Molecular Population Genetics of Redundant Floral-Regulatory Genes in *Arabidopsis thaliana*

Richard C. Moore, Sarah R. Grant,† and Michael D. Purugganan**

*Department of Genetics, North Carolina State University, Raleigh; and †Department of Biology, University of North Carolina, Chapel Hill

Functional redundancy between duplicated genes is predicted to be transitory, as one gene either loses its function or gains a new function, or both genes accrue degenerative, yet complimentary mutations. Yet there are many examples where functional redundancy has been maintained between gene duplicates. To determine whether selection is acting on functionally redundant gene duplicates, we performed molecular evolution and population genetic analyses between two pairs of functionally redundant MADS-box genes from the model plant *Arabidopsis thaliana*: *SEPALLATA1* (*SEP1*) and *SEPALLATA2* (*SEP2*), involved in floral organ identity, and *SHATTERPROOF1* (*SHP1*) and *SHATTERPROOF2* (*SHP2*), involved in seed shattering. We found evidence for purifying selection acting to constrain functional divergence between paralogous genes. The protein evolution of both pairs of duplicate genes is functionally constrained, as evidenced by K_a/K_s ratios of 0.16 between paralogs. This functional constraint is stronger in the highly conserved DNA-binding and protein-binding MIK region than in the C-terminal region. We also assayed the evolutionary forces acting between orthologs of the *SEP* and *SHP* genes in *A. thaliana* and the closely related species, *Arabidopsis lyrata*. Heterogeneity analyses of the polymorphism-to-divergence ratio indicate selective sweeps have occurred within the transcriptional unit of *SHP1* and the promoter of *SHP2* in the *A. thaliana* lineage. Similar analyses identified a significant reduction in polymorphism within the *SEP1* locus, spanning the 3′ region of intron 1 to exon 3, that may represent an intragenic sweep within the *SEP1* locus. We discuss whether the evolutionary forces acting on *SEP1* and *SEP2* versus *SHP1* and *SHP2* vary according to their position in the floral developmental pathway, as found with other floral-regulatory genes.

Introduction

Functional redundancy is often invoked when the genetic knockout of a single gene gives no discernible mutant phenotype, especially when a related duplicate gene exists (Pickett and Meeks-Wagner 1995). The widespread absence of mutant phenotypes in gene knockout experiments in model organisms suggests functional redundancy may play an important role in the evolution and development of eukaryotic organisms (Winzeler et al. 1999; Giaever et al. 2002; Kamath et al. 2003; Simmer et al. 2003). Furthermore, the prevalence of functional redundancy suggests that it is an important component of genetic robustness, a process that buffers an organism's phenotype against deleterious mutations (Krakauer and Nowak 1999; Wagner 1999; Gu 2003). The compensatory activities of both related duplicated loci and unrelated genes in alternate pathways contribute to genetic robustness (Wagner 1999; Kitami and Nadeau 2002; Gu 2003). Although the precise contribution of these two modes of functional compensation to robustness still needs to be resolved, the abundance of duplicate genes in eukaryotic genomes suggests gene duplication plays a crucial role (Wolfe and Shields 1997; *Caenorhabditis elegans* Sequencing Consortium 1998; Adams et al. 2000; Arabidopsis Genome Initiative 2000; Venter et al. 2001). Indeed, in a viability comparison of knockouts in single copy and duplicate copy genes in yeast, 59% of the knockouts that yield a weak or a null phenotype are in duplicated loci (Gu et al. 2003).

If redundant genes provide no fitness advantage, then redundancy is predicted to be evolutionarily unstable (Nowak et al. 1997). Although redundancy is often the immediate fate of duplicated loci, it is usually followed by the loss of one gene or the functional diversification of one or both of the duplicate loci (Clark 1994; Walsh 1995; Wagner 1998; Lynch and Force 2000; Lynch et al. 2001; Walsh 2003). Indeed, the transient nature of functional redundancy has prompted many to address whether functional redundancy can be selectively maintained (Nowak et al. 1997; Wagner 1999), Selective retention of redundancy is often closely linked to the role duplicate genes play in genetic robustness. This is especially the case for developmental genes, where redundancy guards against developmental error. If the mutation rates of both duplicate genes are lower than the rate of developmental error associated with their duplicate, the fitness cost of developmental error will be substantially reduced, and redundancy will be selectively maintained (Cooke et al. 1997; Nowak et al. 1997; Krakauer and Nowak 1999). Moreover, a developmental pathway with multiple, redundant genes might be more resistant to deleterious mutations than one with only a few such genes (Wagner 1999).

An example of a developmental pathway with multiple layers of redundancy is the floral developmental pathway as described for the model plant, *Arabidopsis thaliana*. Redundant duplicate genes act at all stages during floral development in this species, from the initial stages of floral meristem identity to the specification of organ identity and seed shattering (Kempin, Savidge, and Yanofsky 1995; Ferrandiz et al. 2000; Liljegren et al. 2000; Pelaz et al. 2000, 2001; Pinyopich et al. 2003). To assess how selection acts on functionally redundant duplicate developmental genes, we analyzed the molecular population genetics of two paralogous pairs of floral-regulatory genes from *A. thaliana*: *SEPALLATA1* (*SEP1*; also *AGL2*, At5g15800) and *SEPALLATA2* (*SEP2*; also *AGL4*, At3g02310) (Pelaz et al. 2000, 2001) and *SHATTERPROOF1* (*SHP1*; also *AGL1*,

Key words: *SEPALLATA*, *SHATTERPROOF*, paralogs, functional redundancy, floral development.

E-mail: rcmoore@unity.ncsu.edu.

Mol. Biol. Evol. 22(1):91–103. 2005
doi:10.1093/molbev/msh261
Advance Access publication September 15, 2004

At3g58780) and *SHATTERPROOF2* (*SHP2*; also *AGL5*, At2g42830) (Liljegren et al. 2000; Pinyopich et al. 2003).

SEP1/*SEP2* and *SHP1*/*SHP2* paralogs are members of the large family of type II MADS-box transcriptional regulators (Alvarez-Buylla et al. 2000; Parenicova et al. 2003). Both gene pairs are considered to be functionally redundant; single-gene knockouts give no observable phenotypes, and paralogs share similar expression domains (Ma, Yanofsky, and Meyerowitz 1991; Flanagan and Ma 1994). Although both gene pairs are involved in floral development, they act at different stages of the reproductive developmental pathway. *SEP1* and *SEP2*, along with the partially redundant and evolutionarily related *SEPALLATA3* (*SEP3*), are involved in the establishment of floral organ identity (Pelaz et al. 2000, 2001), whereas *SHP1* and *SHP2* function downstream of *SEP1* and *SEP2* in reproductive development, controlling valve margin and dehiscence zone differentiation in the *A. thaliana* silique, or seedpod (Liljegren et al. 2000). *SHP1* and *SHP2* also share partially overlapping function in carpal and ovule identity with the closely related floral homeotic genes *AGAMOUS* (*AG*) and *SEEDSTICK* (*STK*) (Pinyopich et al. 2003).

To characterize the evolutionary forces acting on these functionally redundant duplicate genes, we first conducted a molecular evolution analysis of protein divergence between paralogous duplicate genes. Despite their relatively long coexistence (~26 Myr), we find no evidence that either gene is evolving new functions or becoming a pseudogene. Instead, we found evidence that purifying selection is the primary evolutionary force acting on both sets of paralogs, possibly because of functional constraints on protein evolution. This observation is consistent with developmental genetic analyses, which indicate each gene can substitute for a loss of function in its paralog. We also analyzed patterns of within-species and between-species sequence variation at each duplicate locus. *SEP1* and *SEP2* exhibit average levels of nucleotide polymorphism compared with other *A. thaliana* nuclear genes. However, a heterogeneity analysis of the polymorphism-to-divergence ratio indicates that *SEP1* has significantly reduced levels of polymorphism in a region spanning intron 1 to exon 3, suggestive of an intragenic selective sweep. In contrast to the *SEPALLATA*s, both *SHP1* and *SHP2* have relatively low levels of intraspecific polymorphism. Heterogeneity analyses indicate this reduction in polymorphism is the consequence of contrasting regional selective sweeps, one in the transcriptional unit of *SHP1* and another in the promoter of *SHP2*. Because patterns of molecular evolution of floral-regulatory genes differ according to their position in the floral developmental pathway (Olsen et al. 2002), we also address the hypothesis that the evolutionary forces acting on the *SEPALLATA*s and *SHATTERPROOF*s vary according to their operational position in the floral developmental pathway.

Materials and Methods

Isolation and Sequencing of Alleles

Genomic DNA was isolated from young leaves of 25 *A. thaliana* accessions (table 1 in Supplementary Material online) and one *Arabidopsis lyrata* individual using the Plant DNeasy Mini Kit (Qiagen, Valencia, Calif.). The *A. lyrata* individual was grown from seed isolated from a Karhumaki, Russia population and was provided by O. Savolainen (University of Oulu, Oulu, Finland) and Helmi Kuittinen (University of Barcelona, Barcelona, Spain). Approximately 1 kb of the promoter, the 5′ untranslated region (UTR), all exons and introns, and less than 250 bp of the 3′ UTR were amplified and sequenced for all genes.

