Complex Rearrangements Lead to Novel Chimeric Gene Fusion Polymorphisms at the *Arabidopsis thaliana MAF2-5* Flowering Time Gene Cluster

Ana L. Caicedo, Christina Richards,† Ian M. Ehrenreich,†‡ and Michael D. Purugganan†*

*Biology Department, 221 Morrill Science Center, University of Massachusetts; †Department of Biology, Center for Genomics and Systems Biology, New York University; and ‡Department of Genetics, North Carolina State University

Tandem gene clusters of multigene families are rearrangement hotspots and may be a major source of novel gene formation. Here, we report on a molecular population genetic analysis of the *MAF2-5* gene cluster of the model plant species, *Arabidopsis thaliana*. The *MAF2-5* genes are a MADS-box multigene family cluster spanning ∼24 kbp on chromosome 5. We find heterogeneous evolutionary dynamics among these genes, all of which are closely related to the floral repressor, *FLC*, and are believed to play a role in the control of flowering time in *A. thaliana*. Low levels of nonsynonymous single nucleotide polymorphism (SNP) observed for *MAF4* and *MAF5* suggest purifying selection and conservation of function. In contrast, high levels of nonsynonymous SNPs, insertion–deletion, and rearrangements are observed for *MAF2* and *MAF3*, including novel gene fusions that persist as a moderate-frequency polymorphism in *A. thaliana*. These fused genes, involving *MAF2* and portions of *MAF3*, are expressed, resulting in the production of chimeric, alternatively spliced transcripts of *MAF2*. Association studies support a correlation between the described *MAF2–MAF3* gene rearrangements and flowering time variation in the species. The finding that complex rearrangements within gene clusters, such as those observed for *MAF2*, might play a role in the generation of ecologically important phenotypic variation, emphasize the need for emerging high throughput genotyping and sequencing techniques to correctly reconstruct gene chimeras and other complex polymorphisms.

Introduction

The classical model of gene evolution proposes that new genes arise, in large part, through duplication of existing loci or genomic regions, which diverge to acquire new functions (Ohno 1970). Established examples of gene origin through duplication include the α and β chains of hemoglobin (Itano 1957) and primate opsin-encoding genes (Yokoyama S and Yokoyama R 1989). Access to extensive molecular data in the 1990s revealed many other molecular mechanisms contributing to the origin of genes, such as gene fusion and fission, retroposition, and lateral gene transfer (reviewed in Long et al. 2003). That these mechanisms could give rise to genes in recent timescales was revealed by description of the first young gene, *Jingwei*, a chimeric gene (i.e., containing portions of multiple genes) found in African *Drosophila* species, which arose through a combination of exon shuffling, retroposition, and gene duplication (Long and Langley 1993; Wang et al. 2000). Gene duplication and many of the other genomic mechanisms associated with the rise of new genes can be greatly accelerated in tandem gene clusters of multigene families. The homologous relationships among gene family members facilitate recombination, giving rise to further duplications and/or fusion of genes. As a result, tandem gene clusters are rearrangement hotspots and may be a major source of novel gene formation.

A particularly important and well-studied gene family within plants is the MADS-box family, characterized by a conserved ∼180 base pair (bp) region that encodes a DNA-binding domain (Schwarz-Sommer et al. 1990; Purugganan et al.1995, 1997). MADS-box genes encode transcription factors, whose members play roles in control of development and signal transduction in all eukaryotes (Theissen et al. 1996; Ng and Yanofsky 2001; Becker and Theissen 2003). MADS-box genes have been found to be involved in various plant developmental processes, such as determination of organ identity (Theissen 2001), flowering time (e.g., Lee et al. 2000), and fruit and seed development (e.g., Liljegren et al. 2000). Because of their regulatory roles, polymorphism in MADS-box genes and birth of novel MADS-box loci have the potential to greatly impact developmental and morphological variation within and between plant species. Here, we report on a novel chimeric MADS-box gene that is segregating at the population level in the model plant species *Arabidopsis thaliana*. The gene belongs to the *MAF* MADS-box subfamily (*MAF*, MADS Affecting Flowering), in which some members occur as a multigene tandem cluster, underscoring how the dynamics of multigene families can give rise to novel genes.

The *MAF* genes are a set of six genes in *A. thaliana* that belong to a monophyletic clade within the MADS-box gene family (Alvarez-Buylla et al. 2000). This clade includes the well-characterized *FLC* (Michaels and Amasino 1999; Sheldon et al. 1999) and *FLM* (Scortecci et al. 2001) genes, the latter first described as *MAF1* (Ratcliffe et al. 2001) and four genes referred to as *MAF2-5* (AT5G65050, AT5G65060, AT5G65070, and AT5G65080). Although *FLC* and *FLM* have distinct genomic locations (top of chromosome 5 and bottom of chromosome 1, respectively), the *MAF2-5* genes occur as a tandem gene cluster spanning ∼24 kbp on the bottom of chromosome 5 (Ratcliffe et al. 2001), with individual genes separated by less than 1.5 kbp of intergenic spacers (fig. 1*A*). These four genes are highly similar, sharing 76–91% amino acid (aa) identity within the MADS DNA-binding domain (Ratcliffe et al. 2001).

Genes belonging to the *MAF* gene family have been found to play a role in regulation of flowering time in *A. thaliana*. *FLC* represses flowering by downregulating the expression of the floral pathway integrators, *FT* and *SOC1* (Lee et al. 2000; Hepworth et al. 2002; Rouse et al. 2002). *FLC* expression, in turn, is downregulated

Key words: alternative splicing, gene origin, gene fusion, MADS-box, quantitative trait loci.

E-mail: mp132@nyu.edu.

Mol. Biol. Evol. 26(3):699–711. 2009
doi:10.1093/molbev/msn300
Advance Access publication January 12, 2009

