Nucleotide Sequence Variation at Two Genes of the Phenylpropanoid Pathway, the *FAH1* and *F3H* Genes, in *Arabidopsis thaliana*

Montserrat Aguadé

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Spain

The *FAH1* and *F3H* genes encode ferulate-5-hydroxylase and flavanone-3-hydroxylase, which are enzymes in the pathways leading to the synthesis of sinapic acid esters and flavonoids, respectively. Nucleotide variation at these genes was surveyed by sequencing a sample of 20 worldwide *Arabidopsis thaliana* ecotypes and one *Arabidopsis lyrata* spp. *petraea* stock. In contrast with most previously studied genes, the percentage of singletons was rather low in both the *FAH1* and the *F3H* gene regions. There was, therefore, no footprint of a recent species expansion in the pattern of nucleotide variation in these regions. In both *FAH1* and *F3H*, nucleotide variation was structured into two major highly differentiated haplotypes. In both genes, there was a peak of silent polymorphism in the 5′ part of the coding region without a parallel increase in silent divergence. In *FAH1*, the peak was centered at the beginning of the second exon. In *F3H*, nucleotide diversity was highest at the beginning of the gene. The observed pattern of variation in both *FAH1* and *F3H*, although suggestive of balancing selection, was compatible with a neutral model with no recombination.

Introduction

Phenylpropanoid products play diverse roles in the responses of plants to different biotic and abiotic stimuli, and their synthesis can be triggered by such stresses. Flavonoids and other phenylpropanoids have long been thought to protect plants against UV radiation, because they accumulate primarily in the epidermal layers of leaves and stems (Day 1993) and strongly absorb light in the UV-B range (Chapple et al. 1994). In *Arabidopsis thaliana*, the concentration in leaves of products such as kaempferol conjugates and sinapoate esters increases in response to UV irradiation (Li et al. 1993; Lois 1994), which supports the proposed protective role of these compounds against UV-B damage. It has recently been shown that both flavonoids and sinapic acid derivatives are involved in UV protection, as mutations blocking the synthesis of these compounds confer high sensitivity to UV. Some of the *transparent testa* (for colorless seed coats), or *tt*, mutants of *A. thaliana*, which contain no detectable levels of leaf flavonoids, have been characterized, and the corresponding single-copy genes have been cloned (Feinbaum and Ausubel 1988; Shirley, Hanley, and Goodman 1992; Pelletier and Shirley 1996; Wisman et al. 1998). These genes encode different enzymes in the general phenylpropanoid pathway, such as chalcone synthase or CHS, chalcone isomerase or CHI, flavanone-3-hydroxylase or F3H, and dihydroflavonol-reductase or DFR, which can be considered upstream enzymes in the anthocyanin pathway. On the other hand, the *fah1* mutant has allowed establishment of the importance of sinapic acid esters in UV protection (Landry, Chapple, and Last 1995; Meyer et al. 1996). This mutation affects the gene encoding ferulate-5-hydroxylase or F5H, which catalyzes the irreversible hydroxylation in the branch pathway that diverts ferulic acid toward sinapic acid derivatives such as sinapoate malate and other sinapic esters.

DNA sequences contain valuable information about the evolutionary history of the particular region of the genome and/or species studied. According to the neutral theory of molecular evolution (Kimura 1983), the levels of polymorphism within species and divergence between species at different genomic regions should be positively correlated if populations are in mutation-drift equilibrium. The neutral theory also makes definite predictions about the frequency spectrum of variants segregating in natural populations (Watterson 1974). Deviations from equilibrium due to demographic events such as population bottlenecks or expansions will affect the level and pattern of polymorphism in all regions of the genome. On the other hand, directional or balancing selection in a particular region leaves a characteristic footprint on that region's pattern and level of neutral variation. Therefore, unlike demographic events, selection affects only specific parts of the genome. Comparison of the level and distribution of nucleotide variation within and between species has been successfully used, for example in Drosophila, to establish the role of natural selection, as opposed to genetic drift, in shaping molecular variation within species and thus in determining molecular evolution in particular regions of the genome (Hudson, Kreitman, and Aguadé 1987; Tajima 1989; McDonald and Kreitman 1991; Fu and Li 1993; McDonald 1996, 1998).

Variation at one of the genes of the phenylpropanoid pathway, the *CHI*, or chalcone isomerase, gene, was previously studied in different ecotypes of the self-fertilizing species *A. thaliana* and in its close relative, the outcrossing species *Arabidopsis lyrata* ssp. *petraea* (Kuittinen and Aguadé 2000). The *CHI* gene encodes the second enzyme in the pathway leading to the synthesis of flavonoids and anthocyanins. The frequency spectrum of nucleotide variants in the *CHI* gene region was skewed toward an excess of low-frequency polymorphisms, which was consistent with the suggested recent expansion of the species (Price, Palmer, and Al-Shehbaz 1994). Variation in this region did not suggest

Key words: polymorphism, divergence, *Arabidopsis thaliana, Arabidopsis lyrata.*

Address for correspondence and reprints: Montserrat Aguadé, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645, 08071 Barcelona, Spain. E-mail: aguade@bio.ub.es.

