Haplotypic Divergence Coupled With Lack of Diversity at the *Arabidopsis thaliana* Alcohol Dehydrogenase Locus: Roles for Both Balancing and Directional Selection?

Ursula Hanfstingl[*,†] **Andrew Berry**,[‡] **Elizabeth A. Kellogg**,[‡] **James T. Costa III**,[‡] **Wolfhart Rüdiger**[†] **and Frederick M. Ausubel**[*]

[*]Department of Genetics, Harvard Medical School and Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, [‡]MCZ Laboratories, Harvard University, Cambridge, Massachusetts 02138, and [†]Institut für Botanik, 80638 München, Federal Republic of Germany

Manuscript received April 3, 1994
Accepted for publication July 20, 1994

ABSTRACT

We designate a region of the alcohol dehydrogenase locus (*Adh*) of the weedy crucifer, *Arabidopsis thaliana*, as "hypervariable" on the basis of a comparison of sequences from ecotypes Columbia and Landsberg. We found eight synonymous and two replacement mutations in the first 262 nucleotides of exon 4, and an additional two mutations in the contiguous region of intron 3. The rest of the sequence (2611 bp) has just three mutations, all of them confined to noncoding regions. Our survey of the hypervariable region among 37 ecotypes of *A. thaliana* revealed two predominant haplotypes, corresponding to the Columbia and Landsberg sequences. We identified five additional haplotypes and 4 additional segregating sites. The lack of haplotype diversity is presumably in part a function of low rates of recombination between haplotypes conferred by *A. thaliana*'s tendency to self-fertilize. However, an analysis in 32 ecotypes of 12 genome-wide polymorphic markers distinguishing Columbia and Landsberg ecotypes indicated levels of outcrossing sufficient at least to erode linkage disequilibrium between dispersed markers. We discuss possible evolutionary explanations for the coupled observation of marked divergence within the hypervariable region and a lack of haplotype diversity among ecotypes. The sequence of the region for closely related species argues against the possibility that one allele is the product of introgression. We note (1) that several loss of function mutations (both naturally and chemically induced) map to the hypervariable region, and (2) the presence of two amino acid replacement polymorphisms, one of which causes the mobility difference between the two major classes of *A. thaliana Adh* electrophoretic alleles. We argue that protein polymorphism in such a functionally significant part of the molecule may be subject to balancing selection. The observed pattern of extensive divergence between the alleles is consistent with this explanation because balancing selection on a particular site maintains linked neutral polymorphisms at intermediate frequencies.

STUDIES of the distribution of molecular polymorphism in natural populations provide a powerful means of identifying the evolutionary forces affecting a locus. The alternative, an experimental approach designed to detect fitness differences among alleles, may lack the necessary sensitivity: whereas selection coefficients as small as $1/N_e$(N_e is the effective population size of a species) can influence gene frequencies, experimental approaches are only able to detect fitness differences on the order of 1 or a few percent. Because polymorphism studies were pioneered in Drosophila and *Escherichia coli* [*e.g.*, LEWONTIN and HUBBY (1966) and MILKMAN (1973), respectively), nucleotide polymorphism studies of nuclear DNA have concentrated on these taxa. KREITMAN's (1983) studies of alcohol dehydrogenase (*Adh*) in Drosophila initiated this work and, with the introduction of statistical tools to distinguish between the actions of neutral and adaptive processes (HUDSON *et al.* 1987; MCDONALD and KREITMAN 1991), his work has remained the standard (KREITMAN and HUDSON 1991).

There are few such studies in plants [though this situation is changing fast: see THOMAS *et al.* (1993) and GAUT and CLEGG (1993a,b)]. Most plant work has focused on *Adh*, especially on the allozymes of one of the two maize *Adh* genes, *Adh1* (*e.g.*, DENNIS *et al.* 1984; SACHS *et al.* 1986). GAUT and CLEGG (1993a) compared the sequences of one *Adh* allele from each of two teosintes (*Zea luxurians* and *Zea diploperennis*), *Adh-1F*, *Adh-1S*, *Adh1-C^m* (OSTERMAN and DENNIS 1989) and three additional alleles of *Zea mays* and found no evidence of either balancing selection or a selective sweep at the locus. In their study of pearl millet (*Pennisetum glaucum*), GAUT and CLEGG (1993b) analyzed 20 alleles of *Adh1* from 10 individuals, 6 of which were wild-collected, and using the tests of TAJIMA (1989) and of FU and LI (1993), they were again unable to reject the hypothesis that all polymorphisms were neutral.

We describe here an analysis of naturally occurring sequence variation in *Adh* in the weedy crucifer, *Arabidopsis thaliana*, which, unlike many plants (GOTTLIEB

1982), has only one *Adh* locus. Under normal circumstances *Adh* is a non-essential gene in plants. However, *Adh* is strongly induced under anaerobic conditions, such as flooding (when it is essential for seedlings) (SACHS *et al.* 1980; DOLFERUS *et al.* 1985) and low temperature (JARILLO *et al.* 1993). Both types of stress stop mitochondrial ATP production and the cells shift to fermentation. It is expressed generally in seeds, young seedlings and pollen, but activity declines rapidly in green tissue of mature plants (DOLFERUS and JACOBS 1984).

DOLFERUS and JACOBS (1984) studied electrophoretic variation at *Adh* among 65 ecotypes (geographical isolates) of *A. thaliana* and classified 6.2% as slow, 46.1% as fast and 47.7% as super fast. They further characterized the different allozymes for pH optima, thermolability and, in the case of super fast, for substrate specificity, finding slight variation among the three allozymes. The gene for one of these allozymes, the super fast allele of the Landsberg *erecta* ecotype, has been sequenced (CHANG and MEYEROWITZ 1986) and is homologous to and structurally similar to *Adh* genes sequenced from other plants. DOLFERUS *et al.* (1990) sequenced two representatives of the Bensheim ecotype and found them to be identical to the Landsberg sequence. Because of this sequence identity, we cannot exclude the possibility that Bensheim is derived evolutionarily from the same stock as Landsberg. These cannot therefore be considered to be independent samples of diversity at this locus and, for this reason, we will refer to Landsberg as the single sequenced ecotype.

We have completely sequenced the *Adh* gene in the Columbia ecotype of *A. thaliana.* Comparing this to the published Landsberg ecotype sequence revealed a restricted cluster of mutations (or "hypervariable region") while the rest of the sequence was nearly identical between the two ecotypes. This clustered pattern of polymorphism is strikingly different from that seen in either Zea or Pennisetum. We therefore undertook a set of experiments to explore the nature and history of this variation. (1) We sequenced the hypervariable region in one plant of each of 37 additional ecotypes from diverse geographical locations. We found a total of only seven haplotypes (including the original Columbia and Landsberg ones), which fall into either Columbia- or Landsberg-like classes. (2) We gauged electrophoretic mobilities of allozymes for the same plants to determine the relationship between electrophoretic mobility and sequence variation in the hypervariable region. (3) We studied spontaneous and artificially induced deficiency mutants of *Adh* to determine the nature and position of the mutations. The molecular basis of the loss of ADH function has now been identified for a total of six mutants (DOLFERUS *et al.* 1990; this study) and we find that the three mutations that do not introduce a stop codon fall in the hypervariable region, suggesting that it encodes a functionally important part of the protein. (4) We sequenced the hypervariable region in several other members of the Brassicaceae and constructed gene trees which indicate that the allelic variation arose within *A. thaliana.* Further cladistic analysis of 12 polymorphic sites scattered throughout the genome for 34 ecotypes reveals no evidence of a historical or geographical dichotomy within the species. (5) Finally, we argue that these patterns are consistent with a model of balancing selection acting on the *Adh* polymorphism in *A. thaliana.*