PCR primers were designed based on the Col-0 gene sequences using Primer3 (http://www-genome.wi.mit.edu/genome_software/other/primer3.html). Primers were designed to be specific to each paralog, and, for *A. thaliana*, at least one primer was designed to anneal to noncoding regions (tables 2 and 3 in Supplementary Material online). PCR of *A. thaliana* and *A. lyrata* samples was performed with *Taq* DNA polymerase (Roche, Indianapolis, Ind.) using manufacturer's protocols. DNA fragments amplified from *A. thaliana* were purified using the QIAquick Gel Extraction Kit (Qiagen) and directly sequenced. Amplified *A. lyrata* products were subcloned using the TA TOPO PCR Cloning Kit (Invitrogen, Carlsbad, Calif.), and plasmid DNA from five to six independent clones was sequenced. DNA sequencing was conducted at the North Carolina State University Genome Research Laboratory with a Prism 3700 96-capilary automated sequencer (Applied Biosystems, Foster City, Calif.). All polymorphisms were visually confirmed, and ambiguous polymorphisms were rechecked by PCR reamplification and sequencing. GenBank accession numbers for these genes are AY72 7576 to AY727670.

Molecular Evolution and Population Genetic Data Analysis

Sequences were visually aligned against the *A. thaliana* sequence previously identified in the *Arabidopsis* whole-genome sequence (Arabidopsis Genome Initiative 2000). The *A. lyrata* ortholog was used as the out-group in the analyses. MEGA version 2.1 (Giaever et al. 2002) was used to estimate interspecific nucleotide sequence divergence distances from synonymous sites and nonsynonymous sites with the Nei-Gojobori model and from noncoding, silent sites using the Kimura two-parameter model, with standard error determined from 500 bootstrap replicates. Divergence times for paralogs were determined using the divergence distances at synonymous sites and the methods described in Li (1997). Levels of silent-site nucleotide diversity per site were estimated as π (Nei 1987) and θ_W (Watterson 1975). Polymorphism and sliding window K_a/K_s analyses were conducted using DnaSP version 3.99 (Rozas and Rozas 1999). DnaSP 3.99 was also used to perform tests of selection, including Tajima's *D* statistic (Tajima 1989) and the McDonald-Kreitman test (McDonald and Kreitman 1991).

Codon-based models of selection were analyzed using the codeml program of PAML (Yang 1997). Maximum-likelihood values for the discrete M0 (no selection) and M3 (with selection) models, as well as continuous beta distribution models M7 (no selection) and M8 (with selection) were obtained and a maximum-likelihood ratio test was used to compare the significance of the M0 versus

Table 1
Sequence Distances Between Orthologs and Between Paralogs

	K_{sil}	K_s	K_a	K_a/K_s
Orthologs				
A. thaliana SEP1 vs. *A.lyrata SEP1*	0.112±0.006	0.108±0.024	0.001±0.001	0.01
A. thaliana SEP2 vs. *A. lyrata SEP2*	0.099±0.007	0.059±0.017	0.009±0.004	0.15
A. thaliana SHP1 vs. *A.lyrata SHP1*	0.097±0.006	0.077±0.021	0.009±0.004	0.12
A. thaliana SHP2 vs. *A lyrata SHP2*	0.078±0.005	0.090±0.022	0.008±0.004	0.09
Paralogs				
A. thaliana SEP1 vs. *A.thaliana SEP2*	n.d.	0.409±0.040	0.065±0.010	0.16
A. thaliana SHP1 vs. *A.thaliana SHP2*	n.d.	0.438±0.044	0.071±0.012	0.16

NOTE.—K_{sil} = average distance between groups at silent sites in introns; K_s = average distance between groups at synonymous sites; K_a = average distance between groups at nonsynonymous sites (standard error determined by 500 random bootstrap replicates); n.d.= not determined because of unalignable intron sequences.

M3 and M7 versus M8 models. Differential rates of codon evolution were tested across trees containing five taxa that included each paralog, their orthologs from *A. lyrata*, and either *SEPALLATA3*(*SEP3*) for the *SEP* paralogs or *AGAMOUS* (AG) for *SHP* paralogs as an out-group. Sequences were aligned using ClustalX version 1.8 (Thompson et al. 1997) and trees were assembled in PAUP* version 4.0 (Swofford 2002).

The Hudson-Kreitman-Aguade (HKA) test (Hudson, Kreitman, and Aguade 1987) was conducted using silent-site differences. Individual HKA tests were conducted for each locus against six neutrally evolving reference loci using DnaSP 3.99. The following loci were chosen as the reference loci in these tests: *AP1* (Olsen et al. 2002), *AP3* and *PI* (Purugganan and Suddith 1999), *CAL* (Purugganan and Suddith 1998), and *F3H* and *FAH1* (Aguade 2001). Probabilities of each of these tests were corrected using the Simes method for combining probabilities from multiple tests (Simes 1986). Tests of heterogeneity in the polymorphism-to-divergence ratio across each gene were performed using DNA Slider (McDonald 1998). For each test, 1,000 simulations at R values of 2, 4, 8, 16, and 32 were run, and the highest *P* value was reported.

Results

SEP and *SHP* Paralogs Arose During the Last Whole-Genome Duplication Event

We dated the timing of the duplication events that gave rise to *SEP1*/*SEP2* and *SHP1*/*SHP2* paralogs to determine how long these paralogs have existed in a functionally redundant state. The silent-site nucleotide divergence between orthologs is comparable between duplicate gene pairs, ranging from 0.078 to 0.112 (table 1). Thus, the underlying neutral mutation rate is roughly equivalent between duplicates.

The timing of the duplication events leading to *SEP1*/*SEP2* and *SHP1*/*SHP2* paralogs was estimated using synonymous-site nucleotide divergence (K_s), as intron sequences could not be reliably aligned between paralogs. The levels of K_s are comparable between both paralogous gene pairs ($K_s = 0.409 \pm 0.040$ for *SEP1* versus *SEP2*, and 0.438 ± 0.044 for *SHP1* versus *SHP2* [table 1]). The molecular clock was calibrated using the levels of synonymous-site nucleotide divergence between the orthologs of these genes in *A. thaliana* and *A. lyrata* (table 1) and using a divergence date of 5.2 MYA for these two species (Koch, Haubold, and Mitchell-Olds 2000).

Using this clock rate, the duplication events that led to the establishment of these redundant gene pairs appear to have occurred within the same time frame, with a divergence date of 25.2 ± 10.1 MYA for the *SEP1*/*SEP2* divergence, and 27.7 ± 10.7 MYA for the *SHP1*/*SHP2* split (Li 1997). Given the similarities in divergence dates between these two floral gene duplicates and their dispersed locations within the genome (*SEP1* on chromosome V, *SEP2* on III, *SHP1* on chromosome III, and *SHP2* on chromosome II), it is likely that a single large-scale duplication event, followed by chromosomal reshuffling, gave rise to both sets of paralogs. In accordance with this possibility, each paralogous pair is found colinearly in large duplicated segments of the *Arabidopsis* genome (Arabidopsis Genome Initiative 2000; Vision, Brown, and Tanksley 2000; Bowers et al. 2003; Blanc and Wolfe 2004). Moreover, this estimate of divergence time is congruent with some of the lower estimates of the last genome duplication event of *A. thaliana*, which was believed to have occurred close to the emergence of crucifers, 24 to 48 MYA (Blanc, Hokamp, and Wolfe 2003; Bowers et al. 2003; Ermolaeva et al. 2003). Thus, *SEP* and *SHP* paralogs have coexisted in a redundant state since their emergence, most likely, during the last whole-genome duplication event.

Purifying Selection Acts on *SEP1/2* and *SHP1/2* Proteins

It is notable that functional redundancy has been maintained between *SEP* and *SHP* paralogs for approximately 26 Myr, since redundancy is predicted to be lost on the road to either gene diversification or gene loss. It is possible that subtle differences in protein function may exist between paralogous proteins that are undetectable in loss-of-function mutation analyses but are detectable by molecular evolutionary analyses. Therefore, we examined patterns of protein evolution between orthologs and paralogs of the *SEP* and *SHP* proteins to determine if there is relaxed and/or selective divergence in amino acid sequence.

Orthologs of these genes were sequenced from one individual of *A. lyrata* to allow for interspecific sequence

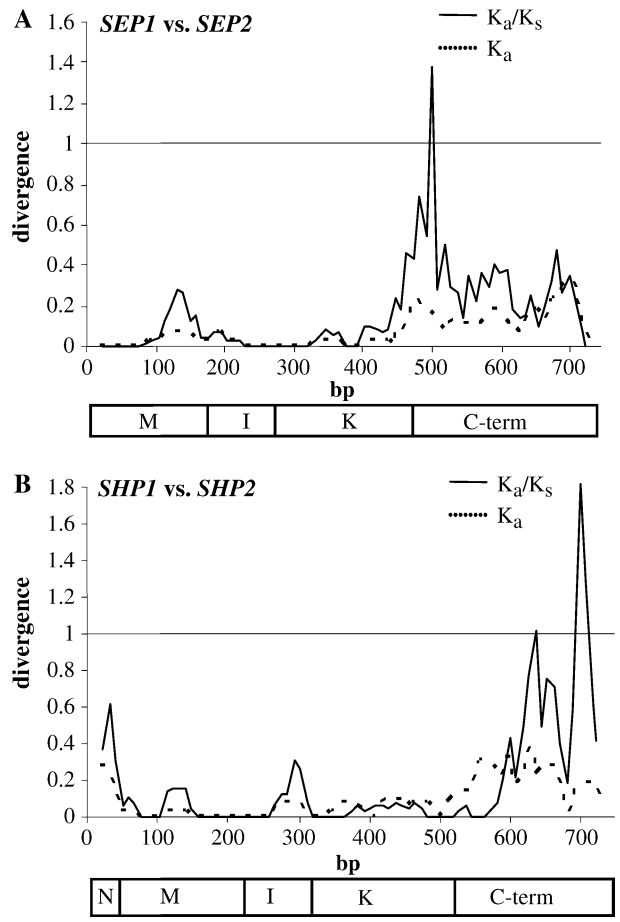


FIG. 1.—Sliding window plots of average K_a and K_a/K_s for all pairwise comparisons between alleles of (*A*) *SEP1* and *SEP2* and (*B*) *SHP1* and *SHP2*. The boundaries of the MADS-box (M), intervening region (I), K domain (K), C-terminal domain (C-term), and, for *SHP*, the flanking N-terminal domain (N) are indicated below the *x*-axis. Window size = 45 bp; step size = 9 bp.