Mol. Biol. Evol. 18(1):1–9. 2001

Table 1
***Arabidopsis thaliana* Ecotypes Studied**

Name	Origin	Accession No.[a]
CAN-0	Canary Islands, Spain	N1064
CHA-O	Champex, Switzerland	N1068
COL-2	Landsberg, Poland	N907
COND	Condara, Khurmatov, Tadjikistan	N916
CVI-0	Cape Verdi	N902
GR-5	Graz, Austria	N1206
ITA-0	Ibel Tazekka, Morocco	N1244
KAS-1	Kashmir, India	N903
LA-0	Landsberg, Poland	N1298
ME-0	Mechtshausen, Germany	N1364
MH-0	Muehlen, Poland	N904
MR-0	Monte/Tosso, Italy	N1372
NC-1	Ville-en-Vermois, France	N1388
PER-1	Perm, Russia	N1444
RI-0	Richmond, British Columbia, Canada	N1492
RSCH-0	Rschew/Starize, Russia	N1490
RUB-1	Rubezhnoe, Ukraina	N927
TUL-0	Turk Lake, Fla.	N1570
WS-0	Vasljevici/Drijepr, Byelorussia	N1602
YO-0	Yosemite, Calif.	N1622

[a] At the Nottingham Stock Center.

the action of natural selection. Different parts of a metabolic pathway may differ in their responsiveness to selection. Also, the relative importance of sinapoate esters and flavonoids as UV-protectants is not well established. Nucleotide variation at one gene in each of these pathways (the *FAH1* and *F3H* genes) has been surveyed by sequencing these regions for a similar sample of 20 ecotypes of *A. thaliana* and for one individual of the closely related species *A. lyrata* ssp. *petraea.*

Materials and Methods

Plant Materials

Twenty *A. thaliana* ecotypes sampled worldwide were used in the present study (table 1). Sixteen of these ecotypes were a subsample of those studied for the *CHI* region (Kuittinen and Aguadé 2000). Seeds were obtained from the Arabidopsis Nottingham Stock Center. For the interspecific comparison, one *A. lyrata* ssp. *petraea* stock from Karhumäki (Russia) was used; this stock was collected and kindly provided by O. Savolainen and H. Kuittinen. For each accession or stock, one plant was grown in a greenhouse under natural light.

DNA Extraction, PCR Amplification, and Sequencing

Genomic DNA was extracted from leaves with a modification of the CTAB procedure (Rogers and Bendich 1985) or with the DNAeasy plant extraction kit (Qiagen). For each region, a pair of 20-nt-long oligonucleotides was designed based on a previously published sequence (accession numbers AF068574 and U33932, respectively): 5′-GAACTTTGCCTCCTGACAAC-3′ and 5′-TTCCACCCCTAATTGACACA-3′ for the *FAH1* region, and 5′-AGCTAGCCGGAGAGTCTAAG-3′ and 5′-ACACCGCGCCTAGCATAATT3-′ for the *F3H* region. These primers were used in PCR reactions to amplify the corresponding regions, which encompassed the coding and 3′ flanking regions of each gene: an ~2.2-kb fragment for the *FAH1* region and an ~1.3-kb fragment for the *F3H* region. PCR products were purified with Qiaquick columns (Qiagen), and both strands were subsequently sequenced using primers spaced on average 350 nt apart. The Dye Terminator chemistry (Perkin-Elmer) was used for sequencing, and products were separated with an ABI 377 Automated DNA Sequencer (Perkin-Elmer). The homologous regions of *A. lyrata* ssp. *petraea* were amplified using the PCR primers designed for *A. thaliana*: for *F3H*, the same region was amplified, while the region amplified for *FAH1* was slightly shorter. The new sequences reported in this paper are deposited in the EMBL sequence database library under accession numbers AJ295566–AJ295607.

Sequence Analysis

Sequences were assembled and aligned with the SeqEd program (Perkin-Elmer), which was also used to check all variable sites. The sequences were edited for further analyses using the program MacClade, version 3.0.6 (Maddison and Maddison 1992). The program DnaSP, version 3.14 (Rozas and Rozas 1999), was used for most intraspecific and some interspecific analyses. Neighbor-joining trees (Saitou and Nei 1987) were built with the TreeCon program (Van de Peer and de Wachter 1994) using genetic distances corrected for multiple hits (Jukes and Cantor 1969).

Nucleotide variation was estimated as nucleotide diversity (π; Nei 1987). The genetic distance between major haplotypes (see below) was estimated as the average number of pairwise differences between haplotypes (k) and as the average per site number of nucleotide differences between haplotypes (D_{xy}; Nei 1987). The four-gamete test (Hudson and Kaplan 1985) was used to infer the minimum number of recombination events in the history of each sample. The recombination parameter $C = 4Nc$ was estimated by the method of Hudson (1987), which is based on the variance in the number of pairwise differences. Linkage disequilibrium or gametic association was measured as the correlation coefficient R^2 (Hill and Robertson 1968) using only informative sites (for which the rarest variant is present more than once); the statistical significance of the associations was established by the χ^2 test. The Z_{nS} statistic (Kelly 1998) was also estimated, and its statistical significance was established by computer simulations based on the coalescent process without recombination (Hudson 1990) and with recombination.

Coalescent Simulations

Computer simulations were used to contrast whether the presence of two subsets of highly differentiated sequences in a sample of size n is consistent with the equilibrium neutral model (J. Rozas, personal communication). The test statistic used is based on the number of mutations fixed between the two subsets given the total number of segregating sites (S) in the sample. Here, these subsets refer to a partition of the genealogical tree

