the plant journal

SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2009) **59**, 987–1000 doi: 10.1111/j.1365-313X.2009.03928.x

Intronic regulatory elements determine the divergent expression patterns of *AGAMOUS-LIKE6* subfamily members in Arabidopsis

Stephen E. Schauer[1], Philipp M. Schlüter[1], Ramarmurthy Baskar[1,2,†], Jacqueline Gheyselinck[1,‡], Arturo Bolaños[1], Mark D. Curtis[1] and Ueli Grossniklaus[1,2,*]

[1]*Institute of Plant Biology & Zürich-Basel Plant Science Center, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland, and*

[2]*Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA*

Received 9 March 2009; revised 7 May 2009; accepted 12 May 2009; published online 22 June 2009.

[*]For correspondence (fax +41 44 634 8204; e-mail grossnik@botinst.uzh.ch).

[†]Present address: Department of Biotechnology, Indian Institute of Technology-Madras, Chennai-600 036, India.

[‡]Present address: Département de Biologie Moléculaire Végétale, Bâtiment Biophore, Université de Lausanne, CH-1015 Lausanne, Switzerland.

SUMMARY

The screening of enhancer detector lines in *Arabidopsis thaliana* has identified genes that are specifically expressed in the sporophytic tissue of the ovule. One such gene is the MADS-domain transcription factor *AGAMOUS-LIKE6* (*AGL6*), which is expressed asymmetrically in the endothelial layer of the ovule, adjacent to the developing haploid female gametophyte. Transcription of *AGL6* is regulated at multiple stages of development by enhancer and silencer elements located in both the upstream regulatory region and the large first intron. These include a bipartite enhancer, which requires elements in both the upstream regulatory region and the first intron, active in the endothelium. Transcription of the *AGL13* locus, which encodes the other member of the *AGL6* subfamily in Arabidopsis, is also regulated by elements located in the upstream regulatory region and in the first intron. There is, however, no overlapping expression of *AGL6* and *AGL13* except in the chalaza of the developing ovule, as was shown using a dual gene reporter system. Phylogenetic shadowing of the first intron of *AGL6* and *AGL13* homologs from other Brassicaceae identified four regions of conservation that probably contain the binding sites of transcriptional regulators, three of which are conserved outside Brassicaceae. Further phylogenetic analysis using the protein-encoding domains of *AGL6* and *AGL13* revealed that the MADS DNA-binding domain shows considerable divergence. Together, these results suggest that *AGL6* and *AGL13* show signs of subfunctionalization, with divergent expression patterns, regulatory sequences and possibly functions.

Keywords: regulatory introns, *AGAMOUS-LIKE6*, enhancer, phylogenetic shadowing, gene regulation, MADS-domain transcription factors.

INTRODUCTION

Despite the sequencing of the *Arabidopsis thaliana* genome, the role of most of the identified genes has yet to be uncovered (The Arabidopsis Genome Initiative 2000). This is especially true for genes with late functions in reproductive development, due to both developmental epistasis (the use of a gene at multiple stages of development) and genetic redundancy (compensation of a loss-of-function mutation by the activity of another gene). One way to circumvent these effects is to screen enhancer detector (or enhancer trap, ET) and gene trap (GT) lines for specific late spatiotemporal expression patterns (Bellen *et al.*, 1989; Grossniklaus *et al.*, 1989; Klimyuk *et al.*, 1995; Sundaresan *et al.*, 1995; Bellen, 1999).

Enhancer detection in particular is a powerful tool for understanding genetic regulation. Enhancers are positive transcriptional regulatory elements that increase the transcription rate of target genes, and whose genomic position is not constrained: enhancers can be found upstream or downstream of the transcribed region of the gene – or in intronic regions – and their function is independent of

orientation (reviewed in Kleinjan and van Heyningen, 2005). Silencers are similar to enhancers, but they are bound by transcriptional repressors rather than activators. Enhancer detector constructs contain a reporter gene [usually the *uidA* gene of *Escherichia coli*, which encodes a β-glucuronidase (GUS; Jefferson, 1989)] under the control of a minimal promoter [usually from the cauliflower mosaic virus 35S RNA promoter (*min35S*; Benfey and Chua, 1990)]. In the system established by Sundaresan *et al.* (1995), used in this study, the *min35S:GUS* reporter was engineered inside a modified maize *Dissociator* transposable element (*Ds*). When this *Ds* element is integrated into the genome, any adjacent enhancer/silencer element can regulate the transcription of the *min35S:GUS* gene present on *Ds*. While a large number of enhancer detector lines have been isolated, the nature and position of the enhancer in – or relative to – the gene has rarely been investigated (Yang *et al.*, 2005).

While studies have illuminated the role of enhancer and silencer elements inside intronic regions of plant genes (Reddy and Reddy, 2004; Kim *et al.*, 2006; Searle *et al.*, 2006), the first Arabidopsis gene shown to have regulatory elements inside an intron was the MADS-domain transcription factor *AGAMOUS* (*AG*; Sieburth and Meyerowitz, 1997), and enhancer and silencer elements appear to be critical for *AG* evolution (Causier *et al.*, 2005, 2008). Although most introns in Arabidopsis are between 70 and 120 bp in length, all MIKC^c-type MADS-domain genes – with the exception of *APETALA3* and *AGAMOUS-LIKE15* (*AGL15*, Heck *et al.*, 1995; Jack *et al.*, 1992) – contain at least one intron that is larger than 500 bp. Sieburth and Meyerowitz (1997) showed that spatiotemporal regulatory elements contained in the large, second intron of *AG* were necessary for both gene activation and repression. Consistent with this finding, the MADS-domain transcription factor genes *FLOWERING LOCUS C* (or *AGL25*) and *SEEDSTICK* (or *AGL11*) also have important regulatory elements inside their large first intron (Sheldon *et al.*, 2002; Kooiker *et al.*, 2005). The roles of large introns in the regulation of other *AGL*s have not been thoroughly examined. Specifically, it is unknown to what extent the regulatory elements inside the large introns of *AGL*s are sufficient to control their spatiotemporal gene expression (i.e. act as independent enhancers or silencers), or if they require additional upstream regulatory elements for activity (as bipartite enhancers or silencers).

In this paper we report the isolation of enhancer detector line ET447, which is inserted inside *AGL6* (*At2g45650*). *AGL6* is a member of the poorly investigated *AGL6* subfamily of MADS-domain proteins, with unknown function (Becker and Theißen, 2003). Like other *AGL*s, *AGL6* has a K-domain (necessary for protein-protein interactions, Fan *et al.*, 1997) and an intervening domain (I-domain) between the K and MADS DNA-binding domains (Ma *et al.*, 1991). The Arabidopsis genome also contains the additional *AGL6* subfamily member *AGL13* (*At3g61120*), which is located on a genomic block that arose from a genome duplication event 35–85 million years ago (Ma; Blanc *et al.*, 2003; Bowers *et al.*, 2003). To date, no mutant phenotype for either gene has been identified in Arabidopsis (Ma *et al.*, 1991; Rounsley *et al.*, 1995). Importantly, only the first of the seven introns in *AGL6* and *AGL13* is longer than 120 bp, implying that the large, first intron of *AGL6* and/or *AGL13* could contain regulatory elements necessary for proper spatiotemporal expression.

The role of changes in spatiotemporal expression in the evolution of duplicated genes has been much researched (reviewed in Li *et al.*, 2005; Rijpkema *et al.*, 2007). The duplication–degeneration–complementation (DDC) model of Force *et al.* (1999) predicts three outcomes of duplicated genes: nonfunctionalization, neofunctionalization, and subfunctionalization. In nonfunctionalization, null mutations occur in one paralog, ultimately generating a non-expressed pseudogene. In neofunctionalization, one of the paralogs mutates to gain a novel function (perhaps due to changes in regulatory elements), and is established by positive selection. In subfunctionalization, there is a reciprocal loss of some regulatory elements from each paralog by 'degenerative mutations' (Force *et al.*, 1999), such that the paralogs' combined expression pattern replicates the ancestral expression pattern. While some gene pairs (for example *pax6a* and *pax6b* in vertebrates; Kleinjan *et al.*, 2008) are consistent with the DDC model, the *hoxb5a* and *hoxb5b* gene pair from teleosts suggests a more complicated story: there was no evidence for a simple reciprocal loss of regulatory elements and the interactions between regulatory elements was deemed critical (Jarinova *et al.*, 2008). While subfunctionalization has been invoked to explain many of the duplicated MADS-domain genes (Rijpkema *et al.*, 2007), verification of the reciprocal loss of regulatory elements, which is critical for the DDC model, is lacking in plants. Here, we show that *AGL6* and *AGL13* are regulated through complex interactions of enhancer and silencer elements located in both the 5′-upstream regions and the large first introns of these genes. Phylogenetic analyses suggest that *AGL6* and *AGL13* are undergoing subfunctionalization, probably involving the attenuation of enhancer elements.

RESULTS

The expression of *AGL6* is controlled by elements both upstream and inside the first intron

The reporter activity of ET447, a line uncovered in a screen for ovule-specific expression patterns, is initiated in the presumptive endothelial layer while the female gametophyte is at the two-nuclear stage, before the morphological