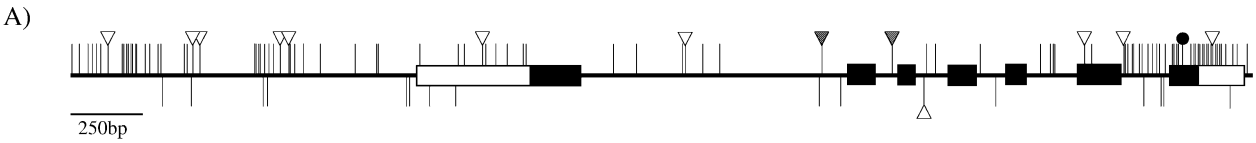
comparisons in relative levels of the replacement (K_a) and synonymous (K_s) site nucleotide divergence rate (table 1). A ratio of K_a to K_s greater than 1, the neutral expectation, serves as an indicator of positive selection, whereas a ratio less than 1 indicates purifying selection caused by functional constraint on protein evolution. The resulting K_a/K_s ratio for the *SEP2* orthologs ($K_a/K_s = 0.15$) is an order of magnitude greater than that for the *SEP1* orthologs ($K_a/K_s = 0.01$), which suggests that the protein sequence of *SEP2* evolved under weaker selective constraint in these two Brassicaceae species. Both loci, however, have K_a/K_s values much lower than 1, indicating amino acid replacements between orthologs are functionally constrained. The K_a/K_s values for *SHP1* ($K_a/K_s = 0.12$) and *SHP2* ($K_a/K_s = 0.09$) orthologs are similar to the value for *SEP2*. Thus, amino acid replacements between *SHP1* and *SHP2* orthologs are similarly constrained. In accordance with the neutral theory of molecular evolution, the ratio of nonsynonymous to synonymous mutations between species should be equivalent to the ratio of nonsynonymous to synonymous changes within species (McDonald and Kreitman 1991). This relationship can be tested with the McDonald-Kreitman test of protein evolution (McDonald and Kreitman 1991). This test is not significant for all genes ($P < 0.18$ to 0.61; data not shown), suggesting orthologous protein sequences are evolving neutrally.

We also explored the divergence of amino acid sequence between paralogs occurring after the duplication events that established the redundant gene pairs. The average K_a/K_s ratio between *SEP1* versus *SEP2* and *SHP1* versus *SHP2* paralogs is 0.16 (table 1), indicating purifying selection, in the form of functional constraint on amino acids, is acting to suppress protein divergence. The K_a/K_s ratios between *SEP* and *SHP* paralogs, however, reflect the average value across the entire coding sequence. It is still possible that certain functional regions, and even specific codons, exhibit the molecular signature of diversifying selection.

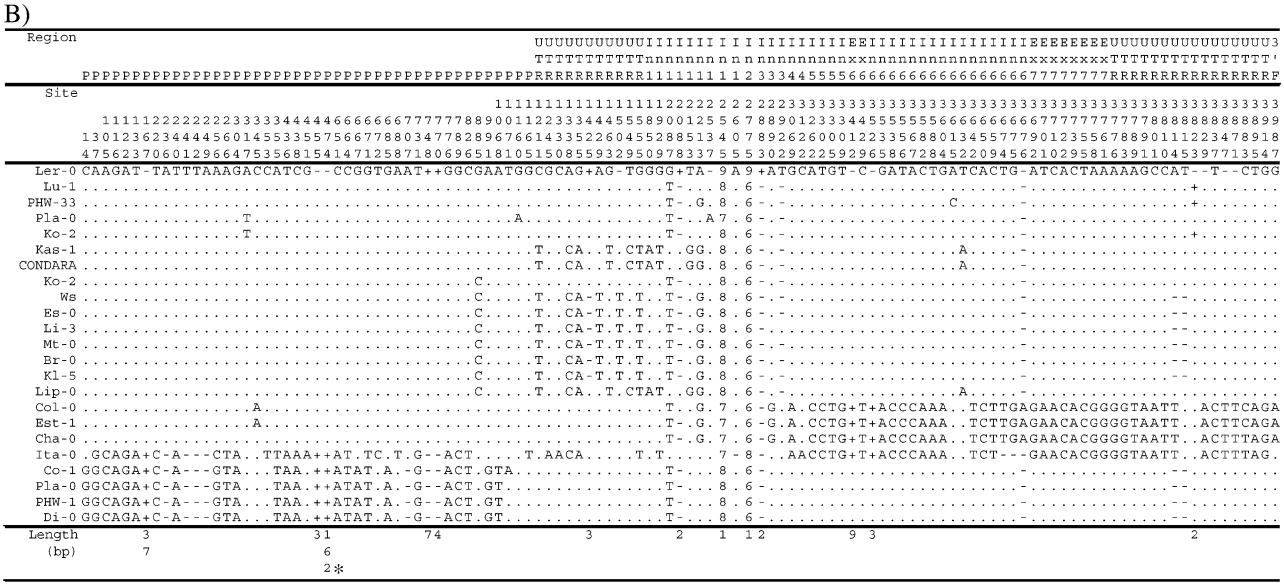
To test this possibility, we performed a sliding window analysis to detect fluctuations in K_a/K_s in the different functional domains of these MADS-box genes. We detected heterogeneity in this ratio between both pairs of paralogs (fig. 1). Between *SEP* paralogs, the K_a/K_s ratio is elevated in the C-terminal domain, which contains the transcriptional activation domain, relative to the DNA-interacting and protein-interacting MIK region. This ratio exceeds 1, the neutral expectation, near the junction between the K and C-terminal domains. A similar increase in the K_a/K_s ratio was found in the C-terminal domain between *SHP* paralogs, with a sharp increase of the K_a/K_s ratio to greater than 1 at the C-terminus. There is also an increase in K_a/K_s between *SHP* paralogs at the N-terminus in a region of uncharacterized function that flanks the MADS-box domain.

Maximum-likelihood analysis of codon-based models of evolution did not detect any specific amino acid residues under diversifying selection in either *SEP* or *SHP* paralogs, although for both sets of paralogs, the M3 model, which allows for multiple, discrete ω ($= K_a/K_s$) classes, fit the data significantly better than does the M0 model, which allows for only one ω class ($P < 0.001$). For *SEP* paralogs, the largest proportion of sites (72%) are under purifying selection ($\omega_0 = 0.012$), and a smaller proportion of sites exhibit relaxed functional constraint ($\omega_{1,2} = 0.308$). For *SHP* paralogs, an approximately equal proportion of sites (52%) are subject to strong purifying selection ($\omega_0 = 0.0001$) compared with the proportion (48%) that exhibit relaxed evolutionary constraint ($\omega_{1,2} = 0.247$). For *SEP* paralogs, 39% of codons in the C-terminal region exhibited relaxed functional constraint, compared with 14% of codons in the MIK region. For *SHP* paralogs, 74% of codons found in either the N-terminal region (N) or C-terminal region exhibit relaxed constraint compared with 16% of codons found in the MIK region. These results signify a relaxation of functional constraint in the C-terminal regions of these paralogous transcription factors, consistent with previous studies that show the C-terminus tends to be more functionally divergent than the more highly conserved MIK region (Kramer, Dorit, and Irish 1998; Lawton-Rauh, Buckler, and Purugganan 1999; Purugganan and Suddith 1999; Lamb and Irish 2003; Litt and Irish 2003; Vandenbussche et al. 2003). However, it is evident that purifying selection is the

A)



B)



C) Miniature inverted repeat (MITE) sequence:

TTA*GGGTGTGAATGGTGA*CATACACCGCACCCAAAACCACATCAATAGTGTTGTAACAATATTAAAATCTTTACATATACATATGTATCTATGTATTTTCGttACTGTTTGAACCGCACCCCACTTCTCCCGCACCACAAAACGCAA*TCACCATTCACACCC***TTA**

FIG. 2.—(*A*) Gene model and distribution of polymorphic sites for *SEP1*. Black boxes indicate exons; white boxes indicate 5′ and 3′ untranslated regions. Polymorphic sites found in two or more accessions are indicated by lines above the gene model, whereas single polymorphisms are indicated by lines below the gene model. Indels are indicated by white triangles and microsatellites by gray triangles. Nonsynonymous polymorphisms are indicated by a black circle. (*B*) Polymorphism table for *SEP1*. The nucleotide position and region of each polymorphism are indicated (P = promoter, UTR = untranslated region, In = intron, and Ex = exon). The length of indels is indicated below each occurrence. The number of repeat units within microsatellites is indicated by a number within the table, and the base length of one repeat unit is indicated below each occurrence. A dot represents an equivalent bp relative to the reference sequence. A plus is an insertion of more than 2 bp, and a minus represent a deletion. Asterisk indicates 162-bp MITE insertion. (*C*) Sequence of MITE insertion from Co-1. Terminal-direct repeats are in bold; inverted repeats are underlined.

dominant evolutionary force acting on these duplicate gene pairs.

SEP1 and *SEP2* Have Average Levels of Intraspecific Nucleotide Diversity

We performed a molecular population genetic analysis of within-species nucleotide polymorphism and between-species divergence for *SEP1*/*SEP2* and *SHP1*/*SHP2* to better understand the evolutionary forces acting on these redundant genes at the population-level and the species-level. Importantly, these analyses are not limited to protein-coding sequences, unlike molecular evolutionary studies of protein divergence.