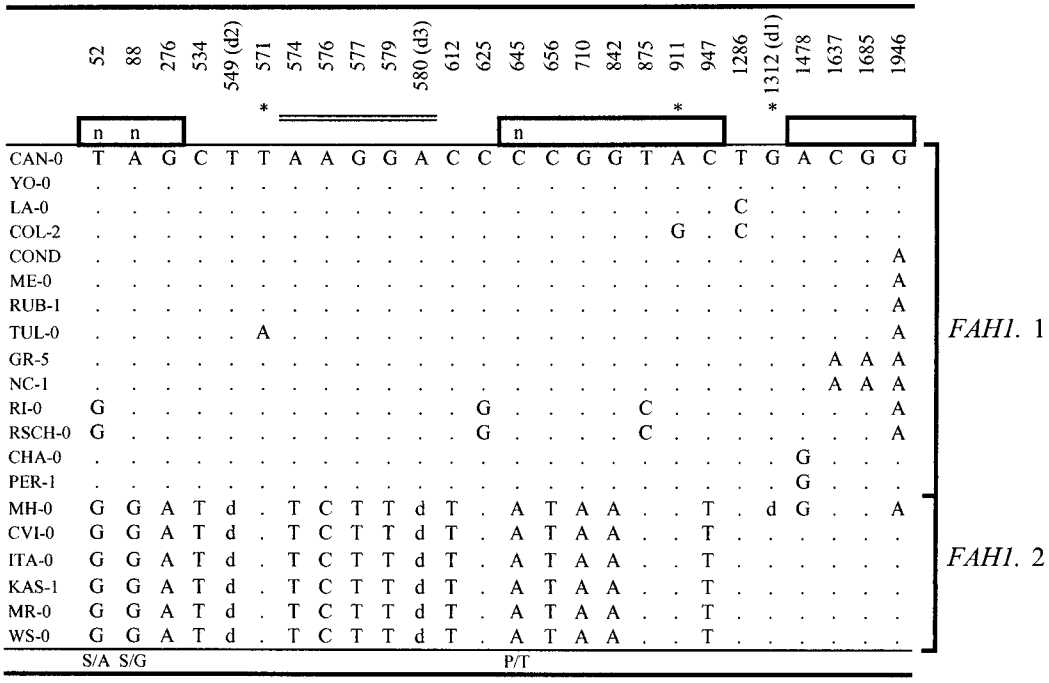

FIG. 1.—Nucleotide and length polymorphisms in the *FAH1* gene of *Arabidopsis thaliana.* Nucleotides are numbered from the initiation codon. A dot indicates the same nucleotide as in the first sequence, and a hyphen indicates the presence of the corresponding deletion. n = nonsynonymous change; d# = deletion of # bp; * = singleton. *FAH1*.1 and *FAH1*.2 refer to the two major haplotypes in the region. Open boxes indicate exons, and a double bar above an intronic region indicates a complex mutational event. For each nonsynonymous polymorphism, the one-letter symbols of the corresponding amino acids are given in the lower part of the column; the first symbol indicates the amino acid present in the Can-0 ecotype.

where the two lineages descending from the root have a and $n - a$ sequences (a, $n - a$ partition). Processes like balancing selection, population subdivision, or population decline should generate a larger number of fixed differences than those expected under neutrality. Population expansion should, on the other hand, generate a smaller number of fixed differences than expected under stationarity.

The empirical distribution of the number of fixed differences between two subsets of sequences was obtained by computer simulation (1,000 replicates) based on the coalescent process with no recombination (Kingman 1982*a*, 1982*b*; Hudson 1990). The genealogical samples were generated conditioned on n, S and a particular partition (a, $n - a$) of sequences. Random genealogies were generated using conventional procedures, but samples with partitions other than (a, $n - a$) were discarded.

Results

Nucleotide Polymorphism

Figure 1 summarizes the distribution of nucleotide sequence variation in the *FAH1* gene. A total of 23 nucleotide polymorphisms and 3 length polymorphisms were detected in the 2,199-bp region studied (2,193 bp after excluding alignment gaps). Four nucleotide polymorphisms and one length polymorphism (a 3-bp deletion) in the first intron can be considered part of a complex mutational event (fig. 1); to be conservative, only one of the four nucleotide changes involved was considered in further analyses. Another length polymorphism was a 2-bp indel in a TA microsatellite in the first intron, while a third one was a 1-bp indel in the second intron. Table 2 gives the estimates of nucleotide variation for the whole *FAH1* region sequenced and for its different functional parts.

Figure 2 summarizes the distribution of nucleotide sequence variation in the *F3H* gene. Thirty nucleotide polymorphisms and no length polymorphisms were detected in the 1,242-bp region studied. As previously detected in the *CHI* gene (Kuittinen and Aguadé 2000), in *F3H* there are also some variants present only in ecotypes from the same geographical area. Nucleotide variation estimates are given in table 2, both for the whole region and for its different functional parts.

In both genes, the proportion of singletons, or polymorphic sites where the rarest variant is present only once in the sample, is rather low (table 2). For *FAH1*, the frequency spectrum of polymorphic sites was skewed toward an excess of sites at which the two variants were present at intermediate frequencies. Consequently, for *FAH1*, Tajima's D statistic was positive (0.826), although not significantly different from 0. For *F3H*, the value of this statistic was positive but rather low (0.252) and again compatible with the neutral hypothesis. Similar results were obtained for the Fu and Li statistics D^* and F^* (results not shown).

Two recombination events were detected in the evolutionary history of *FAH1* by the four-gamete test (Hudson and Kaplan 1985). Both events occurred in the

Table 2
Polymorphism and Divergence in the *FAH1* and *F3H* Genes

	Coding Region				
	Synonymous	Nonsynonymous	Noncoding	All Silent Sites	Total
FAH1					
No. of sites	362.6	1,197.4	630	992.6	2,190
S	11 (1)	3 (0)	6 (1)	17 (2)	20 (2)
π	0.010	0.0012	0.003	0.005	0.003
K	0.129	0.008	0.096	0.111	0.048
F3H					
No. of sites	233.1	798.9	210	443.1	1,242
S	18 (2)	5 (1)	7 (5)	25 (7)	30 (8)
π	0.029	0.0015	0.005	0.0175	0.007
K	0.208	0.008	0.113	0.164	0.058

Note.—*S* = number of segregating sites; π = nucleotide diversity. *K* = nucleotide divergence. Singletons are given in parentheses.

rightmost part of the region studied (between polymorphic sites 947 and 1478 and between sites 1478 and 1946, respectively). Although this test detected only one recombination event in the history of the *F3H* sample (between sites 57 and 204), the partition of the ecotypes for the different variants in this region indicates that ecotypes ME-0, RSCH-0, CHA-0, and RUB-1 are recombinants. The recombination parameter, *C*, estimated from the sampled sequences was 3.2 for *FAH1* and 1.1 for *F3H*.