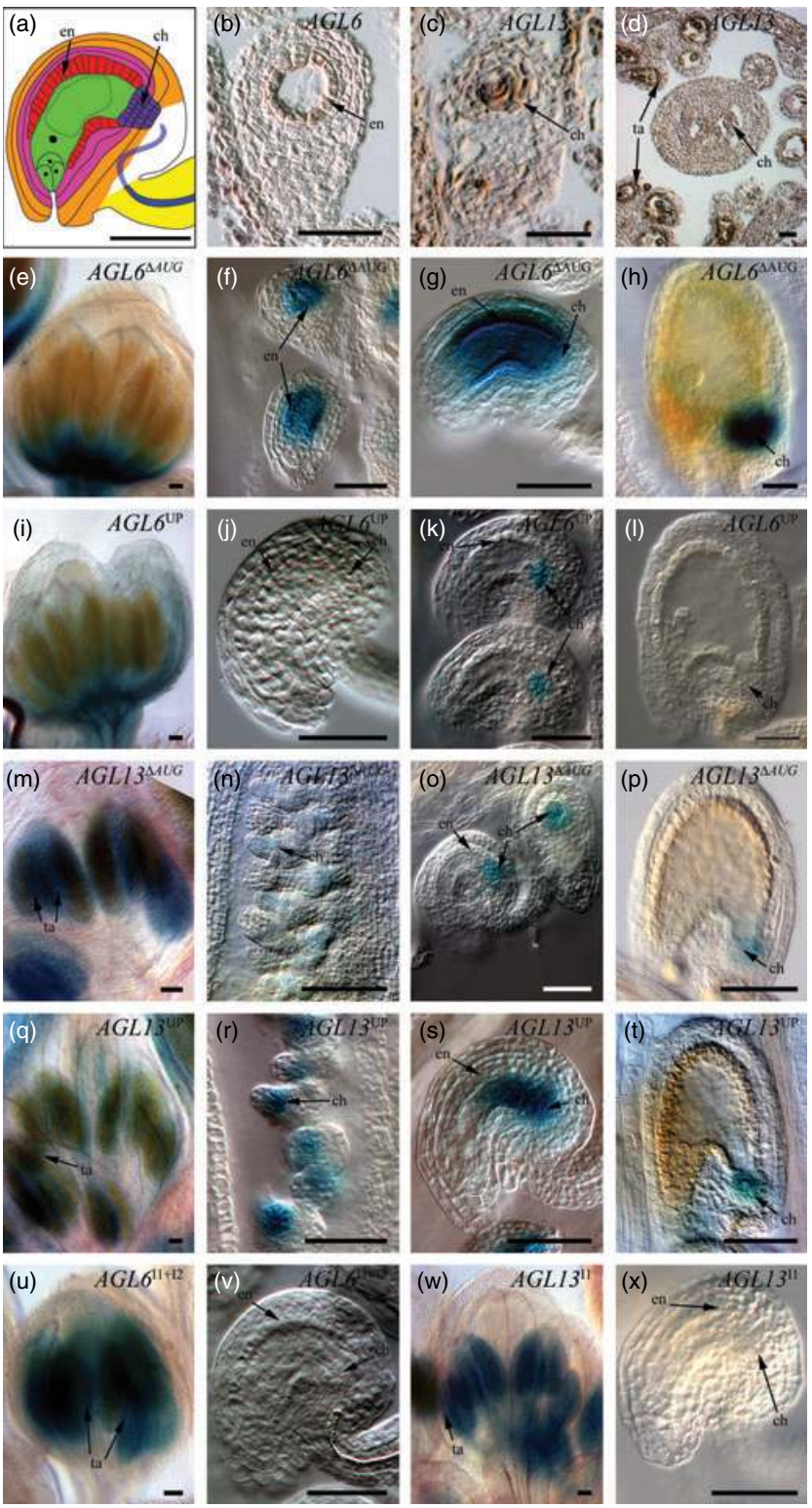

Figure 1. Expression patterns of *AGL6* and *AGL13* require regulatory elements located inside their first introns.

(a) Diagram of a mature ovule. The female gametophyte is shown in green, surrounded by the endothelium (shown in red) and chalaza (in purple). The other two cell layers of the inner integument are shown in pink, while the cell layers of to the outer integument are shown in orange. The vasculature is shown in blue, and the funiculus in yellow.

(b–d) *In situ* RNA hybridization showing the localization of *AGL6* in the endothelium (b) and *AGL13* in the chalaza (c) and the anther tapetum (d).

(e–x) *Promoter:GUS* fusions. *AGL6*$^{\Delta AUG}$*:GUS* expression in stage 11 flowers (e), four-nuclear stage of female gametophyte development (f), mature ovules (g) and developing seeds at globular stage of embryogenesis (h). *AGL6*^{UP}*:GUS* expression in stage 11 flowers (i), four-nuclear stage of female gametophyte development (j), mature ovules (k) and developing seeds at the two-cell stage of embryogenesis (l). *AGL13*$^{\Delta AUG}$*:GUS* expression in stage 11 flowers (m), integument initiation stage of ovule development (n), mature ovules (o) and developing seeds at the two-cell stage of embryogenesis (p). *AGL13*^{UP}*:GUS* expression in stage 11 flowers (q), integument initiation stage of ovule development (r), mature ovules (s), and developing seeds at the two-cell stage of embryogenesis (t). *AGL6*$^{I1+I2as}$*:min35S:GUS* expression in stage 11 flowers (u), and mature ovules (v). *AGL13*^{I1as}*:min35S:GUS* expression in stage 11 flowers (w), and mature ovules (x). All bars represent 50 μm; ch, chalaza; en, endothelial layer; ta, tapetum.

formation of the endothelial layer (Schneitz *et al.*, 1995; de Folter *et al.*, 2006) (Figure 1). Reporter activity was detected throughout the endothelial layer until fertilization, after which it became restricted to the endothelium at the chalazal end of the seed (Figure S1 in Supporting Information). By the heart stage of embryo development (Meinke and Sussex, 1979), ET447 reporter activity was no longer detected.

The *Ds* insertion in ET447 is 14 bp upstream of the 5′ splice donor site of exon 1 of *AGL6*, which encodes the MADS DNA-binding domain. Not only does the *Ds* insertion introduce two in-frame stop codons, but also *AGL6* mRNA was not detected by reverse transcriptase (RT)-PCR in homozygous plants (data not shown). ET8885, an additional ET line with an insertion inside intron 1 of *AGL6* (267 bp downstream of the 5′ splice donor site of exon 1), reproduced the expression pattern of ET447 (Figure S1). Both expression patterns are consistent with the *in situ* RNA hybridization data for *AGL6* (Figure 1b).

To identify the location of the enhancer identified by both ET447 and ET8885, a series of *promoter:GUS* constructs were made that contained: (i) the upstream sequence in combination with the first intron, (ii) the upstream sequence alone, and (iii) the first intron upstream of *min35S* (Figures 2, S2 and Appendix S1). The $AGL6^{\Delta AUG}$*:GUS* lines (which contain the upstream region in combination with the first intron but lack the translational initiation site of *AGL6*) replicated the expression pattern of ET447 and ET8885, implying that the endothelium-specific enhancer is contained in this genomic region (Figures 1f–h and S1m). In order to investigate the synergistic effects of upstream and intronic elements, the $AGL6^{\Delta AUG}$ sequence was divided into upstream ($AGL6^{UP}$) and intronic regions ($AGL6^{I1+12}$). In plants containing $AGL6^{UP}$*:GUS*, reporter activity was not detected early in ovule development (Figure 1j). In mature ovules, reporter activity was restricted to the chalazal region, and did not expand into the endothelial cell layer (Figure 1k), although this chalaza-specific expression was not maintained later during seed development (Figure 1l). This suggests that the endothelium-specific enhancer element is located in the first intron. However, plant lines containing the $AGL6^{I1+I2s}$*:min35S:GUS* or $AGL6^{I1+I2as}$*:min35S:GUS* constructs, which have the intron upstream of a minimal promoter in either sense or antisense orientation, showed no reporter activity throughout ovule development (Figure 1v). Taken together, this analysis suggests a bipartite endothelium-specific enhancer at the *AGL6* locus that requires the activity of regulatory elements that are located upstream of the translational initiation site as well as inside the first intron (Figure 2).

Additionally, reporter activity was detected for both $AGL6^{I1+I2s}$*:min35S:GUS* and $AGL6^{I1+I2as}$*:min35S:GUS* constructs in the tapetum, starting in stage 8 flowers (Smyth *et al.*, 1990) and ending before stage 12 (Figure 1u), indicating the presence of a tapetum-specific enhancer inside intron 1 of *AGL6*. However, tapetal expression was not seen

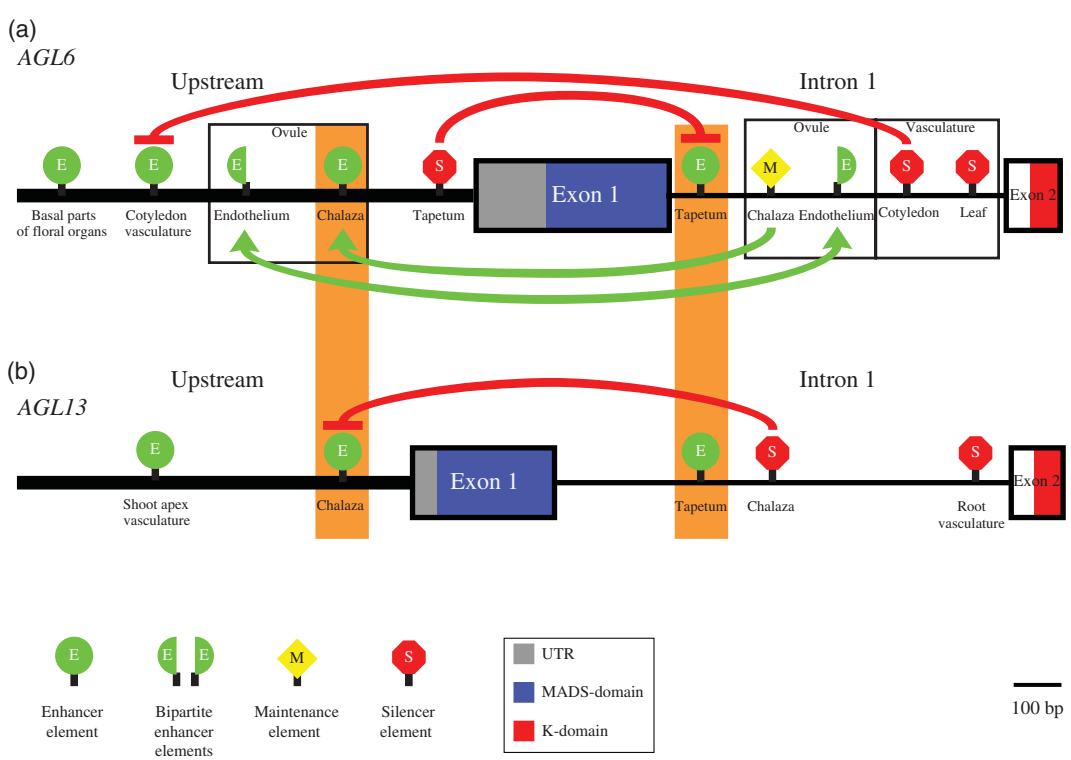

Figure 2. Summary of the regulatory elements of the *AGL6* subfamily members in Arabidopsis.
A summary of the enhancer and silencer elements demonstrated in Figures 1, S1 and S3 relative to their position in either the upstream or the intronic regions of *AGL6* (a) and *AGL13* (b). The exact position of the elements is unknown. Green arrows indicate positive interactions between elements, while red bars indicate attenuation. Elements with similar expression patterns in *AGL6* (a) and *AGL13* (b) are indicated in orange boxes.

in ET447, ET8885, *AGL6*$^{\Delta AUG}$*:GUS* or *AGL6*^{UP}*:GUS* lines. Instead, reporter activity was detected in the basal parts of all the floral organs during these floral stages, ultimately restricted to the abscission zone (Figure 1e,i).

