Nucleotide Variation at the *CHALCONE ISOMERASE* Locus in *Arabidopsis thaliana*

Helmi Kuittinen and Montserrat Aguadé

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, 08071 Barcelona, Spain

Manuscript received October 15, 1999
Accepted for publication February 14, 2000

ABSTRACT

An ~1.9-kb region encompassing the *CHI* gene, which encodes chalcone isomerase, was sequenced in 24 worldwide ecotypes of *Arabidopsis thaliana* (L.) Heynh. and in 1 ecotype of *A. lyrata* ssp. *petraea*. There was no evidence for dimorphism at the *CHI* region. A minimum of three recombination events was inferred in the history of the sampled ecotypes of the highly selfing *A. thaliana*. The estimated nucleotide diversity ($\theta_{TOTAL} = 0.004$, $\theta_{SIL} = 0.005$) was on the lower part of the range of the corresponding estimates for other gene regions. The skewness of the frequency spectrum toward an excess of low-frequency polymorphisms, together with the bell-shaped distribution of pairwise nucleotide differences at *CHI*, suggests that *A. thaliana* has recently experienced a rapid population growth. Although this pattern could also be explained by a recent selective sweep at the studied region, results from the other studied loci and from an AFLP survey seem to support the expansion hypothesis. Comparison of silent polymorphism and divergence at the *CHI* region and at the *Adh1* and *ChiA* revealed in some cases a significant deviation of the direct relationship predicted by the neutral theory, which would be compatible with balancing selection acting at the latter regions.

DNA sequences accumulate information that can reveal the important factors contributing to the evolutionary dynamics of a species. Thus far, most data of nucleotide variation are from Drosophila species, while only a few data sets are on sequence variation in plants. Compared to obligatorily outcrossing animals, plants provide an additional variable, breeding system, which affects the level and pattern of genetic variability in many ways. The level of neutral genetic variability in an inbreeder is expected to be low as compared to a similar, outcrossing species. Theoretically, the smaller effective population size of the inbreeder (Pollak 1987), together with the effects of background selection (Charlesworth *et al.* 1993) and hitchhiking (Hedrick 1980) should result in comparatively reduced levels of variability. In case of balancing selection acting at a locus, the signal in an inbreeder should be strong due to the genome-wide reduced variation caused by background selection (Nordborg *et al.* 1996). Reduced migration and the maintenance of allele combinations at different loci (Allard *et al.* 1968) as well as the relatively high probability of ultimate survival of new favorable recessive mutations (Pollak 1987) contribute to the great potential for local adaptation of inbreeding species. The selfing mode of reproduction also has an effect on the distribution of deleterious mutations. Inbreeding species should be free of strongly deleterious mutations because of purging (Husband and Schemske 1996). On the other hand, mutations of small deleterious effect can easily become fixed in small populations due to drift (Lynch and Walsh 1998). In fact, drift is thought to play an important role in inbreeding species not only because of the smaller effective population sizes but also because inbreeding species may go through severe bottlenecks through their ability to establish a new population from a single seed.

Arabidopsis thaliana is highly inbreeding (Abbot and Gomez 1989). It inhabits sandy or rocky soil, road sides, and other disturbed areas. Due to its life history traits, *A. thaliana* is vulnerable to rapid colonization and extinction cycles. This can result in the observed lack of variability in populations at the pheripheral areas of its distribution (Kuittinen *et al.* 1997). *A. thaliana* is thought to have gone through rapid expansion over the world from its suggested center of distribution in Central Himalaya (Price *et al.* 1994). This is supported by the lack of association between the geographical origin and genetic distance between ecotypes (*e.g.*, Ullrich *et al.* 1997). On the basis of dimorphism of alleles at different loci, it has been speculated that before the expansion two isolated *A. thaliana* populations were fused or that there has been introgression from another species (Innan *et al.* 1996). Some selective scenarios could also explain the observed results.

Products of the flavonoid pathway, flavonoids and tannins, are pigments in flowers, leaves, fruits, and seeds. These compounds protect the plant against microorganisms and herbivores, and they also act as UV-light ab-

Corresponding author: Helmi Kuittinen, Department of Biology, University of Oulu, PL3000, 90401 Oulu, Finland.
E-mail address: helmi.kuittinen@oulu.fi

sorbants. Some of these compounds are involved in symbiotic relationships between plants and microorganisms (Chapple *et al.* 1994). *CHALCONE FLAVANONE ISOMERASE* (*CHI*) codes for the second enzyme in the branch leading from the general phenylpropanoid pathway to flavonoid and tannin biosynthesis. While in many plants CHI is encoded by a small gene family, in *A. thaliana* it is a single-copy gene (Shirley *et al.* 1992). CHI seems to be one of the key enzymes involved in protection against ultraviolet-B (UVB) light, as mutants at this gene are among the flavonoid mutants most sensitive to elevated levels of UV-B (Li *et al.* 1993). Mutants at *CHI* have reduced levels of both leaf flavonoids and sinapate esters, products of an earlier branch from the general phenylpropanoid pathway, suggesting a link between these two branches (Li *et al.* 1993). There is natural variation in the production of flavonoids (Ziska *et al.* 1992) and in the sensitivity to UV-B between plants from different elevations (Sullivan *et al.* 1992). *A. thaliana* occurs in areas with variable levels of natural UV-B irradiation, such as high and low elevations and different latitudes, and thus variation at the *CHI* gene might be contributing to the adaptive phenotypic variation. Here we study the level and pattern of nucleotide variation at the *CHI* gene in a sample of natural *A. thaliana* ecotypes to ascertain the relevant evolutionary factors influencing the population dynamics of this annual selfing weed.