The proportion of pairwise comparisons with a significant association between variants, i.e., sites in linkage disequilibrium, was high for both *FAH1* and *F3H* (table 3). These percentages remained above 30% after Bonferroni correction for multiple comparisons. Values of the Z_{nS} statistic (Kelly 1998) were high: 0.383 and 0.501 for *FAH1* and *F3H*, respectively. The probabilities of obtaining values higher than those observed were low but were not significant assuming no recombination (0.154 and 0.053, respectively). When the *C* value estimated for each gene was used, the probabilities were 0.047 and 0.031, respectively. The overall level of significant disequilibrium in the two regions therefore probably cannot be explained by mutation-drift equilibrium.

Pattern of Nucleotide Variation

For *FAH1* and *F3H*, 10 and 12 different haplotypes were detected, respectively, in the sample of 20 ecotypes. In both regions, two highly differentiated sets of haplotypes could be identified (figs. 1 and 2). For *FAH1*,

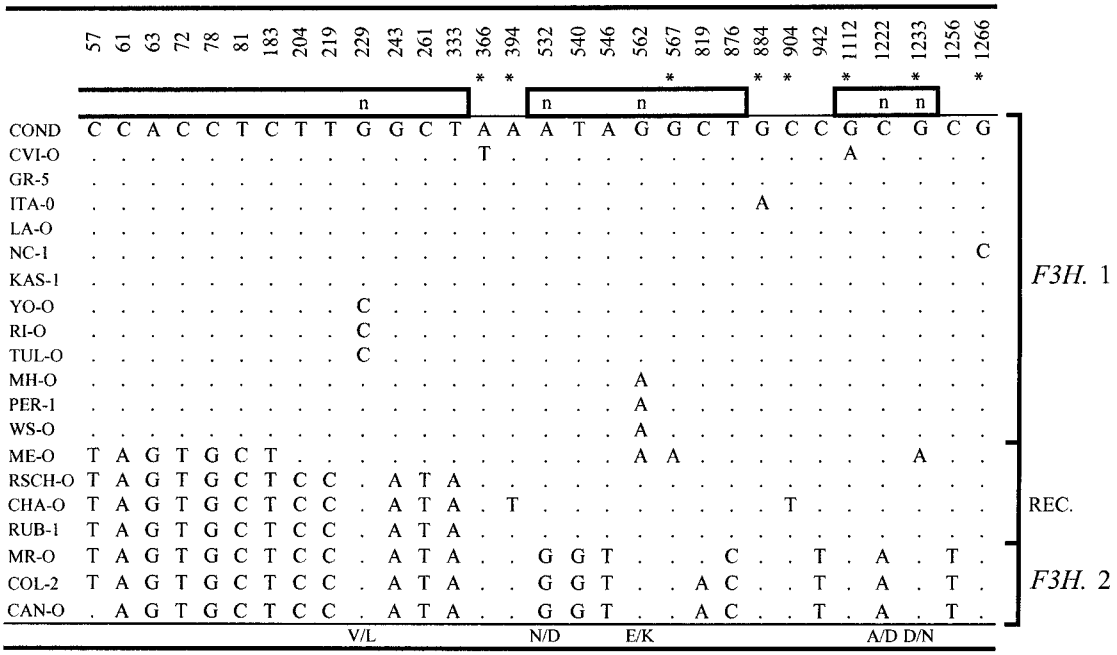

Fig. 2.—Nucleotide polymorphisms in the *F3H* gene of *Arabidopsis thaliana*. Nucleotides are numbered from the initiation codon. A dot indicates the same nucleotide as in the first sequence. n = nonsynonymous change; * = singleton; REC. = recombinant haplotypes. *F3H*.1 and *F3H*.2 refer to the two major haplotypes in the region. Open boxes indicate exons. For each nonsynonymous polymorphism, the one-letter symbols of the corresponding amino acids are given in the lower part of the column; the first symbol indicates the amino acid present in the Cond ecotype.

Table 3
Linkage Disequilibrium

	No. of Informative Sites	No. of Comparisons	No. of Significant Comparisons[a]	% Significant Comparisons[b]
FAH1	18	153	57 (47)	30.7
F3H	22	231	182 (82)	35.1

[a] Number of significant comparisons ($P < 0.05$) using the χ^2 test without and with (in parentheses) the Bonferroni correction for multiple comparisons ($\alpha = 0.05$).

[b] After applying the Bonferroni correction.

the two major haplotype classes (*FAH1*.1 and *FAH1*.2, with 14 and 6 ecotypes, respectively) differed at nine fixed nucleotide positions, one complex difference, and one indel. Two of the fixed differences were nonsynonymous differences. All fixed differences between haplotypes were located in the 5′ half of the region studied (between sites 1 and 1100 out of the 2,199 nt sequenced). The ecotypes with the less frequent haplotype came from both Europe and Asia. Most of the variation in this gene was due to the differences between these two haplotypes, and the level of variation within haplotypes was much lower than that between haplotypes. For the complete region studied, the average numbers of pairwise differences were 2.66 for haplotype *FAH1*.1, 0.67 for haplotype *FAH1*.2, and 12.52 between haplotypes (1.08, 0, and 11.3, respectively, when only the 5′ half of the region studied was considered), and the same was true for silent variation (fig. 3).

For the *F3H* gene, there were also two highly differentiated haplotypes (haplotypes *F3H*.1 and *F3H*.2), differing at 18 sites (two of them nonsynonymous), and the differences extended across the whole region studied. Unlike the *FAH1* result, one of the two major *F3H* haplotypes exhibited rather low frequency and was present in three ecotypes from both Europe (Italy and Germany) and the Canary Islands. In this gene, there were also recombinant sequence types between the two major haplotypes (fig. 2). As in the *FAH1* gene, most variation in the whole sample was due to differences between haplotypes: the average numbers of pairwise differences were 1.4 for haplotype *F3H*.1 (present in 13 ecotypes) and 1.3 for haplotype *F3H*.2 (present in three ecotypes), but 20.1 between haplotypes.

