topic of active research. Computational tools are designed to uncover novel regulatory sites, where nothing is assumed a priori of the transcription factor or its preferred binding sites. Over the past two decades, numerous tools have become available for this task. In fact, with increasing power and accuracy of computational algorithms the *de novo* detection of cis-elements has become a major research area for the analysis of regulatory binding sites.

In comparison with known site identification, the lack of a priori knowledge makes *de novo* discovery a more challenging task and deduces in general a high false positive rate (Guhathakurta 2006). In order to reduce the complexity of site search and increase the specificity of detection, two important assumptions based on evolutionary and functional conservation of regulatory motifs have been employed by most *de novo* approaches. The idea of evolutionary conservation suggests that, orthologous sequences that are likely to have critical functional roles are significantly more similar than what would be expected under a reasonable model of neutral evolution due to negative (purifying) selection (Siepel et al, 2005). Hence, the function and DNA binding preferences of transcription factors are well conserved between evolutionary closely related species. The technique based on this assumption is well known as phylogenetic footprinting (Blanchette et al, 2002; Blanchette and Tompa, 2002) and in general uses comparative genome analysis to look for non-coding sequences that are conserved across species. The traditional large-scale comparative genomics methods for finding regulatory elements have relied on detecting locally conserved motifs within global alignments of orthologous upstream sequences (e.g. Cliften et al, 2003; Kellis et al, 2003). Detected motifs are in general presented in a manner of position-specific weight matrix and evaluated by information content as predictive confidence (see 1.2.1). Although powerful and straightforward, these approaches are highly dependent on alignment accuracy and evolutionary relationship of involved species. Motif detection based on these methods can fail when, on one hand, upstream regions are highly divergent because of large evolutionary distance or have undergone genomic rearrangements. On the other hand, true binding sites may not be overrepresented from non-functional surrounded regions due to high similarity among the sequences compared, if the corresponding species are too closely related. Several alternative algorithms, e.g., the principle of network-level conservation (Elemento et al, 2005) and FootPrinter (Blanchette et al, 2003), have been implemented based on word-search criteria instead of alignments to overcome these limitations. The former one is simple, fast and comprehensive which discovers globally conserved regulatory elements. However, the utilization is restricted to comparison between two genomes. The latter method, also known as an application of phylogenetic shadowing (Boffelli et al, 2003), can consider a large number of orthologous sequences and take their evolutionary relationships into account.

Although significant advances have been made in the past several years, de novo binding site discovery based on evolutionary conservation of functional non-coding elements seems inapplicable in some cases. For example, species-specific binding sites have no significant matches in other species. Detecting these sites by phylogenetic footprinting is impossible unless a large number of closely related species are available. Approaches utilizing sequence information from single species have been developed to solve such problems. The basic idea is to identify cis-regulatory motifs in a given set of genes in the same species that are likely to be regulated by the same (group of) transcription factor(s), i.e. co-regulated genes. Discovery, therefore, is based on a shared feature or function instead of common evolutionary history. In this approach co-regulated genes are assumed to be co-expressed and the information to uncover co-expressed genes by their shared features or functions is typically collected from microarray experiments or the recently developed next generation sequencing technology. Cis-regulatory elements are identified in promoter sequences of co-expressed genes by their statistical over-representation. MotifSampler (Thijs et al, 2001), MEME (Bailey et al, 2009), AlignACE (Hughes et al, 2000) and FIRE (Elemento et al, 2007) are among the well-established and widely applied methods following this principle. These approaches based on functional conservation of cis-elements possess a prominent advantage for the de novo analysis of single organism, if no reference genomes with proper evolutionary distance are available. The limitations of such algorithms, however, are also obvious. Detection depends entirely on the quality of microarray experiments which can be enormously various from diverse platforms and laboratories. Data integration, processing and comparison among different experiments are still problematic. Also, such approach is based on the assumption that co-regulated genes are co-expressed, and vice versa. However, microarrays measure steady-state mRNA levels which are affected by cis-active transcriptional control as well as by post-transcriptional mechanisms regulating mRNA stability. Some experimental setups may violate the assumption of a simple regulatory circuitry causing the observed cluster structures. Mixed cell and tissue samples, long-term or pleiotropic effects triggering transcription factor cascades may considerably complicate the interpretation of expression clusters. Thus, co-expressed genes may often not be transcriptionally co-regulated. Furthermore, technical challenges like the gene clustering methods applied and statistical tests for determination of over-represented patterns can influence the cis-element detections (Guhathakurta 2006).

Leveraging on the fact that transcription factor binding sites are in general both more over-represented across co-regulated genes and more conserved across species, a number of recent developments, including PhyloCon (Wang et al, 2003), PhyloGibbs (Siddharthan et al, 2005), PhyME (Sinha et al, 2004), have reported with equal or higher performances compared to the methods that use only one of the two properties of cis-elements (evolutionary conservation or over-representation in co-regulated genes). These approaches reflect a recent tendency to integrate diverse resources of

evidences for motif discovery. Thanks to improvements of microarray and sequencing technologies within last decade, dramatically increasing number of both whole-genome sequences and high-quality genome-wide microarray experiments has made such analysis feasible. There are valid arguments regarding the disadvantages of using a predetermined alignment result, under which these methods generally operate (Sosinsky et al, 2007; Zhou et al, 2007). Nevertheless, in practice these approaches generate the most reliable analytical outcomes.

Several tools have been developed with additional supports to further improve the de novo cis-element discovery. GibbsModule, for example, integrates evidence of cis-regulatory modules (Xie et al, 2008), and FIRE takes constraints of position and orientation of motifs into account (Elemento et al, 2007). Other tools like regulatory sequence analysis tool (RSAT) bundle independent methods to investigate and integrate multiple evidences (Thomas-Chollier et al, 2008). Moreover, basic inputs for all computational finders are genomic sequences of core, proximal and even distal promoters. A general problem for all motif detections is false or missing annotations of promoter and gene boundaries that may hide and even corrupt motif discoveries. Hence, the improvement of genetic element annotations for the interested genome can make significant advances of computational motif detection.

There are still numerous challenges of de novo motif prediction, although remarkable improvements have been achieved in the past decade. For example, all identification algorithms have very limited power in the search for motifs in distal promoters and enhancers, because iterative and stochastic searches can easily be trapped into local maxima when the search space is large. Also, motifs with small size may possess low information content and can not be identified by widely applied statistical models. Furthermore, assessments of prediction accuracy for the diverse motif finders revealed that there is no single best method for motif discovery (Prakash et al, 2005; Tompa et al, 2005). Each program has its own strength and weakness for the wide range of existing diverse biological questions. To address a particular problem, experimental setup or expected result type, attempts of diverse computational algorithms with combinatorial evidences and custom-made adjustments were highly recommended to undertake in order to achieve optimal results.

1.3 Evolution of cis-regulatory element

“The art of progress is to preserve order amid change and to preserve change amid order.”

- Alfred North Whitehead

Cis-regulatory elements are commonly predicted based on the assumption of their evolutionary conservation. However, the exact nature of their conservation presents a complex picture and the fundamental assumption of regulatory comparative genomics

has been challenged (Moses et al, 2003; Emberly et al, 2003). Ideas about the evolutionary significance of cis-regulatory non-coding mutations are nearly as old as the discovery of regulatory sequences themselves. Soon after the discovery of *lac* operon (Jacob et al, 1961), Jacob and Monod speculated about the unique role that mutations in operators might have during the course of evolution (Monod et al, 1961). Two hypotheses have argued that regulatory mutations make a qualitatively distinct contribution to phenotypic evolution compared to mutations of coding sequences. The first hypothesis, also considered as one potential mechanism of cis-element evolution (see above), argues that selection operates more efficiently on cis-regulatory elements to easily achieve various regulatory states (Stern, 2000; Wilkins, 2002; Wray et al, 2003). Also, the modular organization of some regulatory regions means that a mutation in one module might affect only one part of the overall transcription profile to achieve a spatial-temporal regulation (Stern, 2000; Wilkins, 2002). By contrast, most non-synonymous coding mutations change the resulting protein no matter where it is expressed. Reduced pleiotropy allows selection to operate more efficiently on regulatory regions. This might be particularly relevant for genes expressed in a variety of cell types and tissues. Secondly, some types of phenotypic differences, especially traits associated with dynamic processes such as reproduction, development, behavior and immune responses might be expected to evolve to some extent more readily through regulatory rather than coding mutations (Davidson, 2006). However, Direct testing and evaluation of these hypotheses is not straightforward. Nevertheless, for decades, numerous studies towards various organisms including human, fruit fly, fish, worm and plant have identified cis-regulatory mutations with functionally significant consequences for morphology, physiology and behavior (studies reviewed by Wray, 2007). Remarkably, these analyses of trait divergence have demonstrated evolutionary changes at loci for which functional coding changes have been ruled out and functional cis-regulatory mutations have been implicated or directly demonstrated at the molecular level, revealing that evolution of cis-regulatory elements is sufficient to account for changes in gene regulation within and between closely related species.

Over the course of evolution, regulatory elements of non-coding sequences can evolve during mutations of individual sites including substitution, insertion or deletion. It can also be realized by changing the number, affinity or topology (orientation, spacing, order) of component sites within a cis-regulatory module. Although the mechanisms of cis-elements evolution are still unclear, two hypotheses are widely argued. Firstly, weak selection on individual sites and neutral binding sites might be a rich source for the de novo creation of regulatory elements from non-functional sequence, even if the expression pattern of the target gene is conserved (Ludwig et al, 2006). The effort of ENCODE project consortium included many elements that are expressed or functional according to standard genomics approaches yet selectively neutral (The ENCODE project consortium, 2007). Secondly, it has recently been shown that there are thousands of transposable elements insertions near developmentally regulated human genes and that the transposable element-derived sequences are under strong purifying selection (Lowe et al, 2007). Some of these sequences are believed to be functional

cis-regulatory elements (Bejerano et al, 2006). These findings indicate that the enormous numbers of transposable elements in many animal genomes and thus potentially also for plant genomes are presumable a source of new functional cis-elements.

The number of well-documented cases in which cis-element mutations have contributed to interesting trait differences has grown rapidly during the past few years. However, attempts to characterize the evolutionary patterns of regulatory sequences used a few well-studied cis-elements and were limited in their scope (Wray, 2007). Large-scale and more systematic study of cis-element evolution is still challenged, as the analysis requires extensive data on sets of orthologous non-coding sequences and a large collection of experimentally verified cis-element sequences. The availability of 12 *Drosophila* species (*Drosophila* 12 Genomes Consortium, 2007) and a large collection of experimentally verified *Drosophila* regulatory sequences (Gallo et al, 2006) enable a large-scale study of cis-element mutations. Nevertheless, only limited analysis has been undertaken for other taxa – including plants - due to lack of genome sequences/annotations or knowledge of regulatory elements. Moreover, quantitative estimation of the strength of selection on binding sites has rarely been made. Several open questions like how binding affinities affect selective pressure of binding sites and how the presence of other sites in the neighborhood influences the probability of mutations are far from being solved.

1.4 Goals of this study

Similar to other higher organisms, higher plants need to integrate a large amount of developmental signals to regulate complex patterns of gene expression during their development and differentiation. Moreover, plants have unique needs and strategies for responding to changes in their environments, including light, heat, drought stimuli, abiotic and biotic stresses which cause a range of genes to be activated as part of the plant-defenses and stress responses. Therefore, tight and dynamic gene regulation is necessary in higher plants. However, less analysis has been addressed on the general transcriptional machinery in plants in comparison to animals and yeast and the regulation of gene expression is still far from being understood.

In particular, large-scale detection, characterization and evolutionary analysis of cis-regulatory elements in plants are in its infancy due to the lack of their genome sequences/annotations, high quality genome-wide expression data and well-documented regulatory elements. Until recently, only complete genome sequences of *Arabidopsis thaliana* and *Oryza sativa* (rice) with their annotation were available which are considered as model plants for dicots and monocots, respectively (*Arabidopsis* Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005). The limited application of regulatory comparative genomics in the plant world can now be overcome with the completion of several further genomes including *Populus trichocarpa* (Tuskan et al, 2006), *Sorghum bicolor* (Paterson et al, 2009),

Arabidopsis lyrata (Hu et al, 2011). In addition, thanks to advanced microarray and efforts of worldwide laboratories within the last decade, large sets of microarray datasets from *Arabidopsis thaliana* are publicly available. Recently developed array platforms are currently being used to generate genome-wide expression profiles for several crop species (reviewed by Rensink et al, 2005). The availability of an increasing number of expression datasets provides the opportunities to collect co-expressed gene groups and genome-widely monitor transcriptional profiles and under diverse tissues and stimuli.

Leveraging on whole genome sequences and genome-wide high quality expression data sets, transcriptional machinery of higher plants can now be systematically studied based on comparative and functional genomics. Sorghum and rice, for example, belong to two different grass subfamilies, the Panicoideae and Bambusoideae, respectively, that diverged approximately 60 million years ago (Mya) (Paterson et al, 2009). The rice genome is already available since 2005 while the sorghum genome was recently finished (International Rice Genome Sequencing Project, 2005; Paterson et al, 2009). Though genome sizes differ twofold, gene number and order are similar: about 60% of sorghum genes are located in syntenic regions to rice and orthologous relationships are well established by genetic markers as well as whole genome comparisons (Paterson et al, 2009; Bowers, 2003). In addition, transcriptome data for rice that monitor genome-wide expression levels of many thousands rice genes have become available in recent years (Ma, 2005; Nakano, 2006; Zhou, 2007). This now, for the first time, allows us to analyze conserved sequence elements on a genome scale and to detect candidates for transcription factor binding sites between monocotyledonous species. Moreover, as models of dicotyledonous plants, *Arabidopsis thaliana* and *Arabidopsis lyrata* whose genome was recently completed diverged approximately 10 Mya (Hu et al, 2011). On one hand, the *A. thaliana* genome with 125 Mb is much smaller than that of *A. lyrata*, which is more than 200 Mb. On the other hand, apart from rearrangement, some 90% of the *A. thaliana* and *A. lyrata* genomes have remained syntenic, with the vast majority in highly conserved collinear arrangements (Hu et al, 2011). Thus, comparison of these two close related species with apparent genome size difference can be undertaken not only for the large-scale discovery of cis-regulatory elements but also for the systematic analysis of their evolutionary processes.

As another case study of cis-element evolution, investigation of specific compatibility situations between plastid and nuclear genomes in the flowering plant genus *Oenothera* subspecies *Oenothera* (= *Euoenothera*) were undertaken. In higher plants, one of the disturbances in the development of the resulting cybrids or hybrids caused by co-evolution of the intracellular genetic compartments is the genome-plastome incompatibility (PGI). Particular combinations of nucleus and plastids can affect the development of chloroplast and the chlorophyll synthesis (Stubbe, 1959). PGI has been observed and studied in various higher plants (Pandey et al, 1987; Przywara et al, 1989; Metzlauff et al, 1982; Arisumi 1985), especially well characterized in the

flowering plant *Oenothera* (Renner, 1924, 1936; Stubbe, 1955, 1959). A unique combination of genetic features and complete description of compatibilities has allowed a comprehensive study of PGI in *Oenothera*. Though detailed described, the mechanism and determinants for PGI in *Oenothera* are still far from clear. The search for such genomic determinants is restricted due to a lack of both plastid and nucleus genomes. In 2008, Greiner and co-worker represent the complete nucleotide sequences of representative of the 5 basic *Oenothera* plastomes (I, II, III, IV, V; table 16 in material and methods), their comparison, evolutionary relationship and temporal relation (Greiner et al, 2008). This gave the opportunity for large-scale search of plastid-specific genetic determinants causal to PGI, e.g., changes of protein coding genes or mutations in promoters and polymerase binding sites among the 5 plastid genomes.

This study aims to large-scale detection and characterization of cis-regulatory elements between the dicotyledonous plants, *Arabidopsis thaliana* and *Arabidopsis lyrata*, as well as for the first time between the economically and agriculturally highly important monocotyledonous plants, rice and sorghum. Several computational approaches based on phylogenetic footprinting and co-expression have been applied to systematically identify experimentally verified and novel regulatory motifs, and highlight promising methods for cis-elements discovery in grasses and in higher plants in general. Moreover, evolutionary imprints on non-coding genomic sequences in plants were characterized. Polymorphism of regulatory sites generated by speciation and/or gene duplication of *Arabidopsis thaliana* and *Arabidopsis lyrata* was discussed. Also, diverse situations of compatibilities between plastid genome and nuclear genome of *Oenothera* were studied which are considered as being caused by mutation of particular plastid gene promoters. These analyses revealed that evolution of regulatory elements between close related species might lead to phenotype changes and indicated the evolutionary importance of promoters and cis-regulatory elements in higher plants.

2. Results

2.1 Detection and evaluation of *cis*-regulatory element in higher plants

Like in other organisms, discovery of *cis*-regulatory elements and characterization of promoters in higher plants is an important focus of research since decades. However, analysis is strongly restricted due to limitation in the availability of complete genome sequences. Genome-wide characterization of *cis*-elements is therefore a challenging study. Until recently, the limitation can be overcome with the completion of several further genomes including *Sorghum bicolor* and *Arabidopsis lyrata*. Together with already existing genome sequences of rice and *Arabidopsis thaliana* and associated with their numerous genome-wide expression datasets, regulatory comparative genomics in higher plants can now be addressed. In this study, large-scale identification of *cis*-regulatory elements between the dicotyledonous plants, *A. thaliana* and *A. lyrata*, as well as for the first time between the economically and agriculturally highly important monocotyledonous plants, rice and sorghum, has been undertaken. Several complementary methods have been applied to provide a large-scale collection of *cis*-elements and provide promising approaches for *cis*-element discovery in higher plants in general.

2.1.1 *Cis*-regulatory elements detection in rice and sorghum by combination of co-expression and evolutionary conservation

Sorghum and rice are economically and agriculturally important cereals. They belong to two different grass subfamilies, the Panicoideae and Bambusoideae, respectively, that diverged approximately 60 Mya and demonstrate high similarity of gene numbers and order (see introduction). With the recently completion of the sorghum genome and the already existing rice genome (international rice genome sequencing project, 2005; Paterson et al, 2009), large-scale discovery of *cis*-regulatory elements in sorghum and rice can now be addressed using comparative genomics. In this study, concepts of both evolutionary conservation and overrepresentation were combined (see introduction). To identify candidate *cis*-regulatory elements in rice and sorghum, transcriptional networks in rice were firstly derived from correlation matrices of three independent rice expression data sets (see below). Groups of co-expressed rice genes are obtained as maximal cliques of these networks and each gene of a clique is complemented by its sorghum ortholog. The *PhyloCon* approach developed by Wang and Stormo (2003) was applied to this data set to detect motifs in upstream sequences that are both overrepresented in co-expressed genes and conserved between orthologs (see introduction). Figure 2 depicts an overview of the motif discovery schema. The schema uses *PhyloCon* as analytical and algorithmic basis for the analysis.

2.1.1.1 Preparation of genomic sequence datasets

PhyloCon uses evolutionary conserved footprints for motif discovery. To determine

the genomic search space, promoter regions of orthologous genes between the genome of *Oryza sativa ssp. japonica* (International Rice Genome Sequencing Project, 2005) and *Sorghum bicolor* (Paterson et al, 2009) were determined in a first step. Syntenic regions between Sorghum and rice detected by Paterson et al (2009) were applied to identify orthologous gene pairs. Tandem gene duplications occur frequently in plant genomes and comprise approximately one fifth of all genes (International Rice Genome Sequencing Project, 2005; *Arabidopsis* Genome Initiative, 2000). Considering tandem duplicated genes pronouncedly increases the genomic search space. Hence, to avoid the complication of tandem duplication, only sorghum-rice gene pairs that were detected as bidirectional best *Blastp* hits within syntenic blocks were considered as orthologous gene pairs. In total, 15,773 orthologous gene pairs were identified. In addition, since comparison of incorrect upstream regions caused by erroneously annotated gene starts can strongly impair motif discovery, the analysis was restricted to gene pairs for which pairwise alignments included regions before the 15th amino acid of either protein sequences. After this filtering step, 12,192 orthologous gene pairs fulfilled the criteria and were selected for further analysis.

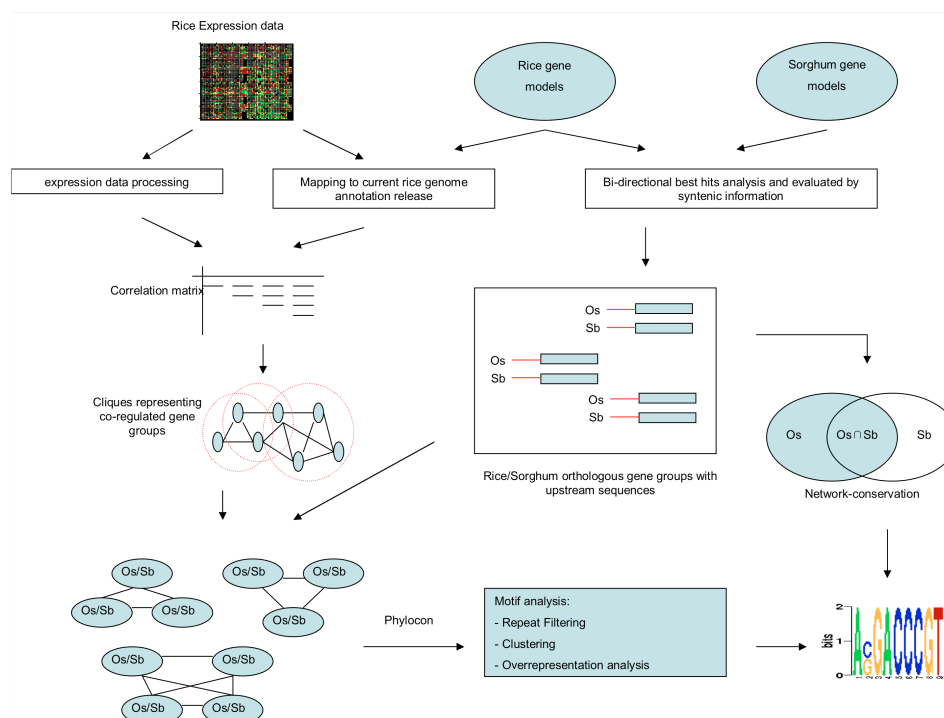


Figure 2: Workflow of cis-element discovery in rice and sorghum.

Two complementary approaches have been employed for motif discovery in rice and sorghum. Starting from orthologous gene pairs between rice and sorghum, orthologous upstream sequences have been isolated. These sequences were used to identify motifs with a high conservation rate between syntenic pairs compared to their single genome frequencies (network-level conservation approach). For PhyloCon, orthologous pairs that were supported by co-expression were combined to orthologous groups and were subjected to a PhyloCon analysis. Motifs were consecutively filtered for repeats and merged by clustering. Statistical significance has been re-evaluated for each of the clustered motifs. For further details, see text.

Promoter regions were defined as genomic sequences from the start codon to the start of the upstream preceding gene, with a maximal distance of 3kb which models current knowledge of plant promoter sizes.

2.1.1.2 Rice expression data processing

Besides phylogenetic conservation, overrepresentation within co-expressed gene groups is utilized as another information resource to discover cis-elements by *PhyloCon*. The next step for *PhyloCon* analysis, thus, was detection of co-expressed gene groups in the rice and the sorghum genome which are typically collected from expression data generated by microarray or deep sequencing experiments. As large scale expression data of sorghum were not available, rice was selected as target species to collect co-expressed gene groups in this study.

Three large-scale rice expression data sets were used to identify co-expressed rice gene groups in this study: massively parallel signature sequencing data and two oligonucleotide array experiments, denoted as MPSS (Nobuta, 2007), YALE-1 (Ma et al, 2005) and YALE-2 (Zhou et al, 2007), respectively. The MPSS data set comprises 249,990 distinct 17-base sequence “signatures” representing rice transcripts (The Institute for Genomic Research, TIGR version 4.0). Gene transcriptional activities were measured based on normalized expression values of these signatures in “TPM” (transcripts per million). The libraries include 12 diverse untreated tissues (e.g. young and mature root and leaf) with some replicates and six abiotic stress treatments (Nobuta, 2007). For the datasets of YALE-1 and YALE-2, a 70-mer oligo corresponding to the sequence within the coding regions of gene models based on a phase II rice genome assembly version (<http://rice.genomics.org.cn>) was designed and printed as two-slide microarray set. The YALE-1 dataset used the array to evaluate gene expression levels at representative developmental stages during the rice life cycle. 42 diverse developmental stages and tissues including seedling shoots, tillering-stage and roots, heading and filling-stage panicles were selected. For YALE-2, genome-level responses to drought and high-salinity stress in rice were elucidated using the same oligo microarray. Gene expression was profiled in differentiation, i.e. ratios, between control and each treated samples including rice shoot, flag leaf and panicle under various drought or high-salinity conditions. Experiments and arrays of each data set are summarized in the material and methods (table 17).

In order to balance the dye-bias, remove the print-tip effect within each array and experimental variances across replicate arrays, raw expression data sets need to be normalized. This process for MPSS, YALE-1 and YALE-2 data sets has been accomplished by Nobuta, 2007, Ma et al, 2005 and Zhou et al, 2007, respectively. The generated normalized data sets were adopted in this study for subsequently analysis. However, several reasons underlie the necessity for further processing and filtering of expression data. Firstly, microarray probes are frequently erroneous due to annotation

updates and improved gene modeling. Thus, the 70mer oligonucleotides of the YALE-1 and YALE-2 arrays as well as the 17mer MPSS tag sequences have been remapped to current rice gene models of the RAP2 annotation (see material and methods). In total, 22,271 MPSS signatures and 27,887 oligonucleotides probes were identified which were unambiguously mapped exactly on one gene and used for the subsequent analysis. Secondly, systematic errors can result in abnormally low expression level of some array probes compared to background. In order to remove such effect caused by systematic errors, the remapped probes were further filtered following the methods of Rinn et al (2003) with minor revision (see material and methods). Overall 19,396, 13,904 and 20,633 reliable and significant probes were selected for MPSS, YALE-1 and YALE-2, respectively, and expression data of these probes were used to detect co-expressed gene groups. As it is highly problematic to combine expression data derived from different platforms, gene groups were identified separately for each of the three expression data sets.

2.1.1.3 Determination of co-expressed groups

Cis-regulatory elements are assumed to be overrepresented in a set of genes that are under similar regulation under particular conditions (see introduction). Hence, co-expressed gene groups were determined and used for the subsequent survey of overrepresented cis-elements in their promoters. The goal of analyzing the rice expression data is to discover such co-expressed gene groups. To measure the expression similarity of two genes, the Pearson correlation coefficient was calculated according to expression data of all included experimental conditions. For each of the three expression data sets, co-expressed genes were defined as pairs whose Pearson correlation exceeded the 99%-quantile of the background distribution of all correlation coefficients. Background and quantiles were estimated from the all-against-all Pearson correlation matrix. In total, co-expressed gene pairs covering 16,426, 13,223 and 18,820 distinct genes from MPSS, YALE-1 and YALE-2 were selected, respectively. A gene group was considered as co-expressed if all of its gene pairs are co-expressed. An undirected graph was constructed with nodes representing genes and edges between them if genes were co-expressed (see material and methods). From this graph, co-expressed gene groups were extracted as maximal cliques for each node where each gene is connected with the other genes. To avoid clusters with broad or unspecific expression patterns, we restricted our analysis to nodes with ≤ 100 edges. Finally, in total 7,456 co-expressed groups covering 11,412 genes, 6,677 groups comprising 15,146 genes and 6,681 groups including 8,793 genes were determined from the YALE-1, MPSS and YALE-2 graph, respectively.

2.1.1.4 *PhyloCon* motif discovery and analysis

PhyloCon was applied to discover phylogenetic footprints between rice and sorghum in each rice co-expressed gene group. Thus, the rice genes contained by the selected co-expressed gene groups were further filtered for assigned orthologous rice-sorghum

gene pairs as determined and described in the section 2.1.1.1. As a result, 4683, 4263 and 2185 rice co-expressed gene groups including 6,667, 4,395 and 2,379 rice-sorghum gene pairs were included by the MPSS, YALE-1 and YALE-2 data set, respectively, and were subjected to *PhyloCon* analysis. Given a co-expressed rice gene group, *PhyloCon* detects suboptimal alignments for each rice-sorghum orthologous pairs. The alignments from different orthologous pairs are compared and merged into new profiles (details see Wang et al, 2003). The final top scored profiles are reported as common motifs for the corresponding co-expressed group and represented as multiple sequence alignment matrices of sorghum and rice instances (fig. 3, left panel).

After application of *PhyloCon* for each co-expressed gene group, overall 17,068, 14,754 and 5,337 alignment matrices from the MPSS, YALE-1 and YALE-2 data sets were discovered, respectively. The average sizes of motifs detected on the three datasets are similar (approx. 20 base pairs; table 1). Moreover, for each co-expressed clique, on average more than its 70% rice-sorghum gene pairs cover the corresponding motifs. This is consistent with overrepresentation of detected rice-sorghum conserved motifs in rice co-expressed gene groups. In order to further quantitatively characterize detected motifs, the model position-specific weight matrix (PWM) was chosen. This presentation can measure the degree of motif variability and specificity by directly using the probabilities in an alignment matrix. An alternative model, the word/k-mer based motif representation, has no statistical power and is hardly able to provide quantitative description of alignment-based motifs (see introduction). Each alignment matrix was transformed to a PWM and the corresponding information content suggesting motif (binding) specificity against background was calculated (see material and methods). Average information contents of detected motifs are in the range of 14.6 by MPSS up to 20 by YALE-2 (table 1) which indicates their high binding specificity.

Detected motifs show on average high overrepresentation in co-expressed gene groups and high information content (table 1). However, occurrences in cliques, binding specificities and sizes of individual motifs vary dramatically within each dataset. The motif lengths range from 5 bases up to more than 100. The maximal size of motif detected by YALE-1 even reaches 208 (table 1; variance=380.6). A similar size distribution was also observed for motif occurrences in cliques and their specificities. The fraction of genes in co-expressed cliques that contain the corresponding motifs vary between 3% and 100% (variance=418.5), and information contents are in the range of 5 up to 200 (variance=275.1) by YALE-1 (table 1). The large motif variations are due to the algorithm applied by *PhyloCon* that merges profiles from diverse orthologous gene pairs and discover more informative sub-profiles (details see Wang et al, 2003). The profile recruitments are terminated, until no more optimal sub-profiles can be generated. Hence, information content, size of a final motif and the fraction of covered genes within a particular clique are determined by the cycles of profile comparison and the number of recruited gene members. However, correlation of these motif features has been observed. For the detected motifs in each data set, information content raises with the increase of motif size ($p < 0.01$), while negative correlation between size and occurrence rate ($p < 0.01$) as well as information content and occurrence rate ($p < 0.01$) were found.

		PhyloCon	Over. Test	Clustering
Number (#)	<i>MPSS</i>	17,068	4,951	1,622
	<i>YALE-1</i>	14,754	5,885	1,501
	<i>YALE-2</i>	5,337	3,731	866
Length (bp)	<i>MPSS</i>	17 (5-172)	32 (10-172)	23 (7-93)
	<i>YALE-1</i>	21 (5-208)	35 (10-182)	22 (8-108)
	<i>YALE-2</i>	22 (5-148)	26 (10-148)	20 (7-74)
Occurrence (%)	<i>MPSS</i>	74 (7-100)	60 (7-100)	58 (12-100)
	<i>YALE-1</i>	71 (3-100)	60 (7-100)	61 (13-100)
	<i>YALE-2</i>	75 (28-100)	66 (28-100)	66 (33-100)
Information content	<i>MPSS</i>	14.6 (4.3-165.9)	28 (9.8-156.7)	19.3 (7.8-81.4)
	<i>YALE-1</i>	17.9 (5.5-198.7)	30 (10.4-155.8)	18.4 (7.7-93.7)
	<i>YALE-2</i>	20 (5.7-123.8)	23.6 (9.4-123.8)	17.2 (7.6-63.8)

Table 1: Summary of PhyloCon detected motifs

Table summarizes several features of PhyloCon detected motifs for each step of analysis. “PhyloCon”, “Over. Test” and “Clustering” indicate initial motifs deduced by PhyloCon, motifs after filtering by overrepresentation test and final motifs after clustering, respectively (details see text). “Occurrence” represents the fraction of gene members of PhyloCon defined gene groups (in percent) which contains their respective motifs. Information content was calculated according to Stormo, 1998. The numbers left to and in brackets indicate the average value and range of values, respectively.

Functional elements of non-coding sequences are overrepresented in upstream sequences of co-expressed as well as orthologous gene groups compared to all upstream regions, the so called background distribution. However, motif detection by *PhyloCon* is completely restricted to locally confined groups, i.e., upstream sequences of rice co-expressed gene groups with orthologous sorghum genes, and disregards background distribution. Thus, the detected profiles derived from widely distributed elements (e.g. the TATA box) may also present high occurrence rate in all rice upstream sequences. Therefore, profiles need to be filtered to avoid such global or unspecific overrepresentations. Thus, for each PWM, a cumulative binomial probability distribution was employed to test statistical significance of its overrepresentation within the respective co-expression groups in comparison to all rice upstream sequences (see material and methods). 4,950, 5,850 and 3,731 motifs for *MPSS*, *YALE-1* and *YALE-2*, respectively, were obtained after the overrepresentation test which demonstrated significantly higher information contents and larger size compared to initially detected motifs (t-test $p < 0.01$; table 1). In contrast, the average occurrence rate of these motifs in respective cliques decreased (table 1). These differences suggest that the applied overrepresentation test filtered a large portion of initially detected short motifs that may display global overrepresentation, i.e., overrepresentation both within co-expressed gene group and in all rice upstream sequences. Thus, a list of more specific and informative motifs is generated.

Transcription regulation is commonly achieved by binding several transcription factors to a set of cis-acting elements (see introduction). Such combinatorial attachment indicates that, on one hand, two genes with single/or only few shared cis-elements may not reach a similar expression level and on the other hand, two differentially expressed genes may still share individual cis-elements. However, motif detection by *PhyloCon* was applied for each predefined co-expressed clique separately. This may deduce redundant motifs among diverse gene groups (fig. 3a). Moreover, depending on the order of profile recruitments, *PhyloCon* can generate motifs with large overlapping even within the same co-expressed gene group (fig. 3b). Thus, with the goal to limit redundancy and combine motifs, clustering of previously collected motifs was subsequently applied. Similarity of two motifs was measured by their alignment generated from all involved instances (see material and methods). Motifs with significant high similarity were clustered and gap-free regions of their multiple sequence alignments were extracted as new motif profiles (see material and methods).