MATERIALS AND METHODS

Plant material: Ecotypes were supplied by the Arabidopsis Information Service (AIS), Frankfurt, Federal Republic of Germany, and by the Nottingham Arabidopsis Stock Centre, University of Nottingham, United Kingdom (Table 1). Remaining ecotypes of *A. thaliana* (Lex, Mv, Phil, Ct, Fr and Jp) were isolated from the wild. The origins of the ecotype Bou (AUSUBEL laboratory) and of the isolate C24 (obtained from H. GOODMAN, Harvard Medical School, Boston, Massachusetts) could not be traced. We have doubts that Fl-3 (Finland) is a true diploid isolate of *A. thaliana*: it has a growth pattern and the large seeds characteristic of hybrids or polyploid *A. thaliana* derivatives. Nevertheless, Fl-3 yielded sequence and CAPS patterns (see below) typical of *A. thaliana.*

The ecotypes Landsberg *erecta* (Ler) and Columbia (Col-0) are standard laboratory strains, widely used for classical and molecular genetic analysis. Landsberg *erecta* and Columbia were originally isolated from a phenotypically heterogeneous seed stock (reviewed in RÉDEI 1992). Seeds from other species in the Arabidopsis group (PRICE *et al.* 1994), *Arabidopsis pumila* (synonym *Arabidopsis griffithiana*) ($2n$, $4n = 16, 32$), *Arabidopsis wallichii* ($2n = 14, 16, 18$) and *Arabidopsis suecica* ($2n = 26$) were acquired from AIS. The DNA for *Cardaminopsis arenosa* ($2n$, $4n = 16, 32$) was kindly provided by R. PRICE, University of Georgia, Athens, as were seeds from *Halimolobos diffusa* var. *jaegeri.* ($2n = 16$) (D. W. TAYLOR s.n., 10 September 1988). *Capsella bursa-pastoris* ($2n$, $4n = 16, 32$) was isolated from the wild. Arabidopsis plants were grown in the green house in a soil-mix medium under fluorescent lights (16-hr day).

Sequence analysis of the Columbia ecotype *Adh* gene: A Columbia genomic DNA library in lambda GEM 11 (Promega) was kindly provided by J. T. MULLIGAN and R. W. DAVIS, Stanford University. Probing with a 1.8-kb *Bsa*AI fragment of plasmid clone pMY417 (constructed by M. YANOFSKY) containing the transcribed region of the Landsberg *erecta Adh* gene resulted in the isolation of a phage clone, λUH17, with a complete Columbia *Adh* coding sequence [as determined by restriction mapping and comparison to the Landsberg clone λfAt3102 (CHANG and MEYEROWITZ 1986)]. We sequenced the entire coding region plus all but 272 bp of the flanking region sequenced in Landsberg *erecta Adh* (CHANG and MEYEROWITZ 1986), using a polymerase chain reaction (PCR)-based approach consisting of an initial amplification from phage template followed by cycle-sequencing (Promega) of the purified fragment [for sequencing strategy, see Figure 1B and HANFSTINGL (1994)]. The sequence was verified by sequencing the complementary strand or by repeatedly sequencing the same segment; in all cases where the sequence differed from the published Landsberg sequence, both strands were sequenced.

Sequence analysis of different ecotypes, related species and Adh-deficient mutants: To compare the 200-bp region found to be hypervariable between Columbia and Landsberg with the sequence of the same region in other ecotypes, genomic DNA was isolated from a single plant of each ecotype (DELLAPORTA

TABLE 1

Hypervariable region sequence for 39 ecotypes

	Variable sites														Allozyme type
Position[a]	1641	1644	1649	1657	1659	1701	1713	1734	1761	1779	1800	1821	1835	1839	
Polymorphic sites between Ler and Col	1	2			3	4	5	6	7	8	9	10			
Triplet position	3	3	2	1	3	3	3	3	3	3	3	3	2	3	
Amino acid change	No	Asp/Glu	Arg/Pro	Gln/Lys	Gln/His	No	No	No	No	No	No	No	Pro/Arg	Asp/Glu	
Side chain difference		Both negatively charged	Positively charged → uncharged	Uncharged → pos. charged	Uncharged → pos. charged or uncharged								Neutral → positively charged	Both negatively charged	
Ler haplotype:	A	T	G	C	G	A	T	T	C	G	T	C	C	T	Super fast[b]
1. Landsberg-type ecotypes															
Ag-0 France	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Be-0 Germany	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast[b]
Bur-0 Ireland	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Bou Unknown	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Ct USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Fr USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Ge-0 Swiss	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Gre-0 USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Kin-0 USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Lex USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Mv USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Nd-0 Germany	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Phil USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Tul-0 USA	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast
Zü-0 Swiss	—	—	—	—	—	—	—	—	—	—	—	—	—	—	Super fast[c]
Bla-1 Spain	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	—	—	—	—	—	Super fast[d]
Bus-0 Norway	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	—	—	—	—	—	Super fast
C24 unknown	—	—	—	$\boxed{\text{A}}$	—	—	$\boxed{\text{G}}$	—	—	—	—	—	—	—	Fast
Est-0 Estland	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	—	—	—	—	—	Super fast[b]
Fe-1 Germany	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	—	—	—	—	—	Super fast[b]
Jp USA	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	—	—	—	—	—	Super fast
N916 Tadjikistan	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	$\boxed{\text{T}}$	—	—	—	$\boxed{\text{A}}$	Super fast[e]
Rsch-0 Russia	—	—	—	—	—	—	$\boxed{\text{G}}$	—	—	$\boxed{\text{T}}$	—	—	—	$\boxed{\text{A}}$	Super fast[c]
2. Columbia-type ecotypes															
Col-0	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast[c]
Aa-0 Germany	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Ba-1 UK	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Co-4 Portugal	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Cvi-0 Capv. Is.	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Esc-0 Spain	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
F1-3A Finland	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
F1-3B Finland	G	G	$\boxed{\text{C}}$	—	C	C	G	A	T	$\boxed{\text{—}}$	A	T	—	—	Fast[e]
Hau-0 Denmark	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Kas-1 India	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
L1-0 Spain	G	G	—	—	C	C	G	A	T	T	A	T	$\boxed{\text{G}}$	—	Slow[d]
Mh-0 Poland	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Ms-0 Russia	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
No-0 Germany	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Po-1 Germany	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast
Ws-0 Ukraine	G	G	—	—	C	C	G	A	T	T	A	T	—	—	Fast

Dashes denote concordance with the published Landsberg sequence (CHANG and MEYEROWITZ 1986); the substituted base is given for deviant sites. Deviations from "type" haplotypes (either Columbia or Landsberg haplotypes) are in boxes.

[a] Numbering after CHANG and MEYEROWITZ (1986).

[b,c,d] The allozyme tested as type super fast ([b]), fast ([c]) or slow ([d]) in DOLFERUS and JACOBS (1984).

[e] Allozyme assayed for plants other than those sequenced.

