Copyright © 1999 by the Genetics Society of America

Molecular Population Genetics of Floral Homeotic Loci: Departures From the Equilibrium-Neutral Model at the *APETALA3* and *PISTILLATA* Genes of *Arabidopsis thaliana*

Michael D. Purugganan and Jane I. Suddith

Department of Genetics, North Carolina State University, Raleigh, North Carolina 27695

Manuscript received May 12, 1998
Accepted for publication October 30, 1998

ABSTRACT

Molecular variation in genes that regulate development provides insights into the evolutionary processes that shape the diversification of morphogenetic pathways. Intraspecific sequence variation at the *APETALA3* and *PISTILLATA* floral homeotic genes of *Arabidopsis thaliana* was analyzed to infer the extent and nature of diversity at these regulatory loci. Comparison of *AP3* and *PI* diversity with three previously studied genes revealed several features in the patterning of nucleotide polymorphisms common between Arabidopsis nuclear loci, including an excess of low-frequency nucleotide polymorphisms and significantly elevated levels of intraspecific replacement variation. This pattern suggests that *A. thaliana* has undergone recent, rapid population expansion and now exists in small, inbred subpopulations. The elevated intraspecific replacement levels may thus represent slightly deleterious polymorphisms that differentiate distinct ecotypes. The distribution of replacement and synonymous changes in AP3 and PI core and noncore functional domains also indicates differences in the patterns of molecular evolution between these interacting floral regulatory genes.

THE evolutionary genetic basis of morphological differences between species remains one of the central issues in evolutionary biology (Gould 1977; Raff 1996). It has become increasingly clear from developmental genetic studies that classes of regulatory genes control morphological differentiation, and recent approaches have led to suggestions that evolutionary diversification at these developmental loci may contribute to interspecies differences in structure (Palopoli and Patel 1996; Shubin *et al.* 1997; Purugganan 1998). Despite its central relevance to the study of morphological diversification, little is known about the molecular population genetics of developmental pathways and the genes that comprise them.

Levels of genetic diversity at regulatory loci may govern the rates of morphological divergence and limit the degree to which selection at these genes can shape evolutionary change (Palopoli and Patel 1996; Richter *et al.* 1997). The fate of mutations that arise in regulatory genes is governed, not only by selective pressures at specific loci, but also by the species' life history and population structure. The interplay between population history, breeding system, and selection determines the nature of molecular variation and delineates the mechanisms that lead to evolutionary diversification. Molecular population genetic analysis provides a context for examining how historical, demographic, and genetic processes are interwoven to shape evolutionary variation in genes (Avise 1994; Aquadro 1997).

We are studying the molecular population genetics of the floral developmental pathway in *Arabidopsis thaliana*, probing the nature of variation in regulatory genes that control flower differentiation in plants. Our focus has been on the floral homeotic loci that are members of the plant MADS-box regulatory gene family of sequence-specific DNA-binding transcriptional activators (Ma *et al.* 1991; Yanofsky 1995; Riechmann and Meyerowitz 1997). Two of these homeotic genes, *APETALA3* and *PISTILLATA*, are partially responsible for petal and stamen specification in developing Arabidopsis flowers, and mutations at these loci result in the transformation of petals to sepaloid organs and stamens to carpel-like structures (Jack *et al.* 1992; Goto and Meyerowitz 1994). *AP3* and *PI* are related by a gene duplication event that took place prior to the origin of angiosperms (Kramer *et al.* 1998), and their protein products appear to interact directly to control organ formation (McGonigle *et al.* 1996; Riechmann *et al.* 1996).

Recent molecular studies have delimited the domains within both APETALA3 and PISTILLATA that are essential for the regulatory functions of these proteins (Riechmann and Meyerowitz 1997). Deletion and domain-swapping analyses have defined the core region of these proteins as consisting of the MADS-box, the I-region, and approximately the first 16 amino acids of the K-region (Krizek and Meyerowitz 1996a,b; Riechmann *et al.* 1996; see Figure 1). This core sequence is necessary for DNA-binding and dimerization activity of

Corresponding author: Michael Purugganan, Department of Genetics, Box 7614, North Carolina State University, Raleigh, NC 27695. E-mail: michael_purugganan@ncsu.edu

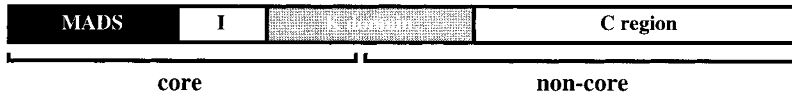

Figure 1.—Schematic of structural and functional domains in AP3 and PI proteins.

both AP3 and PI proteins and also appears to provide functional specificity. The other half of the protein, the noncore region, includes the rest of the K-box and the C-terminal sequence and may be involved in strengthening specific dimerization activities and in direct contact with the transcriptional machinery (Riechmann and Meyerowitz 1997).

Mutant analysis indicates that both *AP3* and *PI* perform overlapping, nonredundant functions in stamen and petal development, and the comparable developmental roles played by these regulatory loci in floral morphogenesis suggest they may both evolve with similar dynamics. In this article, we determine the levels and distribution of nucleotide polymorphisms in *AP3* and *PI* among different ecotypes of *A. thaliana.* Our results, coupled with those from other unlinked nuclear loci, provide insights into the history and population structure of this wild weed and allow us to explore the evolution of these two floral homeotic genes.

MATERIALS AND METHODS

Isolation and sequencing of alleles: The *A. thaliana* ecotypes were obtained from single-seed-propagated material provided by the Arabidopsis Biological Resource Center (see Table 1). The Kent, Bretagne, Lisse, and Corsacalla seed stocks were from the population collection of P. H. Williams maintained at the Arabidopsis Biological Resource Center. *A. lyrata* seed was provided by C. H. Langley.