Clustering of motifs can potentially alter its occurrence significance. Figure 3c depicts one example where each individual motif members in the cluster is overrepresented while the clustered motif is not. Thus, statistical overrepresentation of the newly formed motifs was retested as described previously to filter motifs with global overrepresentation. In total, 1,622, 1,500 and 866 motifs were derived in MPSS, YALE-1 and YALE-2 data sets, respectively. Compared to unclustered motifs, no large difference of motif occurrence rate in cluster was observed which suggests more instances were involved in the clustered motifs. However, the average size of new motifs considerably decreased (t-test $p < 0.01$; table 1). This may result in relative higher motif degeneration rate and reduce the information content (table 1). Nevertheless, a final list of low redundant motifs with reliable information contents was collected from all the three experiment data sets and supported both by evolutionary conservation as well as by co-expression. The detected motifs and their occurrence in each rice and sorghum gene are provided in the appendixes 1-6. Table 2 shows part of results derived by MPSS data set for rice.

Gene	Site	Position	Cluster ID
Os01g0100900	CCCAATCCCCCGGCCGAA	-35	750
Os01g0102000	CCGCACTGCTCGCCC	-2529	1596
Os01g0102950	CGCCGCACACACG	-20	1655
Os01g0104400	CACCACCCAGCAGTCTC	-2024	1573
	CGTCGTGCGCACTGCCCGGCTCCAGCCGCCGGCTGCTG	-2078	1248
Os01g0104800	CGACGCCGCCTCCGTGAGCAGGGAG	-1903	318
Os01g0104900	CCACTGACGGCC	-818	267
Os01g0106900	CACACAATCTTCTTCTCCTCT	-118	166
	CCCCATTGCCCCGTTCTGCATCCCGACCAGAC	-65	941
Os01g0107000	GCGCCTCGATCCCCTTCTC	-2893	76
Os01g0108000	CCCAAAGCCACCT	-98	1081
	TCCTCTCTTGACTCCACTCGCCCT	-42	1401
Os01g0108800	ACCTGCCCCAATTCTCTCCTCCA	-170	1460
	CCCAATTCTCTCCT	-164	1081
	CCTCACAACCAGTCTTCTCCTC	-196	1112
	CCTCCACCTCGCCCAATTCTCTCCTCCAC	-175	825
	CGCCCAATTCTCTCCTCCA	-166	362
	CGGCTCCACCTC	-147	362
	CTCCTCACAACCA	-198	1763
	CTTCTTCTCCTC	-184	1763
	TTCCCCCTCCCCTGGCTG	-2170	612
	TTCTCTCCTCCACGGCTCCACCTCT	-159	1401
Os01g0109700	CACCCATCCTTCCCTCTCGCGTC	-915	266
	CCCTCTTCGCGCCGCGCTCGGGGGTCTCGC	-828	454
	CGAGGTTTGAGAGGATCGGCGGCGGCGGCGGCGGC	-755	454
	CGATTAGCCGCTCTCCTCCCCGCGCG	-876	266
	CGGTGTACGT	-769	454
	CTCCTCCCCGCGCGTGC	-863	454
	GCAGATCCGCGC	-782	454
	TCTTCTTCTTTCGCTCGCCCTTGCTCGTTCC	-951	266

Table 2: Motif sites detected by PhyloCon in rice derived from MPSS.

Table shows part of motif sites detected by PhyloCon in rice genes derived from MPSS. The position of sites shown in the third column indicates the distance to the start codon. The complete lists of motif sites for rice and sorghum are represent in the appendixes 1-6.

2.1.1.5 Validation of PhyloCon detected motifs

In this study, computational de novo discovery of cis-regulatory elements is based on several assumptions of their features, e.g., overrepresentation and evolutionary

conservation. Thus, authentic biological functionality of detected motifs needs to be verified in order to provide confidence of collected cis-elements and to test reliability of applied computational approaches. In previous studies, the functionality of motifs has been confirmed by a variety of approaches (Wray, 2007). One of the straightforward methods is experimental verification of function of each individual detected element. For example, alteration of gene transcription activities can be monitored by mutating its putative cis-regulatory sites. Such experimental validation provides direct evidence of function of potential sites. However these tests are economic inefficient and not applicable if the amount of elements to be verified is large. To overcome the experimental limitation, validation of putative elements was undertaken by computational methods utilizing prior experimental acknowledges of cis-regulatory elements. One method is the survey for existence of previously verified sites among discovered elements. These sites can be collected from public databases or individual studies. Despite limited availability and redundancy of known sites in grasses (see discussion), overall 76 known sites were obtained from two databases TRANSCFAC and PLACE and mapped to detected motifs (see material and methods). Table 3 shows several examples from numerous detected sites which displayed high similarities to the verified elements including many variants of ACGT motifs like G-box or ABA response element as well as several hormone response factors.

Genes associated with the same biological process may be co-expressed according to the “guilt-by-association” rule (Quackenbush, 2003) and potentially regulated by similar regulatory system. For large-scale analysis in previous studies, gene ontology annotations and metabolic pathways were correlated with particular motifs (Maleck, 2000). In details, functional motifs were enriched in upstream sequences of gene sets associated with particular biological process in comparison to all upstream sequences. This principle can in general be used to validate detected motifs and their functionalities, yet could not be applied in this study due to the limitation of the current rice GO annotation (see discussion).



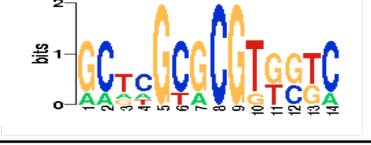

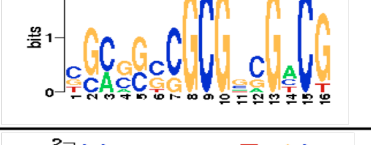

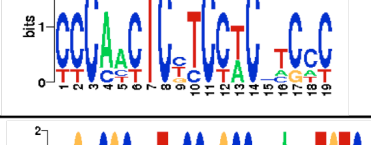
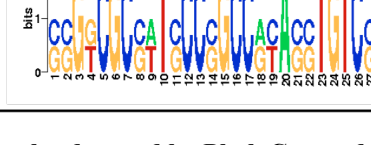
Dataset	Cluster ID	Sequence Logo	Known Site	Description
MPSS	1774		CACGTGG	G-box plus G
MPSS	1849		TATCCA	alpha-amylase promoters of rice
YALE-1	1274		CGCGTGG	"Motif III" in promoter of rice rab16B gene
YALE-1	1963		CGGCGGCCTCGCCACG	"region 1" ABRE-like sequence found in rice Osem gene
YALE-1	74		GCCGCGTGGC	"Motif III" in promoter of rice rab16B gene
YALE-2	424		CCAGGTGG	"Site I" of rice proliferating cell nuclear antigen
YALE-2	622		CAACTC	CAREs (CAACTC regulatory elements)
YALE-2	858		TACGTGTC	ABRE motif A

Table 3: Motif examples detected by PhyloCon and matching to reported known sites

Motifs detected by PhyloCon were remapped to reported regulatory elements and their similarity was manually inspected. Table shows several examples in sequence logos with their matching known sites.

Sizes of cis-elements in higher plants are comparable to non-plant species and typically range is between 6 and 12 base pairs (Matys et al, 2003; Bryne et al, 2007). However, the mean size of *PhyloCon* profiles detected in this study is 22 base pairs, which is considerably longer than expected. Hence, profiles likely represent concrete conserved sites rather than conserved motifs, i.e., statistical models for transcriptional factor binding sites. Such large sizes for evolutionary conserved sites in grasses are

consistent with a previous study of 228 maize and rice pairwise and 56 rice, maize and sorghum three-way comparisons, in which a minimum motif size ≥ 20 bp was considered as significant (Guo et al, 2003). Such long sites of detected profiles can be composed of several motifs whose close proximity is required to realize particular function of the respective co-expressed group. Alternatively, some of the detected sites could represent signals associated with transcriptional gene activity like mRNA stability signals or miRNA target sites for which longer sizes have been reported (Vazquez F, 2006).

In summary, a list of motifs with high specificity was detected as potential cis-regulatory elements in rice and sorghum in this study. The mean size of detected motifs is considerably larger than expected in higher plants and likely represent concrete conserved sites rather than conserved motifs. *PhyloCon*, on one hand, utilizes both information resources of evolutionary conservation and overrepresentation in co-expressed group to increase the power of cis-element detection. On the other hand, the prediction was restricted only in strictly defined groups. Identification of “globally” functional elements, thus, was difficult to realize (details see discussion). Nevertheless, numerous matches to experimentally verified sites provide a high confidence of detected motifs and reliability of the used pipeline.

2.1.2 Cis-regulatory element detection in rice and sorghum by “global” evolutionary conservation

Cis-element discovery by *PhyloCon* is restricted in co-expression gene groups with orthologous gene partners and highly depends on the quantity and quality of expression data. In contrast to motif detection with regard to expression data or from a confined or user-selected set of genes, “network-level conservation” detects globally conserved motifs from comparison between two genomes. Functional motifs are identified by their higher conservation in orthologous promoter pairs in comparison to a random expectation from the occurrence in the respective genome. An alignment-free implementation of the network-level conservation principle, FASTCOMPARE, has been successfully employed to motif discovery in yeast, nematodes, fruit flies and human (Elemento et al, 2005). In this study, FASTCOMPARE was adopted to investigate network-level conservation in sorghum and rice with some modifications. Figure 4 depicts the principle of network-level conservation applied in this study for motif discovery and was applied as a complementary approach to the *PhyloCon* method for cis-element detection in sorghum and rice.

2.1.2.1 Motif detection by network-level conservation: implementation

The “Network-level conservation” principle discovers sequence motifs that are overrepresented in the orthologous promoter pairs compared to promoters of an individual genome. The same sets of 12,129 orthologous promoter pairs from rice and

sorghum identified in section 2.1.1.1 for PhyloCon analysis were used as motif search space. Firstly, a list of all possible k-mers with size range from 6 to 12mers was generated by permutation of the nucleotide alphabet. Such size range was chosen, as it corresponds to the typical size range in higher plants (Matys et al, 2003; Bryne et al, 2007). However, due to large computational time costs for the further analysis including motif degeneration and removal of redundant motifs (see below), motifs in the range of 6 to 9 base pairs were considered in this study. Secondly, for a given k-mer, its occurrences in orthologous promoters of each genome and co-occurrence between orthologous promoters were identified based on simple string match. Expected co-occurrence was determined and ratio between observed and expected co-occurrence was calculated as a test score representing the genome-wide evolutionary conservation rate of the respective k-mer (see material and methods).

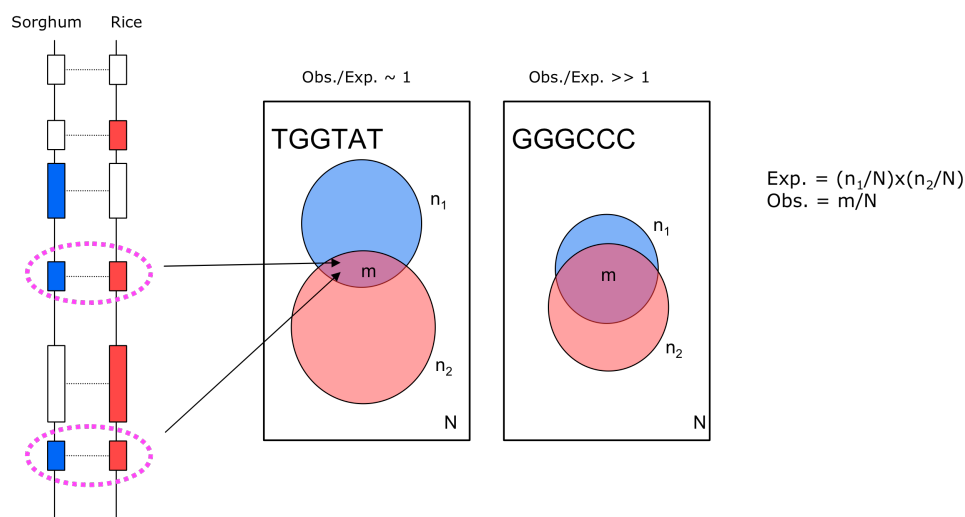


Figure 4: Principle of “network-level conservation” approach

“Network-level conservation” detects globally conserved motifs from comparison between two genomes. Functional motifs are identified by their unusually high retention in orthologous promoter pairs in comparison to the expectation from single genome occurrence.

Moreover, as binding to cis-regulatory motifs commonly tolerates some degree of variability for particular site positions, degenerated motif detection was subsequently included in the analysis. Degenerated motifs were represented as regular expression patterns (table 6) and generated from k-mer sites by substitution of the nucleotide alphabet at any position to multiple letters. A heuristic degeneration process was applied in this study (see material and methods), as comprehensive enumeration of all possible variations at each position for each k-mer is computationally infeasible.

During this process, a more generalized motif with a higher score, i.e. genome-wide conservation rate, replaced the one from which it has been generated.

Top scored motifs demonstrate high evolutionary conservation rate. Detected motifs, thus, need to be ranked by comparing their scores. However, background occurrences of motifs in each genome and their co-occurrences in orthologous promoter pairs vary dramatically among different sizes. This may result in pronounced score distribution bias (fig. 5 upper panel). Therefore, scores were subsequently normalized and transformed to z-scores for each motif size in order to realize comparison of motifs among different sizes (fig. 5 lower panel). Motifs with an overrepresentation of two standard deviations above the mean ($z\text{-score} \geq 2$) were considered as candidate motifs that are supported by their significantly high global conservation rate within orthologous promoter regions between rice and sorghum. In total 7,340 motifs corresponding 2.4% of all degenerated k-mers in the range of 6 to 9 base pairs were discovered and regarded as highly evolutionary conserved between rice and Sorghum. Table 4 displays the number and average z-score of these significant motifs for each k-mer size. On one hand, the fraction of significant motifs decreases with increase of motif size (5.3% and 1.83% for 6- and 9-mer, respectively). On the other hand, it has been observed that longer motifs also circumvent motifs with extremely higher conservation rate (table 5). The highest z-score of detected 9-mers is 93.7 which is around 15-fold higher than top z-score of 6-mers.

Motif enumeration and degeneration were undertaken separately in each motif set with the same size. The lack of pattern comparison among motifs from diverse sizes can result in redundancy of detected motifs. Many motifs may be derived from size (or regular expression) variations of one common motif. For example, several size and degeneracy variants of reported telomere repeats (AAACCCT) $_n$ in *Arabidopsis thaliana* (Tremousaygue et al, 1999; 2003), like AACCCCTA, AAACCC, AACCCCTAG, were highly overrepresented and identified as candidate motifs. In order to avoid redundancy, motifs with size variations showing word/string overlapping were merged by removal of motifs derived from other higher scored motifs (see material and methods). In total 3,985 motifs remained and demonstrated higher average conservation rate compared to redundant motifs (table 4). Large fractions of motifs with small size were removed (table 4), as they are frequently

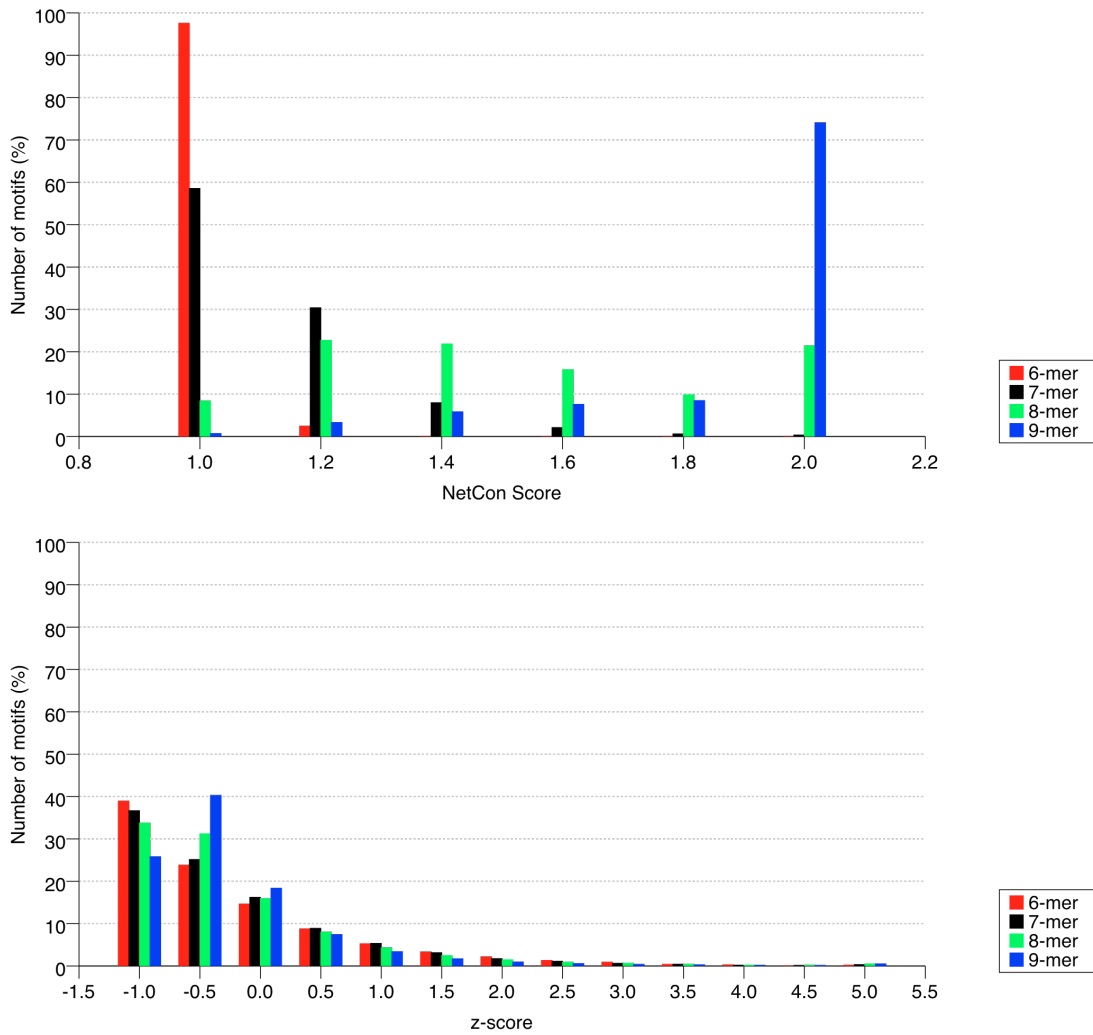


Figure 5: Distribution of conservation scores and z-scores for 6-9mer motifs

Figure displays distribution of conservation score directly calculated by “network-level conservation” approach applied in this study (upper panel) and z-scores normalized and transformed from conservation score (lower panel). Number of motifs (in percent) of each score bin is shown on the y-axis.

		initialDeg	zscoreFilter	merge
6-mer	#	4096	217	6
	%	100	5.29	2.76
	z-score	0	2.89	3.61
7-mer	#	16384	753	138
	%	100	4.59	18.3
	z-score	0	3.06	3.13
8-mer	#	52421	2184	1218
	%	100	4.16	55.77
	z-score	0	3.3	3.4
9-mer	#	228768	4186	2623
	%	100	1.83	62.66
	z-score	0	3.83	4.28
total	#	301669	7340	3985
	%	100	2.43	54.29
	z-score	0	3.57	3.98

Table 4: Summary of detected motifs by network-level conservation

Table summarizes several features of detected motifs based on “network-level conservation” approach. The analysis procedures include “initialDeg”, “zscoreFilter” and “merge” indicate initial degenerated motifs discovered by network-level conservation, motifs after filtering using z-scores and final motifs after removing redundant ones among different sizes, respectively (details see text). The percentage numbers represent the remaining fraction of all motifs from previous analysis procedure.

6-mer		7-mer		8-mer		9-mer	
A:C:G:C:G:T	6.54	T:A:A:C:G:C:G	8.54	C:T:T:A:C:G:C:G	25.59	C:G:C:C:T:T:A:G:A	93.68
T:A:C:G:C:G	6.05	C:G:C:C:T:A:T	8.19	A:C:G:T:T:A:C:G	15.89	C:A:C:G:T:T:A:C:G	44.24
G:A:C:C:C:G	5.91	T:A:C:G:C:G:C	7.75	G:A:C:C:G:T:T:G	15.42	G:T:C:A:A:C:G:C:G	34.73
C:G:T:A:C:G	5.48	G:G:T:C:A:A:C	7.68	G:C:G:C:C:T:T:A	14.22	G:A:C:C:G:T:T:G:C	28.63
G:G:G:C:C:C	5.34	G:T:A:C:G:C:G	6.95	C:C:G:T:T:A:T:C	14.09	C:G:A:C:C:G:T:T:G	28.59
C:C:C:G:G:G	5.27	G:G:A:C:C:C:G	6.76	A:A:C:G:G:T:C:G	13.31	C:C:G:G:A:T:A:A:C	27.50
C:G:C:G:A:C	5.03	C:G:C:G:T:A:A	6.71	C:G:C:C:T:T:A:G	13.00	C:G:T:T:G:A:C:T:C	24.66
C:G:C:G:T:A	5.03	C:T:G:T:C:G:G	6.67	A:G:C:G:C:G:T:A	12.95	T:G:A:C:C:G:G:A	24.41
C:G:C:G:T:C	4.78	T:A:A:C:G:G:C	6.61	A:T:A:C:G:C:C:C	12.32	C:G:C:G:C:T:A:T:A	21.25

Table 5: Top motifs of each k-mer size discovered in rice and sorghum

For each k-mer analyzed in the study, 10 motifs with top z-scores are listed in the table.

derived from longer ones and show lower scores. Moreover, 3,353 (84%) of the deduced motifs are specific sites without any degenerated positions. Several reasons can be considered to result in such specificity of motif detection based on the “network-level conservation” principle and are addressed in the discussion part. Appendix 7 summarizes the detected motifs.

2.1.2.2 Detection of dyadic motif

Cis-elements with close proximity can bind transcriptional factor simultaneously as motif pattern to ensure a coherent regulation of respective genes (Davidson, 2006). To simulate such *cis*-regulatory structure, the model of dyadic motifs with patterns of type $\{X\}_a\{N\}_b\{X\}_a$ were also investigated in the study, where X represents a specific letter, N represents any letter from the nucleotide alphabet, and a and b ranging from 2 to 6 and 6 to 12, respectively. Similar principle as described previously with minor modification was employed to detect dyadic motifs. Instead of degeneration of a specific letter, a greedy schema was applied to test whether more specified versions of the initially unspecific spacer sequence results in a higher scoring motif (details see material and methods).

Overall 441 dyadic motifs were detected (table 7; appendix 8). Notably, 158 (36%) of them were converted to non-dyadic ones with higher conservation scores. In fact, for dyad motifs in the range of 10 to 18-mers, significantly smaller number of spacers in final dyadic motifs (numbers in the table 6) were observed compared to the expected range of spacer in initial dyadic motifs (highlighted in pink cells of table 6). This indicates that large fraction of spacer for numerous dyadic motifs were converted to specific letters and showed high specificity. Moreover, several dozens of these motifs show a very rare occurrence ($3 \leq n < 10$) in the 12,129 orthologous upstream sequences of rice and sorghum but all or almost all of occurrences in single species are conserved between orthologous pairs (table 7; appendix 8; highlighted in red).

	0	1	2	3	4	5	6	7	8	9	10	11	12	total
10	108	11	0	0	0	0	0	0	0	0	0	0	0	119
11	30	7	0	0	0	0	0	0	0	0	0	0	0	37
12	9	10	10	26	0	0	0	0	0	0	0	0	0	55
13	9	3	5	5	3	0	0	0	0	0	0	0	0	25
14	1	2	2	5	12	19	1	0	0	0	0	0	0	42
15	0	2	4	2	3	2	3	1	0	0	0	0	0	17
16	0	0	1	1	0	3	2	4	2	0	0	0	0	13
17	0	0	0	1	2	0	4	3	5	0	0	0	0	15
18	0	1	0	0	0	1	2	3	0	2	0	0	0	9
19	0	0	0	0	0	0	0	0	9	2	15	1	0	27
20	0	0	0	0	0	0	0	0	0	1	2	7	1	11
21	0	0	0	0	0	0	0	0	0	1	3	5	52	61
22	1	0	0	0	0	0	0	0	0	0	0	4	3	8
23	0	0	0	0	0	0	0	0	0	0	0	0	2	2
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0
total	158	36	22	40	20	25	12	11	16	6	20	17	58	441

Table 6: Number of dyadic motifs for diverse motif length and spacer length

Dyadic motifs were initially generated based on the criteria that the allowed number of specific letters and spacer is in the range of 2 to 6 and 6 to 12, respectively. Hence, according to specification process applied in this study (details see text and methods), length of dyadic motifs can be in range from 10 to 24 (rows in table), while length of spacer is between 0 and 12 (columns in table). Depending on all possible length of dyadic motifs and length of spacer regions, table shows the obtained numbers of dyadic motifs (numbers in cells) and expected regions where motifs could appear. Pink regions are the possible length of spacer for dyad motifs with respective lengths.

Motif	#rice	#sorghum	#common	Zscore
T:C:G:C:N:N:N:G:N:N:G:C:G:N:N:A:G:G	17	16	7	36.49
A:A:C:C:N:T:C:C:A:G:A:A	16	7	3	32.35
C:A:C:G:N:G:N:N:N:N:N:N:C:N:N:A:N:T:G:G	28	26	5	24.42
C:T:A:G:G:G:N:N:N:G:G:T:T	19	28	6	18.58
G:G:T:A:C:G:A:T:C	13	8	3	18.51
T:A:G:C:G:C:G:T:C:T:G:A:C:T:T:C:A:G:A:T:C:A:G:A:A	5	3	3	18.51
T:T:C:N:N:G:G:N:N:G:G:T:T:C	18	15	3	18.3
G:C:G:G:A:T:A:A:G:C:C	4	4	3	17.34
G:A:A:N:C:N:G:C:N:T:N:C:A:C	15	19	3	17.31
C:C:A:C:G:C:N:N:N:N:C:N:N:N:N:C:A:G	14	18	3	16.7
T:T:C:G:N:T:C:C:N:N:N:G:C:A	19	26	5	16.63

Table 7: Dyadic motifs detected by network-level conservation

Examples for long specific motifs are shown that emerged from dyad motifs with initially unspecified spacer sequences (denoted as N). Note that many motifs (marked in red) have a low occurrence rate in rice and sorghum; however, most or all occurrences are conserved between orthologous pairs. The complete list is present in the appendix 8.

2.1.2.3 Validation of detected motifs

Several approaches utilizing diverse information resources were applied to evaluate motif discovery based on genome-wide evolutionary conservation between rice and sorghum. Firstly, similar to *PhyloCon* analysis in 2.1.1.6, mapping detected motifs to the verified sites was undertaken to verify identified motifs. Out of 76 distinct sequences extracted from public database PLACE and TRANSFAC (see 3.1.1.6), 21 known regulatory sites have been mapped by 416 “network-level” detected motifs. Table 8 displays several examples of top scored short motifs with their experimental evidences. In addition, using literature searching, motifs extracted from literature but not present in databases have been detected, for instance a perfect match to ethylene response element GGGCCC and motifs highly similar to the telo-box AAACCCTA reported in *Arabidopsis thaliana* (see above; Tremousaygue et al, 1999; 2003).

Notably, one known regulatory pattern which contains two cis-elements, ABRE and CE3, and confers transcriptional ABA responses (Hobo, et al, 1999) has been detected as an example of co-conservation. These two elements are highly conserved both in sequence and in position among two rice upstream sequences of Os01g0705200 and Os05g0542500 with their respective sorghum orthologous partners (fig. 6). Such conserved cis-element arrangement may play an essential role to maintain the transcriptional mechanism of the two rice genes both of which have been annotated as late embryogenesis abundant (LEA) proteins.

Limited availability of prior experimental knowledge of cis-regulatory sites and gene functional annotation in grasses restricts the validation of putative elements. Hence, other information resources like expression data were utilized to evaluate “network-level conservation” analysis by monitoring transcriptional activities and expression correlation for the genes related to corresponding motifs. Responses to environmental changes and expression patterns in higher eukaryotes frequently result from the combinatorial actions of two or more transcription factors that bind to

Motif	#rice	#sorghum	Exp.	Obs.	Z-score	Known sites	Motif Name
G:C:G:G:A:A:A	177	196	2.86	22	6.5334	G:C:G:G:A:A:A	re2f-1 element
C:C:T:T:A:T:C:C	390	315	10.13	75	6.2174	C:T:T:A:T:C:C	GATA/SBX element
A:C:G:C:G:T:G:T:C	37	31	0.09	3	4.8835	A:A:C:G:C:G:T:G:T:C	CE3 (Coupling element)
C:A:C:G:T:G:A	950	873	68.38	156	4.7977	C:A:C:G:T:G	G-box
G:G:A:C:G:T:C:A	116	104	0.99	6	4.7054	A:C:G:T:C:A	hexamer motif
T:T:A:A:T:G:CG:C:G	95	53	0.42	9	4.2548	T:T:A:A:T:G:G	Target of WUS
C:C:A:C:G:T:G	1577	1099	142.89	308	4.2168	C:C:A:C:G:T:G:G	G-box
G:T:A:C:G:T	2788	2859	657.18	874	3.9756	G:T:A:C:G:T:G	ACGT motif
A:C:C:G:A:C:G	880	800	58.04	117	3.5723	A:C:C:G:A:C	DRE
C:G:C:A:T:A:T:C	129	96	1.02	5	3.4547	C:A:T:A:T:C	I-Box
A:C:G:T:G:G:C	1408	1006	116.78	231	3.3984	A:C:G:T:G:G:C:G	ABRE
G:C:A:A:C:G:T:G:A	49	44	0.18	4	3.2268	C:A:A:C:G:T:G	OsBP-5 binding site
A:A:C:C:G:A:C	714	715	42.09	79	2.9321	A:C:C:G:A:C	DRE
T:T:T:C:C:C:G:C	248	244	4.99	33	2.9232	T:T:T:C:C:C:G:C	E2F binding site
G:G:G:C:C:C	3813	3394	1066	1472	2.8948	G:G:G:C:C:C	ERE
T:A:G:C:C:G:C:C:T	56	55	0.25	5	2.7213	A:G:C:C:G:C:C	AGC box
G:C:G:G:TATA:T:T	53	44	0.19	3	2.7055	G:C:G:G:T:A:A:T:T	GT2 binding site
G:C:A:C:G:T:G:G	258	219	4.66	19	2.5531	C:A:C:G:T:G:G	G-box plus G
T:A:A:C:C:C:T:A	432	337	12	48	2.4654	A:A:C:C:C:T:A	Telo-box
A:C:T:T:T:G:C:G	114	134	1.26	5	2.4333	A:C:T:T:T:G	T-box
T:A:C:G:T:A:C	1119	1210	111.63	194	2.2906	T:A:C:G:T:A	A-box
T:A:G:C:C:G:C:C:A	68	62	0.35	6	2.285	A:G:C:C:G:C:C	AGC box
C:A:A:C:G:T:G:G	249	178	3.65	14	2.2805	C:A:A:C:G:T:G	OsBP-5 binding site
G:G:G:T:A:A:T:CT:G	62	44	0.22	3	2.128	G:G:T:A:A:T:T	GT2 binding site

Table 8: Motif samples in rice and sorghum detected by network-level conservation

Table gives motif examples with a high conservation rate in rice-sorghum orthologs. The number of genes (out of 12,129 syntenic genes) in rice and sorghum containing the respective motif in their upstream sequence are shown in the columns '#rice' and '#sorghum', respectively. The following columns show the number of expected and observed co-occurrences in syntenic pairs, as well as the z-score. Matches to known sites/motifs are indicated in the last two columns.

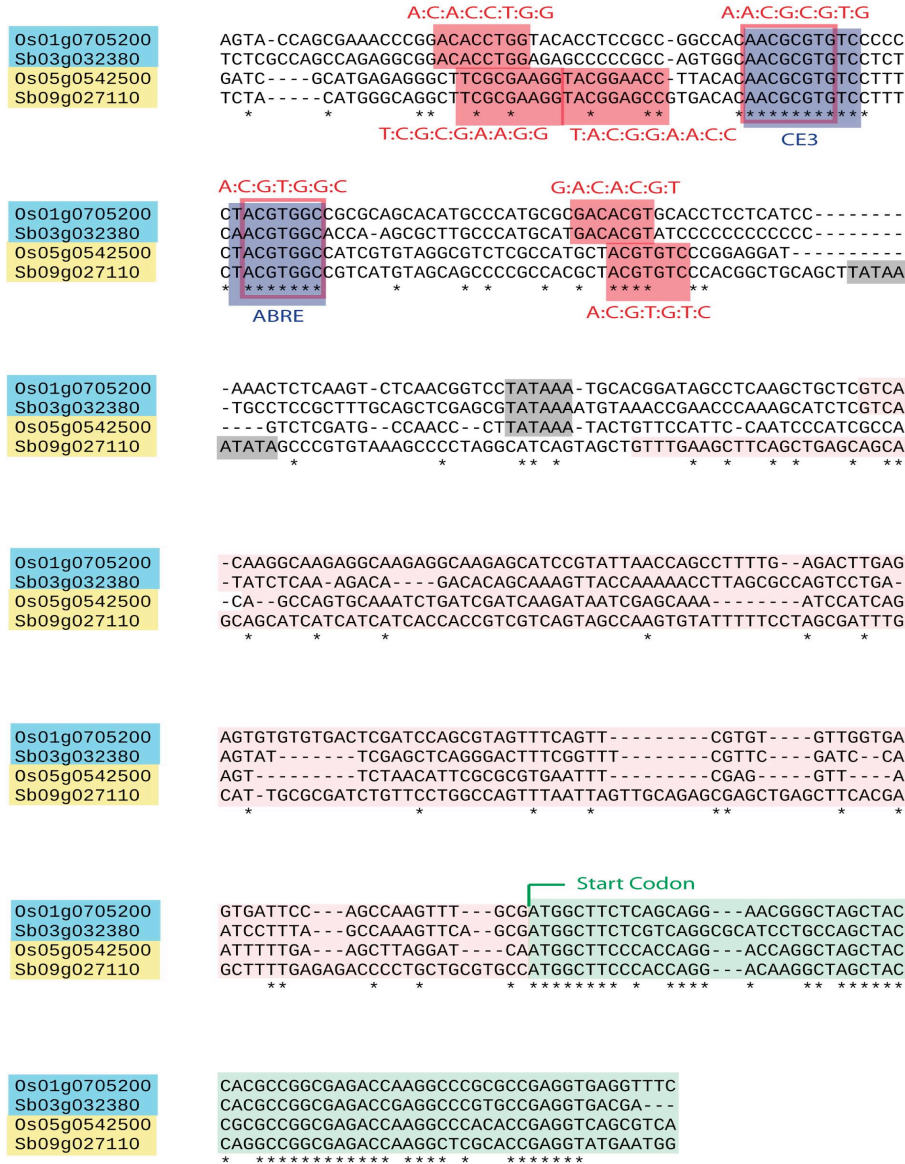


Figure 6: Co-conservation of two ABA-response elements in LEA promoters.