MATERIALS AND METHODS

Sequencing: *A. thaliana* ecotypes were obtained from the Nottingham Stock Center or collected from the field. The ecotypes were randomly sampled from a wide range of geographical origins (Table 1). One seed from each of 23 accessions was used. Seeds were grown and DNA was extracted from the leaves. The *CHI* gene region was amplified with primers designed from a published sequence of the Landsberg *erecta* (L*er*) laboratory line (GenBank accession no. M86358). The analyzed region corresponds to bases 51–1974 of the L*er* sequence, including the complete coding region (741 bp), 610 bp of the 5′ flanking region, three introns (426 bp), and 147 bp of the 3′ flanking region. The amplified fragment was purified with Qiaquick (QIAGEN, Chatsworth, CA) columns and used as a template for direct cycle sequencing with fluorescent dideoxy terminators according to the manufacturers' instructions (Pelkin-Elmer, Norwalk, CT; Amersham, Buckinghamshire, England). Both strands were sequenced and the alleles were aligned manually. The published L*er* sequence was included in the analyses. The *A. lyrata* ssp. *petraea* [syn. *Arabis petraea* (L.) Lam] sequence corresponding to bases 440–1934 of the L*er* sequence was obtained in a similar manner. The newly determined sequences will appear in the EMBL, GenBank, and DDBJ sequence databases under accession nos. AJ287299–AJ287322.

Analyses: A neighbor-joining tree was constructed with the TREECON version 1.3b program (Van de Peer and De Wachter 1994) using the number of nucleotide differences as a distance measure and no outgroup. The minimum number of recombination events was calculated according to Hudson and Kaplan (1985). Their method uses the four-gamete test and assumes no homoplasy. This is a reasonable assumption because of the low frequency of polymorphic sites in the analyzed set of sequences. The significance of pairwise linkage disequilibrium was tested with Fisher's exact test, excluding noninformative sites (singletons).

The expected numbers of pairwise nucleotide differences were calculated assuming no recombination. This was done for situations of both constant population size (Watterson 1975; Slatkin and Hudson 1991; Rogers and Harpending 1992) and of allowing the population to grow or decline (Rogers and Harpending 1992). The value of θ after population growth or decline was set to infinity to enable us to estimate the initial value of θ and $2ut$ from the data (Rogers 1995). The smoothness of the pairwise difference distribution was quantified with the raggedness statistic (Harpending 1994) and its value was tested with computer simulations using the coalescent algorithm.

Tajima's (1989) and Fu and Li's (1993) tests were applied to the data using the total number of mutations to contrast whether the sample was in mutation-drift equilibrium. Significant nonzero values of the test statistics can appear due to a recent bottleneck or expansion, or if selection is acting at the locus. Under a constant-rate neutral model the levels of polymorphism and interspecific divergence at a given locus should be directly related as both measures are dependent on the neutral mutation rate. The Hudson, Kreitman, and Aguadé test (HKA test; Hudson *et al.* 1987) was performed to test this prediction. For the interspecific comparison, the sequence of *A. lyrata* ssp. *petraea* was used.

All the analyses, except for construction of a neighbor-joining tree, were performed using the DnaSP software version 3.0 package (Rozas and Rozas 1997).

RESULTS

Twenty different haplotypes were found in the sample of 24 ecotypes. One haplotype was present three times and another two were present twice in the sample. There were 41 nucleotide, 7 indel and 1 A/T repeat polymorphisms (Figure 1). The number of singletons was 21 after removing the redundant haplotypes from the analysis. In the COL-2 ecotype there was a coincidental duplication of a 10-bp sequence and deletion of 59 bp. There were only five polymorphic sites in the coding region. Of the 3 nonsynonymous polymorphisms, 2 involved a conservative amino acid change (Ile/Val and Leu/Val, respectively), while the third one was a change in the stop codon. The rare variant of this last polymorphism caused the extension of the translated region by two residues (Arg and Glu). Only the Leu/Val polymorphism segregated at intermediate frequencies in the sample. The levels of both total and silent nucleotide polymorphism were low (Table 2). The level of synonymous polymorphism was similar to that in the introns. These estimates were, however, lower than those for both the 5′ and 3′ flanking regions. In contrast, divergence estimates for the noncoding regions (5′ and 3′ flanking and introns) were rather similar (Table 2), but lower than synonymous divergence in the coding region.

The minimum number of recombination events in the history of the set of sequences was three. No association between geographical origin and proximity in the

TABLE 1

Names and origins of ecotypes used in this study

Name	Origin, country	Longitude/latitude	Accession no.[a]
CAN-0	Canary Islands, Spain	W15–16/N28	N1064
CHA-0	Champex, Switzerland	E7/N46	N1068
COL-2	Landsberg, Germany[b]	E11/N48	N907
CON	Condara, Khurmatov, Tadjikistan	E68–75/N36–40	N916
CVI-0	Cape Verdi Islands, Portugal	W23–25/N15–17	N902
ED1	Eden, Sweden	E18–19/N63–64	
GR-5	Graz, Austria	E15–16/N45–47	N1206
GRAN1	Granollers, Spain	E2–3/N41–42	
GRAP1	Graponne, France	E4–5/N45–46	
ITA-0	Ibel Tazekka, Morocco	W4/N35	N1244
KAS-1	Kashmir, India	E74–80/N34–36	N903
L*er*	Landsberg, Poland[c]	E11/N48	
LU1	Lulep Istjak, Sweden	E16–17/N66	
ME-0	Mechtshausen, Germany	E9–10/N51–52	N1364
MH-0	Muehlen, Poland	E20–21/N53–54	N904
MR-0	Monte/Tosso, Italy	E9–10/N44–45	N1372
PER-1	Perm, Russia	E56/N58	N1444
RI-0	Richmond, BC, Canada	W123/N49–50	N1492
RUB-1	Rubezhnoe, Ukraine	E38–39/N49–50	N927
RV1	Ruds Vedby, Denmark	E11–12/N55–56	
TV1	Tvärminne, Finland	E23–24/N59–60	
TUL-0	Turk Lake, FL, USA	W80–81/N25–26	N1570
WLP1	Wilp, Netherlands	E5–6/N52–53	
YO-0	Yosemite, CA, USA	W119–120/N38	N1622