Miniprep DNA was isolated from young leaves as previously described (Ausubel 1992). PCR was performed with 40 cycles of 1 min at 95°, 1 min at 52°, and 3 min at 72° followed by 15 min at 72°. The error-correcting recombinant Tth polymerase XL formulation (Perkin Elmer, Norwalk, CT) was used to minimize nucleotide misincorporation. The error rate for this polymerase formulation, on the basis of multiple amplification and resequencing of known genes, is less than 1 in 7000 bp (M. D. Purugganan and J. I. Suddith, unpublished results). We estimate that the nonsampling variance of nucleotide diversity due to PCR misincorporation, $Var_{PCR}(\pi)$, is negligible [$Var_{PCR}(\pi)/Var(\pi) \sim 0.14$] and does not significantly affect the frequency distribution of polymorphisms. The *AP3*-specific primers AP3F (for exon 1 forward), 5′-GAATATGGCGAGAG GGAAGATCC-3′, and AP3R (for exon 7 reverse), 5′-GCC TTTAATTATTCAAGAAGATGG-3′, and the PI-specific primers PI-1F (for exon 1 forward), 5′-GAGAAAGATGGGTAGAG GAAG-3′, and PI-1R (for exon 6 reverse), ATCTCGATGAT CAATCGATGACC-3′, were used in PCR reactions to amplify alleles from *A. thaliana* and *A. lyrata.* The isolation of *A. lyrata* sequences will be reported elsewhere (A. L. Lawton Rauh, E. S. Buckler and M. D. Purugganan, unpublished results). Amplified DNA was cloned into pCR2.1 using the TA cloning kit (Invitrogen, San Diego). DNA sequencing for both genes was conducted with the ABI377 automated sequencer using a series of nine nested internal sense and antisense primers. All sequence polymorphisms were visually rechecked from chromatograms, with special attention to low-frequency polymorphisms (Hamblin and Aquadro 1997). The DNA se-

TABLE 1

Ecotypes/field strains

Ecotype/field strain	Locality	ABRC seed accession
Basel-1	Basel, Switzerland	CS996
Blanes/Gerona-1	Blanes/Gerona, Spain	CS970
Bretagne-2[a]	Bretagne, France	CS6096
Bretagne-3[a]	Bretagne, France	CS6098
Burghaun/Rhon-0	Burghaun/Rhon, Germany	CS1006
Burghaun/Rhon-2	Burghaun/Rhon, Germany	CS1008
Cape Verde-1	Cape Verde Islands	CS902
Chisdra-1	Chisdra, Russia	CS1074
Coimbra-1	Coimbra, Potrugal	CS1084
Columbia	Landsberg, Germany	CS20
Corsacalla-1[a]	Corsacalla, Italy	CS6042
Graz-3	Graz, Austria	CS1202
Jelinka-1	Vranov u Brno, Czechoslovakia	CS1248
Kashmir-1	Kashmir, India	CS903
Kent-2[a]	Kent, Great Britain	CS6054
Kent-3[a]	Kent, Great Britain	CS6059
Landsberg erecta	Landsberg, Germany	—
Limburg-3	Limburg, Germany	CS1316
Limburg-8	Limburg, Germany	CS1332
Lisse-1[a]	Lisse, Netherlands	CS6090
Lisse-2[a]	Lisse, Netherlands	CS6092

[a] Field strains from the Williams collection (ABRC). The name is not an official ecotype designation.

quences are available from GenBank (accession numbers AF115798 to AF115830).

Data analysis: Sequences used in this study were visually aligned. Phylogenetic analyses were conducted using PAUP 3.1 (maximum parsimony; Swofford 1992). The heuristic search algorithm was utilized using the tree bisection-reconnection procedure, with the *A. lyrata* orthologues as the outgroup. Node support is assessed with 500 bootstrap replicates of the data. The polymorphism data was analyzed using the SITES (Hey and Wakeley 1997) and DNASP (Rozas and Rozas 1997) programs. Levels of nucleotide diversity were estimated as mean pairwise differences (π) and number of segregating sites (Θ) (Nei 1987). Association between polymorphic nucleotide sites was tested using Fisher's exact test, corrected for multiple tests using the Bonferroni procedure (Sokal and Rohlf 1981). Identification of possible recombinants utilized the four-gamete test (Hudson and Kaplan 1985). The Tajima (Tajima 1989) and Fu and Li (Fu and Li 1993) tests for distribution of nucleotide polymorphisms were conducted without specifying an outgroup. Contingency tests for independence of mutational categories, referred to as the McDonald-Kreitman test (McDonald and Kreitman 1991), were conducted using Fisher's exact test to evaluate significance. The coding region variation was also partitioned into core and noncore functional domains (Riechmann *et al.* 1996) for separate contingency analyses (Templeton 1996).

RESULTS

Nucleotide variation at the Arabidopsis *APETALA3* and *PISTILLATA* genes: A total of 19 *APETALA3* and 16 *PISTILLATA* alleles was isolated from a collection of 21 distinct, mostly European, *A. thaliana* ecotypes. Around 1.68 kb was sequenced for each *APETALA3* allele, spanning exons one to seven and including 4 bp of the 5′-untranslated region (UTR) and 10 bp of the 3′ flanking region of the gene. This sequence encompasses the entire *AP3* coding region. Approximately 2.05 kb of sequence was obtained from *PISTILLATA* alleles. The *PI* sequences include the entire coding region (from exons 1 to 6 and intervening introns), 7 bp of the 5′-UTR, and 13 bp of the 3′ flanking region of the gene. The *APETALA3* and *PISTILLATA* genes encode proteins of 232 and 208 amino acids in length, respectively.

Both the *APETALA3* and *PISTILLATA* genes in *A. thaliana* display considerable amounts of nucleotide variation (see Figures 2 and 3). For *AP3*, a total of 78 nucleotide polymorphisms are present in the sampled alleles. These include 20 replacement polymorphisms that result in amino acid variation between alleles, while only 8 polymorphisms within the coding region are synonymous. There are 7 conservative and 13 radical replacement polymorphisms at *AP3.* The only insertion/deletion variations are associated with a $(TG)_n$ microsatellite repeat in intron 5 of the gene, which differs in repeat length from 8 to 13 between alleles.

The *PI* alleles reveal a total of 67 nucleotide changes, of which 16, 12 replacement and 4 synonymous, polymorphisms are found within the coding region. There are 5 conservative and 7 radical replacement polymorphisms in this gene. Five distinct insertion/deletion polymorphisms were observed in intron sequences, including two 1-bp indels. The Lisse-1 allele contains a unique 22-bp insertion in the first intron. The other two indels are associated with an interrupted microsatellite $(TG)_3TCAG(TG)_n$, where $n = 6$ to 8.