As a result of recombination, different parts of a given sequence have different evolutionary histories. Nevertheless, the presence of two highly differentiated haplotypes in both genes studied was clearly detectable in the corresponding neighbor-joining trees of the 20 ecotypes when the sequence of *A. lyrata* ssp. *petraea* was used as the outgroup (results not shown). For the *FAH1* region, where recombination was detected in the less variable region, both clusters of ecotypes were supported by high bootstrap values (86% and 100%). On the other hand, for the *F3H* region, only the clade corresponding to the *F3H*.1 haplotype showed a high bootstrap value (95%).

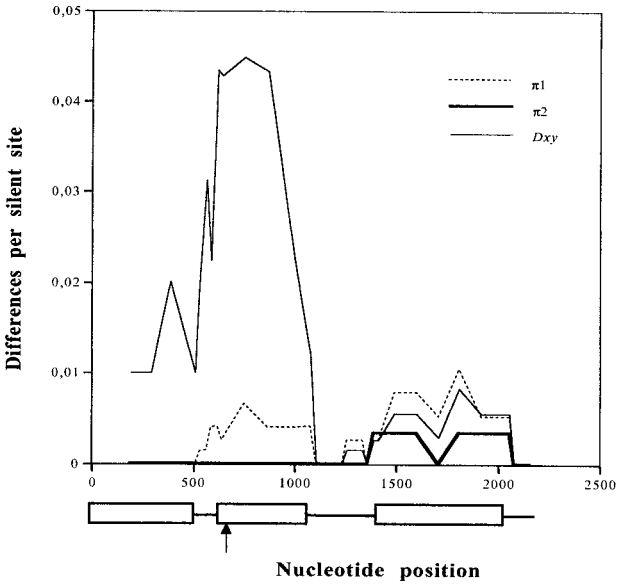

FIG. 3.—Sliding-window plot of silent nucleotide diversity within (π) and between (D_{xy}) the two major haplotypes of the *FAH1* gene region. Windows include 100 silent sites, with successive displacements of 25 sites. The structure of the *FAH1* gene is represented below the plot. The arrow below the second exon indicates the nonsynonymous polymorphism at site 645.

Amino Acid Replacement Polymorphism and Divergence

Three replacement polymorphisms, resulting in three amino acid haplotypes (fig. 1), were detected in the N-terminal half of the deduced F5H protein; none of these polymorphisms involved any charge change. At the corresponding residues, *A. lyrata* presented an alanine, a glycine, and a threonine, respectively, as in FAH1.2. Five amino acid polymorphisms were detected in the F3H protein (fig. 2); these polymorphisms resulted in five amino acid haplotypes, four of which differed in the total charge of the corresponding protein. At each of these residues, *A. lyrata* presented the most common amino acid in the *A. thaliana* ecotypes except at the second residue, where *A. lyrata* had an aspartic acid like ecotypes MR-0, Col-2, and Can-0.

Under neutrality, the ratio between nonsynonymous and synonymous changes should be the same within and between species. The McDonald and Kreitman (1991), or MK, test examines this prediction using a 2×2 contingency table where the observed differences are classified as synonymous or nonsynonymous and also as polymorphic within species or fixed between species. No significant deviation from neutrality was detected in either the *FAH1* or the *F3H* coding regions (results not shown).

Silent Nucleotide Polymorphism and Divergence

Silent divergence in the *FAH1* and *F3H* genes was estimated both for the different functional regions (table 2) and by sliding a window of 100 silent sites across each of the regions studied (fig. 4). In both genes, silent

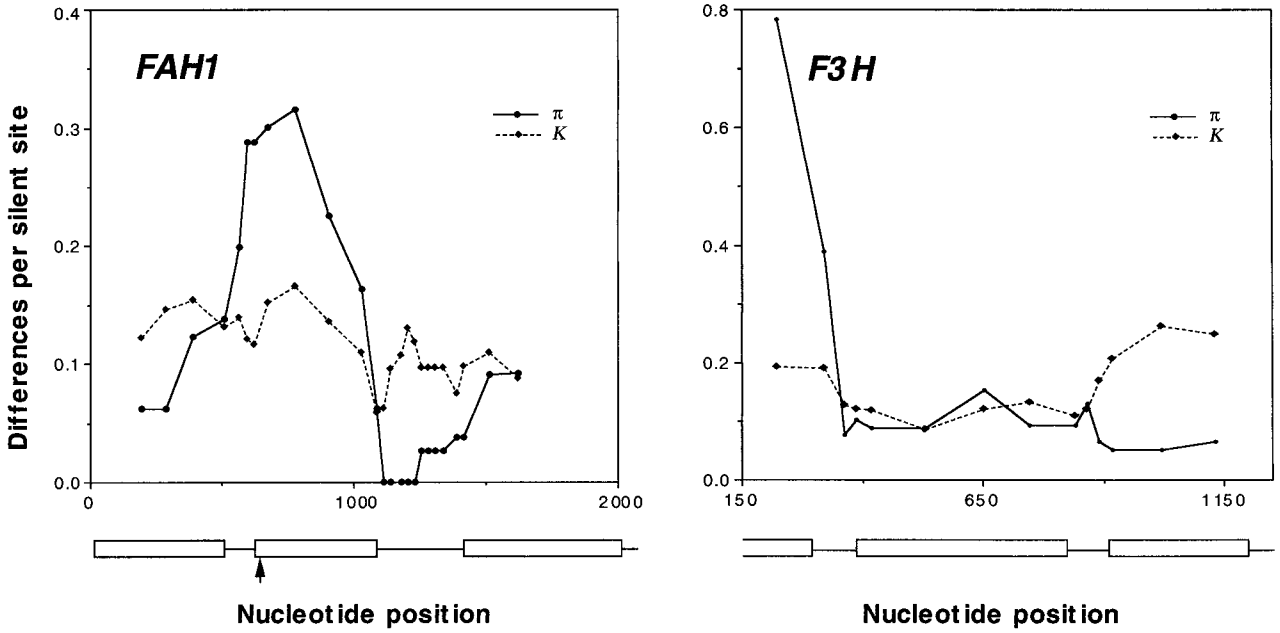

FIG. 4.—Sliding-window plot of silent nucleotide diversity (π) in the *FAH1* and *F3H* gene regions in *Arabidopsis thaliana* and silent divergence (K) between *A. thaliana* and *Arabidopsis lyrata* ssp. *petraea*. Windows include 100 silent sites, with successive displacements of 25 sites. Silent polymorphism was scaled by the divergence time $T + 1$, as estimated when applying the Hudson-Kreitman-Aguadé test. The structure of the *FAH1* and *F3H* genes is represented below the corresponding plot. The arrow below the second exon of *FAH1* indicates the nonsynonymous polymorphism at site 645.