The alignment of two rice upstream sequences with their respective sorghum syntenic partners is shown. Pairs are marked as light blue and yellow frames. The rice genes Os01g0705200 and Os05g0542500 have been annotated as late embryogenesis abundant (LEA) proteins. Two known ABA response elements, ABRE and CE3 (dark blue frames), are highly conserved in position and inter-motif distance between all four promoters. Functionality of a similar motif arrangement has been reported for other ABA responsive rice genes. Additional motifs that are conserved between a rice-sorghum pair but not in all four promoters are shown as red frames and may indicate different responses between the pairs. Gray frames, pink and light green fragment depict potential TATA box, annotated transcript and annotated coding region, respectively.

several distinct cis-elements within a promoter (e.g. Levine et al, 2003). For functional elements that control transcriptional activities, it is therefore expected that the number of shared elements of a gene pair will positively correlate with its expression similarity. For rice promoter pairs, the relation of co-occurrence of “network-level” detected motifs and their expression congruency was analyzed. Expression similarity between a rice gene pair was measured by the Pearson correlation coefficient. To determine particular candidate motifs for a rice gene, all significant motifs that were present in both upstream sequences of a rice gene and its respective orthologous sorghum partner were selected. Pairs were binned according to the number of motifs they have in common (table 9 for MPSS data set), and for each bin the mean Pearson correlation from its members was determined. As shown in figure 7, a positive association between the number of shared motifs and the Pearson correlation coefficient for MPSS and YALE2 but not for YALE1 data was detected. Chi-square tests show significant deviations from independency (df=72; chi-square sums 3128 and 6046 for MPSS and YALE-2, respectively, p-value $< 10^{-16}$). Positive correlation was confirmed by a non-parametric, one-sided Wilcoxon rank test (p-value $p_{MPSS} < 10^{-16}$, $p_{YALE-2} < 10^{-16}$). YALE-1 was not significant (p-value $p_{YALE-1} \sim 1.0$). One explanation of weak correlation observed in YALE-1 data set could be that the amount of experiments included in the analysis was much larger and covered more diverse conditions and tissues compared to YALE-2 and MPSS data set. This might weaken the correlation that could potentially be indicated by subset of YALE-1 experiments. Nevertheless, the positive correlation between the number of shared elements of a gene pair and its expression similarity observed in YALE-2 and MPSS data sets is consistent with the combinatorial nature of transcription regulation and strongly indicates that a large fraction of detected motifs are associated with control of transcription.

	0	1	2	3	4	5	6	7	8
"-1 <> -0.8"	2	0	0	0	0	0	0	0	0
"-0.8 <> -0.6"	1431	272	55	6	1	1	0	0	0
"-0.6 <> -0.4"	132028	23001	3702	546	79	14	3	0	0
"-0.4 <> -0.2"	2539865	438897	63332	9676	1600	279	63	15	6
"-0.2 <> 0"	7679759	1364883	195354	29871	5163	1013	234	49	20
"0 <> 0.2"	5061393	924923	128443	19470	3180	679	138	31	15
"0.2 <> 0.4"	2819248	525909	73393	11014	1952	390	71	13	12
"0.4 <> 0.6"	1411328	267432	37523	5865	935	182	33	22	6
"0.6 <> 0.8"	585643	113000	15855	2627	446	81	21	11	6
"0.8 <> 1"	173582	33326	4943	845	166	44	11	2	5

Table 9: Number of rice gene pairs with respect number of their shared motifs associated with expression similarity

Columns of table show the number of detected motifs two rice genes can share, while rows indicate pearson correlation coefficient of rice gene pair measured using MPSS data set. Numbers of rice gene pairs are binned for the number of motifs they share associated with diverse intervals of pearson correlation coefficient.

Among the putative dyadic motifs, several dozens of long and unusually highly conserved motifs were discovered (appendix 8). These may provide highly specific site or several binding sites in close proximity to ensure a coherent regulation of the respective genes, at least in one biological process or response. For example, the detected motif CACGNGNTTTGAC is conserved in two WRKY transcription factors and a seven-helix transmembrane protein homolog to the Mlo1 gene from barley. WRKY transcription factors as well as the Mlo1 gene have been experimentally linked to primary pathogen responses (Panstruga, 2005; Eulgem et al, 1999). In another group, histones H2A and H2B as well as a high mobility group I/Y-2 are present. All these proteins are known to build or dynamically interact with chromatin structures. For the motif GCTCTNCNCNAAGA, conserved occurrences are found for enzymes of the phenylpropanoid- and lignin metabolism, two hydroxyanthranilate- hydroxycinnamoyltransferases and a ferulate-5-hydroxylase.

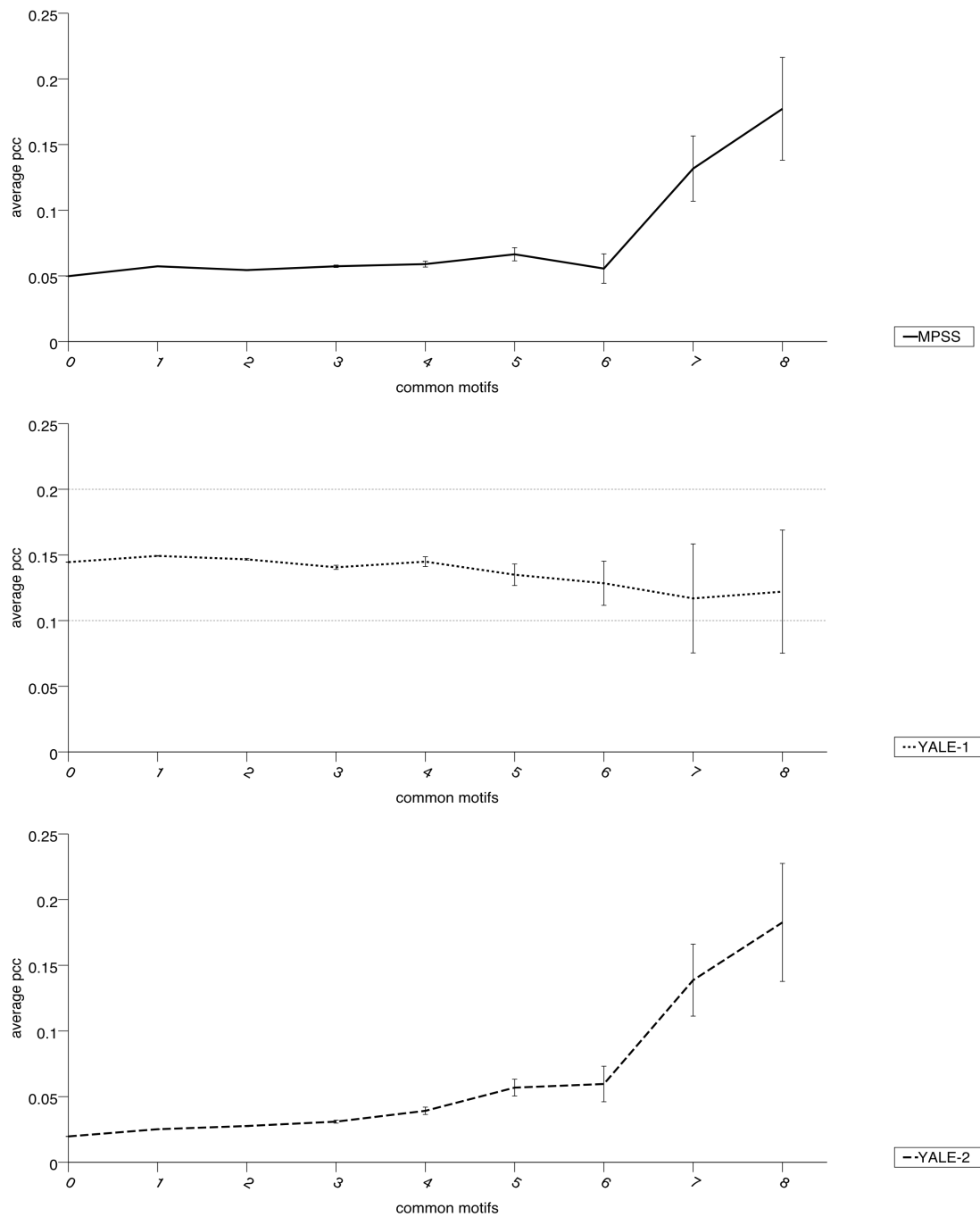


Figure 7: Dependency between number of shared motifs and expression similarity.

On the x-axis, the shared numbers of network-level conserved motifs for all pairwise comparisons of the 12,129 rice genes are shown. Common motif numbers are binned into 0, 1, 2.. up to 8 or more motifs shared between two rice upstream sequences. Expression similarities are provided on the y-axis as mean Pearson correlation coefficients with error bars for each bin. Figure 3 show results for expression data sets of MPSS, YALE-1 and YALE-2. Positive trends in MPSS and YALE-2 are significant.

In summary, both short and dyadic long candidate motifs were retained based on “network-level conservation” principle where motifs with an overrepresentation of

two standard deviations above the mean ($z\text{-score} > 2$) are considered as significantly evolutionary conserved between rice and sorghum. Short motif detection employed degeneration process while specification was applied to discover conserved dyadic long motifs. In total, 4,441 significant motifs (including 455 dyadic motifs) were detected of which several hundreds mapping to the verified regulatory binding sites have been obtained. The positive correlation between the number of shared elements of a gene pair and its expression similarity was observed. These evidences suggest the approach based on “network-level conservation” serve as a paradigm to cis-element detection in grasses and even in higher plants based on genome-wide evolutionary conservation between close related species.

2.1.3 Cis-regulatory element detection in *Arabidopsis thaliana* by “mutual” information among co-expressed gene groups

Transcription-related functional elements of non-coding sequences are commonly overrepresented in co-expressed gene groups to achieve their similar regulation of transcription (see introduction). This principle has been considered as a complementary approach to phylogenetic footprinting and has been widely applied to detect cis-regulatory elements. In particular, due to the lack of genome sequences in higher plants until recently, methods based on this principle can be employed for large-scale functional elements discovery within one completed genome, as long as associated expression data are available.

2.1.3.1 “Mutual” information and its application for cis-element detection by *FIRE*

Numerous approaches have been developed based on the principle of overrepresentation in co-expressed gene groups (see introduction). In this study, the motif detection package *FIRE* (finding informative regulatory elements) developed by Elemento et al (2007) was applied based on the concept of “mutual” information described firstly by Cover et al (2006). Instead of quantifying occurrence of a given motif in each co-expressed gene group separately according to classic approaches, *FIRE* estimates distribution of motif presence/absence among diverse co-expressed gene groups and analyses whether it is over-/underrepresented in one or several groups. Such relative representation is regarded as “mutual” information. In particular, those motifs whose patterns of presence/absence cross promoters of all members of particular gene group are most informative about the expression of the corresponding genes. Compared to general approaches, the advantages of *FIRE* based on mutual information are twofold. Firstly, functional motifs which may not overrepresented by surveying single gene group can be statistically prominent respecting overall groups, while widely distributed non-functional elements can be considered and avoided by investigating the distribution of motif occurrence among diverse co-expressed gene groups. This can increase the sensitivity and selectivity of cis-element prediction. Secondly, *FIRE* takes the information of motif absence into account which increases

the prediction power.

Briefly, *FIRE* starts by k-mers from all possible nucleotide alphabet combinations and for each k-mer, its mutual information was quantified as a mutual information score (Elemento et al, 2007). Next, all k-mers are sorted by their mutual information scores and the most informative ones are regarded as “seeds”. Other k-mers that provide little information over gene expression are disregarded to avoid redundant output. The optimization procedure is then carried out for each seed of k-mers by its degeneration and extension. Only changes that lead to more informative motifs were retained (Elemento et al. 2007). This procedure is repeated until no more informative motifs can be generated. The returned motifs are considered as final detected motifs for k-mers.

2.1.3.2 *Cis*-regulatory element detection in *Arabidopsis thaliana* using *FIRE*

FIRE has been proved as powerful tool for cis-element discovery for various types of experimental data from several organisms including yeast, fly, mouse and human (Elemento et al, 2007). For higher plants, detection of cis-elements in *Arabidopsis thaliana* has been undertaken, yet restricted by extremely limited expression data sets (Elemento et al, 2007). In this study, *FIRE* analysis using a much more comprehensive collection of microarray data for *Arabidopsis thaliana* was undertaken to extend and improve cis-element discovery by Elemento (2007). In total more than 2000 chips were collected from the *NASC* microarray database (see material and methods). They were designed to uncover the whole transcriptomes of *Arabidopsis thaliana* under different developmental stages, growth condition treatments, pathogen infection, stress series and hormone treatments. For each experiment, similarly expressed genes can be grouped and *FIRE* can be applied to discover candidate cis-elements that may be related to particular functional responses to the respective condition. However, similar experimental conditions are contained in the data sets and motif detection by *FIRE* for similar conditions may result in redundant analysis and results. Hence, according to their condition similarities, chips were manually grouped into 157 experimental clusters for the subsequently analysis (see material and methods; table 18). Expression data were subsequently normalized for each experimental cluster separately and remapped to the current genome template (TAIR8), in order to remove systematical errors, cross hybridization and avoid outdated probe mapping (see material and methods). Finally, overall 20,305 unique remapped genes are obtained which are significantly expressed in at least one experimental cluster, around half of which (10,442) are expressed in all of the 157 experimental groups (see material and methods; table 18).

The *FIRE* package was used for each experimental group separately to identify cis-regulatory elements that are functionally related with the respective condition. Genes with similar expression level were grouped, so that mutual information among gene groups monitoring diverse transcriptional activities can be captured by *FIRE* for

each k-mer motif. Its mutual information score can be calculated to represent functional conservation rate for the respective condition. In this study, the Chinese Restaurant Cluster (*CRC*) method, a model-based Bayesian clustering algorithm following iterative chinese restaurant process, was applied to identify co-expressed gene groups (see material and methods). The numbers of gene clusters in each experiment groups were summarized in the material and methods (table 18).

FIRE discovers motifs in promoter regions of 20,305 related genes. The search space was extracted as genomic sequences from their start codon to the start of their upstream preceding genes with a maximal distance of 2kb according to the TAIR8 *Arabidopsis thaliana* annotation. *FIRE* package applied for each experiment group to identify the most mutual informative motifs in promoter regions with respect of previously collected co-expressed gene groups. As the *FIRE* analysis integrate extension process to detect more informative motifs (see above), the sizes of investigated k-mers were set from 4 to 8 base pairs so that final detected motifs after motif extension can correspond with the size range of general plant cis-elements which is 6 up to 12 base pairs long (Matys et al, 2003; Bryne et al, 2007).

For a given experiment group, *FIRE* analysis calculated a bit score indicating mutual information for each possible k-mer (Elemento et al, 2007). As the analysis was undertaken for each k-mer size with the individual experimental groups, a direct comparison of scores among different motif sizes or diverse experiments is not applicable. To overcome this limitation, *FIRE* allows for a transformation process of bit scores to z-scores (Elemento et al, 2007) with which motifs of diverse sizes and predicted from different experiments were comparable. Overall 2080 candidate motifs that passed a randomization test were collected from 157 experiments. The length range of motifs is between 6 to 10 which corresponds with the general size of cis-elements in higher plants. The minimal and median z-score is 6.2 and 11.9, respectively, suggesting high mutual information of candidate motifs. Only 44 out of 2080 (2%) detected motifs represent specific sites and are enriched in motifs with short sizes (6- and 7-mer), while the overwhelming majority of motifs (98%) contains variability (table 10). 1048 (50%) motifs even contain position(s) with complete enumeration of nucleotide alphabets. Moreover, Chi-square test showed significant deviations from independency between grade of variability and size (table 10; p-value $< 10^{-16}$) and indicates that the *FIRE* analysis simultaneously favors degeneration and extension to discover long motifs with high variability.

	1	2	3	4	Total
6	22	30	21	7	80
7	15	72	90	61	238
8	5	87	166	189	447
9	2	58	233	328	621
10	0	12	219	463	694
Total	44	259	729	1048	2080

Table 10: Binned numbers of FIRE motifs with respect of motif length and degeneration degree

Table summarizes the binned numbers of FIRE detected motifs according to their sizes associated with the maximal number of alphabets motifs have reached at any position

FIRE analysis was applied for diverse sizes and under different experimental conditions separately. Due to overlapping of seed lengths and repeated detections from different experiments, detected motifs may show redundancy. Similar filtering criteria as employed for motif detection by the “network-level conservation” approach in section 2.1.2.1 was used to reduce the redundancy. Motifs were ordered according to their mutual information score and all motifs that were derived from a higher scoring motif were removed (see material and methods). Finally 271 motifs in the size range of 7 to 10 bp, i.e. 13% out of a total of 2080 detected motifs, were found to represent non-redundant motifs. Several reasons might explain the high redundancy rate of the *FIRE* analysis for *Arabidopsis thaliana* in this study. Firstly, experiment clusters grouped manually from single arrays may still share similar treatment conditions. This can result in redundant gene cluster partitions and informative motifs detected by the analysis. In fact the comparison of all experiment groups against the results from all comparison indicate a 26% motif overlap on average (fig. 8). Secondly, prediction of motifs undertaken for size variants, e.g., 4 to 8-mer in this study, is a resource for motif redundancy and well known as a shortcoming of approaches based on word/k-mer analysis for cis-element detection (Guhathakurta 2006). Thirdly, as *FIRE* detects motif over-/underrepresentation among different gene clusters, the quality of the clustering process plays an essential role. However, pronounced unequal distribution of gene cluster sizes was derived within each of the 157 experiments. Figure 9 shows that a large portion of genes were grouped into one single cluster while other genes were partitioned into small clusters. This can lead to gene cluster overlapping and redundant informative motifs among different experiments. Moreover, leaving the matter of possibly problematic experiment and co-expressed gene grouping, genes and/or gene modules can indeed be shared by different experiments and redundant cis-elements are therefore expected in such cases.

Functionalities of individual cis-regulatory elements may be affected by their distance to transcription start sites (Beer et al, 2004). Thus, constraint on positions of particular motifs is expected in their functional related co-expressed gene groups. Survey of such constraint can be achieved by *FIRE* analysis which identifies whether position

bias of detected motifs is present in particular co-expressed gene groups compared to in others. In total 55 out of 271 predicted motifs demonstrate a position bias according to *FIRE* analysis.

Appendix 9 summarizes *FIRE* detected 271 motifs in *Arabidopsis thaliana* with their z-scores of mutual information and position bias. Table 11 shows several examples. The minimal and median z-score is 6.2 and 9.9, respectively. Though still highly

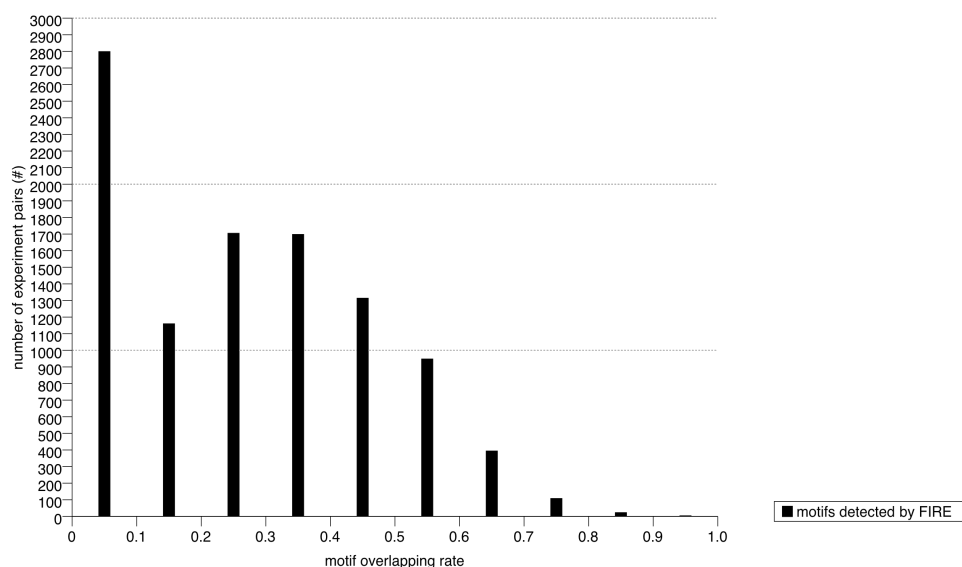


Figure 8: Distribution of motif overlapping rate for motifs detected by FIRE from diverse experiments

Bins of motif overlapping rate depicted on x axis is defined as the ratio between the shared number and union number of motifs detected by two experiments. Overlapping rate has been calculated for each experiment pair with all-against all comparison. Number of experiment pairs of each bin of motif overlapping rate is shown on the y-axis.

Motif	Z-score	Position bias	Known site mapping	
			<i>AGRIS</i>	<i>PLACE</i>
GT:AC:C:A:C:G:T:AG	66.565	Y	Y	Y
CT:A:C:G:T:G:GT:CT:ACG	66.751	Y	Y	Y
ACT:CT:A:C:G:T:G:GT:CT:ACG	66.704	Y	Y	Y
AGT:A:A:T:A:A:T:ACT:G:AGT	9.029			
ACGT:ACGT:A:C:C:C:AG:G:AT:ACGT	12.601		Y	Y
ACT:CG:T:AC:C:G:T:C:A:ACT	9.719		Y	Y
ACGT:A:G:C:T:C:C:ACGT	10.313			
ACT:C:T:T:A:T:C:C:AGT:ACGT	16.433	Y		Y
ACGT:T:A:C:AG:C:AC:G:CG:ACT	9.367		Y	Y

Table 11: Motifs of *A. thaliana* detected by FIRE

Table lists several motifs of *A. thaliana* detected by FIRE with their respective z-scores. Overrepresentation of position bias and match to known sites of AGRIS and PLACE databases are signed as “Y” in the corresponding columns of the table. The complete list is present in the appendix 9.

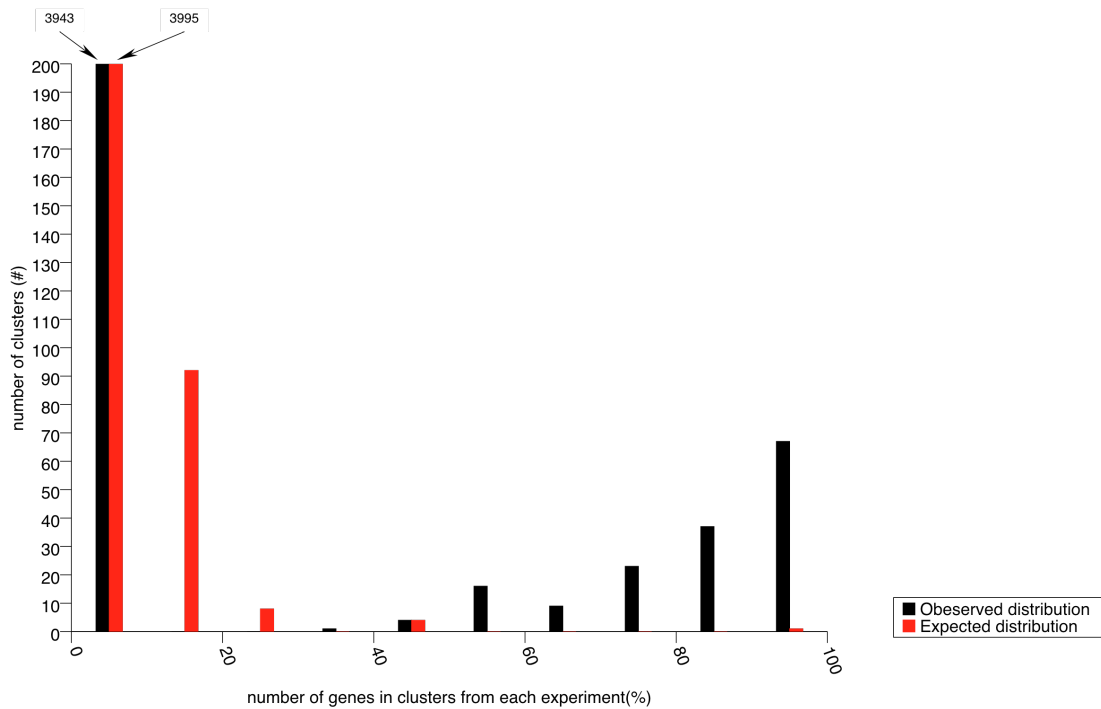


Figure 9: Distribution of expected and observed gene cluster sizes

Observed size of a cluster is represented as the fraction of all genes which are involved in particular experiment and contained in the respective cluster determined by CRC clustering approach for the respective experiment (in percent and depicted on x-axis), while expected size of a cluster is estimated under assumption that genes are equally distributed among gene clusters. Observed and expected number of clusters with different sizes from all experiments in this study is shown on the y axis in black and red bars, respectively.

informative, a bias towards lower z-scores was observed compared to initially deduced motifs by *FIRE* (fig. 10; one sided wilcox rank sum test $p=5 \times 10^{-13}$). In addition, all or large fraction of initial motifs with small sizes, i.e. 6 and 8-mers, have been removed and the final motifs contain a higher enrichment in the range of higher degeneration and larger sizes (table 12).

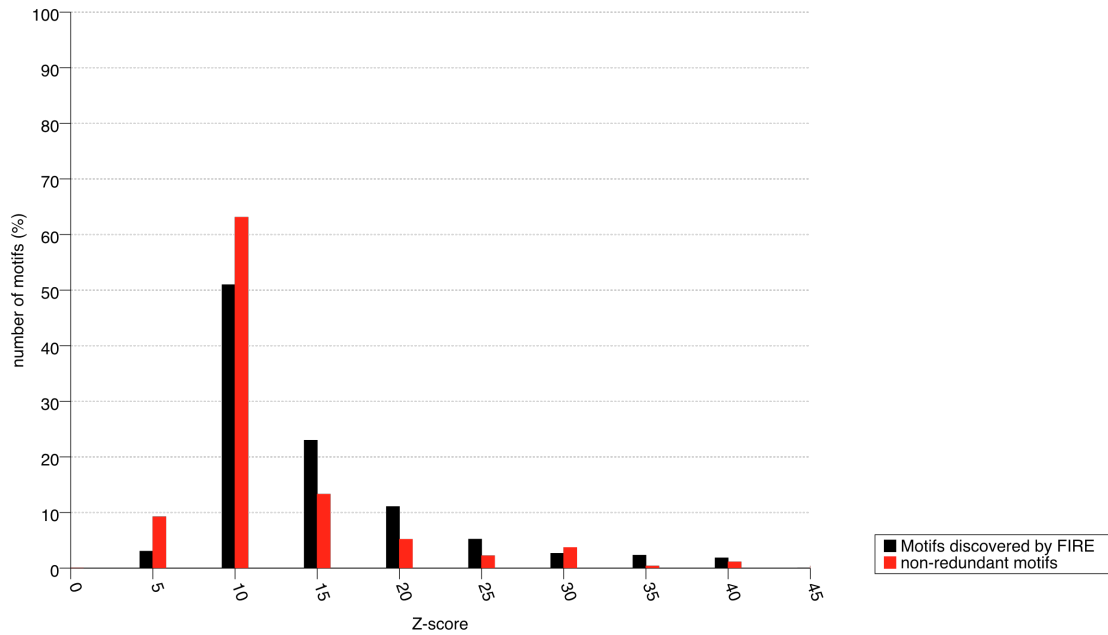


Figure 10: Distribution of z-scores for motifs initially deduced by FIRE and non-redundant motifs

Bins of z-scores are depicted on x axis. With respect of each z-score bin, fraction of motifs initially predicted by FIRE and non-redundant motifs is shown on y axis in black and red bars, respectively.

	1	2	3	4	Total
6	0	0	0	0	0
7	0	0	1	1	2
8	1	3	4	12	20
9	1	2	23	30	56
10	0	6	59	128	193
Total	2	11	87	171	271

Table 12: Binned numbers of non-redundant motifs with respect of motif length and degeneration degree

Table summarizes the binned numbers of final non-redundant motifs according to their sizes associated with the maximal number of alphabets motifs have reached at any position.

2.1.3.3 Validation of detected motifs

Similar to previous analysis of cis-elements in rice and sorghum, *FIRE* detected motifs were firstly evaluated by their mapping to experimentally verified bindings sites. In contrast to grass genomes, *Arabidopsis thaliana* has been studied more comprehensively and more evidences of its binding elements are available. In this study, evidences were extracted from the public AGRIS (Davuluri et al, 2003) and PLACE (Higo et al, 1999) databases. The former one contains specific functional

binding sites while the latter database contains more generalized binding models which reflect possible attachment variability of transcription factors (e.g. G-box: MCACGTGGC, M=A/C). After removing identical sites, overall 407 sites from AGRIS and 135 motifs from PLACE were used to map and compare motifs detected by the *FIRE* analysis. Employing similar criteria of mapping process as used for evaluation of “network-level conservation” analysis of rice and sorghum, 60 (22.1%) and 152 (56.1%) out of the 271 predicted motifs matched 72 verified sites from AGRIS database and 66 known motifs from PLACE database, respectively. These include e.g. light-responsive elements like CDA-1 binding site in dark response element f of chlorophyll a/b-binding protein2 gene in *Arabidopsis*, or stress-related elements like binding sites responsible for drought, low-temperature or high-salt stress. In general observations suggested that, *FIRE* analysis leads to high-confident *Arabidopsis thaliana* motif detection and is prominent on the selectivity, while motif detection by “network-level conservation” approach favors sensitivity over selectivity where overall 3,809 motifs were discovered but only 559 (14.7%) matched to known sites. The motifs mapped to known sites were summarized in appendix 9 (table 11). In addition, motifs which were mapped to reported motifs display significantly higher information content (higher z-scores) in comparison with the ones matching no known sites (fig 11: $p=5.884 \times 10^{-15}$ for AGRIS and $p=9.457 \times 10^{-10}$ for PLACE; wilcox rank sum test). This is consistent with the fact that compared to randomly or nonspecific distributed occurrence, verified motifs were highly over-/underrepresented in defined gene clusters under particular condition in order to realize regulation of transcriptional activities of corresponding coherent genes.

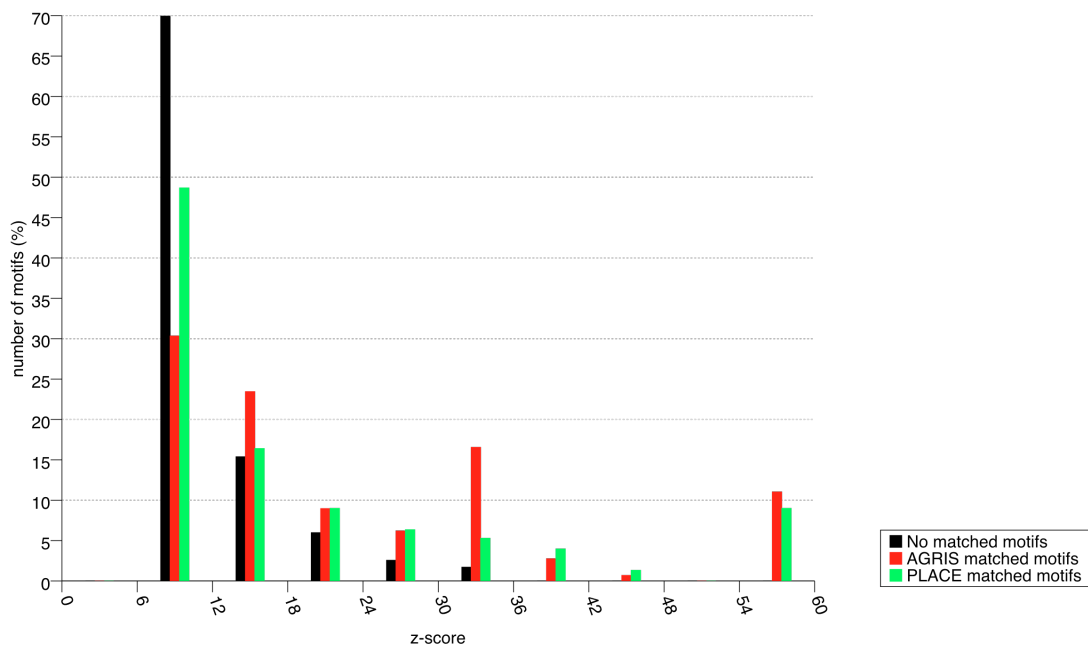


Figure 11: Distribution of z-scores for motifs matching and not matching to reported sites
 Bins of z-scores are depicted on x axis. With respect of each z-score bin, fraction of motifs matching and not matching to reported sites is shown on y axis in black and red/green bars, respectively.

As discussed in 2.1.1.6, the principle of gene ontology (GO) enrichment which is widely used to verify detected motifs can not be applied for evaluation of cis-elements in rice and sorghum due to the limited availability of their genome-wide GO annotations. However, this principle can be used to validate detected motifs of *Arabidopsis thaliana* thanks to its rich and highly curated gene ontology annotation. GO annotation was imported from the GO databases according to TAIR8 annotation (Berardini et al, 2004). Overall 2,917 GO terms that belong to the category ‘biological processes’ cover 23,860 genes (73% of all genes). Employing a cumulative binomial probability distribution test with adjustment applying the Benjamini-Hochberg method (Benjamini et al, 1995) to correct for a false discovery rate of 5%, 24 discovered motifs were overrepresented in at least one GO term (Table 13). Notably, more than half of the overrepresented motifs have been successfully matched to verified sites and dozens indicate exactly the functionalities these verified sites reported. For instance, the detected motif CT:A:C:G:T:G:GT:CT:ACG has been matched to the ABRE-like binding site TACGTGTC (Shinozaki et al, 2000) and was enriched in the gene sets associated with the biological process category classified as response to low abscisic acid stimulus (table 13). Moreover, detected motifs mapping to reported sites tend to represent enrichment in more than one biological process which are strongly functional-related. The predicted motif AG:C:C:G:A:C:ACG:T:ACG:ACGT, for example, has mapped to LTRE promoter motif ACCGACA which are related to differential expression of low-temperature-induced genes in *Arabidopsis thaliana* (Nordin et al, 1993). This motif indicates overrepresentation in particular biological processes (GO:0009266, GO:0009409, GO:0009414, GO:0009415, GO:0009631) that describe response to temperature stimulus and water deprivation (Table 13). Interestingly, several CGT:C:GT:A:G:A:A:C:ACGT like motifs are enriched for that describe genes responding to light, heat or hydrogen peroxide treatments do not match to known sites and may indicate novel functional motifs.

