Promoter Analysis of MADS-Box Genes in Eudicots Through Phylogenetic Footprinting

Stefanie De Bodt, Guenter Theissen,† and Yves Van de Peer**

*Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology,
Ghent University, Ghent, Belgium; and †Department of Genetics, Friedrich-Schiller-University, Jena, Germany

The MIKC MADS-box gene family has been shaped by extensive gene duplications giving rise to subfamilies of genes with distinct functions and expression patterns. However, within these subfamilies the functional assignment is not that clear-cut, and considerable functional redundancy exists. One way to investigate the diversity in regulation present in these subfamilies is promoter sequence analysis. With the advent of genome sequencing projects, we are now able to exert a comparative analysis of *Arabidopsis* and poplar promoters of MADS-box genes belonging to the same subfamily. Based on the principle of phylogenetic footprinting, sequences conserved between the promoters of homologous genes are thought to be functional. Here, we have investigated the evolution of MADS-box genes at the promoter level and show that many genes have diverged in their regulatory sequences after duplication and/or speciation. Furthermore, using phylogenetic footprinting, a distinction can be made between redundancy, neo/nonfunctionalization, and subfunctionalization.

Introduction

Promoters integrate information about the status of the cell in which they reside and alter the rate of transcriptional initiation of a gene accordingly. They function in the form of noncoding DNA near genes and transcription factors binding to that DNA in a sequence-specific manner. The identification of these transcription factor–binding sites (TFBS) is far from straightforward (Qiu 2003; Rombauts et al. 2003; Pavesi, Mauri, and Pesole 2004; Wasserman and Sandelin 2004; Tompa et al. 2005). As a consequence, not much is known about the average number of binding sites present in a typical promoter, the distance over which they are spread, or the number and position of TFBS and the nature of the gene product or the mode of expression. Based on relatively few well-characterized promoters such as the one of the *ENDO16* gene, which encodes a large extracellular protein playing a role in cell adhesion, it is estimated that 10–50 binding sites for 5–15 different transcription factors are present in an average promoter, comprising 10%–20% of the regulatory region (Wray et al. 2003). TFBS have an average size of 5–12 bp, corresponding to a protected footprint of 10–20 bp (Wray et al. 2003; Zhang and Gerstein 2003). Spacing between TFBS varies enormously, from partial overlap to distances of tens of kilobases, and spacing can greatly influence gene expression (Buchanan et al. 1997). Relative positioning and orientation of TFBS are determined by the steric hindrance on the one hand and the cooperative binding of transcription factors on the other hand, as well as by DNA bending. The combined action of different transcription factors enables a diverse range of expression patterns and different regulatory modules, defined as clusters of binding sites, produce discrete aspects of the total transcription profile. On average, such modules contain 6–15 binding sites that bind four to eight different transcription factors (Fickett and Wasserman 2000; Wray et al. 2003). While most promoters contain two or more clearly distinct modules, some apparently lack modular organization altogether (Wray et al. 2003). In conclusion, as far as we know, no strict rules concerning the size, positioning, and number of regulatory elements and modules can be applied for their detection.

In addition, little is known about the evolution of TFBS residing in promoters. In general, TFBS are thought to be quite variable and most binding sites tolerate at least one, but often more, specific nucleotide substitutions without losing functionality (Latchman 1998; Courey 2001). According to Collado-Vides, Magasanik, and Gralla (1991), in *Escherichia coli*, different sites for the same transcription factor differ by about 20%–30% of the bases relevant for binding. Moreover, Stone and Wray (2001) predict that new binding sites arise and become fixed within populations via local point mutations on microevolutionary timescales. Widespread turnover of TFBS and heterogeneity in substitution rates across binding sites have been observed when comparing human-primate and human-rodent promoters (Dermitzakis and Clark 2002). In addition, the constraints on the binding sites seem to be generally dependent on the phylogenetic lineage (e.g., mammalian or rodent) and not so much on the importance of the site for the expression of the gene (Dermitzakis and Clark 2002). Therefore, in practice, it is difficult to estimate the degree of conservation between regulatory elements and to choose an adequate number of allowed nucleotide substitutions for the detection of TFBS. Moreover, often inversions cause rearrangements of regulatory elements (Chuzhanova et al. 2000). These rearrangements, at least in vertebrate promoters, increase in number with increasing evolutionary distance (Chuzhanova et al. 2000). Consequently, readily aligning promoter sequences with global alignment procedures to detect a substantial portion of the conserved regulatory elements will often fail. Although it is largely unknown how often these rearrangements occur in plants, motif detection methods need to be able to cope with such rearrangements.

In phylogenetic footprinting, regions that are conserved between orthologous regulatory sequences are thought to be of functional importance (Wasserman et al. 2000; Bulyk 2003; Weitzman 2003; Zhang and Gerstein 2003). Possible TFBS can be regarded as islands of conservation

Key words: phylogenetic footprinting, MADS-box genes, eudicots, evolution, promoter.

E-mail: yves.vandepeer@psb.ugent.be.

Mol. Biol. Evol. 23(6):1293–1303. 2006
doi:10.1093/molbev/msk016
Advance Access publication March 31, 2006

surrounded by nonconserved sequences. Therefore, by comparing the promoter regions of genes across related taxa, TFBS that confer conserved expression patterns can be identified (Wasserman et al. 2000). For identifying TFBS by phylogenetic footprinting, it is important that the genes compared have not diverged too much so that conserved elements can still be recognized. On the other hand, divergence between genes has to be large enough to assure that nonfunctional regions have accumulated enough mutations so that they can be distinguished from functional, conserved regions. Until now, the application of phylogenetic footprinting mainly focused on promoters of animals and yeasts using conventional sequence alignment programs (Wasserman et al. 2000; Bulyk 2003; Weitzman 2003; Zhang and Gerstein 2003), on the identification of conserved noncoding sequences (CNS) in limited sets of grass promoters (Guo and Moose 2003; Inada et al. 2003), and on the detection of regulatory regions of a few genes from closely related eudicot plant species belonging to the *Brassicaceae* (Koch et al. 2001; Hong et al. 2003) or the *Cucurbitaceae* (Ayre, Blair, and Turgeon 2003). For instance, in the study of Ayre, Blair, and Turgeon (2003), local alignments of at least 8 nt with 90% similarity were identified in the *GAS1* (*GALACTINOL SYNTHASE 1*) promoter regions of several *Cucurbitaceae* species using the Gibbs sampler algorithm of the MACAW software package (Schuler, Altschul, and Lipman 1991; Lawrence et al. 1993). Nevertheless, it remains largely unknown if the phylogenetic footprinting technique is widely applicable to the detection of novel binding sites in promoters of relatively distantly related eudicot plant species, such as *Arabidopsis thaliana* and *Populus trichocarpa*, for which the complete genome sequences are now available. Furthermore, it also remains to be investigated how the large number of paralogs in these genomes affects the performance of phylogenetic footprinting techniques.

Indeed, gene and genome duplications have been shown to be particularly prominent in plant genomes and have greatly influenced their organization and evolution (Otto and Whitton 2000; Wendel 2000; Simillion et al. 2002; De Bodt et al. 2005; Maere et al. 2005). These duplication events also have profound effects on gene function and regulation. Gene duplication, as an important driving force for generating evolutionary novelty, enables genes to diverge in function (or expression) through neofunctionalization, where the gene acquires a new function (or expression pattern) or subfunctionalization, a process in which functions (or expression patterns) are partitioned between duplicate genes (Ohno 1970; Prince and Pickett 2002). Thus, if one wants to study the effects of gene duplication and their influence on gene regulation, it is necessary to identify the correct orthologous and paralogous relationships (Theissen 2002) and take these into account when detecting regulatory elements. Differences in gene regulation are believed to be a major source of diversity in higher eukaryotes. More in particular, the divergence of promoter regions of transcriptional regulators is suggested to be of major importance in the evolution of structural complexity of extant flowering plants (Doebley and Lukens 1998; Tautz 2000; Kellogg 2004). Alteration in the expression profiles or in the binding properties of developmental transcription factors can yield particularly interesting phenotypes in organ structures (e.g., the increase in carpelloid structures), homeotic transformations (causing a body part to develop in an inappropriate position in an organism), or novel morphologies. Therefore, it is interesting to exert a comparative analysis of the promoters of MADS-box genes, which are key players in the regulation of plant development, and to investigate the possible occurrence of functional divergence according to the subfunctionalization (or duplication-degeneration-complementation) model, which tries to explain the evolution of duplicated genes on the one hand and developmental and morphological diversity on the other (Force et al. 1999).