A total of 23 *SEP1* and *SEP2* alleles were isolated from a collection of *A. thaliana* ecotypes sampled primarily from Europe. Approximately 4.0 kb of each *SEP1* allele was sequenced, spanning exons 1 to 7 and including 1.2 kb of the promoter, 350 bp of the 5′ UTR, and 50 bp of the 3′ UTR (fig. 2*A*). Approximately 3.4 kb of sequence was obtained for each *SEP2* allele, including 900 bp of the promoter, 350 bp of the 5′ UTR, and the entire coding region (fig. 3*A*).

These two redundant floral developmental genes show similar levels of nucleotide polymorphism. In *SEP1*, there are a total of 88 polymorphic nucleotide sites (fig. 2). Of these, there is one replacement polymorphism and eight synonymous polymorphisms. There are also 11 insertion/deletion (indel) polymorphisms, all in noncoding regions, which range in size from 2 bps to a 162 bp miniature inverted-repeat transposable element (MITE [Casacuberta et al. 1998; Casacuberta and Santiago 2003]) (fig. 2*B* and *C*). This MITE is flanked by TTA direct repeats and inverted repeats that are 60% identical to those of the *Tourist* family of MITEs, previously identified in cereals (Bureau and Wessler 1992, 1994). A Blast search of this 162-bp MITE against the *Arabidopsis* genome yielded 47 unique sequences with an E value less than 10^{-10}. We detected no similarity of this MITE to genomes outside of *Arabidopsis*. Within the *A. thaliana* genome, all highly similar sequences were found within 3 kb upstream of coding sequence, and 18 of these were found within the 1-kb upstream region. Aside from these indels, there are also two mononucleotide microsatellite repeats found in the first and second introns. The *SEP2* alleles have a total of 80 nucleotide polymorphisms, with one replacement and four

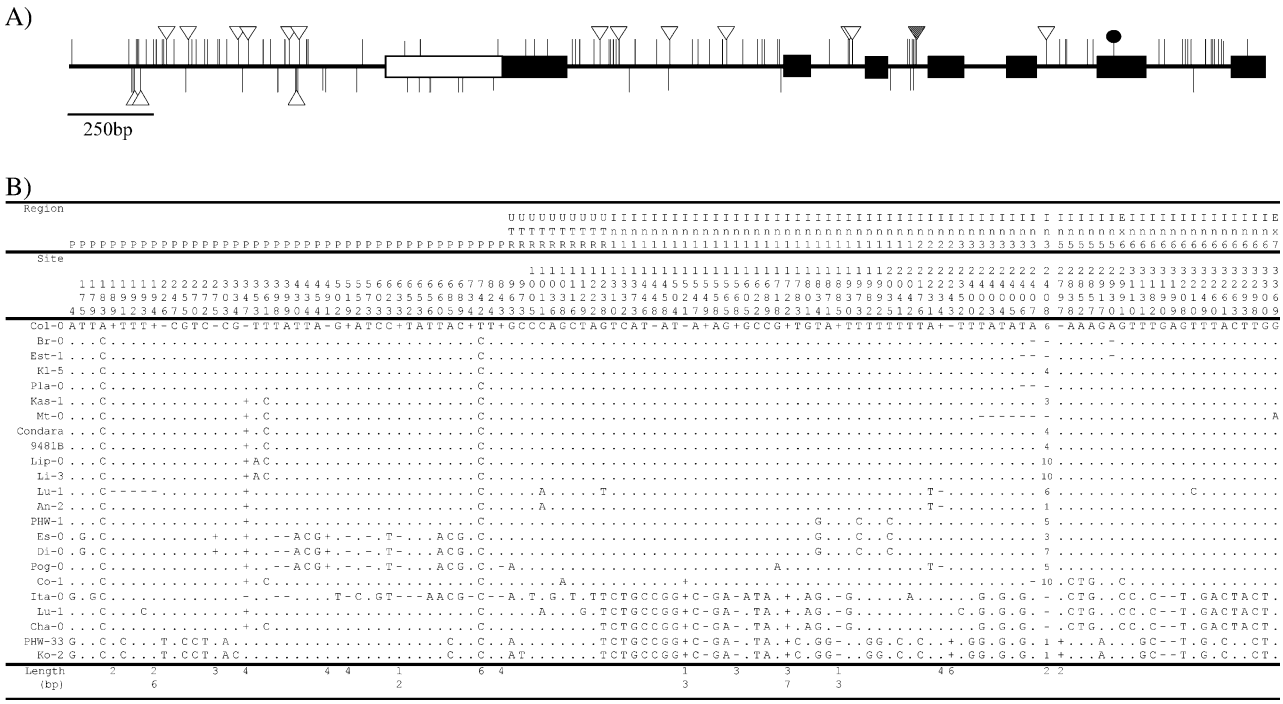


FIG. 3.—(*A*) Gene model and distribution of polymorphic sites for *SEP2*. Symbols are the same as described in figure 1. (*B*) Polymorphism table for *SEP2*. Arrangement and symbols are the same as described for figure 1.

synonymous polymorphisms (fig. 3). Sixteen indel polymorphisms occur in *SEP2* in the noncoding region, ranging in size from 2 to 37 bps, as well as one dinucleotide microsatellite (fig. 3).

Estimates of silent-site nucleotide variation are comparable between these two redundant genes (table 2). The estimate of *SEP1* intraspecific nucleotide diversity for silent sites, π, is 0.0080, whereas *SEP2* has a π of 0.0067 (table 2). These estimates are comparable to the mean nucleotide diversity ($\pi = 0.0074$) calculated for a collection of previously published nuclear genes in *A. thaliana* (Yoshida et al. 2003). This level of nucleotide diversity is characteristic of neutrally evolving genes in *A. thaliana* (Purugganan and Suddith 1998; Purugganan and Suddith 1999; Aguade 2001; Olsen et al. 2002).

SHP1 and *SHP2* have Reduced Levels of Intraspecific Nucleotide Diversity

A total of 24 *SHP1* and 25 *SHP2* alleles were isolated from *A. thaliana* ecotypes. As with the *SEP* alleles, sequences from the entire coding region, a portion of the promoter, the 5′ UTR, and the 3′ UTR were obtained for each gene. Approximately 4.6 kb of each *SHP1* allele was sequenced, spanning exons 1 to 7 and including 1.2 kb of the promoter, 600 bp of the 5′ UTR, and 50 bp of the 3′ UTR (fig. 4*A*). Approximately 4.9 kb of sequence was obtained for each *SHP2* allele, including 1.2 kb of the promoter, 200 bp of the 5′ UTR, the entire coding region, and 100 bp of the 3′ UTR (fig. 5*A*).

In *SHP1*, there are a total of 36 polymorphic nucleotide sites, all of which are found in noncoding sequence (fig. 4). There are also four indel polymorphisms, all in introns, which range in size from 2 to 35 bp, as well as one dinucleotide and two mononucleotide microsatellites, all of which are found in the first intron. The *SHP2* alleles have a total of 60 nucleotide polymorphisms, with two replacement and two synonymous polymorphisms (fig. 5). Fifteen indel polymorphisms occur in *SHP2*, all in introns, and range in size from 2 to 32 bp. There are also two dinucleotide microsatellites and one mononucleotide microsatellite, all in the first intron.

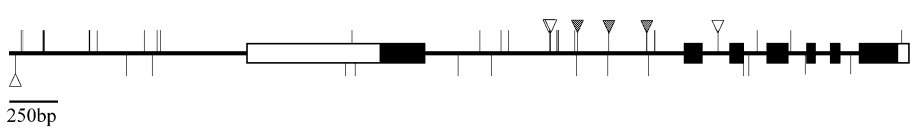
Estimates of silent-site nucleotide diversity for these two redundant floral developmental genes are twofold to fivefold lower than that found at the *SEP1* and *SEP2* loci (table 2). The value of π for *SHP1* is 0.0015, and *SHP2* has a π of 0.0035 (table 2). Both of these values are lower than the mean nucleotide diversity of 0.0074 reported for other *A. thaliana* genes (Yoshida et al. 2003). Reduced levels of nucleotide diversity can indicate positive selection for an advantageous haplotype in the form of a selective sweep that eliminates neutral variation linked to the advantageous mutation.

Table 2
Summary of Nucleotide Diversity

Gene	n	Length (bp)	S	π_s	π_{syn}	π_{nonsyn}	θ_W	Tajima's *D* Statistic
SEP1	23	3987	88	0.0080	0.0145	0.0005	0.0076	0.133 (NS)
SEP2	23	3424	80	0.0067	0.0068	0.0005	0.0081	−0.687 (NS)
SHP1	25	4629	36	0.0015	0.0000	0.0000	0.0024	−1.48 (NS)
SHP2	25	4921	60	0.0035	0.0033	0.0003	0.0037	−0.268 (NS)

NOTE.—S = number of nucleotide polymorphisms; π_s = estimated pairwise silent-site diversity; π_{syn} = estimated pairwise synonymous-site diversity; π_{nonsyn} = estimated pairwise nonsynonymous-site diversity; θ_W = estimated nucleotide diversity based on number of segregating sites; NS = not significant.