divergence was more homogeneously distributed than silent polymorphism. In the *FAH1* gene, there was a peak of silent nucleotide diversity that, like the peak of between-haplotype silent diversity (measured as D_{xy}; fig. 3), was centered at the beginning of the second exon and, more specifically, at the nonsynonymous polymorphism at site 645. In the *F3H* gene, the estimated silent nucleotide diversity was highest at the beginning of the region studied (fig. 4).

The Hudson-Kreitman-Aguadé (HKA) test (Hudson, Kreitman, and Aguadé 1987), which tests for decoupling between polymorphism and divergence in a particular region, did not detect any significant deviation from neutrality for either the *FAH1* or the *F3H* genes when they were divided into two equal-sized fragments (results not shown). The *FAH1* region was also divided into two fragments according to the distribution of fixed differences between haplotypes (see above). Also in this case, the HKA test detected no deviation from neutrality (results not shown).

The different tests of heterogeneity in the ratio of silent polymorphism to divergence across a given DNA region developed by McDonald (1996, 1998), which do not require any a priori partition of the region studied, were also applied to the *FAH1* and *F3H* regions. The mean sliding G test revealed some heterogeneity in the ratio of polymorphism to divergence both for *FAH1* (with probabilities ranging between 0.057 and 0.085 for the different recombination values used) and for *F3H* (with probabilities ranging between 0.040 and 0.057). In the *F3H* gene, the Kolmogorov-Smirnov test also revealed some possible heterogeneity (with probabilities ranging between 0.045 and 0.067).

Discussion

Patterns of Nucleotide Variation in *A. thaliana*

Initial surveys of nucleotide variation in *A. thaliana* ecotypes revealed the presence of two highly differentiated haplotypes, or dimorphism, in both the *Adh* (Innan et al. 1996) and the *ChiA* (Kawabe et al. 1997) regions. Dimorphism was, however, restricted to a few nucleotide differences at the subsequently analyzed MADs box genes (*CAL, PI,* and *AP3*; Purugganan and Suddith 1998, 1999). If dimorphism is found to be a genomewide phenomenon, as argued by Purugganan and Suddith (1999), it could be due to the recent admixture of two differentiated populations.

The genomewide presence of dimorphism can, however, be questioned because two clear patterns of variation emerge from the genes thus far surveyed. In the *Adh* (Innan et al. 1996), *ChiA* (Kawabe et al. 1997), *ChiB* (Kawabe and Miyashita 1999), and *Rpm1* (Stahl et al. 1999) regions, variation is structured into two highly differentiated haplotypes, and most nucleotide diversity can be attributed to the between-haplotype differences, with relatively little variation present within haplotypes. On the other hand, in the *CAL, PI, AP3* (Purugganan and Suddith 1998, 1999), and *CHI* (Kuittinen and Aguadé 2000) gene regions, there is no clear evidence for two major haplotypes. Variation in the two genes studied here conforms to the first pattern.

The presence of recombinants between the two highly differentiated haplotypes in some genomic regions (*Adh, ChiA, ChiB, Rpm1, FAH1, F3H*) requires that the two haplotypes of each region segregated in the same population at some point in the past. Most prob-

Table 4
Nucleotide Diversity in Different Gene Regions

Locus	π	π_{sil}	π_a
Adh	0.0080	0.0115	0.0021
ChiA	0.0104	0.0149	0.0037
ChiB	0.0091	0.0136	0.0009
FAH1	0.0031	0.0055	0.0012
F3H	0.0072	0.0175	0.0015
CAL	0.0074	0.0080	0.0054
AP3	0.0064	0.0075	0.0040
PI	0.0053	0.0060	0.0030
CHI	0.0040	0.0054	0.0011

NOTE.—π = nucleotide diversity; π_{sil} = silent (noncoding and synonymous at coding regions) nucleotide diversity; π_a = nonsynonymous nucleotide diversity. Data sources for the different gene regions are referenced in the text.

ably, they were present in the ancestral *A. thaliana* populations of Central Asia prior to the suggested worldwide expansion of the species (Price, Palmer, and Al-Shehbaz 1994). The general lack of association between the distribution of the different haplotypes and their geographical origin would also favor the view that recombination had occurred before the species expanded its range. The contrasting patterns of variation in different gene regions seem to preclude admixture of two highly differentiated populations prior to the expansion. Also, a genomewide amplified fragment length polymorphism analysis of variation in *A. thaliana* using 38 ecotypes sampled worldwide (Miyashita, Kawabe, and Innan 1999) gives no support to the admixture hypothesis. However, neither of these observations constitutes clear evidence against population structure in the ancestral population.

Level of Nucleotide Variation in *A. thaliana* and Demography

Despite the presence of two highly differentiated haplotypes in both *FAH1* and *F3H*, silent nucleotide diversity was lower in the former gene (tables 2 and 4). This is concordant with the observed shorter extent of the region showing two major haplotypes in *FAH1* (figs. 1 and 2). Except for this gene, silent nucleotide diversity was generally higher in those regions with two major haplotypes (*Adh, ChiA, ChiB, FAH1,* and *F3H*) than in the other regions (table 4). In fact, the average silent nucleotide diversity in the former regions was 0.0126, while the average value for *CAL, AP3, PI,* and *CHI* was 0.0067.