The MADS-box gene family has been the subject of extensive studies that try to unravel the structural complexity of extant flowering plants. The majority of MIKC-type MADS-box genes is involved in the determination of flowering time, floral meristem, and floral organ identity. According to the ABC model, A class genes (*APETALA1* and *APETALA2*) specify sepals in the first whorl, A and B class genes (*APETALA3* and *PISTILLATA*) petals in the second whorl, B and C class genes (*AGAMOUS*) stamens (male reproductive organs) in the third whorl, and C class genes specify carpels (female reproductive organs) in the fourth whorl of a typical *Arabidopsis* flower (see fig. 1) (Coen and Meyerowitz 1991; Weigel and Meyerowitz 1994). In addition to ABC class genes, D class genes (*SEEDSTICK*), important for ovule development, and E class genes (*SEPALLATA1*, *SEPALLATA2*, *SEPALLATA3*, *SEPALLATA4*), which encode additional transcription factors playing a role in determining the identity of all four whorls, have been described which led to an extension of the ABC model (Theissen 2001). Moreover, evidence is growing on the formation of multimers, possibly tetramers according to the quartet model, consisting of proteins encoded by these different MADS-box gene classes and regulating different aspects of floral organ development (Honma and Goto 2001; Theissen and Saedler 2001; de Folter et al. 2005; Kaufmann, Melzer, and Theissen 2005). Several other MADS-box genes function upstream (flowering time genes: *FLOWERING LOCUS C*, *SHORT VEGETATIVE PHASE*, *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1*; meristem identity genes: *APETALA1*, *CAULIFLOWER*, *FRUITFUL*; and intermediate genes: *AGAMOUS*) or downstream (*FRUITFUL*, *SHATTERPROOF1*, *SHATTERPROOF2*, *AGL13*) of these floral organ identity genes (see fig. 1) or regulate different aspects of root, leaf, and embryo development (*ANR1*, *AGL12*, *AGL17*, *AGL15*) (Ng and Yanofsky 2001).

In many cases, (partial) functional redundancy between paralogous MADS-box genes (e.g., *AP1-CAL*, *SHP1-SHP2*, and *SEP1-SEP2-SEP3-SEP4*) has been described (Ferrandiz et al. 2000; Pinyopich et al. 2003; Ditta et al. 2004; Malcomber and Kellogg 2005). Nevertheless, changes in expression patterns after duplication of MADS-box genes have occurred as well. For example, the *Petunia PhDEF* gene (belonging to the euAP3 lineage) is expressed in petals and stamens, like its *Arabidopsis* homolog, while the *Petunia PhTM6* gene (belonging to the TM6 lineage) is expressed in stamens and carpels, like the tomato *LeTM6*

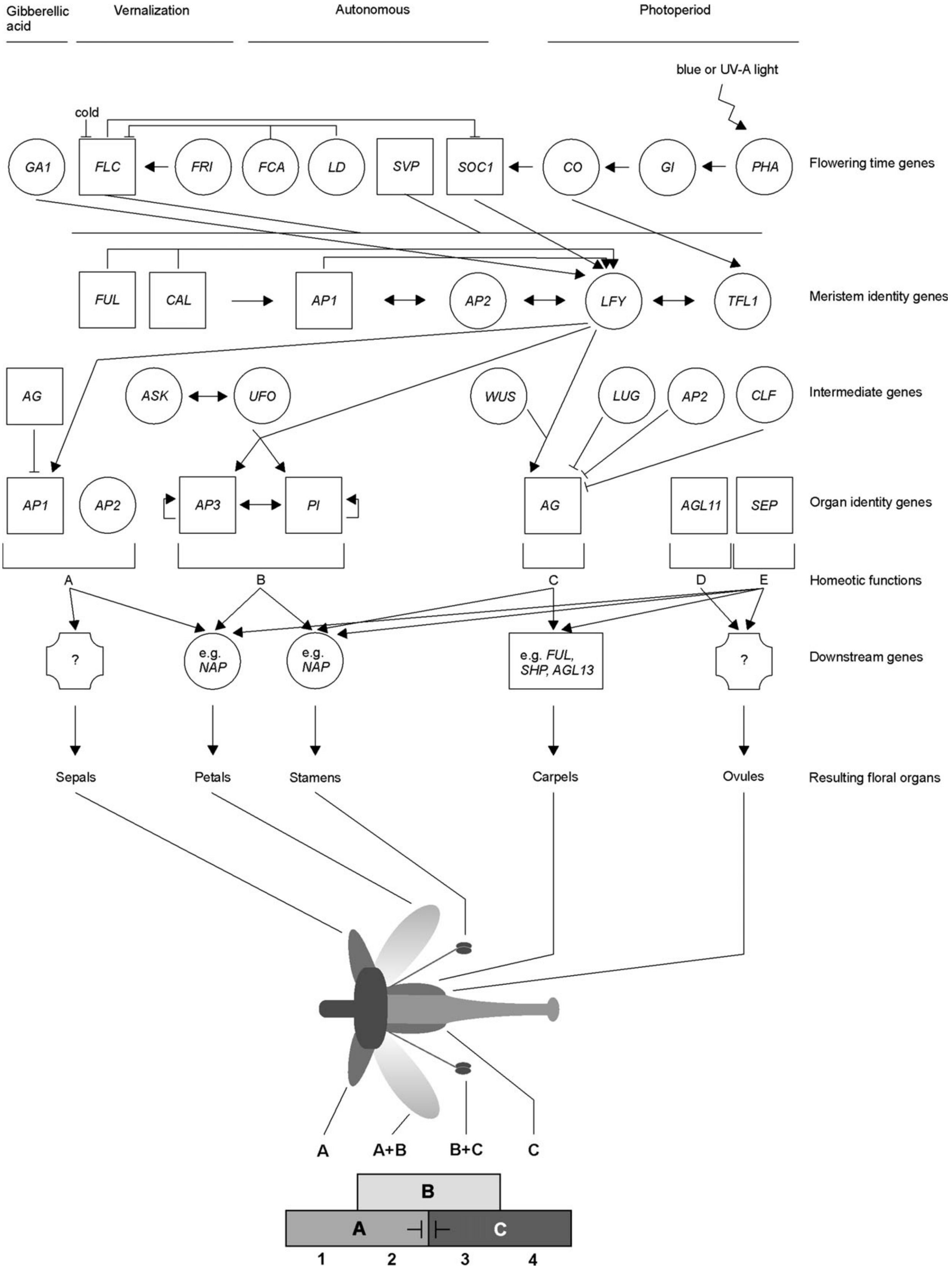

FIG. 1.—A simplified depiction of the genetic hierarchy that controls flower development in the eudicot model plant *Arabidopsis thaliana* (modified from a figure in Soltis et al. [2002]). Some regulatory interactions between the genes are symbolized by arrows (activation), double arrows (synergistic interaction), or barred lines (inhibition, antagonistic interaction). MADS-box genes are depicted by squares.

gene, hinting to divergence of the promoters of these MADS-box genes after the duplication leading to the euAP3 and TM6 lineage (Vandenbussche et al. 2004).