A)



B)

Region	P	P	P	P	P	P	P	P	P	P	P	P	P	P	U T R	U T R	U T R	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	I n 1	U T R
Site															1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	4	4	4	4
					1	1	1	4	4	5	6	6	8	8	8	8	7	7	7	3	4	4	5	5	7	7	8	8	9	9	9	0	0	2	2	3	6	7	8	8	0	1	3	5
	3	6	6	7	7	7	1	1	5	0	9	3	6	7	3	6	8	0	4	5	8	3	6	8	8	2	2	2	1	2	2	8	8	8	8	2	5	7	0	5	2	0	6	9
	5	2	7	5	6	7	4	5	7	4	6	9	3	7	6	8	0	4	5	9	9	9	2	7	1	2	3	1	1	3	0	1	1	9	4	3	8	6	0	2	1	7	7
Col-0	G	A	C	A	G	T	A	A	G	T	C	G	C	G	T	C	C	G	A	A	A	A	−	+	C	G	G	−	G	8	G	18	13	C	−	+	A	A	C	C	C	T	T
Mt-0	+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	8	.	18	13	.	.	.	.	.	.	.	.	.	.
PHW-1	−	.	.	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	−	T	A	A	.	.	9	.	16	14	.	.	.	.	.	.	.	.	.	C
Est-1	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	−	T	A	A	.	.	7	.	13	14	T	.	.	T	.	.	.	.	.	C
Li-3	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	8	.	16	14	.	.	.	.	.	.	.	.	.	C
Lu-1	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	9	.	16	14	.	.	.	.	.	.	.	.	.	C
P0g-0	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	10	.	16	14	.	.	.	.	.	.	.	.	.	C
Tsu-1	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	8	.	16	14	.	.	.	.	.	.	.	.	.	C
PHW-33	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	9	.	16	14	.	.	.	.	.	.	.	.	.	C
Kl-5	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	A	.	.	C	G	.	.	.	.	.	.	.	8	.	15	14	.	.	.	.	.	.	.	.	.	C
Kas-1	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	A	10	.	19	14	.	.	.	.	.	.	.	.	.	C
Condara	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	16	.	19	14	.	.	.	.	.	.	.	.	.	C
9481B	+	.	A	.	.	.	.	−	.	.	.	.	.	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	8	.	17	14	.	.	.	.	.	.	.	.	.	C
Di-0	+	.	A	.	.	.	.	−	.	.	.	.	T	.	.	.	.	.	.	.	C	G	.	.	.	.	.	.	.	8	.	17	14	.	.	.	.	.	.	.	.	.	C
Ler-0	+	.	A	.	.	.	.	−	.	.	.	.	T	.	.	.	.	.	.	.	C	G	.	.	.	.	.	T	.	10	.	17	14	.	.	.	.	.	.	.	.	.	C
Ws	+	.	A	.	.	.	.	−	.	.	.	.	T	.	.	.	.	.	.	.	C	G	.	.	.	.	.	T	.	10	.	17	16	.	.	.	.	.	T	T	.	.	C
Es-0	+	.	A	.	.	.	.	−	.	.	.	.	T	.	.	.	.	.	.	.	C	G	.	.	.	.	.	T	.	9	.	17	15	.	.	.	.	.	T	T	.	.	C
Br-0	+	.	A	.	.	.	.	−	.	.	.	.	T	.	.	.	.	.	.	.	C	G	.	.	.	.	.	T	.	10	.	9	14	.	.	.	.	.	T	T	.	.	C
Co-1	+	.	A	.	.	.	T	−	A	.	.	.	.	A	.	T	.	.	T	.	.	G	.	.	.	.	.	T	.	19	.	22	14	.	A	−	.	.	.	.	.	.	C
Pla-0	+	.	A	.	.	.	T	−	A	.	.	.	.	A	.	T	.	.	T	.	.	G	.	.	.	.	.	T	.	16	.	21	14	.	A	−	.	.	.	.	.	.	C
An-2	+	.	A	.	.	.	T	−	A	.	.	.	.	A	.	T	.	.	T	.	.	G	.	.	.	.	.	T	.	14	.	19	13	.	A	−	.	.	.	.	T	.	C
Lip-0	+	T	A	.	.	.	.	−	.	.	A	A	.	A	.	.	.	C	T	.	.	G	+	.	.	.	.	.	.	14	.	15	18	.	.	.	.	.	.	.	.	.	C
Ita-0	+	T	A	.	.	.	.	−	.	.	A	.	.	A	.	.	.	.	T	.	.	G	+	.	.	.	.	.	.	9	.	18	14	.	.	.	.	T	.	.	.	G	C
Ko-2	+	.	A	T	T	A	.	−	.	G	.	.	.	.	G	.	.	.	T	T	.	G	.	.	.	.	.	.	.	13	C	14	15	.	.	.	.	.	.	.	.	.	C
Length (bp)	9																						5	3 5					2		1	1			2								

FIG. 4.—(*A*) Gene model and distribution of polymorphic sites for *SHP1*. Symbols are the same as described in figure 1. (*B*) Polymorphism table for *SHP1*. Arrangement and symbols are the same as described for figure 1, except that there are no nonsynonymous polymorphisms.

Heterogeneity Analyses Detect an Intragenic Sweep in *SEP1*

The levels of polymorphism at the *SEP1* and *SEP2* loci are similar to the levels observed for most *A. thaliana* nuclear genes and are comparable to other neutrally evolving genes. For neutrally evolving loci, the degree of intraspecific polymorphism is positively correlated with the level of interspecific divergence. This correlation can be tested using the HKA test of selection, which compares the levels of polymorphism to divergence of a test locus with those of an unlinked, neutrally evolving locus (Hudson, Kreitman, and Aguade 1987). The levels of polymorphism to divergence for these two loci were compared with six other *A. thaliana* nuclear genes; these latter six genes have levels of polymorphism that are consistent with neutral evolutionary expectations (see *Materials and Methods*). HKA tests against these six reference loci do not show any significant deviation from the expectations of the neutral theory for either *SEP1* or *SEP2* ($P < 0.22$ to 0.87 [table 3]). The HKA tests are also not significant even if we partition the genes into the promoter and the transcriptional unit (TU), which includes exons, introns, and 5′ and 3′ UTRs ($P < 0.38$ to 0.98; data not shown).

Nevertheless, a runs test (McDonald 1998) indicates that there exists significant heterogeneity in the frequency of polymorphic sites relative to fixed differences across the entire sequenced region of one of these two redundant genes, *SEP1* (table 4). This test analyses a sliding window of polymorphism levels relative to fixed differences between species across the gene of interest. In our study, we used the ortholog from *A. lyrata* for the interspecific divergence comparison. The neutral equilibrium model predicts a constant ratio of polymorphism to divergence across the gene. Valleys formed from a low frequency of polymorphic sites relative to divergent sites may be indicative of positive selection in the form of localized selective sweeps, whereas peaks may be associated with balanced polymorphisms. The runs test utilizes several different test statistics, including G_{mean} and G_{max}, which are most sensitive to detecting runs with one or two peaks/valleys in the polymorphism-to-divergence ratio. For *SEP1*, the G_{mean} and G_{max} statistics are highly significant ($P < 0.001$ for each statistic [table 4]), and may arise from a valley of reduced polymorphism to divergence between the 3′ end of intron 1 and exon 3, a region that is flanked by a peak at the 3′ end of the gene (fig. 6*A*). There are no amino acid differences in this region between orthologs; therefore, the putative target of selection is a regulatory *cis*-element. As the large first or second intron of MADS-box genes is necessary for proper transcriptional regulation (Sieburth and Meyerowitz 1997; Sheldon et al. 2002; Hong et al. 2003), this possibility is not unreasonable. In contrast, there is no significant heterogeneity in the levels of polymorphism and divergence across the *SEP2* locus (fig. 6*B* and table 4).

Selective Forces Act Differentially on the Promoter and Transcriptional Units of *SHP1* and *SHP2*

Unlike the *SEP1* and *SEP2* genes, the *SHP1* and *SHP2* loci show reduced levels of nucleotide variation

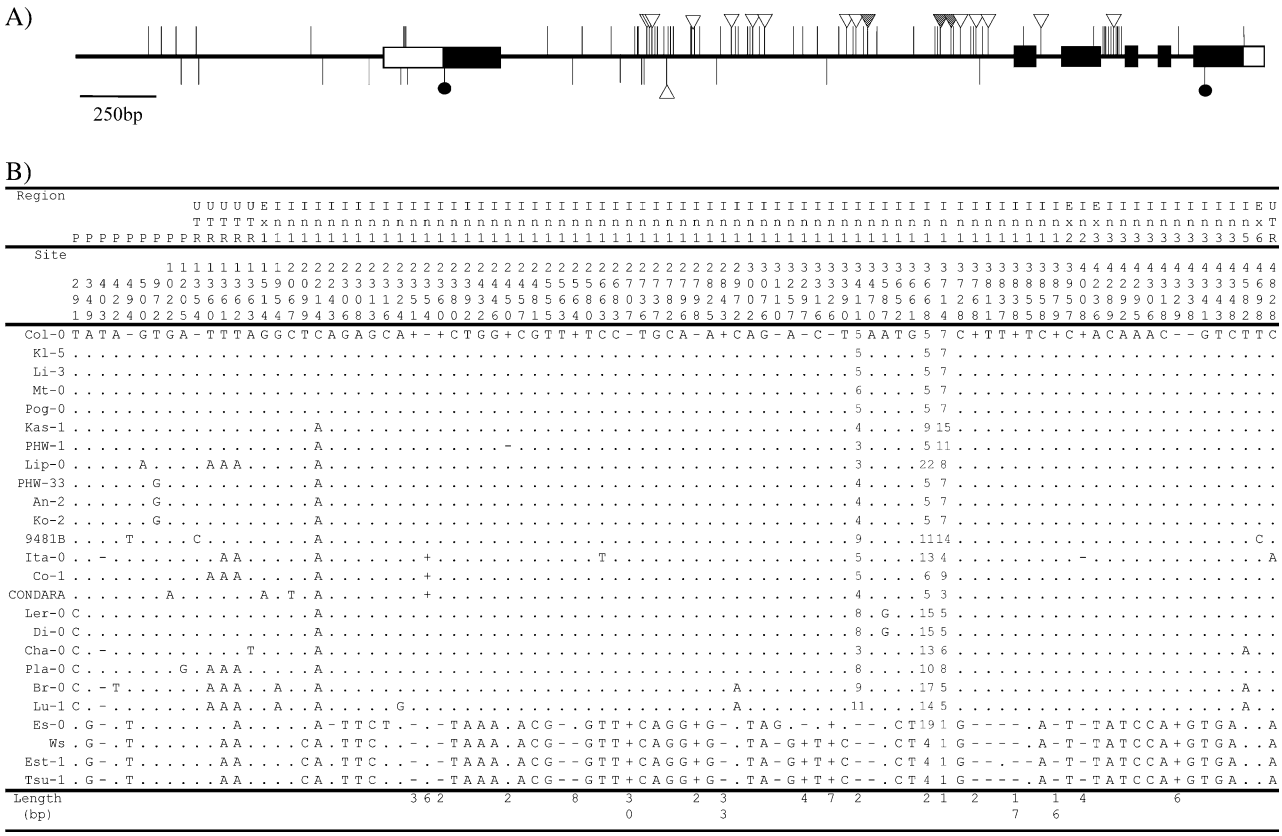


FIG. 5.—(*A*) Gene model and distribution of polymorphic sites for *SHP2*. Symbols are the same as described in figure 1. (*B*) Polymorphism table for *SHP2*. Arrangement and symbols are the same as described for figure 1.