Motif	Known site	GO id	GO description	P-value
ACGT:ACT:ACG:G:G:C:C:C:A:ACGT	G:G:G:C:C	GO:0006412	translation	0.00186909
		GO:0006996	organelle organization	0.00374111
		GO:0010467	gene expression	0.01898328
		GO:0022613	ribonucleoprotein complex biogenesis and assembly	1.14E-07
		GO:0034961	cellular biopolymer biosynthetic process	0.04944211
ACGT:ACT:C:CT:T:A:T:C:C:ACGT	-	GO:0042254	ribosome biogenesis	5.03E-08
		GO:0015979	photosynthesis	0.00061349
ACGT:ACT:CT:T:C:C:C:G:C:CGT	T:T:T:C:C:C:G:C	GO:0006139	nucleobase	
		GO:0006259	DNA metabolic process	2.89E-13
		GO:0006260	DNA replication	8.88E-15
		GO:0006261	DNA-dependent DNA replication	3.57E-10
ACGT:AT:A:A:G:T:C:A:A:ACT	-	GO:0006270	DNA replication initiation	5.99E-08
		GO:0031347	regulation of defense response	0.02602009
ACT:ACG:G:G:C:C:C:A:ACT:ACGT	G:G:G:C:C	GO:0006412	translation	1.67E-14
		GO:0009059	macromolecule biosynthetic process	2.62E-06
		GO:0010467	gene expression	1.72E-11
		GO:0022613	ribonucleoprotein complex biogenesis and assembly	0.00044511
		GO:0034645	cellular macromolecule biosynthetic process	1.80E-06
		GO:0034960	cellular biopolymer metabolic process	3.67E-07
		GO:0034961	cellular biopolymer biosynthetic process	1.69E-10
		GO:0042254	ribosome biogenesis	1.87E-05
ACT:AG:G:A:T:A:AG:G:AG:ACGT	G:A:T:A:A:G	GO:0043284	biopolymer biosynthetic process	2.75E-10
		GO:0044249	cellular biosynthetic process	0.04865509
ACT:AG:G:C:C:A:C:A:ACT	A:G:C:C:A:C	GO:0015979	photosynthesis	5.04E-05
ACT:C:CT:T:A:T:C:C:ACGT	-	GO:0015979	photosynthesis	0.00014579
ACT:C:G:T:A:G:F:A:A:AG	-	GO:0015979	photosynthesis	0.00061349
ACT:C:G:T:A:G:F:A:A:AG	-	GO:0015976	carbon utilization	0.02953305
ACT:C:T:T:A:T:C:C:AGT:ACGT	-	GO:0015979	photosynthesis	0.00327716
ACT:CT:A:C:G:T:G:GT:CT:ACG	T:A:C:G:T:G:T:C	GO:0009414	response to water deprivation	0.00454429
		GO:0009415	response to water	0.01145528
ACT:CT:G:C:C:A:C:CG	-	GO:0006091	generation of precursor metabolites and energy	0.0204558
		GO:0009767	photosynthetic electron transport chain	0.00434638
		GO:0009773	photosynthetic electron transport in photosystem I	0.01440273
		GO:0015979	photosynthesis	8.19E-19
		GO:0019684	photosynthesis, light reaction	0.0111936
		GO:0022900	electron transport chain	0.00391057
AG:C:C:G:A:C:ACG:T:ACG:ACGT	A:C:C:G:A:C:A	GO:0055114	oxidation reduction	0.01527634
		GO:0009266	response to temperature stimulus	0.00095763
		GO:0009409	response to cold	1.47E-05
		GO:0009414	response to water deprivation	0.00202607
		GO:0009415	response to water	0.00323445
AGT:A:AC:AC:A:C:G:T:G:ACGT	C:C:A:C:G:T:G:G	GO:0009631	cold acclimation	0.00018648
		GO:0009719	response to endogenous stimulus	0.0384034
		GO:0009753	response to jasmonic acid stimulus	1.41E-08
		GO:0016137	glycoside metabolic process	0.00032956
AGT:AC:C:A:C:G:T:G:ACGT:ACGT	C:C:A:C:G:T:G:G	GO:0019757	glycosinolate metabolic process	0.00032956
		GO:0019760	glucosinolate metabolic process	0.00032956
		GO:0015979	photosynthesis	0.00013185
AGT:AG:A:A:C:C:C:T:A:AG	A:A:A:C:C:C:T:A:A	GO:0006412	translation	0.00019668
		GO:0006996	organelle organization	0.00319668
		GO:0010467	gene expression	4.14E-08
		GO:0016070	RNA metabolic process	0.00134329
		GO:0022613	ribonucleoprotein complex biogenesis and assembly	9.79E-07
		GO:0034960	cellular biopolymer metabolic process	1.20E-05
		GO:0034961	cellular biopolymer biosynthetic process	0.00128152
		GO:0042254	ribosome biogenesis	2.31E-05
AGT:GT:AC:C:A:C:G:T:AC:ACGT	A:G:C:C:A:C	GO:0043284	biopolymer biosynthetic process	0.00238509
		GO:0009266	response to temperature stimulus	0.02138697

CGT:ACGT:G:A:T:C:GT:G:AG:AGT	-	GO:0016192	vesicle-mediated transport	0.00011675
CGT:C:GT:A:G:A:A:C:ACGT	-	GO:0000302	response to reactive oxygen species	0.00164759
		GO:0009266	response to temperature stimulus	0.00011961
		GO:0009408	response to heat	5.38E-08
		GO:0009642	response to light intensity	0.00469168
		GO:0009644	response to high light intensity	0.0008604
CGT:GT:AG:C:C:G:A:C:ACGT:T	T:G:G:C:C:G:A:C	GO:0042542	response to hydrogen peroxide	0.00050167
CGT:T:C:G:A:G:AC:A:AC	-	GO:0009631	cold acclimation	0.00210073
		GO:0000302	response to reactive oxygen species	0.04200023
		GO:0009644	response to high light intensity	0.0259315
CT:A:C:G:T:G:GT:CT:ACG	T:A:C:G:T:G:T:C	GO:0042542	response to hydrogen peroxide	0.01739061
		GO:0009414	response to water deprivation	0.0004613
GT:AC:C:A:C:G:T:AG	C:A:C:G:T:G	GO:0009415	response to water	0.00125418
		GO:0009737	response to abscisic acid stimulus	0.0004551
		GO:0009414	response to water deprivation	0.00812357
T:T:A:A:T:T:A:ACT	-	GO:0009415	response to water	0.02112891
		GO:0009737	response to abscisic acid stimulus	0.00019711
T:T:A:A:T:T:A:ACT	-	GO:0009719	response to endogenous stimulus	0.01377222
		GO:0009725	response to hormone stimulus	0.03924843

Table 13: FIRE motifs in *A. thaliana* associated with biological processes

Table lists 24 motifs in *A. Thaliana* detected by FIRE which are enriched in gene set associated with at least one biological process. The respective GO identity, functional description and p-values are shown in the last three columns. The second column indicates reported elements which are mapped by particular detected motifs.

2.1.3.4 Evolutionary conservation of functional candidate motifs detected by FIRE

FIRE analysis discovers the motifs that are over-/underrepresented in co-expressed genes within one species and analyses their potential function of similar transcription regulation under particular conditions. According to the principle of phylogenetic footprinting, the functional candidate motifs detected by *FIRE* are expected to tend to be maintained during evolution and conserved in closely related species compared to non-functional ones. In order to investigate if the functional candidate motifs of *Arabidopsis thaliana* detected by *FIRE* are maintained during evolution, motifs that are evolutionary conserved were analysed between *Arabidopsis thaliana* and the recently completed genome of *Arabidopsis lyrata*, and then mapped by the *FIRE* motifs so that their evolutionary conservation rates can be measured.

Recently the genome of *Arabidopsis lyrata* has been finished (Hu et al, 2011). It is evolutionary closely related to *Arabidopsis thaliana*. The two species diverged about 10 Mya, but the *A. thaliana* genome with 125 Mb is much smaller than that of *A. lyrata* which is more than 200 Mb (Hu et al, 2011). Despite of an apparent shrinkage in genome size, overall sequence identity between both species exceeds 80% and some 90% of genomes have remained syntenic with the vast majority in highly conserved collinear arrangements (Hu et al, 2011). The syntenic relationship allows to use the genomes of *Arabidopsis thaliana* and *Arabidopsis lyrata* for comparative genomic analysis. *FASTCOMPARE* based on “network-level conservation” described in section 2.1.2.1 was applied to identify phylogenetic footprints between *A. thaliana*

and *A. lyrata* (see material and methods). Similar to the motif detection in rice and sorghum, z-scores of given motifs transformed by ratios between observed and expected co-occurrence corresponds to genome-wide conservation rate between *A. thaliana* and *A. lyrata* and represents the conservation rate of corresponding detected motifs. As a complementary estimator to the z-score, conservation of predicted motifs can be accessed by comparing their ratio scores to the scores of all corresponding k-mers, i.e. background scores. A conservation index of a given motif is defined as the fraction of all corresponding k-mers that contained lower ratio scores than the score of the corresponding motif. For instance, the conservation index of 0.75 from detected motif G:C:C:A:C:G:T:A implies that this motif is more conserved, i.e., higher scored, than 75% of all 8-mers.

Based on sequence similarity, motifs detected by *FIRE* were subsequently mapped to *FASTCOMPARE* motifs so that their evolutionary conservation rate can be monitored. A successful match was defined if a *FIRE* detected motif contains all words of a *FASTCOMPARE* motif. If more than one match was derived, the *FASTCOMPARE* motif with the highest score was retained. Following the criteria, All 271 *FIRE* predicted motifs were mapped to at least one *FASTCOMPARE* detected motif (Table 14; appendix 10). Further comparisons between all motifs predicted by *FASTCOMPARE* as background and the ones mapped to the *FIRE* motifs were undertaken using z-scores and conservation indexes. Figure 12 shows that among all evolutionary conserved motifs between *A. thaliana* and *A. lyrata*, the ones matching to *FIRE* motifs display significantly higher conservation rate ($p < 2.2e-16$ of Wilcoxon rank sum test for both score types). This indicates that candidate motifs detected by *FIRE* indeed favor to be preserved during evolution compared to background non-coding sequences. Therefore, their functionalities are strongly suggested.

<i>FIRE</i>		<i>FASTCOMPARE</i>			
Motif	Z-score	Motif	Score	Z-score	Conservation index
GT:AC:C:A:C:G:T:AG	66.57	G:C:C:A:C:G:T:A	38.73	0.28	0.769673938
AT:A:C:G:T:T:C:T:ACGT	7.89	T:A:C:G:T:T:C:T:G	104.66	-0.01	0.660162761
AGT:ACGT:AT:C:T:T:A:T:C:CT	14.56	G:T:A:C:T:T:A:T:C:C	480.97	0.80	0.862648503
ACGT:AGT:AG:C:A:G:A:T:C:ACGT	7.73	C:G:A:C:A:G:A:T:C:G	1251.5	3.63	0.990349775
AGT:C:A:G:A:G:A:A:AGT	7.92	G:C:A:G:A:G:A:A:T	33.68	-0.58	0.261126656
AG:A:A:G:A:G:AG:G:AT:GT	9.01	G:A:A:G:A:G:G:T:G	194.29	-0.25	0.537141307
ACGT:A:G:A:T:T:C:C:CT:ACGT	10.08	G:A:G:A:T:T:C:C:C:C	842.36	2.13	0.963195191
ACGT:G:ACT:T:T:T:C:CG:CG:AC	11.74	C:G:T:T:T:T:C:G:G:C	730.04	1.72	0.948703702

Table 14: Mapping FIRE detected motifs to motifs discovered by network-level conservation

Table lists motifs in *A. thaliana* detected by *FIRE* (left panel) and their corresponding motifs in *A. thaliana* and *A. lyrata* discovered base on network-level conservation principle (right panel). The complete list is represent in the appendix 10.

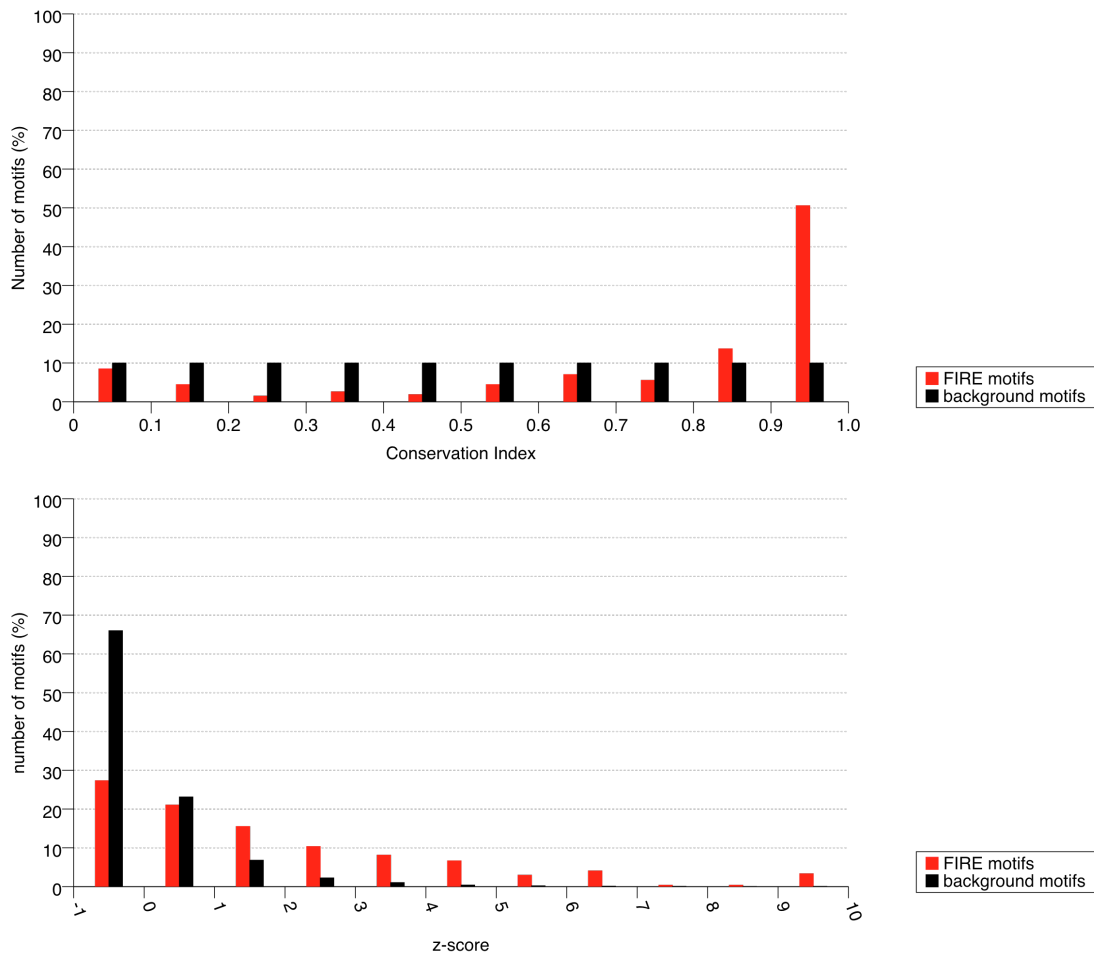


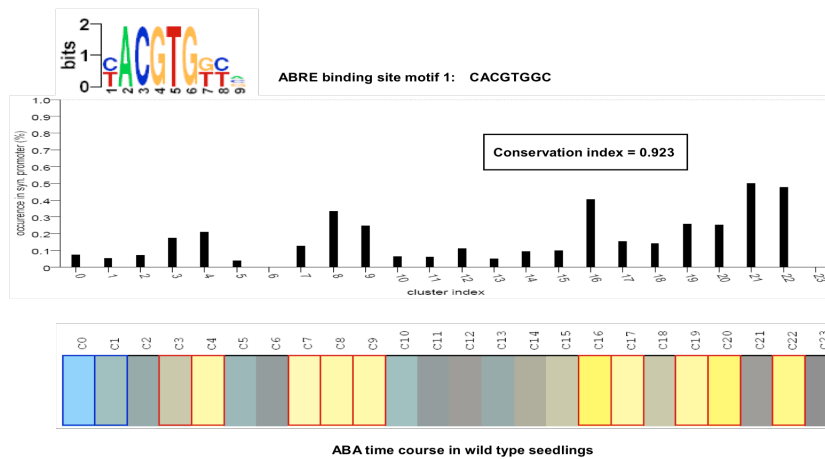
Figure 12: Distribution of evolutionary conservation rate

Figure shows the distribution of evolutionary conservation rate for FASTCAMPARE motifs which have also been detected by FIRE (red bars) and all FASTCOMPARE motifs (black bars). Fraction of motifs is depicted on y axis with respect of each bin of conservation index (upper panel) and z-score (lower panel).

Remarkably, functional candidate motifs predicted by *FIRE* that show a tendency of “global” evolutionary conservation between *A. thaliana* and *A. lyrata* do not demonstrate an equally high conservation rate in each co-expressed gene group. A differential rate of conservation was observed among co-expressed gene groups for most *FIRE* motifs. In fact, most candidate motifs show higher evolutionary conservation rate in the co-expressed gene group in which they have been found to be overrepresented, while less conservation is expected in the gene group where they are underrepresented. This positive correlation between evolutionary conservation and level of representation was confirmed by one-sided correlation test using Pearson’s product moment correlation coefficient ($p < 10^{-16}$). For instance, as shown in figure 13A, the known cis-regulatory element CACGTGGC that is reported as ABRE binding site motif was detected from the experiment “ABA time course in wt seedlings” by *FIRE*. Co-occurrence of the motif between promoter regions of *A.*

thaliana and *A. lyrata* was calculated within each gene group and used to estimate the respective conservation rate. Figure 13A shows that, with the exception of gene cluster 21, degrees of overrepresentation in different functional gene groups were consistent with the conservation rate from corresponding gene clusters. The similar observation was obtained even for relatively lower “global” evolutionary conserved motifs. For instance, the predicted telo-box promoter motif AAACCCTAA showed such correlation of conservation and overrepresentation, although a relatively lower “global” conservation was observed (conservation index = 0.663) (figure 13B). Such positive correlation between overrepresentation and evolutionary conservation within respective co-expressed gene groups suggests that functional candidate motifs predicted by FIRE are not only under constraints of selection, but also favor to be preserved between *A. thaliana* and *A. lyrata* in the gene groups where they occur frequently and may play an important role for transcription regulation of respective genes.

A)



B)

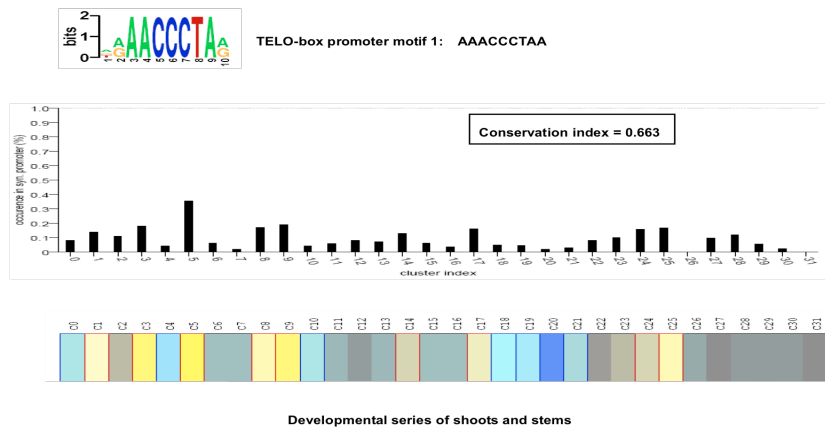


Figure 13: two examples of FIRE detected motifs and distribution of their conservation rate and overrepresentation

Two examples of FIRE detected motifs that have been mapped to experimentally verified regulatory elements are shown in the figure. Conservation rate between *A. thaliana* and *A. lyrata* of both motifs were inspected in each gene cluster determined by FIRE from respective experiment (lower panel in both examples). Conservation rate of motif was estimated by its co-occurrence within a gene cluster which is defined as fraction of syntenic pairs in the respective cluster of which both species contain the motifs. Conservation rate for each cluster is depicted on y axis in the middle panel of both examples and compared to degree of motif over-/underrepresentation (deduced by FIRE analysis with different color representing diverse degrees of representation) for the respective gene cluster (lower panel in both examples).

2.2. Evolutionary analysis of cis-regulatory elements in plants

For decades, numerous studies on various organisms have identified cis-regulatory mutations with functionally significant consequences for morphology, physiology and behavior (see introduction). These analyses of trait divergence have demonstrated evolutionary changes at loci for which functional changes in the protein sequence have been ruled out and functional cis-regulatory mutations have been implicated or directly demonstrated. This demonstrates that evolution of cis-regulatory elements contributes to phenotypic changes among closely related species. With increasing knowledge of verified cis-regulatory elements and previous prediction of novel cis-regulatory elements, computational analysis in this study addressed a genome-wide survey of cis-element mutations and their potential functional consequences for trait changes in higher plants. The changes of cis-acting regions in duplicated loci generated during evolution (see below) were surveyed in the study, as duplicated genes with high sequence similarity suggest minor functional coding changes and trait difference may mainly be caused by cis-element mutations to alter expression of the respective genes.

Duplicated (paralogous) loci can be generated by a gene duplication event. Orthologous genes on the other hand are created from a speciation event and define functional equivalent genes among different species. Evolution of duplicate genes is considered as a source of phenotypic changes among species during speciation or generation of evolutionary novelties like neo- or subfunctionalization after gene duplication (Ohno, 1970). On the other hand, a pair of duplicated genes can also diverge in function as a result of changes in regulatory elements (Force et al, 1999). Instead of alterations of protein function and/or structure, phenotypic changes and generation of evolutionary novelties can be achieved by mutation of various sites. These sites, such as splice sites or cis-regulatory elements, govern the functional characteristics of the respective gene based on the duplication-degeneration-complementation (DDC) model (Force et al, 1999). The temporal or spatial gene expression under different stimuli can be changed, although the coding regions were preserved during evolution. Mutation of functional binding sites in promoter regions of duplicated loci, both orthologs and paralogs, were therefore analyzed in this study to unveil the potential effect of their cis-elements mutations on phenotypic variations.

Two plant models have been selected to survey the evolution of cis-regulatory element and promoters in this study. In the first case, the evolutionary closely related species *A. thaliana* and *A. lyrata* were chosen as models to address large-scale analysis of cis-element evolution in higher plants. Their evolutionary history and genomic collinear arrangement have been well characterized (Hu et al, 2011), so that identification of paralogs and orthologs can be achieved. Furthermore, numerous predicted and experimentally verified cis-regulatory sites in *A. thaliana* and *A. lyrata* have been reported (see 2.1.3) which can be used for comprehensive survey of

divergence and conservation of cis-elements in orthologous and paralogous gene pairs. In addition, a wealth of *A. thaliana* expression data sets (see 2.1.3.2) makes it feasible to monitor changes of transcriptional activities of duplicated genes and to relate to cis-element evolution take place. As another case study of cis-element evolution, investigation of specific compatibility situations between plastid and nuclear genomes in the flowering plant genus *Oenothera* were undertaken. Polymorphism of diverse genetic determinants, e.g., promoter or protein coding regions, in close related *Oenothera* species caused by speciation were uncovered which can be seen as candidates to plastome-genome incompatibility. The analysis for the first time addresses a large-scale search for plastid-specific genetic determinants that are likely to be causal to compartmental co-evolution in higher plants.

2.2.1 Evolution of cis-regulatory elements during duplication and speciation in *A. thaliana* and *A. lyrata*

Evolution of cis-regulatory elements in higher plants was studied by comparing the *A. thaliana* genome and the recently completed *A. lyrata* genome (*Arabidopsis* Genome Initiative, 2000; Hu et al, 2011). Their phylogenetic relationship has been well characterized. The speciation process took place about 10 million years ago (Hu et al, 2011). Overall high sequence similarity and synteny between both genomes have been discovered (see introduction) which provides a large number of orthologous gene pairs (Hu et al, 2011). Furthermore, ancient polyploidization events that took place in the lineage have been well characterized (Simillion, et al., 2002; Blanc, et al., 2003; Bowers, et al., 2003). Two polyploidization events, called as *Arabidopsis* alpha and beta duplications, occurred approximately 14.5-86 Mya and 112-235 Mya, respectively (Simillion, et al., 2002; Blanc, et al., 2003; Bowers, et al., 2003). Since gene duplication by both polyploidization events occurred earlier than speciation of *A. thaliana* and *A. lyrata*, a higher sequence conservation rate between *A. thaliana* and *A. lyrata* orthologs compared to paralogous copies in the *A. thaliana* and *A. lyrata* genomes can be expected.

To survey evolution of cis-elements in *A. thaliana* and *A. lyrata*, orthologous and paralogous gene pairs were identified. Cis-element motifs of *A. thaliana* detected by the analysis detailed in 2.1.3.2 were used as target sites and mapped to promoters of detected orthologous and paralogous gene pairs. The conservation rate of each motif was estimated in orthologous and paralogous gene pairs by analyzing the number of their shared motifs. The study provides a first start for investigating evolution of cis-regulatory motifs with respect of gene duplication and speciation process.

2.2.1.1 Identification of orthologous and paralogous genes in *A. thaliana* and *A. lyrata*

Both orthologous and paralogous gene pairs were determined by sequence homology-based searches. 17,521 orthologous gene pairs identified by the work of Hu

et al in 2011 were used as orthologous gene set between *A. thaliana* and *A. lyrata*. To determine the paralogous genes, duplicated gene groups were separately selected from *A. thaliana* and *A. lyrata*. The groups were either within segmental duplications originating from ancient polyploidization event (Simillion, et al., 2002; Blanc, et al., 2003; Bowers, et al., 2003) or consisted of genes organized in tandem arrays. Computational selection of segmental and tandem duplications followed stringent similarity thresholds. The *FASTA3* package, a DNA and/or protein homology search tool and based on algorithm of global alignment (Pearson et al, 1988), was applied to estimate the similarities among gene pairs (see material and methods). Tandemly duplicated gene groups were determined if all gene pairs of corresponding groups exceeded a similarity threshold and were separated by less than five intermediate unrelated genes (see material and methods), while segmentally duplicated groups consisted of gene pairs which exceeded the similarity threshold and were located in the duplicated segments determined previously by Hu et al (2011).

In total, 803 tandemly and 736 segmentally duplicated groups were selected from *A. lyrata* comprising 1619 and 1660 genes, respectively. For *A. thaliana*, 575 tandemly and 760 segmentally duplicated groups comprising 1196 and 1665 genes were obtained, respectively. The average size of tandemly and segmentally duplicated groups was 2.04 and 2.22, respectively. This indicates that most of the tandem duplicates (1029 out of 1378) consisted of two members and most of the segmental duplicates (1319 out of 1496) show 1-to-1 duplicated relation, although groups with large size were observed (sizes of tandem and segmental duplicates were up to 14 and 12, respectively). Furthermore, 169 and 147 genes have been detected as both segmentally and tandemly duplicated in *A. lyrata* and *A. thaliana*, respectively.

2.2.1.2 cis-element conservation during speciation and gene duplication

To obtain a first view about the evolution of functional non-coding sequences, conservation of cis-elements in paralogous and orthologous genes were analyzed. 271 cis-element motifs from *A. thaliana* that were detected by the FIRE based analysis (2.1.3.2) were considered as target cis-elements and mapped to the promoter regions of genes comprised in orthologous and paralogous gene groups selected previously. Promoter regions are extracted as genomic sequences from their start codon to the start of their upstream preceding genes with a maximal distance of 2kb according to the TAIR8 *Arabidopsis thaliana* annotation. A motif was considered as located in a promoter if at least one word of the respective motif was contained in the promoter region. The number of common motifs presented in each orthologous and paralogous gene pairs were calculated (figure 14). The conservation rates of cis-element in paralogous and orthologous gene pairs were compared to the background distribution of conservation as observed from 10,000 randomly selected gene pairs in *A. thaliana/A. lyrata* and between *A. thaliana* and *A. lyrata*, respectively. The results demonstrated that orthologous genes (blue bars in the top panel) shared significantly more common functional motifs than random gene pairs (black bars in the top panel)

(p-value $< 10^{-16}$ for wilcoxon rank test). This is consistent with the observations that transcriptional regulation tends to be preserved in orthologs to realize their similar transcription activities among closely related species (Blanchette et al, 2002; Blanchette and Tompa, 2002). Also, shared motifs in paralogous gene pairs show a bias towards a higher number of conservation elements compared to random gene pairs (red and green bars in the middle and bottom panels; p-value $< 10^{-16}$ for wilcoxon rank test). This indicates that control of regulation of gene transcription tends to be conserved after gene duplication to increase the robustness of genetic networks. Notably, detected motifs were more conserved in orthologous genes of *A. thaliana* and *A. lyrata* (blue bars in the top panel) than paralogous genes in both species (red and green bars in the middle and bottom panels; p-value $< 10^{-16}$ for wilcoxon rank test). This observation reflects the evolutionary time scale of *Arabidopsis* that duplication took place earlier than speciation (2.2.1) and is consistent with the expectation that higher preservation of regulatory elements exists in orthologs than in paralogs.

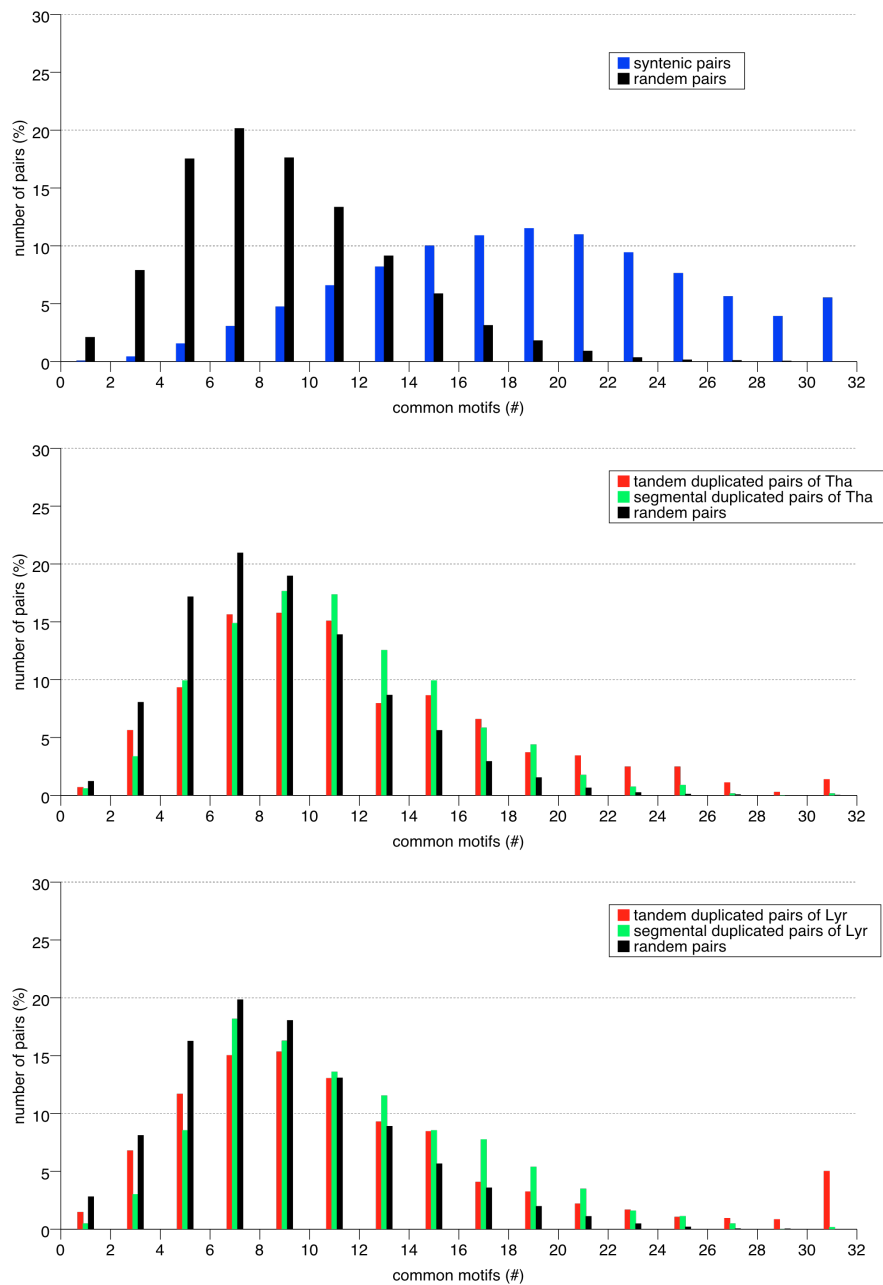


Figure 14: Distribution of common motifs shared by orthologous and paralogous gene pairs in *A. thaliana* and *A. lyrata*

X-axis displays the binned number of common motifs. The number of orthologous gene pairs (blue bars in the upper panel), tandem (red bars) and segmental (green bars) duplicated gene pairs for *A. thaliana* (middle panel) and *A. lyrata* (below panel) for each bin of shared motifs are shown on the y-axis. The background distribution of common motifs is observed from 10,000 randomly selected gene pairs (black bars) within *A. thaliana/A. lyrata* (middle and below panel, respectively) and between *A. thaliana* and *A. lyrata* (upper panel).

2.2.1.3 Evolution of cis-elements in complex paralog-ortholog gene networks

The time scale of evolutionary events in *Arabidopsis* (see 2.2.1) indicates that gene duplication in the *Arabidopsis* lineage happened much earlier than speciation of *Arabidopsis thaliana* and *Arabidopsis lyrata*. Thus, loci could be duplicated in the ancestral genome before speciation (fig 15). Along with this evolutionary process, gene groups exist for which each member has both orthologous and paralogous partner. Such gene groups that are preserved during evolution indicate essential functionality of each member. On the other hand, polymorphisms, especially changes in the promoters, of one or several members are noteworthy and of special interest. Based on the DDC model, mutable multiple regulatory regions responsible for transcriptional regulation of these genes may play an essential role in alteration of expression for one or several members of these gene families and result in neo-/subfunctionalities (Force et al, 1999). To analyse this in more detail, evolution of cis-regulatory elements in these gene groups was surveyed in this study.