[a] Accession number at the Nottingham Arabidopsis Stock Center.
[b] Selected from La-0 by Redei and taken through single seed by Shauna Somerville.
[c] Selected from La-0 after X-ray mutagenesis.

neighbor-joining tree could be observed except for the identical haplotypes (CON and KAS-1; YO-0, RI-0, and TUL-0; Ler and WLP1) that originated from nearby regions (Figure 2). Tests of linkage disequilibrium revealed a significant association in 27 of the 171 pairwise combinations (16%). Only four pairwise combinations were significant when the Bonferroni correction for multiple comparisons was applied.

The observed distribution of pairwise nucleotide differences was compared to the distributions expected under the assumptions of either constant population size or population growth/decline (Figure 3). The gradually declining curve expected with constant population size is an average of what is expected in a large number of populations, *i.e.*, a single population is not expected to follow that curve, but rather to have several peaks. In contrast, if a population has experienced a rapid growth the distribution tends to have one peak "wave" (Rogers and Harpending 1992). The raggedness statistic, $r = 0.0145$, gives support for one peak as $P(r < r_{obs}) = 0.032$. Concordant with this result, Tajima's test and Fu and Li's tests indicated an excess of singletons as expected under the rapid growth hypothesis. However, none of these tests were significant: Tajima's D was -0.98 ($P > 0.10$) and Fu and Li's D^* and F^* were -0.94 ($P > 0.10$) and -1.12 ($P > 0.10$), respectively.

Figure 4 presents a comparison of the level of silent polymorphism in *A. thaliana* and the degree of silent divergence between *A. thaliana* and *A. lyrata* ssp. *petraea* in the different functional regions. The relationship between the level of polymorphism in *A. thaliana* and the degree of divergence between the two species does not follow the expectation of a constant-rate neutral model, as revealed by the HKA test (Table 3).

DISCUSSION

Weak support for dimorphism: Previous studies of nuclear genes in *A. thaliana* had revealed dimorphism characterized by the presence of two groups of sequences where blocks of variants are shared within a group and fixed between groups. The neighbor-joining tree based on *CHI* sequences was bifurcated with a rather low (34%) bootstrap value at the basal branching point, and the branches leading to the first nodes in the tree were short. Among the alleles at *CHI*, three recombination events could be inferred by the four-gamete test. If there were dimorphism, the two parental types should be in the two different groups and the recombinants should be distributed among the groups. However, parental and recombinant types could not be identified based on partitioning of polymorphisms.