In addition to this kind of gene expression analyses, it is now possible to compare the promoter sequences of MADS-box genes from different plants because an increasing number of plant genomes are being sequenced and to study the divergence of MADS-box genes after duplication and speciation at the regulatory level. Here, we investigate the promoters of all duplicated MIKC-type MADS-box genes by phylogenetic footprinting of homologous (both orthologous and paralogous) *Arabidopsis* and poplar genes.

Materials and Methods

Identification of *Arabidopsis* and Poplar MADS-Box Homologs

Based on the complete set of annotated *Arabidopsis* MIKC-type MADS-box genes (De Bodt et al. 2003; Parenicova et al. 2003), we searched for homologous genes in the genome of *P. trichocarpa* (Department of Energy [DoE] Joint Genome Institute and Poplar Genome Consortium). The Eugene prediction (Schiex, Moisan, and Rouzé 2001) of the poplar genome was used; however, some gene annotations needed manual correction. Moreover, when scanning the raw genomic sequence for MADS-box genes, additional genes could be identified, after which they were manually annotated. Sequences can be found in the Supplementary Material online. To define the correct orthologous or co-orthologous relationships between the *Arabidopsis* and poplar MADS-box genes, phylogenetic trees were built based on the alignment of all MADS-domain proteins as well as on (larger) alignments of individual gene subfamilies to verify the correct *Arabidopsis*-poplar relationships (see fig. 2 and Supplementary Material online). Some differences can be observed between the complete gene family tree and the subfamily trees for nodes that are weakly supported in the family tree. The *Arabidopsis*-poplar gene relationships were always deduced from the subfamily trees. These trees are thought to be more reliable as they are based on more phylogenetic informative positions which are not shared by all superfamily members. The studied groups are indicated on figure 2; one-to-one relationships are marked in black, two-to-one relationships are marked in light gray, and one-to-many or many-to-many relationships in dark gray. Only gene duplicates arisen after the divergence of *Arabidopsis* and poplar (based on the phylogenetic trees) are investigated. Phylogenetic trees were constructed based on Poisson distances and calculating 500 bootstrap samples using TREECON (Van de Peer and De Wachter 1994). Only bootstrap values above 50% are shown in the phylogenetic tree in figure 2; nodes with lower bootstrap values are

⟶

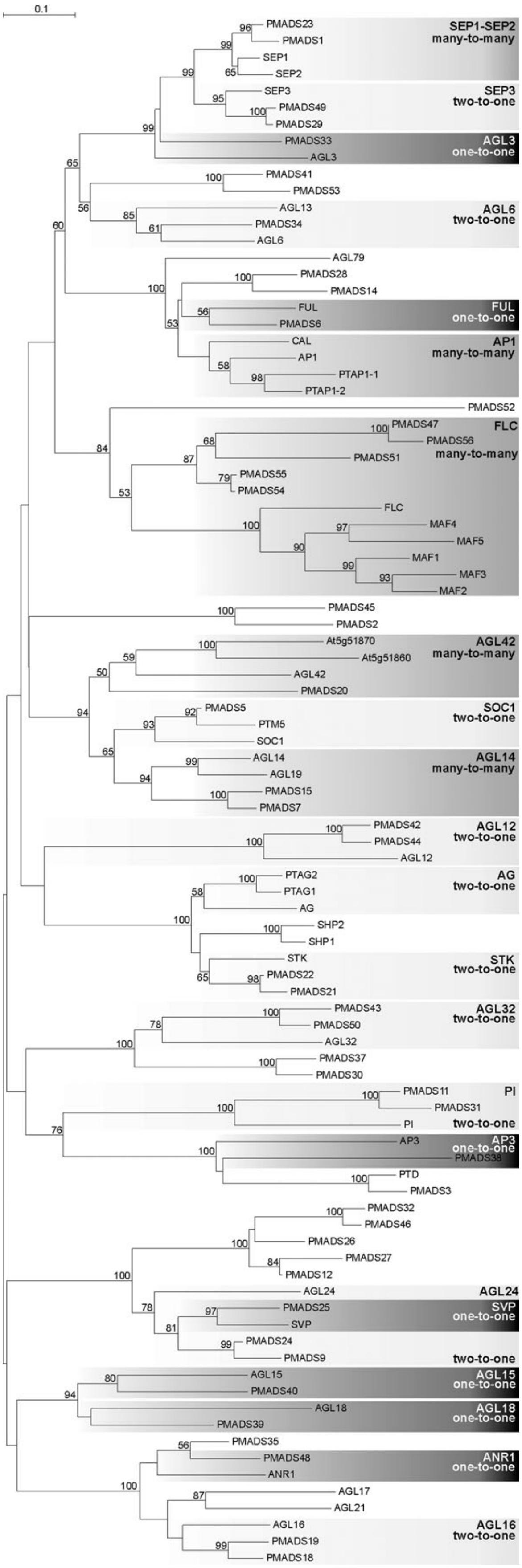

FIG. 2.—Phylogenetic tree of *Arabidopsis* and poplar MADS-box genes. One-to-one relationships indicated in black, two-to-one in light gray, and one-to-many and many-to-many in dark gray (see text for details and *Materials and Methods* for more information on the construction of the phylogenetic tree). Bootstrap values below 50% are not shown.

thought to be poorly supported and should be interpreted with caution.

Promoter Comparison of *Arabidopsis* and Poplar MADS-Box Genes

For every *Arabidopsis* and poplar MADS-box gene, 1 kb upstream of the translation start site was considered as promoter sequence, unless the intergenic region was smaller than 1 kb, which was only the case for the *ANR1* promoter. Although we are aware that evidence is growing on the presence of TFBS in introns, we here chose to limit our analysis to the 1 kb upstream region of every gene. Ideally, the upstream regions should start from the transcription initiation site rather than the translation start site. However, certainly in the case of poplar genes, these transcription initiation sites are difficult to identify, partly due to lack of full-length cDNA sequences of these MADS-box genes. Per subfamily (as indicated in fig. 2), promoters were analyzed by FootPrinter (Blanchette and Tompa 2003). This tool takes into account the evolutionary relationships and distances between the genes compared (based on a phylogenetic tree). Consequently, one can not only choose the number of mutations that are allowed between the conserved motifs but can also indicate the number of mutations that are allowed to occur in function of the evolutionary distance (see later). To choose adequate motif sizes, for which the number of resulting false positive motifs would be minimal, we counted the number of motifs detected in randomized data sets. Randomized data sets were made through shuffling of the homologous promoter sequences. Results can be found in the Supplementary Material online (Table S1). From these analyses, we can conclude that, in general, motifs of short length (up to 12 bp) cannot be detected in a statistically significant way when comparing only two sequences. When using restrictions based on these randomizations, a large fraction of biologically relevant binding sites is omitted from the final results because, for instance, motifs of sizes smaller than 12 have often been reported in the literature (e.g., Hong et al. 2003; PlantCare [Lescot et al. 2002]; Place [Higo et al. 1999]). Therefore, we chose motif sizes that were slightly smaller than is suggested by the randomization studies. However, as a result, a higher number of false positives are detected and postprocessing of the motifs is necessary. This postprocessing involves the calculation of the occurrence of a certain detected motif in all *Arabidopsis* promoters using MatInspector (Quandt et al. 1995). All *Arabidopsis* promoters were scanned with the detected motifs, and identical matches between the motif and the *Arabidopsis* promoter were counted to determine the occurrence of that motif, reflecting the probability of detecting the particular motif by chance. The occurrence of the motifs in the whole *Arabidopsis* promoterome is shown in the Supplementary Material online. An additional postprocessing step was made for FootPrinter analyses where motif losses were allowed. In these cases, we selected for motifs common between different species rather than motifs that are species-specific because the latter motifs are likely to represent noise (cases in which species-specific motifs are omitted in the figures are indicated in the Supplementary Material online). The number of mutations that are allowed per motif and per branch was kept constant, namely two mutations per motif and one mutation per branch. When allowing motif losses (to detect motifs specific for certain lineages and absent in others), more detailed mutation parameters, describing the number of mutations allowed over a certain evolutionary distance, need to be given. As these evolutionary distances can differ considerably between subfamilies, we have adjusted these parameters for every subfamily separately. Finally, a subregion size can be chosen which restricts the detection of motifs in corresponding subsequences of the homologous promoters. This parameter was set to 50. However, increasing this value did not drastically affect the results, although in some cases, keeping this parameter small can lower the amount of false positive motifs (data not shown). For all other parameters, default values were used.