The estimates of sequence diversity at these two regu-

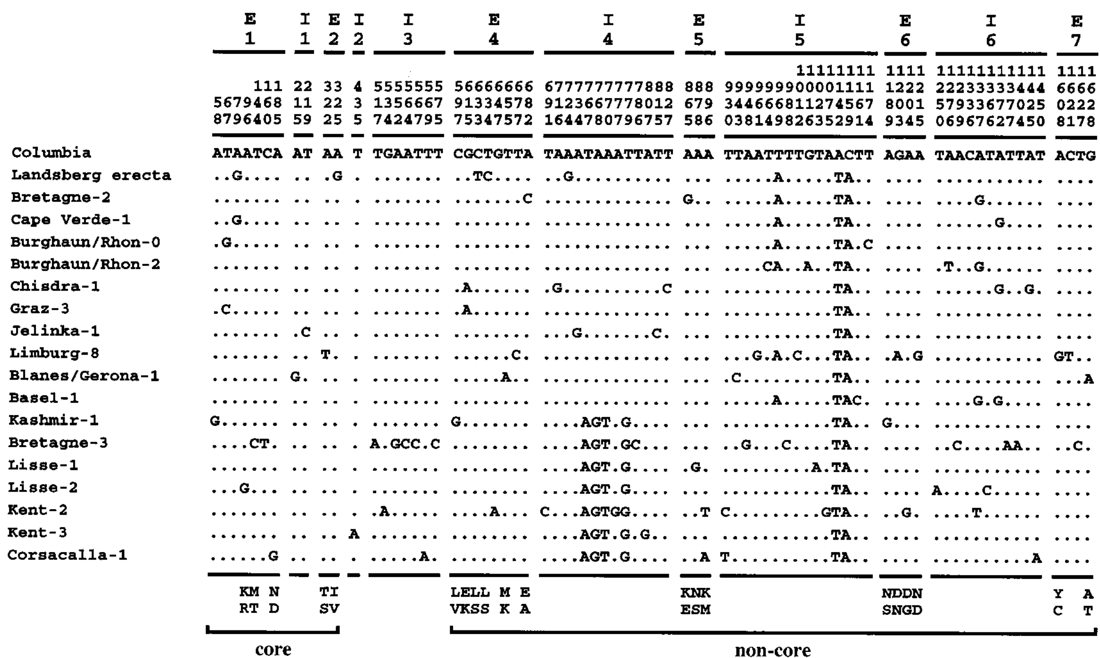

Figure 2.—Sequence of *AP3* alleles from different Arabidopsis ecotypes. All allele sequences are compared to the reference allele from the Columbia ecotype. The position of the polymorphic sites and their locations in introns and exons are indicated at the top. The different amino acids encoded by replacement polymorphisms are shown below, with the most predominant amino acid shown first. The core and noncore regions are indicated.

```
                     E                    I                    E     I      E   I E I E    I     E
                     1                    1                    2     2      3   3 4 4 5    5     6

                                                      11111111 11 11111111 1111 1 11 1 1 1111111 2
                       1122 2233344445555566777888899901111222 23 33334444 4445 6 66 7 7 8888889 0
                     592523 6658848891358937269012809922366288 93 56990022 7892 0 44 2 8 2445880 1
                     830850 3955930429625593408179122820902001 45 17191501 1443 2 15 0 7 4474080 9

Landsberg erecta     ATTAAT GCACAATTAATACAATTAGTCAGATGTGACATTC AA TAATACAA TATA T GA A C TAACGCG T
Cape Verde-1         .C.... .T............G..........A........ .. .....A.. .... . .. . . ...T.T. C
Graz-3               ...... .T.....C.G................A........ G. .....A.. ...G . .. . . ..GT.T. .
Limburg-3            ...... .TG......................A........ .. .....AG. .... . .. . . ...T.T. .
Chisdra-1            ...... .T......G......G..A......A..T..... .. C...TA.. .... . .. . . C..T.TA .
Lisse-2              ...... .T........C....G..A......A........ .. CG..TA.. C... . .. . . ...T.TA .
Corsacalla-1         ...... .T.A...........G..A......A........ .. C...TA.. .... . A. . . ...T.TA .
Kent-2               G..... AT.............G..A..T...A.A...C.. .. ...CTA.. .... . .G . A ...T.T. .
Kent-3               .....C .T...........G.GC.A..T...A........ .. ...CTA.C .... . .. . . ...T.T. .
Bretagne-3           ...... .T.............G..A..T...A......C. .. ..GCTA.. .... . .. . . ...T.T. .
Coimbra-1            ..C.T. .T...G......T..G..A..T...A........ .G ...CTA.. .... . .. . . ...T.T. .
Lisse-1              ...G.. .T................ACT....A...GT... .. ....TA.. .G.. . .. . . .G.T.TA .
Columbia             ...... .T.............G.TA......AA....... .. ....TA.. ..C. A .. . . ...TAT. .
Burghaun/Rhon-0      ...... .T.........T...G..A..T..CA........ .. ....TA.. .... . .. . . ...T.T. .
Burghaun/Rhon-2      ...... .T.........T...G..A..T..CA........ .. ....TA.. .T.. . .. . . ...T.T. .
Jelinka-1            ...... .T..T.C........G..A...CG.A.......G .. ....TA.. .... . .. G . ...T.TA .

                      VLNKS                                    N            D E    MI   L         M
                      APDMP                                    D            G G    IV   I         T

                                         core                                      non-core
```

Figure 3.—Sequence of *PI* alleles from different Arabidopsis ecotypes. All allele sequences are compared to the reference allele from the Landsberg erecta ecotype. The positions of the polymorphic sites and their locations in introns and exons are indicated at the top. The different amino acids encoded by replacement polymorphisms are shown below, with the most predominant amino acid shown first. The core and noncore regions are indicated.

latory loci are comparable, although *PI* shows slightly less variation. The overall estimates of species-wide nucleotide diversity, π, for *AP3* and *PI* are 0.0064 ± 0.0008 and 0.0053 ± 0.0004, respectively. The estimate of Θ for *AP3* is 0.01319, with an upper bound of 0.0179 and a lower bound of 0.01169 at 95% confidence. For *PI*, the estimate for Θ is 0.010 (upper bound is 0.0137, and lower bound is 0.00878).

The four-gamete test also indicates that at least one intragenic recombination has occurred in both *AP3* and *PI*. In *AP3*, the Basel-1 allele may have originated from recombination between positions 1337 and 1372 in intron 6. The pattern of variation at the *PI* Lisse-1 allele is also consistent with recombination between positions 673 and 1900 (see Figures 2 and 3).