Finally, examination of 12-day-old seedlings identified a silencer activity located inside the first intron. No reporter activity was detected for either ET447 or ET8885 lines at this stage. Weak reporter activity, however, was detected in the vasculature of the cotyledons of *AGL6*$^{\Delta AUG}$*:GUS* lines (Figure S3b). This expression domain expanded to the vasculature of the developing leaves in the *AGL6*^{UP}*:GUS* lines (Figure S3g), but no expression was detected in the *AGL6*$^{I1 + I2s}$*:min35S:GUS* or *AGL6*$^{I1 + I2as}$*:min35S:GUS* lines, indicating that there is a leaf vasculature-specific silencer present inside the first intron that prevents expression in seedlings.

A different set of upstream and intronic regulatory elements controls the expression of *AGL13*

To address whether a similar complex set of intronic elements are found in the other *AGL6* subfamily member, the regulatory regions of *AGL13* were examined. Although it was not isolated in the initial screen, the enhancer detector line ET5830, with the *Ds* element inserted inside intron 1 of *AGL13* (189 bp upstream of the 3′ splice acceptor site of exon 2), was examined. While the insertion does not disrupt the coding sequence, no *AGL13* mRNA was detected by RT-PCR in homozygous plants (data not shown). In contrast to ET447 and ET8885, no ET5830 reporter gene activity was observed throughout ovule or seed development (Figure S1). Consistent with the *in situ* hybridization pattern of *AGL13* (Figure 1d), ET5830 showed reporter gene activity in the tapetum, in a spatiotemporal expression pattern that overlapped with the intron-encoded tapetum-specific enhancer activity of *AGL6*. To narrow down the location of the tapetum-specific enhancer identified by ET5830, a series of *promoter:GUS* constructs were made, similar to those described above (Figure S2). The *AGL13*$^{\Delta AUG}$*:GUS* construct replicated the expression pattern of ET5830 in the tapetum (Figure 1m), suggesting that the enhancer is located in this genomic region. *AGL13*^{UP}*:GUS* lines, however, failed to show the broad tapetum-specific expression pattern (Figure 1q), while both *AGL13*^{I1s}*:min35S:GUS* and *AGL13*^{I1as}*:min35S:GUS* constructs reproduced the expression pattern of ET5830 (Figure 1w).

The *AGL13 promoter:GUS* constructs, however, showed some marked differences from the expression pattern observed in the enhancer detector line. Unlike ET5830, *AGL13*$^{\Delta AUG}$*:GUS* was expressed in the developing ovule, consistent with the observed *in situ* RNA hybridization pattern of *AGL13* (Figure 1c; Rounsley *et al.*, 1995). Reporter activity was detected early in the chalaza, directly adjacent to the megaspore mother cell (Figure 1n), and continued there throughout the rest of ovule development (Figure 1o); it was subsequently detected at the chalazal end of the developing seed coat (Figure 1p). In plants containing *AGL13*^{UP}*:GUS*, which lacks the first intron, the reporter activity showed the same pattern of expression as *AGL13*$^{\Delta AUG}$*:GUS*, although the level of expression was higher at all stages of development (Figure 1n–p). This suggests the presence of a silencer element inside the first intron. Consistent with this hypothesis, plant lines containing either the *AGL13*^{I1s}*:min35S:GUS* or the *AGL13*^{I1as}*:min35S:GUS* construct showed no reporter activity throughout ovule development (Figure 1x), similar to the ET5830 enhancer detector line.

As the regulatory elements located in the first intron of *AGL6* were also active in 12-day-old seedlings, the expression patterns in seedling development were examined. Similar to ET447 and ET8885, no expression was detected in the *AGL13* enhancer detector line ET5830. For the *AGL13*$^{\Delta AUG}$*:GUS* reporter line, however, reporter activity was detected in the vasculature underlying the developing shoot apical meristem (Figure S3i). Reporter activity was also detected in the root vasculature of *AGL13*^{UP}*:GUS* seedlings (Figure S3p), suggesting the presence of a root-specific silencer element inside the first intron of *AGL13* (Figure 2).

Expression of *AGL6* and *AGL13* overlaps in the chalaza

To investigate the overlap in expression of *AGL6* and *AGL13* in the developing ovule, plant lines expressing the plant codon-optimized version of the *Clostridium perfringens* neuraminidase *nanH* (or *NAN*; Kirby and Kavanagh, 2002) reporter gene under the control of the *AGL13* regulatory elements (*AGL13*$^{\Delta AUG}$*:NAN*) were crossed to *AGL6*$^{\Delta AUG}$*:GUS*. *AGL13*$^{\Delta AUG}$*:NAN* reporter activity initiated early in the chalaza, before any *AGL6*$^{\Delta AUG}$*:GUS* activity was detected (Figure 3a). However, around the two-nuclear stage, *AGL6*$^{\Delta AUG}$*:GUS* activity initiated in the presumptive endothelial layer and the chalaza, which also showed *AGL13*$^{\Delta AUG}$*:NAN* activity (Figure 3b). Throughout the later stages of ovulogenesis, *AGL13*$^{\Delta AUG}$*:NAN* activity was restricted to the chalaza, while *AGL6*$^{\Delta AUG}$*:GUS* activity was detected in both the chalaza and the endothelium (Figure 3c,d).

Identification of conserved regions of homology inside the first intron

As subfunctionalization predicts the reciprocal loss of regulatory elements between paralogs, the first introns of *AGL6* and *AGL13* were investigated to identify paralog-specific putative binding sites for transcription factors. AthaMap and TRANSFAC® Patch scans of the intron, however, revealed a large number of putative transcription factor-binding sites, hindering the identification of the important factors (Matys *et al.*, 2006; Galuschka *et al.*, 2007). To reduce the number of putative sites, phylogenetic shadowing, which utilizes sequence information from homologous genes of related

species (e.g. Hong *et al.*, 2003), was used to identify regions conserved between several *A. thaliana* relatives. The first intron sequences were aligned for the *AGL6* and *AGL13* homologs from *A. thaliana*, *Arabidopsis halleri*, *Arabidopsis lyrata*, *Boechera gunnisoniana*, *Boechera stricta* and, for *AGL6*, *Brassica oleracea* var. *botrytis* (where an *AGL13* homolog was not amplified). Subsequent analysis identified several conserved areas, which could be loosely grouped into four regions containing several conserved predicted transcription factor-binding sites (Figures 4a and S4–S6 and Table S1).

To assess if these areas of homology were conserved between *AGL6* and *AGL13*, dot-plot analysis was performed, which revealed no obvious sequence similarity between intron 1 of *AGL6* and *AGL13* of *A. thaliana* (Figure 4b). On a dot-plot, a diagonal line indicates a stretch of similar nucleotide sequence. To examine if the conserved intronic regions identified in *AGL6* match those identified in *AGL13*, we looked at the diagonal density, which is a scaled representation of the number of diagonals of a given size class among a pair of sequences. Regions of sequence similarity would be expected to have higher diagonal densities with potentially longer diagonals, while regions with dissimilar sequences would be indistinguishable from the null distribution for diagonal densities (derived from randomly resampled input sequences). The non-conserved regions of intron 1 of *AGL6* and *AGL13* were not statistically different from the null distribution, consistent with overall sequence divergence. When comparing the conserved areas, an excess of small regions of sequence similarity (short diagonals) was observed (Figure 4c). With the exception of a putative intron-mediated enhancement site (IME; Rose *et al.*, 2008), however, none of the predicted transcription factor-binding sites inside the conserved areas was present in both *AGL6* and *AGL13*. Similar dot-plot analysis of the first intron of *PMADS34* (the *Populus trichocarpa AGL6* subfamily member) and *VvMADS3* (the *Vitis vinifera AGL6* subfamily member) revealed that both sequences had significant similarity to parts of Region B of *AGL6*, parts of Region D of *AGL6* and parts of Region B of *AGL13* (Figures 5 and S7). While the aforementioned parts of *AGL6* Region D do not correspond to any predicted transcription factor-binding sites, the identified parts of Regions B of *AGL6* and *AGL13* contain binding sites for a different predicted homeodomain transcription factor, ZmHox2a (Table S1).

Divergence in different domains of *AGL6* subfamily members

AGL6 and *AGL13* are located on duplicated genomic blocks that arose during the most recent genome duplication event that occurred 35–85 Ma in the common ancestor of Brassicaceae (Blanc *et al.*, 2003; Bowers *et al.*, 2003). This is consistent with a phylogenetic analysis of *AGL6* and *AGL13* coding sequences, in which the Brassicaceae *AGL13* homologs formed a sister group to the Brassicaceae *AGL6* homologs, but were nested within the greater *AGL6* subfamily of MADS-domain genes (Figures 6 and S8). When specific domains of the genes were used, the combined I- and K-domains gave similar trees to when using the full-length (or MADS + I + K) sequences. However, the MADS-domains of the *AGL13* homologs showed higher divergence from the rest of the *AGL6* subfamily members, as indicated by a longer branch leading to the *AGL13* homologs (Figures 6 and S9). The longer branch length (that is, more substitutions) suggested that *AGL13* might have been less constrained in its MADS-domain sequence than *AGL6*.