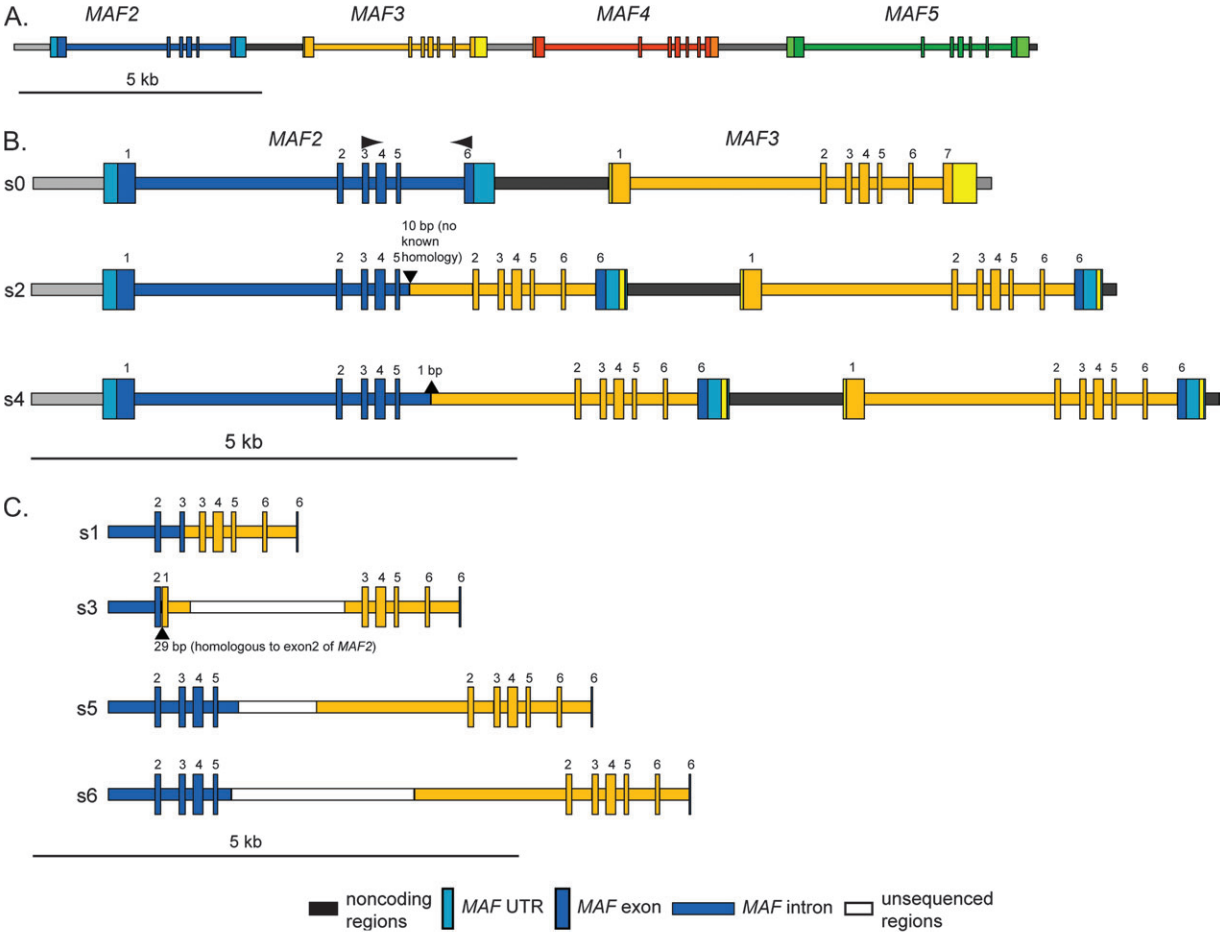

FIG. 1.—Gene models for the sequenced *MAF2-5* genomic region. (*A*) Gene model based on Col-0 genome sequence. Note that multiple alternative splice variants have been reported for all *MAF2-5* genes; the gene models used in this diagram correspond to TAIR gene models AT5G65050.1 for *MAF2*, AT5G65060.1 for *MAF3*, AT5G65070.1 for *MAF4*, and AT5G65080.1 for *MAF5*. (*B*) Gene models of accessions with large inserts in *MAF2*, for which the *MAF2-5* genomic region was sequenced in its entirety, compared with model of sequences lacking a *MAF2* insert. For simplicity, only the regions pertaining to *MAF2* and *MAF3* are shown in the model. Colors denote gene homology, and numbers identify exons of highest homology in the pertaining gene. Note that the *MAF2* region homologous to exon 6 of *MAF3* is not always transcribed and is not represented in our model. "s" codes denote insert type, with s0 denoting absence of insert. Arrows represent placement of primers used for amplification of cDNA. (*C*) Gene models for additional *MAF2* insert types sequenced. Only the region amplified and sequenced for each type is shown. Colors and numbers denote regions of highest homology in *MAF2* and *MAF3*.

by vernalization, a prolonged exposure to cold, providing a mechanism to ensure flowering after an extended winter period (Michaels and Amasino 1999; Sheldon et al. 1999). *FLM* has also been shown to act as a floral repressor, but, unlike *FLC*, is not affected by vernalization (Ratcliffe et al. 2001; Scortecci et al. 2001). Given their sequence similarity to *FLC* and *FLM*, the *MAF2-5* genes on chromosome 5 are suspected to be involved in the regulation of flowering time, and *MAF2* has been shown to encode a floral repressor that may prevent a vernalization response under short periods (10–21 days) of cold (Ratcliffe et al. 2003). The roles of the other *MAF* genes remain to be determined; however, *MAF3* and *MAF4* expression is downregulated, whereas *MAF5* is upregulated, by exposure to cold treatment (Ratcliffe et al. 2003), suggesting that all genes may be involved in modifications of the vernalization response. All genes in the *MAF* clade have also been shown to be subject to extensive alternative splicing (Ratcliffe et al. 2001, 2003; Scortecci et al. 2001; Caicedo et al. 2004), possibly giving rise to multiple protein products with different functional roles within single plants.

Here, we report on a molecular population genetic analysis of the *MAF2-5* cluster in *A. thaliana*, which reveals several complex rearrangements in this genomic region occurring in the species. These rearrangements result in the generation of novel gene fusions involving *MAF2* and *MAF3* that persists as a moderate-frequency polymorphism in *A. thaliana*. These fused genes are expressed, resulting in the production of chimeric, alternatively spliced transcripts. Various recent quantitative trait loci (QTL) mapping analyses suggest that a locus affecting natural variation in flowering time is also localized to a genomic region spanning the *MAF2-5* gene cluster at the bottom of chromosome 5 (Ungerer et al. 2002, 2003; El-Lithy et al. 2004, 2006; Simon et al. 2008), and we show that association studies support a correlation between several of the described *MAF2–MAF3* gene rearrangements and flowering time variation in the species. These results suggest that the *MAF2-5* genes play a role in regulation of flowering time in *A. thaliana* and that tandem gene clusters can serve as generators of novel gene fusions that may lead to the formation of complex loci and/or inactivation of genes that can affect ecologically relevant traits.

Materials and Methods

DNA Sequencing and Molecular Population Genetic Analyses

A set of 16 *A. thaliana* ecotypes was obtained from the Arabidopsis Biological Resource Center (ABRC) (see supplementary table 1, Supplementary Material online). Genomic DNA was isolated from young leaves using a Plant DNeasy Mini kit (Qiagen, Valencia, CA). Previous QTL mapping analyses in the laboratory using L*er*-2 × Cvi-0 recombinant inbred lines (RILs) suggest that natural variation in flowering time between these two ecotypes is due, in part, to polymorphism in the genomic region spanning the *MAF2-5* cluster (Engelmann K, Ungerer M, Purugganan MD, unpublished data, Ungerer et al. 2002, 2003); thus, DNA from 1.2.5QZ, a fine-mapping line containing the L*er*-2 *MAF2-5* genomic region in a Cvi-0 background, was also included in the sample set. For all accessions, the complete ~24 kbp *MAF2-5* genomic region on chromosome 5 was amplified in overlapping fragments using Ex-Taq DNA polymerase (Takara, Otsu, Japan). The amplified region includes the open reading frames of *MAF2*, *MAF3*, *MAF4*, and *MAF5*, all intergenic regions, a ~900-bp region upstream of *MAF2* and a ~600-bp region downstream of *MAF5*. Polymerase chain reaction (PCR) products were purified using QIAquick gel extraction kits (Qiagen) and sequenced directly with Big Dye Terminator cycle sequencing kits (Applied Biosystems, Foster City, CA). Primers used for PCR amplification as well as internal primers used for DNA sequencing are listed in supplementary table 2, Supplementary Material online. Sequencing was carried out at the NC State Genome Research Laboratory with an ABI Prism 3700 automated sequencer (Applied Biosystems).

All sequences were aligned and edited against the published Col-0 genomic sequence (http://www.arabidopsis.org/), using the Phred and Phrap programs (Codon Code, Dedham, MA) and BioLign Version 2.09.1 (Tom Hall, NC State University). Estimates of polymorphism, nucleotide diversity, and Tajima's *D* for the *MAF2-5* region were carried out using DnaSP 4.10 (Rozas J and Rozas R 1999). Estimates of synonymous and nonsynonymous nucleotide diversity (π_a/π_s), with missing data, including estimates of bootstrapped data sets, were calculated using libsequence 1.6.5 (Thornton 2003). Relationships among *MAF* genes were obtained from maximum parsimony analyses carried out in MEGA 4 (Tamura et al. 2007), and analyses of branch specific d*N*/d*S* estimates were carried out as implemented in PAML (Yang 1997). The possibility of gene conversion among *MAF* genes was explored with the program GENECONV (Sawyer 1999); analyses were carried out with no mismatches allowed, and global *P* values were assigned based on 10,000 permutations and corrected for multiple comparisons. Linkage disequilibrium (LD) estimates among selected polymorphisms of the *MAF* genes were obtained with Fisher's Exact Tests.

Expression Analyses

RNA was extracted from selected accessions using an RNAEasy kit (Qiagen) and treated with DNA-free DNase (Ambion, Austin, TX). Complementary DNA (cDNA) was synthesized from approximately 1 μg of total RNA with the Retroscript reverse transcription kit (Ambion). Reverse-transcription PCR (RT-PCR) was performed using primers that amplify *MAF2* transcription products using standard PCR amplification protocols (supplementary table 2, Supplementary Material online). Products were run in ethidium-stained 2% agarose gels, purified using QIAquick gel extraction kits (Qiagen), and sequenced directly with Big Dye Terminator cycle sequencing kits (Applied Biosystems) or cloned into plasmid vectors using the TA Cloning Kit (with pCR 2.1 vector) with One Shot TOP10 Chemically Competent *Escherichia coli* (Invitrogen, Carlsbad, CA). Sequencing was carried out using an ABI 3100 automated sequencer at the NYU Center for Genomics and Systems Biology.

Genotyping and Association Tests

The entire sequenced *MAF2-5* genomic region was examined for single nucleotide polymorphisms (SNPs) or indel polymorphisms that could potentially affect *MAF* gene function. Putative functionally important polymorphisms were genotyped in a panel of 169 ecotypes (supplementary table 3, Supplementary Material online). Selected polymorphisms were genotyped either by 1) direct sequencing of short PCR fragments, 2) comparison of lengths of PCR products during electrophoresis, 3) differential amplification by high-specificity primers, or 4) the derived cleaved amplified polymorphisms (dCAPs) method (Neff et al. 2002). Four polymorphisms occurring at moderate frequency in the genotyped sample (>10%) were chosen to carry out association analyses with flowering time: 1) the presence or absence of a large insert in the 3′ region of *MAF2*, 2) a nonsynonymous SNP in exon 1 of *MAF3*, 3) a nonsynonymous SNP in exon 7 of *MAF3*, and 4) three linked SNPs in exon 7 of *MAF3* (supplementary tables 2 and 3, Supplementary Material online).

Flowering time and rosette leaf number (RLN) under long-day and short-day growth chamber conditions were obtained for each accession as previously described (Olsen et al. 2004). Association analysis of flowering time and RLN with genotype were conducted using a mixed linear model (MLM) as implemented in Tassel (Bradbury et al. 2005; Yu et al. 2006). Population structure in our accession sample was assessed using the program *structure* (Pritchard et al. 2000) based on previously published SNP markers (Schmid et al. 2006); the highest likelihood score was obtained for a model with five populations as published in Korves et al. (2007). The relative kinship matrix for our sample using these SNP markers was obtained using the software package SPAGeDi 1.2 (Hardy and Vekemans 2002).