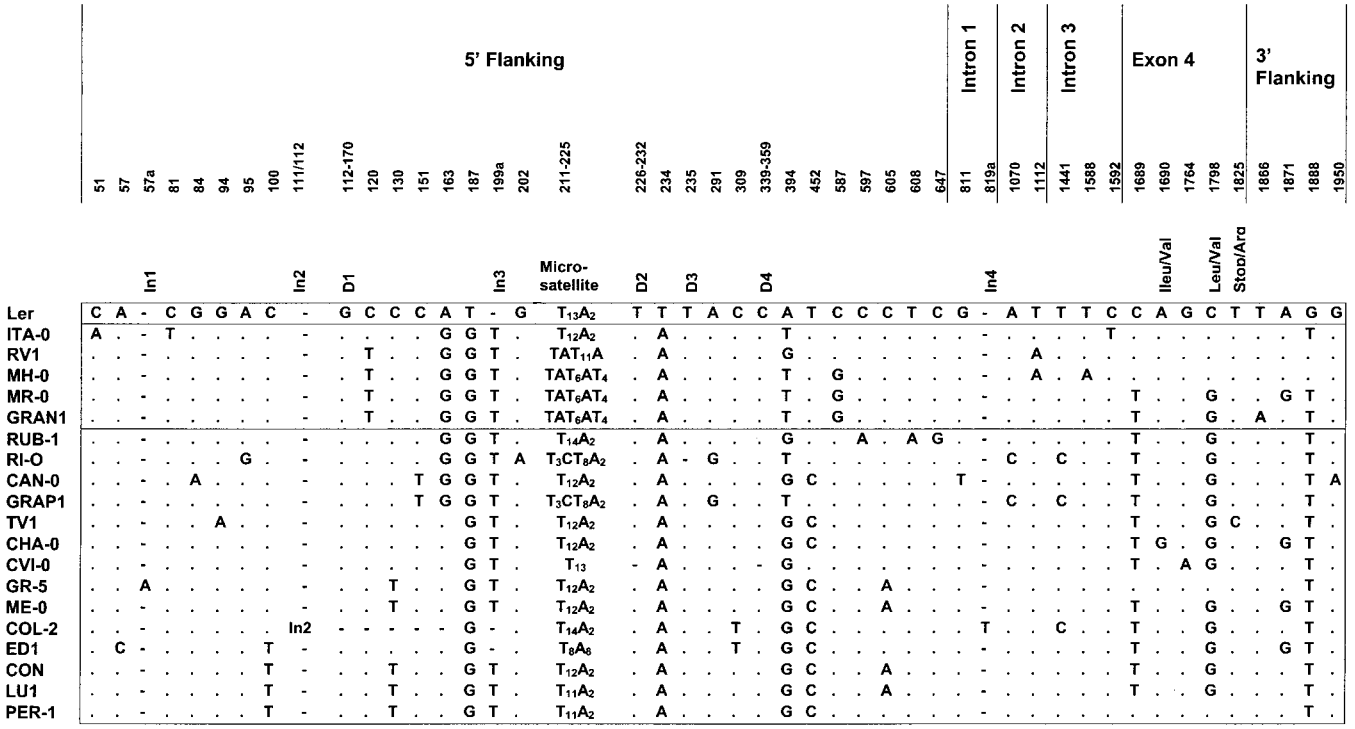

Figure 1.—Sequence polymorphisms. According to the four-gamete test, recombinations have occurred between sites 100 and 130, sites 151 and 291, and sites 587 and 1112. In, insertion; D, deletion. A dot represents the same nucleotide as L*er*; a hyphen represents the absence of the corresponding insertion or nucleotide(s). In2 = CTTTGATTTT. Three haplotypes were represented by more than one ecotype: Ler = WLP1, RI-0 = YO-0 = TUL-0, CON = KAS-1. The horizontal line dividing the ecotypes into two groups reflects the bifurcation of the neighbor-joining tree.

Concordantly, the division of the ecotypes in the two groups by the neighbor-joining procedure could not be tracked easily from the table of polymorphisms. In fact, there were no sites or groups of sites that would give nearly the same partitioning as the neighbor-joining procedure (Figure 1). In all other published data sets, it had been possible to identify sites that divided the sequences in two parental classes. However, in *CAL*, *AP3*, and *PI*, the dimorphism involved only a few sites that in the case of *AP3* extended over a short distance (Purugganan and Suddith 1998, 1999), and in fact the overall patterns of nonrandom associations resembled that found in the *CHI* gene region. In *Adh1* and *ChiA* the dimorphism was clear (Innan *et al.* 1996; Kawabe *et al.* 1997).

Kawabe *et al.* (1997) suggested that the whole genome of *A. thaliana* could be dimorphic. The frequencies of the two allele classes at the nuclear loci do not seem homogenous, as the ratio varies from 5:3 at *Adh1* (blocks 1 and 6 from Innan *et al.* 1996) to 13:2 at *ChiA* (Kawabe *et al.* 1997). In an amplified fragment length polymorphism (AFLP) study of a worldwide sample of *A. thaliana*, Miyashita *et al.* (1999) found that the frequency distribution of AFLP bands followed that expected under a stochastic model with one basal population, except for a high occurrence of singletons and a low frequency of fixed bands. No excess of any intermediate class was evident in the analysis. Thus, there is no support for dimorphism at least with homogenous frequencies.

The occurrence of dimorphism at *Adh1* and *ChiA* was explained either by balancing selection or by historical events like the fusion of two divergent populations or introgression from another species (Innan *et al.* 1996; Kawabe *et al.* 1997). If the latter cases are true, the other allele type at *CHI* has disappeared due to drift or selection at *CHI* or another linked locus. This would be in concordance with the relatively low levels of variation found in the *CHI* region but, on the other hand, perhaps a more extreme Tajima's *D* value would be expected in that case. Another explanation would be frequent recombination breaking nonrandom associations after the putative introgression. However, with only one base population, drift combined with low levels of recombination might also be able to produce the kinds of patterns observed, at least at genes other than *Adh1* and *ChiA*.

Variance in the distribution of pairwise nucleotide differences: The smooth bell-shaped distribution of the pairwise number of nucleotide differences at *CHI* and the excess of rare variants are concordant with the expansion model. However, the distribution of pairwise nucleotide differences can be affected not only by expansion but also by recombination and selection. Se-

TABLE 2

Summary of nucleotide polymorphism and divergence at the *CHI* region

	Total coding	Nonsyn.	Syn.	Total noncoding	5′ flanking	Introns	3′ flanking	Syn. and noncoding	Total
No. of sites	741	563.4	177.6	1088	515	426	147	1265.6	1829
S	5	3	2	31	21	6	4	33	36
π	0.0015	0.0011	0.0029	0.0058	0.0087	0.0025	0.0051	0.0054	0.0040
θ	0.0018	0.0014	0.0030	0.0079	0.0114	0.0038	0.0073	0.0072	0.0053
K	0.0845	0.0494	0.1962	0.1009	0.0808	0.1123	0.0992	0.1193	0.0927

Nonsyn., nonsynonymous; syn., synonymous; *S*, number of segregating sites; *K*, average proportion of nucleotide differences between *A. thaliana* and *A. lyrata* ssp. *petraea*.

lective sweeps and recombination both reduce the variance of the pairwise nucleotide differences, while balancing selection would increase the variance. Expansion should affect all genes equally, so the same kind of "wave" should be seen in all genes. In contrast, effects of selection and recombination could differ from region to region resulting in variable patterns of nucleotide polymorphism. For *CAL*, *PI*, and *AP3*, the unimodal distribution of the number of pairwise nucleotide differences seems also concordant with the expansion hypothesis, but not for *Adh1*, *ChiA*, and *ChiB* with multimodal distributions (Figure 3). Although the discordancy among the pictures of the four genes would not support expansion, other aspects of the data seem to favor that hypothesis. For the *CHI* region, identical haplotypes were found in nearby locations: three identical haplotypes originated from America, two from India and Tadjikistan, and two from The Netherlands and Germany, indicating strong recent founder effects. Otherwise, no geographical associations were found, similar to the previous studies. This, together with the negative Tajima's *D* values found in all the studied regions, can be taken as strong evidence for the recent expansion hypothesis. Also, an extensive AFLP study (Miyashita *et al.* 1999) revealed a starlike phylogeny of the different ecotypes, supporting a recent expansion. If expansion is taken as a fact, and the wavelike distribution of pairwise differences found for AFLP fragments and for the *CHI*, *CAL*, *PI*, and *AP3* genes is due to it, *Adh1*, *ChiA*, and *ChiB* could be considered outliers.