Additionally, the average of the ratios of non-synonymous substitutions (K_a, altering the amino acid sequence) to synonymous substitutions (K_s, preserving the amino acid sequence) from the Brassicaceae *AGL6* subfamily

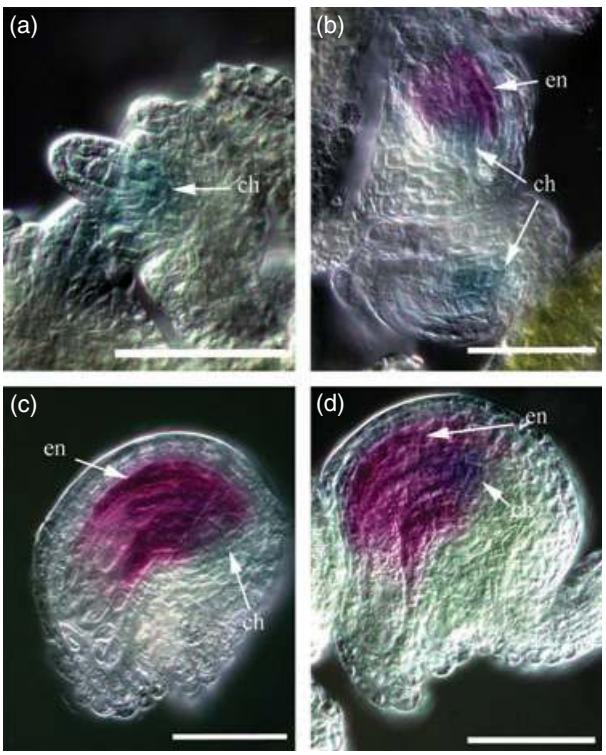

Figure 3. Overlapping expression of the *AGL6* and *AGL13* promoters in the chalaza of the developing ovule.
AGL13$^{\Delta AUG}$*:NAN* activity is detected as an indigo blue precipitate, *AGL6*$^{\Delta AUG}$*:GUS* activity is detected as a magenta precipitate, and overlapping expression is detected as a purple precipitate. *AGL13*$^{\Delta AUG}$*:NAN* activity is first detected in the chalaza at around megasporogenesis (a). At around the four-nuclear stage of female gametophyte development (b), *AGL6*$^{\Delta AUG}$*:GUS* starts to be expressed in the presumptive endothelial layer and *AGL6*$^{\Delta AUG}$*:GUS* and *AGL13*$^{\Delta AUG}$*:NAN* activity overlap in the chalaza. This overlapping expression in the chalaza continues throughout the rest of ovule development (c, d). All bars represent 50 μm; ch, chalaza; en, endothelial layer.

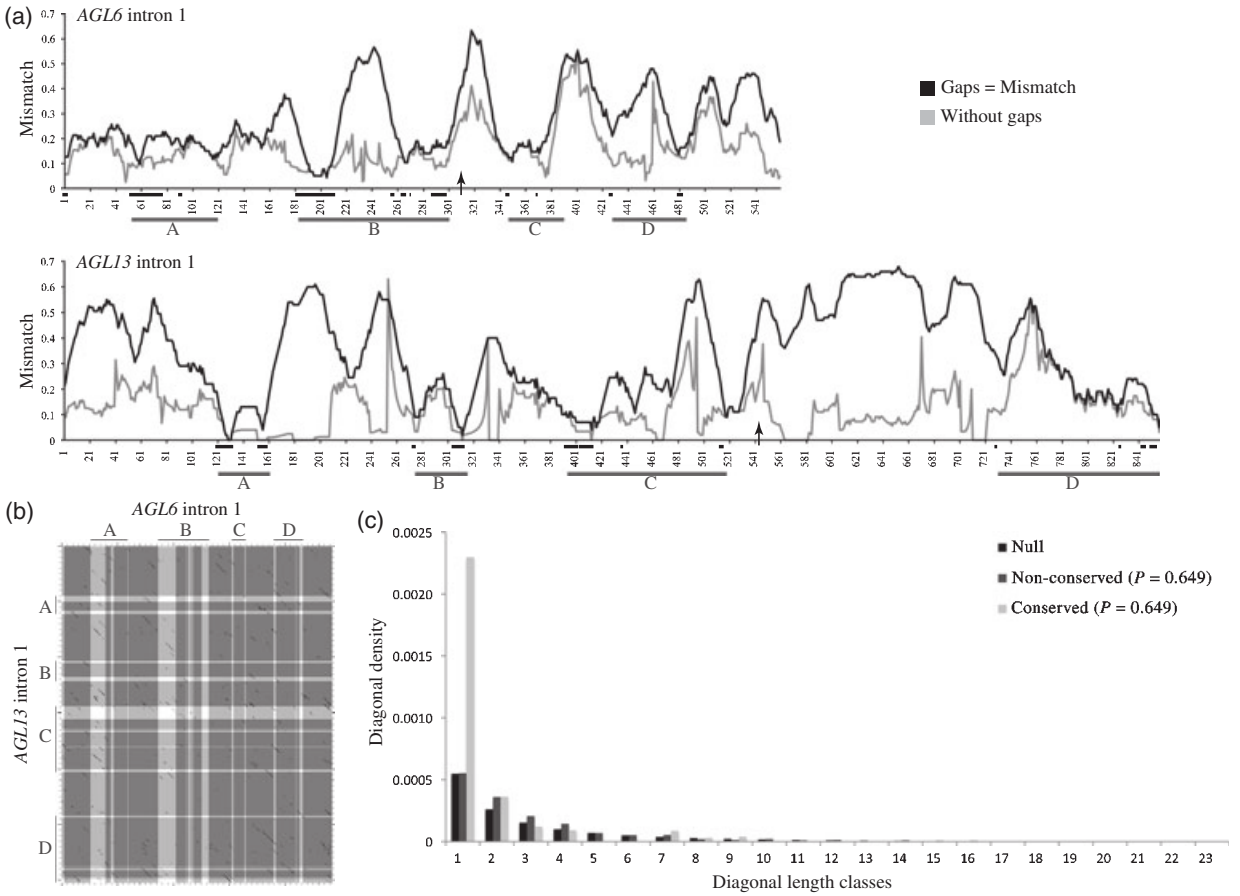

Figure 4. Regions of conservation inside the first intron of *AGL6* and *AGL13*.
(a) A 20-nucleotide sliding window analysis of the aligned sequences showing the average pairwise mismatch for the first intron of *AGL6* (top) and *AGL13* (bottom). The black line on the graph represents an analysis where a gap in the sequence was treated as a mismatch, while the light gray line represents an analysis where gaps were not considered. The *x*-axis indicates the position of the sliding window along the aligned first intron sequence. The conserved areas (as defined from the total of phylogenetic shadowing analyses) are identified by a black line beneath the graph, while the four (loosely defined) regions of homology are identified as dark gray lines. The lower 99% confidence bounds for mismatch values with and without gaps, respectively, were 0.151 and 0.267 for *AGL6*, and 0.122 and 0.332 for *AGL13*. The insertion sites of ET8885 (top) and ET5830 (bottom) are indicated with arrows.
(b) A sliding window dot-plot (window size = 20; allowing eight mismatches per window; step size = 1) comparing the first introns of *AGL6* and *AGL13* of *Arabidopsis thaliana*. Light gray areas and, where overlapping, white boxes indicate the conserved areas of *AGL6* and *AGL13* identified by phylogenetic shadowing, while conserved sliding windows are represented by black dots, with stretches of similarity identified as black diagonal lines.
(c) The distribution of diagonals (i.e. matching sliding windows) identified by dot-plot analysis of intron 1 of *AGL6* and *AGL13*. The black bars correspond to the null distribution of diagonals (as derived from randomly resampled input sequences), the dark gray bars correspond to the distribution of diagonals of the areas of the introns that are not conserved in both *AGL6* and *AGL13*, and the light gray bars correspond to the distribution of diagonals of the areas defined by conservation in both *AGL6* and *AGL13*. The probability of a distribution different from the null distribution (Kolmogorov–Smirnov test) is indicated.

members were compared for the MADS + I + K, the I + K, and the MADS-domains (Figure 6). All average K_a/K_s ratios were <1.0, which is consistent with purifying selection, and all average K_a/K_s ratios were lower in *AGL6* than in *AGL13*. Except for the I + K domain, all K_a/K_s ratios were significantly different between *AGL6* and *AGL13* for the homologous domains. For *AGL6* there was a significant difference between I + K and MADS-domains ($P < 0.001$), whereas for *AGL13* there was no significant difference ($P = 0.29$).

There is evidence for past positive selection using a d_N/d_S-based maximum likelihood ω-estimate, which also uses the ratio of non-synonymous to synonymous substitutions (Figure 6 and Table S4). Branch-specific models that allow for positive selection for the branch prior to genome duplication (ω_a), or leading either to *AGL6* (ω_b) or *AGL13* (ω_c) Brassicaceae homologs, are significantly better than the null model (Table S4), with the sole exception of ω_b for the MADS domain of Brassicaceae *AGL6* homologs. This confirms the K_a/K_s-based analyses but also suggests that selection may already have been acting on the ancestral Brassicaceae *AGL6* homolog prior to the genome duplication event that gave rise to the *AGL6* and *AGL13* paralogs.