Scanning Public Databases of TFBS

In order to compare known TFBS with those detected by our phylogenetic footprinting approach, PlantCare (Lescot et al. 2002), Place (Higo et al. 1999), and TRANSFAC (Wingender et al. 2000) databases were scanned. To account for the variability of TFBS, all motifs were transformed into position-specific frequency matrices (PFM). MotifComparison from the INCLUSive package (Coessens et al. 2003) was used to compare the PFM. No shifts between matrices are allowed, so that the shortest motif overlaps completely with the longest motif. Only motifs having a score higher than 0.4 and having a size greater than 5 are considered.

Results

The identification of TFBS through phylogenetic footprinting has been mainly restricted to the direct comparison of orthologous promoter sequences. However, because both the genomes of *A. thaliana* and *P. trichocarpa* have undergone at least two and likely three genome duplication events in their evolutionary past (Simillion et al. 2002; Bowers et al. 2003; Maere et al. 2005; Sterck et al. 2005), only in rare cases a one-to-one orthologous relationship is expected between both species. Indeed, when inspecting the complete phylogeny of *Arabidopsis* and poplar MADS-box genes, only few one-to-one relationships exist (indicated in black on fig. 2). In other cases, one *Arabidopsis* gene is a homolog to two "co-orthologs," or "inparalogs" of poplar (one-to-two, indicated in light gray), or vice versa (two-to-one, also indicated in light gray), or *Arabidopsis* and poplar homologs show a one-to-many or many-to-many relationship (indicated in dark gray) (see fig. 2). To allow the detection of common as well as duplicate- and species-specific regulatory elements in the promoters of *Arabidopsis* and poplar MADS-box genes, we have applied the phylogenetic footprinting program FootPrinter (Blanchette, Schwikowski, and Tompa 2002; Blanchette and Tompa 2003). When looking for shared TFBS, this program takes into account the evolutionary relationships and distances of the genes of which the promoters are analyzed (see *Materials and Methods*). Moreover, motif losses and inversions can be allowed, which is, for instance, not the case in the detection of conserved CNS through alignment methods like AVID/VISTA (Guo and Moose 2003).

Promoter Divergence After Duplication and Speciation

Genes with One-to-One Relationships in Arabidopsis *and Poplar*

Conserved elements between one-to-one *Arabidopsis*-poplar orthologs (i.e., seven pairs of genes out of the 22 studied subfamilies; see fig. 2 and fig. S1, see Supplementary Material online) were identified. The number of motifs uncovered between two genes can differ greatly between subfamilies, reflecting the difference in the overall degree of conservation (Table S2, see Supplementary Material online). For instance, the promoters of the *FRUITFUL* (*FUL*), *SHORT VEGETATIVE PHASE* (*SVP*), and *APETALA3* (*AP3*) homologs are clearly more conserved than all other promoters showing a one-to-one relationship, as 17, 14, and 10 motifs of size 13 can be uncovered, respectively, while on average, only six conserved motifs can be identified in the other promoters (excluding ANR1 as its promoter sequence is shorter than 1 kb; see Supplementary Material online). When inspecting the organization of the detected motifs, inversions and changes in spacing are often observed. Furthermore, we find most of the (longest) motifs close to the translation start site (see Supplementary Material online). This variation in spacing and high conservation close to the translation start site was also observed by Ayre, Blair, and Turgeon (2003) when studying the *GAS1* promoter regions of several *Cucurbitaceae* species (Ayre, Blair, and Turgeon 2003). Nevertheless, motifs can be identified over the whole length of the promoter sequence (1 kb in this study).

Genes with One-to-Two Relationships in Arabidopsis *and Poplar*

Quite often, one *Arabidopsis* gene is related to two co-orthologous poplar genes (nine out of the 22 studied subfamilies), while the opposite is rarer (one out of the 22 studied subfamilies) (see fig. 2 and fig. S2, see Supplementary Material online). Details on the construction of the phylogenetic tree of all *Arabidopsis* and poplar MADS-box genes shown in figure 2 can be found in the *Materials and Methods*. Again, large differences in the overall degree of conservation can be observed (Tables S3 and S4, see Supplementary Material online). For instance, in the *AGAMOUS* promoters, motifs of up to size 12 can be identified, while in the *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1* (*SOC1*) promoters, only motifs of size 9 can be detected. The duplicated poplar promoters of the *AG* (*AG/PTAG1-PTAG2*), *STK* (*STK/PMADS21-PMADS22*), and *AGL16* (*AGL16/PMADS38-PMADS39*) genes are highly conserved. All three promoters are equally well conserved in each subfamily. As a result, the homologous genes can be expected to show similar expression patterns, which has been experimentally confirmed for *PTAG1* and *PTAG2* (Brunner et al. 2000). In the case of *AG*, it should be remarked that the second intron of this gene has been shown to possess regulatory activity (Hong et al. 2003). Nevertheless, the sequence upstream of the translation start site that is analyzed here shows high conservation and thus possible functional activity, as well. In contrast to the high degree of promoter conservation observed for the genes discussed above, other promoter sequences are poorly conserved. For instance, when comparing the promoters of the *AGL6* homologs (*AGL6-AGL13/PMADS34*), no motifs are unveiled that are conserved between one of the two *Arabidopsis* promoters and the poplar promoter, while a considerable number of motifs can be detected between the *Arabidopsis AGL6* and the *AGL13* promoter. This seems to suggest that the *Arabidopsis* and poplar homologs have acquired distinct expression patterns after the divergence of the two species and thus could hint to either non- or neofunctionalization of the poplar homolog. However, promoter sequence or expression data from other plants will be necessary to distinguish between either two fates.

The promoters of the *AGL32* homologs (*AGL32/PMADS43-PMADS50*) in poplar also show overall low conservation. However, duplicate-specific motifs (motifs present in one gene duplicate but not in the other, e.g., present in the promoter of *PMADS43* but not in *PMADS50*) as well as species-specific motifs can be identified. In the case of the *AGL12* promoters (*AGL12/PMADS42-PMADS44*), even a higher number of duplicate-specific motifs are found, pointing to subfunctionalization of *PMADS42* and *PMADS44* (see fig. 3*a*). Partitioning of motifs has occurred between the two duplicate promoters throughout the 1-kb promoter sequence. As for the promoters of *SEP3* genes (*SEP3/PMADS29-PMADS49*; see Fig. S2, Supplementary Material online), particularly high conservation between the *PMADS49* and *SEP3* promoters in the region between 600 and 1000 bp upstream of the ATG is detected, as well as overall conservation between all three homologous promoters in the region close to the transcription start site (1–600 bp). A similar observation as the one made for the *SEP3* gene can be made when comparing the promoters of the *AGL24* homologs (see fig. 3*b*). Although the region close to the ATG is well conserved between all three promoters (*AGL24*, *PMADS9*, and *PMADS34*), the region between 600 and 900 bp is only well conserved between the promoters of *AGL24* and *PMADS24*. This observation possibly points to a distinct partitioning of the promoter's function, leading to a more restricted expression pattern for one of the duplicates, but which is clearly different from the promoter divergence described for the AGL12 and AGL32 subfamily. In addition, one of the two poplar promoters has remained very similar to the *Arabidopsis* promoter, while in the case of *AGL32* and *AGL12*, both poplar promoters have diverged considerably. An even more drastic divergence between duplicated poplar promoters can be observed for *SOC1* (*SOC1/PTM5-PMADS5*) and *PI* (*PI/PMADS11-PMADS31*) homologs because one of the two poplar promoters (*PMADS5* in the SOC1 subfamily and *PMADS11* in the PI subfamily) shares far more (duplicate-specific) motifs with the *Arabidopsis* ortholog than the other poplar promoters (*PTM5* in the SOC1 subfamily and *PMADS31* in the PI subfamily) (see Fig. S2, Supplementary Material online). Also, this example possibly indicates non- or neofunctionalization of one of the poplar promoters after duplication.