Results

Genetic Variation in the *MAF2-5* Gene Cluster

Our previous QTL mapping results suggest that the variation in flowering time between the Cvi-0 and L*er*-2

A. thaliana accessions under long-day and short-day conditions may be due, in part, to polymorphism in a QTL that spans the *MAF2-5* genomic region on the bottom of chromosome 5 (Engelmann K, Ungerer M, Purugganan MD, unpublished data, Ungerer et al. 2002, 2003). Similar QTL results based on other crosses (L*er* × Kond, L*er* × Sha, L*er* × Kas-2, Blh × Col, Ct × Col, Cvi × Col, and Sha × Col) (El-Lithy et al. 2004, 2006; Simon et al. 2008) have also recently been reported. We sequenced the entire *MAF2-5* gene cluster in Cvi-0 (CS902) and the fine-mapping line 1.2.5QZ, which contains the L*er*-2 *MAF2-5* region in a Cvi-0 genomic background (Engelmann K, Ungerer M, Purugganan MD, unpublished data), to identify possible functional polymorphisms between these two lines.

Excluding gaps and missing data, we aligned nearly 19 kbp of sequence encompassing the *MAF2-5* gene cluster (fig. 1*A*). A total of 63 SNPs were found between the two accessions. Of these, three SNPs lead to amino acid differences between *MAF2-5* gene products in these two accessions: 1) a C to G substitution in the first exon of Cvi-0 *MAF2* (position 1029 in the *MAF2-5* alignment; PopSet Genbank records: EU980626 EU980614) leading to a conservative L to V replacement, 2) a T to A substitution in exon 4 of the Cvi-0 *MAF2* gene (position 3657), leading to a radical V to E amino acid change, and 3) a G to A substitution in exon 7 of Cvi-0 *MAF3* (position 12419), leading to a conservative V to M amino acid change. Additionally, Cvi-0 was found to contain a 146-bp insertion in the first exon of *MAF3* (position 8915 in the alignment) leading to a frameshift and premature stop codon in this gene.

We also sequenced the entire *MAF2-5* genomic region in 15 additional *A. thaliana* accessions to determine the prevalence of the polymorphism found between Cvi-0 and Ler-2 and to check for additional polymorphisms of potentially functional importance. The aligned sequence spans approximately 24 kbp (PopSet Genbank records EU980614–EU980630) in this larger sample set and includes the entire coding regions for the four *MAF* genes, their untranslated region (UTRs) and intergenic regions.

We examined molecular variation in all 17 sequenced accessions and the published Col-0 genome sequence (Initiative 2000). A total of 279 SNPs and over 80 indels, many associated with repeat regions, were observed across the entire *MAF2-5* genomic region. Estimates of total sites nucleotide diversity, π, varied between 0.0008 and 0.0055 across sampled features (table 1), consistently lower than the average nucleotide diversity of 0.0071 observed for the genome of *A. thaliana* (Schmid et al. 2005). Interestingly, nucleotide diversity values were similar across the four *MAF* genes for both silent and total sites ($\pi \sim 0.002-0.003$) (table 1), but somewhat lower than values observed for the closely related paralog, *FLC* ($\pi = 0.0044$; Caicedo et al. 2004). The distribution of polymorphism is also very different in the *MAF* genomic region compared with *FLC*. Consistent negative Tajima's values in the *MAF2-5* genes indicate a predominance of rare alleles (table 1), as is common throughout the *A. thaliana* genome (Nordborg et al. 2005), but unlike the strong dimorphic haplotype structure observed for *FLC* (Caicedo et al. 2004).

Despite similar levels of nucleotide diversity across the *MAF2-5* genes, there seem to be differences in some aspects of the patterning of variation at these loci. A total of 15 nonsynonymous SNPs were found across *MAF2-5*, but these are unevenly distributed among the genes. *MAF2* and *MAF3* contain four and eight nonsynonymous substitutions, respectively (three of which are singletons in each case), whereas *MAF4* and *MAF5* contain two and one nonsynonymous substitutions, all singletons, respectively. Ratios of nonsynonymous/synonymous nucleotide diversity (π_a/π_s) are consistent with *MAF4* and *MAF5* being more evolutionarily conserved and under greater purifying selection than *MAF2* and *MAF3* (table 1), and the π_a/π_s confidence intervals (CI) for these genes, suggest that differences among genes are significant (table 1).

The observed levels of indel polymorphisms also suggest that *MAF2* and *MAF3* are evolving more rapidly than their downstream paralogs. In *MAF2*, two accessions, Bs-1 (CS996) and Gr-3 (CS1202) contain a 5-bp deletion in the third exon, leading to a frameshift and putative early stop codon. Additionally, four of the sequenced accessions were found to contain very large insertions in the last intron of the *MAF2* gene. Curiously, these insertions are highly similar to portions of *MAF3* (fig. 1*B*). One of the large insert types (identified in Bs-1 and Gr-3) is ~2.7 kbp in length and appears to consist of a portion of the *MAF3* open reading frame (ORF), spanning parts of intron 1 to intron 6 and the intervening exons (insert type s4, fig. 1*B*). A second ~1.9-kbp insert (identified in accessions Kas-1 [CS903] and Chi-1 [CS1074]) is similar to the larger 2.7-kbp insert, but has a smaller portion of the *MAF3* intron 1 and is accompanied by a 221-bp deletion of intron 5 of *MAF2* and a 10-bp insertion (not homologous to *MAF3*) at its 5′ end. Both large inserts are accompanied by large deletions (~300 or ~500 bp) at the 3′ end of *MAF2* intron 5 (insert type s2, fig. 1*B*). Accessions with these large insertions in *MAF2* had an intact *MAF3* gene (fig. 1*B*), suggesting that rearrangements involving partial duplication and insertion are responsible for the origin of this sequence. Surprisingly, in accessions with *MAF2* inserts, the last exon of *MAF3* and portions of the *MAF3* 3′ UTR possess homology to corresponding regions of *MAF2*, suggesting further rearrangements or possible gene conversion between the two genes (fig. 1*B*).

MAF3 also harbors potentially functionally important indel polymorphisms. Three accessions (Cvi-0, Ita-0 [CS1244], and PHW-1 [CS6042]) were found to contain different indels in exons 1 or 3, leading to frameshifts and early stop codons. Two accessions (PHW-33 [CS6092] and Co-1 [CS1084] contain a 9-bp in-frame deletion in exon 1, leading to the loss of three amino acids. In contrast, no indel polymorphisms or rearrangements with potential functional consequences were identified in *MAF4* and *MAF5*.

To determine if positive selection may be driving the divergence of the *MAF2-5* genes, we tested for variable d*N*/d*S* ratio among *MAF* gene lineages (Yang 1997). Although d*N*/d*S* ratios for branches leading to *MAF2* and *MAF3* were among the highest estimated (supplementary fig. 1, Supplementary Material online), no ratio exceeded 1, and no significant difference was found between the log likelihoods for a model with variable evolutionary rates among branches and a one-ratio model. Thus, despite the higher levels of nonsynonymous polymorphism and indels found

Table 1
Diversity Statistics in the *MAF2-5* Region

	Region									
	Upstream	*MAF2*	Intergenic 2–3	*MAF3*[a]	Intergenic 3–4	*MAF4*	Intergenic 4–5	*MAF5*	Downstream	Total[b]
N	17	18	18	18	17	18	18	17	15	18
Length	757	6,854	1,171	3,838	1,536	3,842	1,438	4,978	159	24,720
Length—gaps	750	2,983	637	3,226	576	3,040	882	4,310	154	16,558
Missing data excluded										
h	10	12	10	15	5	12	9	15	2	n.a.
S	14	38	16	46	6	41	11	48	1	221
π	0.0055	0.0026	0.0051	0.0033	0.0020	0.0028	0.0018	0.0021	0.0009	n.a.
θ_W	0.0055	0.0037	0.0073	0.0042	0.0031	0.0039	0.0036	0.0034	0.0020	n.a.
Tajima's D	−0.008	−1.262	−1.136	−0.795	−1.159	−1.186	−1.823	−1.541	−1.1595	n.a.
π_s	0.0055	0.0027	0.0051	0.0030	0.0020	0.0032	0.0018	0.0023	0.0009	n.a.
θ_{Ws}	0.0055	0.0038	0.0073	0.0040	0.0031	0.0044	0.0036	0.0037	0.0020	n.a.
π_a/π_s	n.a.	1.227	n.a.	3.147	n.a.	0.204	n.a.	0.098	n.a.	n.a.
Missing data included[c]										
S	14	53	21	52	14	49	17	58	1	279
Π	0.0055	0.0017	0.0036	0.0030	0.0019	0.0028	0.0020	0.0024	0.0008	0.0025
θ_W	0.0055	0.0024	0.0052	0.0040	0.0029	0.0037	0.0035	0.0036	0.0019	0.0034
π_a/π_s[d]	n.a.	1.423 (0.856–1.158)	n.a.	2.812 (2.643–3.271)	n.a.	0.108 (0.100–0.134)	n.a.	0.090 (0.107–0.174)	n.a.	n.a.