TABLE 2

CAPS analysis of ecotypes

	Chromosome locus											
	I	*II*	*II*	*III*	*III*	*III*	*IV*	*IV*	*IV*	*IV*	*V*	*V*
Ecotype	*ADH*	*GPA1*	*m429*	*GapC*	*GAPA*	*GL1*	*GA1*	*AG*	*PG11*	*DHS1*	*DFR*	*LFY3*
Col-0	C	C	C	C	C	C	C	C	C	C	C	C
Ler	L	L	L	L	L	L	L	L	L	L	L	L
Aa-0	C	L	L	L	L	L	C	L	L	L	C	L
Ag-0	L	C	C	L	L	×	ND	L	ND	L	L	L
Ba-1	C	C	L	C	ND	C	ND	L	L	L	C	L
Be-0	L	C	L	C	L	L	L	L	ND	L	C	L
Bus-0	L	C	L	C	L	×	ND	L	L	L	C	L
Bou	L	C	L	L	L	C	C	L	L	L	C	L
Bur-0	L	ND	L	C	L	L	C	L	L	C	C	L
C24	L	C	L	L	L	×	L	L	ND	L	C	L
Ct	L	C	L	L	L	L	C	L	C	L	C	L
Co-4	C	C	C	L	L	L	C	L	L	L	C	L
Cvi-0	C	C	L	C	L	L	C	L	L	L	L	L
Est-0	L	ND	L	L	L	L	C	L	C	L	C	L
Fl-3	C	C	L	C	L	L	C	L	L	L	C	ND
Fr	L	C	L	L	L	×	C	L	ND	L	L	L
Ge-0	L	C	L	C	L	×	C	L	L	L	C	L
Gre-0	L	ND	L	L	ND	L	C	L	ND	L	C	ND
Hau-0	C	C	L	C	L	×	C	C	ND	C	C	L
Jp	L	C	L	L	L	L	C	L	L	ND	C	L
Kas-1	C	C	L	C	L	×	C	L	L	L	C	ND
Kin-0	L	ND	L	L	ND	L	C	L	ND	ND	C	L
Lex	L	C	L	L	L	×	C	L	L	L	L	L
L1-0	C	C	L	L	L	×	C	L	L	L	L	L
Mv	L	C	L	ND	L	L	C	L	C	L	C	L
Ms-0	C	C	C	C	L	L	C	L	L	L	C	L
Mh-0	C	ND	L	C	L	L	C	L	ND	ND	C	L
Nd-0	L	C	C	C	L	L	C	L	ND	L	ND	L
No-0	C	C	L	C	L	L	C	L	ND	L	C	L
Phil	L	ND	L	L	ND	C	C	ND	ND	ND	C	L
Po-1	C	C	C	C	L	L	C	L	L	L	L	L
Sei-0	C	C	L	C	L	L	L	L	ND	L	ND	L
Tul-0	L	ND	L	L	ND	L	C	ND	ND	L	C	ND
Ws-0	C	C	L	L	L	L	C	L	C	L	C	L

C denotes a Columbia genotype; L a Landsberg genotype; × a genotype found in neither; ND that data was unobtainable.

et al. 1983). A 1.6-kb fragment encompassing the hypervariable region [using primers ADH4^{fwd} (1564–1588) and ADH4^{rev} (3148–3125) [numbering after CHANG and MEYEROWITZ (1986)] (Figure 1B)], was amplified from each genomic DNA. Cycle-sequencing was used to determine the hypervariable region sequence using sequencing primers ADH4^{fwd} and ADH401^{rev} (1894–1870). Except for ecotype Fl-3 for which sequences of four separate plants were compared, each sequence was determined from a single plant. Because ecotypes are inbred lines, heterozygosity was not a problem.

We sequenced the same way the hypervariable region of species closely related to *A. thaliana.* In addition to the primers ADH4^{fwd} and ADH401^{rev}, we used the forward primer ADH402 (1765–1786, T at position 1779) for all species and, in *Halimolobos diffusa,* a reverse primer, ADH403 (1869–1849, G, A, T at positions 1863, 1857, 1854, respectively) was used instead of ADH401.

Genomic DNA was also prepared from the ADH deficient mutants and their hypervariable regions sequenced as for the ecotypes. For those mutants lacking a loss-of-function mutation in the hypervariable region, we made no systematic attempt to sequence the entire locus but sequenced instead a number of haphazardly selected coding regions using appropriate amplification and sequencing primers.

PCR-amplification/restriction digest procedure: To study the distribution in each ecotype of other polymorphic markers scattered throughout the genome, we carried out a PCR-amplification/restriction digest procedure (KONIECZNY and AUSUBEL 1993), using 12 of the 18 sets of described primers (Table 2), giving us markers for all 5 chromosomes. We used the same DNA preparations as were used for sequencing *Adh.*

Callus induction and isozyme analysis: Callus cultures were started from sterilized seeds on 0.8% tissue culture agar plates (Hazelton) containing MS salts and vitamins (GIBCO), 20 g sucrose and 0.05 mg/liter kinetin and 1 mg/liter 2,4-dichlorophenoxyacetic acid (2,4-D). When possible, we started callus cultures from seeds collected from the same plant as was used for sequencing. We performed starch gel electrophoresis on material extracted according to DOLFERUS and JACOBS (1984) and stained the gels for ADH activity as described in JACOBS *et al.* (1988).

Production and analysis of *Adh* mutants

Seed material for the selection of natural Adh mutants: In May 1989, 10 g of Columbia wild-type seeds were sown on a 3 m × 2 m plot of former grassland near Bloomington, Indiana. The plants were left unattended, and the seeds were harvested at the end of the summer. The first batch reseeded itself and gave rise to an unknown number of generations, until, in September, all plants were harvested and dried, yielding 170 g of seed.

Mutagenized seedstock for artificial Adh mutants: Separate mutagenesis experiments on Col, Ler, Mh-0 and Sei-0 ecotypes were carried out using (1) ethyl methanesulfonate (EMS) (performed by N. OLSZEWSKI); (2) 1,2:3,4-diepoxybutane

(DEB) (G. STORZ); and (3) γ-rays (D. VOYTAS; B. HAUGE) [see HANFSTINGL (1994) for details].

ADH selection in seeds: Seeds from the natural seedstock and from the M_2 generation of mutagenized populations were screened for loss of ADH function by treating them with allyl alcohol (JACOBS *et al.* 1988). For natural mutants, each of four batches with 10 g of seeds each was suspended in 1 liter of water then incubated in 50 mM allyl alcohol before being washed in water as described by JACOBS *et al.* (1988). The seeds were then mixed with 0.1% agarose to facilitate even spreading and pipetted onto soil in the greenhouse. Two flats (0.3 m × 0.6 m) were used for 10 g of seeds. As a control, 30 seeds of a line which carries the ADH-deficient locus from the EMS induced mutant *Adh102* and the visible mutation *glabra* (allele *gl1*; trichomes (hairs) absent on leaves and stems) from a linkage tester line (cross performed by D. VOYTAS) were mixed with each 10 g seed batch. They were recovered as *glabra* plants that survive allyl selection. When seeds from individual plants or small mutagenized populations were screened, they were incubated in dialysis bags. All mutants were reselected to make sure that the initial selection was uniformly effective. To ensure that the material used for DNA preparations was derived from a pure seedstock, we used plants grown from seeds of a single plant that were tested for ADH under sterile conditions and grown on agar plates (0.8%; MS salts, vitamins and 2% sucrose) before they were transferred to soil.

Genetic analysis and mapping of ADH-deficient mutants: The allyl alcohol screen allows the survival of seedlings with an enzymatically nonfunctional ADH dimer. We performed three genetic crosses with 15 ADH-deficient mutants (14 naturally and 1 DEB-induced). The test for segregation frequencies after allyl alcohol incubation was done under sterile conditions on agar plates (see above).

To determine whether our mutations were in *cis* or *trans*, a complementation cross was performed with the *Adh* mutant R002 as acceptor of the pollen. Since R002 has a known mutation in the coding sequence (DOLFERUS *et al.* 1990), a *trans* mutation among the ADH mutants studied can be complemented and give rise to a single functional *Adh* allele in all individuals of the F1 generation. In a heterozygote individual with one wild-type allele, however, enough functional ADH dimers are produced to convert allyl alcohol into toxic aldehyde which kills the seedling. Seeds of an uncomplemented cross survive. The allyl test was done separately with at least three siliques.

The same 15 mutants were tested for recessive or dominant behavior in the F_1 generation of a cross to Landsberg *erecta*. In the case of a recessive mutation, an allyl alcohol test will result in the death of all seedlings since they are heterozygotes with one wild type allele and this single allele is sufficient to kill the seedling. In the case of a dominant mutation, either all seeds or 50% will germinate depending whether the mutant plant used in the cross was homozygous or heterozygous. At least three siliques from each plant were tested. Since the *erecta* mutation is recessive and mutant pollen was applied to a Landsberg *erecta* stamen, a successful cross can be tested by means of the complemented phenotype (LANDSBERG wild type) of the F1 generation.

To determine whether the mutant loci in all 15 mutants map close to *Adh*, which is located on chromosome *1*, a cross between the 15 mutants and a chromosome *1* markerline (*ch1*, chlorina; *ap1*, apetala; *gl2*, glabra) was performed. The F_1 generation was checked for a successful cross by phenotype complementation of the *erecta* phenotype of the markerline (pollen receiver). For the linkage test, the total seed yield of one F_1 plant (1000–1500 seeds) was first screened in an allyl alcohol test to select for the surviving one fourth of the plants homozygous for ADH deficiency. These seedlings were transplanted into soil and scored for recombination frequency between the phenotypic markers. *Adh* maps to chromosome *1* (CHANG *et al.* 1988) close to the *gl2* locus and a mutation at or near the *Adh* locus will lead to a very low frequency of recombinants between a *glabra* phenotype and ADH deficiency.