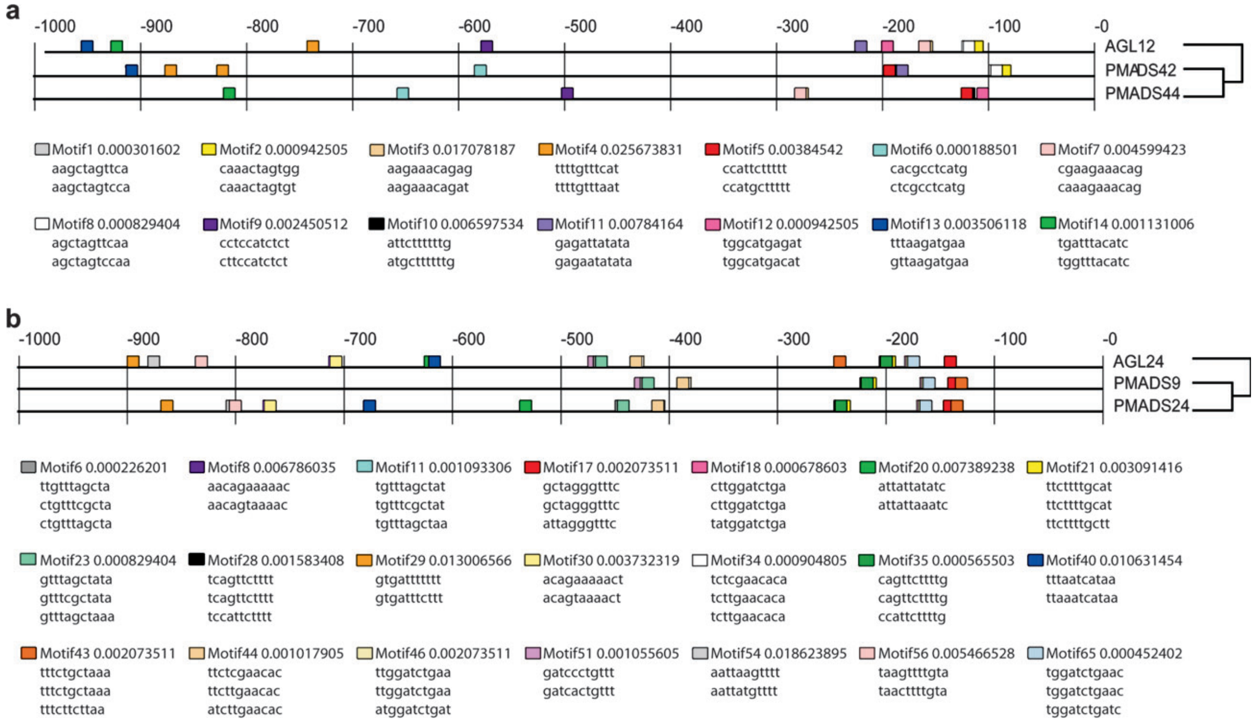

FIG. 3.—Visual representation of motifs detected by FOOTPRINTER in the promoters of *Arabidopsis* and poplar. (*a*) AGL12 MADS-box genes suggesting subfunctionalization after duplication (FOOTPRINTER parameters: motif size: 11, allowed mutations: 2, losses) and (*b*) AGL24 MADS-box genes suggesting neo- or nonfunctionalization of the poplar *PMADS9* promoter in the region between 600 and 1000 bp upstream of the ATG (FOOTPRINTER parameters: motif size: 10, allowed mutations: 2, losses). Next to the motif name, the occurrence of that motif in all *Arabidopsis* promoters is given.

Genes with One-to-Many or Many-to-Many Relationships in Arabidopsis *and Poplar*

When inspecting the results of subfamilies consisting of more than three members (4 out of the 22 studied subfamilies), a decrease in overall conservation is observed, reflecting the growing promoter divergence with an increasing number of duplicates (see fig. 2; Fig. S3 and Tables S5 and S6, Supplementary Material online). Consequently, the detection of motifs becomes more difficult as fewer motifs are shared between different promoters and shared motifs tend to be smaller. Often, only motifs of size up to 9 can be detected. However, when considering these specific subfamilies in more detail, some poplar promoters show more conservation with particular *Arabidopsis* promoters in distinct regions of the promoter, as was also sometimes observed in the two-to-one analyses (see higher). For instance, in the case of the SEP1 subfamily, all four promoters are conserved in the region from 100 to 600 bp upstream of the ATG, but *PMADS1* and *SEP1* are also conserved in the region of 600–1000 bp (see fig. 4), while for the other promoters, no motifs are identified in this region (see Fig. S3, Supplementary Material online). Thus, although *SEP1* and *SEP2* are described to possess highly overlapping functions (Pelaz et al. 2000), based on the motifs detected in their promoter sequences, they do possess some differences on regulatory level (see Fig. S3, Supplementary Material online).

The promoters of the homologs *AGL14* and *PMADS7*, on the one hand, and *AGL19* and *PMADS15*, on the other hand, are particularly well conserved (see Supplementary Material online), reflecting distinct expression patterns of both *Arabidopsis* and poplar paralogs, even though the duplicates arose and evolved independently. When comparing *AGL19* and *PMADS15*, a clear inversion of the region between 250 and 400 bp can be observed, resulting in three motifs in opposite order in the *AGL19* promoter compared to the *PMADS15* promoter (see subset 4 in Fig. S3, Supplementary Material online). The influence of this inversion on the expression of these genes remains to be investigated.

The FLC subfamily contains a high number of paralogs in both *Arabidopsis* and poplar. For simplicity, only the results for a limited set of homologs are shown in the Supplementary Material online. Other combinations and larger sets of homologs were also analyzed, but results are not discussed because they all show a high within-species conservation at the coding level (probably due to the fact that several homologs are recently duplicated genes), as well as partitioning of motifs, for example, between the *FLC* and the *MAF1* promoter, when compared to the promoter of *PMADS51* (see Fig. S3, Supplementary Material online).