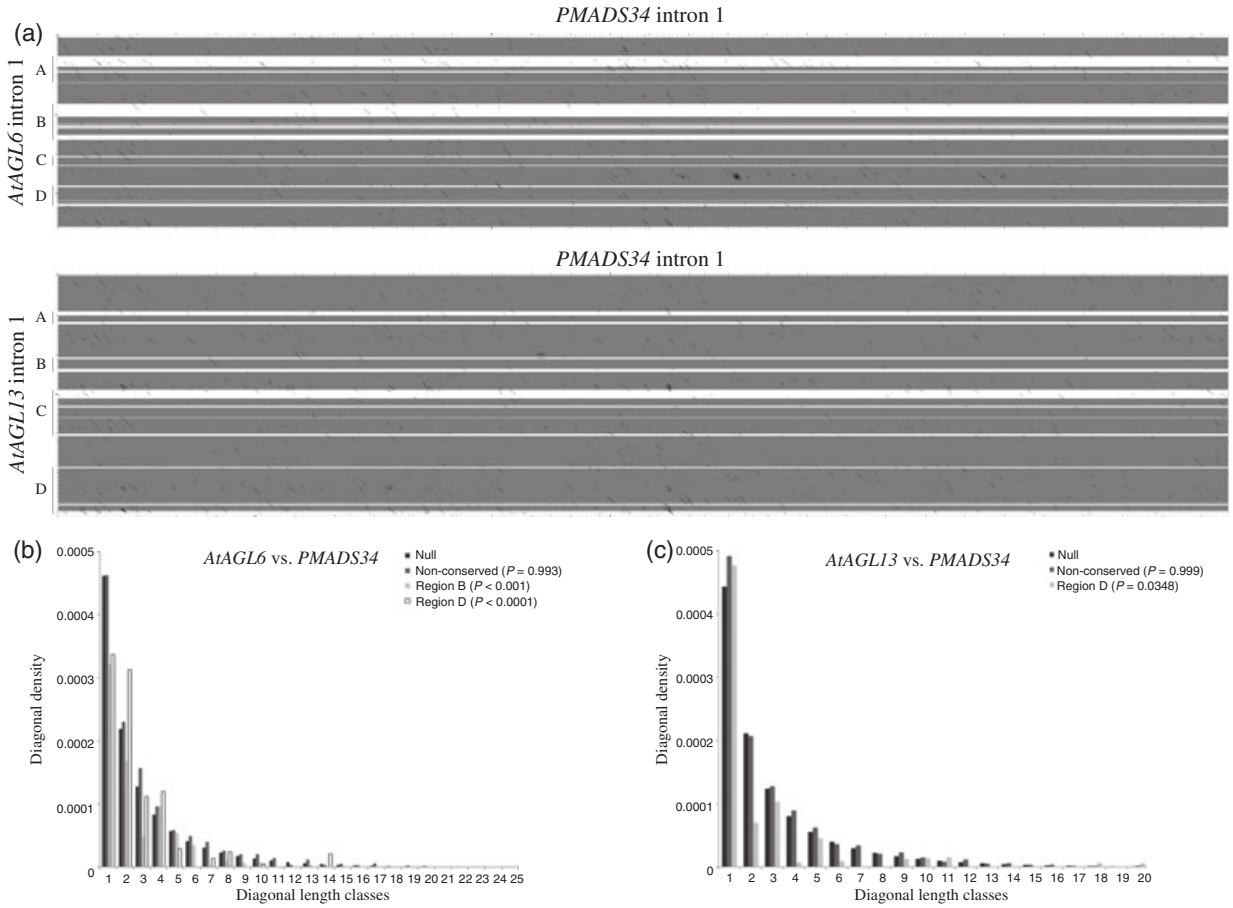

Figure 5. Regions of sequence similarity inside the first intron between *Populus PMADS34* and Arabidopsis *AGL6* subfamily members.
(a) A sliding window dot-plot (window size = 20; allowing eight mismatches per window; step size = 1) comparing the first intron of *PMADS34* of *Populus trichocarpa* and the first introns of *AGL6* (top) and *AGL13* (bottom) of *Arabidopsis thaliana*. White areas indicate the conserved areas of *AGL6* and *AGL13* identified by phylogenetic shadowing, while matching sliding windows are represented by black dots, with stretches of similarity identified as black diagonal lines.
(b, c) The distribution of diagonals identified by dot-plot analysis of intron 1 of *Populus PMADS34* and either Arabidopsis *AGL6* (b) or *AGL13* (c). The black bars correspond to the null distribution of diagonals, the dark gray bars correspond to the distribution diagonals of the non-conserved regions of the *A. thaliana* introns, and the light gray and white bars correspond to the distribution of matching sliding windows of the conserved B and C regions, respectively. Probabilities for a distribution different from the null distribution (Kolmogorov–Smirnov test) are indicated.

Putative ancestral expression pattern of *AGL6*

To distinguish between subfunctionalization and neofunctionalization in the DDC model, the ancestral expression pattern of *AGL6* needs to be established (Figure 7). Gymnosperm *AGL6* homologs are expressed in both male and female reproductive tissues, although expression is initiated earlier and in a broader spatial pattern than for *AGL6* and *AGL13* in Arabidopsis (Mouradov *et al.*, 1998; Shindo *et al.*, 1999; Carlsbecker *et al.*, 2004). A similar pattern occurs in angiosperms, with *AGL6* homologs expressed in both male and female reproductive tissues, although male tissue-specific expression is not detected in some species. Either male-specific expression evolved *de novo* at least three times for the *AGL6* subfamily or it was present ancestrally and then lost. Of the two, it seems most plausible that the ancestral expression pattern for the *AGL6* subfamily was in both the male and female reproductive tissues.

DISCUSSION

Complex regulatory interaction uncovered by enhancer detection screening

The intronic regions of *AGL6* and *AGL13* are necessary for their proper spatiotemporal expression, and contain both enhancer and silencer elements (Figure 2). Moreover, the interactions between upstream and intronic elements are critical. While some of the interactions are relatively simple, such as the bipartite endothelial enhancer of *AGL6* that requires both upstream and intronic elements, other interactions can be more complex. For example, the intron-located tapetum enhancer of *AGL6* is completely attenuated

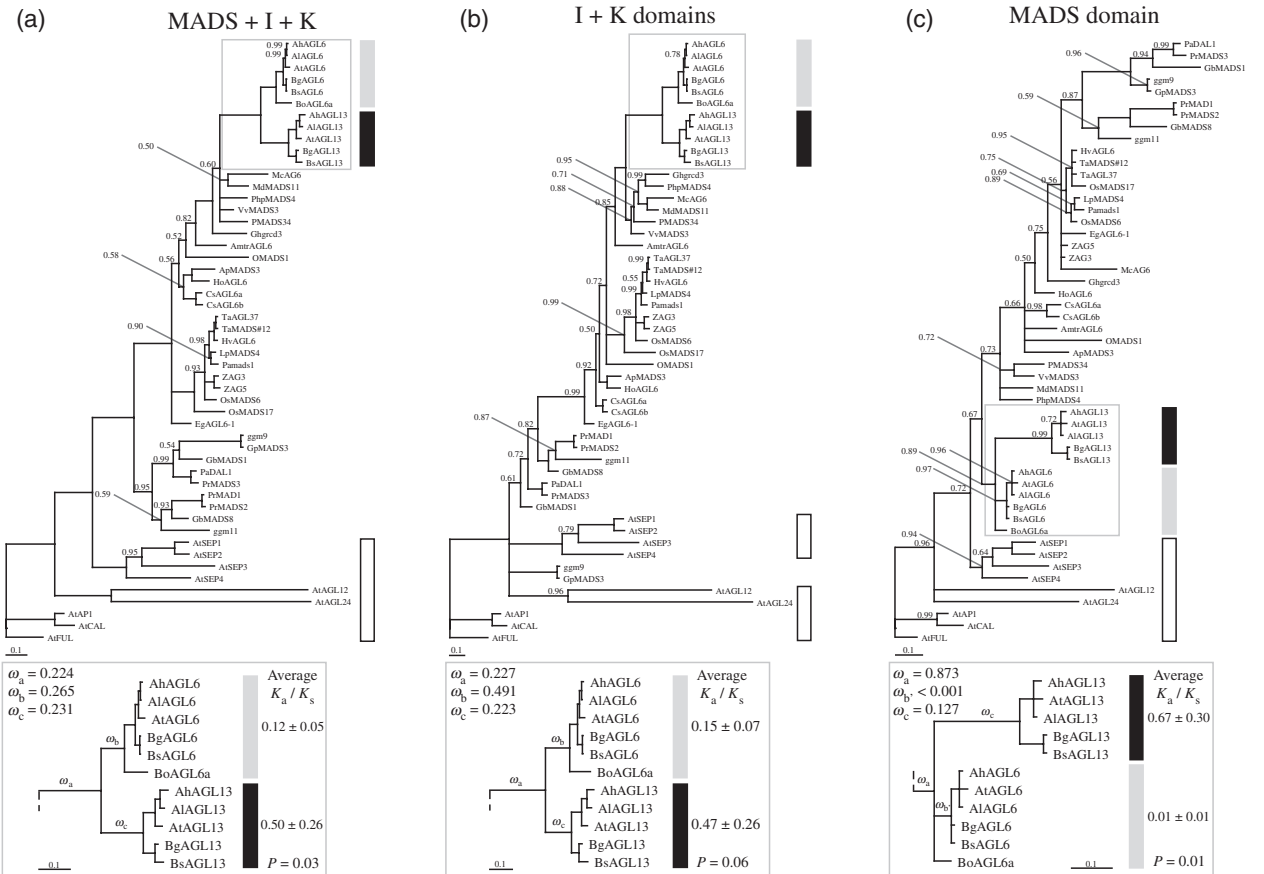

Figure 6. Bayesian inference phylogenetic analyses of domains of *AGL6* subfamily members.
Bayesian inference analysis using (a) the nucleotide sequence of the MADS + I + K domains, (b) the regions encoding the I + K domains and (c) the region encoding the MADS-domain. While the trees in (a) and (b) are similar, the trees using the MADS-domain deviate, with longer branch lengths leading to the *AGL13* homologs, suggesting that their MADS-domains are less constrained than those of *AGL6*. Branch support values (Bayesian posterior probabilities) <1.00 are shown above the branches (support of 1.00 is not indicated). Light gray boxes indicate *AGL6* homologs of Brassicaceae, while dark gray boxes indicate *AGL13* homologs of Brassicaceae. The white boxes indicate related MADS-domain proteins from *Arabidopsis thaliana*. In the insets, the average K_a/K_s values are presented with the standard deviation (*P*-values from a Welsh's *t*-test), and ω-values indicating the strength of positive selection (in the free-ratio branch model M1; Table S4) on three branches of each tree are shown.

by a silencer element located in the upstream region of *AGL6*. Additionally, both *AGL6* and *AGL13* have intron-localized elements that can modify the activity of the upstream-localized chalazal enhancer, although in different ways: in *AGL6* the intronic element maintains expression; while in *AGL13* the intronic element attenuates expression. The complexity of these regulatory interactions might explain why it has been difficult to recapitulate some of the enhancer detector expression patterns with *promoter:GUS* constructs containing only upstream regulatory sequences (R. Basker and U. Grossniklaus, University of Zürich, Switzerland, unpublished data).

While both the *in situ* RNA localization of *AGL6* and *AGL6$^{\Delta AUG}$:GUS* lines largely recapitulated the reporter gene expression of ET447 and ET8885, the reporter gene expression of ET5830 was very different from both the *in situ* RNA localization of *AGL13* and *AGL13$^{\Delta AUG}$:GUS*. For example, both *AGL13* RNA and *promoter:GUS* reporter activity was detected in the chalaza during ovulogenesis, while no reporter activity was detected in the ovules of line ET5830. These results are not surprising, as enhancer detector lines in *Drosophila* do not always reproduce all the expression pattern elements of a given detected gene (Bellen *et al.*, 1989), although this phenomenon has not yet been reported in plants.