Molecular analysis of ADH mutants: For the detection of length variants in the *Adh* gene of the natural mutants and to test for mutations resulting from the insertion of transposable elements, genomic DNA of the mutants was cut with *Eco*RI and *Hin*dIII, run on a 0.8% gel and transferred to a filter (AUSUBEL *et al.* 1989). The blot was probed with the *Sac*I-*Bam*HI insert of plasmid pMY417 and with the pooled *Eco*RI fragments of the insert in λfAt3102 (CHANG and MEYEROWITZ 1986) to check for chromosome rearrangements in a wider area. The Landsberg *Adh* mutants were screened for length mutations by amplification of six contiguous pieces of the *Adh* gene covering the published Landsberg sequence. The products were visualized on a 1.5% agarose gel.

For RNA analysis and anaerobic *ADH* induction, *Adh* mRNA induction through anaerobiosis was performed as described in CHANG and MEYEROWITZ (1986) with slight modifications. The sterilized seeds were kept for 2 days at 4°, and the seedlings were harvested after 4 days. The anaerobic induction was carried out for 4 hr. Total RNA was isolated from frozen material (a) of anaerobically induced and (b) untreated seedlings (as control) and Northern blotted. The blots were probed with the *Sac*I-*Bam*HI insert of plasmid pMY417 or PCR products of λUH17. To standardize for loading differences, the blots were stripped in 2 mM Tris, pH 8.0, 2 mM EDTA at 70° and reprobed with a 6-kb *Eco*RI fragment containing the *Ab140* gene (*Cab1*) of the 3 CAB (light-harvesting chlorophyll *a/b* protein) encoding genes described by LEUTWILER *et al.* (1986).

Analysis: DNA analysis was carried out with programs in the GCG package (Madison, Wisconsin; DEVEREUX *et al.* 1984). Phylogenetic analysis was carried out using PAUP 3.1.1 (SWOFFORD 1993). In addition to standard population tests, we applied a test devised by R. C. LEWONTIN (personal communication) for detecting clustering of polymorphic sites. This is based on the data's fit to the broken stick distribution (MACARTHUR 1957), which describes the distribution of interval lengths between points placed at random along a line. We used two test statistics, maximum interval length and interval length variance, and determined their significance relative to a null hypothesis of a uniform distribution of points (polymorphic sites) by means of Monte Carlo simulation. Tables of critical values of these statistics are available from R. C. LEWONTIN (Harvard University, Cambridge, Massachusetts).

RESULTS

Nucleotide polymorphism between the *A. thaliana* ecotypes Columbia and Landsberg: *Adh* is a single copy gene in *A. thaliana* and was first sequenced and characterized in the ecotype Landsberg by CHANG and MEYEROWITZ (1986). Our 2923-bp sequence of the wild-type Columbia *Adh* gene overlaps almost completely with that published for the Landsberg ecotype (3195 bp) and includes the entire coding sequence. The Columbia sequence has been deposited in the EMBL/GenBank/DDBJ Nucleotide Sequence Databases under the accession no. X77943. The sequence incorporates an open reading frame of 1137 nucleotides (379 amino acids), 567 nucleotides in 6 introns, 835 nucleotides upstream

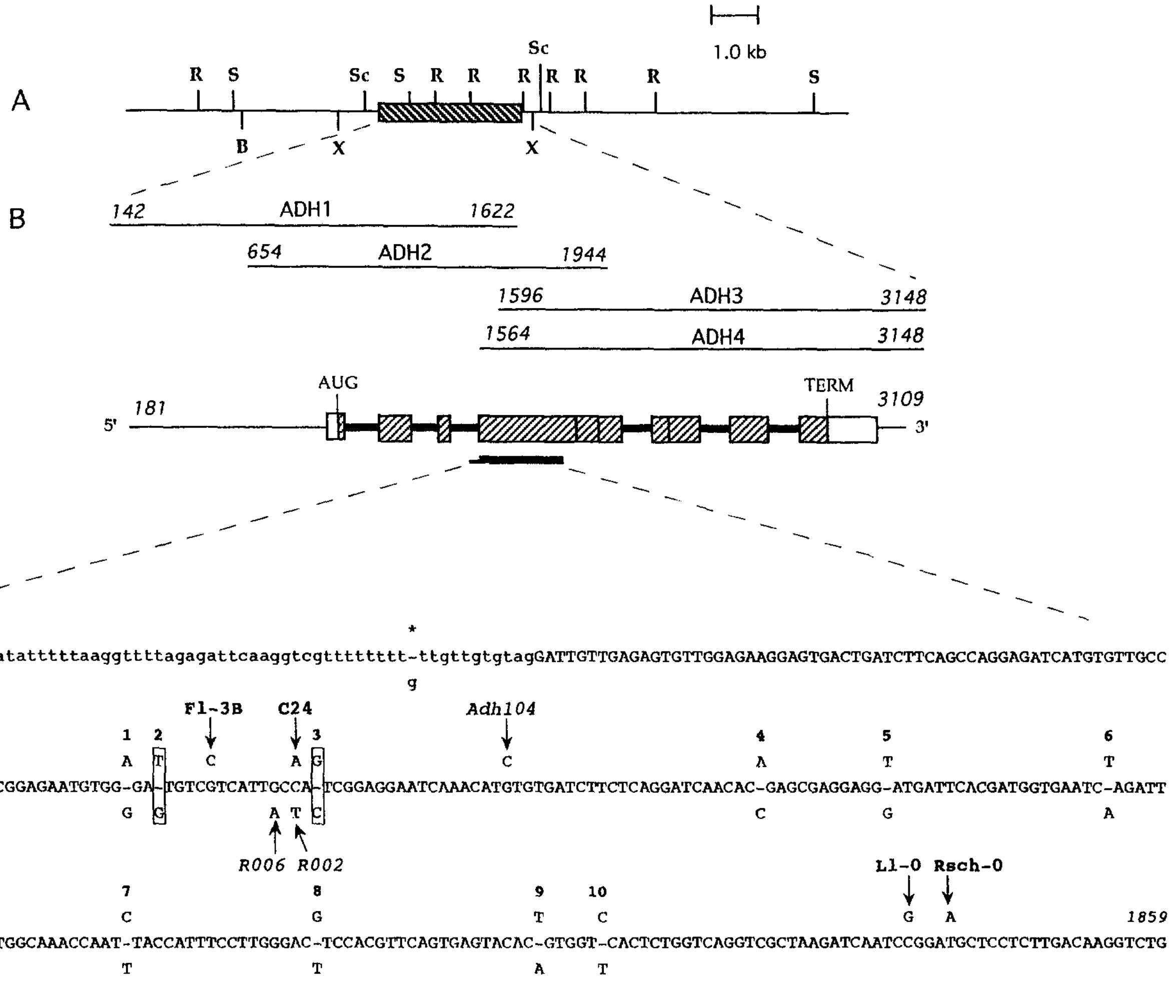
A
R S Sc S R R R Sc R R R S
B X X
1.0 kb
B
142 ADH1 1622
654 ADH2 1944
1596 ADH3 3148
1564 ADH4 3148
AUG TERM
5' 181 3109 3'
C
1501 g *
tcgtttt-gatattttaaggttttagagattcaaggtcgttttttt-ttgttgtgtagGATTGTTGAGAGTGTTGGAGAAGGAGTGACTGATCTTCAGCCAGGAGATCATGTGTTGCC
* g
Adh103 F1-3B C24 Adh104
1 2 3 4 5 6
1620 A A T C A G C A T T
GATCTTTACCGGAGAATGTGG-GA-TGTCGTCATTGCCA-TCGGAGGAATCAAACATGTGTGATCTTCTCAGGATCAACAC-GAGCGAGGAGG-ATGATTCACGATGGTGAATC-AGATT
G G A T C C G A
R006 R002
L1-0 Rsch-0
7 8 9 10
1740 C G T C G A 1859
CTCCATTAATGGCAAACCAAT-TACCATTTCCTTGGGAC-TCCACGTTCAGTGAGTACAC-GTGGT-CACTCTGGTCAGGTCGCTAAGATCAATCCGGATGCTCCTCTTGACAAGGTCTG
T T A T