[a] Silent statistics for *MAF3* were calculated excluding the Cvi-0 insert, as it causes a frameshift.

[b] Diversity statistics were not calculated for the complete genomic region excluding missing data, because failed sequencing for some accessions leads to the exclusion of complete genes.

[c] Analyses that include missing data were carried out considering individual sites; however, DNAsp cannot calculate silent θ_W and π when missing data are included.

[d] π_a/π_s with missing data was calculated with Libsequence; 95% CIs (in parentheses) were obtained from 100 bootstrapped data sets. Note that 36% of *MAF2* data sets and 18% of *MAF3* data sets had infinite π_a/π_s values due to lack of synonymous substitutions. These data sets were excluded from CI estimates.

in *MAF2* and *MAF3*, which suggest that these genes are evolving more rapidly than their downstream paralogs, there is no evidence that selection has driven the divergence of the *MAF* genes.

Given the presence of *MAF3*-like sequence in some inserts observed in *MAF2* alleles, we also assessed the extent of gene conversion among protein-coding regions of the *MAF2-5* paralogs. The large *MAF2* insertions were excluded from this analysis, as they are composed of coding and noncoding data (see below), and their identity with *MAF3* is known. Significant conversion events were detected only between *MAF4* and *MAF5* (supplementary table 4, Supplementary Material online) and were confined to stretches of 30–40 bp at the beginning of these genes. This suggests that, despite the recent divergence of the *MAF* gene family, gene conversion is not playing a large role in the evolution of these genes.

Genotyping of *MAF2-5* Gene Cluster Variation

We determined the prevalence of identified, potentially functional *MAF2-5* polymorphisms in the species by genotyping 169 European accessions of *A. thaliana*. Genotyped polymorphisms include mutations leading to amino acid substitutions, or indel polymorphisms that result in frameshifts or are large enough to potentially disrupt gene function. The polymorphisms that were genotyped were also found at moderate-to-high frequencies in the sequenced accessions set; no singletons were genotyped, unless they characterized a difference between Cvi-0 and L*er*-2. The locations of the genotyped polymorphisms are summarized in supplementary table 3, Supplementary Material online, and primers used are in supplementary table 2, Supplementary Material online.

Several of the genotyped polymorphisms, including all genotyped differences between Cvi-0 and L*er*-2, were found to occur at very low frequency within European accessions of *A. thaliana* (supplementary table 3, Supplementary Material online). For example, the radical nonsynonymous SNP in exon 4 of *MAF2*, characterizing Cvi-0 and Ita-0 was not found to occur in any other accession (from 151 successfully genotyped accessions). Likewise, the 146-bp insertion in the first exon of Cvi-0 *MAF3* was not found in any other genotyped accession (from 128 successfully genotyped accessions). A single exception was a G to A substitution characteristic of Cvi-0 in exon 7 of *MAF3*, which was found in 35% (45 of 129) successfully genotyped accessions (supplementary table 3, Supplementary Material online).

Because *MAF3* was found to harbor several nonsynonymous SNPs and several possible loss-of-function mutations (four of the sequenced accessions carried one of three different indels leading to frameshifts), all exons of this gene were genotyped by sequencing to assess variation within *A. thaliana*. Sequencing efforts for any given exon were discontinued in cases where at least half of the accessions had been sequenced and no polymorphism had been discovered, as polymorphisms occurring at very low frequency in the population provide no power for association analyses. None of the three *MAF3* indel mutations were found to occur with significant frequency in the expanded *A. thaliana* sample set. Only exons 1 and 7 of *MAF3* harbor moderate-frequency polymorphic sites. In exon 1, a C to A substitution leading to a radical amino acid change from asparagine (N) to lysine (K) was found to occur in 29% (35 of 122) successfully genotyped accessions. Exon 7 harbors multiple polymorphisms occurring at moderate frequencies: 1) a G to A substitution present in Cvi-0 and

Table 2
Large Insertions Found in the 5′ Region of *MAF2*

Insert Type	Size (bp)[a]	Frequency[b]
s1	~1,100	2
s2	~1,900	20
s3	~3,100	7
s4	~2,700	8
s5	~3,700	1
s6	~4,700	3
s0	No insert	118
?	?	10
Total		169

[a] Inserts s3, s5, and s6 were not sequenced in their entirety and estimates of insert size are approximate; insert types are ranked according to size of amplicon, but the occurrence of deletions in conjunction with several insert types makes prediction of insert size based on amplicon size difficult.

[b] Observed frequency based on genotyping panel of 169 accessions.

described above leading to a conservative amino acid change (valine [V] to methionine [M]) and 2) three linked mutations (a C to T substitution leading to a radical arginine [R] to tryptophan [W] replacement, a C to G substitution leading to a radical proline [P] to arginine [R] amino acid change, and a synonymous T to C substitution) that were found in 33% (43 of 129) successfully genotyped accessions (supplementary table 3, Supplementary Material online).

Two *MAF2* indels were initially genotyped in the expanded accession set: 1) the 5-bp deletion in exon 3 and 2) the large insertions occurring in the 3′ end of *MAF2*. The 5-bp deletion occurs at low frequency in European *A. thaliana*; only five accessions of the 129 successfully genotyped were found to carry it (supplementary table 3, Supplementary Material online). Large insertions in the 3′ portion of the *MAF2* gene, on the other hand, were found to be relatively common in the sampled population. Forty-one of 159 (25.8%) successfully genotyped accessions have a large insert in the 3′ portion of *MAF2*.

Surprisingly, the expanded genotyping of the *MAF2* insertions revealed the existence of several insert types that vary greatly in size, suggesting either that multiple insertion events or postinsertion rearrangements have occurred (fig. 1*C*; supplementary table3, Supplementary Material online). Six different insert types, referred to as s1 through s6, ranging in size from ~1.1 to ~4.7 kbp, were found occurring at varying frequencies within the *A. thaliana* genotyping set (table 2; fig. 1*B* and *C*; supplementary table 3, Supplementary Material online). The previously described 2.7-kbp insert corresponds to type s4, whereas the 1.9-kbp insertion has been designated as s2 (table 2). The latter is the most common insertion observed, present in 12.6% (20 of 159) of the sample and in nearly half of all accessions that had an insertion in 3′ region of *MAF2*.

Characterizing the Variable *MAF2* Insertions

Given the prevalence of variably sized large insertions in the *MAF2* gene, we attempted to further characterize the structures of these inserts. Representative accessions were chosen to isolate and sequence the 5′ and 3′ portions of each of the six types (s1–s6) of identified inserts (Genbank Popset records EU931627–EU931636).

As expected, two of the *MAF2* large inserts, s2 and s4, match those previously sequenced and described above. The four remaining insertions appear to be novel, but share substantial sequence similarity with s2 and s4. All insertions have a common 3′ end, which is identical to exons 3–6 of *MAF3*, intervening introns, and the entire intron 6 sequence. Although not sequenced in its entirety, accessions bearing all these insertions appear to possess an intact last exon (exon 6) of *MAF2*.

Although all *MAF2* insertions have high sequence similarity in their 3′ ends, the 5′ portion of these inserts varies among the different types, though they do possess some shared features. In particular, several insertions are accompanied by deletions of portions of *MAF2*, including exon regions (fig. 1*B* and *C*). Insertion s1 is accompanied by deletion of a portion of *MAF2* exon 3 and the entire region encompassing introns 3–5; the insert shows homology to regions spanning introns 2–6 of *MAF3*. The other insertions, which include types s2 through s6, seem to contain regions identical to introns 1–6 of *MAF3*, but vary in the length of sequence homologous to *MAF3* intron 1 and the length of deleted *MAF2* portions. The s3 insertion is accompanied by the deletion of *MAF2* introns 3–5 and a 26-bp insertion partially homologous to part of *MAF2* exon 2; this insert type is the only one that was confirmed to contain sequence homologous to a portion of exon 1 of *MAF3*. The s5 and s6 insertions contain large regions identical in sequence to *MAF3* intron 1 and have intact *MAF2* sequence, up until intron 5. Inserts s3, s5, and s6 were not sequenced in their entirety, so portions of the structure of these insertions remain undetermined (fig. 1*C*).

The origin(s) of these large insertions in *A. thaliana MAF2* alleles is perplexing. The shared 3′ insertion boundary suggests that a single genomic rearrangement may have given rise to all insert types. Moreover, all accessions that were fully sequenced for the *MAF2-5* genomic region possess the entire *MAF3* gene downstream of *MAF2*, suggesting that a duplication event must have been involved, which left the original *MAF3* locus intact. Note that for accessions carrying inserts that were only partially sequenced (types s1, s3, s5, and s6), no conclusions can be drawn about the presence of an intact *MAF3* gene, although this seems likely based on the genomic sequence of types s2 and s4. Curiously, there are three linked SNPs in exon 7 of the *MAF3* gene, characteristic primarily of accessions with *MAF2* inserts, that cause exon 7 in these samples to resemble exon 6 of *MAF2* (fig. 1*B*). Thus, in fully sequenced accessions that carry large insertions within *MAF2*, the last exons of *MAF2* and *MAF3* are indistinguishable; this suggests that gene conversion may have occurred between portions of *MAF2* and *MAF3* associated with the origin of the *MAF2* inserts.