Significant excess of low-frequency polymorphisms: The distribution of polymorphic sites in both *AP3* and *PI* is significantly skewed toward rare alleles. In the *AP3* gene, 69 of the 78 nucleotide polymorphisms are found only once in the sample (singletons). The *PI* gene also shows a preponderance of low-frequency alleles—58 of 67 nucleotide polymorphisms at this locus are also singletons.

The skewness in the frequency distribution of polymorphisms is significant in both the Tajima (1989) and Fu and Li (1993) tests. The Tajima test statistic D is -2.1507 for *AP3* ($P < 0.05$); the negative value of the D statistic indicates that sampled alleles have an excess of rare alleles over that expected in an equilibrium population. The Fu and Li test statistic D^* is also significantly negative for this gene ($D^* = -3.3728$, $P < 0.02$). The same pattern is observed for the *PI* locus. For this gene, both the Tajima ($D = -2.0183$, $P < 0.05$) and Fu and Li tests ($D^* = -2.8755$, $P < 0.02$) reveal that the distribution of polymorphisms for this regulatory gene is also significantly biased toward rare alleles.

Intraspecific gene genealogies of the *AP3* and *PI* loci: The genealogies of the naturally occurring alleles for *APETALA3* and *PISTILLATA* are shown in Figure 4. Two classes of *APETALA3* alleles are present in *A. thaliana*, suggesting that this gene exhibits allelic dimorphism in this species (see Figures 2 and 4). The class A alleles, which are found in 12 of 19 sampled ecotypes, form a monophyletic group with 96% bootstrap support in the maximum parsimony tree. The rest of the alleles (referred to as the *AP3* class B alleles) form an unresolved basal group in the phylogeny. The two allele classes are distinguished by four closely linked nucleotide polymorphisms in intron 4 (T to A at position 767, A to G at position 768, A to T at position 770, and T to G at position 779; see Figure 2). The mean pairwise differences (π) within the A and B allele classes are 0.00616 and 0.00448 differences/bp, respectively. The average pairwise difference between the two allelic classes is 0.00797 differences/bp.

There is only weak support for allelic dimorphism at *PI*. Instead, *PI* alleles appear to be structured into four allele groups, which are differentiated from each other by only one to two shared nucleotide polymorphisms (see Figure 4). Both the *AP3* and *PI* gene genealogies do not reveal any clear relationship of alleles with locality, and the differentiation into two allelic classes is not strongly correlated with geography. Moreover, the genealogies of the *AP3* and *PI* alleles are not concordant with one another.

Nonrandom associations between polymorphic sites: The four polymorphisms in *APETALA3* that differentiate the A and B allele classes are in complete linkage disequilibrium with each other. Four segregating sites in a 13-bp region of intron 4 (positions 767 to 779) exhibit significant levels of nonrandom association with each other ($P < 0.001$, corrected for multiple tests; see Figure 5). In the *PI* locus, the strongest nonrandom association is between site 798 in intron 1 and position 1401 in intron 2 ($P < 0.001$, corrected for multiple tests). The polymorphisms at these two sites are also

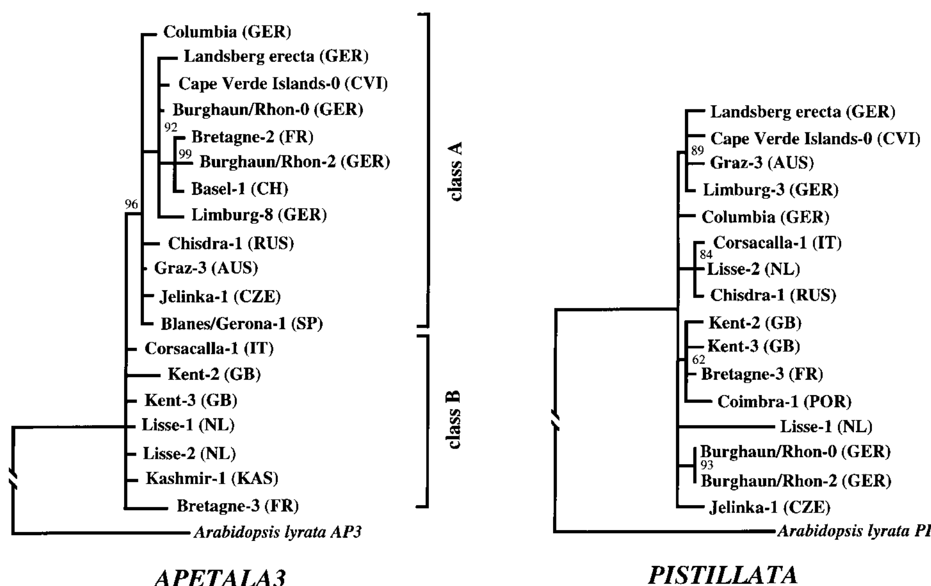

Figure 4.—Gene genealogies of *AP3* and *PI* alleles. All nodes with < 50% bootstrap support are collapsed; the other bootstrap values are indicated next to relevant nodes. Class A and B alleles for *AP3* are indicated.

in complete linkage disequilibrium with each other. Because the products of these two loci physically interact during floral development, we also tested whether there is linkage disequilibrium among sites between the two regulatory genes. Analyses of the joint sequence of *AP3* and *PI*, however, do not reveal any significant intergenic association of polymorphic sites.

Excess intraspecific replacement polymorphisms at *AP3* and *PI*: The protein-coding regions of the *AP3* and *PI* genes are not evolving according to the predictions of the equilibrium neutral hypothesis (see Table 2). The *AP3* gene appears to contain an excess of intraspecific replacement polymorphisms, and the McDonald-Kreitman test rejects the prediction of the equilibrium neutral theory that intraspecific polymorphisms and interspecific divergences are correlated. The *AP3* gene has 20 replacement and only 8 synonymous polymorphisms, and comparison of *A. thaliana* and *A. lyrata* genes reveals only 6 replacement (4 conservative and 2 radical) and 14 synonymous differences. The excess of intraspecific replacement polymorphisms is significant for this locus ($P = 0.005$).