The estimates of nonsynonymous nucleotide diversity in the *FAH1* and *F3H* genes (0.0012 and 0.0015, respectively) were somewhat similar to the estimate for *CHI* (0.0011; Kuittinen and Aguadé 2000), another gene of the phenylpropanoid pathway. They were also similar to the *ChiB* estimate (0.0009; Kawabe and Miyashita 1999), but lower than estimates for other genes thus far studied (table 4). Comparison of synonymous and nonsynonymous variation within and between species revealed an excess of nonsynonymous polymorphisms at the *CAL, PI, AP3* (Purugganan and Suddith 1998, 1999), and *ChiA* (Kawabe et al. 1997) regions, with the rarer variant at most of these sites present in only one of the ecotypes sampled. It was proposed that most nonsynonymous mutations in these regions would be slightly deleterious and would therefore have become fixed in the small *A. thaliana* populations. However, in the *FAH1* and *F3H* coding regions, as in other previously surveyed genes, no excess of nonsynonymous polymorphism was detected. That would indicate that in these regions the selection coefficients against most nonsynonymous mutations would be high enough for those mutations not to behave as neutral even in the small *A. thaliana* populations.

Recent demographic events, such as a population expansion, should affect the pattern of nucleotide variation in all parts of the genome. In regions not subject to balancing selection, the frequency spectrum of variants would be expected to be skewed toward an excess of polymorphisms with rare variants. The four genes for which there was no clear evidence for dimorphism (*CAL, AP3, PI,* and *CHI*) presented an excess of singletons. This excess resulted in a negative value of Tajima's *D* statistic, which was significant only for the first three genes (Purugganan and Suddith 1998, 1999; Kuittinen and Aguadé 2000). This observation has been considered to support a recent increase in the population size of the species.

The two highly differentiated haplotypes present in some regions (*Adh, ChiA, ChiB, Rpm1, FAH1,* and *F3H*) clearly predate the worldwide expansion of the species (see above). Even if in each gene region the two haplotypes were maintained by balancing selection, variation within each haplotype should also reflect the expansion. In all of these regions, the level of within-haplotype diversity was low compared with between-haplotypes diversity, and the low number of within-haplotype polymorphisms in most regions probably precludes the detection of any footprint of the expansion in those regions. This result might otherwise question the suggested expansion of the species.

Dimorphism in the *FAH1* and *F3H* Genes

The two divergent haplotypes present in the *Adh* region of *A. thaliana* suggested the presence of a balanced polymorphism in the fourth exon of this gene, perhaps associated with allozyme variation (Hanfstingl et al. 1994; Innan et al. 1996). In the *Rpm1* gene region, the divergent haplotypes are associated with a phenotypic difference: susceptibility or resistance to a pathogen (Stahl et al. 1999). The significant excess of silent polymorphism detected at the *Rpm1* ''junction'' region was attributed to the action of balancing selection. The similar pattern of variation observed in the other regions with clear dimorphism also suggests balancing selection. However, a dimorphic pattern of variation could also conform to the expectations of a neutral process in an essentially selfing species like *A. thaliana*. In this species, the reported level of outcrossing is less than 1% (Abbott and Gomes 1989), and plants will be mostly homozygous. The scarcity of heterozygous individuals will cause recombination to be effectively very rare.

In a constant-size neutral coalescent process with no recombination (as reviewed in Hudson 1990), the time between two coalescence events is approximately exponentially distributed. Accordingly, the expected time required for the coalescence from *n* to two sequences is nearly equal to the time required for these two sequences to coalesce to a single sequence or the most recent common ancestor (MRCA) of all sequences. Thus, the branch separating the two sets of sequences on each side of the root might accumulate a high number of mutations. Computer simulation of the coalescent process with no recombination (see *Materials and Methods*) was used to test whether the number of silent differences fixed between the two divergent haplotypes in each of the *FAH1* and *F3H* genes was consistent with the neutral process. For the *FAH1* gene, only the 5′ half of the region studied (where all fixed differences between haplotypes were located; see *Results*) was analyzed. For 12 polymorphisms and a partition of 14 and 6 sequences (fig. 1), the probability of having a number of fixed differences equal to or higher than the observed eight differences was 0.084. This probability was 0.061 when the 12 silent nucleotide polymorphisms and one indel were considered, and 0.11 when only 11 nucleotide polymorphisms were considered (those remaining after exclusion of the nucleotide polymorphism associated with the complex mutational event; see *Results*). In the *F3H* gene with 23 silent polymorphisms, the corresponding probability for a partition of 13 and 3 sequences with 16 silent fixed differences (fig. 2) was 0.065. Thus, the pattern of variation observed in both the *FAH1* and the *F3H* genes would seem to be compatible with a constant-size neutral process with no recombination.

In *A. thaliana*, recombination might be very low but not entirely absent. Recombination would decrease the probabilities of observing the actual numbers of fixed differences between the two divergent haplotypes present in both the *FAH1* and the *F3H* genes. This, together with the observed heterogeneous distribution of silent polymorphic and fixed changes across both genes, may indicate that processes other than genetic drift (e.g., selection) are contributing to the generation of the observed patterns of variation in these genes. Also, when considering all regions thus far studied in this species, it seems difficult to envisage that a neutral process with population expansion might be causing the contrasting patterns of variation detected, i.e., the presence of dimorphism in some genes and a starlike phylogeny in the rest. The number of regions surveyed is, however, not very large, but it is rapidly increasing. The joint analysis of variation in a large number of regions might be the most promising way to establish the role played by demographic events and drift, as opposed to selection, in the evolutionary history of *A. thaliana.*

Acknowledgments

I am grateful to Helmi Kuittinen and Outi Savolainen for the *A. lyrata* ssp. *petraea* seeds from Karhumäki, to Helmi Kuittinen for sharing her expertise in the Arabidopsis system, to the Nottingham Arabidopsis Stock Center for *A. thaliana* seeds, to Serveis Científico-Tècnics at Universitat de Barcelona for automated sequencing facilities, and to Julio Rozas for sharing version 3.37 of the DnaSP program and for simulations. Special thanks are given to Julio Rozas and Carmen Segarra for comments and to David Salguero for his excellent technical assistance. This work was supported by grants PB97-0918 from Dirección General de Investigación Científica y Técnica, Spain, and 1997SGR-59 from Comissió Interdepartamental de Recerca i Tecnologia, Generalitat de Catalunya.