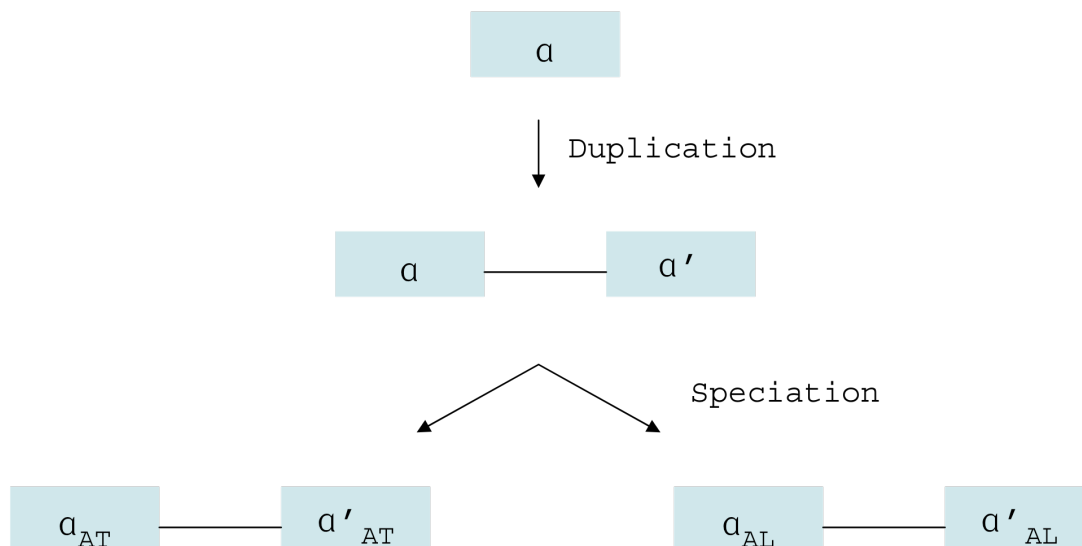


Figure 15: Evolutionary events of *Arabidopsis*

The time scale of evolutionary events for *Arabidopsis* including polyploidization and speciation were characterized. Figure simulates the process for a gene example a . a and a' are considered as duplicated genes deduced by duplication of a gene, while a_{AT} - a_{AL} and a'_{AT} - a'_{AL} represent orthologous gene pairs generated by speciation process, where AT and AL mean *A. thaliana* and *A. lyrata*, respectively.

In order to systematically investigate evolution of regulatory regions from orthologous/paralogous gene groups, orthologous gene pairs between *A. thaliana* and *A. lyrata* identified by Hu et al (2011) and tandemly/segmentally duplicated groups in *A. thaliana* and *A. lyrata* detected in 2.2.1.1 were used and their promoter regions were analyzed. Considering orthologous and paralogous relationship, a gene can contain only orthologous gene partner or only (one or more) paralogous partner(s). It can also have both orthologous and paralogous gene partners. Moreover, partners can

have further orthologous/paralogous gene partners. In order to represent such complex combination of orthologous/paralogous gene relationship, gene networks were identified and classified of which nodes are genes and edges represent paralogous or orthologous relationship. In a network, completeness of paralogy/orthology for a gene in that context to defined as a gene that has both paralogous and orthologous partner genes, while partialness is defined as a gene contain either paralogous or orthologous partner. Figure 16 depicts several networks based on their completeness and partialness of paralog/orthology relationship. The simplest case is a pair of orthologous or paralogous gene of which regulation are either preserved or changed (fig. 16 a). Networks with three and four nodes can be generated which demonstrate partial and complete paralogy/orthology relationship, respectively (fig 16 b-c). As a first focus, the networks containing four genes with complete paralogy/orthology relationship were analyzed in this study (fig 16 c).

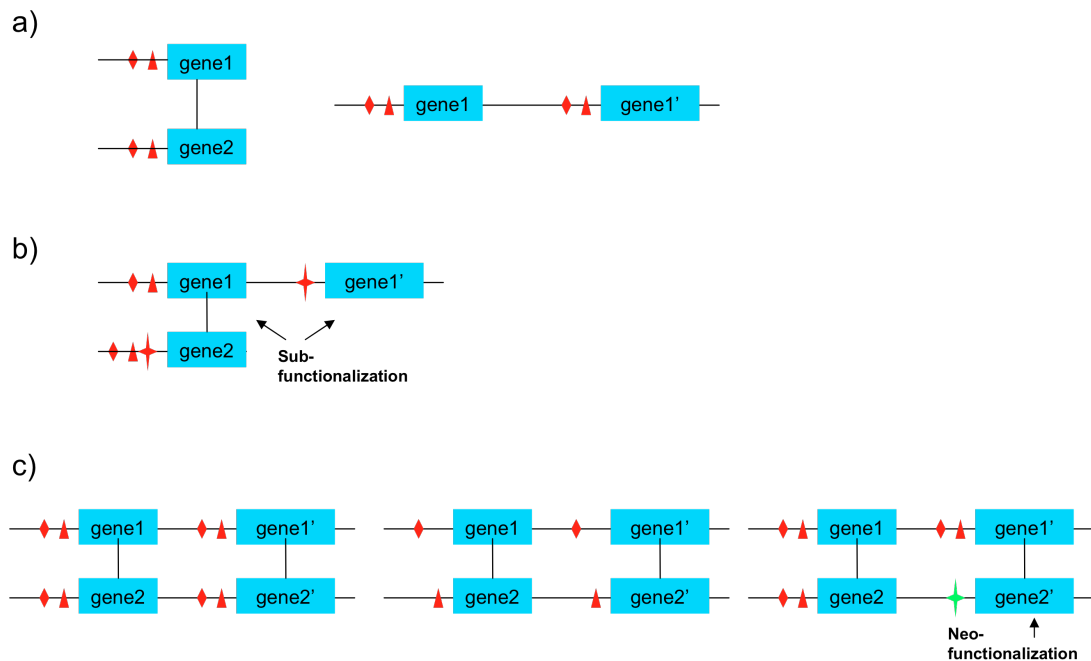


Figure 16: Paralogy-orthology gene networks

Figure shows various combinations of gene partners based on their completeness and partialness of paralog/orthology relationship. Gene1-gene2 and gene1'-gene2' represent orthologous gene pairs while gene1-gene1' and gene2-gene2' demonstrate paralogous gene pairs. Diverse red signs represent different cis-regulatory elements for respective gene. Figure a) represents networks comprising either an orthologous or a paralogous gene pairs, while networks with three (figure b) and four members (figures c) demonstrate partial and complete paralogy/orthology relationship, respectively. Various relevant situations of cis-element conservation and divergence associated with network structures are illustrated.

Overall 341 gene groups with such quartettes structure were determined (see material and methods). 271 motifs detected based on *FIRE* were used and the analysis described in 2.2.1.2 was undertaken to gain insight the conservation/divergence of

cis-elements in the promoters of genes in the respective quartettes. The number of shared motifs for the paralogous and orthologous gene pairs within quartettes was compared to the pairs not associated with quartettes and to the background pairs. Functional motifs were more conserved in orthologous gene pairs within quartettes compared to the pairs not associated with quartettes (fig. 17, upper panel; p-value < 10^{-16} for wilcox rank test). However, such conservation bias was not observed in paralogous gene pairs within quartettes compared to the pairs not associated with quartettes (fig. 17, middle panel: p-value=0.7, lower panels: p-value=0.75). This indicates, on one hand, that gene groups that are preserved during evolution, e.g. quartettes in this study, are functionally important. The highly conservation rate of the respective regulatory elements likely contributes to realize their functionality during evolution. On the other hand, gene duplication in the *Arabidopsis* lineage by polyploidization occurred much earlier than the speciation of *A. thaliana* and *A. lyrata*. Thus, differences in the sequence divergence rate can not be observed between duplicated genes that have orthologous gene partners as compared to those without orthologous gene partners.

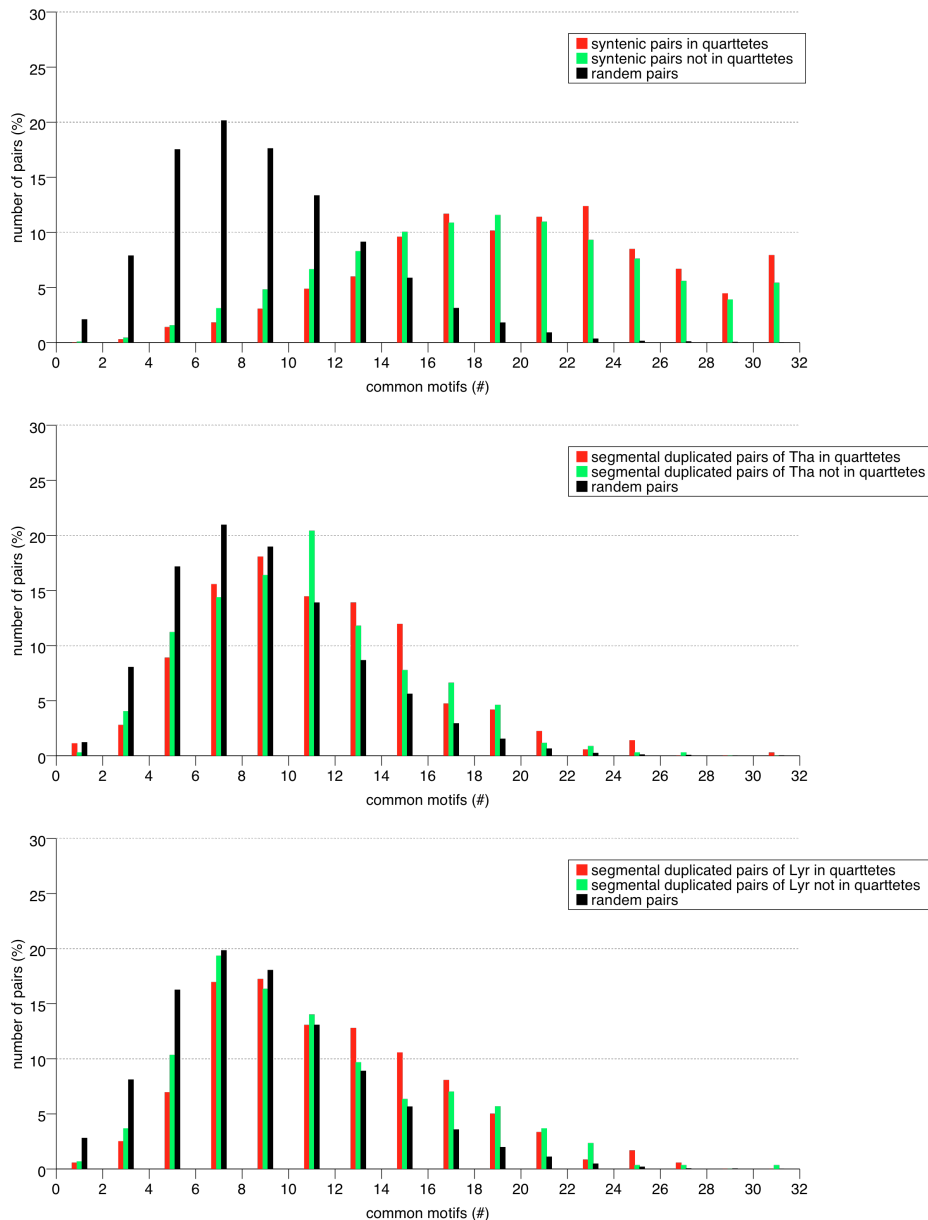


Figure 17: Distribution of common motifs shared by orthologous and paralogous gene pairs in quartettes and the pairs not associated with quartettes

X axis displays the binned number of common motifs. Y axis of the upper panel shows the number of orthologous gene pairs for each bin of shared motifs, while the middle and below panels illustrate the number of segmental duplicated gene pairs in *A. thaliana* and *A. lyrata*, respectively. Red bars and green bars in the figures represent the distribution of common motifs of gene pairs in quartettes (red bars) and not associated with quartettes (green bars). The background distribution of common motifs is observed from 10,000 randomly selected gene pairs (black bars) within *A. thaliana/A. lyrata* (middle and below panel, respectively) and between *A. thaliana* and *A. lyrata* (upper panel).

One particular situation of cis-element divergence in quartettes is that regulatory elements are conserved in three gene members while missing in the fourth one (fig. 16 c; right panel). Presence and detection of such polymorphisms that are potentially

caused by speciation and/or a duplication event is significant, since in these situations mutation of functional regulatory elements is restricted to one member of the gene set for which the protein coding regions are evolutionary preserved. Such scenarios suggest that the alteration of transcriptional regulation for the respective gene potentially results in a novel function mediated through altered regulation. Motifs mapped to known sites (see 2.1.3.3) were used to address this question, since their functions have been assigned and can be utilized to test for the generation of novel function. Their presence/absence in the promoters of each gene in quartettes was monitored. Moreover, gene ontology terms of *A. thaliana* associated with biological processes (see 2.1.3.3) were used to survey functional divergence of *A. thaliana* genes in quartettes. *A. lyrata* lacks expression data and functional annotation. Thus, quartettes have been used to analyze which polymorphism, i.e. presence versus absence, of verified regulatory motifs and functional difference exist between respective *A. thaliana* paralogous gene pairs. In total 294 out of 341 detected quartettes contain absence/presence polymorphism of individual motifs between *A. thaliana* gene pairs and for 118 of them, both motif polymorphism and functional divergence were observed. After manual inspection, 45 of these quartettes show consistence between mutated motifs and a discribed functional difference. For example, the motif ACGT:ACT:ACG:G:G:C:C:C:A:ACGT matches to the ethylene response element GGGCCC which has also been linked to stress responses. It has been preserved in the *A. thaliana* gene AT1G13230, *A. lyrata* genes Al_scaffold_0005_382 and fgenes2_kg.1__1433__AT1G13230 but missing in the *A. thaliana* gene AT3G25670, the fourth member of the quartette. Interestingly, AT1G13230 has been assigned as corresponding to biotic stimulus response while such functionality was not indicated for gene AT3G25670. AT3G25670, however, was annotated as containing with relations to signaling cascades. The indication of mutated motifs to contribute to diverse functionalities of orthologous/paralogous gene sets, of which protein-coding regions are evolutionary preserved, suggests alteration of transcriptional regulation for the respective gene and novel functionality generated during evolution.

2.2.1.4 Summary

Duplication of gene loci during are regarded as a major source for the generation of evolutionary novelties, e.g., neo-/subfunctionalization (Ohno, 1970). This can be achieved by change in protein and/or mutations of functional important non-coding regions such as regulatory elements. In this study, paralogous and orthologous gene pairs were determined in *A. thaliana* and *A. lyrata*, and conservation/divergence of previously discovered cis-regulatory motifs was surveyed in diverse types of paralog-ortholog gene networks. One particular situation of cis-element divergence within quartettes has been discussed where regulatory elements are conserved in promoter regions of three gene members while missing in the fourth one. In order to analyze this special case, mutations of experimentally verified regulatory elements within quartettes were determined and functional difference of *A. thaliana* genes

within respective quartettes were then inspected thereby utilizing the functional annotation available for *A. thaliana*. Consistence between mutated motifs and differences of the respective gene functional annotation has been observed in several dozens of quartettes which indicates the potential effect of cis-element mutations on changes of gene functionality. This case study can be considered as a first start for large-scale characterization of cis-regulatory element evolution in closely related species. Polymorphism related to diverse features of cis-regulatory elements associated with different types of complex paralogy-orthology gene networks can be further studied to in-depth investigate the evolution of cis-regulatory elements after gene duplication and/or speciation (see discussion).

2.2.2. Identification of genetic determinants for genome-plastome incompatibility of flowering plant genus *Oenothera*, subspecies *Oenothera* (= *Euoenothera*)

Detection of cis-regulatory elements is commonly based on phylogenetic footprinting which assumes that functional non-coding sequences favor to be preserved during evolution. However, numerous functional regulatory elements have been observed to be divergent even in evolutionary closely related species and been proven to play an essential role in speciation and generation of phenotypic variations (see introduction). Characterization of cis-element mutations among evolutionary closely related species has therefore become major topic in biology. However, due to limited description of phenotypic variations in close related species and lack of their genomic sequence data including orthologous non-coding sequences and a large collection of experimentally verified cis-elements, attempts to characterize the evolutionary patterns of regulatory sequences on a genome scale and their roles is still challenged and limited in their scope (Wray, 2007). In this study, a well-characterized case of phenotypic changes in higher plants, the differential genome-plastome compatibility situations of the flowering plant *Oenothera*, subspecies *Oenothera* (= *Euoenothera*) (material and methods; Table 16) was analyzed and dozens of genomic variations in coding sequences, promoters and intergenic regions among five genetically distinct plastomes of *Oenothera* (material and methods; table 16) have been detected as putative molecular determinants responsible for genome-plastome incompatibilities.

2.2.2.1. Compartmentalized genetic system and co-evolution of intracellular genetic compartments

The evolution of eukaryotic genomes originated in endosymbiotic cell conglomerates and was based on a conversion and an extension of the genetic potentials of initially free-living partner cells that coevolved into a single, integrated compartmentalized genetic system. To date, this machinery, with nucleus/cytosol and mitochondria in animals and fungi, and in addition with plastids in plants, is regulated spatiotemporally and quantitatively in its entirety (summarized in Herrmann 1997). Mitochondria and chloroplasts possess only small genomes because they have lost a large fraction of their ancestral genes many of which by transfer to the nucleus

(Martin 1998 and 2003). However, much of the nuclear coding potential, in the order of 25–30%, is required for the management of the energy-transducing organelles (Herrmann 1997). This illustrates both their tight genetic and metabolic integration and the importance of the compartmentalized genetic system in the control of the principal energy supply for the cell. Co-evolution of the intracellular genetic compartments can affect eukaryotic evolution on long and short timescales (Herrmann and Westhoff 2001; Herrmann et al. 2003). In the latter case, it becomes obvious that interspecific organelle exchanges, e.g., between plastids and nuclei, even between closely related species frequently cause serious disturbances in the development of the resulting cybrids or hybrids (Stubbe 1989; Schmitz-Linneweber et al. 2005). Compartmental co-evolution is characteristic of eukaryotic organisms and an important, often neglected, element in the speciation processes (Levin 2003).

2.2.2.2 *Oenothera* and its diverse genome-plastome compatibilities

In higher plants, one of the disturbances in the development of the resulting cybrids or hybrids caused by co-evolution of the intracellular genetic compartments is the plastome-genome incompatibility (PGI). Particular combinations of nucleus and plastids can affect the development of chloroplast and the chlorophyll synthesis (Stubbe, 1959). PGI has been observed and studied in various higher plants (Pandey et al, 1987; Przywara et al, 1989; Metzloff et al, 1982; Arisumi 1985), especially well characterized in the flowering plant *Oenothera* (Renner, 1924, 1936; Stubbe, 1955, 1959).

PGI has been extensively studied in *Oenothera* for several reasons. Firstly, *Oenothera* genetics includes a unique combination of features, such as biparental transmission of plastids (Chiu et al, 1988), a general interfertility of species, viable and fertile hybrid offspring, as well as permanent translocation of heterozygotic genomes, generally operating in combination with a system of gametophytic or sporophytic lethal factors (reviewed by Cleland, 1972). Together they allow the exchange of plastids and nuclei as well as the substitution of entire haploid chromosome sets or of individual (or more) chromosome pairs between species. Such exchanges frequently result in developmentally impaired, though fertile, plastome–genome incompatible hybrids (Cleland 1972; Stubbe and Raven 1979; Stubbe 1989; Harte 1994; Dietrich et al. 1997). Secondly, geographical distribution and genomic constitution of *Oenothera* has been characterized (Dietrich et al, 1997; Fig. 18). Subsection *Oenothera* (= *Euoenothera*) is 1 of the 5 subsections of the section *Oenothera* on which most of the experimental work in this taxon has been undertaken. The genetic study of Stubbe (1959) on this subsection categorizes plastids of this subsection into five basic, genetically distinguishable plastome types (I, II, III, IV, V; material and methods, table 16). A complete description of compatibilities of all combinations of these plastomes associated with three basic nuclear genomes (A, B and C), which occur in homozygous AA, BB, CC as well as in stable (complex) heterozygous AB, BC, AC combinations, have been performed (Stubbe, 1959; Fig. 19). The degree of

compatibility between plastomes and genome in *Oenothera* is defined by the ability of a plastid to become fully pigmented in a given nuclear background. The incompatible genome/plastomes combinations result in hybrid bleaching as well as in quantitative differences in chlorophyll content (Schoertz et al, 1964). In summary, a unique combination of genetic features and complete description of compatibilities has allowed a comprehensive study of PGI in *Oenothera*.

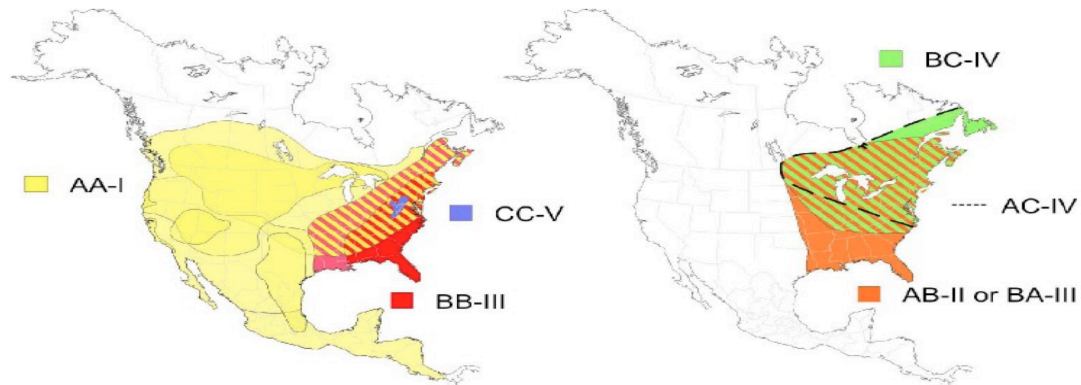


Figure 18: Distribution of the 11 North American species of subsection *Oenothera* of the genus *Oenothera*.

The map summarizes data presented in Dietrich et al. (1997) and includes information about the 6 basic nuclear genotypes containing the 3 haploid genomes, A, B and C, and their associated plastome types (I-V) of that subsection. Yellow and red gradations designated the distribution of distinct AA-I and BB-III genotypes. The left map show the areas populated by homozygous species, the right one that of their hybrids. Note that all genotypes overlap geographically.

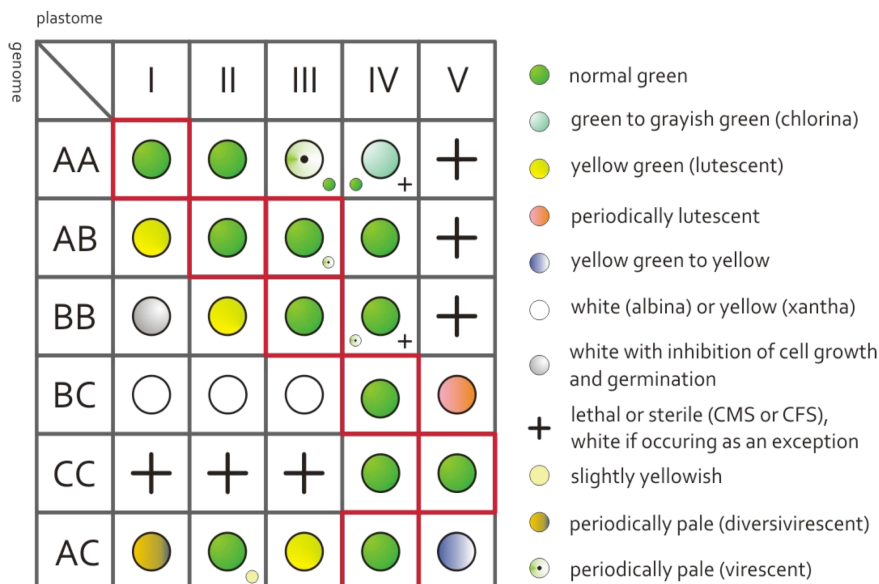


Figure 19: Plastome-genome compatibility/incompatibility in the subsection *Oenothera*.

A, B and C represent the basic haploid nuclear genomes, I-V the 5 genetically distinguishable plastomes. Genotypes boxed in bold represent naturally occurring species. Minor symbols indicate exceptions noted for some nuclear subgenotypes.

2.2.2.3 Search for putative genetic determinants responsible for PGI in *Oenothera*

Though detailed described, the mechanism and determinants for PGI in *Oenothera* are still far from clear. Many factors can be causative for PGI (see discussion). At the genomic sequence level, putative factors can be polymorphisms, including insertion, deletion and translocations, of coding regions and/or regulatory elements in both the plastomes and the genomes (Kochevenko et al, 1999). However, the search for such genomic determinants was restricted until recently due to a lack of both plastid and nucleus genomes. In 2008, Greiner and co-workers represent the complete nucleotide sequences of representatives of the 5 basic *Oenothera* plastomes, their comparison, evolutionary relationships, and temporal relation (Greiner et al, 2008). All 5 genomes are perfectly collinear. Close evolutionary relationship among the five plastomes (between 0.8 to 1 Mya) was identified and high sequence similarity between 96.3% to 98.6% was found (Greiner et al, 2008). Deletions, rearrangements as well as duplications of entire genes were not detected. Such sequences from closely related, but morphologically distinct and still interbreeding species for which organelle and nuclear genetics including cybrid technology is available enabled large-scale search of plastid-specific genetic determinants that might be causal to compartmental co-evolution, e.g. PGI in this study. As mentioned previously, putative factors could be polymorphisms of coding regions and/or regulatory elements in plastid genome. However, numbers of single mutations like small insertions and deletions or replacements of nucleotides are in the range of several thousands if each pairwise plastome comparison is considered. To delineate candidate PGI determinants, single- or small-scale molecular differences between orthologous genetic elements, notably coding sequences and predicted functional elements in intergenic regions, were analyzed in this study. Two categories of polymorphisms were considered including predicted or known polypeptide variance, and mutations of promoters and polymerase binding sites. Moreover, the evolutionary sequence filtering approach was based on a comparison of sequence differences with the incompatibility chart of subsection *Oenothera* listing all possible genome–plastome combinations, which are either compatible or incompatible (fig. 19). For instance, genetic polymorphisms could be excluded to be causative for incompatibility when any other plastome containing polymorphic sequences identical to the incompatible plastome was compatible in the respective nuclear background. On the other hand, polymorphisms in genes, which are experimentally proven not to explain an incompatible phenotype (an example discussed in 2.2.2.4), were excluded from the list of potential candidates for PGI.

In protein-coding genes, two approaches were undertaken to survey the impact of plastome-specific differences. First, a putative functional impact of non-synonymous sites in polypeptides was estimated from three aspects: the degree of evolutionary conservation among five *Oenothera* plastomes and plastomes from reference species, the putative changes of biochemical properties, and location on the protein in relation to known functional domains indicating potential effects on function. Second,

protein-coding genes with length polymorphisms which may generate variant polypeptides are of intrinsic interest for PGI. Genes with length variations among the evening primrose plastomes include reading frame shifts and alternative stop codons caused by single-base pair insertion/deletion were regarded as potential plastome-specific candidates responsible for PGI in *Oenothera*. Leveraging on these selection criteria, 33 significant mutations in coding regions of 19 loci for single amino acid exchanges and 11 insertions/deletions in 7 genes causing variant polypeptides have been identified for the distinct plastome-genome combinations in the compatibility scheme (fig. 19) and listed by Greiner et al (2008).

In a complementary approach, mutations in promoters and polymerase binding sites have been analyzed. These may affect transcriptional activities of genes nearby and can be considered as candidates causative for PGI. Polymerase binding sites in promoter sequences were firstly predicted and polymorphism in detected binding sites among five plastome genomes were then compared with the generically determined compatibility relationships in interspecific hybrids (fig. 19) in order to classify them and to pinpoint potential determinants contributing to PGI.

Transcription of plastid genes is performed by at least two RNA polymerases in higher plants: (a) a single-subunit type enzyme related to those of T-odd phages and mitochondria (Lerbs-Mache, 1993; Hedtke et al, 1997) and (b) a multi-subunit form which resemble those of bacteria and eukaryotic nuclei (Cahoon et al, 2001; Hajdukiewicz et al, 1997; Sugita et al, 1996; Maliga, 1998). The former one is produced by nuclear gene(s) (nuclear-encoded phage-type plastid RNA polymerase; NEP), while the latter one, which is characterized as plastid-encoded plastid RNA polymerase (PEP) because it is encoded by chloroplast genome, contains a catalytic core consisting of bacterial RNA polymerase-encoded α , β , and β' homologous subunits. The PEP enzyme also contains a nuclear-encoded subunit, the σ factor, responsible for promoter specificity and initiation of transcription (Ishihama, 1988; Lonetto et al, 1992; Homann et al, 2003; Kanamaru et al, 2004). The PEP is the primary enzyme for photosynthesis-related gene transcription in plastids. The polymerases recognize distinct sites in plastid promoters. In order to identify their binding sites, a pattern search for reported consensus sequences was employed to predict functional RNA polymerase binding sites of PEP and NEP. The consensus sequence TYRMNN(N)₁₆₋₂₀WANNWT which covers most experimentally reported PEP sigma factors (-35 and -10 boxes) (Homann et al, 2003; Kanamaru et al, 2004) and ATA₀₋₁N₀₋₁GAA(N)₁₅₋₂₃YRT which represents NEP type Ib promoters (Silhavy et al, 1998; Kapoor et al, 1999) were selected as target patterns and have been searched in the intergenic regions which are delimited either by the 5' neighboring gene or a maximum size of 600 base pairs. The other two NEP promoter types have been excluded from the analysis as the consensus of NEP type Ia promoters – YRTa – is too short and generalized for computational prediction and NEP type II promoters are characterized from just a single case (Liere et al, 2006). As a result, among 113 genes reported by Greiner (Greiner et al, 2008), 75 genes contain at least one putative sigma

factor binding site in their promoter regions, whereas 69 genes contain potential NEP type Ib promoters. In total, at least one polymerase binding site, either NEP or PEP, was detected in 88 genes, and both types of binding sites can be detected in 56 genes.

Only predicted binding sites that represent variations among all five *Oenothera* plastomes can potentially contribute to PGI. Thus, 25 NEP- and 38 PEP-binding sites altered in at least one of the five *Oenothera* plastomes remain for further analysis. Their polymorphism among five plastome genomes were compared with the distinct plastome-genome combinations in the compatibility scheme (fig.19) in order to classify them and to pinpoint potential determinants contributing to particular PGI situations. For instance, a large deletion in the plastome I in the intergenic region between gene *clpP* and *psbB* compared to the other four plastomes was discovered. In that case certain RNA polymerase binding sites are missing. Such a deletion in the plastome I may be the candidate for the AB-I incompatibility compared to completely compatible between genome AB and plastomes II, III, IV (For detail analysis, see 2.2.2.4). Candidates of binding site polymorphism that are potentially causative for particular compatibilities relationship are summarized in the table 15.

Binding of PEP and NEP polymerases has restriction with respect to their position and the numbers of binding sites in the upstream sequence of respective gene. Therefore, predicted binding sites that are potentially causative for particular compatibilities were further evaluated according to three criteria: position to a translational start site, number and similarity to the experimentally reported consensus sequences. The effect of the predicted binding sites was inspected and estimated how likely the corresponding polymorphism may result in particular PGIs (Table 15). Overall 9 putative PEP promoters (Table 15), notably those of *clpP*, *psbB*, *rpl16*, *rpl33*, *rps12*, *rps15*, *trn_{GCC}*, *trn_{LCAA}*, *trn_{UGA}*, and 7 predicted NEP promoters (Table 15), namely those of *atpH*, *clpP*, *ndhG*, *psbB*, *psbK*, *rps4*, and *trn_{GCC}*, were detected as candidates causing PGI. Moreover, the promoters of three genes, *clpP*, *psbB*, and *trn_{GCC}*, contain changes for both polymerase types.

Putative PEP promoter		
Gene	Estimated effect	PGI considered
accD	N/A	N/A
atpB	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
atpF	unlikely	AA-V;AB-V;BB-V
clpP	likely	AB-I;AC-I
ndhD	unlikely	BB-I;BB-II
ndhF	N/A	N/A
ndhG	possible	AB-I;AC-I
petL	N/A	N/A
psaJ	possible	N/A
psbB	likely	AB-I;AC-I
psbD	unlikely	BB-I;BB-II
psbE	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
psbI	possible	AB-I;AC-I
rbcL	possible	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
rpl16	likely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
rpl20	possible	AA-IV
rpl22	possible	AA-IV
rpl32	possible	N/A
rpl33	likely	BB-I;BB-II
rps12	likely	AA-V;AB-V;BB-V
rps15	likely	AA-V;AB-V;BB-V;BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
rps16	unlikely	BB-I;BB-II
rps18	possible	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
rps4	unlikely	AA-IV
trnFGAA	unlikely	AB-I;AC-I
trnGGCC	likely	AA-V;AB-V;BB-V
trnGUCC	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
trnHGUG	unlikely	AA-V;AB-V;BB-V
trnLCAA	likely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
trnLUAG	unlikely	AA-V;AB-V;BB-V;BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
trnPUGG	N/A	N/A
trnQUUG	possible	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
trnSGCU	N/A	N/A
trnSUGA	likely	BB-II
trnTUGU	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
ycf2	possible	N/A
petN	unlikely	AA-IV

Putative NEP Promoters		
Gene	Estimated effect	PGI considered
ycf2	unlikely	AA-III;AC-III
atpB	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
atpF	unlikely	AA-III;AC-III
atpH	likely	AB-I;AC-I
atpI	possible	AA-V;AB-V;BB-V
clpP	likely	AA-III;AC-III;AA-V;AB-V;BB-V
ndhF	possible	AA-V;AB-V;BB-V;BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
ndhG	likely	AB-I;AC-I
psaI	unlikely	BB-II
psaJ	unlikely	AA-IV
psbB	likely	AA-III;AC-III
psbK	likely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
rpl20	unlikely	AA-IV
rpl33	possible	AA-V;AB-V;BB-V
rpoB	unlikely	AA-V;AB-V;BB-V
rps12	possible	AA-V;AB-V;BB-V
rps15	unlikely	AA-IV
rps16	unlikely	AA-III;AC-III
rps18	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
rps4	likely	AA-IV
trnGGCC	likely	AA-V;AB-V;BB-V
trnCAU	possible	AA-III;AC-III
trnPUGG	unlikely	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
trnQUUG	possible	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III
trnTUGU	possible	BC-I;BC-II;BC-III;CC-I;CC-II;CC-III

Table 15: Assessment of putative PEP promoters as candidate loci for plastome-genome incompatibility in *Oenothera*

Table summarizes loci which contain polymorphism in putative PEP and NEP promoters among five *Oenothera* plastomes. Mutation in predicted binding sites causative for particular compatibilities were further evaluated according to different criteria (see text) and effect of the predicted binding sites was manually estimated suggesting how likely (second column) the corresponding polymorphism may result in particular PGIs (third column).