The subfunctionalization of the *AGL6* subfamily in Brassicaceae

In general, *AGL6* and *AGL13* do not appear to have overlapping expression domains, suggesting the reciprocal loss of regulatory elements, which is consistent with subfunctionalization (Figures 2 and S10). While it is unclear how this ultimately affects the transcriptional regulatory functions of AGL6 and AGL13, change in the expression profiles of the

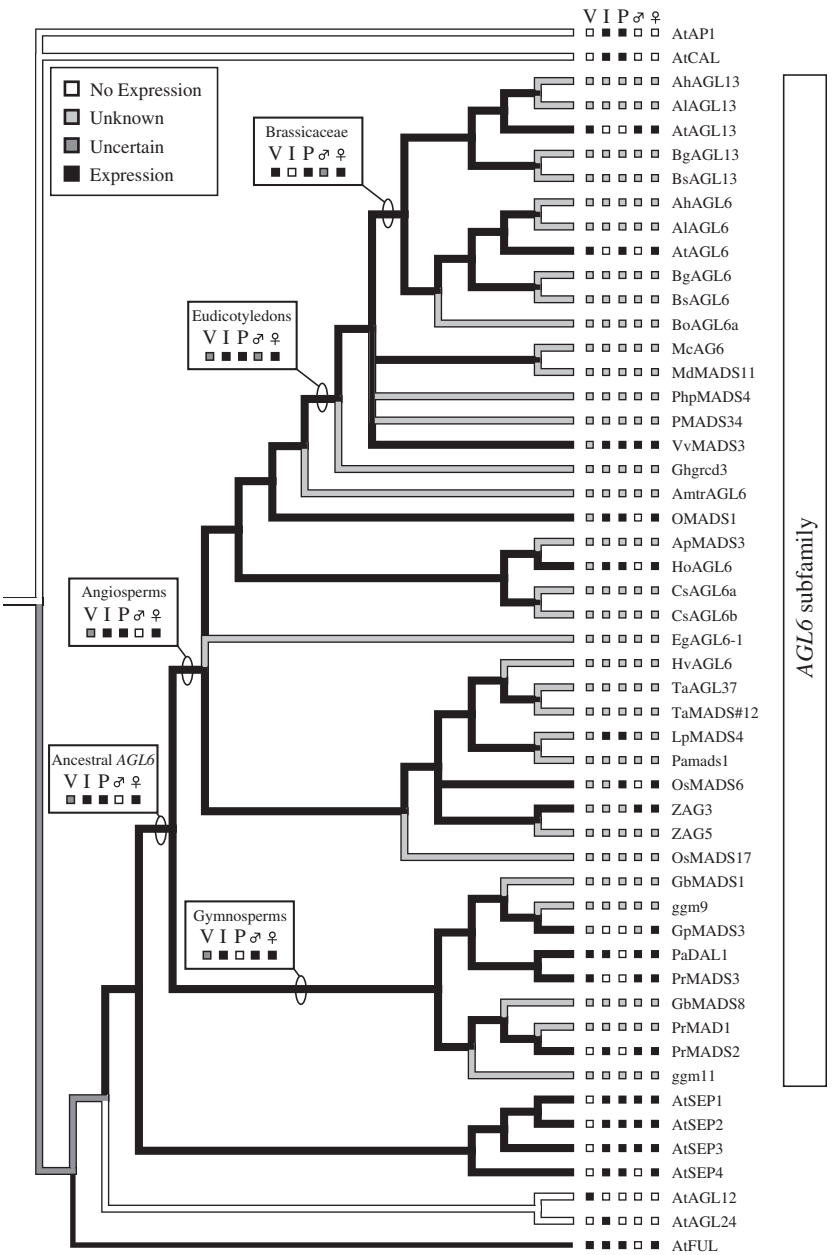

Figure 7. Expression pattern of *AGL6* subfamily members.
A summary of the expression patterns of *AGL6* subfamily plotted on phylogeny (Table S5), with the predicted ancestral states indicated. The expression in female reproductive tissues is indicated on the tree. The expression patterns are annotated as: V, vascular; I, inflorescence; P, perianth; ♂, male reproductive organ; ♀, female reproductive organ.

paralogs is a critical (and sufficient; Force *et al.*, 1999) step in subfunctionalization. The only place where *AGL6* and *AGL13* are co-expressed in the plant is the developing chalaza, suggesting that the spatiotemporal expression patterns of *AGL6* and *AGL13* have diverged. This suggests that the ancestral expression pattern changed during subfunctionalization in Arabidopsis: *AGL13* retained expression in the tapetum and had reduced expression in the ovule, while *AGL6* lost expression in the tapetum and retained expression in the ovule. Consistent with these observations, comparison of steady-state mRNA levels by microarray analysis as well as phylogenetic footprinting of the upstream regions also suggests subfunctionalization of the *AGL6* subfamily (de Bodt *et al.*, 2006; Duarte *et al.*, 2006). However, the mechanism for subfunctionalization does not appear to occur by the loss of regulatory elements but by attenuation due to the activity of *cis*-acting silencer elements, suppressing *AGL6* expression in the tapetum and partially suppressing *AGL13* expression in the chalaza. While this phenomenon appears to have occurred in only a minority of elements, it would provide an additional explanation for the reappearance of 'lost' regulatory elements over evolution (Locascio *et al.*, 2002), due to the gain, and then loss, of attenuating silencer elements. The gain of an attenuating silencer element should be a more infrequent event than the loss of an element by mutation, which might explain why the

process of subfunctionalization can take longer than predicted (compare Force *et al.*, 1999; Jarinova *et al.*, 2008). Interestingly, the expression of *AGL6* subfamily members in male reproductive tissues may be an example in a plant of 'lost' regulatory elements: while *AGL6* subfamily members are expressed in male reproductive tissues in the 'ancestral' gymnosperm, this expression domain was lost in the 'ancestral' angiosperm, only to reappear in eudicotyledons.

Phylogenetic shadowing revealed that the first intron of both *AGL6* and *AGL13* contained four regions of conservation possibly related to expression differences, although these regions were not conserved between *AGL6* and *AGL13*, under the assumption that the expression patterns of *AGL6* and *AGL13* orthologs are conserved in Brassicaceae. As most *AGL6* homologs have similar expression patterns (Figure 7), this assumption seems valid. While the lack of obvious sequence similarity among the first introns is consistent with subfunctionalization, it contrasts with the presence of regulatory elements, such as the tapetum enhancer, that appear to be conserved. It is possible that the regions around the attenuated tapetum-specific enhancer in *AGL6*, for example, could have diverged sufficiently from the tapetum-specific enhancer in *AGL13* that it is not identified in this analysis, yet retained its functional activity. Comparison with *AGL6* subfamily members outside of Brassicaceae, however, indicated sequence similarity to two regions of the intron of *AGL6* and one region of the intron of *AGL13*. This suggests that there has been a reciprocal loss of regulatory elements from intron 1 after gene duplication, consistent with subfunctionalization.

Diversification of the Brassicaceae *AGL6* subfamily

Comparison of the average K_a/K_s ratios for the Brassicaceae *AGL6* subfamily members indicated that both *AGL6* and *AGL13* are undergoing purifying selection, although the strength of selection was higher for *AGL6* than for *AGL13*. Interestingly, the largest difference between *AGL6* and *AGL13* was observed when the average K_a/K_s ratios for the MADS-domain were compared. The MADS-domain of *AGL6* is under strong purifying selection, while selection on the MADS-domain of *AGL13* is relaxed. This is consistent with the ω analysis for this domain, with $\omega_{b'}$ (for the branch leading to *AGL6*) consistent with purifying selection, which is not seen for ω_c (for the branch leading to *AGL13*). Whether this is reflected in different DNA-binding properties between *AGL6* and *AGL13* is unknown. Additionally, the high ω_a-value for the branch leading to the duplication implies that the MADS-domain was under positive selection prior to duplication. After the duplication event, evidence of positive selection is observed on the I + K domains in simple models (with either ω_b or ω_c variable and all other ω-values constant), although this is not observed in more complex models (with both ω_b and ω_c variable; Table S4). While the evolutionary forces driving the diversification of the Brassicaceae *AGL6* subfamily are complex, the lack of unambiguous evidence of positive selection post-duplication, which has been postulated as necessary for neofunctionalization (Force *et al.*, 1999), would argue against neofunctionalization.

Several lines of evidence support the hypothesis that the duplicated genes *AGL6* and *AGL13* are undergoing (or have undergone) subfunctionalization. However, like the *hoxb5a* and *hoxb5b* gene pair (Jarinova *et al.*, 2008), it appears that the situation is more complex than predicted by the DDC model, probably involving the attenuation of enhancer elements. Further research into the activity and nature of the regulatory elements driving the expression of other gene pairs in plants is necessary to establish whether this complexity is the rule rather than the exception.

EXPERIMENTAL PROCEDURES

In situ hybridization

In situ hybridization was performed as published in Golden *et al.* (2002) with probes described in Appendix S1. Briefly, due to the high sequence similarity between *AGL6* and *AGL13*, as well as very low transcript levels, sections were hybridized with a pooled set of probes consisting of small (~90 bp) regions of non-conservation 3′ of the K-domain and a slightly larger (~200 bp) region of the 3′-untranslated region (UTR).

Promoter–reporter gene constructs

To investigate both promoter and enhancer activities, a series of T-DNA vectors containing the *GUS* and/or *NAN* reporter gene(s) were made using new GATEWAY® (Invitrogen, http://www.invitrogen.com/) compatible vectors that have the 5′ *NOPALINE SYNTHASE* promoter of *Agrobacterium tumefaciens* (Shaw *et al.*, 1984) driving a plant-selectable maker (hygromycin phosphotransferase for *GUS* vectors, and phosphinotricin acetyl transferase for *NAN* vectors; see Appendix S1 for more details). These were used to generate a series of *GUS* reporter constructs of the *AGL6* regulatory regions. The first construct (*AGL6*$^{\Delta AUG}$*:GUS*) contained 3.0 kb of genomic sequence upstream of the translational start site fused with 0.6 kb intron 1 containing both the 5′ splice donor site of exon 1 and the 3′ splice acceptor site of exon 2. This construct lacks 157 bp of exon 1, removing AUG start codons from exon 1 that are both in and out of the reading frame, preventing the expression of a truncated AGL protein (containing only the MADS DNA-binding domain). The second construct (*AGL6*^{UP}*:GUS*) contains only the 3.0 kb of upstream genomic sequence. To detect the presence of enhancer elements that can work independently of upstream elements, a 0.8 kb fragment containing the first intron, second exon, and second intron was placed upstream of a *min35S* promoter driving the *GUS* reporter gene. As enhancer elements lack directionality, this fragment was cloned in both a sense and antisense orientation (*AGL6*$^{I1+I2s}$*:min35S:GUS* and *AGL6*$^{I1+I2as}$*:min35S:GUS*). The first construct (*AGL13*$^{\Delta AUG}$*:GUS*) contained 1.4 kb upstream and 0.7 kb of intron 1. Similar to *AGL6*$^{\Delta AUG}$*:GUS*, this construct lacks 177 bp of exon 1, removing AUG start codons that are both in and out of the reading frame. The second construct (*AGL13*^{UP}*:GUS*) contains only the 1.4 kb of upstream genomic sequence. As with *AGL6*, a 0.7 kb fragment containing the first intron of *AGL13* was placed upstream of a *min35S* driving the *GUS* reporter gene. Again, this fragment was cloned in both a sense and antisense orientation (*AGL13*^{I1s}*:min35S:GUS* and *AGL13*^{I1as}*:min35S:GUS*). The *promoter:NAN* construct (*AGL13*$^{\Delta AUG}$*:NAN*) contains the same promoter construct

as *AGL13*$^{\Delta AUG}$*:GUS* upstream of *NAN*. Construction of these vectors is detailed in Appendix S1. All T-DNA vectors were transformed into *A. tumefaciens* strain GV3101, and introduced into both Arabidopsis Landsberg *erecta* and Columbia accessions by floral dipping (Clough and Bent, 1998). Positive transformants were selected either on hygromycin (T-DNAs containing the *GUS* reporter) or Basta (T-DNAs containing the *NAN* reporter). An average of 13 independent primary transformants were examined for each transgenic line (a minimum of six), and no obvious differences were observed in the expression patterns between independent primary transformants.