Phylogenetic Footprinting and the Number of Compared Species and Their Evolutionary Distance

As described above, we were able to detect several putative TFBS in the promoters of *Arabidopsis* and poplar homologs through phylogenetic footprinting. However, in some cases, only very few or very short motifs could be identified. To investigate the effect of evolutionary distance on the uncovering of TFBS by FOOTPRINTER, we have compared sets of promoters from several eudicot species. In

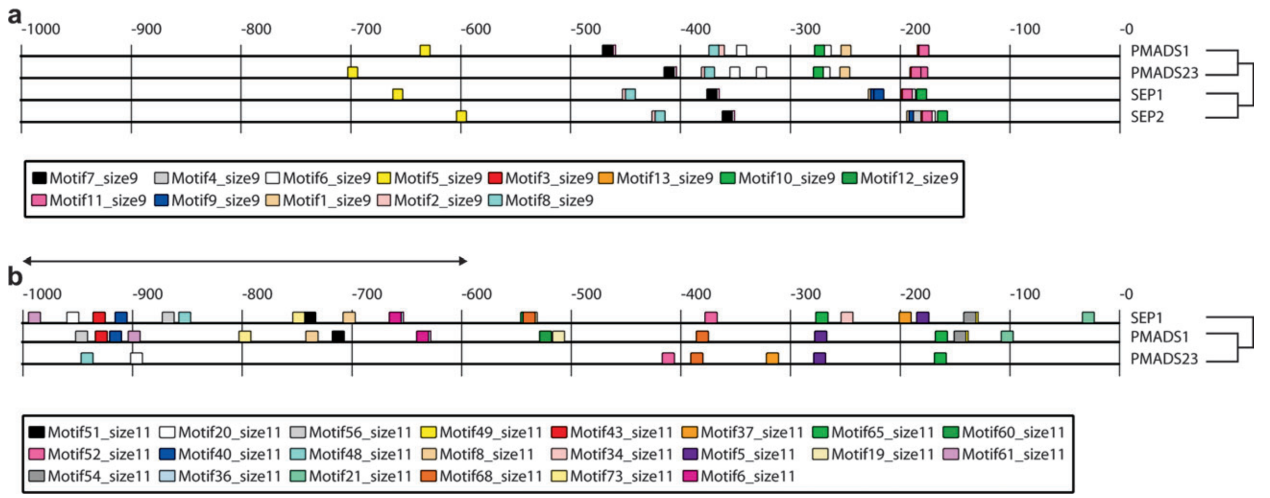

FIG. 4.—Visual representation of motifs detected by FOOTPRINTER in the promoters of *Arabidopsis* and poplar SEP1 MADS-box genes. The region of 600–1000 bp upstream of the ATG is only conserved between the promoters of *SEP1* and *PMADS1* (indicated by an arrow). (*a*) FOOTPRINTER parameters: motif size: 9, allowed mutations: 2 and (*b*) FOOTPRINTER parameters: motif size: 11, allowed mutations: 2, losses, comparison of three out of four promoters).

addition, we have also investigated the effect of increasing the number of species for phylogenetic footprinting. To this end, we studied the APETALA3 subfamily in greater detail. For this gene, several orthologs (and their promoters) in different species have been identified. The following genes were included in the AP3 data set: *A. thaliana AP3* (At3g54340), *Brassica oleracea AP3* (AF043610), *Solanum tuberosum* (potato) *DEF4* (X67511), *Petunia hybrida PMADS1* (AY370519), *Medicago truncatula NMH7* (AF042068), *P. trichocarpa PMADS38* (gw1.II.2311.1), and *Lotus corniculatus var. japonicus AP3* (AP006109). A further advantage of evaluating phylogenetic footprinting on the *AP3* homologs is that this promoter is dissected in great detail in *A. thaliana* (e.g., Hill et al. 1998). As such, the detected motifs can be compared with the already known regulatory elements as a validation of the results.

First, we compiled data sets consisting of three species. The performance of FOOTPRINTER was evaluated for different combinations of species according to the plant lineage they belong to, namely *Arabidopsis*-*Brassica*-*Fabaceae*/*Salicaceae*, *Arabidopsis*-*Brassica*-*Solanaceae*, *Arabidopsis*-*Fabaceae*/*Salicaceae*-*Solanaceae*, *Arabidopsis*-*Fabaceae*/*Salicaceae*, and *Arabidopsis*-*Solanaceae* (see Fig. S4, Supplementary Material online). When including a species such as *B. oleracea*, which is evolutionarily close to *Arabidopsis*, a large number of large motifs (size 11–12) are detected, compared to the other species combinations. Overall, we observe that including a *Solanaceae* species in the data set increases the number of motifs uncovered, and this is most clearly visible for the *Arabidopsis*-*Petunia*-potato comparison. Strikingly, the conservation between the *Arabidopsis AP3* and the *Fabaceae*/*Salicaceae* promoters is relatively low, even though *Fabaceae*/*Salicaceae* species are more closely related to *Brassicaceae species* than *Solanaceae* species. Most probably, this is due to the specific evolution of *AP3* expression in these species, rather than this being a general characteristic. We predict that the *AP3* homologs in *Fabaceae*/*Salicaceae* have a rather different expression pattern than *Brassicaceae* and *Solanaceae AP3* genes.

Second, we compared the promoter sequences of four species. In this case, it is even more obvious that the presence of a close relative (in this case *Brassica*) in the data set is highly profitable as almost no motifs (of size 9) are detected when *Brassica* is omitted (see Supplementary Material online).

Last, phylogenetic footprinting on an even larger group of species was studied. However, in this case, it is necessary that motif loss is allowed; no motifs of size 8 or higher can be detected in more than four promoters from the eudicot species mentioned above. Allowing motif losses, a clear distinction between *Arabidopsis*-*Solanaceae* and *Fabaceae*/*Salicaceae* promoters is again observed.

In conclusion, it should be noted that none of the additional species studied here possess a completely sequenced genome and thus it remains possible that for some species additional homologs exist, which could have introduced additional promoter divergence (see higher).

Known TFBS

The flower developmental pathway, as depicted in figure 1, should provide us with a reasonable amount of information to validate our results. Most flower MADS-box genes are regulated by MADS proteins and as such should possess a CArG-box in their promoter (Riechmann, Wang, and Meyerowitz 1996). In addition, some other transcription factors are known to regulate MADS-box genes as well. However, the link between the transcription factor and the binding site in the promoter of its target genes is often unknown as no strict consensus binding sites exist for a certain type of transcription factor. For instance, when inspecting public databases of known TFBS, such as PlantCare (Lescot et al. 2002), Place (Higo et al. 1999), or TRANSFAC (Wingender et al. 2000), numerous possible binding sites exist for a MYB transcription factor. An important regulator of MADS-box genes is LEAFY, which is thought to play a role in the transition from vegetative to reproductive phase. In a previous study by Hong et al.

(2003), the LEAFY binding site was identified in the second intron of AGAMOUS. LEAFY is thought to bind the consensus sequence CCANTG(T/G) (Parcy et al. 1998; Busch, Bomblies, and Weigel 1999). Adjacent to this site, the WUSCHEL binding site (TTAATGG) is detected in the second intron of *AGAMOUS*. In addition, CArG, CCAAT, and AAGAAT boxes are identified in the second intron of *AGAMOUS* (Hong et al. 2003). Another well-studied MADS promoter is that of the *AP3* gene. Computational analyses of Brassicaceae *AP3* promoters identified CArG, MYB, Dof, zinc finger, and G-boxes (Koch et al. 2001), and experimental analyses showed that discrete regions of the promoter regulate different (spatial and temporal) aspects of the expression profile (Hill et al. 1998). However, these binding sites are generally quite short, and their identification is thus not straightforward. Indeed, the number of false positives that is detected when looking for such short sequences will be high.

We scanned the public databases PlantCare (Lescot et al. 2002), Place (Higo et al. 1999), and TRANSFAC (Wingender et al. 2000) to identify possible known binding sites in our data set. To this end, we transformed our and publicly available binding sites into PFM and applied the tool MotifComparison from the INCLUSive package (Coessens et al. 2003) (see *Materials and Methods* for details). A table of putative TFBS can be found in Supplementary Material online (Table S7). CCAAT, AGAAA (pollen specific), and CACGTG (G-box) boxes can often be found in the promoter of MADS-box genes. However, these boxes cannot be easily linked to the regulatory hierarchy known to control flower development. LEAFY and WUSCHEL binding sites cannot be found in the MADS promoters if stringent parameters are used. When allowing deviations and/or shifts (see *Materials and Methods*) between the PFM, the WUS binding site (TTAATGG) can be identified in the promoters of the AP3, AGL32, and ANR1 subfamilies. TFBS that were identified in MADS promoters previously, such as the CArG-boxes in the AP3 promoter (Place; Riechmann, Wang, and Meyerowitz (1996), could also be uncovered in our analysis.