2.2.2.4 The *clpP-psbB* intergenic region is a potential AB-I incompatibility determinant

In order to investigate how divergent intergenic regions may result in PGI, in this study AB-I incompatibility was chosen as case and the candidates of intergenic regions and their promoter architecture that are potentially causative of this incompatibility were discussed. AB-I displays a yellow-green phenotype, whereas the plastome/genome combination between AB and II, III, and IV are green (Fig 19). Changes in plastome I shared with, or similar to at least one in plastomes II-IV can therefore be disregarded. Plastome V was excluded from the analysis because the combinations AB-V, AA-V and BB-V are extremely disharmonic and differ substantially from the AB-I phenotype. They are fully bleached, largely pollen sterile and display severe inhibition of cell division (Stubbe 1963). The principle genetic determinants responsible for their phenotypes are presumably complex and different from those causing bleaching of AB-I individuals (Greiner et al, 2008).

According to comparison between diversities of promoter sequence among different *Oenothera* plastomes and different genome/plastome compatibilities, variant intergenic regions of four genes may result in the AB-I incompatibility, namely 5' upstream of *atpH*, *ndhG*, *psbI* and the spacer between *clpP* and *psbB* (Table 15). All regions involving the NADPH complex were disregarded since knockouts of individual NDH subunits in tobacco lack a noticeable phenotype change (Burrows et al, 1998; Kofer et al, 1998). Hence, the 5' upstream of *ndhG* can be excluded. The variant 5' upstream of *atpH* was also disregarded since neither ATP synthase nor the cytochrome complex is affected in AB-I combination as revealed by immunological analysis (Greiner et al, 2008). Furthermore, the 5' upstream of *psbI* is unlikely involved in AB-I phenotype because 1) mRNA levels of *psbI* are not changed between five plastomes, and 2) a knockout of *psbI* showed/s no apparent phenotype (Schwenkert et al, 2006). As a result, the only remaining intergenic region that may be causative for AB-I incompatibility is the intergenic region between *clpP* and *psbB*.

To address the question how plastome I specifically affects the compatibility with genome AB, a phylogenetic footprinting analysis for *clpP-psbB* intergenic regions of *Oenothera* plastomes I-IV was undertaken. *Arabidopsis*, tobacco, spinach, *Atropa*, *Eucalyptus* and *Gossypium* were employed as reference species, as these species are evolutionary closely related to *Oenothera* and their plastid genomes have been sequenced and annotated. A striking difference between plastome I and plastomes II, III, IV is a 148-bp deletion in the plastome I intergenic region (Fig. 20). Among the previously predicted putative binding sites, one putative *clpP* promoter and two potential *psbB* promoters are located in this region and are conserved in the other three *Oenothera* plastomes and reference plastomes (Fig. 21). Another experimentally confirmed NEP promoter of *clpP* is also missing in the plastome I (Fig. 21). Conservation of known and predicted sites in other species but absence in the *Oenothera* plastome I indicates that the missing elements in this region of plastome I

may result in the effect of influencing the expression of *psbB* and *clpP* in the AB-I combination.

If altered expression of *psbB* and/or *clpP* contributes to the bleaching of the AB-I phenotype, decreased PSII activity/levels and eventually pleiotropic effects due to *clpP* overexpression would be expected. Thus, change of PSII activity can be surveyed in order to provide experimentally evidence for the effect of the 148-bp deletion of plastome I in intergenic region between *clpP* and *psbB* discovered bioinformatically. The activity of PSII relative to that of PSI was monitored by spectroscopic analysis (Greiner et al, 2008). Chlorophyll fluorescence analysis is fully consistent with a decreased PSII activity relative to PSI and suggests that the lesions in the incompatible AB-I hybrid primarily affect PSII function (Greiner et al, 2008).

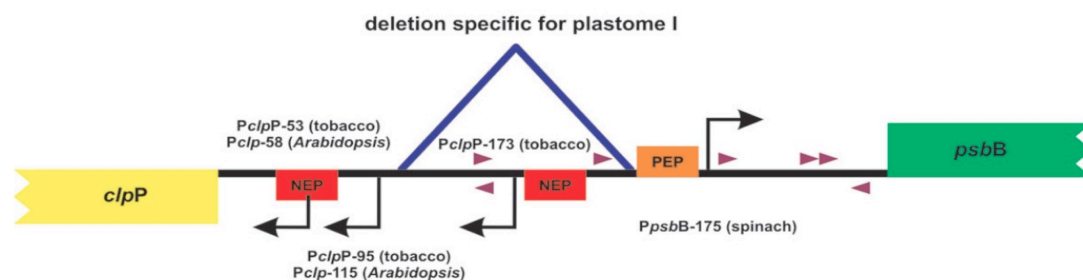


Figure 20: Schematic overview of the *clpP/psbB* spacer region in *Oenothera*, spinach, tobacco, *Atropa*, *Eucalyptus*, *Gossypium*, and *Arabidopsis*.

Positions of the indicated transcription start sites (black arrows) of NEP and PEP promoters (*PclpP* and *PpsbB*) relative to the start codons were determined experimentally in *Arabidopsis*, tobacco, and spinach (Westhoff 1985; Hajdukiewicz et al. 1997; Sriraman et al. 1998; Swiatecka-Hagenbruch et al. 2007). Putative, not experimental verified promoters in *Oenothera* are marked with filled triangles. The experimentally verified *PclpP-173* and *PpsbB-175* are highly conserved and confirmed bioinformatically in *Oenothera* and all references species. The deletion (open triangle) is not present in *Oenothera* plastomes II–V or plastomes of other species sequenced so far and, therefore, specific for plastome I in *Oenothera*.

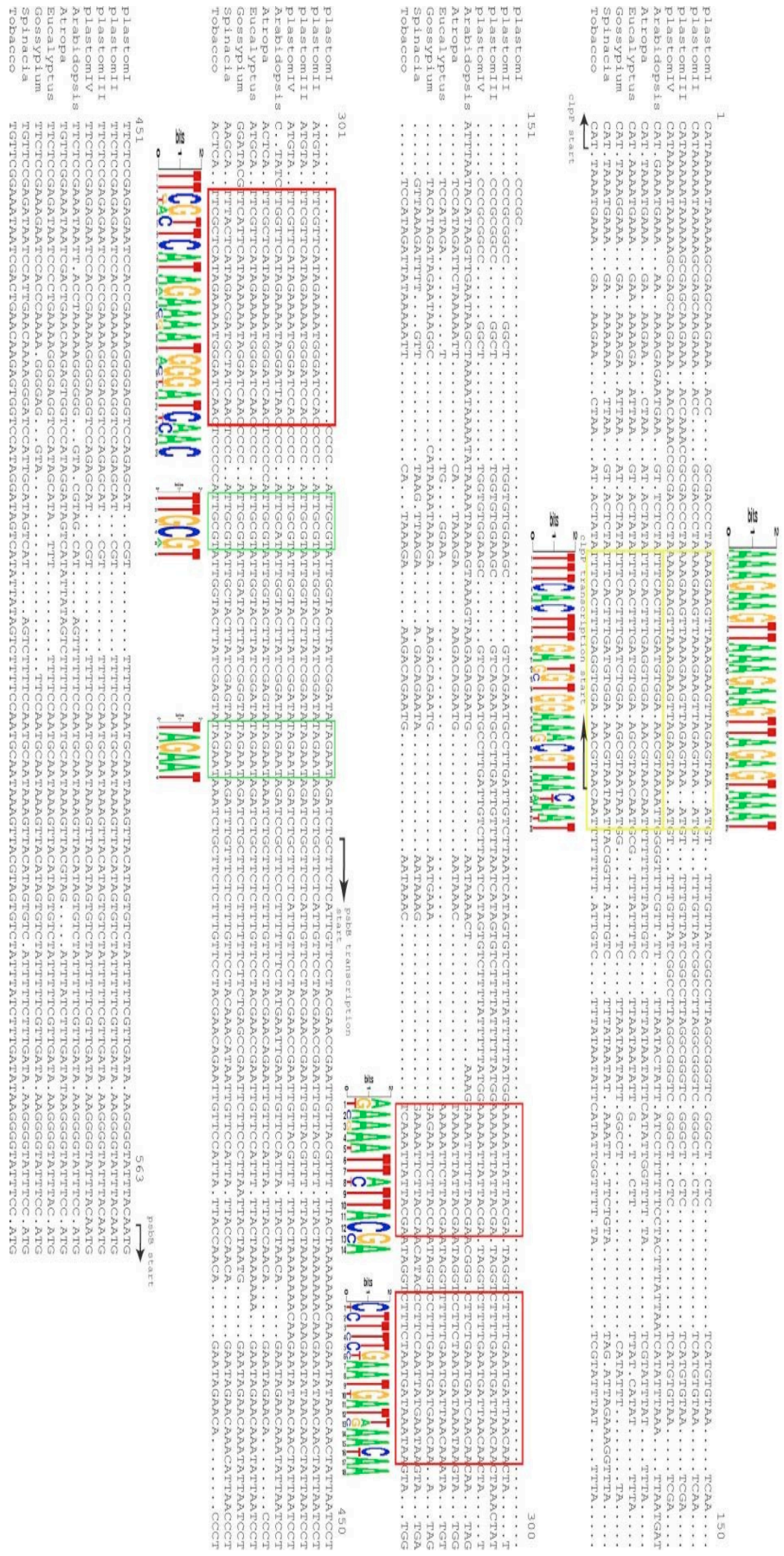


Figure 21: Alignment of intergenic region between clpP and psbB

Figure displays the alignment of spacer region between clpP and psbB from Oenothera plastomes I-IV, Arabidopsis, Atropa, Sycalyptus, Gossypium, spinach and Tobacco. The arrows at both ends of the alignment indicate the start codons of clpP and psbB genes, while the arrow in the middle represents transcription start site of psbB. The Boxed regions demonstrate reported RNA binding sites (details see figure 20) and their respective sequence logos were illustrated. Yellow boxed region represents the element highly conserved among Oenothera plastomes while divergent compared to other reference plastomes. Green boxed regions depict potential binding sites which are completely conserved among all plastomes. Notably, a unique large deletion in plastome I can be observed in the alignment which contains three putative promoter elements (red boxed).

2.2.2.5 Summary

The completion of five genetically distinct *Oenothera* plastomes by Greiner in 2008 for the first time offered the possibility to large-scale search for plastid-specific genetic determinants causal to compartmental co-evolution. In this study, variant sequence regions in plastomes I-V were systematically compared with compatibility/incompatibility patterns to filter for loci in plastomes relevant for PGI. Dozens of candidates in protein-coding and intergenic regions have been identified which are candidates to contribute to particular PGIs. As a case study, the survey finally correlated the AB-I phenotype with a single major locus, the intergenic region *psbB-clpP*, and a 148-bp deletion specific on plastome I has been computationally discovered where several putative and known polymerase binding sites are missing. The plastome I specific deletion has been experimentally proved to result in decreasing PSII activity relative to PSI and eventually the lesions in the incompatible AB-I hybrid (Greiner et al, 2008). The study suggests that evolution of cis-element and promoter contribute directly to speciation and trait difference among evolutionary close related species.

3. Discussion

3.1 Discovery of cis-regulatory elements in higher plants

Based on high quality sequence information and gene annotation deduced by rice genome projects, reliable and robust promoter sequences associated with the respective genes can be extracted. However, only few promoter sequences from other grass genomes have become available until recently. Comparative or computational biology approaches were therefore restricted to studies of individual pairs of interest and limited by the availability of only a few hundreds of grass promoter sequences (Wray, 2007). Thus, our knowledge of *cis*-regulatory elements in monocotyledonous plants is limited to those that have been reported and deposited in plant motif databases. The few dozens of known motifs are in sharp contrast to findings that higher plant genomes typically encode on average more than 1,500 transcription factors (Riechmann, 2000; Xiong 2005).

With the completion of the sorghum genome, a genome-wide assessment of regulatory sites in rice and sorghum upstream sequences has now become feasible. In this survey, approaches based on two different tools, PhyloCon and FASTCOMPARE, were employed. Both tools and approaches have been successfully applied to motif discovery in many non-plant organisms including yeast and mammals (Wang et al, 2003; Elemento et al, 2005). In addition, PhyloCon has previously been applied successfully for *cis*-element analysis in genome survey sequences of *Brassica oleraceae* vs. *Arabidopsis thaliana* (Elemento 2005; Haberer 2006). FASTCOMPARE is based on the 'network-level conservation' principle. This presupposes that regulatory circuitries will be largely conserved between two evolutionary related species and functional network motifs can be detected by their higher global or genome-wide conservation rate compared to non-functional sequences. Evolutionary conservation of functional elements is also assumed for phylogenetic footprinting that discovers motifs from a group of orthologous gene pairs. For the analysis based on PhyloCon, the orthologous groups that are compared and combined result from a prior selection of orthologous mate-pairs by co-expression analysis.

Similar to grass genomes, until recently comparative and computational study of *cis*-elements in dicotyledonous plants was also restricted due to lack of complete genomes with reasonable evolutionary relationship and comprehensive expression datasets. The previous work of large-scale *cis*-element detection in *A. thaliana* and *Brassica oleraceae* (Elemento 2005; Haberer 2006), for example, uses the genomic survey sequences of *Brassica* for the comparative genomics analysis, while the survey of Elemento (2007) used the *FIRE* package for a genome-wide discovery of *cis*-elements in *A. thaliana*, yet restricted by extremely limited expression datasets. In

this study, more than 2000 microarrays covering 157 different experiments and providing detailed expression information for *A. thaliana* were analyzed by *FIRE* for large-scale discovery of cis-regulatory elements in this model of dicotyledonous plants. The *FIRE* package was applied in a genome scale survey to discover functional elements in promoter regions of *A. thaliana* that utilizes “mutual” information deduced by estimating distribution of motif presence/absence among diverse co-expressed gene groups and investigates whether it is over-/underrepresented in one or several groups. This survey was considered as extension and improvement of cis-element discovery compared to the previous work. Moreover, with the completion of the *A. lyrata* genome which demonstrates overall high sequence similarity and highly conserved collinearity to *A. thaliana*, FASTCOMPARE was employed to detect phylogenetic footprints between *A. thaliana* and *A. lyrata* as their conserved cis-elements.

Compared to cis-element detection by PhyloCon that is restricted in well-defined gene groups, motif discovery based on „network-level conservation” and “mutual information” is an effective global, alignment-free approach. Hence, application of these approaches for cis-element detection is considered as complementation rather than redundancy.

Ab initio analysis of cis-regulatory elements is notoriously error-prone due to small motif sizes and motif degeneracy. To increase the specificity of cis-element prediction, the information of evolutionary conservation and co-expression has been taken into account as two axes of independent but complementary resources. Cis-element detection by PhyloCon in rice and sorghum combined both resources. The discovery based on the FASTCOMPARE and the FIRE approaches utilized one information resource but were evaluated by the complementary data axis. In particular, the study was designed to select functional candidate sites and motifs that are associated with transcriptional activity. For the motifs detection in rice, co-expression was derived from correlations exceeding the top 1% of background similarities and the employed clique approach required all group members to have a significant expression correlation with all other group members. For the analysis of *A. thaliana*, co-expressed gene groups were firstly determined and mutual over-/underrepresentation of particular motifs was estimated to uncover their functional specificity and indicate their roles in regulation of transcriptional activity.

However, determination of search space for discovery of regulatory sites is problematic. Search in short upstream sequence results in missing functional elements. Search for distal promoters and enhancers in genomic sequences of large size results in large time-cost and may still fail, as iterative and stochastic searches can easily be trapped into local maxima when the search space is large. In this survey, we restricted motif searches to 5'-upstream sequences of size 2 kb (for k-mer searches including FASTACOMPARE and FIRE) or 3 kb (for PhyloCon) (i) to take current knowledge of plant promoter sizes into account and, (ii) to focus on plant core promoters that

presumably contain most functional elements. Though functional enhancers and *cis*-elements in e.g. mammalian promoters have been reported up to several tens of thousands bases distant to transcription initiation sites (TIS), plant promoters seem to be more compact (Lockton 2005). In addition, chance co-occurrences will strongly increase, in particular, for smaller k-mers and degenerated motifs. Upstream sequences of larger size would thus have adverse effects by accumulating false positives or losing statistical power.

3.1.1 Detected *cis*-regulatory motifs and sites.

Results reported in this study can be divided into two categories: conserved sites and motifs. PhyloCon position specific scoring matrices (PSSMs) are supported by their conservation between orthologous promoters and their simultaneous co-occurrence in genes with expression similarities. The mean size of PhyloCon motifs detected in this study was considerably longer (20-23 bp; table 1). However, sizes of *cis*-elements in plants typically range between 6 and 12 base pairs (Bryne 2007; Matys 2003). Hence, PhyloCon Motifs likely represent concrete conserved sites rather than generalized statistical models for transcription factors. Large sizes for phylogenetic footprints in grasses are consistent with a previous study of maize, rice and sorghum comparisons, in which a minimum motif size ≥ 20 bp was found to be significant (Guo 2003). Such long sites for PhyloCon PSSMs can be composed of two or more motifs and close proximity of these sites is required for functionality in the respective co-expressed group. Alternatively, some of the detected sites could represent signals associated with transcriptional gene activity such as mRNA stability signals or miRNA target sites, for which longer sizes have been reported (Vazquez 2006). Complementary to these long conserved regions, many of the detected *FIRE* and network-level conserved motifs represent candidates for transcription factor binding sites. The size ranges of detected motifs are 6-9 bp and 7-10 bp for FASTCOMPARE and *FIRE* analysis, respectively. This is consistent with typical size range of plant *cis*-elements (Bryne 2007; Matys 2003). Both approaches integrate k-mer degeneration process and detect motifs with a certain of variability. Thus, both methods reported models of functional elements rather than concrete sites.

Analysis of both FASTCOMPARE and *FIRE* approaches are based on the principle to search for k-mer's and detect motifs with comparable size as models of functional *cis*-elements. However, difference in number and degree of variability was found between motifs detected by the two methods. After subjecting the individual detected sites to clustering, in total 4,426 non-redundant motifs (including 441 dyad motifs) were found based on the "network-level conservation" approach. However, the rice genome contains more than 1,600 genes encoding transcription factors and a similar number of distinct *cis*-regulatory motifs could be expected (Xiong 2005). Besides a certain false-positive rate generally existing by each de novo approach for *cis*-element detection, several reasons are considered to explain the fact that more than two-fold the number of elements has been predicted compared to the expected ones. Some of

the detected motifs may still be too specific and one transcriptional factor may bind to several related motifs. Observation of sequence variability for few motifs (see results) indicates that the chosen scoring function, i.e., ratio between observed and expected co-occurrence, favors specific words or overrepresented motifs with an overall low occurrence rate in a genome. Consequently, alternative scoring functions to measure the conservation rate for motifs, e.g. cumulative hypergeometric test score or contingency coefficient of cross tabulations, can generate higher degree of motif degeneracy. Yet, heuristic implementation instead of deterministic algorithm may be another reason that imposes a somewhat artificial limit on the degree of motif degeneration and motif detection power. However, implementation of deterministic algorithms is computational infeasible. Also, many motifs were obtained from dyadic motif searches that converged to motifs with highly specified spacer sequences. For these long motifs, similar considerations may apply as for *PhyloCon* sites discussed above. Taking this into account, the number of motifs reported by *FASTCOMPARE* is close to the number of transcription factors present in rice. *FIRE* analysis, on the other hand, deduced 271 non-redundant motifs in *A. thaliana* that is far less than the number of cis-elements expected in higher plants and the number of sites detected by *PhyloCon* and *FASTCOMPARE*. In contrast, *FIRE* motifs demonstrate higher degree of degeneration compared to those detected through *PhyloCon* and *FASTCOMPARE* analysis. This indicates that *FIRE* analysis favors specificity over sensitivity and leads to high-confident motif detection. In fact, due to possibly missing prediction of transcription factor binding sites by *FASTCOMPARE* that tolerate high degeneracy, *FIRE* can be considered as a complementary method to *FASTCOMPARE* analysis and complete the candidate list of k-mer based cis-element motifs.

3.1.2 Evaluation of detected cis-regulatory elements

Biological functionality of detected motifs needs to be verified in order to provide confidence of collected cis-elements and reliability of applied computational approaches. In this survey, functionality of predicted cis-regulatory motifs has been confirmed by a variety of approaches. Firstly, candidate cis-elements have been mapped to experimentally verified sites available from public databases and literature reports. For motifs from *A. thaliana*, numerous reported cis-acting sites available from the *PLACE* and *AGRIS* databases have been successfully found in the list of candidate motifs. Compared to *Arabidopsis*, only a limited number of previously detected and reported sites is available for grasses. 74 and 55 sequences of experimentally verified rice transcriptional factor binding sites have been extracted from the *PLACE* and *TRANSFAC* databases, respectively (Higo et al, 1999; Matys et al, 2003). Overall 96 known sites were obtained after filtering for redundancy. However, many sites are still redundant due to representation of binding site variations or pronounced sequence overlaps between databases and even within one database. Hence, the exact number of different known motifs is difficult to assess. Despite the limited availability of experimentally verified cis-regulatory elements in grasses, numerous matches to known grass motifs or sites in public databases and

literature reports have been found. This includes many variants of the ACGT motif, like the G-box or the ABA response element as well as ethylene response elements among others. Interestingly, some top-scoring motifs do not match previously published elements and indicate novel cis-regulatory motifs.

Many surveys have reported an association of particular motifs with particular biological processes (Quackenbush, 2003). For large-scale analysis, gene ontology classification or metabolic pathways can be correlated with particular motifs (Maleck et al, 2000). Such association can be applied to analyze for the functional association of candidate elements. 24 out of 271 discovered motifs in *A. thaliana* demonstrate significant association with particular biological processes. Notably, more than half of these motifs have successfully matched to the verified sites and several of them indicate exactly the functionalities these verified sites report. For cis-regulatory motifs in rice, however, only a few such associations can be identified, and all enrichments were in very broad biological categories, e.g. 'transcription' (see Methods, results not shown). The missing associations likely result from limitations of the current rice GO annotation. In our search, we found for only 755 RAP2 rice genes (2.7%) at least one GO term belonging to the category 'biological process'. Similarly, only 1,376 rice genes (4.9 %) could be mapped on KEGG pathways. In total, a functional annotation has been found for less than 5% of all rice genes. The sparse data basis and low resolution of the current rice GO annotation for grass species that mostly assigns top level terms are the most probable causes for the limited success in detecting significant enrichments and application of functional association to confirm discovered cis-elements in rice.

Individual approaches have been employed to confirm candidate motifs depending on the information resource on which cis-elements detection based. For example, similarity of transcription level for rice genes sharing motifs which were detected based on phylogenetic footprinting of rice and sorghum, i.e. “network-level conservation” in this study, has been estimated using rice expression data. The analysis showed that the number of motifs two rice genes have in common positively correlates with their expression similarity. This is consistent with the combinatorial nature of transcription regulation (Davidson 2001; Levine 2003) and strongly indicates that a large fraction of detected motifs are associated with control of transcription. For the candidate motifs in *A. thaliana* discovered based on “mutual” information among diverse co-expressed gene groups, their evolutionary conservation rates between *A. thaliana* and close evolutionary related species *A. lyrata* were calculated. The analysis has indicated that candidate motifs detected by FIRE indeed favor to be preserved during evolution compared to background non-coding sequences and therefore strongly suggested their functionalities. Moreover, FIRE estimated degree of overrepresentation for each motif in each co-expressed gene group. Also, conservation rate of a motif for a particular co-expressed gene group has been calculated as the fraction of orthologous gene pairs between *A. thaliana* and *A. lyrata* in the corresponding co-expressed gene group sharing the respective motif. A

significant positive correlation was found between the degree of overrepresentation and such evolutionary conservation rate. This indicated that functional candidate motifs are not only under constraints of selection, but also favor to be preserved during evolution in the gene groups where they occur frequently and may play an important role for transcription regulation of respective genes.

3.1.3 Summary and future work

Large-scale detection of cis-regulatory elements based on comparative genomics has been addressed in grass species for the first time. The analysis was made possible by the completion of the sorghum genome and the already available rice genome. Based on the transcriptional activity of rice genes, thousands of significant motifs have been detected in this study. This will provide experimental researchers with a prioritized list of candidates for the gene of interest and can guide experimental designs for numerous sorghum and rice genes. Moreover, a comprehensive collection of expression datasets monitoring various experimental conditions made an extension and improvement of genome-wide cis-element characterization in the dicotyledonous plant *Arabidopsis thaliana* compared to previous surveys. Together, the applied approaches for detection of cis-regulatory elements in this study has been evaluated and can serve as paradigm for analysis in further grasses and in higher plants in general (Wang et al, 2009). Additional grass genome projects, for instance *Brachypodium distachyon* (International Brachypodium Initiative, 2010), a wheat relative, and maize (Schnable et al, 2009) have recently been completed and can be expected to deliver important and information-rich comparative genome templates in future studies (Pennisi 2007). This will enable and stimulate whole-genome comparative studies between three and more grass genome sequences. In particular, comparisons between two closely related grasses, maize and sorghum, will allow (i) branch-specific motifs to be accessed and, at the same time, (ii) the identification of motifs common to the monocot clade. In addition, with increasing knowledge of individual cis-elements in higher plants and development of systems biology, in-depth characterization of cis-regulatory modules and their related dynamic transcriptional networks can be addressed in the near future.

3.2 Evolutionary analysis of cis-regulatory elements in higher plants

For decades, numerous studies of trait difference have demonstrated evolutionary changes at loci for which functional coding changes have been excluded and functional cis-regulatory mutations have been implicated or directly demonstrated at the molecular level, revealing evolution of cis-regulatory elements is essential for phenotypic divergence. Though important, attempts to large-scale characterize the evolutionary patterns of regulatory sequences and their roles is still challenged and limited in their scope (Wray, 2007). This was due to limited description of phenotypic variations in close related species, lack of their genomic sequence data including orthologous/paralogous non-coding sequences and comprehensive collection of

experimentally verified cis-elements. In this study, computational analysis has been addressed to genome-wide survey cis-element mutations and their potential functional consequences for trait changes in two dicotyledonous plant models: *Arabidopsis* (*Arabidopsis thaliana* and *Arabidopsis lyrata*) and *Oenothera* (five plastid genomes of subspecies *Euoenothera*, material and method, table 16).

3.2.1 Evolution of cis-regulatory elements in *A. thaliana* and *A. lyrata*

Characterization of regulatory element mutations and their potential roles for alteration of respective gene expression has been surveyed leveraging on the dicotyledonous plant lineage *Arabidopsis*. The recently finished genome sequence of *A. lyrata* is evolutionary closely related to *A. thaliana* and is a suitable comparative template for *A. thaliana* (Hu et al, 2011). The evolutionary events in *Arabidopsis* lineage including gene duplication by polyploidization and speciation of *A. thaliana* and *A. lyrata* have been well characterized. High sequence similarity and conservation of collinear genetic arrangement due to close evolutionary distance have realized determination of a large amount of orthologous and paralogous genes. Also, the potential roles of discovered regulatory mutations can be unveiled thanks to comprehensive functional annotation of genes and numerous reported cis-elements in *A. thaliana*. Together, These features provide the opportunity to investigate cis-element evolution and their potential function in the *Arabidopsis* lineage.

Evolution of duplicate genes is considered as a source of phenotypic changes among species during speciation or generation of evolutionary novelties like neo- or subfunctionalization after gene duplication (Ohno, 1970). On the other hand, a pair of duplicated genes can also diverge in function as a result of changes in regulatory elements (Force et al, 1999). Instead of alterations of protein function and/or structure, phenotypic changes and generation of evolutionary novelties can be achieved by mutation of various sites. These sites, such as splice sites or cis-regulatory elements, govern the functional characteristics of the respective gene based on the duplication-degeneration-complementation (DDC) model (Force et al, 1999). The temporal or spatial gene expression under different stimuli can be changed, although the coding regions were preserved during evolution. Mutation of functional binding sites in promoter regions of duplicated loci, both orthologs and paralogs, were therefore analyzed in this study to unveil the potential effect of their cis-elements mutations on phenotypic variations in *Arabidopsis*.

Paralogous and orthologous gene pairs were firstly determined in *A. thaliana* and *A. lyrata*. Conservation/divergence of previously discovered cis-regulatory motifs by FIRE analysis was surveyed in the promoters of orthologous gene and paralogous gene pairs. Cis-elements demonstrated higher conservation rate in orthologous gene pairs between *A. thaliana* and *A. lyrata* compared to random ones. This is consistent with the observations that transcriptional regulation favors to be preserved in orthologs to realize their similar transcription activities among closely related species

(Blanchette et al, 2002; Blanchette and Tompa, 2002). Similarly, the higher conservation rate of cis-elements has been observed in paralogous gene pairs in *A. thaliana* and *A. lyrata* compared to random ones. This indicates that control of regulation of gene transcription tends to be conserved after gene duplication to increase the robustness of genetic networks. Notably, cis-elements were more conserved in orthologous genes of *A. thaliana* and *A. lyrata* than paralogous genes in both species. This observation reflects the evolutionary time scale of *Arabidopsis* that gene duplication by polyploidization took place earlier than speciation and is consistent with the expectation that higher preservation of regulatory elements exists in orthologs than in paralogs.

Along with this evolutionary process in *Arabidopsis* lineage, gene groups exist for which each member has both orthologous and paralogous partner. Such gene groups that are preserved during evolution indicate essential functionality of each member. On the other hand, polymorphisms, especially changes in the promoters, of one or several members are noteworthy and of special interest. Based on the DDC model, mutable multiple regulatory regions responsible for transcriptional regulation of these genes may play an essential role in alteration of expression for one or several members of these gene families and result in neo-/subfunctionalities (Force et al, 1999). Thus, evolution of cis-regulatory elements in these gene groups was surveyed in this study.

By integrating orthologous and paralogous relationship, gene networks were identified and classified (see material and methods). As a first focus, the networks containing four genes with complete paralogous/orthologous relationship were analyzed in this study (fig 16 c). Overall 341 gene groups with such quartettes structure were determined (see material and methods). Cis-elements detected by *FIRE* analysis were used to discuss the conservation/divergence of cis-elements in the promoters of genes in the respective quartettes. The number of shared motifs for the paralogous and orthologous gene pairs within quartettes was compared to the pairs not associated with quartettes and to the background pairs. Functional motifs were more conserved in orthologous gene pairs within quartettes compared to the pairs not associated with quartettes. However, such conservation bias was not observed in paralogous gene pairs within quartettes compared to the pairs not associated with quartettes. This indicates, on one hand, that gene groups that are preserved during evolution, e.g. quartettes in this case, are functional important and the highly conservation rate of the respective regulatory elements contributes to realize the maintaining of their functionalities during evolution. On the other hand, gene duplication in *Arabidopsis* lineage by polyploidization occurred much earlier than speciation process by *A. thaliana* and *A. lyrata*, so that the difference of sequence divergence rate can not be observed between the duplicated genes having orthologous gene partners and the one not having orthologous gene partners.

In this study, one particular situation of cis-element divergence within quartettes has

been discussed where regulatory elements are conserved in promoter regions of three gene members while missing in the fourth one. Difference between expectation that highly conservation of regulatory regions exists within quartettes and observation that mutation appears in single member of such extraordinary preserved network strongly suggests alteration of transcriptional regulation for the respective gene and potentially results in a novel function. In order to analyze this specific case, mutations of experimentally verified regulatory elements within quartettes were firstly determined and functional differences of *A. thaliana* genes within respective quartettes were then inspected utilizing the functional annotation of the *A. thaliana* genes. Consistence between mutated motifs and difference of respective gene function annotation has been observed in several dozens of quartettes which indicates the potential effect of cis-element mutations on changes of gene functionality.

This case study can be considered as a first start for large-scale characterization of cis-regulatory element evolution in close related species. Complementary to pairs or quartettes, mutations of cis-elements can be further discussed in more complex gene groups, e.g. partial paralog-ortholog gene networks. Consequently, cis-elements evolution of paralog/ortholog networks can be regarded as an important determinant for sub-functionalization according to duplication-degeneration-complementation model (Force et al, 1999; fig 16c). Besides, features in the cis-element sequence itself like orientation, location, copy number, even mutual distance and interaction of individual regulatory sites may also cause differential regulation of respective gene. Hence, complementary to simply inspect presence/absence of regulatory sites, polymorphism of these features can be used for an in-depth investigation of evolutionary process of cis-elements. However, detailed inspection of gene transcription activity and functionality is so far restricted to *A. thaliana*. With increase knowledge of *A. lyrata* and further close related higher plant species by analysis of expression data sets and gene functional annotation, comprehensive survey of phenotypic variation due to the evolution of regulation can be undertaken.

3.2.2 Search for genetic determinants causal to plastome-genome incompatibilities (PGIs) in *Oenothera*

In higher plants, one of the disturbances in the development of the resulting cybrids or hybrids caused by co-evolution of the intracellular genetic compartments is the plastome-genome incompatibility (PGI). Particular combinations of nucleus and plastids can affect the development of chloroplast and the chlorophyll synthesis (Stubbe, 1959). PGI has been observed and studied in various higher plants (Pandey et al, 1987; Przywara et al, 1989; Metzloff et al, 1982; Arisumi 1985), especially well characterized in the flowering plant *Oenothera* (Renner, 1924, 1936; Stubbe, 1955, 1959). The genetic study of Stubbe (1959) on *Oenothera* subsection (*Eu*)*oenothera* categorizes plastids of this subsection into five basic, genetically distinguishable plastome types (I, II, III, IV, V; material and methods table 16). A complete description of compatibilities of all combinations of these plastomes associated with

basic nuclear genomes has been performed (Stubbe, 1959; Fig. 19). Though detailed described, the mechanism and determinants for PGI in *Oenothera* are still far from clear. The search for genomic determinants is restricted due to a lack of both plastid and nucleus genomes.

Recently, the complete nucleotide sequences of representatives of the five basic *Oenothera* plastomes, their comparison, evolutionary relationships, and temporal relation have been published/presented (Greiner, 2008). Such sequences from closely related, morphologically distinct, and still interbreeding species for which organelle and nuclear genetics including cybrid technology is available offered the possibility to undertake large-scale search for plastid-specific genetic determinants causal to compartmental co-evolution.

Compartmental coevolution is accompanied by distinct changes in the 5 available Onagracean organelle chromosomes and the respective nuclear genomes. In subsection *Oenothera*, relationships between plastome and genome are crucial for the vitality of interspecific hybrids (Stubbe 1964; Dietrich et al. 1997). All its species can be crossed with one another, forming seeds with fully developed hybrid embryos that usually germinate and produce fertile progeny. However, the development of such hybrids is frequently disturbed and limited only by incompatibilities between the plastome of one parent plant with the genotype of the other one when the genetic compartments were not coevolved (fig. 19). Reversibility of interspecific compartmental incompatibility is a distinguishing feature to nuclear and plastid mutations affecting the organelle; an incompatible plastid foreign to a nucleus, for instance, will regreen if recombined with its genuine genome. Therefore, PGI is not based on mutations in single genes but in changed interactions of coevolved gene pairs, one of which resides in the chloroplast, the other in the nuclear genome. These PGI gene pairs represent a special case of the Dobzhansky–Muller model underlying their impact in speciation processes (Dobzhansky 1937). However, our knowledge to computationally predictable functional elements from primary sequences is limited, and *Oenothera* nuclear genome sequences - the complementary part on which prespeciation and coevolution processes are acting - are missing. Thus, analysis was restricted to plastid-localized determinants.