GUS staining

For GUS staining, tissue was dissected (when necessary), immersed in GUS buffer [2.9 mM Na_2HPO_4, 2.1 mM NaH_2PO_4, 1 mM EDTA, 0.1% Triton X-100, 1 mM potassium hexacyanoferrate (II), 1 mM potassium hexacyanoferrate (III), 50 μg/μl chloramphenicol, 1 mM 5-bromo-4-chloro-3-indoxyl-β-D-glucuronic acid (X-Gluc; Biosynth, http://www.biosynth.com/)], vacuum infiltrated and incubated at 37°C for several hours, washed with 1× PBS (140 mM NaCl, 2.7 mM KCl, 10.1 mM Na_2HPO_4, 1.8 mM KH_2PO_4), and cleared for over an hour in clearing solution (1× PBS, 20% DL-lactic acid, 20% glycerol). This procedure was the same for the co-localization experiments, although using a slightly different staining buffer [2.9 mM Na_2HPO_4, 2.1 mM NaH_2PO_4, 1 mM EDTA, 0.1% Triton X-100, 2.5 mM potassium hexacyanoferrate (II), 2.5 mM potassium hexacyanoferrate (III), 50 μg/μl chloramphenicol, 1 mM 5-bromo-6-chloro-3-indoxyl-β-D-glucuronic acid (X-GlucM, GUS substrate; Glycosynth, http://www.glycosynth.co.uk/), 0.5 mM 5-bromo-4-chloro-3-indolyl α-D-*N*-acetylneuraminic acid (X-NeuNAc, NAN substrate; Biosynth)].

Phylogenetic shadowing and reconstructions

For identification of conserved regions, the nucleotide sequence of the first introns of the *AGL6* and *AGL13* homologs from various Brassicaceae (primer sequences are described in Table S2) were aligned separately using ClustalX (Thompson *et al.*, 1997; http://www.clustal.org/) and adjusted manually in BioEdit 7 (Hall, 1999; http://www.mbio.ncsu.edu/BioEdit/bioedit.html/; Figures S3 and S4). Analyses were performed using sequence mismatch and, to account for the potential effects of phylogeny on each analysis, maximum parsimony (MP) and maximum likelihood (ML). Average pairwise sequence mismatch was calculated for sliding windows of 20 nucleotides in length along the intron sequence alignment with a step size of one. This was done either excluding or including nucleotide/gap character comparisons for mismatch calculation, using a program written for this purpose. The MP and ML analyses were conducted in PAUP* 4.0b10 (Swofford, 2002; http://paup.csit.fsu.edu/), where ML analyses used the model of nucleotide evolution selected by Modeltest 3.06 (Posada and Crandall, 1998; http://darwin.uvigo.es/software/modeltest.html/) under the Akaike information criterion (AIC; TrN + I model in every case). First, a phylogeny was reconstructed for the first intron. Then, tree scores (tree length for MP, – ln likelihood for ML) for this phylogeny were calculated given the data from each 20-nucleotide-long window. A null distribution for expected mismatch values and tree scores was obtained by conducting the same analyses on 1000 independent 20-nucleotide windows taken from the randomly resampled (bootstrapped) input sequence. The null distribution was used for calculation of mean, standard deviation and 99% confidence interval on mismatch values and tree scores in Microsoft Excel 2007. *AGL6* and *AGL13* tree scores and mismatch were plotted along the sequence. Sequence windows consistently scoring less than the lower 99% confidence boundary in all analyses (and <15% gap characters) were considered to be conserved. Alignments and all software written for these analyses are freely available from http://bot-serv1.uzh.ch/home/grossnik/software/AGL/AGL_index.html.

Comparison of *AGL6* and *AGL13* intron sequences with those of the more distantly related *PMADS34* and *VvMADS3* was done by dot-plot analysis because reliable sequence alignments were not possible. Dot-plot analysis was performed with a window size of 20, allowing eight mismatches, and a step size of one. Consecutive stretches of matching windows, represented by diagonals in the dot-plot, were counted for the entire dot-plot and separately for areas that were considered conserved for *AGL6* or *AGL13*. For every diagonal length class the sum of diagonals (or diagonal fractions) in a given area was then scaled by this area to give a diagonal density. Null distributions were derived from randomly resampled input sequences (keeping their relative length proportions) and a minimum of 500 000 diagonals. Statistical difference among expected and observed distributions of diagonal length classes was assessed by Kolmogorov–Smirnov (KS) tests using R 2.7.2 (R Development Core Team, 2008; http://www.r-project.org/).

Phylogenetic analyses were performed both on amino acid and coding nucleotide sequences (Table S3). Initial sequence alignments were produced using ClustalX (Thompson *et al.*, 1997), and the nucleotide sequence alignment adjusted manually in BioEdit 7 (Hall, 1999), using the amino acid alignment as a guide (Figure S8). Amino acid sequences were subjected to a MP analysis in PAUP* 4.0b10 (Swofford, 2002) using a heuristic search with tree bisection–reconnection swapping and 10 random sequence addition replications, with the MulTrees option and a MaxTrees limit of one million in effect. Bootstrap analyses with 1000 pseudo-replications each were conducted using heuristic searches with the same settings as above, except that the number of sequence-addition replicates was lowered to three. For analysis of the MADS-domain only, it was necessary to perform the analysis without sequence addition replication and with a lower MaxTrees limit of 5000 enforced. Bayesian inference (BI) was used for analyses based on nucleotide sequences, using MrBayes 3.1.2 (Ronquist and Huelsenbeck, 2003; http://www.mrbayes.csit.fsu.edu/) and models of nucleotide evolution selected by MrModeltest 2.3 (Nylander, 2004; http://www.abc.se/~nylander/) under the AIC (GTR + I + Γ for both MADS and I + K domains). For each BI analysis, Markov chain Monte Carlo (MCMC) runs (four chains each) were performed, sampling one tree every 1000th generation, for 10 million generations, after which the chains seemed to have converged, as judged by inspection of interchain distances and the apparent stationarity of log-likelihood scores. In every analysis, the first 500 of the trees retained were discarded as a 'burn-in'. Both MP and BI were performed on MADS and I + K domains separately. In addition, the entire amino acid sequence was analyzed using MP and the MADS + I + K nucleotide sequence with BI. The C-domain was not analyzed due to high sequence divergence.

K_a/K_s ratios for *AGL6* subfamily members from Brassicaceae were calculated using DnaSP version 4.50 (Rozas *et al.*, 2003; http://www.ub.edu/dnasp/). Mean and variance were calculated for *AGL6* and *AGL13* homologs and significance tested using a Welsh's *t*-test. Maximum likelihood-based codon substitution analyses were performed using the codeml program of PAML 4.1 (Yang, 2007; http://abacus.gene.ucl.ac.uk/software/paml.html/), with site, branch and branch-site models using trees from Figure 6. Likelihood ratio tests were calculated as twice the difference in log likelihood for different models and compared against a χ^2-distrubution with degrees of freedom as the difference in free model parameters. Ancestral expression patterns were reconstructed by parsimony in Mesquite 2.6 (Maddison and Maddison, 2009; http://mesquiteproject.org/mesquite/mesquite.

html/), using the phylogeny from Figure 6(a) and coding expression as a 'standard' character.

ACKNOWLEDGEMENTS

This work was supported by the University of Zürich, a NSF International Research Fellowship to SES (INT-0301399); a Schrödinger Fellowship from the Austrian Science Fund (FWF) to PMS (J2678-B16); grants from the Swiss National Science Foundation to MDC (3100A0-100281) and UG (31-112489/1); and a grant from the USDA to UG (98-35304-6412). Seeds for ET8885 and ET5830 were obtained from the Cold Spring Harbor Laboratory Collection. We would like to thank Tony Kavanagh for the pUC19-NAN vector, Juan-Miguel Escobar-Restrepo and Anja Schmidt for plant material/DNA, Kentaro Shimizu and Takashi Tsuchimatsu for comments and suggestions on phylogenetic analyses, and Berit Rosc-Schlüter and Monica Grobei for comments on the manuscript.

SUPPORTING INFORMATION

The following supporting information is available for the article online:

Figure S1. Expression patterns of enhancer trap (ET) lines in reproductive tissues.

Figure S2. A cartoon of *AGL6* and *AGL13 Promoter:GUS* constructs.

Figure S3. Expression of *AGL6* and *AGL13 Promoter:GUS* lines in 12-day-old seedlings.

Figure S4. Alignment of intron 1 of *AGL6* homologs from Brassicaceae.

Figure S5. Alignment of intron 1 of *AGL13* homologs from Brassicaceae.

Figure S6. Regions of conservation inside the first introns of *AGL6* and *AGL13*.

Figure S7. Regions of conservation inside the first intron between *Vitis VvMADS3* and Arabidopsis *AGL6* subfamily members.

Figure S8. Alignment of amino acid sequence of *AGL6* subfamily members.

Figure S9. Maximum parsimony analysis of the *AGL6* subfamily.

Figure S10. Partial summary of observed expression patterns.

Table S1. Conserved predicted transcription factor-binding sites inside intron 1.

Table S2. Primers used to amplify *AGL6* and *AGL13* homologs from Brassicaceae.

Table S3. GenBank accession numbers of *AGL6* subfamily members.

Table S4. d_N/d_S-based maximum likelihood analysis of positive selection.

Table S5. Expression pattern of *AGL6* subfamily members.