LITERATURE CITED

ABBOT, R. J., and M. F. GOMES. 1989. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity **62**:411–418.

CHAPPLE, C. C. S., B. W. SHIRLEY, M. ZOOK, R. HAMMERSCHMIDT, and S. C. SOMERVILLE. 1994. Secondary metabolism in *Arabidopsis.* Pp. 989–1030 *in* E. M. MEYEROWITZ and C. SOMERVILLE, eds. *Arabidopsis.* Cold Spring Harbor Laboratory Press, New York.

DAY, T. A. 1993. Relating UV-B radiation screening effectiveness of foliage to absorbing-compound concentration and anatomical characteristics in a diverse group of plants. Oecologia **95**:542–550.

FEINBAUM, R. L., and R. M. AUSUBEL. 1988. Transcriptional regulation of the *Arabidopsis thaliana* chalcone synthase gene. Mol. Cell. Biol. **8**:1985–1992.

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

HANSFTINGL, U., A. BERRY, E. A. KELLOGG, J. T. COSTA III, W. RUDIGE, and R. M. 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.

HILL, W. G., and A. ROBERTSON. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. **38**:226–231.

HUDSON, R. R. 1987. Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50**:245–250.

———. 1990. Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7**:1–44.

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 **111**:147–164.

HUDSON, R. R., M. KREITMAN, and M. AGUADÉ. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

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.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–120 *in* H. M. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KAWABE, 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.

KAWABE, A., and N. T. MIYASHITA. 1999. DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopis thaliana.* Genetics **153**:1445–1453.

KELLY, J. 1998. A test of neutrality based on interlocus associations. Genetics **146**:1197–1206.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KINGMAN, J. F. C. 1982*a*. On the genealogy of large populations. J. Appl. Prob. **19A**:27–43.

———. 1982*b*. The coalescent. Stochast. Proc. Appl. **13**:235–248.

KUITTINEN, H., and M. AGUADÉ. 2000. Nucleotide variation at the *Chalcone Isomerase* locus in *Arabidopsis thaliana*. Genetics **155**:863–872.

LANDRY, G., C. C. S. CHAPPLE, and R. L. LAST. 1995. Arabidopsis mutants lacking phenolic sunscreens exhibit enhanced ultraviolet-B injury and oxidative damage. Plant Physiol. **109**:1159–1166.

LI, J., T.-M. OU-LEE, R. RABA, R. G. AMUNDSON, and R. L. LAST. 1993. Arabidopsis flavonoid mutants are hypersensitive to UV-B irradiation. Plant Cell **5**:171–179.

LOIS, R. 1994. Accumulation of UV-absorbing flavonoids induced by UV-B radiation in *Arabidopsis thaliana*. Planta **194**:498–503.

MCDONALD, J. H. 1996. Detecting nonneutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **13**:253–260.

———. 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.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade: analysis of phylogeny and character evolution. Version 3.0. Sinauer, Sunderland, Mass.

MEYER, K., J. C. CUSUMANO, C. SOMERVILLE, and C. C. S. CHAPPLE. 1996. Ferulate-5-hydroxylase from *Arabidopsis thaliana* defines a new family of cytochrome P450-dependent monooxygenases. Proc. Natl. Acad. Sci. USA **93**: 6869–6874.

MIYASHITA, N. T., A. KAWABE, and H. INNAN. 1999. DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. Genetics **152**:1723–1731.

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

PELLETIER, M. K., and B. W. SHIRLEY. 1996. Analysis of flavanone 3-hydroxylase in Arabidopsis seedlings. Plant Physiol. **111**:339–345.

PRICE, R. A., J. D. PALMER, and I. A. AL-SHEHBAZ. 1994. Systematic relationships of *Arabidopsis*: a molecular and morphological perspective. Pp. 7–19 *in* E. M. MEYEROWITZ and C. SOMERVILLE, eds. *Arabidopsis*. Cold Spring Harbor Laboratory Press, New York.

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.

ROGERS, S. O., and A. J. BENDICH. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol. Biol. **5**:69–76.

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

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SHIRLEY, B., S. HANLEY, and H. M. GOODMAN. 1992. Effects of ionizing radiation on a plant genome: analysis of two Arabidopsis *transparent testa* mutations. Plant Cell **4**:333–347.

STAHL, M., G. DWYER, R. MAURICIO, M. KREITMAN, and J. BERGELSON. 1999. Dynamics of disease resistance polymorphism at the *Rpm1* locus of Arabidopsis. Nature **400**: 667–671.

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

VAN DE PEER, Y., and R. DE WACHTER. 1994. TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput. Appl. Biosci. **10**:569–570.

WATTERSON, G. A. 1974. The sampling theory of selectively neutral alleles. Adv. Appl. Prob. **6**:463–488.

WISMAN, E., U. HARTMAN, M. SAGASSER, E. BAUMANN, K. PALME, K. HAHLBROCK, H. SAEDLER, and B. WEISSHAAR. 1998. Knock-out mutants from an En-1 mutagenized *Arabidopsis thaliana* population generate phenylpropanoid biosynthesis phenotypes. Proc. Natl. Acad. Sci. USA **95**: 12432–12436.

WOLFGANG STEPHAN, reviewing editor

Accepted September 6, 2000