relative to the average for *A. thaliana* genes. Reduced levels of intraspecific nucleotide variation can be indicative of a selective sweep if there is not a corresponding reduction in interspecific divergence. We, therefore, performed pairwise HKA tests for both *SHP1* and *SHP2* relative to neutrally evolving reference loci. We also conducted a runs test for each locus to examine heterogeneity in the ratio of polymorphism to divergence across these two genes.

For *SHP1*, pairwise HKA tests against all six neutral reference loci are significant ($P < 0.001$ to 0.035 [table 3]), indicating the *SHP1* locus has significantly reduced levels of intraspecific polymorphism relative to divergence. We combined the *P* values for these six nonindependent HKA tests using the Simes method for combined probability (Simes 1986). The Simes probability for *SHP1* ($P < 0.0004$) is significant at the Bonferroni-corrected level. For *SHP2*, the HKA test is significant against *AP3* ($P < 0.031$) and marginally significant for two other genes ($P < 0.070$ to 0.075 [table 3]). The Simes combined probability ($P < 0.025$) across all six tests is not significant at the Bonferroni-corrected level.

The runs test for *SHP1* resulted in a significant runs statistic, K_R ($P < 0.023$ [table 4]). This test statistic is most appropriate at detecting multiple peaks/valleys of heterogeneity in the polymorphism-to-divergence ratio. In

Table 3
Pairwise HKA Probability Values of Duplicate Genes Versus Six Neutral Reference Loci by Region

Gene	Region	Reference Locus						Simes *P*
		AP3	*F3H*	*PI*	*AP1*	*CAL*	*FAH*	
SEP1	Pr + TU	0.22	0.61	0.56	0.74	0.75	0.70	0.13
SEP1	Pr + TU	0.48	0.72	0.67	0.85	0.87	0.61	0.14
SHP1	Pr + TU	0.001^{****}	0.001^{****}	0.002^{***}	0.008^{***}	0.006^{***}	0.035^{**}	0.0004[a]
	Pr	0.10	0.19	0.16	0.21	0.22	0.56	0.044
	TU	0.001^{****}	0.001^{****}	0.001^{****}	0.004^{***}	0.003^{***}	0.012^{**}	0.0002[b]
SHP1	Pr + TU	0.031^{**}	0.075^{*}	0.070^{*}	0.13	0.13	0.44	0.025
	Pr	0.036^{**}	0.069^{*}	0.057^{*}	0.073^{*}	0.081^{*}	0.24	0.016
	TU	0.061^{*}	0.13	0.12	0.19	0.19	0.57	0.039

NOTE.—Pr = promoter; TU = transcriptional unit; Pr + TU = combined promoter and transcriptional unit. $^{*}P < 0.10$; $^{**}P < 0.05$; $^{***}P < 0.01$; $^{****}P \leq 0.001$.
[a] Significant with Bonferonni correction.

Table 4
Application of Heterogeneity Tests to Sequence Polymorphism and Divergence Data

	P			
Gene	G_{mean}	G_{max}	K_R	D_{KS}
SEP1	0.001^{***}	0.001^{***}	0.169	0.022^{*}
SEP2	0.671	0.564	0.340	0.515
SHP1	0.312	0.191	0.023^{*}	0.233
SHP2	0.174	0.089	0.006^{**}	0.038^{*}

NOTE.—*P* = proportion of the simulations, performed with the recombination parameter yielding the greatest *P*, that had values greater than or equal to the observed data. G_{mean} = mean sliding G statistic; G_{max} = maximum sliding G statistic; K_R = runs statistic; D_{KS} = Kolmogorov-Smirnov statistic. * $P < 0.05$; ** $P < 0.01$; $^{***}P \leq 0.001$.

particular, there are two valleys in the transcriptional unit spanning the region across exon 1, as well as a wide region spanning exons 2 to 7, characteristic of a regional selective sweep (fig. 6*C*). One possible target of selection in the transcriptional unit is a fixed, radical substitution of G to R within exon 1 of the *A. thaliana SHP1* alleles compared with the *A. lyrata* allele. This region adjoins the highly conserved DNA-binding MADS-box domain but is not functionally characterized. We cannot rule out the possibility that the target of selection is a regulatory element found within the first intron, as the valleys of significantly reduced polymorphism extend into the 5′ and 3′ regions of this relatively large and transcriptionally relevant intron.

The reduction in polymorphism across the TU of *SHP1* is reflected in HKA tests that partition the promoter and transcriptional unit. Pairwise HKA tests of the *SHP1* promoter region against the six reference neutral loci are not significant ($P < 0.10$ to 0.56 [table 3]). In contrast, HKA tests that compare the transcriptional unit against all six neutral reference genes are highly significant ($P < 0.001$ to 0.012 [table 3]). The Simes combined probability across all tests for the *SHP1* TU ($P < 0.0002$) is significant at the Bonferroni-corrected level. These results indicate that the transcriptional unit, but not the promoter, of *SHP1* has a significantly reduced level of polymorphism, which is consistent with a selective sweep.

For the *SHP2* runs test, both the K_R ($P < 0.006$) and D_{KS} ($P < 0.038$) statistics are significant (table 4), indicating multiple peaks/valleys of heterogeneity in the frequency of polymorphism relative to fixed sites across the sequenced region (fig. 6*D*). Unlike *SHP1*, there is relatively low polymorphism in the promoter region of *SHP2*. Pairwise HKA tests of the *SHP2* promoter against the six reference neutral loci are significant for *AP3* ($P < 0.036$) and marginally significant against four other reference genes ($P < 0.057$ to 0.081 [table 3]). The Simes combined probability across all six tests ($P < 0.016$) is marginally significant at the Bonferroni-corrected level. In contrast, HKA comparisons between the transcriptional unit and the reference loci are only marginally significant against *AP3* ($P < 0.061$) and are not significant against all other neutral loci (table 3). The Simes combined probability for the *SHP2* TU ($P < 0.039$) is not significant at the Bonferroni-corrected level. These results suggest that there is a marginally significant reduction in polymorphism levels at the promoter of *SHP2* that may be associated with positive selection. Because the putative sweep is found in the promoter region, the target of selection is presumably a regulatory *cis*-element.

It is possible that selection may be acting on flanking regions of these putative sweeps. However, contrasting patterns of evolutionary forces between promoter and

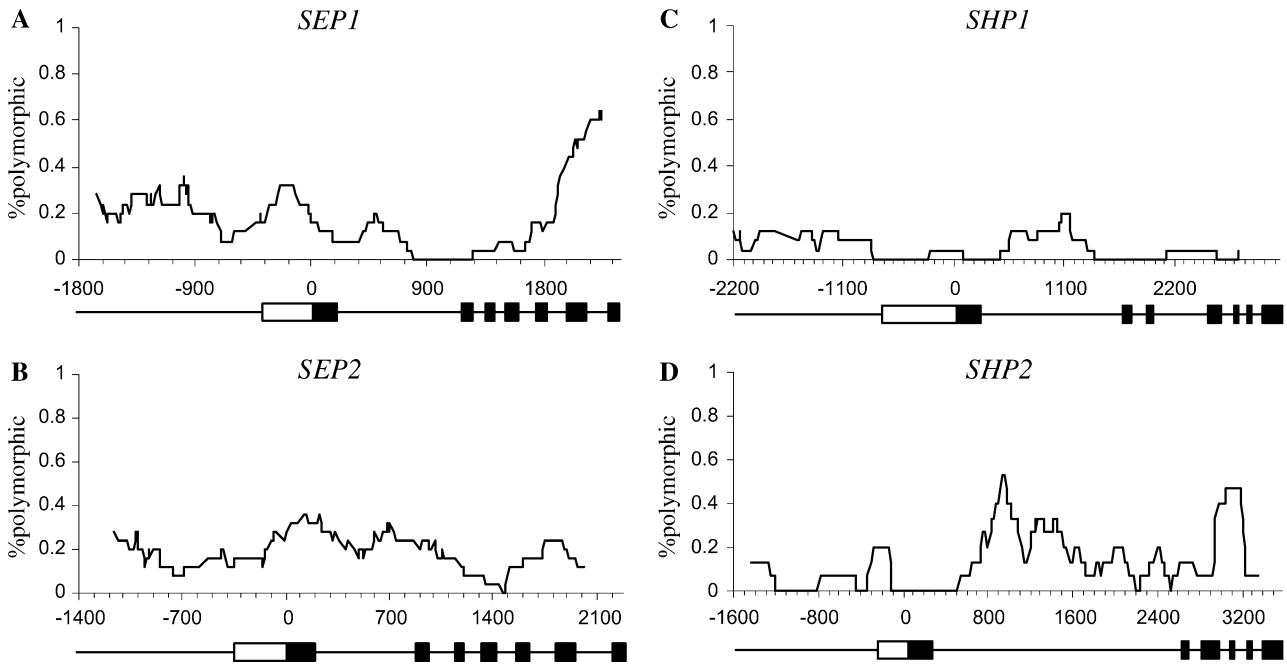


FIG. 6.—Sliding window plots showing the distribution of polymorphic sites relative to fixed differences for (*A*) *SEP1*, (*B*) *SEP2*, (*C*) *SHP1*, and (*D*) *SHP2*. The proportion of variable sites for a given window that are polymorphic (% polymorphic) is plotted at the nucleotide position midway between the endpoints of the window. The number of variable sites in each window is 50.

transcriptional unit are not unprecedented in *A. thaliana*, as has been documented for the *A. thaliana* inflorescence architecture gene, *TERMINAL FLOWER 1* (*TFL1* [Olsen et al. 2002]).