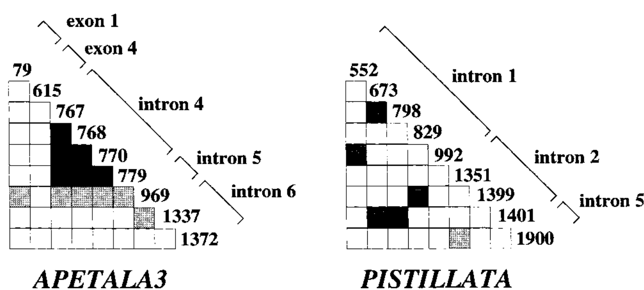

Figure 5.—Linkage disequilibrium between polymorphic sites in *AP3* and *PI*. The matrix indicates the pairs of sites that are nonrandomly associated with each other. The association between sites marked by the black squares is highly significant ($P < 0.001$). The dark gray and light gray squares indicate associations that are significant to the $0.001 < P < 0.01$ and $0.01 < P < 0.05$ levels, respectively.

The *PI* gene also exhibits a significant excess of within-species replacement polymorphisms. There are 12 replacement and 4 synonymous polymorphisms at the *PI* locus of *A. thaliana*, which differs from the *A. lyrata* orthologue by 11 replacement (6 conservative and 5 radical) and 16 synonymous differences. The McDonald-Kreitman test also rejects the predictions of the neutral hypothesis for the *PI* gene ($P = 0.03$).

Distribution of sequence variation between functional domains of *AP3* and *PI*: The partitioning of nucleotide changes between core and noncore functional domains within the two genes (see Figure 1) can be examined by studying the number of within-species polymorphisms and between-species differences (see Table 2). The *AP3* gene has an excess of within-species replacement polymorphisms in both the core (5 replacement to 4 synonymous) and noncore (15 replacement to 4 synonymous) regions, with a threefold increase in replacement variation in the latter domain. The number of fixed replacement differences between *A. thaliana* and *A. lyrata AP3* genes, however, is lower than observed synonymous fixed differences. The core region has 6 replacement and 14 synonymous differences, while the noncore region has 5 replacement and 10 synonymous changes. Contingency tests reveal that the relative ratio of within-species replacement to synonymous differences in the core domain is not significantly different when compared to differences between species ($P = 0.238$). The noncore region, however, contains an excess of intraspecific replacement polymorphisms compared to levels of between-species differences ($P = 0.009$).

The *PI* gene also exhibits an excess in levels of intraspecific replacement to synonymous polymorphisms in both functional domains of the gene. Both core and

TABLE 2

Contingency analysis of mutational categories in the *AP3* and *PI* coding regions

	All[a]		Core[b]		Noncore[b]	
Coding regions	Polymorphic	Fixed	Polymorphic	Fixed	Polymorphic	Fixed
AP3						
R	20	6	5	1	15	5
S	8	14	4	4	4	10
P value[c]	0.005**	0.005**	0.238	0.238	0.009**	0.009**
PI						
R	12	11	6	1	6	10
S	4	16	2	8	2	8
P value[c]	0.030*	0.030*	0.014*	0.014*	0.252	0.252

Replacement (R) and synonymous (S) polymorphisms within *A. thaliana* and fixed differences between *A. thaliana* and *A. lyrata* are indicated for each domain. *, significant at the $P < 0.05$ level; ** significant at the $P < 0.01$ level.

[a] Analysis for the entire protein ("all") is equivalent to the McDonald-Kreitman test.

[b] Core and noncore domains are defined in Reichmann *et al.* (1996).

[c] Probability values for the contingency tables are calculated using Fisher's exact text.

noncore regions have 6 replacement and 2 synonymous changes. The partitioning of the fixed differences, however, differs between core and noncore domains. Comparison of the core region of *PI* between *A. thaliana* and *A. lyrata* reveals only 1 replacement difference and 8 synonymous changes. In contrast, the noncore region also has 8 synonymous differences, but 10 replacement differences are observed between *PI* genes of these two species. For the noncore region, levels of intraspecific polymorphisms are correlated with interspecific replacement changes ($P = 0.252$). The core region of PI, however, displays low levels of between-species replacement difference (1/9) and an excess of within-species replacement polymorphisms (6/8), and this difference is significant ($P = 0.014$).

DISCUSSION

Variation at floral regulatory gene loci: The *APETALA3* and *PISTILLATA* loci of *A. thaliana* play central roles in floral organ development (Jack *et al.* 1992; Goto and Meyerowitz 1994). Assessment of the levels of variation at *AP3* and *PI*, as well as previous work on the floral meristem identity gene *CAULIFLOWER* (Purugganan and Suddith 1998), indicates that these developmental regulatory genes possess appreciable levels of nucleotide variation. The overall estimates of intraspecific variation at these two organ identity genes ($\pi = 0.005$–0.0065) are comparable to the estimate for *CAULIFLOWER* ($\pi = 0.0070$). The levels of diversity of these three regulatory genes are also similar to nucleotide variation estimates for two structural loci (*ChiA* and *Adh*) in Arabidopsis (see Table 3).

Population subdivision and expansion of *A. thaliana*: The extent and patterning of variation in loci are governed by the interplay of selective and demographic processes that shape the diversity of genes and genetic pathways. Evolutionary forces that affect the entire genome leave similar patterns across different unlinked loci and provide information on population history, structure, and dynamics. The specific details of population structuring in this plant species are crucial to understanding the nature of the evolutionary forces that shape diversity at these floral regulatory loci.

A recent study investigated variation in three nuclear loci within and between Arabidopsis populations using restriction site variation (Bergelson *et al.* 1998). Comparison of this study with sequence data on *AP3* and *PI* as well as three previously studied, unlinked loci—the *CAL* (Purugganan and Suddith 1998), *ChiA* (Kamabe *et al.* 1997), and *Adh* (Hanfstingl *et al.* 1994; Innan *et al.* 1996) genes—permits us to evaluate common patterns of variation that presumably arise from their shared evolutionary histories within the *A. thaliana* genome. Table 3 summarizes three features characteristic of most Arabidopsis genes that have been thus far analyzed: (1) excess of low-frequency nucleotide polymorphisms, (2) excess of within-species replacement site variation, and (3) evidence for allelic dimorphism.