A high level of recombination in the selfing *A. thaliana*: A selfing species can be considered as a group of homozygous lines. The outcrossing rate of *A. thaliana* is lower than 0.3% (Abbot and Gomez 1989), and it is thus a true selfer. The level of effective recombination in a highly selfing species would generally be expected to be low, because of the rare occurrence of double heterozygotes. However, in *A. thaliana* each additional

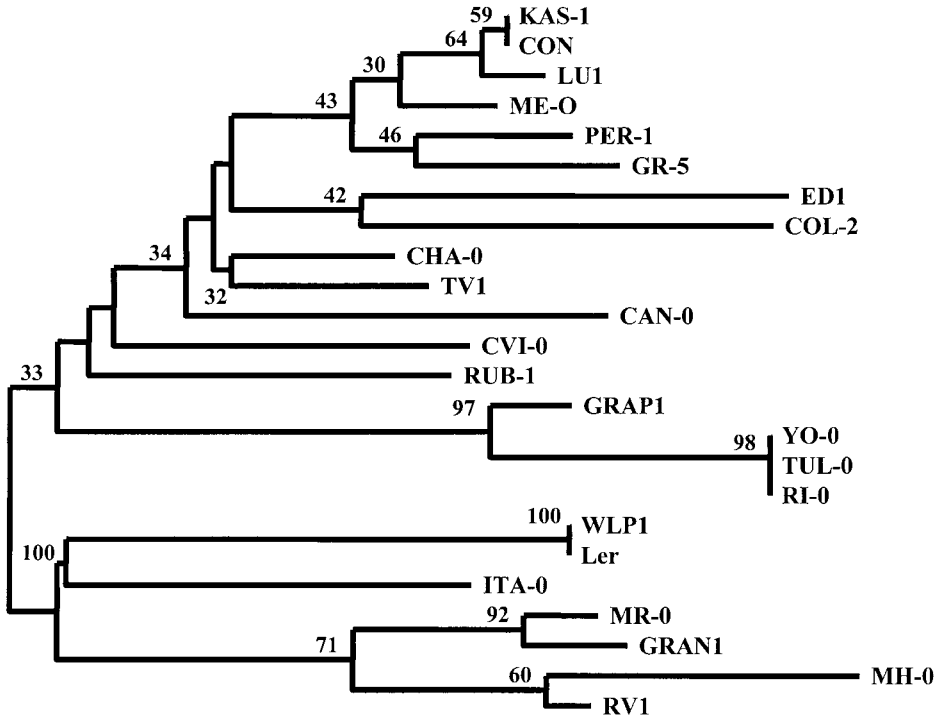

Figure 2.—A neighbor-joining tree of the 24 ecotypes based on the *CHI* gene region sequences. The bootstrap values ≥30% are given next to the node they refer to. The unit for branch length is number of nucleotide differences.