Discussion

Functional redundancy between duplicated genes is predicted to occur immediately after gene duplication and is lost as one gene of the duplicate pair becomes a pseudogene or gains a new function, or both genes accrue degenerative, yet compensatory, mutations (Clark 1994; Walsh 1995; Nowak et al. 1997; Wagner 1998; Lynch and Force 2000; Lynch et al. 2001; Walsh 2003). Despite its temporary nature, functionally redundant duplicate genes exist in abundance, presumably because functional redundancy guards against deleterious mutations and contributes to the genetic robustness of an organism (Pickett and Meeks-Wagner 1995; Wagner 1999; de Visser et al. 2003; Gu 2003). It is, therefore, important to understand what evolutionary forces contribute to the maintenance of functional redundancy. We addressed this question by analyzing the molecular population genetics of two pairs of functionally redundant paralogous genes from the model plant, *A. thaliana*: *SEP1* and *SEP2*, which are involved in floral organ identity (Pelaz et al. 2000, 2001), and *SHP1* and *SHP2*, which are involved in the regulation of seed shattering (Liljegren et al. 2000; Pinyopich et al. 2003).

Molecular clock analyses suggest that these two gene pairs originated at the same time, approximately 26 MYA. They are found within large duplicated segments of the *A. thaliana* genome and are most likely the product of the last whole-genome duplication that is speculated to have occurred between 24 and 48 MYA (Blanc, Hokamp, and Wolfe 2003; Bowers et al. 2003; Ermolaeva et al. 2003). Thus, these paralogs have existed in a state of functional redundancy for a considerable amount of time, considering transcriptional silencing or altered gene expression can occur rapidly after a polyploidization event (Adams et al. 2003). *SEP* and *SHP* paralogs escaped the fate of most duplicate gene pairs that resulted from the last whole-genome duplication event of *Arabidopsis*. Only 30% of *Arabidopsis* genes have duplicate copies that survived since that time (Bowers et al. 2003). However, retention of *SEP* and *SHP* paralogs may be caused in part by their functional role as transcription factors, as an overly disproportionate number of duplicate transcriptional regulators have been maintained since the last whole-genome duplication event (Blanc and Wolfe 2004).

It is unlikely that functional equivalence between *SEP* and *SHP* paralogs is being maintained by drift, given the length of time since their emergence. It is far more likely that one of the redundant loci will become a pseudogene (Bailey, Poulter, and Stockwell 1978; Li 1980; Walsh 1995) or that they will functionally diverge in this timeframe. Up to 20% to 30% of duplicate genes in some genomes exhibit diversifying selection in the amino acid sequences of their proteins (Conant and Wagner 2003), and the majority of duplicate genes in yeast and humans show divergent expression patterns (Gu et al. 2002; Makova and Li 2003). In *Arabidopsis*, the majority of duplicate genes from the last whole-genome duplication has altered expression (57%) or has evidence of functional diversification (62%) (Blanc and Wolfe 2004). The divergence in expression can occur extremely rapidly in humans; more than 70% of surveyed duplicates with a $Ks = 0.06$ have diverged in expression in at least one tissue (Makova and Li 2003). Moreover, even among redundant genes, partial redundancy is predictably more evolutionarily stable than the complete redundancy observed for *SEP* and *SHP* paralogs (Nowak et al. 1997; Wagner 1999). Functional redundancy is expected to be lost over time as the duplicates approach an equilibrium state of partial redundancy that is maintained by mutation-selection balance (Wagner 1999).

We suggest that selection has played a role in the maintenance of redundancy between these developmental duplicate genes in the form of purifying selection. Despite equivalent mutation rates, it is clear none of the genes we surveyed are in the process of becoming a pseudogene. In fact, the protein evolution of both pairs of duplicate genes is functionally constrained, as evidenced by K_a/K_s ratios more than sixfold lower than the neutral expectation of 1. Although the K_a/K_s ratio for both *SEP* and *SHP* paralogs is not much different from other *Arabidopsis* duplicates from the last genome duplication (Zhang, Vision, and Gaut 2002), we conclude from our data that purifying selection, in the form of functional constraint, is the dominant evolutionary force acting on *SEP* and *SHP* paralogs. Given that both duplicate gene pairs are involved in reproductive development, the selective maintenance of functional redundancy may help to reduce the fitness cost associated with developmental error (Cooke et al. 1997; Nowak et al. 1997; Krakauer and Nowak 1999). A survey of the rate of developmental error, and/or fitness cost, associated with only one gene of the redundant pair intact versus the presence of both redundant duplicate loci could address this possibility.

Purifying selection does not, however, act equally across these redundant MADS-box genes and appears to be acting more strongly on the MIK region, which contains the DNA-binding MADS-box domain and the K domain responsible for protein-protein interactions. For *SEP* paralogs, the average K_a/K_s ratio is twofold lower in the MIK region compared with the C-terminal region, and for *SHP* paralogs, there is a fivefold lower reduction of K_a/K_s in the MIK region (figure 1 and data not shown). Moreover, in some regions of the C-terminal domains of these paralogs, the K_a/K_s ratio rises above the neutral expectation of 1 (fig. 1), although diversifying selection on individual amino acids in these regions could not be detected. Between closely related, yet functionally divergent, MADS-box genes, the MADS and K domains are highly conserved, and it is the C-terminal domain that exhibits the greatest degree of amino acid diversification. The C-terminal domain is necessary for *trans*-activation of downstream genes and for the formation of ternary or quaternary protein complexes (Riechmann and Meyerowitz 1997; Egea-Cortines, Saedler, and Sommer 1999; Honma and Goto 2001) and is the site of functional specificity of MADS-box transcription factor loci (Lamb and Irish 2003). Previous studies of paralogous MADS-box genes,

such as *AP3* and *PI,* show similar diversification of the C-terminal domain (Kramer, Dorit, and Irish 1998; Lawton-Rauh, Buckler, and Purugganan 1999; Purugganan and Suddith 1999; Lamb and Irish 2003); however, in contrast to *SEP* and *SHP* paralogs, these differences also result in functional diversification. This is not unexpected, as the duplication events leading to these paralogs predate the diversification of the angiosperms (~120 MYA [Kramer, Dorit, and Irish 1998]) and are much older than either the *SEP* or *SHP* duplicate loci. Perhaps there is a point at which functional divergence provides a greater selective advantage over functional redundancy, but the *SEP* and *SHP* paralogs do not appear to have reached this stage. Instead, redundancy may actually prime these duplicate genes for functional divergence by ensuring their survival in the genome.

Our analyses also provide further insights into the molecular evolution of the floral developmental pathway (Weigel 1995; Yanofsky 1995; Jack 2001; Theissen 2001). A study of the evolution of this developmental pathway revealed that the signatures of positive and/or balancing selection are observed for three early-acting loci: one inflorescence architecture gene (*TFL1*) and two floral meristem identity (*LFY* and *AP1*) genes (Olsen et al. 2002). In contrast, the two later-acting floral organ-identity genes (*AP3* and *PI*) appear to be evolving according to the neutral-equilibrium model.

Consistent with this study, our results indicate that two other organ-identity loci (*SEP1* and *SEP2*) also appear to be evolving neutrally, despite some evidence from the runs test that there may be a limited intragenic selective sweep within *SEP1*. These floral organ-identity genes are expected to have strongly constrained developmental functions because whorl organ identities are highly conserved within the Brassicaceae family.

In contrast, positive selection, in the form of selective sweeps, is observed for the later-acting *SHP1* and *SHP2* developmental genes. Both *SHP1* and *SHP2* exhibit significantly reduced levels of intraspecific polymorphism relative to interspecific divergence, a pattern consistent with the action of recent selective sweeps. These latter two are not organ-identity genes, but function in the development of the valve margin of the carpel and developing siliques (Liljegren et al. 2000). It is possible that selection on these genes may be associated with selection for fruit development and seed dispersal, for which there is variation even between *Arabidopsis* species. Interestingly, selective forces may act differently between the *SHP1* and *SHP2* genes within the evolutionary history of *A. thaliana*. The selective sweep occurs in the transcriptional unit of *SHP1* but in the promoter of *SHP2*. Perhaps different, but complementary, functional aspects are being selected upon between *SHP1* and *SHP2* specific to the *A. thaliana* lineage. Only by comparing the functional efficacy of the *A. thaliana* and *A. lyrata* orthologs and their transcriptional regulation can we address this hypothesis. Thus, our analysis of the *SEP1/SEP2* and *SHP1/SHP2* floral-regulatory gene pairs, along with the six previously studied floral and inflorescence developmental loci (Olsen et al. 2002), provide the initial steps in describing the patterns that characterize the diversification of developmental systems and the evolution of developmental regulatory processes.

Acknowledgments

The authors would like to thank Ana Caicedo, Ken Olsen, and Kentaro Shimizu for stimulating discussions and critical readings of the manuscript. This work was funded by an NIH postdoctoral fellowship to RCM and an NSF grant to MDP.