Unlike the similarity in 3′ portions of the insertion sequences, the varying length of the 5′ portion of the *MAF2* insertions suggests that each insert type has had a separate origin. This conclusion is supported by the presence of small amounts of SNP variation among insert types that are homologous to SNPs present in corresponding *MAF3* haplotypes (e.g., sites 6814 and 11931 in the genomic

Table 3
Chimeric transcripts observed for some accessions carrying large insertions within the *MAF2* gene.

			Homology												
			MAF2					*MAF3* (insert region)							*MAF2*
Accession	Transcript	*MAF2* insert type	exon 3	intron 3	exon 4	exon 5	intron 5[a]	exon 2	intron 2	exon 3	intron 3	exon 4	exon 5	exon 6	exon 6
Col	predicted	s0	x		x	x									x
Jl-2	CS6744	s1	x[b]						x[c]	x		x	x	x	x
Ka-0	CS6752	s1	x[b]						x[c]	x		x	x	x	x
Kas-1	CS903.1	s2	x		x	x				x		x	x	x	x
Kas-1	CS903.2	s2	x	x	x	x				x	x	x[d]	x	x	x
Kas-1	CS903.3	s2	x	x	x	x				x		x	x	x	x
Chi-1	CS6665.1	s2	x		x	x				x		x	x	x	x
Chi-1	CS6665.2	s2	x	x	x	x				x		x	x	x	x
Chi-1	CS6665.3	s2	x	x	x	x				x		x[d]	x	x	x
Gr-3	CS6725	s4	x	x	x	x				x		x	x	x	x
Jl-3	CS6745.1	s4	x	x	x	x[e]		x		x		x	x[f]	x	x
Jl-3	CS6745.2	s4	x	x	x	x[e]		x		x		x	x	x	x
Jl-3	CS6745.3	s4	x	x	x	x				x		x	x	x	x
Jl-3	CS6745.4	s4	x	x	x	x				x			x[f]	x	x
Jl-3	CS6745.5	s4	x		x	x				x		x	x	x	x
Pa-2	CS6826.1	s5	x	x	x	x		x		x		x	x	x	x
Pa-2	CS6826.2	s5	x	x	x	x	x	x		x		x	x	x	x
Pa-2	CS6826.3	s5	x	x	x	x[e]	x	x		x		x	x	x	x
Pa-2	CS6826.4	s5	x	x	x	x	x	x		x		x[d]	x	x	x
Pa-2	CS6826.5	s5	x		x	x	x	x		x		x	x	x	x

Note.—Marked columns contain regions observed in expressed product.

[a] Alternative splice products of Col-type *MAF2* contain an additional exon in this intronic region, which is homologous to *MAF2* exon 6

[b] Accessions carrying s1 insert types contain only partial *MAF2* exon 3 sequence (see Fig. 1)

[c] These transcripts contain 16 bp homologous to *MAF3* intron 2

[d] These transcripts contain a longer version of sequence homologous to *MAF3* exon 4

[e] These transcripts contain a 4 bp deletion in the transcribed *MAF2* exon 5

[f] These transcripts contain a 4 bp deletion in the transcribed region homologous to *MAF3* exon 5

alignment and sites 6961 and 12078 in the genomic alignment). Although it is difficult to imagine that six independent insertion events would yield inserts sharing common 3′ regions, it may be that the high levels of sequence similarity between *MAF2* and *MAF3* exons make this genomic region prone to certain types of rearrangements, yielding similar insertions at the same location.

The Insertion of *MAF3* Coding Regions into *MAF2* Results in Gene Fusions That Express Chimeric mRNAs

Because the large insertions in *MAF2* introduce novel coding sequence into this gene, we attempted to determine the effect these would have on *MAF2* expression. We amplified *MAF2* from cDNA obtained from plants grown under long-day conditions. Attempts were made to obtain *MAF2* from at least one accession representing each insert type. *MAF2* and *MAF3* have the highest levels of sequence similarity in the *MAF* gene clade, making the design of specific primers challenging; thus, results were not obtained for all inserts. The placement of expression primers on the third and sixth exons of *MAF2* also prevented any amplification from accessions with an s3 insert type (fig. 1*B*).

We successfully amplified *MAF2* cDNA from seven accessions containing four insert types (s1, s2, s4, and s5), all of which expressed chimeric mRNAs due to the inserted sequence (table 3; Genbank Popset records EU998941–EU998960). All accessions expressed *MAF2* exons, but varied in their expression of *MAF3* type exons in the insert (table 3). Note that we cannot draw any conclusions about the transcription of exons 1 and 2 of *MAF2* due to the localization of our 5′ amplification primer. The *MAF2-5* genes are known to exhibit extensive alternative splicing, with two to five currently characterized variants per gene (Ratcliffe et al. 2001, 2003; Scortecci et al. 2001). Interestingly, we also observed extensive alternative splicing for insert-containing *MAF2* alleles (table 3). Exon skipping, exonization of introns, and partial transcription of given features were all observed in *MAF2* transcripts with inserts (table 3). Among the seven accessions examined, 19 transcripts were observed. We also often amplified known alternatively spliced *MAF2* and *MAF3* products from cDNA obtained from plants grown under standard conditions (data not shown) (Ratcliffe et al. 2001; Scortecci et al. 2001), implying that alternative splicing may be common. It is thus possible that the primers we used for RT-PCR may be biased against some splice products, and the diversity of alternative spliced transcripts may be even larger than the one reported here.

Because every accession with an insert that we examined produced chimeric mRNAs, this gives rise to the possibility that the large insertions may affect *MAF2* protein products. Assuming normal transcription of the first two *MAF2* exons, we attempted to characterize the putative protein products of the observed chimeric transcripts. Because the MADS-box domain is encoded completely within the first exon of *MAF2*, putative proteins are all expected to have a MADs-box domain. Predicted proteins were

```
          1         11        21        31        41        51        61        71        81
          |         |         |         |         |         |         |         |         |
Col       MGRKKVEIKRIENKSSRQVTFSKRRNGLIEKARQLSILCESSIAVLVVSGSGKLYKSASGDNMSKIIDRYEIHHADELEALDLA

                91        101       111       121       131       141       151       161
                |         |         |         |         |         |         |         |
Col       EKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISLEEQLETALSVTRARKTELMMGEVKSLQKTVGKKTFLVIEGDRGMSWEN
CS6665.1  EKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISLEEQLETALSVTRARKTELMMGEVKSLQKTDLAEKIRNYLPHKELLEIV
CS6745.5  EKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISLEEQLETALSVTRARKTELMMGEVKSLQKTDLAEKIRNYLPHKELLEIV
CS903.1   EKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISLEEQLETALSVTRARKTELMMGEVKSLQKTDLAEKIRNYLPHKELLEIV
CS6826.5  EKTRNYLPLKELLEIVQSKLEESNVDNASVDTLISLEEQLETALSVTRARKTELMMGEVKSLQKTENLQREENQTLASQHVKDH
CS6725    ?K?RNYLPRVTRNSPKVSTKTLLSPSSSDEKYFFFSFLLANYEYSKLEESNVDNASVDTLISLEEQLETALSVTRARKTELMMG
CS6744    ???????PLKELWDRITGSCRKNSELSSTQGVTRNSPKIL*-------------------------------------------
CS6752    EKTRNYLPLKELWDRITGSCRKNSELSSTQGVTRNSPKIL*-------------------------------------------
CS903.2   EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS903.3   EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6665.2  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6665.3  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6745.1  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6745.2  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6745.3  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6745.4  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6826.1  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6826.2  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6826.3  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------
CS6826.4  EKTRNYLPLKELLEIVQRLAQRHFYLPLLLMKNTFFFLFFWRIMNTASLKNQMSIMQVWIL*----------------------

            171       181       191       201       211
            |         |         |         |         |
Col       GSGNKVRETLPLLK*-------------------------------
CS6665.1  QRFSNIYGGTARDCSVSN*---------------------------
CS6745.5  QRFSNIYGGTARDCSVSN*---------------------------
CS903.1   QRFSNIYGGTARDCSVSN*---------------------------
CS6826.5  *---------------------------------------------
CS6725    EVKSLQKTDLAEKIRNYLPHKELLEIVQRFSNIYGGTARDCSVSN*
CS6744    ----------------------------------------------
CS6752    ----------------------------------------------
CS903.2   ----------------------------------------------
CS903.3   ----------------------------------------------
CS6665.2  ----------------------------------------------
CS6665.3  ----------------------------------------------
CS6745.1  ----------------------------------------------
CS6745.2  ----------------------------------------------
CS6745.3  ----------------------------------------------
CS6745.4  ----------------------------------------------
CS6826.1  ----------------------------------------------
CS6826.2  ----------------------------------------------
CS6826.3  ----------------------------------------------
CS6826.4  ----------------------------------------------
```

FIG. 2.—Predicted proteins for observed *MAF2* chimeric mRNA products. Protein IDs refer to transcript IDs in table 3. Shaded regions correspond to amino acids translated from *MAF3*-like insert regions; thus, predicted proteins with shaded regions are chimeric due to the inserts. The predicted MAF2 protein for Col, which does not carry an insert, is shown for comparison. No attempt has been made to align the predicted protein products, due to very limited homology at the protein level. Note that protein predictions were made assuming normal translation of the first 84 MAF2 amino acids, as this region was not amplified from cDNA from accessions with inserts in *MAF2*.

considered chimeric if they contained translated amino acids from *MAF2*- and *MAF3*-type sequences. All seven accessions we examined, containing s1, s2, s4, and s5 insert types, were predicted to produce chimeric protein variants, and a total of four chimeric protein variants were observed (fig. 2). However, the most common protein prediction, observed for 12 chimeric transcripts belonging to four accessions with s2, s4, and s5 type inserts, corresponded to a shortened, nonchimeric product, in which transcription of intron 3 led to a frameshift and early stop codon prior to the transcribed *MAF3* sequence (table 3, fig. 2).