The alignment of sequence differences between plastomes of closely related, interbreeding species to predict genetic elements combined with filtering by their evolutionary and functional relevance as well as combinatorial logics is a promising strategy to pinpoint potential plastid-localized determinants involved in compartmental coevolution since predictions are testable. At a molecular level, coevolution of polypeptides with their interaction partners, polypeptides with polypeptides or polypeptides with nucleic acid molecules, is a well-known phenomenon (Goh et al. 2000). Basically, it could reflect a regulatory and/or a structural basis for the PGIs. Diverging loci are therefore of intrinsic interest but the vast majority of such loci found in *Oenothera* either does not seem to be involved in

interspecific compartmental incompatibility or does not contribute in a simple way. Applying filtering criteria, several dozens of amino acid changes and mutations in promoter candidates have been identified which point to particular PGIs and are subject for further experimental testing.

The case study of the hybrid AB-I has proven that the strategy of systematic filtering on genetically well-defined material is useful and attests to the power of the method. It correlated the bleached AB-I phenotype with a distinct major locus, a plastome I-specific deletion in the *clpP*–*psbB* intergenic region with reduced PSII activity in AB-I. Bioinformatic and biophysical data are consistent with a primary lesion in PSII and reminiscent to PSII downregulation in a bleached *Arabidopsis* mutant with severe deficiencies of *psbB* transcripts and CP47 proteins (Meurer et al. 1996, 2002). This mutant also displays comparable fluorescence characteristics. Loci such as the *clpP*–*psbB* intergenic region deduced by this approach are therefore potential candidates that deserve further study of underlying molecular mechanisms of PGI.

4. Summary & outlook

In higher organisms, gene promoters and their cis-regulatory element composition play an essential role for the tight spatial-temporal regulation of gene transcription. Hence, identification and functional characterization of cis-elements has become a major task for a comprehensive understanding of the regulatory mechanisms and circuits. In this work, various complementary approaches that make use of phylogenetic approaches and/or the combination of genome and transcriptome data have been applied. Large-scale detection of cis-regulatory elements based on comparative genomics has been addressed in grass species for the first time. The analysis was made possible by the completion of the sorghum genome and the already available rice genome. Based on the transcriptional activity of rice genes, thousands of significant motifs have been detected in this study. This will provide experimental researchers with a prioritized list of candidates for the gene of interest and can guide experimental designs for numerous sorghum and rice genes. Moreover, a comprehensive collection of expression datasets monitoring various experimental conditions made an extension and improvement of genome-wide cis-element characterization in the dicotyledonous plant *Arabidopsis thaliana* compared to previous surveys. Together, the applied approaches for detection of cis-regulatory elements in this study has been evaluated and can serve as paradigm for analysis in further grasses and in higher plants in general. Additional grass genome projects, for instance *Brachypodium distachyon* (International Brachypodium Initiative, 2010), a wheat relative, and maize (Schnable et al, 2009) have recently been completed and can be expected to deliver important and information-rich comparative genome templates in future studies. This will enable and stimulate whole-genome comparative studies between three and more grass genome sequences. In particular, comparisons between two closely related grasses, maize and sorghum, will allow (i) branch-specific motifs to be accessed and, at the same time, (ii) the identification of motifs common to the monocot clade. In addition, with the improvement of the next-generation sequencing technology, ChIP-seq and RNA-seq data can be used for the cis-element discovery. ChIP-seq allows to genome-wide mapping of cis-elements, while co-expressed gene groups determined by RNA-seq data can be used to identify functional conserved cis-elements. Compared to ChIP-chip, ChIP-seq produced data with a much finer resolution at reasonable costs. Generating ChIP-seq and RNA-seq data are independent on genome annotation for prior probe selection and avoids biases introduced during hybridization of microarrays. Therefore, utilizing next-generation sequencing data can experimentally and computationally improve the genome-wide characterization of cis-elements. Moreover, numerous studies have shown that mutations of cis-regulatory elements in various organisms can cause functional significant consequences for morphology, physiology and behaviour. In this study, evolutionary analysis of novel and verified cis-elements has been undertaken in *Arabidopsis* lineage and, with a focus on cis-element evolution and divergence, among five distinct platid genomes of *Oenothera* species. Divergence of regulatory elements has been analyzed with respect to gene duplication by polyploidization and speciation

events in *A. thaliana* and *A. lyrata*. In *Oenothera*, footprints of functional transcription factor binding sites among five plastomes and their mutations potentially causing plastome-genome-incompatibilities have been discussed. In this study, a foundation on the path was laid to develop a compendium of functional elements for plants, including their compositions, evolutionary mechanisms and functions on phenotypic changes. With increasing number of completed plant genome sequences, comprehensive and high quality expression data and accurate functional gene annotations, the approaches for cis-element detection and the principles for evolutionary analysis of regulatory elements developed during this study can be applied to study the complex regulatory mechanisms and the evolution of cis-regulatory elements. Moreover, with increasing knowledge of individual cis-elements in higher plants and development of systems biology, in-depth characterization of cis-regulatory modules and their related dynamic transcriptional networks can be addressed in the forthcoming future.

5. Material and methods

5.1 Array lists included in each data set

Table 17 lists experiments designed for each array and data set used in this study.

MPSS data set
Young root: 14 days
Mature root: 60 days
Young leaf: 14 days
Mature leaf: 60 days
Etiolated seedling: dark grown seedling: 10 days
Germinating seed: 12h day/night cycle: 3 days
Stem: 60 days
Meristematic tissues: 60 days
Mature pollen
Mature stigma and ovary
Immature panicle: 90 days
Callus: 35 days
Root - Salt: 250mM salt for 24h
Leaf - Salt: 250mM salt for 24h
Root - Drought: 5 days
Leaf - Drought: 5 days
Root - Cold: 4°C for 24h
Leaf - Cold: 4°C for 24h

YALE-2 data set
Flag leaf drought stage 1
Flag leaf drought stage 2
Flag leaf drought stage 3
Flag leaf water recovery after drought stage 3
Flag leaf high-saltinity stage 1
Flag leaf high-saltinity stage 2
Flag leaf high-saltinity stage 3
Shoot drought stage 1
Shoot drought stage 2
Shoot drought stage 3
Shoot water recovery after drought stage 3
Shoot high-saltinity stage 1
Shoot high-saltinity stage 2
Shoot high-saltinity stage 3
Panicle drought stage 1
Panicle drought stage 2
Panicle drought stage 3
Panicle water recovery after drought stage 3
Panicle high-saltinity stage 1
Panicle high-saltinity stage 2
Panicle high-saltinity stage 3

YALE-1 data set
Embryo - Coleoptile - 12hr post imbibition
Embryo - Coleoptile - 24hr post imbibition
Embryo - Coleoptile - Dry Seed
Embryo - Epiblast - 12hr post imbibition
Embryo - Epiblast - 24hr post imbibition
Embryo - Epiblast - Dry Seed
Embryo - Plumule - 12hr post imbibition
Embryo - Plumule - 24hr post imbibition
Embryo - Plumule - Dry Seed
Embryo - Radicle - 12hr post imbibition
Embryo - Radicle - 24hr post imbibition
Embryo - Radicle - Dry Seed
Embryo - Scutellum - 12hr post imbibition
Embryo - Scutellum - 24hr post imbibition
Embryo - Scutellum - Dry Seed
Seedling - Leaf blade (2nd leaf) - Bulliform
Seedling - Leaf blade (2nd leaf) - Bundle sheath
Seedling - Leaf blade (2nd leaf) - Epidermal Long Cell
Seedling - Leaf blade (2nd leaf) - Mesophyll
Seedling - Leaf blade (2nd leaf) - Stomata
Seedling - Leaf blade (2nd leaf) - Vein
Seedling - Root elongation zone - Central metaxylem
Seedling - Root elongation zone - Cortex
Seedling - Root elongation zone - Endodermis
Seedling - Root elongation zone - Epidermis
Seedling - Root elongation zone - Stele/Vascular bundle
Seedling - Root maturation zone - Cortex
Seedling - Root maturation zone - Endodermis
Seedling - Root maturation zone - Epidermis
Seedling - Root maturation zone - Stele/Vascular bundle
Seedling - Root tip - Central metaxylem precursor
Seedling - Root tip - Cortex
Seedling - Root tip - Lateral root cap
Seedling - Root tip - Vascular bundle
Seedling - Shoot - Apical meristem
Seedling - Shoot - Axillary meristem
Seedling - Shoot - Axillary primordium
Seedling - Shoot - P1
Seedling - Shoot - P2
Seedling - Shoot - P3
Tissue - Whole leaf blade (fresh)
Tissue - Whole root (fresh)

5.2 Mapping probes of expression data to current annotation in PhyloCon analysis

Due to annotation updates and improved gene modeling, oligonucleotide and tag sequence mapping are frequently erroneous. Therefore, the 70mer oligonucleotides of the YALE-1 and YALE-2 arrays as well as the 17mer MPSS tag sequences have been remapped to current rice gene models of the RAP2 annotation. The sequence analysis software *Vmatch* (Kurtz, 2010) based on index structure has been used to deduce the position of probe matches in genomic sequences. For both type of probe sets, only probes that unambiguously identify exactly one gene were used for the analysis. Due to the technique of massively parallel signature sequencing, if more than one probe were mapped to one transcript, only the probe which is 3' most located to transcript was analyzed, as this is expected to be the most informative (Meyers et al, 2004).

5.3 Expression data processing and filtering in PhyloCon analysis

Filters were employed to identify signatures or oligonucleotide probes with abnormally low expression level generated by systematic errors. Starting from 22,271 successfully remapped MPSS signatures, a total of 19,396 reliable and significant signatures were selected as described in Meyers et al (2004). For YALE-1 data which have been normalized (Ma et al, 2005), significantly expressed probes were determined following the methods of Rinn et al (2003). Background expression is derived from all measurements of 58,404 oligonucleotide probes in 42 experiments. For each expression intensity level, percentage of measurements exceeding the respective expression level has been determined for two classes: (i) probes for which less than three replicates are higher than the respective expression level and (ii) probes for owning higher intensities for three or more replicates. As shown in the Figure 22, the expression intensity level of 410 exceeding which more than 95% of all measurements belong to class (ii) was considered as intensity cutoff. The probes expressed higher than this cutoff in at least one experiment were selected as significantly expressed. In total, 13,904 genes/probes showed significant expression levels from the 27,887 mapped probes. For YALE-2, filter results from the original analysis (Zhou et al, 2007) were adopted which derives 20,633 reliable expressed genes/probes from the 27,887 mapped probes. Median of intensities from 3 replicates of each experiment was firstly determined and then transformed to log₂ ratio. These ratios were considered as expression level of each experiment for YALE-2 probes.

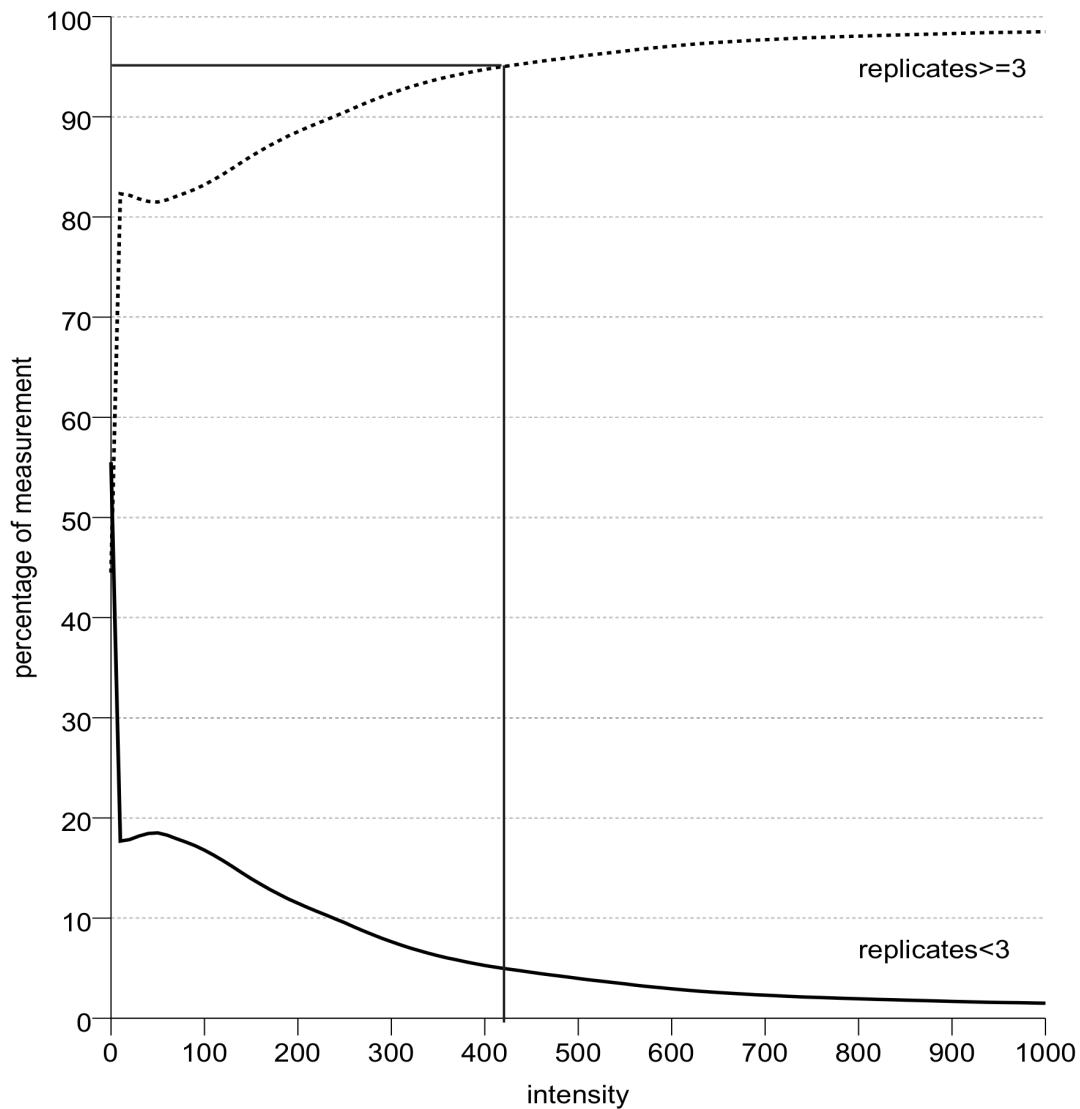


Figure 22: Determination of significant and reliable expression levels in YALE-1.

Significantly expressed probes have been determined according to Rinn et al (2003). Background expression is derived from all measurements of 58,404 oligonucleotide probes in 42 experiments. For each expression intensity, the percentage of measurements exceeding the respective expression level has been determined for two classes: (i) probes for which two or more replicates fell below the respective threshold and (ii) probes having higher intensities for 3 or more replicates. The x-axis depicts expression levels measured as the intensity of Cy5 dye, the y-axis the percentage of total measurements. For YALE-1, we found an expression intensity of 410 corresponding to the top 5% of all measurements.

5.4 Determination of co-expression gene groups in PhyloCon analysis

For each of the three expression data sets, co-expressed genes were defined as pairs whose pearson correlation exceeded the 99%-quantile of the background distribution of all correlation coefficients. Background and quantiles were estimated from the all-against all pearson correlation matrix. Thresholds for pearson correlation coefficients were 0.79, 0.88 and 0.93 for MPSS, YALE-1 and YALE-2, respectively

(figure 23). Gene pairs whose pearson correlation exceeded the threshold in the corresponding data set were considered as co-expressed. In total, co-expressed gene pairs covering 16,426, 13,223 and 18,820 distinct genes from MPSS, YALE-1 and YALE-2 were selected, respectively. A gene group was considered as co-expressed if all of its gene pairs are co-expressed. Graphic theory was employed to detect co-expressed gene groups by applying networkx package of python program (<https://networkx.lanl.gov/wiki>). Undirected graphs have been constructed with nodes representing genes and edges between gene pairs if they were co-expressed. From the graphs, co-expressed gene groups were extracted as maximal cliques for each node, the so called “anchor gene”. To avoid clusters with broad or unspecific expression patterns, analysis to nodes with ≤ 100 edges was restricted. As most nodes deduced tens to hundreds maximal cliques which overlap strongly, this increased the number of total maximal cliques dramatically and made further analysis for each co-expressed gene group infeasible. To solve this problem, only one clique for each anchor gene which contained the most members and showed the highest average pearson correlation was maintained. Then, all identical cliques derived from different anchor genes were removed to obtain a non-redundant list of co-expressed groups.

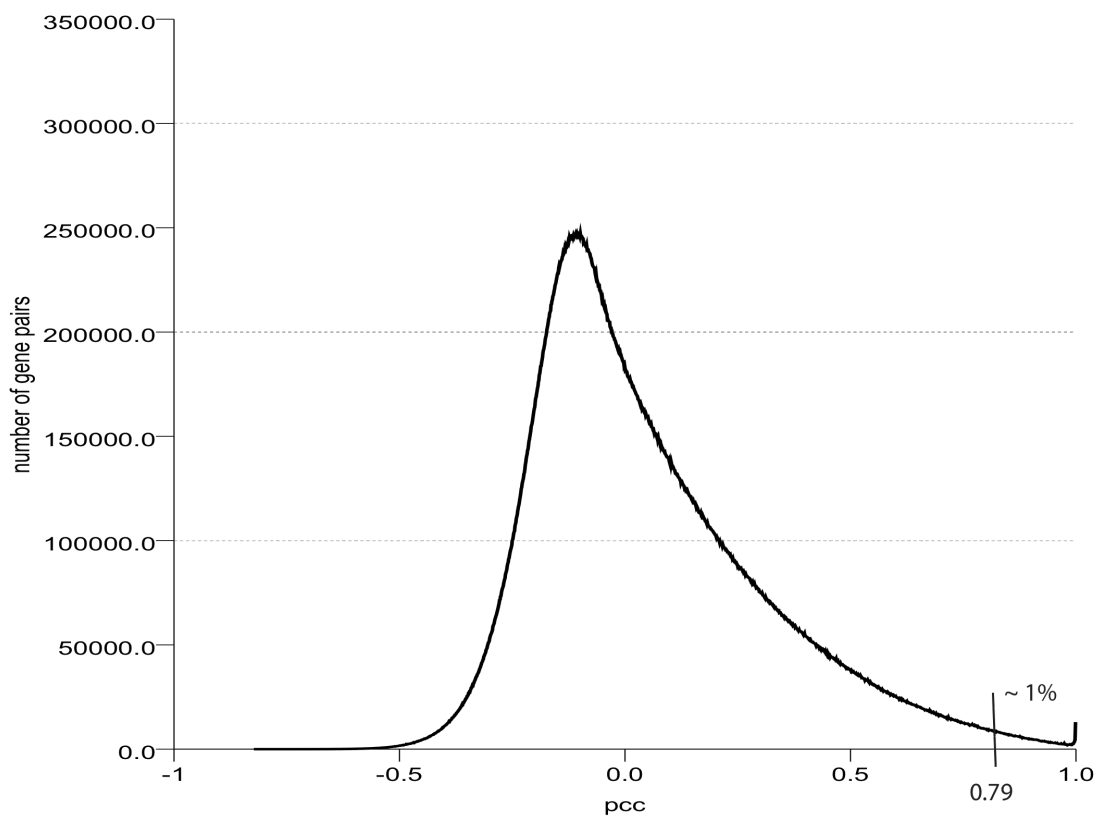


Figure 23: Background distribution for MPSS expression data.

Pearson correlations have been calculated for each gene versus all other genes. The correlation matrix has been used as background distribution for genome-wide expression similarities. The 99%-quantile has been numerically determined as significance level for co-expression of a gene pair. As an example, additional file 2 shows the background distribution for the MPSS expression data. X-axis depicts Pearson Correlation Coefficients, y-axis the number of gene pairs. The line marks the obtained 99%-quantile for MPSS at $r=0.79$.

5.5 *PhyloCon* motif discovery and analysis

PhyloCon was downloaded from <http://ural.wustl.edu/~twang/PhyloCon/> (Wang et al, 2003). A maximum of 10 intermediate matrices per cycle was retrieved. 200 temporary matrices per cycle were allowed and the number of SDS was set to 0.5 (for details, see Wang et al, 2003). Only forward strands of upstream sequences were analyzed.

To undertake further analysis for detected profiles including overrepresentation test and motifs clustering, alignment matrices were firstly transformed into position weight matrices (PWMs) to generate a scoring function for sequence instances and determine the threshold scores. 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 $j = 1,2,\dots,m$. The following formula has been applied to transform this $(4 \times m)$ -count/frequency matrix into a $(4 \times m)$ -PWM (Hertz, 1999):

$$a_{ij} = \log \frac{(n_{ij} + p_i) / (N + 1)}{P_i}$$

where N is the number of instances in the alignment, p_i the rice 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 is the respective score for nucleotide i at position j . Such PWMs assign a unique score for each sequence instance with length m . Let $S = s_1, s_2, \dots, s_m$ be a sequence of length m with each $s_j = 1, \dots, m$ representing a letter in the alphabet. The score to sequence S is calculated by summing up all $j = 1, 2, \dots, m$ single cell scores a_{ij} , where i corresponds to letter s_j . A cutoff score is specific for each PWM and derived from the instance in the original *PhyloCon* alignment matrix with the lowest score. A sequence is considered as an instance of a PWM if its score exceeds its cutoff score.

For each PWM, the statistical significance of its overrepresentation within the respective co-expression groups in comparison to all rice upstream sequences was tested. Cumulative binomial probability distribution was employed to test the significance of overrepresentation:

$$P - value = \sum_{k=1}^n \binom{n}{k} \cdot p^k \cdot (1 - p)^{n-k}$$

Where p , considered as expected frequency, is ratio between the number of rice upstream sequences containing motifs and the total number of sequences, while n is size of co-expressed gene groups and k is the number of genes in the groups which contain the motif. Occurrences in the rice gene sets of co-expressed groups generated observed frequencies of motif occurrences. P-values were adjusted for multiple hypotheses testing applying the Benjamini-Hochberg method (Benjamini et al, 1995)

to correct for a false discovery rate of 5%. Overall 17,068, 14,754 and 5,337 PWMs were obtained from the MPSS, YALE-1 and YALE-2 datasets, respectively ($P \leq 0.01$).

To obtain a set of non-redundant motifs for each data set, similar PWMs were subsequently merged. To estimate similarity of two profiles, multiple alignments of all their instances detected by *PhyloCon* were generated using *ClustalW* 1.74. As gaps rarely appear within cis-elements, parameters of *ClustalW* were set as gap opening penalty 1000 and gap extension penalty 0.001 to avoid intermediate gaps. Column score of alignments was the sum of all pairs of scores (match, mismatch and gap scoring 1, -1 and -2, respectively) and alignment score was the sum of column scores and normalized by size. A similarity matrix based on all-against all pairwise PWM alignments were constructed and hierarchical ‘bottom-up’-clustering has been performed by *hclust* in the R package. Clusters were determined by the cutoff corresponding to one-sided 5% significance level that has been deduced from the entire similarity matrix (used as background distribution). All instance sequences of profiles comprised in such clusters were extracted and used to rebuild multiple alignments. Gap-free regions were extracted as new motif profiles.

5.6 Validation of PhyloCon motifs

To validate detected motifs, they have been compared to previous reports in rice. 74 and 55 sequences of experimentally verified rice transcriptional factor binding sites have been extracted from PLACE and TRANSFAC databases, respectively (Higo et al, 1999; Matys et al, 2003). Overall 96 known sites were obtained after removing identical ones. As described previously, to compare the predicted profiles with reported sites, a multiple alignment between a known site and all instances included in the corresponding profile was generated and alignment scores were calculated. For each known site, the predicted profile with the highest alignment score was considered as match to the respective site. Visual inspection by comparing reported site to the sequence logo of best hit motif was undertaken to estimate the mapping quality.

5.7 Motif detection in rice and sorghum by network-level conservation

To detect occurrence of a particular motif, Overall 12,192 syntenic gene pairs determined by previous analysis in the section 3.1.1.1 were used and promoter regions were defined as genomic sequences from the start codon to the start of the upstream preceding gene, with a maximal distance of 2kb. String matching was employed to map motifs against promoter regions. The observed and expected co-occurrences of a particular motif were estimated as following formulas:

$$O_{\text{exp}} = (n_{\text{rice}} / N) \cdot (n_{\text{sorghum}} / N)$$

$$O_{\text{obs}} = m / N$$

Where $N = 12,192$, the total number of syntenic promoter pairs, m is the number of

syntenic promoter pairs in which both rice and sorghum promoter regions contain the motif, n_{rice} and n_{sorghum} represent the total number of rice and sorghum promoter regions, respectively, of syntenic promoter pairs which contain the motif. Genome-wide conservation scores were calculated as ratios between observed and expected co-occurrences.

Degenerated motifs were represented as regular expression patterns and generated from k-mer sites previously by substitution of single alphabet at any position to multiple letters. A heuristic degeneration process was applied in this study, as comprehensive enumeration of all possible variations at each position for each k-mer is computationally infeasible. For each exact word, a randomly chosen letter from the alphabet was added to one randomly selected position. For one round of degeneration, this procedure was repeated four times, i.e. up to four positions could be degenerated for one word. For each exact word, 100 independent iterations were analyzed. For each round of degeneration, the former generalized motif was replaced by the new one if score of the new generalized motif was higher.

In order to avoid redundancy, motifs with size variations showing word/string overlapping were merged by removal of motifs derived from other higher scored motifs. A motif is defined to be derived from another motif if it constitutes a word of this motif. For variable positions, both motifs had to have overlapping specificities, i.e. they had to share at least one letter from the alphabet. To reduce redundancy, motifs were ordered according to their score and all motifs from this list that were derived from a higher scoring motif were removed.

5.8 Dyad motif detection by network-level conservation

Cis-elements with close proximity can bind transcriptional factor simultaneously as motif pattern to ensure a coherent regulation of respective gens. To simulate such *cis*-regulatory structure, the model of dyadic motifs with patterns of type $\{X\}_a\{N\}_b\{X\}_a$ were also investigated in the study, where X represents a specific letter, N represents any letter from the nucleotide alphabet, and a and b ranging from 2 to 6 and 6 to 12, respectively. A greedy scheme was applied to test whether more specified versions of the initially unspecific spacer sequence results in a higher scoring motif. For each position in the spacer, the highest scoring representation of all 15 subset variations of the nucleotide alphabet was determined, for instance $\{A\}, \{AC\}, \{AG\}$ and so on. Next, each spacer position was replaced by its locally highest scoring letter representation starting from the position with the highest score improvement to the second, third and so on improvements. Motifs were re-scored after each replacement. Iterations were repeated as long as the total score of the motif improved, eventually resulting in dyadic motifs whose spacers have been completely substituted.

5.9 Motif detection by FIRE analysis

The prediction power of *FIRE* approach based on mutual information primarily depends on the wealth of expression data sets. In higher plants, *Arabidopsis thaliana*

has been comprehensively investigated and abundant expression profiling responding to a broad range of developmental stages and stimuli has been generated from a variety of microarray platforms. As comparison among different platforms may be problematic (The Toxicogenomics Research Consortium, 2005), measurements from a single gene chip, the Affymatrix ATH1 GeneChip (<http://www.affymetrix.com/products/arrays/index.affx?Arabidopsis>), were used in this study. This also represents currently the largest and most comprehensive expression data set of this species. Transcriptomic data sets have been extracted from the Nottingham *Arabidopsis* Stock Center's (*NASC*) microarray database (CD-ROM release as of Craigon et al, 2004) that mainly collects *AtGeneExpress* data sets (Schmid et al., 2005) and some experiments of individual laboratories

In total more than 2000 chips have been collected from *NASC* microarray database. They were designed to uncover the whole transcriptome of *Arabidopsis thaliana* under different developmental stages, growth condition treatments, pathogen infection, stress series and hormone treatments. RMA approach included in R package has been employed to normalize and processing data sets deduced by Affymetrix chips. Moreover, probes have been remapped to current CDF (chip definition file) to avoid erroneous and/or outdated mapping. Due to large number of chips which share similar experiment conditions, these 2000 microarray data sets have been manually grouped into 157 experiment clusters based on similarity of experimental condition (table 18).

Experiment	#Gene	#Cluster
ABA_timecourse_in_wt_seedlings	16120	24
Alex_McCormac	17245	38
Aluru_immutans_Variation_mutant_of_At	15439	18
Birnbaum_A_gene_expression_map_of_the_Arabidopsis_root	18071	50
Brendan_Davies	15278	17
Bryan_Pickett	14984	11
Cain_polycomb_binding_protein	14935	6
Capper_Effect_of_CaM_overExp_on_Arabidopsis_transcriptome	15177	11
Casson_laser_capture_micro_dissected_embryonic_tissues	16347	51
Chris_West	15234	20
Corrina_Hampton	15481	17
David_Brown	15232	18
David_Honys	16408	41
De_Grauwe	15687	18
De_Veylder_Arabidopsis_E2F_target_genes	15770	35
Dorthe_Villadsen_Hexose_signalling	15052	9
Edwards_Circadian_gene_Exp_under_different_light_treatments	17015	34
Edwards_Identifying_targets_of_FLC	14977	7
Edwards_genes_modelling_the_Arabidopsis_circadian_clock	15341	17
Eulgem_wt_and_constitutively_active_At_MAPK2	14941	7
Evans_membrane_fluidity_in_low_temperature_perception	15822	25
Finch-Savage_understanding_seed_dormancy	16721	27
Frank_Millenaar	15320	19
Fukuda_In_vitro_tracheary_element_transdifferentiation	16447	36
GA3_timecourse_wt_GA1_mutant_seedlings	16452	24
Gareth_Warren	17697	38
Gema_2	16907	39
Gema_Vizcay_Barrena	16587	39
Gould_circadian_clock_temperature_buffering_mechanism	15976	20
Greco_Raffaella	14959	7
Haruko_Okamoto	15280	19
Heinekamp_Low_chronic_exposure_to_Cs137	16812	41
Helenius_ABA-treatment	15056	16
Hennig_Early_Reproductive_Stages	15437	33
Hsiu-Ling-Yap_AM_signalling_pathways	14943	4
Hsiu_Ling_Yap	15121	14
IAA_timecourse_in_wt_seedlings	15361	16
Irene_Bramke	15545	32
Jeremy_Pritchard	15014	9
Jodi_Swidzinski	15862	47
Johanna_Cornah	16449	32
Johanna_Cornah_Glyoxylate_cycle	15606	23
John_Hammond_3	15488	28
John_Hammond_4	17998	44
Joy_Boyce	15280	20
Julie_Bruggemann	15556	27
Julie_Lloyd_pho3_mutation	15573	23
Kadalayil_a_new_Arabidopsis_SH2_domain_containing_gene	14924	2
Karen_Greville	16194	23
Knight_Dark-induced_gene_expression_in_sfr6	15326	23
Laura_Heggie	16282	44
Laura_Heggie_Stomatal_development	16273	30
Lindsey_laser-capture_micro-dissected_embryonic_tissues	17113	53
Maike_Rentel	15123	14
Malcolm_Campbell	16479	30
Marc_Knight	15168	18
Marchant_comp_auxin_res_axr4_mutant_wt_col0	14909	2
Mark_Diamond	15300	17
Mark_Jones	15287	20
Mark_Jones_AtrbohC_mutation	15103	17
Marocco_temperature_sensitive_mutants	14924	6
Marta_de_Torres_Zabala	17847	40

Menges At cell suspension	15597	22
Mike Wheeler	16668	39
Mittler Hydrogen peroxide stress and Zat12 over-Exp	15499	21
Mittler Over-expression of MBF1c enhances stress tolerance	15232	13
Nick Jordan	15050	11
Nicky Evans	15659	40
Paul Jarvis 2	15011	10
Peter Urwin Nematode susceptibility	14978	8
Pieterse pathogen and insect attack	16733	44
Pourtau sugar acc during early leaf senescence	15830	32
Pracharoenwattana Peroxisomal mdh mutant	14946	8
Robin Walters	15120	13
Rosalia Deeken Tumour development	15454	29
Sally Ward Meristem activity	14943	8
Scarse-Field wt and 2 ko alleles of At CAMTA	14975	8
Schmid FRI FLC combos	15313	15
Schroeder Response to potassium starvation in roots	15033	10
Shane Murray	14965	9
Shirras Env Genomics of Calcicole-calcifuge physiology	15492	38
Shirras env genom calcicole calcifuge physiology	15296	21
Simon Turner	16254	40
Somerville Tissue Type Arrays of Columbia-0	16063	41
Sophie Filleur	15316	15
St Clair Exp Level Polymorphism Project ELP Col-0	16566	30
St Clair Exp Level Polymorphism Project ELP Cvi-1	16792	27
St Clair Exp Level Polymorphism Project ELP Kin-0	17153	35
St Clair Exp Level Polymorphism Project ELP Mt-0	17037	30
St Clair Exp Level Polymorphism Project ELP Tsu-1	17223	32
St Clair Exp Level Polymorphism Project ELP Van-0	17299	35
St Clair Expression Level Polymorphism Project ELP Est	16910	31
Steve Smith	16668	37
Susannah Bird	15158	15
Thorlby Gene Expression During Recovery from Freezing	15085	18
Thornton Growth promotion Trichoderma hamatum	14924	1
Ulm UV-B response of Arabidopsis	15277	16
Underwood At P.syringae pv. tomato DC3000 interaction	18438	40
Vicky Buchanan Wollaston	16261	38
Vogel Response to CBF2 expression	15016	12
Vogel Response to ZAT12 expression	15059	14
Vogel Response to cold plate grown plants	15845	35
Vogel Response to cold soil grown plants	15513	27
Weigel Floral transition and early flower development	15870	25
Werner mycotoxin treatment wt altered sens mutant	15269	19
William Willats Pectin Biosynthesis	16466	35
Yang Mutant array	15862	27
Yinbo Gan	14924	4
Zeatin timecourse in wt seedlings	15288	13
acc time course in wildtype seedlings	15182	12
arr21c overexpression	15559	22
basic hormone treatment of seeds	16145	23
bergua funct genom shoot meristem dormancy	17978	38
brassinolide timecourse wt det2 seedlings	16196	22
comparison of plant hormone related mutants	15544	21
cytokinin treatment of seedlings	16304	47
dev series flower and pollen	19129	43
dev series leaves	18091	35
dev series mutants and other ecotypes	15871	25
dev series roots	16841	36
dev series seedlings and whole plants	16977	35
dev series shoots and stems	18094	32
dev series siliques and seeds	17775	44
different temperature treatment of seeds	15116	19
effect GA inhibitors on seeds	15613	13

effect_brassinosteroid_inhibitors_seed	15433	16
effect_ibuprofen_SA_daminozide_seedlings	15412	15
effect_of_ABA_during_seed_imbibition	17546	36
effect_of_auxin_inhibitors_on_seedlings	15595	19
effect_of_brassinosteroids_in_seedlings	16021	16
effect_of_cycloheximide_on_seedlings	15634	35
effect_of_ethylene_inhibitors_on_seedlings	15734	25
effect_of_photosynthesis_inhibitor_PNO8_on_seedlings	16020	22
effect_of_proteasome_inhibitor_MG13_on_seedlings	15001	10
exp_profiling_early_germinating_seeds	15690	24
light_treatments	16208	22
methyl_jasmonate_timecourse_in_wt	15515	20
newbury_zinc_tolerance_and_accumulation_a_halleri	17287	53
pseudomonas_half_leaf_injection	16338	24
respons_bac_oomycete_elicitors	17632	38
response_to_Botrytis_cinerea_infection	16490	38
response_to_erysiphe_orontii_infection	16582	22
response_to_phytophthora_infestans	16978	43
response_to_sulfate_limitations	15813	19
response_vir_avir_sec_def-host_nonhost_bac	18159	41
short_functional_genomics_of_ozone_stress_in_Arabidopsis	15874	35
stress_treatments_cold_stress	18283	52
stress_treatments_control_plants	18217	38
stress_treatments_drought_stress	18241	42
stress_treatments_genotoxic_stress	18081	42
stress_treatments_heat_stress	18662	48
stress_treatments_osmotic_stress	18204	49
stress_treatments_oxidative_stress	18205	41
stress_treatments_salt_stress	18320	46
stress_treatments_uvB_stress	18356	43
stress_treatments_wounding_stress	18240	42
yang_silique_senescence	15817	36

Table 18: Experiments with detailed description and corresponding numbers of genes and gene clusters

Table summarizes all 157 experiments surveyed in this study with description. The number of genes and gene clustered included in each experiment are listed in the last two columns.