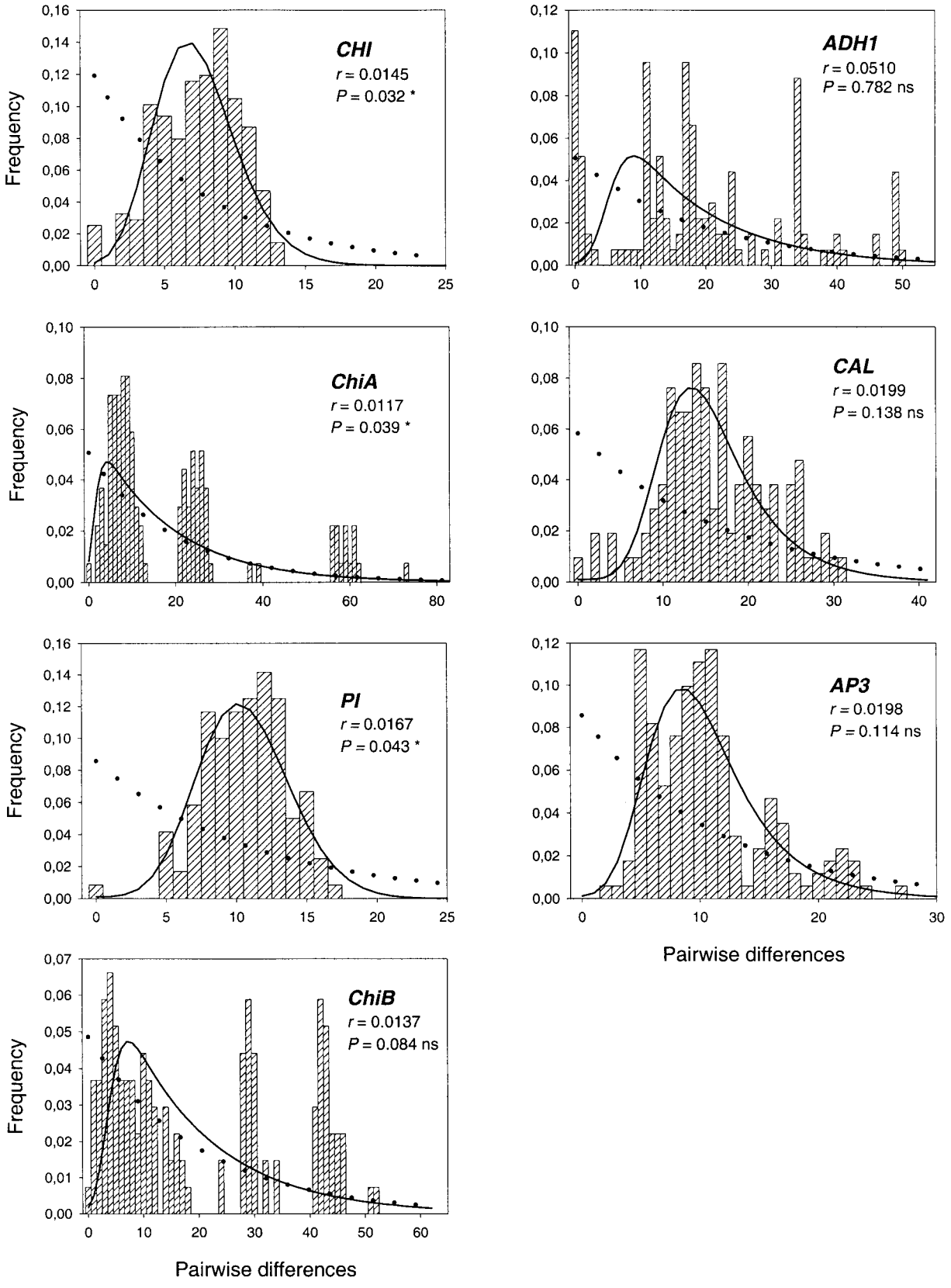

Figure 3.—Observed (columns) and expected distribution of the number of pairwise nucleotide differences under the assumption of constant population size (dotted line) or the assumption of population growth/decline (solid line) at seven loci. Distributions for *Adh1*, *ChiA*, *CAL*, *PI*, *AP3*, and *ChiB* were produced from sequences reported in Innan *et al.* (1996), Kawabe *et al.* (1997), Purugganan and Suddith (1998, 1999), and Kawabe and Miyashita (1999). For each region, both the raggedness statistic *r* and its associated *P* value are given. ns, not significant. *$0.01 < P < 0.05$.

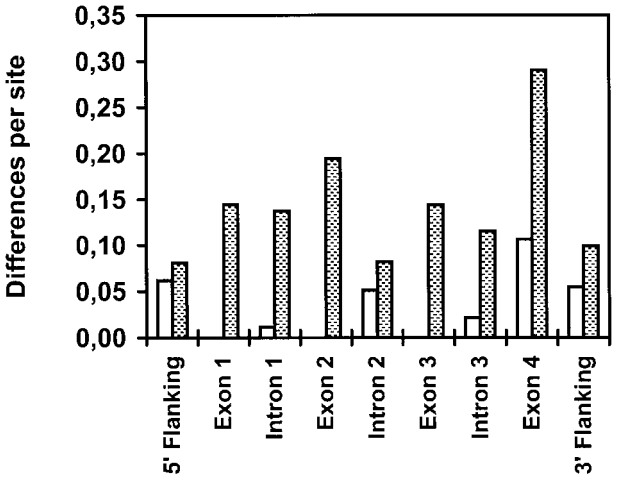

Figure 4.—Level of silent (π_{SILENT}) polymorphism (open bars) within *A. thaliana* compared to the silent (K_{SILENT}) divergence (shaded bars) between *A. thaliana* and *A. lyrata* ssp. *petraea*, in different functional regions of the *CHI* gene.

sequence data set is showing that there has been recombination in the history of the sampled sequences.

At the *CHI* region, the minimum number of recombination events (R_M) inferred with the four-gamete test in the history of the sample was three. Also, in other genes (*Adh1*, *ChiA*, *CAL*, *PI*, *AP3*, *ChiB*) several recombinations had been inferred (Table 4). The power to detect recombination events is dependent on the number of informative sites, which varies in the different regions. After scaling with the number of informative sites, R_M is very similar among the different regions studied in *A. thaliana*, indicating homogeneity in recombination frequency. On the other hand, the estimated recombination parameter, or C value ($C = 4Nc$; Hudson 1987), and the $4Nc/4N\mu$ ratio, which indicates the number of recombination events per mutation event (Hudson *et al.* 1994), vary considerably between the different regions studied in *A. thaliana* (Table 4). The highest values are found at the *CHI* and *PI* regions. Estimation of the recombination parameter is sensitive to violations of the mutation-drift equilibrium assumption due to bottlenecks, expansion, and population subdivision. As all these processes affect equally all genes, the differences between estimates of the recombination parameter among different genes in *A. thaliana* should reflect differences either in the recombination frequency or in selection. The homogeneity of the scaled R_M values in the different regions suggests that the observed differences among genes would rather be due to differences in selection than in the recombination frequency. With the current data the possibility of differences in recombination cannot be excluded either. However, there would be no *a priori* reasons to expect different recombination rates at the different loci studied, because none of them seems to be located in a region of low recombination such as centromeric or telomeric regions.

The question arises of why there is so much recombination in a selfing species like *A. thaliana*. Either the observed recombinations just accumulate over the historical time and the number in the present sample is not higher than what would be expected, or the frequency of recombinants has been increased because of some evolutionary process. One possible explanation could lie in the frequent fixation of slightly deleterious mutations in the small *A. thaliana* populations. If slightly deleterious (partly) recessive mutations become fixed, one would expect some heterosis when crossing plants from different populations. There is indeed some evidence of heterosis in *A. thaliana* (Griffing 1989). The excess of nonsynonymous polymorphism detected in some genes also supports the abundance of slightly deleterious mutations in this species (Kawabe *et al.* 1997; Purugganan and Suddith 1999). In that case, individuals resulting from rare outcrossing would be heterozygous for the slightly deleterious mutations derived from the parental lines. Among the progeny of these hetero-