For both the *AP3* and *PI* floral regulatory genes, the Tajima and Fu and Li tests reveal significantly negative D estimates, indicating a bias toward rare polymorphisms at these loci. A significant excess of singleton changes is also observed for *ChiA* (Kamabe *et al.* 1997), while *CAL* shows negative D^* statistics that are significant at the 10% level ($P < 0.057$ in the Fu and Li test; Purugganan and Suddith 1998). Excess in low-frequency alleles can arise from hitchhiking effects that accompany selective sweeps, which reduce linked sequence variation at specific loci (Aquadro 1997). The regulatory genes *AP3* and *CAL* (Purugganan and Suddith 1998), however, as well as the structural loci

TABLE 3

Features of sequence variation at *A. thaliana* genes

Gene	Map position[a]	π	Dimorphic	Tajima's *D*	Fu and Li D^*	McDonald-Kreitman
AP3	3 (79.9)	0.0064	Yes	−2.1507 ($P < 0.05$)**	−3.3728 ($P < 0.02$)**	$P = 0.005$***
PI	5 (28.4)	0.0053	Yes[b]	−2.0183 ($P < 0.05$)**	−2.8755 ($P < 0.02$)**	$P = 0.03$**
CAL	1 (46.0)	0.0070	Yes	−1.6605 ($P < 0.075$)*	−2.371 ($P < 0.057$)*	$P = 0.036$**
ChiA	NA[c]	0.0104	Yes	−1.8700 ($P < 0.05$)**	NA[c]	$P = 0.097$*[d]
Adh	1 (112.2)	0.0080	Yes	−0.8600 (ns)	−1.0500 (ns)	NA[c,e]

*, significant at the $P < 0.1$ level; **, significant at the $P < 0.05$ level; ***, significant at the $P < 0.01$ level.
[a] Position given as chromosome number, map position in parentheses.
[b] *PI* is only weakly dimorphic.
[c] Not available.
[d] Calculated for eliminating anomalous Cvi-0 sequence.
[e] McDonald-Kreitman test not applied, but number of replacement polymorphisms is less than synonymous polymorphisms.

ChiA (Kamabe *et al.* 1997) and *Adh* (Innan *et al.* 1996), all possess intermediate frequency polymorphisms that differentiate allele classes; this variation would presumably have been eliminated by genetic hitchhiking during species-wide fixation of selected sites.

The excess in low-frequency polymorphisms may be explained by the background selection hypothesis (Charlesworth *et al.* 1993), which suggests that rare variants may be present in population samples as a result of linkage to deleterious mutations. This effect should be marked in regions of low recombination or in primarily selfing species such as Arabidopsis. The background selection hypothesis, however, does not explain the excess in within-species replacement polymorphisms common to most of the Arabidopsis nuclear loci that have been examined (see Table 3).

Population subdivision and/or recent population expansion are mechanisms that may explain, however, both the excess in low-frequency allelic variation for Arabidopsis nuclear loci and the excess of within-species replacement polymorphisms (Aquadro 1997). Despite occasional cross-pollination, low outcrossing rates in *A. thaliana* may have resulted in population subdivision across the species range as a result of limited gene flow between inbred subpopulations (Allard *et al.* 1968; Charlesworth *et al.* 1997; Nordborg 1997). Inbreeding plant populations generally exhibit low within-population heterozygosities, given the low effective population sizes of primarily selfing organisms (Allard *et al.* 1968). Reduced levels of within-population variation have been observed for allozyme (Abbott and Gomes 1989), microsatellite (Todokoro *et al.* 1996; Kuittinen *et al.* 1997), and mitochondrial and nuclear restriction site studies (Bergelson *et al.* 1998) in Arabidopsis populations. Inbreeding plants, however, also display elevated levels of between-population differentiation as alternate genotypes are fixed in different populations by both drift and selection (Allard *et al.* 1968). The excess of low-frequency polymorphisms observed in this species-wide survey of *AP3* and *PI* nucleotide variation in *A. thaliana* may result from overdispersed sampling of distinct, inbred populations.

Differentiated, inbred populations of low effective population size would also explain the widespread occurrence of excess intraspecific replacement polymorphisms in nuclear loci of *A. thaliana.* The McDonald-Kreitman tests indicate that both *AP3* and *PI,* as well as *CAL* (Purugganan and Suddith 1998) and *ChiA* (Kamabe *et al.* 1997), have significantly high levels of intraspecific replacement polymorphisms compared to levels of interspecific divergence. The nearly neutral theory suggests that the intensity of purifying selection is relaxed and slightly deleterious mutations may be fixed at a higher rate in small populations (Ohta 1992). If the ecotypes sampled represent distinct, inbred populations of small Ne, then this would suggest that high levels of observed species-wide intraspecific replacement polymorphisms result from slightly deleterious mutations that are confined to individual ecotype populations.

Recent, rapid population expansion (Cummings and Clegg 1998) could also account for the excess in low-frequency and replacement site polymorphisms observed in *A. thaliana.* It is likely that both population subdivision and expansion have played roles in structuring the molecular diversity within this selfing plant species. Because population subdivision is associated with selfing plant species, one may expect that some of the features observed in *A. thaliana* may be shared by other inbreeding plant groups. There are indications that at least some of the features of nucleotide variation found in Arabidopsis characterize nuclear gene diversity in other selfing plants as well. The *A. gemmifera Adh* gene, for example, also displays a significant excess of within-species low-frequency polymorphisms and intraspecific replacement variation (Miyashita *et al.* 1996). Population subdivision in this selfing species has also been invoked to explain the patterning of diversity observed at this locus. The *Adh* locus from the self-fertilizing *H. vulgare* also displays a significant excess of within-

species replacement polymorphisms and a greater (though not significant) frequency of singleton variation (Cummings and Clegg 1998).

Dimorphic variation in Arabidopsis nuclear genes: Four of the five Arabidopsis genes examined display varying degrees of allelic dimorphism (Innan *et al.* 1996; Kamabe *et al.* 1997; Purugganan and Suddith 1998). The observed dimorphism for most genes is based on only a few sites; in the *AP3* gene, for example, class A and B alleles are differentiated by only four linked polymorphisms in intron 4. The widespread occurrence of allelic dimorphism in Arabidopsis nuclear loci has been ascribed to either introgression from related species or the breakdown of barriers between two previously isolated populations (Hanfstingl *et al.* 1994; Kamabe *et al.* 1997).

It is unclear what mechanisms have contributed to the continued maintenance of polymorphisms that differentiate allelic classes in *A. thaliana* nuclear genes. Association analysis reveals that the four polymorphisms that differentiate the *AP3* allele classes are in complete linkage disequilibrium with each other. In *CAL*, the polymorphic sites that differentiate two allelic classes are spread out over a 260-bp region and include a distinctive replacement polymorphism that may account for phenotypic variation in floral homeotic function. Linkage among sites, coupled with the low outcrossing rates for *A. thaliana*, may have led to the persistence of these sites in linkage disequilibrium.