FIGURE 1.—(A) Restriction map of Columbia *Adh* λ clones using *Eco*RI (R), *Sal*I (S), *Sac*I (Sc), *Bam*HI (B), *Xba*I (X). A shaded box indicates the sequenced region. (B) PCR sequencing strategy for analysis of the Columbia *Adh* gene and of the hypervariable region in ecotypes, ADH-deficient mutants and related species. The PCR amplified fragments ADH1, ADH2 and ADH3 served as templates for PCR sequencing of the Columbia *Adh* gene. Fragment ADH4 was used as template to sequence the hypervariable region in ecotypes, ADH-deficient mutants and related species. The sequenced region is shown in a schematic drawing (adapted from CHANG and MEYEROWITZ (1986)); shaded boxes represent exons, white boxes 5'- and 3'-untranslated regions, and thick lines introns. They are drawn in their relative proportions. A black bar below the drawing indicates the hypervariable region spanning parts of intron 3 and exon 4. (C) Distribution of mutations found in the hypervariable region of ecotypes and ADH-deficient mutants. Intron 3 (lowercase) and exon 4 (uppercase) are shown between nucleotides 1501 and 1859 (numbering after CHANG and MEYEROWITZ 1986). Polymorphic sites between Columbia and Landsberg are shown as dashes; the Landsberg variant is shown above the sequence and the Columbia one below it. The polymorphic sites in the coding region are numbered from 1–10 and the sites where amino acid changes occur between Columbia and Landsberg in boxes. Asterisks denote deletions. Arrows point to additional changes in ADH-deficient mutants (italic) and in ecotypes (bold).

of the translation initiation point and 384 nucleotides beyond the stop codon at the 3' end, as illustrated in Figure 1B. Twelve of the 15 differences between the Columbia and Landsberg sequences occur in a cluster spread over 299 bp of the 2923 bp of compared sequence. We designate this region "hypervariable"; its location and sequence are shown in Figure 1, B and C. Note that 10 mutations are found in exon 4 (referred to subsequently as variable sites 1–10); these are spread over only 180 nucleotides and include 2 amino acid replacements. In addition to the hypervariable cluster, the two sequences differ in three additional sites: two deletions in Columbia, one of 5 bp (nucleotides 249–253; numbering follows Landsberg sequence) and one of one base pair (nucleotide 215), are close together at the 5' end of the 835-bp region upstream of the start codon ATG. The third difference is a mutation (nucleotide 1114) in intron 1. The two λ clones containing the complete Columbia *Adh* gene were mapped with the enzymes *Eco*RI, *Bam*HI, *Sal*I, *Sac* I, and *Xba*I over a region of 16 kb (Figure 1A) and found to be identical to Landsberg *Adh* (CHANG and MEYEROWITZ 1986) except within the hypervariable region. Figure 2 summarizes the pattern of polymorphism between the Columbia and Landsberg *Adh* genes: they are virtually identical except in the hypervariable region where they are highly divergent. The broken stick test described above demonstrates this distribution of polymorphic sites to be

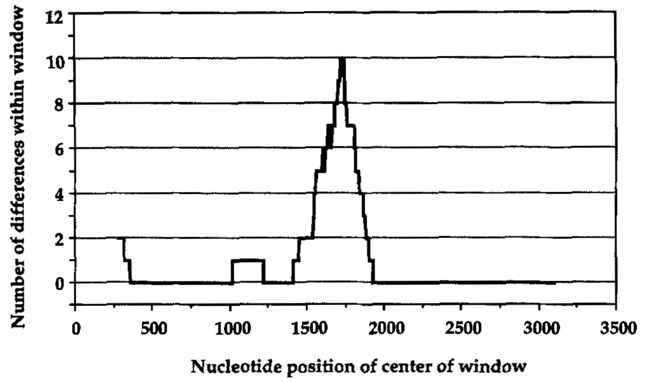

FIGURE 2.—Sliding window plot of divergence between the Columbia and Landsberg haplotypes. Window size 200; window "stepped" 5′-3′ in 5-bp increments.

significantly clustered: $P < 0.01$ for both longest interval length and interval length variance.

Polymorphism in ecotypes: *A. thaliana* ecotypes are widespread geographically and often differ phenotypically and genetically. We investigated patterns of polymorphism in the hypervariable region of the *Adh* gene in an additional 37 geographically dispersed ecotypes. All plants sequenced were homozygous for the hypervariable region. A fragment of the *Adh* gene, containing the 10 translated nucleotide substitutions at the beginning of exon 4, was sequenced for each ecotype and the results are presented in Table 1. All haplotypes were either identical to the Columbia or Landsberg sequences or differed by no more than three mutations from either Columbia or Landsberg; all haplotypes may therefore be classified as Landsberg- or Columbia-like.

There are two forms of deviation from the strict Columbia or Landsberg haplotype. First, we find ecotypes with, for example, "Landsberg" nucleotides at 9 out of 10 polymorphic sites and the "Columbia" nucleotide at the other one. We describe these as ecotypes with a 9L:1C haplotype. These base-pair switches between haplotypes are all silent and are found either as 9:1 or 8:2 chimeras (see Table 1). Although these may conceivably represent instances of parallel mutation, we take them, on grounds of parsimony, either as the products of recombinational exchange of mutations between Columbia and Landsberg haplotypes or as residual "ancestral" haplotypes. The second kind of deviation is an amino acid mutation specific to that ecotype at a position where Columbia and Landsberg sequences are identical; there are four such mutations occurring in five ecotypes. In some ecotypes both kinds of deviations occur. For example, ecotype C24, 9L:1C, has a nucleotide change at position 1657 which changes glutamine to lysine, and the ecotypes N916 and Rsch-0 with the same 8L:2C haplotype share a mutation at nucleotide 1839, which replaces aspartic acid with glutamic acid.

Because ecotype Fl-3's growth habit, large seeds, big flowers and higher chromosome number (E. SCHOTT; personal communication) suggest it to be a hybrid, we sequenced four individuals of this ecotype to look for evidence of polyploidy. We found 3 Fl-3 plants to be homozygous for the Columbia haplotype (Fl-3A) and one plant (Fl-3B) to be homozygous for a 9C:1L haplotype with a substitution at position 1649, introducing a proline for an arginine (Table 1). Variation thus continues to segregate in this inbred line but our sequence analysis revealed no evidence of polyploidy.

Allozyme characterization and correlation to the hypervariable region sequence: *Adh* is constitutively expressed in callus on media with the artificial auxin 2,4-D (DOLFERUS *et al.* 1985). We grew and electrophoresed callus from seeds harvested from the plant that provided material for sequencing (Table 1). In a few cases the sequenced plant had been lost and it was necessary to use seeds from another individual. Note that three of the nine ecotypes assayed for *Adh* allozyme in both this study and that of DOLFERUS and JACOBS (1984) yield different electromorphs in each study. We presume that this is due to contamination and/or confusion of stocks. Unfortunately, because the original material was lost, we cannot assign with confidence an electromorph type to Fl-3B, which has a charge-changing amino acid substitution which might accordingly affect its mobility. Because the amino acid sequences of Columbia (fast) and Landsberg (super fast) differ at only two positions, one or both of those substitutions must be responsible for the differences in their mobilities. The change at position 2 results in no charge change; however, at position 3, the substitution of histidine (which may be positively charged) in Columbia for the neutral glutamine of Landsberg can confer a charge difference; the more positively charged protein migrates as expected less rapidly to the anode. The protonation of histidine is sensitive to changes of pH within the physiological range and we therefore predict that the difference in mobility between Columbia and Landsberg allozymes would be eliminated by running the gels at a higher pH (our running buffer was at pH 6.3). Because we lack complete sequence information for other ecotypes, we cannot unequivocally attribute changes in mobility to observed amino acid differences. However, it is worth noting that the most positively charged protein, Ll-0, which has a substitution of positively charged arginine for neutral proline, on the basis of its hypervariable region sequence, is also the slowest anode-migrating protein in the study. Also, the substitution of lysine (positively charged) for glutamine (neutral) in ecotype C24 apparently retards the allozyme from super fast to fast. The N916/Rsch-0 haplotype carries a same-charge substitution and we detected no change in mobility.