TABLE 3

Summary of the results of HKA tests

	Segregating sites within *A. thaliana*	Divergence between *A. thaliana* and *A. lyrata* ssp. *petraea*/*lyrata*	*P* value
CHI_{TOTAL}	20	136.25	
$Adh1_{CODING}$[a]	19	49.60	0.0639
$ChiA_{CODING}$[a]	44	46.88	0.0003***
CHI_{SILENT}	18	108.71	
$Adh1_{SYNONYMOUS}$[a]	13	33.12	0.1103
$ChiA_{SYNONYMOUS}$[a]	25	23.94	0.0004***
$CHI_{SYNONYMOUS}$	2	25	
$Adh1_{SYNONYMOUS}$[a]	13	33.12	0.0348*
$ChiA_{SYNONYMOUS}$[a]	25	23.94	0.0012**

* $0.01 < P < 0.05$; ** $0.001 < P < 0.01$; *** $P < 0.001$.

[a] The numbers of segregating sites and divergence for *Adh1* and *ChiA* were taken from Kawabe *et al.* (1997).

TABLE 4

Recombination in different regions

Gene region	Location	R_M	No. of informative sites	R_M/no. of informative sites	$C = 4Nc$/gene	C per nucleotide	$\theta = 4N\mu$/gene	$4Nc/4N\mu$
CHI	III-85 cM	3	20	0.15	132	0.0688	7.4	17.8
ChiA	V	2	19	0.11	0.001	0.0000	18.7	0.00005
Adh1	I-113 cM	6	37	0.16	3.3	0.0014	18.7	0.18
CAL	I-46 cM	3	25	0.12	53.2	0.0242	16.2	3.3
PI	V-28 cM	1	9	0.10	*	—	10.7	—
AP3	III-81 cM	1	9	0.10	38.6	0.0231	10.7	3.6
ChiB	III-19 cM	5	60	0.08	1.3	0.0006	19.6	0.066

R_M, minimum number of recombination events in the history of the sample. * >10,000.

zygous individuals, recombinants with fewer slightly deleterious mutations would be favored. This might lead to a higher level of observed recombination. However, if that were the main reason for the relatively high impact of recombination in a selfer, a similar increase would be expected in other selfing species. However, in *Hordeum vulgare* with an estimated outcrossing rate of 1.6% (Brown *et al.* 1978) the recombination parameter C for the *Adh1* region was zero (Cummings and Clegg 1998). The discrepancy could be due to differences in the fixation probability of slightly deleterious mutations, *e.g.*, through differences in population size or other life history traits, or differences in the outcrossing rate. Overall, the surprisingly high impact of recombination in a selfing species remains an interesting question.

Level and pattern of polymorphism and divergence: The level of variability at the *CHI* region, measured as total and silent nucleotide polymorphism, was at the lower part of the range of variation at the regions studied thus far in *A. thaliana*. At the *CHI* region, the levels of synonymous and intron polymorphism were equally low and lower than in the flanking areas. This is in contrast with the general pattern observed in different species of Drosophila, where the level of synonymous polymorphism is usually higher than in introns or in 5′ or 3′ flanking areas (Moriyama and Powell 1996). Introns and 5′ and 3′ flanking regions are considered to be relevant in the regulation of gene expression, leading to higher selective constraints at these regions than at synonymous sites. However, at the *CHI* gene, only a short regulatory sequence, probably mediating the response to UV light, has been identified ~150 bp upstream of the initiation codon (Li *et al.* 1993).

In previous studies an excess of nonsynonymous polymorphism had been found (Kawabe *et al.* 1997; Purugganan and Suddith 1998, 1999), which was attributed to the fixation of mildly deleterious mutations in small populations. In our data set, no significant excess of replacement polymorphisms was detected. However, in this case the McDonald-Kreitman test (McDonald and Kreitman 1991) has very little power due to the low number of polymorphisms (five) in the coding region of *CHI*.

The synonymous divergence between *A. thaliana* and *A. lyrata* was 0.11, 0.12, and 0.15 for *AP3*, *CAL*, and *PI*, respectively (Purugganan and Suddith 1998) and 0.15 for *Adh1* (Savolainen *et al.* 2000). The average proportion of nucleotide differences at synonymous sites between *A. lyrata* and *A. thaliana* was somewhat higher for *CHI* (0.19) than for these other nuclear genes. However, the ratio of replacement to synonymous divergence at *CHI* (0.25) was within the range observed for *AP3*, *CAL*, and *PI* (0.148–0.271; Purugganan and Suddith 1998).

Rausher *et al.* (1999) have estimated nonsynonymous replacement rates at *CHI* between one monocot and two dicot species. Divergence in the Zea-Ipomoea/Antirrhinum comparison was 0.290 (nonsynonymous substitutions per nonsynonymous site); this results in a rate of 1.2×10^{-9} considering 120 million years as the divergence time between monocots and dicots. The rate estimated from the divergence between Ipomoea and Antirrhinum (0.197) would be $2.5–2 \times 10^{-9}$, considering 40–50 million years as their divergence time. The exact divergence time of *A. lyrata* and *A. thaliana* is not known. It has been estimated as 3.8–5.8 million years based on the rates of synonymous substitutions at *Adh1* and *CHS* in the genus Arabis, calibrating the rates with the fossil pollen records of the genus Rorippa (Marcus Koch, personal communication). Assuming 6 million years as the divergence time, the rate of nonsynonymous substitution estimated from the *A. lyrata*/*A. thaliana* comparison, 4×10^{-9}, would be twofold higher than in deeper lineages. On the other hand, the estimated per year rate of synonymous substitutions (1.6×10^{-8}), given the 6-million year divergence time between *A. lyrata* and *A. thaliana*, is similar to that in the Ipomoea-Antirrhinum comparison ($1.4–1.8 \times 10^{-8}$). Thus, there seems to be heterogeneity among the nonsynonymous substitution rates between lineages, although these results are

dependent on the correct estimation of the divergence time.