Selective forces may also operate to maintain these linked polymorphisms. The background selection model suggests that balancing selection in selfing plants such as *A. thaliana* would lead to linkage among neutral variants and selected sites (Nordborg *et al.* 1996). A similar pattern should also be observed under local selection in subdivided populations (Charlesworth *et al.* 1997). Given their location within introns, however, the selective advantages of the linked changes responsible for the dimorphism at *AP3* are not immediately apparent. Studies on evolution of intron sites suggest that some selective pressures related to the formation of pre-mRNA hairpin structures may be involved in gene regulation (Kirby *et al.* 1995; Parsch *et al.* 1997). RNA secondary structure analysis indicates that a hairpin associated with these *AP3* polymorphic sites may possess different melting temperatures in the two allele classes (M. D. Purugganan, unpublished observations). Moreover, molecular studies in the related MADS-box floral homeotic gene *AGAMOUS* reveal that regulatory sites within introns may modulate expression of this regulatory gene (Sieburth and Meyerowitz 1997). No comparable studies for *AP3* have been reported, although it is possible that *cis*-acting regulatory sequences may also be present in the introns of this floral homeotic gene.

Evolutionary dynamics of functional domains within *APETALA3* and *PISTILLATA*: The distribution of replacement variation between distinct functional domains of *APETALA3* and *PISTILLATA* suggests that different regions of these regulatory proteins experience contrasting evolutionary pressures. These differing evolutionary forces between domains are reflected in the distribution of replacement and synonymous changes between gene regions (Templeton 1996). On the basis of a comparison of the patterning of coding region variation among several genes, the expectation for *A. thaliana* is (1) an excess of within-species replacement polymorphisms as a result of an increased tolerance for slightly deleterious mutations in small, inbred populations, and (2) a lower ratio of interspecific replacement-to-synonymous differences arising from purifying selection that prevents fixation of polymorphisms across species.

Contingency analyses can incorporate functional information and partition the mutations between distinct gene domains (Riechmann and Meyerowitz 1997). We can examine whether the pattern of replacement to synonymous variation is similar between the core and noncore domains of *A. thaliana AP3* and *PI* genes (see Table 2). The patterns of replacement and synonymous differences between *A. thaliana* and *A. lyrata* indicate that both core and noncore domains of *APETALA3* are constrained against amino acid changes. In these two domains, levels of replacement differences are lower than synonymous ones, suggesting that purifying selection prevents the fixation of amino acid polymorphisms.

This trend is reversed, however, in the distribution of within-species polymorphisms. The noncore region of *AP3* has 15 replacement and only 5 synonymous polymorphisms, which suggests that this domain can tolerate a high level of possibly slightly deleterious amino acid changes. Contingency tests on levels of replacement and synonymous changes in the *AP3* noncore region reveal the expected (standard pattern) of significant excess in intraspecific replacement polymorphisms. The *AP3* core region, however, shows a lower level of replacement to synonymous polymorphisms (5 replacements to 4 synonymous); this ratio is not significantly different from observed interspecies differences. Given the low overall numbers of polymorphisms and differences in the *AP3* core region, the contingency test applied may have restricted power. Otherwise, this test does indicate that the core region of *AP3*, which performs central functions in this floral regulatory protein, is less tolerant of replacement mutations than the noncore domain.

This contrast in evolutionary dynamics between core and noncore regions is also exhibited by the *PISTILLATA* protein-coding region. Contingency testing suggests that in *PI* it is the core region that is behaving according to the expected standard pattern—high levels of intraspecific replacement polymorphisms and reduced levels of interspecific amino acid change. The noncore region, however, deviates from this expected pattern and indicates that a large number of replace-

ment substitutions have been fixed in the divergence between the *A. thaliana* and *A. lyrata PI* locus. The difference between the pattern of interspecific divergence in core and noncore regions is significant ($P = 0.042$ using Fisher's exact test).

The interspecific amino acid changes in the *PI* noncore region are centered primarily in the C-terminal region, which previous molecular evolutionary analyses have indicated is the most rapidly evolving structural domain of plant MADS-box proteins (Purugganan *et al.* 1995). This high level of amino acid change could arise from a relaxation in the functional constraint at *PISTILLATA* after the separation of these two Brassicaceae species or adaptive fixation of selectively favored amino acid changes. AP3 and PI proteins form a heterodimer necessary for nuclear localization and DNA-binding (Riechmann *et al.* 1996), and the differences in the evolutionary dynamic between domains may arise as a result of some degree of functional complementation between proteins. It should be noted that there are no dramatic differences in the floral morphologies between *A. thaliana* and *A. lyrata*, although the petals of the latter are about twice as large (3–4 mm in *A. thaliana vs.* 7–10 mm in *A. lyrata*). The growth trajectories of petals and stamens should also differ between the two taxa as the selfing syndrome evolved in *A. thaliana.*

The molecular population genetics of the floral homeotic genes *APETALA3* and *PISTILLATA* provide insights into both the population structure of *A. thaliana* and the dynamics of regulatory gene variation. Population structuring could account for the excess of both singleton variation and replacement polymorphisms in species-wide surveys, and the subdivision of *A. thaliana* populations determines, in part, the fate of mutational variation at these regulatory genes. The pattern of variation indicates that functional domains evolve differently from one another within these two floral regulatory genes, which suggests that members of the floral developmental pathway are subject to distinct evolutionary forces.

We would like to acknowledge Ed Buckler, Katy Simonsen, and Elizabeth Friar for a critical reading of this paper and helpful discussions and two anonymous reviewers for suggested improvements. We also gratefully acknowledge Elliot Meyerowitz for providing us with the original *AP3* and *PI* genomic sequences. This work was funded with a grant from the U.S. Department of Agriculture National Research Initiative Plant Genetic Mechanisms Grant 97-35301-4688 and an Alfred P. Sloan Foundation Young Investigator Award in Molecular Evolution to M.D.P.

LITERATURE CITED

Abbott, R. J., and M. F. Gomes, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana.* Heredity **62:** 411–418.

Allard, R. W., S. K. Jain and P. L. Workman, 1968 The genetics of inbreeding populations. Adv. Genet. **14:** 55–131.