Literature Cited

Adams, K. L., R. Cronn, R. Percifield, and J. F. Wendel. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. USA **100**:4649–4654.

Adams, M. D., S. E. Celniker, R. A. Holt et al. 2000. The genome sequence of *Drosophila melanogaster*. Science **287**: 2185–2195.

Aguade, M. 2001. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the *FAH1* and *F3H* genes, in *Arabidopsis thaliana*. Mol. Biol. Evol. **18**:1–9.

Alvarez-Buylla, E. R., S. Pelaz, S. J. Liljegren, S. E. Gold, C. Burgeff, G. S. Ditta, L. Ribas de Pouplana, L. Martinez-Castilla, and M. F. Yanofsky. 2000. An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc. Natl. Acad. Sci. USA **97**:5328–5333.

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

Bailey, G. S., R. T. Poulter, and P. A. Stockwell. 1978. Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. Proc. Natl. Acad. Sci. USA **75**:5575–5579.

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

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

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

Bureau, T. E., and S. R. Wessler. 1992. *Tourist*: a large family of small inverted repeat elements frequently associated with maize genes. Plant Cell **4**:1283–1294.

———. 1994. Mobile inverted-repeat elements of the *Tourist* family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. USA **91**:1411–1415.

Caenorhabditis elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282**:2012–2018.

Casacuberta, E., J. M. Casacuberta, P. Puigdomenech, and A. Monfort. 1998. Presence of miniature inverted-repeat transposable elements (MITEs) in the genome of *Arabidopsis thaliana*: characterisation of the Emigrant family of elements. Plant J. **16**:79–85.

Casacuberta, J. M., and N. Santiago. 2003. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene **311**:1–11.

Clark, A. G. 1994. Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA **91**:2950–2954.

Conant, G. C., and A. Wagner. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. **13**:2052–2058.

Cooke, J., M. A. Nowak, M. Boerlijst, and J. Maynard-Smith. 1997. Evolutionary origins and maintenance of redundant gene expression during metazoan development. Trends Genet. **13**:360–364.

de Visser, J., J. Hermisson, G. P. Wagner et al. 2003. Perspective: evolution and detection of genetic robustness. Evolution **57**:1959–1972.

Egea-Cortines, M., H. Saedler, and H. Sommer. 1999. Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in *Antirrhinum majus*. EMBO J. **18**:5370–5379.

Ermolaeva, M. D., M. Wu, J. A. Eisen, and S. L. Salzberg. 2003. The age of the *Arabidopsis thaliana* genome duplication. Plant Mol. Biol. **51**:859–866.

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

Flanagan, C. A., and H. Ma. 1994. Spatially and temporally regulated expression of the MADS-box gene *AGL2* in wild-type and mutant *Arabidopsis* flowers. Plant Mol. Biol. **26**: 581–595.

Giaever, G., A. M. Chu, L. Ni et al. 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418**: 387–391.

Gu, X. 2003. Evolution of duplicate genes versus genetic robustness against null mutations. Trends Genet. **19**:354–356.

Gu, Z., D. Nicolae, H. H. Lu, and W. H. Li. 2002. Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **18**:609–613.

Gu, Z., L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W. H. Li. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature **421**:63–66.

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

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

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Jack, T. 2001. Relearning our ABCs: new twists on an old model. Trends Plant Sci. **6**:310–316.

Kamath, R. S., A. G. Fraser, Y. Dong et al. 2003. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature **421**:231–237.

Kempin, S. A., B. Savidge, and M. F. Yanofsky. 1995. Molecular basis of the cauliflower phenotype in *Arabidopsis*. Science **267**:522–525.

Kitami, T., and J. H. Nadeau. 2002. Biochemical networking contributes more to genetic buffering in human and mouse metabolic pathways than does gene duplication. Nat. Genet. **32**:191–194.

Koch, M. A., B. Haubold, and T. Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol. Biol. Evol. **17**:1483–1498.

Krakauer, D. C., and M. A. Nowak. 1999. Evolutionary preservation of redundant duplicated genes. Semin. Cell Dev. Biol. **10**:555–559.

Kramer, E. M., R. L. Dorit, and V. F. Irish. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics **149**: 765–783.

Lamb, R. S., and V. F. Irish. 2003. Functional divergence within the *APETALA3/PISTILLATA* floral homeotic gene lineages. Proc. Natl. Acad. Sci. USA **100**:6558–6563.

Lawton-Rauh, A. L., E. S. Buckler, and M. D. Purugganan. 1999. Patterns of molecular evolution among paralogous floral homeotic genes. Mol. Biol. Evol. **16**:1037–1045.

Li, W. H. 1980. Rate of gene silencing at duplicate loci: a theoretical study and interpretation of data from tetraploid fishes. Genetics **95**:237–258.

———. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Liljegren, S. J., G. S. Ditta, Y. Eshed, B. Savidge, J. L. Bowman, and M. F. Yanofsky. 2000. *SHATTERPROOF* MADS-box genes control seed dispersal in *Arabidopsis*. Nature **404**: 766–770.

Litt, A., and V. F. Irish. 2003. Duplication and diversification in the *APETALA1/FRUITFULL* floral homeotic gene lineage: implications for the evolution of floral development. Genetics **165**:821–833.

Lynch, M., and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–473.

Lynch, M., M. O'Hely, B. Walsh, and A. Force. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics **159**:1789–1804.

Ma, H., M. F. Yanofsky, and E. M. Meyerowitz. 1991. *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. Genes Dev. **5**: 484–495.

Makova, K. D., and W. H. Li. 2003. Divergence in the spatial pattern of gene expression between human duplicate genes. Genome Res. **13**:1638–1645.

McDonald, J. H. 1998. Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15**:377–384.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**: 652–654.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Nowak, M. A., M. C. Boerlijst, J. Cooke, and J. M. Smith. 1997. Evolution of genetic redundancy. Nature **388**:167–171.

Olsen, K. M., A. Womack, A. R. Garrett, J. I. Suddith, and M. D. Purugganan. 2002. Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. Genetics **160**:1641–1650.

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

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

Pelaz, S., R. Tapia-Lopez, E. R. Alvarez-Buylla, and M. F. Yanofsky. 2001. Conversion of leaves into petals in *Arabidopsis*. Curr Biol **11**:182–184.

Pickett, F. B., and D. R. Meeks-Wagner. 1995. Seeing double: appreciating genetic redundancy. Plant Cell **7**:1347–1356.

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

Purugganan, M. D., and J. I. Suddith. 1998. Molecular population genetics of the *Arabidopsis CAULIFLOWER* regulatory gene: nonneutral evolution and naturally occurring

variation in floral homeotic function. Proc. Natl. Acad. Sci. USA **95**:8130–8134.

———. 1999. Molecular population genetics of floral homeotic loci: departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics **151**:839–848.

Riechmann, J. L., and E. M. Meyerowitz. 1997. MADS domain proteins in plant development. Biol. Chem. **378**:1079–1101.

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

Sheldon, C. C., A. B. Conn, E. S. Dennis, and W. J. Peakcock. 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.

Sieburth, L. E., and E. M. Meyerowitz. 1997. Molecular dissection of the *AGAMOUS* control region shows that *cis* elements for spatial regulation are located intragenically. Plant Cell **9**:355–365.

Simes, R. J. 1986. An improved bonferroni procedure for multiple tests of significance. Biometrika **73**:751–754.

Simmer, F., C. Moorman, A. M. Van Der Linden, E. Kuijk, P. V. Van Den Berghe, R. Kamath, A. G. Fraser, J. Ahringer, and R. H. Plasterk. 2003. Genome-wide RNAi of *C. elegans* using the hypersensitive rrf-3 strain reveals novel gene functions. PLoS Biol. **1**:E12.

Swofford, D. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585–595.

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

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24**:4876–4882.

Vandenbussche, M., G. Theissen, Y. Van de Peer, and T. Gerats. 2003. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Res. **31**:4401–4409.

Venter, J. C., M. D. Adams, E. W. Myers et al. 2001. The sequence of the human genome. Science **291**:1304–1351.

Vision, T. J., D. G. Brown, and S. D. Tanksley. 2000. The origins of genomic duplications in *Arabidopsis*. Science **290**:2114–2117.

Wagner, A. 1998. The fate of duplicated genes: loss or new function? Bioessays **20**:785–788.

———. 1999. Redundant gene functions and natural selection. J. Evol. Biol. **12**:1–16.

Walsh, B. 2003. Population-genetic models of the fates of duplicate genes. Genetica **118**:279–294.

Walsh, J. B. 1995. How often do duplicated genes evolve new functions? Genetics **139**:421–428.

Watterson, G. A. 1975. Number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7**:256–276.

Weigel, D. 1995. The genetics of flower development: from floral induction to ovule morphogenesis. Annu. Rev. Genet. **29**:19–39.

Winzeler, E. A., D. D. Shoemaker, A. Astromoff et al. 1999. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science **285**:901–906.

Wolfe, K. H., and D. C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yanofsky, M. F. 1995. Floral meristems to floral organs—genes controlling early events in *Arabidopsis* flower development. Annu. Rev. Plant Phys. **46**:167–188.

Yoshida, K., T. Kamiya, A. Kawabe, and N. T. Miyashita. 2003. DNA polymorphism at the *ACAULIS5* locus of the wild plant *Arabidopsis thaliana*. Genes Genet. Syst. **78**:11–21.

Zhang, L., T. J. Vision, and B. S. Gaut. 2002. Patterns of nucleotide substitution among simultaneously duplicated gene pairs in *Arabidopsis thaliana*. Mol. Biol. Evol. **19**:1464–1473.

Kenneth Wolfe, Associate Editor

Accepted September 1, 2004