Although this analysis is based on only partial sequences, we note that we can explain all observed mobilities solely on the basis of the amino acid sequence of the hypervariable region. Thus, our allozyme analysis, which tests for charge differences in the complete ADH

TABLE 3

Analysis of mutations in ADH deficient mutants

	Adh mutations					
	In exon 2		In the hypervariable region			
Mutant name	*Adh101*	*Adh102*	*Adh103*	R006[a]	Roo2[a]	*Adh104*
Nucleotide position[b]	1190	1199	1626	1655	1657	1679
Mutation	point del.	G → A	T → A	G → A	C → T	G → C
Triplet-position	2	2	3	2	1	2
Amino acid change	Ala	Trp	Phe	Cys	Gln	Cys
	↓	↓	↓	↓	↓	↓
	Val, then Stop	Stop	Leu	Tyr	Stop	Ser
Background	Col-0	Col-0	Mh-0	Be-0	Be-0	Col-0
Mutagen	Natural	EMS	γ-Rays	EMS	EMS	Natural
RNA (%)[c]	<10	<50	<50	90	5	90
Protein (%)	0[d]	0[d]	10[d]	60[e]	10[e]	0[d]

[a] Data from DOLFERUS *et al.* (1990) and JACOBS *et al.* (1987).
[b] Numbering of nucleotide positions follows CHANG and MEYEROWITZ (1986).
[c] Values given in percent of wild-type level RNA after anaerobic induction.
[d] Percent of wild-type level of staining in allozyme gels.
[e] Percent of ADH-like CRM (cross reacting material; JACOBS *et al.* 1987); values not strictly comparable with [d].

protein, may indicate that, outside the hypervariable region, if any amino acid mutations occur, they either involve same-charge amino acid substitutions or multiple substitutions resulting in no net charge change. However, there may be a large amount of cryptic protein-mobility variation which our methods of electrophoresis have failed to pick up. Except in the case of the Columbia and Landsberg difference, we therefore make no strong claim to have assayed amino acid sequence variation for the entire protein.

Mutations in natural and artificial *Adh* mutants are located in the polymorphic region: The screen for *Adh* deficient mutants acts at the germinating seed stage such that only plants with a non-functioning enzyme survive. It is thus efficient in isolating mutants from several hundred thousand individuals in one experiment. We used this assay to screen a total of 1,800,000 seeds from a pool of 9,500,000 seeds of a Columbia population. Genetic analysis revealed all of the 14 natural *Adh* mutants isolated to be recessive, in *cis* and at least linked to the *Adh* locus on chromosome one. RNA analysis of anaerobically induced seedlings determined that 13 isolates expressed a very low level of *Adh* mRNA while one mutant had near wild-type level *Adh* mRNA (Table 3).

To date, only two *Adh* mutants, R002 and R006, both EMS-induced and in the Bensheim background, have been fully sequenced (DOLFERUS *et al.* 1990). Because both had mutations in the "hypervariable" region of the *Adh* gene (Table 3), we sequenced this region in all natural *Adh* mutants, and found that one of our natural mutants, *Adh104*, has a mutation in the hypervariable region. *Adh104* has near wild-type levels of *Adh* mRNA, and has serine substituted for one of the 4 cysteines that coordinate the zinc atom apparently responsible for the structural stability of the enzyme (Table 3). However, the larger group of *Adh* deficient mutants with low mRNA had the expected Columbia sequence for the hypervariable region. Further sequencing determined them to be siblings (*Adh101*) since all had the same point deletion at nucleotide 1190 of residue 13 in the second exon. The induced frameshift leads to a UGA stop codon 10 amino acids further on (Table 3). As only two mutants were recovered from 1,800,000 seeds, we can make no estimate of the spontaneous mutation rate of the *Adh* locus of *A. thaliana.*

We extended the sequence analysis of the polymorphic region to artificially induced *Adh* mutants, of which 3 were induced by EMS, 1 by DEB and 14 by γ-rays. Except in the case of the hypervariable region, which was sequenced for all mutants, we made no systematic attempt to identify the molecular lesions causing loss of ADH function. However, sequence of the hypervariable region showed one γ-ray-induced mutant, *Adh 103*, to have an amino acid substitution in the hypervariable region (Table 3). Additionally, an EMS induced mutant, *Adh 102*, had a substitution in exon 2 that introduces a stop codon (Table 3). Protein and RNA titer assay results for all mutants whose mutations were identified are given in Table 3. This shows there to be a total of four loss-of-function mutations that map to the hypervariable region. The disproportionately high concentration of such mutations (including three substitutions and one stop codon) in this region is in part a reflection of the intensity with which each part of the gene was searched for mutations. From our point of view, however, it is important to note that the enzyme is highly sensitive to amino acid changes in this region (all three of the loss-of-function mutations that do not introduce a stop codon are located here), suggesting that it encodes a functionally important part of the protein.

Do we see evidence of the Columbia/Landsberg split throughout the genome?: Our data for the sequence of the hypervariable region for ecotypes reveals two distinct classes of haplotypes, one Columbia-like and the other Landsberg-like. We investigated whether this Columbia/Landsberg dichotomy is the product of historical factors such as an ancient geographical split in the population. Note that such an historical scenario predicts uniform divergence throughout the genome, so, in such a case, we may see two divergent lineages at all loci free to vary, possibly with genome-wide linkage disequilibrium among these variants (if admixture of lineages was recent and/or rates of recombination are low). To study genome-wide divergence and linkage disequilibrium we used a method developed by KONIECZNY and AUSUBEL (1993) to facilitate the mapping of mutations to chromosomes in recombinants between Columbia and Landsberg ecotypes. This method, named CAPS for *c*leaved *a*mplified *p*olymorphic *s*equences, is based on the PCR amplification of DNA fragments followed by digestion at polymorphic restriction enzyme sites. In this case, we used CAPS corresponding to known polymorphisms between Columbia and Landsberg ecotypes that are distributed across all five chromosomes. If all ecotypes are derived from two lineages of which Landsberg and Columbia are representatives, and linkage disequilibrium is high, phylogenetic analysis would reveal the restriction sites as clustering in two main groups. Table 2 shows the digest pattern for most of the ecotypes (although this was not exactly the same set of ecotypes as was used for the sequence analysis, these nevertheless represent a reasonable sample of species-wide diversity). Note that our failure to amplify the appropriate regions in some cases (*i.e.*, where "ND" is entered in Table 2) may reflect sequence divergence at the primer sites. A cladistic analysis using the restriction sites as characters found more than 1000 trees of length 35 steps, consistency index (CI) = 0.333 and retention index (RI) = 0.645. The strict consensus of these trees, with branch lengths proportional to number of changes, is shown in Figure 3a. The lack of hierarchical structure indicates that there is little correlation (linkage) of the restriction site characters. An Adams consensus of the same 1000 trees (Figure 3b) pinpoints ecotypes whose position on the tree is unstable by moving them to the lowest node in common among their various placements. For example the placement of Bus-0, Bur-0, Kas-1, Ge-0 and Mh-0 at the bottom of the tree indicates that their placements vary widely among the 1000 trees; they are therefore among the taxa responsible for the complete lack of hierarchy in the strict consensus tree. The smaller groups (clades) in the Adams consensus tree are those that appear in all trees, although there may be additional taxa included in some of the trees. Mapping the *Adh* haplotypes onto the Adams consensus tree shows that (1) there are not two discrete lineages in the species and (2) the Landsberg and Columbia haplotypes are intermixed throughout the tree. The analysis therefore suggests that the two *Adh* haplotypes are not the product of an ancient split in the species, though, because the effect of any such split would be obscured by recombination between recently admixed lineages, this approach offers only a very weak test of such a historical hypothesis.