As revealed by the HKA test, the relationship between the level of polymorphism in *A. thaliana* and the level of divergence between *A. thaliana* and *A. lyrata* was not always concordant at different loci. The ratio of polymorphism to divergence was significantly lower at *CHI* than in the other two regions (*Adh1* and *ChiA*) analyzed. There is therefore a decoupling between levels of polymorphism and divergence, which affects differentially the *Adh1* and *ChiA* regions, and the *CHI* region. Polymorphism at the *CHI* region was reduced, which could be explained by the hitchhiking effect of an advantageous mutation. This hypothesis would be supported by the negative values of Tajima's *D* and by the bell-shaped distribution of the number of pairwise nucleotide differences. However, both features could also be reflecting the recent expansion of the species (see above). Another possibility would be that both *Adh1* and *ChiA* had increased levels of polymorphism. It has been suggested that balancing selection might be acting at *Adh1* (Hanfstingl *et al.* 1994; Innan *et al.* 1996; Kawabe *et al.* 1997), which could give rise to high intraspecific variability. There is, however, no indication of balancing selection at *ChiA*, although this gene presents, like *Adh1*, a clear dimorphism in the surveyed ecotypes. In *A. thaliana*, the level of genome-wide variation revealed by AFLP analysis (Miyashita *et al.* 1999) is similar to that reported for the *Adh1* and *ChiA* regions and higher than that for the *CHI* region. This would seem to support that it is this last region that presents reduced polymorphism. However, AFLP analysis might be overestimating nucleotide variation as it does not differentiate nucleotide and length variation (Miyashita *et al.* 1999), which does not seem to be negligible in *A. thaliana* as revealed by the sequencing surveys. Only the study of additional regions may allow discrimination between demographic and selective hypotheses.

We are grateful to Outi Savolainen, Peter van Tienderen, Jon Ågren, and Bernard Godelle for sharing *A. thaliana* seeds, to Serveis Científico-Tècnics from Universitat de Barcelona for sequencing facilities, and to Claus Vogl for helpful discussions. This study was supported by postdoctoral fellowships from the Environmental and Natural Resources Reseach Council of Finland (grant 41815) and the European Science Foundation to H.K. and 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 to M.A.

LITERATURE CITED

Abbot, R. J., and M. F. Gomez, 1989 Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. 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.

Brown, A. H. D., D. Zohary and E. Nevo, 1978 Outcrossing rates and heterozygosity in natural populations of *Hordeum spontaneum* koch in Israel. Heredity **41:** 49–62.

Chapple, C. C. S., B. W. Shirley, M. Zook, R. Hammerschmidt and S. C. Somerville, 1994 Secondary metabolism in Arabidopsis, pp. 989–1030 in *Arabidopsis*, edited by E. M. Meyerowitz and C. Somerville. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

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

Cummings, M. P., and M. T. Clegg, 1998 Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): 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.

Griffing, B., 1989 Genetic analysis of plant mixtures. Genetics **122:** 943–956.

Hanfstingl, U., A. Berry, E. A. Kellogg, J. T. Costa III, W. Rudiger *et al.*, 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.

Harpending, H., 1994 Signature of ancient population growth in a low resolution mitochondrial DNA mismatch distribution. Hum. Biol. **66:** 591–600.

Hedrick, P. W., 1980 Hitch-hiking: a comparison of linkage and partial selfing. Genetics **94:** 791–808.

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

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.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

Husband, B., and D. W. Schemske, 1996 Evolution of the magnitude and timing of inbreeding depression in plants. Evolution **50:** 54–70.

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.

Kawabe, A., and N. T. Miyashita, 1999 DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*. Genetics **153:** 1445–1553.

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

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

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.

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

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

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.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13:** 261–277.

Nordborg, M., B. Charlesworth and D. Charlesworth, 1996 Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proc. R. Soc. Lond. Ser. B **263:** 1033–1039.

Pollak, E., 1987 On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics **117:** 353–360.

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 *Arabidopsis*, edited by E. M. Meyerowitz

and C. Somerville. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
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.
Purugganan, M. D., and J. I. Suddith, 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.
Rausher, M., R. Miller and P. Tiffin, 1999 Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. Mol. Biol. Evol. **16:** 266–274.
Rogers, A. R., 1995 Genetic evidence for a pleistocene population. Evolution **49:** 608–615.
Rogers, A. R., and H. Harpending, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol. Biol. Evol. **9:** 552–569.
Rozas, J., and R. Rozas, 1997 DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.
Savolainen, O., C. H. Langley, B. P. Lazzarro and H. Freville, 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *A. thaliana*. Mol. Biol. Evol. **17:** 645–655.
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.
Slatkin, M., and R. R. Hudson, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555–562.
Sullivan, J. H., A. H. Teramura and L. H. Ziska, 1992 Variation in UV-B sensitivity in plants from a 3,000-m elevational gradient in Hawaii. Am. J. Bot. **79:** 737–743.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Ullrich, H., K. Lättig, A. Brennicke and V. Knoop, 1997 Mitochondrial DNA variations and nuclear RFLPs reflect different genetic similarities among 23 *Arabidopsis thaliana* ecotypes. Plant Mol. Biol. **33:** 37–45.
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., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.
Ziska, L. H., A. H. Teramura and J. H. Sullivan, 1992 Physiological sensitivity of plants along an elevational gradient to UV-B radiation. Am. J. Bot. **79:** 863–871.

Communicating editor: A. G. Clark