Due to frameshifts introduced by the inserts, predicted chimeric proteins do not always represent a fusion between MAF2 and MAF3 amino acid sequence, but rather novel proteins (fig. 2). Transcripts CS6665.1 (s2), CS6745.5 (s4), CS903.1 (s2), CS6826.5 (s5), and CS6725 (s4) give rise to putative protein products with an intact MAF2 K-box domain (fig. 2), whereas downstream amino acid sequence stemming from the insert is homologous to the portion of a splice variant of *MAF3* (variant III, Ratcliffe et al. 2003). However, no significant similarity to any protein domain was found for *MAF3*-type transcribed sequence for transcripts CS6744, CS6752, and CS6826.5.

Although only chimeric transcripts were observed for accessions carrying large *MAF2* inserts, our results suggest that putative protein products of these accessions vary greatly. Alternative splicing of the first two exons and introns of *MAF2* could add further levels of variability to the possible proteins produced by accessions with inserts. Despite this variation, it seems that the most common outcomes

Table 4
Mixed Linear Model Association Results for Genotyped *MAF2-5* Polymorphisms

	Locus	*F*	Error df	Error MS	R^2 Model	R^2 Marker	*P* Value
Long day							
Time to flower	*MAF2* insert	5.932	152	38.204	0.163	0.033	0.016*[c]
Time to flower	*MAF3* exon 1	3.260	115	28.782	0.277	0.021	0.074
Time to flower	*MAF3* exon 7 (1)[a]	0.004	123	37.354	0.198	0.000	0.947
Time to flower	*MAF3* exon 7 (2)[b]	0.001	123	37.355	0.198	0.000	0.981
RLN	*MAF2* insert	17.905	152	9.339	0.150	0.100	**<0.0001***
RLN	*MAF3* exon 1	12.729	115	8.457	0.215	0.087	**<0.0001***
RLN	*MAF3* exon 7 (1)	3.296	123	9.225	0.195	0.022	0.072
RLN	*MAF3* exon 7 (2)	3.894	123	9.025	0.213	0.025	0.051
Short day							
Time to flower	*MAF2* insert	17.215	151	54.365	0.263	0.084	**<0.0001***
Time to flower	*MAF3* exon 1	8.104	115	58.044	0.191	0.057	0.0052*
Time to flower	*MAF3* exon 7 (1)	1.902	123	71.850	0.137	0.013	0.170
Time to flower	*MAF3* exon 7 (2)	5.075	123	69.645	0.163	0.035	0.026*
RLN	*MAF2* insert	27.823	151	13.235	0.226	0.143	**<0.0001***
RLN	*MAF3* exon 1	19.678	115	9.763	0.443	0.095	**<0.0001***
RLN	*MAF3* exon 7 (1)	10.550	123	14.486	0.152	0.073	**0.0015***
RLN	*MAF3* exon 7 (2)	24.563	123	10.599	0.380	0.124	**<0.0001***

[a] SNP genotyped in position 12419 of alignment.

[b] Three linked SNPs genotyped in exon 7 of *MAF3* (positions 12467, 12495, and 12506 in alignment).

[c] *P* values marked with asterisks are significant at the nominal level; *P* values in bold remain significant after a strict Bonferroni correction.

involve shortened proteins, some of which contain some novel sequence. It may be that chimeric transcripts are targeted for degradation, thus rendering the *MAF2* gene nonfunctional in accessions with inserts. However, if chimeric transcripts carry intact MADS-box coding sequence, as assumed here, the shortened proteins produced could compete with normal MAF or FLC proteins for DNA-binding sites.

Insertion Alleles in the *MAF2-5* Gene Cluster are Associated with Variation in Flowering Time

Mapping results from Cvi-0 and L*er*-2 RILs indicate that a QTL for flowering time includes a genomic region that encompasses the *MAF2-5* gene cluster on the bottom of chromosome 5 (Engelmann K, Ungerer M, Purugganan MD, unpublished data, Ungerer et al. 2002). There are nonsynonymous polymorphisms that differentiate the Cvi-0 and L*er*-2 *MAF2-5* genes, though most occur at very low frequency in our expanded *A. thaliana* sample set. Based on the polymorphisms found (see above), the 146-bp insertion in the first exon of Cvi-0 *MAF3*, which results in a frameshift and the formation of an early stop codon, may be a strong candidate for flowering time differences between Cvi-0 and L*er*-2. This insertion, however, is found only in the Cape Verde Island accession.

Other QTL mapping studies based on crosses between various accessions have also recently reported flowering time QTL localizing to the *MAF2-5* region (El-Lithy et al. 2004, 2006; Simon et al. 2008), suggesting that loci in this region may have important phenotypic effects. Several potential function-altering polymorphisms occur at moderate frequency in the *MAF2-5* genomic region in *A. thaliana* and could have an effect on flowering time. Four polymorphisms in particular occur at sufficient frequency in our sample to permit an evaluation through association analyses (i.e., $>10\%$): 1) the large insertions in the 3′ region of *MAF2*, 2) the radical SNP in exon 1 of *MAF3* (position 8851 in the alignment), 3) a conservative nonsynonymous substitution in *MAF3* exon 7 (position 12419 in the alignment), and 4) three linked SNPs in exon 7 of *MAF3* (positions 12467, 12495, and 12506 in the alignment). We tested to see whether any of these moderate-frequency polymorphisms were significantly associated with days to bolting and rosette leaf number at bolting (RLN) variation in *A. thaliana*, both common measures of flowering time, and highly genetically correlated traits (Ungerer et al. 2002). We used an MLM approach (Yu et al. 2006), which takes into account population ancestry and familial relatedness among accessions to decrease the probability of spurious associations due to population structure. The genotyping sample set used consists of a geographically restricted set of *A. thaliana* accessions, which aids in further decreasing the probability of false-positive associations. Flowering time and RLN phenotypes were obtained under long-day and short-day conditions as described in Olsen et al. (2004). The three linked SNPs at exon 7 of *MAF3* were in perfect linkage disequilibrium and thus considered a single polymorphism. All large insertions in *MAF2* were considered equivalent; given their positions and structure, it is likely that any phenotypic effects they have are similar, and there is little power to separately evaluate different insert types.

A significant association was detected from several evaluated polymorphisms and flowering time and/or RLN variation under both short-day and long-day conditions (table 4). Notably, however, the large *MAF2* insertions were the only polymorphism consistently associated with both flowering time and RLN variation across both day length conditions; *MAF2* inserts remained associated with at least one measure of flowering time under both short and long days, even after a strict Bonferroni correction (table 4). In all these cases, *MAF2* insertion alleles are associated with earlier flowering or fewer rosette leaves at flowering (table 5). These results suggest that *MAF2*

Table 5
Mean Flowering Phenotypes Associated with the *MAF2* and *MAF3* Genotyped Polymorphisms

		Long Day		Short Day	
Polymorphism		Time to Flower	RLN	Time to Flower	RLN
MAF2 insert	No insert	45.17 (7.13)[c]	12.72 (2.99)	56.12 (8.84)	19.56 (3.99)
	Insert	41.6 (3.86)	10.26 (3.3)	49.01 (3.67)	15.59 (2.53)
MAF3 exon 1	a	41.61 (3.17)	10.47 (3.45)	49.13 (3.4)	15.81 (2.77)
	c	44.92 (6.79)	12.87 (2.84)	55.98 (8.82)	19.7 (4.01)
MAF3 exon 7 (1)[a]	a	43.77 (4.44)	12.8 (2.75)	56.26 (8.79)	20.31 (3.65)
	g	44.79 (7.63)	11.98 (3.57)	53.89 (8.97)	17.86 (4.02)
MAF3 exon 7 (2)[b]	cct	44.16 (4.57)	12.54 (2.58)	56.25 (8.82)	19.86 (3.71)
	tgc	44.98 (9.68)	11.71 (4.43)	51.65 (8.47)	16.42 (3.77)

[a] SNP genotyped in position 12419 of alignment.

[b] Three linked SNPs genotyped in exon 7 of *MAF3* (positions 12467, 12495, and 12506 in alignment).

[c] Numbers in parentheses correspond to standard deviations.

insertion alleles may be good candidate polymorphisms for affecting flowering phenology in *A. thaliana*.