Discussion

Here, we present a gene family-wide comparison of *Arabidopsis* and poplar MADS promoters. Through phylogenetic footprinting we have tried to identify conserved motifs, and thus putative functional TFBS, between homologous *Arabidopsis* and poplar regulatory regions. We have shown that, although in general the promoters of *Arabidopsis* and poplar MADS-box genes are highly diverged, FootPrinter is an adequate tool that can detect inverted as well as lineage- or duplicate-specific motifs. Consequently, the detection of regulatory elements in these eudicot plant promoters is not hampered by the presence of a high number of duplicates. Moreover, it seems to be an ideal tool to study promoters in a comparative way and to investigate the fate of promoters after duplication and speciation. We observe that, based on this limited data set, the degree of promoter divergence can be linked to the number of duplicates present in a certain subfamily. Accordingly, Huminiecki and Wolfe (2004) describe that there is a general trend for paralogous genes to become more specialized in their expression patterns, with decreased breadth and increased specificity of expression as gene family size increases. Thus, gene duplication seems to be a major driving force behind the emergence of divergent gene expression patterns. Indeed, a large fraction of duplicated genes have been shown to possess diverged gene expression patterns (Blanc and Wolfe 2004; Casneuf et al. 2006). According to Force et al. (1999), the probability of subfunctionalization increases with the size and regulatory complexity (i.e., number of enhancers with distinct spatiotemporal expression) of genes relative to their coding regions. As transcription factors are thought to possess high regulatory complexity (Wray et al. 2003), one would expect that this class of genes is particularly prone to subfunctionalization. However, in the MADS-box gene family, both cases of subfunctionalization as well as redundancy and neo/nonfunctionalization seem to occur based on promoter sequence and expression analysis (Ferrario, Immink, and Angenent 2004). Whereas the redundancy of promoters can be necessary to provide (developmental) backup to the organism (Nadeau and Sankoff 1997; Kafri, Bar-Even, and Pilpel 2005), sub- and neo/nonfunctionalization can enable diversification in developmental pathways generating interspecific variation (Doebley and Lukens 1998).

From our study of the *AP3* homologous promoters, we can conclude that adding additional species will enhance the detection of regulatory motifs through phylogenetic footprinting. In particular, we expect the genome sequencing efforts of tomato, *Medicago*, and *Lotus* to be rewarding in this respect. However, including species that are very closely related to *Arabidopsis*, such as *B. oleracea*, can also improve the detection power of a phylogenetic footprinting tool like FootPrinter.

Through comparison of our motifs detected with FootPrinter and those present in publicly available TFBS databases, we could show that this approach is indeed able to detect both known motifs and novel motifs. Some of these novel motifs show similarity with motifs detected in other plant promoters. However, it remains to be seen if these boxes represent active TFBS, which transcription factors bind to them, and which expression patterns they confer. Moreover, the public TFBS databases are far from complete, which inevitably results in a large fraction of the motifs in our data set for which no similarity can be found with motifs in the databases. In addition, reliably identifying motif similarity is not straightforward. Attempting to minimize the number of false positives (see *Materials and Methods*) also entails that short or degenerated binding sites cannot be identified using the approaches discussed here. For instance, due to its degenerated composition, the CArG-box, which is the binding site for MADS transcription factors, could not be identified in most promoters, even though they have been shown to regulate several MADS-box genes. To link the observed promoter sequence divergence with gene expression changes, aiming to assess the contribution of regulatory changes to the evolution of plant development, it will be necessary to perform a broad range of expression studies in different plant species.

Supplementary Material

Tables S1–S7, Figs. S1–S3, and other supplementary materials are available at *Molecular Biology and Evolution* Online (http://www.mbe.oxfordjournals.org/). Supplementary material can also be found at http://bioinformatics.psb.ugent.be.

Acknowledgments

We thank Kerstin Kaufmann, Pia Nutt, and Klaas Vandepoele for suggestions and advice and Lieven Sterck and Stephane Rombauts for technical help. We acknowledge the DoE Joint Genome Institute and Poplar Genome Consortium for providing the genome sequence of *P. trichocarpa*. S.D.B. is indebted to the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) and the European Science Foundation "Integrated Approaches for Functional Genomics Program." This work was supported by a grant from the Fund for Scientific Research, Flanders (3G031805).

Literature Cited

Ayre, B. G., J. E. Blair, and R. Turgeon. 2003. Functional and phylogenetic analyses of a conserved regulatory program in the phloem of minor veins. Plant Physiol. **133**:1229–1239.

Blanc, G., and K. H. Wolfe. 2004. Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16**:1679–1691.

Blanchette, M., B. Schwikowski, and M. Tompa. 2002. Algorithms for phylogenetic footprinting. J. Comput. Biol. **9**:211–223.

Blanchette, M., and M. Tompa. 2003. FootPrinter: a program designed for phylogenetic footprinting. Nucleic Acids Res. **31**:3840–3842.

Bowers, J. E., B. A. Chapman, J. Rong, and A. H. Paterson. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422**:433–438.

Brunner, A. M., W. H. Rottmann, L. A. Sheppard, K. Krutovskii, S. P. DiFazio, S. Leonardi, and S. H. Strauss. 2000. Structure and expression of duplicate AGAMOUS orthologues in poplar. Plant Mol. Biol. **44**:619–634.

Buchanan, K. L., E. A. Smith, S. Dou, L. M. Corcoran, and C. F. Webb. 1997. Family-specific differences in transcription efficiency of Ig heavy chain promoters. J. Immunol. **159**:1247–1254.

Bulyk, M. L. 2003. Computational prediction of transcription-factor binding site locations. Genome Biol. **5**:201.

Busch, M. A., K. Bomblies, and D. Weigel. 1999. Activation of a floral homeotic gene in Arabidopsis. Science **285**:585–587.

Casneuf, T., S. De Bodt, J. Raes, S. Maere, and Y. Van de Peer. 2006. Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana. Genome Biol. **7**:R13.

Chuzhanova, N. A., M. Krawczak, L. A. Nemytikova, V. D. Gusev, and D. N. Cooper. 2000. Promoter shuffling has occurred during the evolution of the vertebrate growth hormone gene. Gene **254**:9–18.

Coen, E. S., and E. M. Meyerowitz. 1991. The war of the whorls: genetic interactions controlling flower development. Nature **353**:31–37.

Coessens, B., G. Thijs, S. Aerts, K. Marchal, F. De Smet, K. Engelen, P. Glenisson, Y. Moreau, J. Mathys, and B. De Moor. 2003. INCLUSive: a web portal and service registry for microarray and regulatory sequence analysis. Nucleic Acids Res. **31**:3468–3470.

Collado-Vides, J., B. Magasanik, and J. D. Gralla. 1991. Control site location and transcriptional regulation in Escherichia coli. Microbiol. Rev. **55**:371–394.

Courey, A. J. 2001. Cooperativity in transcriptional control. Curr. Biol. **11**:R250–R252.

De Bodt, S., S. Maere, and Y. Van de Peer. 2005. Genome duplication and the origin of angiosperms. Trends Ecol. Evol. **20**:591–597.

De Bodt, S., J. Raes, Y. Van de Peer, and G. Theissen. 2003. And then there were many: MADS goes genomic. Trends Plant Sci. **8**:475–483.

de Folter, S., R. G. Immink, M. Kieffer et al. (12 co-authors). 2005. Comprehensive interaction map of the Arabidopsis MADS box transcription factors. Plant Cell **17**:1424–1433.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. **19**:1114–1121.