Aquadro, C. F., 1997 Insights into the evolutionary process from patterns of DNA sequence variability. Curr. Opin. Genet. Dev. **7:** 835–840.

Ausubel, F., 1992 *Short Protocols in Molecular Biology.* John Wiley, New York.

Avise, J. C., 1994 *Molecular Markers, Natural History and Evolution.* Chapman-Hall, New York.

Bergelson, J., E. Stahl, S. Dudek and M. Kreitman, 1998 Genetic variation within and among populations of *Arabidopsis thaliana.* Genetics **148:** 1311–1323.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effect of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155–174.

Cummings, M. P., and M. T. Clegg, 1998 Nucleotide sequence diversity at the *Adh1* locus in wild barley: an evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA **95:** 5637–5642.

Fu, Y.-X., and W. H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Goto, K., and E. M. Meyerowitz, 1994 Function and regulation of the *Arabidopsis* floral homeotic gene *PISTILLATA.* Genes Dev. **13:** 1548–1560.

Gould, S. J., 1977 *Ontogeny and Phylogeny.* Belknap Press, Cambridge, MA.

Hamblin, M. T., and C. F. Aquadro, 1997 Contrasting patterns of nucleotide sequence variation at the *Gld* locus in different populations of *Drosophila melanogaster.* Genetics **145:** 1053–1062.

Hanfstingl, U., A. Berry, E. A. Kellogg, J. T. Costa and F. Ausubel, 1994 Haplotypic divergence coupled with lack of diversity at the *Arabidopsis thaliana alcohol dehydrogenase* locus—roles for both balancing and directional selection. Genetics **138:** 811–828.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **120:** 831–840.

Innan, H., F. Tajima, R. Terauchi and N. T. Miyashita, 1996 Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana.* Genetics **143:** 1761–1770.

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

Kamabe, A., H. Innan, R. Terauchi and N. T. Miyashita, 1997 Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana.* Mol. Biol. Evol. **14:** 1303–1315.

Kirby, D., S. V. Muse and W. Stephan, 1995 Maintenance of pre-mRMA secondary structure by epistatic selection. Proc. Natl. Acad. Sci. USA **92:** 9047–9051.

Kramer, E., R. Dorit and V. 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.

Krizek, B. A., and E. M. Meyerowitz, 1996a Mapping the protein regions responsible for the functional specificities of the *Arabidopsis* MADS domain organ identity proteins. Proc. Natl. Acad. Sci. USA **93:** 4063–4070.

Krizek, B. A., and E. M. Meyerowitz, 1996b The *Arabidopsis* homeotic genes *APETALA3* and *PISTILLATA* are sufficient to provide the class B organ identity function. Development **122:** 11–22.

Kuittinen, H., A. Matilla and O. Savolainen, 1997 Genetic variation at marker loci and in quantitative traits in natural populations of *Arabidopsis thaliana.* Heredity **79:** 144–152.

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.

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

McGonigle, B., K. Bouhidel and V. F. Irish, 1996 Nuclear localization of the *Arabidopsis APETALA3* and *PISTILLATA* homeotic gene products depends on their simultaneous expression. Genes Dev. **10:** 1812–1821.

Miyashita, N. T., H. Innan and R. Terauchi, 1996 Intra- and interspecific variation at the *Adh* locus region in wild plants *Arabi-*

dopsis gemmifera and *Arabidopsis thaliana.* Mol. Biol. Evol. **13:** 433–436.
Nei, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.
Nordborg, M., 1997 Structured coalescent processes on different timescales. Genetics **146:** 1501–1514.
Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. London Ser. B **263:** 1033–1039.
Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.
Palopoli, M. F., and N. Patel, 1996 Neo-Darwinian developmental evolution—can we bridge the gap between pattern and process? Curr. Opin. Genet. Dev. **6:** 502–508.
Parsch, J., S. Tanda and W. Stephan, 1997 Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster.* Proc. Natl. Acad. Sci. USA **94:** 928–933.
Purugganan, M. D., 1998 The molecular evolution of development. Bioessays **20:** 700–711.
Purugganan, M. D., and J. Suddith, 1998 Molecular population genetics of the *Arabidopsis CAULIFLOWER* regulatory gene: nonneutral evolution and wild variation in floral homeotic function. Proc. Natl. Acad. Sci. USA **95:** 8130–8134.
Purugganan, M. D., S. D. Rounsley, R. J. Schmidt and M. F. Yanofsky, 1995 Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics **140:** 345–356.
Raff, R. A., 1996 *The Shape of Life: Genes, Development and the Evolution of Animal Form.* University of Chicago Press, Chicago.
Richter, B., M. Long, R. C. Lewontin and E. Nitasaka, 1997 Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in *Drosophila.* Genetics **145:** 311–323.
Riechmann, J. L., and E. M. Meyerowitz, 1997 MADS domain proteins in plant development. Biol. Chem. **378:** 1079–1101.
Riechmann, J. L., M. Q. Wang and E. M. Meyerowitz, 1996 DNA-binding properties of *Arabidopsis* MADS domain homeotic proteins *APETALA1*, *APETALA3*, *PISTILLATA* and *AGAMOUS.* Nucleic Acids Res. **24:** 3134–3141.
Rozas, J., and R. Rozas, 1997 DnaSP version 2.0: a novel software package for extensive molecular population genetic analysis. Comput. Appl. Biosci. **13:** 307–311.
Shubin, N., C. Tabin and S. Carroll, 1997 Fossils, genes and the evolution of animal limbs. Nature **388:** 639–648.
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.
Sokal, R. R., and F. J. Rohlf, 1981 *Biometry.* W. H. Freeman, San Francisco.
Swofford, D., 1992 *Phylogenetic Analysis Using Parsimony 3.1.* Illinois Natural History Survey, Champaign, IL.
Tajima, F., 1989 Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Templeton, A. R., 1996 Contingency tests of neutrality using intra/interspecific gene trees—rejection of neutrality for the evolution of the mitochondrial *cytochrome oxidase II* gene in the hominoid primates. Genetics **144:** 1263–1270.
Todokoro, S., R. Terauchi and S. Kawano, 1996 Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. Jpn. J. Genet. **70:** 543–554.
Yanofsky, M. F., 1995 Floral meristems to floral organs—genes controlling early events in *Arabidopsis* flower development. Annu. Rev. Plant Physiol. Plant Mol. Biol. **46:** 167–188.

Communicating editor: A. G. Clark