Appendix S1. Detailed descriptions of the construction of both the *Promoter:GUS* and *Promoter:NAN* vectors and all promoter constructs used.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

REFERENCES

Becker, A. and Theißen, G. (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylogenet. Evol.* **29**, 464–489.

Bellen, H.J. (1999) Ten years of enhancer detection: lessons from the fly. *Plant Cell*, **11**, 2271–2281.

Bellen, H.J., O'Kane, C.J., Wilson, C., Grossniklaus, U., Pearson, R.K. and Gehring, W.J. (1989) P-element-mediated enhancer detection: a versatile method to study development in *Drosophila*. *Genes Dev.* **3**, 1288–1300.

Benfey, P.N. and Chua, N.H. (1990) The cauliflower mosaic virus 35S promoter: combinatorial regulation of transcription in plants. *Science*, **250**, 959–966.

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

de Bodt, S., Theißen, G. and van de Peer, Y. (2006) Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.* **23**, 1293–1303.

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

Carlsbecker, A., Tandre, K., Johanson, U., Englund, M. and Engstrom, P. (2004) The MADS-box gene *DAL1* is a potential mediator of the juvenile-to-adult transition in Norway spruce (*Picea abies*). *Plant J.* **40**, 546–557.

Causier, B., Castillo, R., Zhou, J., Ingram, R., Xue, Y., Schwarz-Sommer, Z. and Davies, B. (2005) Evolution in action: following function in duplicated floral homeotic genes. *Curr. Biol.* **15**, 1508–1512.

Causier, B., Bradley, D., Cook, H. and Davies, B. (2008) Conserved intragenic elements were critical for the evolution of the floral C-function. *Plant J.* **58**, 41–52.

Clough, S.J. and Bent, A.F. (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743.

Duarte, J.M., Cui, L., Wall, P.K., Zhang, Q., Zhang, X., Leebens-Mack, J., Ma, H., Altman, N. and dePamphilis, C.W. (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. *Mol. Biol. Evol.* **23**, 469–478.

Fan, H.Y., Hu, Y., Tudor, M. and Ma, H. (1997) Specific interactions between the K domains of AG and AGLs, members of the MADS domain family of DNA binding proteins. *Plant J.* **12**, 999–1010.

de Folter, S., Shchennikova, A.V., Franken, J., Busscher, M., Baskar, R., Grossniklaus, U., Angenent, G.C. and Immink, R.G. (2006) A Bsister MADS-box gene involved in ovule and seed development in petunia and Arabidopsis. *Plant J.* **47**, 934–946.

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

Galuschka, C., Schindler, M., Bulow, L. and Hehl, R. (2007) AthaMap web tools for the analysis and identification of co-regulated genes. *Nucleic Acids Res.* **35**, D857–D862.

Golden, T.A., Schauer, S.E., Lang, J.D., Pien, S., Mushegian, A.R., Grossniklaus, U., Meinke, D.W. and Ray, A. (2002) *SHORT INTEGUMENTS1/SUSPENSOR1/CARPEL FACTORY*, a Dicer homolog, is a maternal effect gene required for embryo development in Arabidopsis. *Plant Physiol.* **130**, 808–822.

Grossniklaus, U., Bellen, H.J., Wilson, C. and Gehring, W.J. (1989) P-element-mediated enhancer detection applied to the study of oogenesis in *Drosophila*. *Development*, **107**, 189–200.

Hall, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95–98.

Heck, G.R., Perry, S.E., Nichols, K.W. and Fernandez, D.E. (1995) AGL15, a MADS domain protein expressed in developing embryos. *Plant Cell*, **7**, 1271–1282.

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

Jack, T., Brockman, L.L. and Meyerowitz, E.M. (1992) The homeotic gene *APETALA3* of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens. *Cell*, **68**, 683–697.

Jarinova, O., Hatch, G., Poitras, L., Prudhomme, C., Grzyb, M., Aubin, J., Berube-Simard, F.A., Jeannotte, L. and Ekker, M. (2008) Functional resolution of duplicated *hoxb5* genes in teleosts. *Development*, **135**, 3543–3553.

Jefferson, R.A. (1989) The *GUS* reporter gene system. *Nature*, **342**, 837–838.

Kim, M.J., Kim, H., Shin, J.S., Chung, C.H., Ohlrogge, J.B. and Suh, M.C. (2006) Seed-specific expression of sesame microsomal oleic acid desaturase is controlled by combinatorial properties between negative *cis*-regulatory elements in the *SeFAD2* promoter and enhancers in the 5′-UTR intron. *Mol. Genet. Genomics*, **276**, 351–368.

Kirby, J. and Kavanagh, T.A. (2002) *NAN* fusions: a synthetic sialidase reporter gene as a sensitive and versatile partner for *GUS*. *Plant J.* **32**, 391–400.
Kleinjan, D.A. and van Heyningen, V. (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32.
Kleinjan, D.A., Bancewicz, R.M., Gautier, P. *et al.* (2008) Subfunctionalization of duplicated zebrafish *pax6* genes by *cis*-regulatory divergence. *PLoS Genet.* **4**, e29.
Klimyuk, V.I., Nussaume, L., Harrison, K. and Jones, J.D. (1995) Novel *GUS* expression patterns following transposition of an enhancer trap *Ds* element in Arabidopsis. *Mol. Gen. Genet.* **249**, 357–365.
Kooiker, M., Airoldi, C.A., Losa, A., Manzotti, P.S., Finzi, L., Kater, M.M. and Colombo, L. (2005) *BASIC PENTACYSTEINE1*, a GA binding protein that induces conformational changes in the regulatory region of the homeotic Arabidopsis gene *SEEDSTICK*. *Plant Cell*, **17**, 722–729.
Li, W.H., Yang, J. and Gu, X. (2005) Expression divergence between duplicate genes. *Trends Genet.* **21**, 602–607.
Locascio, A., Manzanares, M., Blanco, M.J. and Nieto, M.A. (2002) Modularity and reshuffling of *Snail* and *Slug* expression during vertebrate evolution. *Proc. Natl Acad. Sci. USA*, **99**, 16841–16846.
Ma, H., Yanofsky, M.F. and Meyerowitz, E.M. (1991) *AGL1-AGL6*, an Arabidopsis gene family with similarity to floral homeotic and transcription factor genes. *Genes Dev.* **5**, 484–495.
Maddison, W.P. and Maddison, D.R. (2009) *Mesquite: a modular system for evolutionary analysis*. Version 2.6. http://mesquiteproject.org.
Matys, V., Kel-Margoulis, O.V., Fricke, E. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110.
Meinke, D.W. and Sussex, I.M. (1979) Embryo-lethal mutants of *Arabidopsis thaliana*. A model system for genetic analysis of plant embryo development. *Dev. Biol.* **72**, 50–61.
Mouradov, A., Glassick, T.V., Hamdorf, B.A., Murphy, L.C., Marla, S.S., Yang, Y. and Teasdale, R.D. (1998) Family of MADS-Box genes expressed early in male and female reproductive structures of monterey pine. *Plant Physiol.* **117**, 55–62.
Nylander, J.A.A. (2004) *MrModeltest v2*. Program distributed by the author. Uppsala, Sweden: Evolutionary Biology Centre, Uppsala University.
Posada, D. and Crandall, K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.
R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation of Statistical Computing: Vienna, Austria.
Reddy, V.S. and Reddy, A.S. (2004) Developmental and cell-specific expression of *ZWICHEL* is regulated by the intron and exon sequences of its gene. *Plant Mol. Biol.* **54**, 273–293.
Rijpkema, A.S., Gerats, T. and Vandenbussche, M. (2007) Evolutionary complexity of MADS complexes. *Curr. Opin. Plant Biol.* **10**, 32–38.
Ronquist, F. and Huelsenbeck, J.P. (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.
Rose, A.B., Elfersi, T., Parra, G. and Korf, I. (2008) Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. *Plant Cell*, **20**, 543–551.
Rounsley, S.D., Ditta, G.S. and Yanofsky, M.F. (1995) Diverse roles for MADS box genes in Arabidopsis development. *Plant Cell*, **7**, 1259–1269.
Rozas, J., Sánchez-DelBarro, J.C., Messeguer, X. and Rozas, R. (2003) DnaSP, DNA polymorphism analysis by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497.
Schneitz, K., Hulskamp, M. and Pruitt, R.E. (1995) Wild-type ovule development in *Arabidopsis thaliana* – a light-microscope study of cleared whole-mount tissue. *Plant J.* **7**, 731–749.
Searle, I., He, Y., Turck, F., Vincent, C., Fornara, F., Krober, S., Amasino, R.A. and Coupland, G. (2006) The transcription factor *FLC* confers a flowering response to vernalization by repressing meristem competence and systemic signaling in Arabidopsis. *Genes Dev.* **20**, 898–912.
Shaw, C.H., Carter, G.H. and Watson, M.D. (1984) A functional map of the nopaline synthase promoter. *Nucleic Acids Res.* **12**, 7831–7846.
Sheldon, C.C., Conn, A.B., Dennis, E.S. and Peacock, W.J. (2002) Different regulatory regions are required for the vernalization-induced repression of *FLOWERING LOCUS C* and for the epigenetic maintenance of repression. *Plant Cell*, **14**, 2527–2537.
Shindo, S., Ito, M., Ueda, K., Kato, M. and Hasebe, M. (1999) Characterization of MADS genes in the gymnosperm *Gnetum parvifolium* and its implication on the evolution of reproductive organs in seed plants. *Evol. Dev.* **1**, 180–190.
Sieburth, L.E. and Meyerowitz, E.M. (1997) Molecular dissection of the *AGAMOUS* control region shows that *cis* elements for spatial regulation are located intragenically. *Plant Cell*, **9**, 355–365.
Smyth, D.R., Bowman, J.L. and Meyerowitz, E.M. (1990) Early flower development in Arabidopsis. *Plant Cell*, **2**, 755–767.
Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H. and Martienssen, R. (1995) Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes Dev.* **9**, 1797–1810.
Swofford, D.L. (2002) *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*. Version 4. Sunderland, MA, USA: Sinauer Associates.
The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882.
Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.
Yang, W., Jefferson, R.A., Huttner, E., Moore, J.M., Gagliano, W.B. and Grossniklaus, U. (2005) An egg apparatus-specific enhancer of Arabidopsis, identified by enhancer detection. *Plant Physiol.* **139**, 1421–1432.

Accession numbers for the sequence data: GQ117273, GQ117274, GQ117275, GQ117276, GQ117277, GQ117278, GQ117279, GQ-117280.