Ditta, G., A. Pinyopich, P. Robles, S. Pelaz, and M. F. Yanofsky. 2004. The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity. Curr. Biol. **14**:1935–1940.

Doebley, J., and L. Lukens. 1998. Transcriptional regulators and the evolution of plant form. Plant Cell **10**:1075–1082.

Ferrandiz, C., Q. Gu, R. Martienssen, and M. F. Yanofsky. 2000. Redundant regulation of meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER. Development **127**:725–734.

Ferrario, S., R. G. Immink, and G. C. Angenent. 2004. Conservation and diversity in flower land. Curr. Opin. Plant Biol. **7**:84–91.

Fickett, J. W., and W. W. Wasserman. 2000. Discovery and modeling of transcriptional regulatory regions. Curr. Opin. Biotechnol. **11**:19–24.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**:1531–1545.

Guo, H., and S. P. Moose. 2003. Conserved noncoding sequences among cultivated cereal genomes identify candidate regulatory sequence elements and patterns of promoter evolution. Plant Cell **15**:1143–1158.

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

Hill, T. A., C. D. Day, S. C. Zondlo, A. G. Thackeray, and V. F. Irish. 1998. Discrete spatial and temporal cis-acting elements regulate transcription of the Arabidopsis floral homeotic gene APETALA3. Development **125**:1711–1721.

Hong, R. L., L. Hamaguchi, M. A. Busch, and D. Weigel. 2003. Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell **15**:1296–1309.

Honma, T., and K. Goto. 2001. Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. Nature **409**:525–529.

Huminiecki, L., and K. H. Wolfe. 2004. Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res. **14**:1870–1879.

Inada, D. C., A. Bashir, C. Lee, B. C. Thomas, C. Ko, S. A. Goff, and M. Freeling. 2003. Conserved noncoding sequences in the grasses. Genome Res. **13**:2030–2041.

Kafri, R., A. Bar-Even, and Y. Pilpel. 2005. Transcription control reprogramming in genetic backup circuits. Nat. Genet. **37**:295–299.

Kaufmann, K., R. Melzer, and G. Theissen. 2005. MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene **347**:183–198.

Kellogg, E. A. 2004. Evolution of developmental traits. Curr. Opin. Plant. Biol. **7**:92–98.

Koch, M. A., B. Weisshaar, J. Kroymann, B. Haubold, and T. Mitchell-Olds. 2001. Comparative genomics and regulatory evolution: conservation and function of the Chs and Apetala3 promoters. Mol. Biol. Evol. **18**:1882–1891.

Latchman, D. S. 1998. Transcription factors: an overview. Academic Press, San Diego, Calif.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262**:208–214.

Lescot, M., P. Dehais, G. Thijs, K. Marchal, Y. Moreau, Y. Van de Peer, P. Rouze, and S. Rombauts. 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.

Maere, S., S. De Bodt, J. Raes, T. Casneuf, M. Van Montagu, M. Kuiper, and Y. Van de Peer. 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. USA **102**:5454–5459.

Malcomber, S. T., and E. A. Kellogg. 2005. SEPALLATA gene diversification: brave new whorls. Trends Plant Sci. **10**: 427–435.

Nadeau, J. H., and D. Sankoff. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147**:1259–1266.

Ng, M., and M. F. Yanofsky. 2001. Function and evolution of the plant MADS-box gene family. Nat. Rev. Genet. **2**:186–195.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin, Germany.

Otto, S. P., and J. Whitton. 2000. Polyploid incidence and evolution. Annu. Rev. Genet. **34**:401–437.

Parcy, F., O. Nilsson, M. A. Busch, I. Lee, and D. Weigel. 1998. A genetic framework for floral patterning. Nature **395**:561–566.

Parenicova, L., S. de Folter, M. Kieffer et al. (12 co-authors). 2003. Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell **15**:1538–1551.

Pavesi, G., G. Mauri, and G. Pesole. 2004. In silico representation and discovery of transcription factor binding sites. Brief. Bioinform. **5**:217–236.

Pelaz, S., G. S. Ditta, E. Baumann, E. Wisman, and M. F. Yanofsky. 2000. B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature **405**:200–203.

Pinyopich, A., G. S. Ditta, B. Savidge, S. J. Liljegren, E. Baumann, E. Wisman, and M. F. Yanofsky. 2003. Assessing the redundancy of MADS-box genes during carpel and ovule development. Nature **424**:85–88.

Prince, V. E., and F. B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. **3**:827–837.

Qiu, P. 2003. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochem. Biophys. Res. Commun. **309**:495–501.

Quandt, K., K. Frech, H. Karas, E. Wingender, and T. Werner. 1995. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res. **23**:4878–4884.

Riechmann, J. L., M. Wang, and E. M. Meyerowitz. 1996. DNA-binding properties of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. Nucleic Acids Res. **24**:3134–3141.

Rombauts, S., K. Florquin, M. Lescot, K. Marchal, P. Rouze, and Y. van de Peer. 2003. Computational approaches to identify promoters and cis-regulatory elements in plant genomes. Plant Physiol. **132**:1162–1176.

Schiex, T., A. Moisan, and P. Rouzé. 2001. EuGene: an eukaryotic gene finder that combines several sources of evidence. Pp. 111–125 *in* O. Gascuel and M.-F. Sagot, eds. Computational biology, LNCS 2066. Springer, Heidelberg, Germany.

Schuler, G. D., S. F. Altschul, and D. J. Lipman. 1991. A workbench for multiple alignment construction and analysis. Proteins **9**:180–190.

Simillion, C., K. Vandepoele, M. C. Van Montagu, M. Zabeau, and Y. Van de Peer. 2002. The hidden duplication past of Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA **99**: 13627–13632.

Soltis, D. E., P. S. Soltis, V. A. Albert, D. G. Oppenheimer, C. W. dePamphilis, H. Ma, M. W. Frohlich, and G. Theissen. 2002. Missing links: the genetic architecture of flower and floral diversification. Trends Plant Sci. **7**:22–31.

Sterck, L., S. Rombauts, S. Jansson, F. Sterky, P. Rouzé, and Y. Van de Peer. 2005. EST data suggest that poplar is an ancient polyploid. New Phytol. **167**:165–170.

Stone, J. R., and G. A. Wray. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol. **18**:1764–1770.

Tautz, D. 2000. Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10**:575–579.

Theissen, G. 2001. Development of floral organ identity: stories from the MADS house. Curr. Opin. Plant Biol. **4**:75–85.

———. 2002. Secret life of genes. Nature **415**:741.

Theissen, G., and H. Saedler. 2001. Plant biology. Floral quartets. Nature **409**:469–471.

Tompa, M., N. Li, T. L. Bailey et al. (25 co-authors). 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. **23**:137–144.

Vandenbussche, M., J. Zethof, S. Royaert, K. Weterings, and T. Gerats. 2004. The duplicated B-class heterodimer model: whorl-specific effects and complex genetic interactions in Petunia hybrida flower development. Plant Cell **16**:741–754.

Van de Peer, Y., and R. De Wachter. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Appl. Biosci. **10**:569–570.

Wasserman, W. W., M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence. 2000. Human-mouse genome comparisons to locate regulatory sites. Nat. Genet. **26**:225–228.

Wasserman, W. W., and A. Sandelin. 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. **5**:276–287.

Weigel, D., and E. M. Meyerowitz. 1994. The ABCs of floral homeotic genes. Cell **78**:203–209.

Weitzman, J. B. 2003. Tracking evolution's footprints in the genome. J. Biol. **2**:9.

Wendel, J. F. 2000. Genome evolution in polyploids. Plant Mol. Biol. **42**:225–249.

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

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–1419.

Zhang, Z., and M. Gerstein. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J. Biol. **2**:11.

Pekka Pamilo, Associate Editor

Accepted March 29, 2006