Determination of co-expressed gene groups was carried out by CRC analysis. This program is designed to cluster microarray gene expression data collected from multiple experiments and was downloaded from <http://www.sph.umich.edu/csg/qin/CRC/index.shtml>. 20 chains and cycles were allowed and probability cutoff was set as 0.9.

5.10 Motif detection between *A. lyrata* and *A. thaliana* by network-level conservation

FASTCOMPARE was applied to discover phylogenetic footprints between *A. thaliana* and *A. lyrata* with minor changes compared to its application in rice and sorghum. All possible motifs with size range from 6 to 10mers were generated by any combination of nucleotide alphabet to cover the length range of *FIRE* detected motifs. To detect occurrence of a particular motif, Overall 17,521 syntenic gene pairs determined by Hu et al (2011) were used and promoter regions were defined as

genomic sequences from the start codon to the start of the upstream preceding gene, with a maximal distance of 2kb. Z-scores of given motifs transformed by ratios between observed and expected co-occurrence corresponds to genome-wide conservation rate between *A. thaliana* and *A. lyrata* and represents the confidence of corresponding detected motifs.

5.11 Determination of tandem and segmental duplication in *A. thaliana* and *A. lyrata*

Computational selection of segmental and tandem duplications followed stringent similarity thresholds. FASTA3 package, considered as DNA or protein homology search tool based on algorithm of global alignment (Pearson et al, 1988), was applied to estimate the similarities of gene pairs. Default parameters were used. The ratio between opt score and self score generated by FASTA3 was calculated for each gene pair and regarded as the corresponding similarity. Ratio of 0.3 was set as the similarity threshold and gene pairs whose similarities were below the threshold were ruled out. Along the filtering criteria, segmentally duplicated groups consisted of gene pairs which exceeded the threshold and located in the duplicated segments determined previously by Hu et al (2011). Tandemly duplicated gene groups were selected as groups of which all gene pairs within corresponding groups exceeded the threshold and were separated by less than five intermediate unrelated genes. Similar to the PhyloCon analysis in rice and sorghum described previously, undirected graphs have been constructed with nodes representing genes and edges between gene pairs if they shared significant similarity and were separated by less than five intermediate unrelated genes. From the graphs, tandemly duplicated gene groups were extracted as maximal cliques.

5.12 Determination of complete paralog-ortholog gene networks

The networks containing four genes with complete paralogy/orthology relationship were analyzed in this study (fig 15 c). The gene groups with such quartettes structure were firstly determined based on previously detected paralogs and orthologs. The orthologous gene pairs were scanned and the pairs were firstly selected in which both members possess tandemly or segmentally duplicated partners within their own species selected previously. Quartettes were then determined if orthologous relation was identified between these duplicated partners.

5.13 The flowing plant genus *Oenothera* subsection (*Eu*)*oenothera*

Five genetically distinguishable plastid genomes of *Euoenothera* analyzed in this study are listed in the table 16.

Plastome I	<i>Oenothera elata</i> subsp. <i>hookeri</i> strain <i>johansen</i>
Plastome II	<i>Oenothera biennis</i> strain <i>suaveolens</i> <i>Grado</i>
Plastome III	<i>Oenothera glazioviana</i> strain <i>rr-lamarckiana</i> Sweden
Plastome IV	<i>Oenothera parviflora</i> strain <i>atrovirens</i>
Plastome V	<i>Oenothera argillicola</i> strain <i>douthat</i> 1

6. Reference

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.

Arisumi T (1985). Rescuing abortive *Impatiens* hybrids through aseptic culture of ovules. *J. Am. Hortic. Sci.* 110: 273-276.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14 (3): 283–91.

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37: W202-208.

Banerji J, Rusconi S, Schaffner W (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* 27: 299-308.

Beer M, Tavazoie S (2004). Predicting gene expression from sequence. *Cell* 117: 185–198.

Bejerano G, Lowe CB, Ahituv N, King B, Siepel A, Salama SR, Rubin EM, Kent WJ, Haussler D (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* 441: 87-90.

Ben-Tabou DS, Davidson EH (2007). Gene regulation: gene control network in development. *Annu. Rev. Biophys. Biomol. Struct.* 36:191

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B. Methodol.* 57: 289-300.

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LM, Yoon J, Doyle A, Lander G, Moseyko N, Yoo D, Xu I, Zoeckler B, Montoya M, Miller N, Weems D, Rhee SY (2004). Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* 135(2):1-11.

Blanc G, Hokamp K, Wolfe KH (2003). A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* 13: 137–144.

Blanchette M, Tompa M (2003). FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.* 13: 3840-3842.

Blanchette M, Schhwikowski B, Tompa M (2002). Algorithms for phylogenetic footprinting. *J. Comput. Biol.* 9: 211-223.

Blanchette M, Tompa M (2002). Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* 12: 739-748.

Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299: 1331-3.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, et al. (2007). Divergence of transcription factor binding sites across related yeast species. *Science* 317: 815-819.

Bowers JE, Abbey C, Anderson S, Chang C, Draye X, Hoppe AH, Jessup R, Lemke C, Lenington J, Li Z, Lin YR, Liu SC, Luo L, Marler BS, Ming R, Mitchell SE, Qiang D, Reischmann K, Schulze SR, Skinner DN, Wang YW, Kresovich S, Schertz KF, Paterson AH (2003). A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. *Genetics* 165: 367-86.

Bowers JE, Chapman BA, Rong J, Paterson AH (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422: 433-438.

Bray N, Dubchak I, Pachter L (2003). AVID: a global alignment program. *Genome Res.* 13: 97-102.

Brown CT, Callan CG Jr. (2004). Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli*. *Proc Natl Acad Sci USA* 101: 2404-2409.

Bryne JC, Valen E, Tang ME, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A (2007). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* 36: D102-106.

Burrows PA, Sazanov LA, Svab Z, Maliga P, Nixon PJ (1998). Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid *ndh* genes. *EMBO J.* 17: 868-876.

Cahoon AB, Stern DB (2001). Plastid transcription: a ménage à trois. *Trends Plant Sci.* 6(2): 45-6.

Carroll SB, Grenier JK, Weatherbee SD (2001). From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Science. Oxford, Malden, Mass.

Carter D, Chakalova L, Osborne CS, Dai YF, Fraser P (2002). Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* 32: 623-626.

Cereghini S, Saragosti S, Yaniv M, Hamer DH (1984). SV40-a-globulin hybrid minichromosomes. Differences in DNaseI hypersensitivity of promoter and enhancer sequences. *Eur. J. Biochem.* 144: 545-553.

Chiu WL, Stubbe W, Sears BB (1988). Plastid inheritance in *Oenothera*: organelle genome modifies the extent of biparental plastid transmission. *Current Genetics.* 13:181-189

chlorophyll a-apoprotein of the photosystem II reaction centre from spinach. *Mol. Gen. Genet.* 201: 115–123.

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

Cleland RE (1972). *Oenothera*: cytogenetics and evolution. Academic Press, New York.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003). Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301: 71-76.

Cover T, Thomas J (2006). *Elements of Information Theory*. Second Edition (Hoboken, NJ: Wiley-Interscience).

Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004). NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.* 32: D575-7.

Davidson EH (2006). *The regulatory genome: gene regulatory networks in development and evolution*. Academic, Burlington.

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003). AGRIS: *Arabidopsis* gene regulatory information server, an information resource of *Arabidopsis* cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4: 25.

Day WH, McMorris FR (1992). Critical comparison of consensus methods for molecular sequences. *Nucl. Acids Res.* 20: 1093-1099.

de Vetten NC, Ferl RJ (1995). Characterization of a maize G-box binding factor that is induced by hypoxia. *Plant J.* 7(4):589-601.

Dermitzakis ET, Bergman CM, Clark AG (2003). Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. *Mol. Biol. Evol.* 20: 703-714.

Dermitzakis ET, Clark AG (2002). Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19: 1114-1121.

Dietrich W, Wagner WL, Raven PH (1997). Systematics of *Oenothera* section *Oenothera* subsection *Oenothera* (Onagraceae). In: Anderson C, editor. Systematic botany monographs. Laramie (WY): The American Society of Plant Taxonomists. p. 1-123.

Doniger SW, Fay JC (2007). Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3: e99.

Drosophila 12 Genomes Consortium (2007). Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450(7167): 203-18

Elemento O, Slonim N, Tavazoie S (2007). A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell.* 28(2): 337-50.

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.

Emberly E, Rajewsky N, Siggia ED (2003). Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics* 4: 57.

Eulgem T, Rushton PJ, Schmelzer E, Hahlbrock K, Somssich IE (1999). Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. *EMBO J.* 18: 4689-99.

Felsenfeld G (2003). Quantitative approaches to problems of eukaryotic gene expression. *Biophys. Chem.* 100: 607-613.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531-1545.

Loots G, Ovcharenko I (2004). rVista 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.* 32(Web Server Issue), W217-W221

Gallo SM, Li L, Hu Z, Halfon MS (2006). REDfly: a regulatory element database for *Drosophila*. *Bioinformatics* 22(3): 381-383.

Gräff J, Kim D, Dobbin MM, Tsai LH (2010). Epigenetic regulation of gene expression in physiological and pathological brain processes. *Physiol. Rev.* 91: 603-649

Greiner S, Wang X, Rauwolf U, Silber MV, Mayer K, Meurer J, Haberer G, Herrmann RG (2008). The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution. *Nucleic Acids Res.* 36: 2366–2378.

Greiner S, Wang X, Herrmann RG, Rauwolf U, Mayer K, Haberer G, Meurer J (2008). The complete nucleotide sequences of the 5 genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: II. A microevolutionary view using bioinformaticis and formal genetic data. *Mol. Biol. Evol.* 25(9): 2019-2030

Gross DS, Garrard WT (1998). Nuclease hypersensitive sites in chromatin. *Annu.Rev.Biochem.* 57: 159-197.

Guhathakurta D (2006). Computational identification of transcriptional regulatory elements in DNA sequence. *Nucl. Acid. Res.* 34: 3585-3598

Guo H, Moose SP (2003). Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. *Plant Cell* 15: 1143-58.

Hajdukiewicz PT, Allison LA, Maliga P (1997). The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *EMBO J.* 16: 4041–4048.

Hajdukiewicz PTJ, Allison LA, Maliga P (1997). The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. *EMBO J.* 16: 4041–4048.

Hedtke B, Börner T, Weihe A (1997). Mitochondrial and chloroplast phage-type RNA polymerases in *Arabidopsis*. *Science* 277: 809–811.

Hertz GZ, Stormo GD (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15: 563–577.

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

Hobo T, Asada M, Kowyama Y, Hattori T (1999). ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. *The Plant Journal* 19(6): 679-689.

Homann A, Link G (2003). DNA-binding and transcription characteristics of three cloned sigma factors from mustard (*Sinapis alba* L.) suggest overlapping and distinct roles in plastid gene expression. *Eur. J. Biochem.* 270: 1288–1300.

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottillar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet.* 2011 43(5):476-81.

Hughes JD, Estep PW, Tavazoie S, Church GM (2000). Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* 296(5): 1205-14.

International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463(7282): 763-8.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436: 793-800.

Ishihama A (1988). Promoter selectivity of prokaryotic RNA polymerase. *Trends Genet.* 4: 282-286.

Jacob F, Monod J (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318-356.

Kanamaru K, Tanaka K (2004). Roles of chloroplast RNA polymerase sigma factors in chloroplast development and stress response in higher plants. *Biosci. Biotechnol.*

Biochem. 68: 2215–2223.

Kapoor S, Sugiura M (1999). Identification of two essential sequence elements in the nonconsensus type II PatpB-290 plastid promoter by using plastid transcription extracts from cultured tobacco BY-2 cells. *Plant Cell* 11(9): 1799-810.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*: 241-254.

Kim J (2001). Macro-evolution of the hairy enhancer in *Drosophila* species. *J. Exp. Zool.* 291: 175-185.

Koch MA, Weisshaar B, Kroymann J, Haubold B, Mitchell-Olds T (2001). Comparative genomics and regulatory evolution: conservation and function of the *Chs* and *Apetal3* promoters. *Mol. Biol. Evol.* 18: 1882-1891.

Kochevenko AS, Ratushnyak YI, Korneev DY, Stasik OO, Shevchenko VV, Kochubey SM, Gleba YY (1999). Study of the state of photosynthetic apparatus in cybrid tomato plants possessing traits of nuclear-cytoplasmic incompatibility. *Russ. J. Plant Physiol.* 46: 474-481.

Kofer W, Koop HU, Wanner G, Steinmueller K (1998). Mutagenesis of the genes encoding subunits A, C, H, I, J and K of the plastid NAD(P)H-plastoquinone-oxidoreductase in tobacco by polyethylene glycol-mediated plastome transformation. *Mol. Gen. Genet.* 258: 166–173.

Kuo MH, Allis CD (1999). In vivo cross-linking and immunoprecipitation for studying dynamic Protein:DNA associations in a chromatin environment. *Methods* 19: 425-433.

Latchman DS (1997). Transcription factors: an overview. *Int. J. Biochem. Cell. Biol.* 29: 1305-1312.

Leon P, Arroyo A (1998). Nuclear control of plastid and mitochondrial development in higher plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 49: 453-480.

Lerbs-Mache S (1993). The 110-kDa polypeptide of spinach plastid DNA-dependent RNA polymerase: single-subunit enzyme or catalytic core of multimeric enzyme complexes? *Proc. Natl. Acad. Sci. USA* 90: 5509–5513.

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P,

Rombauts S (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30: 325-327.

Levine M, Tjian R (2003). Transcription regulation and animal diversity. *Nature* 424: 147-51.

Lockton S, Gaut BS (2005). Plant conserved non-coding sequences and paralogue evolution. *Trends Genet.* 21: 60-5.

Lonetto M, Gribskov M, Gross CA (1992). The sigma 70 family: sequence conservation and evolutionary relationships. *J. Bacteriol.* 174: 3843-3849.

Lowe CB, Bejerano G, Haussler D (2007). Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci. USA* 104: 8005-8010.

Ludwig MZ, Bergman C, Patel N, Kreitman M (2000). Evidence for stabilizing selection in a eukaryotic cis-regulatory element. *Nature* 403: 564-567.

Ludwig MZ, Patel N, Kreitman M (1998). Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development* 125: 949-958.

Ludwig MZ (2002). Functional evolution of noncoding DNA. *Curr Opin Genet Dev* 12: 634-639.

Ma L, Chen C, Liu X, Jiao Y, Su N, Li L, Wang X, Cao M, Sun N, Zhang X, Bao J, Li J, Pedersen S, Bolund L, Zhao H, Yuan L, Wong GK, Wang J, Deng XW, Wang J (2005). A microarray analysis of the rice transcriptome and its comparison to *Arabidopsis*. *Genome Res.* 15: 1274-1283.

Maclsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006). An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7: 113.

Maeda RK, Karch F (2011). Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Current Opinion in Genetics & Development* 21: 187-193

Mahony S, Carcoran DL, Feingold E, Benos PV (2007). Regulatory conservation of protein coding and microRNA gene in vertebrates: lessons from the opossum genome.

Genome Biol 8: R84.

Maleck K, Levine A, Eulgem T, Morgan A, Schmid J, Lawton KA, Dangl JL, Dietrich RA (2000). The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nat. Genet.* 26(4): 403-10.

Maliga P (1998). Two plastid RNA polymerases of higher plants: an evolving story. *Trends Plant Sci.* 3: 4–6.

Matys V et al (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucl. Acids Res.* 31: 374-378.

McGregor AP, Shaw PJ, Hancock JM, Bopp D, Hediger M, Wratten NS, Dover GA (2001). Rapid restructuring of bicoid-dependent hunchback promoters within and between Dipteran species: implications for molecular coevolution. *Evol. Devel.* 3: 397-407.

Metzlaff M, Pohlheim F, Börner T, Hagemann R (1982). Hybrid variegation in the genus *Pelargonium*. *Curr. Genet.* 5: 245-249.

Meyers BC, Galbraith DW, Nelson T, Agrawal V (2004). Methods for Transcriptional profiling in plants. Be fruitful and replicate. *Plant Physiology* 135: 637-652.

Meyers BC, Tej SS, Vu TH, Haudenschild CD, Agrawal V, Edberg SB, Ghazal H, Decola S (2004). The use of MPSS for whole-genome Transcriptional analysis in *Arabidopsis*. *Genome Res.* 14: 1641-1653.

Molina C, Grotewold E (2005). Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics.* 6(1):25.

Monod J, Jacob F (1961). General conclusions – teleonomic mechanisms in cellular metabolism. Growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* 26: 389-401.

Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.* 3: 19.

Moses AM, Chiang DY, Pollard DA, Lyer VN, Eisen MB (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.* 5: R98.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li X-Y, et al. (2006). Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Comput Biol* 2: e130.

Narlikar L, Ovcharenko I (2009). Identifying regulatory elements in eukaryotic genomes. *Briefing in functional genomics and proteomics* 8: 215-230.

Nobuta K, Venu RC, Lu C, Belo A, Vermaraju K, Kulkarni K, Wang W, Pillay M, Green PJ, Wang GL, Meyers BC (2007). An expression atlas of rice mRNAs and small RNAs. *Nature Biotechnology* 25: 473-477.

Nordin K, Vahala T, Palva ET (1993). Differential expression of two related, low-temperature-induced genes in *Arabidopsis thaliana* (L.) Heynh. *Plant Mol. Biol.* 21(4):641-53.

O'Brien TP, Bult CJ, Cremer C, Grunze M, Knowles BB, Langowski J, McNally J, Pederson T, Politz JC, Pombo A, Schmahl G, Spatz JP, van Driel R (2003). Genome function and nuclear architecture: from gene expression to nanoscience. *Genome Res.* 13: 1029-1241.

Ohno S (1970). *Evolution by Gene Duplication*. Springer-Verlag, New York.

Pandey KK, Grant JE, Williams EG (1987). Interspecific hybridisation between *Trifolium repens* and *T. uniflorum*. *Aust. J. Bot.* 35: 171-182.

Panstruga R (2005). Serpentine plant MLO proteins as entry portals for powdery mildew fungi. *Biochem. Soc. Trans.* 33: 389-92.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberler G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457(7229): 551-556.

Pearson WR, Lipman DJ (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444-2448.

Pennisi E (2007). Genome sequencing. The greening of plant genomics. *Science* 317: 317.

Prakash A, Tompa M (2005). Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.* 23: 1249-1256.

Przywara L, White DWR, Sanders PM, Maher D (1989). Interspecific hybridization of *Trifolium repens* with *T. hybridum* using in ovulo embryo and embryo culture. *Ann. Bot.* 64: 613-624.

Quackenbush J (2003). Micorarrays - guilt by association. *Science* 302(5643): 240-1.

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E (2000). Genome-wide location and function of DNA binding proteins. *Science* 290: 2306-2309.

Renner O (1924). Die Scheckung der Oenotherenbastarde. *Biol. Zentralblatt* 44: 309-336.

Renner O (1936). Zur Kenntnis der nichtmendelnden Buntheit der Laubblätter. *Flora* 30: 218-290.

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G (2000). *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290: 2105-2110.

Rinn JL, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M (2003). The transcriptional activity of human chromosome 22. *Genes & Dev.* 17: 529-540.

Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J (2001). RegulonDB(version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* 29:72-4.

Jeong SY, Rebeiz M, Andolfatto P, Werner T, True J, Carroll SB (2008). The evolution of gene regulation underlies a morphological difference between two drosophila sister species. *Cell* 132: 783-793.

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

Schnable et al (2009). The B73 maize genome: complexity, diversity and dynamics.

Science 326(5956): 1112-1115

Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188: 415-431.

Schoertz F, Bathelt H (1964). Pigmentanalytische Untersuchungen an *Oenothera*. II. Der Albivelutina-Typ. *Planta* 62: 171-190.

Schwenkert S, Umate P, Dal Bosco C, Volz S, Mlcxochova' L, Zoryan M, Eichacker LA, Ohad I, Herrmann RG, Meurer J (2006). PsbI affects the stability, function, and phosphorylation patterns of photosystem II assemblies in tobacco. *J. Biol. Chem.* 281: 34227–34238.

Segal JA, Barnett JL, Crawford DL (1999). Functional analysis of natural variation in Sp1 binding sites of a TATA-less promoter. *J. Mol. Evol.* 49: 736-749.

Seipel K, Georgiev O, Schaffner W (1992). Different activation domains stimulate transcription from remote ('enhancer') and proximal ('promoter') positions. *EMBO J.* 11: 4961-4968.

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002). Functional Annotation of a Full-Length *Arabidopsis* cDNA Collection. *Science* 296(5565): 141-5.

Shinozaki K, Yamaguchi-Shinozaki K (2000). Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signaling pathways. *Curr. Opin. Plant Biol.* 3(3):217-23.

Siddharthan R, Siggia ED, van Nimwegen E (2005). PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* 1(7):e67.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Roesnbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15: 1034-1050.

Silhavy D, Maliga P (1998). Mapping of promoters for the nucleus-encoded plastid RNA polymerase (NEP) in the iojap maize mutant. *Curr. Genet.* 33: 340–344.

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y (2002). The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 99:

13627–13632.

Sinha S, Blanchette M, Tompa M (2004). PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 5: 170.

Smale T, Kadonaga T (2003). The RNA polymerase II core promoter. *Annual review of biochemistry* 72: 449-479.

Sosinsky A, Honig B, Mann RS, Califano A (2007). Discovering transcriptional regulatory regions in *Drosophila* by a nonalignment method for phylogenetic footprinting. *Proc. Natl. Acad. Sci.* 104: 6305-6310.

Sriraman P, Silhavy D, Maliga P (1998). The phage-type PclpP-53 plastid promoter comprises sequences downstream of the transcription initiation site. *Nucleic Acids Res.* 26: 4874–4879.

Staden R (1989). Methods for calculating the probabilities of finding patterns in sequences. *Comput. Appl. Biosci.* 5: 89-96.

Stern DL (2000). Evolutionary developmental biology and the problem of variation. *Evolution* 54: 1079-1091.

Stormo GD (1988). Computer methods to identify recognition sequences. *Ann. Rev. Biophys. Biophys. Chem.* 17: 241-263.

Stormo GD (1990). Consensus patterns in DNA. *Methods Enzymol.* 183: 211-221.

Stormo GD (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16(1): 16-23.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucl. Acids Res.* 14: 6661-6679.

Stubbe W (1955). Erbliche Chlorophylldefekte bei *Oenothera*. *Photo. Wiss.* 4: 3-8.

Stubbe W (1959). Genetische Analyse des Zusammenwirkens von Genom und Plastom bei *Oenothera*. *Z. Vererbungsl.* 90: 288-298.

Sucena E, Stern DL (2000). Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of

ovo/shaven-baby. Proc. Natl. Acad. Sci. USA 97: 4530-4534.

Sugita M, Sugiura M (1996). Regulation of gene expression in chloroplasts of higher plants. Plant Mol. Biol. 32: 315–326.

Swiatecka-Hagenbruch M, Liere K, Boerner T (2007). High diversity of plastidial promoters in *Arabidopsis thaliana*. Mol. Genet. Genomics 277: 725–734.

Tanaka T, Koyanagi KO, Itoh T (2009). Highly diversified molecular evolution of downstream transcription start sites in rice and *Arabidopsis*. Plant Physiol. 149:1316-1324.

The ENCODE project consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

The Toxicogenomics Research Consortium (2005). Standardizing global gene expression analysis between laboratories and across platforms. Nat. Methods 2: 351-356.

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, Moreau Y (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17: 1113-1122.

Williams TM, Selegue JE, Werner T, Gompel N, Kopp A, Carroll SB (2008). The regulation and evolution of a genetic switch controlling sexually dimorphic traits in drosophila. Cell 134: 610-623.

Thomas MC, Chiang CM (2006). The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology 41: 105-178

Thomas-Chollier M, Sand O, Turatsinze J, Janky R, Defrance M, Vervisch E, Brohée S, van Helden J (2008). RSAT: regulatory sequence analysis tools. Nucleic Acids Res. 36: W119-127.

Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavese G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotech. 23: 137-144.

Trémousaygue D, Garnier L, Bardet C, Dabos P, Hervé C, Lescure B (2003). Internal

telomeric repeats and 'TCP domain' protein-binding sites co-operate to regulate gene expression in *Arabidopsis thaliana* cycling cells. *Plant J.* 33: 957-66.

Tremousaygue D, Manevski A, Bardet C, Lescure N, Lescure B (1999). Plant interstitial telomere motifs participate in the control of gene expression in root meristems. *Plant J.* 20: 553-61.

Tuskan et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313: 1596-1604.

Vazquez F (2006). *Arabidopsis* endogenous small RNAs: highways and byways. *Trends Plant Sci.* 11: 460-8.

Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B (2006). A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* 34: 95-97.

Vokes SA, Ji H, McCuine S, Tenzen T, Giles S, Zhong S, Longabaugh WJ, Davidson EH, Wong WH, McMahon AP (2007). Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning. *Development* 134: 1977-1989.

Wang T, Stormo GD (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19: 2369-2380.

Wang X, Haberer G, Mayer KF (2009). Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. *BMC genomics* 10:284

Wasserman WW, Fickett JW (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J.Mol.Biol.* 278: 167-181.

Wilkins AS (2002). The evolution of developmental pathways. Sinauer Associates, Sunderland

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüss M, Reuter I, Schacherer F (2000). TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* 28: 316-319.

Wittkopp PJ (2006). Evolution of cis-regulatory sequence and function in Diptera. *Heredity* 97: 139-147.

Wray GA (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8: 206-216.

Xie D, Cai J, Chia NY, Huck HN, Zhong S (2008). Cross-species de novo identification of cis-regulatory modules with GibbsModule: Application to gene regulation in embryonic stem cells. *Genome Res.* 18: 1325-1335.

Xiong Y, Liu T, Tian C, Sun S, Li J, Chen M (2005). Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. *Plant Mol. Biol.* 59: 191-203.

Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J (2007). Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucl. Acids Res.* 35(18): 6219-26.

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.

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J (2009). Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *Plant J.* 60:350-362.

Yao JL, Cohen D (2000). Multiple gene control of plastome-genome incompatibility and plastid DNA inheritance in interspecific hybrids of *Zantedeschia*. *Theor. Appl. Genet.* 101: 400-406.

Yao JL, Cohen D, Rowland RE (1994). Plastid DNA inheritance and plastome-genome in-compatibility in interspecific hybrids of *Zantedeschia* (Araceae). *Theor. Appl. Genet.* 88: 255-260.

Yao JL, Cohen D, Rowland RE (1995). Interspecific albino and variegated hybrids in the genus *Zantedeschia*. *Plant Sci.* 109: 199-206.

Zhang W, Ruan J, Ho TD, You Y, Yu T, Quatrano RS (2005). Cis-regulatory element based targeted gene finding: genome-wide identification of ABA- and abiotic stress-responsive genes in *Arabidopsis thaliana*. *Bioinformatics* 21(14): 3074-81.

Zhou J, Wang X, Jiao Y, Qin Y, Liu X, He K, Chen C, Ma L, Wang J, Xiong L, Zhang Q, Fan L, Deng XW (2007). Global genome expression analysis of rice in response to drought and high-salinity stresses in shoot, flag leaf, and panicle. *Plant Mol. Biol.* 63: 591-608.

Zhou Q, Wong WH (2007). Coupling hidden markov models for the discovery of cis-regulatory modules in multiple species. *Ann. Appl. Stat.* 1: 36-65.

7. Appendixes

Appendix 1: Motif sites detected by PhyloCon in rice derived from MPSS.

Table shows motif sites detected by PhyloCon in rice genes derived from MPSS. The position of sites shown in the third column indicates the distance to the start codon.

Appendix 2: Motif sites detected by PhyloCon in rice derived from YALE-1.

Table shows motif sites detected by PhyloCon in rice genes derived from YALE-1. The position of sites shown in the third column indicates the distance to the start codon.

Appendix 3: Motif sites detected by PhyloCon in rice derived from YALE-2.

Table shows motif sites detected by PhyloCon in rice genes derived from YALE-2. The position of sites shown in the third column indicates the distance to the start codon.

Appendix 4: Motif sites detected by PhyloCon in sorghum derived from MPSS.

Table shows motif sites detected by PhyloCon in sorghum genes derived from MPSS. The position of sites shown in the third column indicates the distance to the start codon.

Appendix 5: Motif sites detected by PhyloCon in sorghum derived from YALE-1.

Table shows motif sites detected by PhyloCon in sorghum genes derived from YALE-1. The position of sites shown in the third column indicates the distance to the start codon.

Appendix 6: Motif sites detected by PhyloCon in sorghum derived from YALE-2.

Table shows motif sites detected by PhyloCon in sorghum genes derived from YALE-2. The position of sites shown in the third column indicates the distance to the start codon.

Appendix 7: Motif detected by network-level conservation in rice and sorghum

Table shows motif detected by network-level conservation in rice and sorghum genes. They are sorted descending according to their z-scores.

Appendix 8: Dyadic motifs detected by network-level conservation

Long specific motifs are shown that emerged from dyad motifs with initially unspecified spacer sequences (denoted as N). Note that many motifs (marked in red) have a low occurrence rate in rice and sorghum; however, most or all occurrences are conserved between orthologous pairs.

Appendix 9: Motifs of *A. thaliana* detected by FIRE

Table lists FIRE detected motifs of *A. thaliana* with their respective z-scores. Overrepresentation of position bias and match to known sites of AGRIS and PLACE databases are signed as “Y” in the corresponding columns of the table.

Appendix 10: Mapping FIRE detected motifs to motifs discovered by network-level conservation

Table lists motifs in *A. thaliana* detected by FIRE (left panel) and their corresponding motifs in *A. thaliana* and *A. lyrata* discovered base on network-level conservation principle (right panel).

8. Publications

Part of the work presented in this thesis has been published in peer reviewed scientific journals:

- Greiner S*, Wang X*, Rauwolf U, Silber MV, Mayer K, Meurer J, Haberer G, Herrmann RG (2008). **The complete nucleotide sequences of the five genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: I. Sequence evaluation and plastome evolution.** *Nucleic Acids Res.* 36: 2366–2378. (* Joint first authors)

- Greiner S*, Wang X*, Herrmann RG, Rauwolf U, Mayer K, Haberer G, Meurer J (2008). **The complete nucleotide sequences of the 5 genetically distinct plastid genomes of *Oenothera*, subsection *Oenothera*: II. A microevolutionary view using bioinformatics and formal genetic data.** *Mol. Biol. Evol.* 25(9): 2019-2030. (* Joint first authors)

- Wang X*, Haberer G*, Mayer KF (2009). **Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation.** *BMC genomics* 10:284 (* Joint first authors)

- Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottillar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL (2011). **The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change.** *Nat. Genet.* 43(5):476-81.

9. Acknowledgments

Though only my name appears on the cover of this dissertation, a great many people have contributed to its production. I owe my gratitude to all those people who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever. First of all, I would like to thank Prof. Hans-Werner Mewes for his support and guidance. My deepest gratitude is to my advisor, Dr. Klaus Mayer. I have been amazingly fortunate to have an advisor who gave me the freedom to do research on my own and at the same time the guidance to recover when my steps faltered. His insightful comments and constructive criticisms at different stages of my research were thought-provoking which helped me to develop my ideas. I am grateful to him for holding me to a high research standard and enforcing strict validations for each research result, and thus teaching me how to do research. My co-advisor, Dr. Georg Haberer, has been always there to listen and give advice. Georg taught me how to question thoughts and express ideas. I am deeply grateful to him for the long discussions that helped me sort out the technical details of my work. I am also thankful to him for encouraging the use of correct grammar and consistent notation in my writings and for carefully reading and commenting on countless revisions of this manuscript. His patience and support helped me overcome many crisis situations and finish this dissertation. I am also indebted to cooperation together with the research partners from department biology I Botany of University Munich. Prof. Dr. Reinhold Herrmann, Dr. Jörg Meurer, Dr. Stephan Greiner and their colleagues commented on my views and helped me to understand and to enrich my ideas for the work of the *Oenothera* project. I would also like to appreciate meaningful support from all members of the plant group. Their valuable discussions have improved my knowledge in various related research areas. Most importantly, none of this would have been possible without the love and patience of my family. I warmly appreciate the constantly moral support from my parents. My wife, to whom this dissertation is dedicated to, has been a constant source of love, concern, support and strength for all these years. I would like to express my heart-felt gratitude to my wife without whose generosity and understanding the dissertation could never be possibly accomplished.