Although the *MAF2* inserts were the most consistent polymorphism associated with flowering time, all polymorphisms evaluated were associated with at least one of the phenotypic traits measured. In particular, all polymorphic sites had a significant association with RLN under short-day conditions, even after a strict Bonferroni correction (table 4). Linkage disequilibrium (LD) among the polymorphisms examined may account for the various significant associations with the *A. thaliana* flowering time traits. We checked for LD between each of the polymorphisms used for association analyses with exact tests of gametic disequilibrium. All mutational groups are under high linkage disequilibrium with each other (supplementary table 5, Supplementary Material online) as expected from their physical proximity. The patterns of polymorphism suggest that LD is due primarily to the nesting of alleles into certain mutational categories (i.e., alleles have arisen in particular backgrounds, and associations have not been broken up by recombination). Notably, the large *MAF2* inserts seem to occur in a limited genetic background.

This nonrandom distribution of allelic genotypes suggests that all evaluated polymorphisms must be considered as candidates for affecting flowering time in *A. thaliana*. Likewise, polymorphism in linked unexamined regions could also play a role. However, the possible disruption of *MAF2* function caused by the inserted sequence, and the strength of the association of their presence with an earlier flowering time and morphology, reinforce this polymorphism as a candidate for a life history trait–variation in *A. thaliana*.

Discussion

As knowledge of organisms' genome contents grow, it has become apparent that copy number variation (often dubbed CNV) can contribute enormously to polymorphism in many species (e.g., Redon et al. 2006; Dopman and Hartl 2007; Smartt et al. 2007). Genomic rearrangements are common in genomes, and are known to contribute to natural variation in human disease traits (Stankiewicz and Lupski 2002) and can be subject to various forms of selection (Emerson et al. 2008). Multigene families with tandemly located gene members are especially prone to rearrangements including duplications, deletions, and more complex rearrangements that can give rise to chimeric genes, thus forming a hotspot for the origin (and loss) of new functionalities and novel genes.

In plant genomes, the dynamics of gene cluster evolution and formation of chimeras have especially been documented for resistance genes, that is, genes involved in pathogen defense (e.g., Song et al. 1997; Sun et al. 2001; Caicedo and Schaal 2004; Kuang et al. 2004; Shiotani et al. 2007). There are also a few examples of chimeric genes found in *A. thaliana* that contribute to phenotypic polymorphism within the species. Notable among these is the *DM1* locus (Bomblies et al. 2007), implicated in hybrid necrosis in certain crosses of *A. thaliana* accessions, the *MAM1* and *MAM2* genes (Kroymann et al. 2003), involved in resistance to generalist herbivores, and the S locus *SRK* and *ARK* genes (Sherman-Broyles et al. 2007), which are involved in the loss of self-incompatibility within the species. These documented cases of chimeric gene occurrence suggest that gene chimeras may play a role in ecologically important phenotypic variation in *A. thaliana* and other species. Here, we have shown a complex gene rearrangement within the *MAF2-5* family of MADS-box genes in the plant *A. thaliana*, which gives rise to chimeric transcripts and may give rise to chimeric protein products. Surprisingly, this complex polymorphism segregates at moderate frequency within the species, suggesting that it does not have a negative impact on plant fitness, or it is mutationally generated at high enough frequency, that it is not efficiently purged by selection (e.g., Repping et al. 2003). Moreover, it is possible that these complex rearrangements play a role in flowering time variation within *A. thaliana*.

The *MAF* clade within the MADS-box gene family has garnered attention in recent years due to the contribution of two of its members, *FLC* (Michaels and Amasino 1999; Sheldon et al. 1999) and *FLM* (Scortecci et al. 2001), to variation in flowering time within *A. thaliana* (Werner, Borevitz, Uhlenhaut, et al. 2005; Werner, Borevitz, Warthmann, et al. 2005; Korves et al. 2007; Scarcelli et al. 2007). Much less is known, however, about the genetic variation or functional significance of the other four members of the clade, *MAF2-5*. We have shown that the *MAF2-5* gene cluster is evolving in a heterogeneous manner. The low levels of nonsynonymous SNPs observed for *MAF4* and *MAF5* contrast with the high levels of nonsynonymous SNPs, indels, and rearrangements observed for *MAF2* and *MAF3*, which are especially striking given the physical linkage in this region. These results suggest very different evolutionary dynamics for these genes, all of which are believed to play a role in the control of flowering time in *A. thaliana* (Ratcliffe et al. 2003). Purifying selection may be conserving *MAF4* and *MAF5* function, whereas potentially disrupting indels as well as nonsynonymous mutations in *MAF2* and *MAF3* seem to be tolerated and are possibly contributing to natural flowering time variation within the species.

Further clues about the role of the *MAF* genes in flowering time variation within *A. thaliana* could be obtained from study of the geographical distribution of *MAF2-5* alleles and correlations with environmental variables. Our genotyping panel is not ideally suited to this type of analysis, as it was created from accessions from a restricted geographic area with known population ancestry, to reduce the probability of false positives in association studies; localities of origin for ~80% of our accessions occur in northwestern Europe, between 44° and 54° north latitude and west of 22° longitude. However, we explored whether differences in latitude or longitude of origin might be associated with *MAF2* insert presence in our limited *A. thaliana* panel. A significant longitudinal difference was detected ($t = -5.007$, $P < 0.0001$), with accessions with *MAF2* inserts originating preferentially in more easterly locations ($MAF2_{insert}$ mean longitude = 15.48° east; $MAF2_{no\ insert}$ mean longitude 7.8° east). The ecological significance of this result is currently unknown, but, with the caveat that longitudinal variance is high for the *MAF2* insert group (85.51 vs. 27.03), and there is geographical overlap between the two haplotype types, this may suggest that *MAF2* inserts are favored in regions with more continental weather.

Our association analysis results, which are replicated by various QTL mapping studies (Ungerer et al. 2002; El-Lithy et al. 2004, 2006; Simon et al. 2008), suggest that the *MAF2* insertions, or linked polymorphism, reduce the time to flowering under long-day and short-day conditions. Given that *MAF2* is thought to be a flowering repressor, this suggests an impairment of *MAF2* function may occur for alleles with inserts. Association analyses for flowering time under field conditions, where both day length and temperature changes likely occur, are needed to determine whether the *MAF2* inserts or associated polymorphisms are associated with the same phenotypic effect in the wild. If these associations hold, however, it is interesting to note that *MAF2* inserts are derived mutations associated with shorter flowering time, a situation that is reminiscent of that of another flowering gene, *FRIGIDA* (*FRI*). Numerous derived loss-of-function mutations, which are common throughout the species, have been observed for the *FRI* gene in *A. thaliana*, all of which are associated with early flowering and loss of a vernalization requirement (Johanson et al. 2000; Gazzani et al. 2003; Shindo et al. 2005). In another major flowering time locus, *FLC*, disruptive insertions in the first intron occur at extremely low frequency in the species (Caicedo et al. 2004) but have also been found to lead to early flowering (Gazzani et al. 2003; Michaels et al. 2003). Thus, derived mutations decreasing time to flowering, and perhaps favoring a rapid cycling life history, may be common in flowering time genes in *A. thaliana*.

Supplementary Material

Supplementary tables 1–5 and supplementary figure 1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank members of the Purugganan laboratory for a critical reading of the manuscript, as well as K. Engelmann and M. Ungerer for sharing unpublished results. Special thanks to Jonathan Flowers for assistance with bootstrapping scripts. This work was supported in part by a grant from the NSF Frontiers in Biological Research Program.

Literature Cited

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, Pouplana LRd, Martinez-Castilla L, Yanofsky MF. 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci USA.

Becker A, Theissen G. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phylogen Evol. 29:464–489.

Bomblies K, Lempe J, Epple P, Warthmann N, Lanz C, Dangl JL, Weigel D. 2007. Autoimmune response as a mechanism for a Dobzhansky–Muller-type incompatibility syndrome in plants. Plos Biol. 5:e236.

Bradbury P, Kroon D, Zhang Z, Casstevens T, Buckler EE. 2005. Tassel: software for analyzing trait associations, evolutionary patterns, and linkage disequilibrium. SNP Markers Symposium: Discovery, Development, Mapping, Utilization. ASA-CSSA-SSSA International Annual Meetings, Salt Lake City, UT

Caicedo AL, Schaal BA. 2004. Heterogeneous evolutionary processes affect *R* gene diversity in natural populations of *Solanum pimpinellifolium*. Proc Natl Acad Sci USA. 101: 17444–17449.

Caicedo AL, Stinchcombe JR, Olsen KM, Schmitt J, Purugganan MD. 2004. Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. Proc Natl Acad Sci USA. 101:15670–15675.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. Proc Natl Acad Sci. 104:19920–19925.

El-Lithy ME, Bentsink L, Hanhart CJ, Ruys GJ, Rovito D, Broekhof JLM, van der Poel HJA, van Eijk MJT, Vreugdenhil D, Koornneef M. 2006. New *Arabidopsis* recombinant inbred line populations genotyped using SNPWave and their use for mapping flowering-time quantitative trait loci. Genetics. 172:1867–1876.