In addition to phylogenetic analysis, we looked for evidence of pairwise linkage disequilibrium between CAPS sites. Because the CAPS sites were designated as polymorphic on the basis of differences between the Col and Ler ecotypes, all the variants in these ecotypes are by definition in linkage disequilibrium with each other. Col and Ler therefore artificially inflate estimates of linkage disequilibrium and we excluded them from the analysis. We calculated D' values (LEWONTIN 1964) for each pairwise comparison and determined whether each one differs significantly from 0 (*i.e.*, linkage equilibrium) by means of a Fisher exact test. Because of (1) the relatively small sample size (maximally 32; frequently less in individual comparisons owing to missing data), and (2) the low frequency of most segregating sites (only two sites, *Adh* and *Gap-C*, are segregating at frequencies greater than 30%), our power to detect significant associations is low. We find only one significant association, between *Adh* and *Gap-C* ($D' = 0.4464$; $P = 0.047$). Note that this is the only comparison for which we have reasonable power to reject the null hypothesis.

Sequence analysis of related species: One possible source of the extreme divergence observed between Columbia and Landsberg haplotypes in the hypervariable region is introgression (infiltration of alleles from related species). For this reason we have sequenced the same region from a number of closely related species. This also provides both phylogenetic information and an interspecific comparison for future tests of selective neutrality (*i.e.*, HUDSON *et al.* 1987). Our sequence analysis of species related to *A. thaliana* includes *C. arenosa*, *H. diffusa*, *C. bursa-pastoris*, *A. wallichii* and polyploid *Arabidopsis* species *A. suecica* and *A. pumila*. We amplified and sequenced the same fragment as previously described for ecotypes. The data are shown in Figure 4. Where multiple sequences are detected (as heterozygous base pairs in the sequencing ladder. Although some species sequenced are polyploid, we observed no more than two variants segregating at a single nucleotide position), these "variable positions" cannot be assigned to a given haplotype so, for phylogenetic analysis, these sites were coded as variable. The only exception is *A. suecica*, which was not coded as variable because it is likely to represent the reticulation of two lineages: HYLANDER (1957) hypothesized on morphological and karyological grounds that *A. suecica* ($2n = 2x = 26$) is an amphidiploid hybrid of diploid *A. thaliana* and diploid *C. arenosa* ($2n = 2x = 16$). Examination of the *A. suecica* sequence revealed the presence of all sites

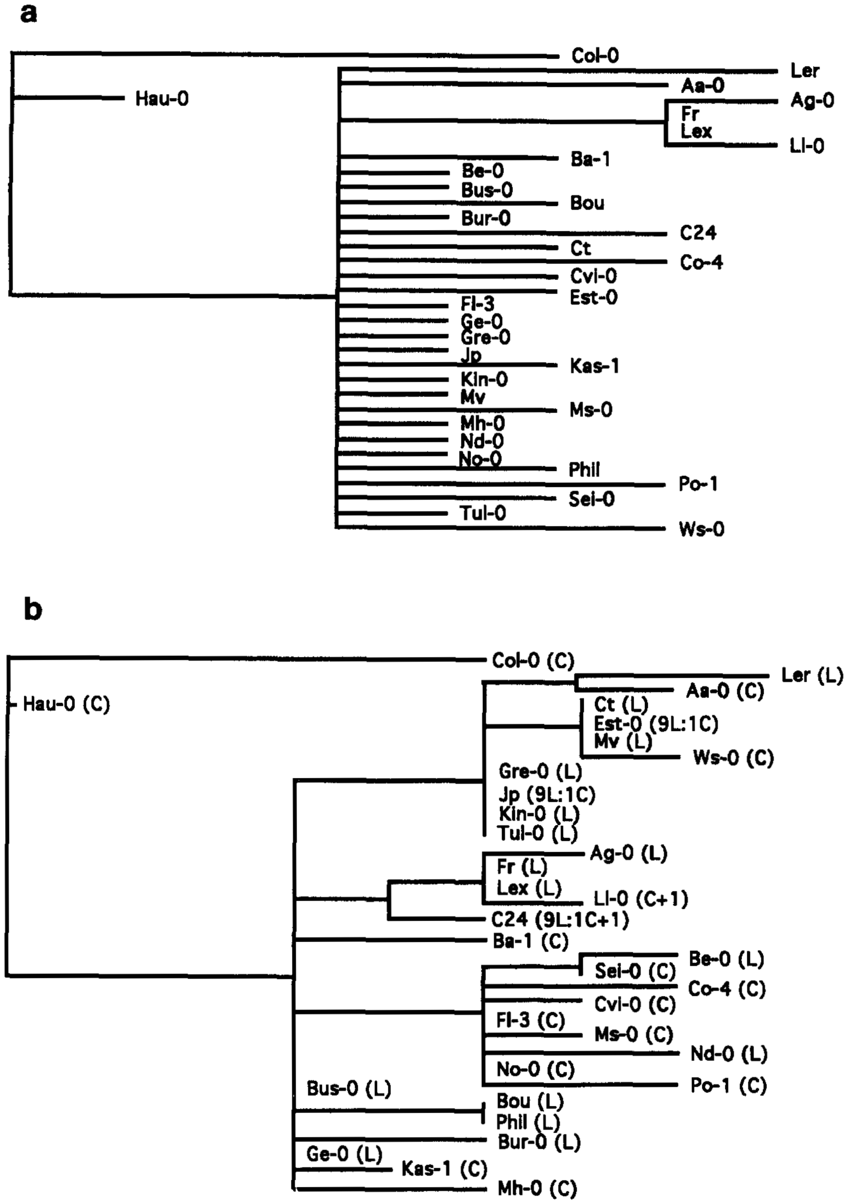

FIGURE 3.—Consensus trees generated by cladistic analysis of 12 CAPS markers in 37 ecotypes. These were computed from 1000 trees of length 35 steps (CI = 0.333, RI = 0.645). (a) Strict consensus, showing only those groups that appear in all trees. (b) Adams consensus, showing groups that appear in all trees, whether or not additional ecotypes are included in those groups. Ecotypes whose position is variable among trees are placed at the lowest node in common among all their placements. The *Adh* haplotype is given for each ecotype: "C" denotes a Columbia sequence, "L" a Landsberg sequence. Ratios refer to the number of Columbia- or Landsberg-characteristic variants segregating at the hypervariable region; additional variant sites are designated "+1."

comprising an *A. thaliana* Columbia haplotype; by removing this "haplotype," we found 16 of the remaining 18 *A. suecica* sites that differ from the *A. thaliana* reference sequence were also present in Cardaminopsis. This we took as support of Hylander's hypothesis and adequate grounds for analyzing two fixed *A. suecica* sequences, one *A. thaliana*-like and the other Cardaminopsis-like. A branch-and-bound search, which is guaranteed to find all most parsimonious trees, found three trees of length 70 steps, CI = 0.778, and RI = 0.881. The strict consensus is shown in Figure 5; bootstrap values were computed from 100 replicates. *A. thaliana* is shown to be monophyletic, and *A. suecica* polyphyletic, as expected. The genus *Arabidopsis* is also polyphyletic, corroborating results from chloroplast DNA (PRICE *et al.* 1994). The important result from our point of view is that *A. thaliana* haplotypes do not group with those from another species (except *A. suecica*), implying that the observed polymorphism arose within *A. thaliana.*

Comparing haplotypes of *A. thaliana* ecotypes with sequences of related species indicates that most substitutions in rare haplotypes (*e.g.*, Ll-0, Rsch-0, C24) are unique, as would be expected. The substitution in ecotype Fl-3B at nucleotide position 1649 (resulting in $\text{Arg}^{103} \rightarrow \text{Pro}^{103}$) is, however, found in at least one allele in *A. suecica*, Cardaminopsis and Capsella. This may be an instance of parallel evolution at the nucleotide sequence level; alternatively, as mentioned above, Fl-3 is a phenotypically anomalous ecotype and it may indeed be the product of hybridization.