El-Lithy ME, Clerkx EJM, Ruys GJ, Koornneef M, Vreugdenhil D. 2004. Quantitative trait locus analysis of growth-related traits in a new *Arabidopsis* recombinant inbred population. Plant Physiol. 135:444–458.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science. 320:1629–1631.

Gazzani S, Gendall AR, Lister C, Dean C. 2003. Analysis of the molecular basis of flowering time variation in *Arabidopsis* accessions. Plant Physiol. 132:1107–1114.

Hardy OJ, Vekemans X. 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes. 2:618–620.

Hepworth SR, Valverde F, Ravenscroft D, Mouradov A, Coupland G. 2002. Antagonistic regulation of flowering-time gene SOC1 by CONSTANS and FLC via separate promoter motifs. EMBO J. 21:4327–4337.

Initiative TAG. 2000. Analysis of the genome of the flowering plant *Arabidopsis thaliana*. Nature. 408:796–815.

Itano HA. 1957. the human hemoglobins: their properties and genetic control. Adv Prot Chem. 12:216–268.

Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C. 2000. Molecular analysis of FRIGIDA, a major determinant of natural variation in *Arabidopsis* flowering time. Science. 290:344–347.

Korves TM, Schmid KJ, Caicedo AL, Mays C, Stinchcombe JR, Purugganan MD, Schmitt J. 2007. Fitness effects associated with the major flowering time gene FRIGIDA in *Arabidopsis thaliana* in the field. Am Nat. 169:E141–E157.

Kroymann J, Donnerhacke S, Schnabelrauch D, Mitchell-Olds T. 2003. Evolutionary dynamics of an *Arabidopsis* insect resistance quantitative trait locus. Proc Natl Acad Sci USA. 100:14587–14592.

Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore KF. 2004. Multiple genetic processes result in heterogeneous rates of evolution within the major cluster resistance genes in lettuce. Plant Cell. 16:2870–2894.

Lee H, Suh SS, Park E, Cho E, Ahn JH, Kim SG, Lee JS, Kwon YM, Lee I. 2000. The AGAMOUS-LIKE 20 MADS domain protein integrates floral inductive pathways in *Arabidopsis*. Genes Dev. 14:2366–2376.

Liljegren SJ, Ditta GS, Eshed Y, Savidge B, Bowman JL, Yanofsky MF. 2000. *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. Nature. 404:766–770.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.

Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science. 260:91–95.

Michaels SD, Amasino RM. 1999. FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell. 11:949–956.

Michaels SD, He YH, Scortecci KC, Amasino RM. 2003. Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in *Arabidopsis*. Proc Natl Acad Sci USA. 100:10102–10107.

Neff MM, Turk E, Kalishman M. 2002. Web-based primer design for single nucleotide polymorphism analysis. Trends Genet. 18:613–615.

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

Nordborg M, Hu TT, Ishino Y, et al. (23 co-authors). 2005. The pattern of polymorphism in *Arabidopsis thaliana*. Plos Biol. 3:1289–1299.

Ohno S. 1970. Evolution by gene duplication. Berlin: Springer.

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD. 2004. Linkage disequilibrium mapping of *Arabidopsis* CRY2 flowering time alleles. Genetics. 167:1361–1369.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics. 155:945–959.

Purugganan MD. 1997. The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. J Mol Evol. 45:392–396.

Purugganan MD, Rounsley SD, Schmidt RJ, Yanofsky MF. 1995. Molecular evolution of flower development – diversification of the plant Mads-Box regulatory gene family. Genetics. 140: 345–356.

Ratcliffe OJ, Kumimoto RW, Wong BJ, Riechmann JL. 2003. Analysis of the *Arabidopsis* MADS AFFECTING FLOWERING gene family: mAF2 prevents vernalization by short periods of cold. Plant Cell. 15:1159–1169.

Ratcliffe OJ, Nadzan GC, Reuber TL, Riechmann JL. 2001. Regulation of flowering in *Arabidopsis* by an FLC homologue. Plant Physiol. 126:122–132.

Redon R, Ishikawa S, Fitch KR, et al. (43 co-authors). 2006. Global variation in copy number in the human genome. Nature. 444:444–454.

Repping S, Skaletsky HJ, Brown L, et al. (12 co-authors). 2003. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. Nat Genet. 35:247–251.

Rouse DT, Sheldon CC, Bagnall DJ, Peacock WJ, Dennis ES. 2002. FLC, a repressor of flowering, is regulated by genes in different inductive pathways. Plant J. 29:183–191.

Rozas J, Rozas R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics. 15:174–175.

Sawyer SA. 1999. GENECONV: a computer package for the statistical detection of gene conversion. Department of Mathematics, Washington University in St. Louis.

Scarcelli N, Cheverud JM, Schaal BA, Kover PX. 2007. Antagonistic pleiotropic effects reduce the potential adaptive value of the FRIGIDA locus. Proc Natl Acad Sci USA. 104:16986–16991.

Schmid KJ, Ramos-Onsins S, Ringys-Beckstein H, Weisshaar B, Mitchell-Olds T. 2005. A multilocus sequence survey in *Arabidopsis thaliana* reveals a genome-wide departure from a neutral model of DNA sequence polymorphism. Genetics. 169:1601–1615.

Schmid KJ, Törjék O, Meyer RC, Schmuths H, Hoffmann M, Altmann T. 2006. Evidence for a large-scale population structure of *Arabidopsis thaliana* from 5 genome-wide SNP markers. Theor Appl Genet. 112:1104–1114.

Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H. 1990. Genetic control of flower development by homeotic genes in *Antirrhinum majus*. Science. 250:931–936.

Scortecci KC, Michaels SD, Amasino RM. 2001. Identification of a MADS-box gene, FLOWERING LOCUS M, that represses flowering. Plant J. 26:229–236.

Sheldon CC, Burn JE, Perez PP, Metzger J, Edwards JA, Peacok WJ, Dennis ES. 1999. The *FLF* MADS box gene: a repressor of flowering in *Arabidopsis* regulated by vernalization and methylation. Plant cell. 11:445–458.

Sherman-Broyles S, Boggs N, Farkas A, Liu P, Vrebalov J, Nasrallah ME, Nasrallah JB. 2007. S locus genes and the evolution of self-fertility in *Arabidopsis thaliana*. Plant Cell. 19:94–106.

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M, Dean C. 2005. Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of *Arabidopsis*. Plant Physiol. 138:1163–1173.

Shiotani H, Fujikawa T, Ishihara H, Tsuyumu S, Ozaki K. 2007. A pthA homolog from *Xanthomonas axonopodis* pv. citri responsible for host-specific suppression of virulence. J Bacteriol. 189:3271–3279.

Simon M, Loudet O, Durand S, Berard A, Brunel D, Sennesal F-X, Durand-Tardif M, Pelletier G, Camilleri C. 2008. Quantitative trait loci mapping in five new large recombinant inbred line populations of *Arabidopsis thaliana* genotyped with consensus single-nucleotide polymorphism markers. Genetics. 178:2253–2264.

Smartt J, Graubert TA, Cahan P, et al. (11 co-authors). 2007. A high-resolution map of segmental DNA copy number variation in the mouse genome. PLoS Genet. 3:e3.doi:10.1371.

Song W-Y, Pi L-Y, Wang G-L, Gardner J, Holsten T, Ronald PC. 1997. Evolution of the rice *Xa21*disease resistance gene family. Plant cell. 9:1279–1287.

Stankiewicz P, Lupski JR. 2002. Genome architecture, rearrangements and genomic disorders. Trends Genet. 18:74–82.

Sun Q, Collins NC, Ayliffe M, Smith SM, Drake J, Pryor T, Hulbert SH. 2001. Recombination between paralogues at the Rp1 resistance locus in maize. Genetics. 158:423–438.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol. 24:1596–1599.

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

Theissen G, Kim J, Saedler H. 1996. Classification and phylogeny of the MADS-box multigene family suggests defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. J Mol Evol. 43:484–516.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 19:2325–2327.

Ungerer MC, Halldorsdottir SS, Modliszewski JL, Mackay TFC, Purugganan MD. 2002. Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. Genetics. 160:1133–1151.

Ungerer MC, Halldorsdottir SS, Purugganan MD, Mackay TFC. 2003. Genotype–environment interactions at quantitative trait loci affecting inflorescence development in *Arabidopsis thaliana*. Genetics. 165:353–365.

Wang W, Zhang J, Alvarez C, Llopart A, Long M. 2000. The origin of the *Jingwei* gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. Mol Biol Evol. 17:1294–1301.

Werner JD, Borevitz JO, Uhlenhaut NH, Ecker JR, Chory J, Weigel D. 2005. FRIGIDA-independent variation in flowering time of natural *Arabidopsis thaliana* accessions. Genetics. 170:1197–1207.

Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR, Chory J, Weigel D. 2005. Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci USA. 102:2460–2465.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS. 13: 555–556.

Yokoyama S, Yokoyama R. 1989. Molecular evolution of human visual pigment genes. Mol Biol Evol. 6:186–197.

Yu J, Pressoir G, Briggs W, et al. (11 co-authors). 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 38:203–208.

Kenneth Wolfe, Associate Editor

Accepted December 22, 2008