DISCUSSION

To date, the study of polymorphism in *A. thaliana* has been limited to restriction fragment length polymorphism (RFLP) (CHANG *et al.* 1988; NAM *et al.* 1989; KING *et al.* 1993), CAPS (KONIECZNY and AUSUBEL 1993) and random amplified polymorphic DNAs (RAPD) (REITER

```
                                                                1  2                    3
               -  -  -     -  -  -     -  -  -     -  -  -    - Asp -     Pro -   - Gln/Lys-  -     -  -  -
           87 GlyAspHis   ValLeuPro   IlePheThr   GlyGluCys   GlyGluCys   ArgHisCys   HisSerGlu   GluSerAsn
Col(Fast)1603 GGAGATCAT   GTGTTGCCG   ATCTTTACC   GGAGAATGT   GGGGAGTGT   CGTCATTGC   CACTCGGAG   GAATCAAAC
Ll-0 (Slow)   ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
Fl-3B (Fast)  ---------   ---------   ---------   ---------   ---------   -C-------   ---------   ---------
Ler (S.Fast)  ---------   ---------   ---------   ---------   --A--T---   ---------   --G------   ---------
Rsch-0(S.Fast)---------   ---------   ---------   ---------   --A--T---   ---------   --G------   ---------
C24 (Fast)    ---------   ---------   ---------   ---------   --A--T---   ---------   A-G------   ---------
Est-0 (S.Fast)---------   ---------   ---------   ---------   --A--T---   ---------   --G------   ---------

A. wallichii  ---------   --T--A---   -----C--T   ---------   --A--C---   -----C--T   ---------   -----C---
A. pumila A   ---------   -----A---   -----C---   ---------   --T--T--C   -----C--T   ---------   -----C---
          B   ---------   --T--A---   -----C--T   ---------   --T--T--C   -----C--T   ---------   -----C---
A. suecica A  ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
           B  ---------   --------C   -----C---   ---------   --A--T---   -C-------   ---------   ---------
Cardaminopsis ---------   --------C   -----C---   ---------   --A--T---   -C------T   ---------   ---------
           B  ---------   --------C   -----C---   ---------   --A--T---   -C------T   ---------   ---------
Capsella A    ---------   --T--A--C   -----C---   -----G---   -----T---   -C---C--T   -----C---   -----C---
         B    ---------   --T--A--C   -----C---   -----G---   -----T---   -C---C--T   -----C---   -----C--T
Halimolobos   ---------   --T--A---   -----C---   ---------   --A--T---   -----C--T   --G------   -----C---

                                              4                 5                          6
               -  -  -     -  -  -     -  -  -     -  -  -     -  -  -     -  -  -     -  -  -     -  -  -
          111 MetCysAsp   LeuLeuArg   IleAsnThr   GluArgGly   GlyMetIle   HisAspGly   GluSerArg   PheSerIle
Col(Fast)1675 ATGTGTGAT   CTTCTCAGG   ATCAACACC   GAGCGAGGA   GGGATGATT   CACGATGGT   GAATCAAGA   TTCTCCATT
Ll-0 (Slow)   ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
Fl-3B (Fast)  ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
Ler (S.Fast)  ---------   ---------   --------A   ---------   --T------   ---------   -----T---   ---------
Rsch-0(S.Fast)---------   ---------   --------A   ---------   ---------   ---------   -----T---   ---------
C24 (Fast)    ---------   ---------   --------A   ---------   ---------   ---------   -----T---   ---------
Est-0 (S.Fast)---------   ---------   --------A   ---------   ---------   ---------   -----T---   ---------

A. wallichii  ---------   -----A---   --------A   --A-----C   ---------   -----C---   -----T---   ---------
A. pumila  A  ---------   ---------   -----T---   ---A-----   ---------   --T--C---   ---------   ---------
           B  ---------   ---------   -----T---   ---A-----   ---------   --T--C---   ---------   ---------
A. suecica A  ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
           B  ---------   ---------   ---------   ---------   ---------   ---------   -----G---   --------C
Cardaminopsis ---------   ---------   ---------   ---------   ---------   ---------   -----G---   --------C
          B   ---------   ---------   -----T--A   ---------   --A------   ---------   -----G---   --------C
Capsella A    ---------   ---------   ---------   ---------   ---------   -----C---   -----T---   ---------
         B    ---------   ---------   ---------   -----T---   --A------   -----C---   -----T---   ---------
Halimolobos   ---------   ---------   ---------   --A-----C   ---------   -----C---   -----T---   ---------

                               7                       8                          9
              Lys -  -     -  - His    -  -  -     -  -  -     -  -  -     -  -  -     -  -  -     -  -  -
          135 AsnGlyLys   ProIleTyr   HisPheLeu   GlyThrSer   ThrPheSer   GluTyrThr   ValValHis   SerGlyGln
Col(Fast)1747 AATGGCAAA   CCAATTTAC   CATTTCCTT   GGGACTTCC   ACGTTCAGT   GAGTACACA   GTGGTTCAC   TCTGGTCAG
Ll-0 (Slow)   ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
Fl-3B (Fast)  ---------   ---------   ---------   -----G---   ---------   ---------   ---------   ---------
Ler (S.Fast)  ---------   -----C---   ---------   -----G---   ---------   --------T   ---------   ---------
Rsch-0(S.Fast)---------   -----C---   ---------   ---------   ---------   --------T   ---------   ---------
C24 (Fast)    ---------   -----C---   ---------   -----G---   ---------   --------T   ---------   ---------
Est-0 (S.Fast)---------   -----C---   ---------   -----G---   ---------   --------T   ---------   ---------

A. wallichii  --A------   --G--AC-T   ---------   --T------   -----T--C   --------T   ---------   ---------
A. pumila A   ---------   --G---C-T   ---------   ---------   -----T---   --------T   ---------   ---------
          B   ---------   --G---C-T   ---------   ---------   -----T---   --------T   ---------   --C------
A. suecica A  ---------   ---------   ---------   ---------   ---------   ---------   ---------   ---------
           B  ---------   ------C-T   ---------   ---------   -----T---   --------T   ---------   --A------
Cardaminopsis ---------   --------T   ---------   ---------   -----T---   --------T   ---------   ---------
          B   ---------   ------C-T   ---------   ---------   -----T---   --------T   ---------   ---------
Capsella A    ---------   --C---C-T   ---------   ---------   -----T---   --------T   ---------   -----C---
         B    ---------   --C---C-T   ---------   ---------   -----T---   --------T   ---------   -----C---
Halimolobos   ---------   --G---C-T   ---------   ---------   -----T---   --------T   --T------   ---------

              10
               -  -  -     -  - Arg   Glu -  -    Ile -  -     -  -  -
          159 ValAlaLys   IleAsnPro   AspAlaPro   LeuAspLys   ValCysIle
Col(Fast)1819 GTTGCTAAG   ATCAATCCG   GATGCTCCT   CTTGACAAG   GTCTGTATT
Ll-0 (Slow)   ---------   -------G-   ---------   ---------   ---------
Fl-3B (Fast)  ---------   ---------   ---------   ---------   ---------
Ler (S.Fast)  --C------   ---------   ---------   ---------   ---------
Rsch-0(S.Fast)--C------   ---------   --A------   ---------   ---------
C24 (Fast)    --C------   ---------   ---------   ---------   ---------
Est-0 (S.Fast)--C------   ---------   ---------   ---------   ---------

A. wallichii  --C------   ---------   ---------   --------A   --T-----C
A. pumila  A  --C------   ---------   ---------   --------A   --T--C---
           B  --A------   ---------   ---------   --------A   --T--C---
A. suecica A  ---------   ---------   ---------   ---------   ---------
           B  ---------   ---------   ---------   --------A   --T--C---
Cardaminopsis ---------   ---------   ---------   --------A   --T--C---
          B   --C------   ---------   ---------   --------A   --T--C---
Capsella A    --C------   -----C---   ---------   --------A   ---------
         B    --C------   -----C---   ---------   --------A   ---------
Halimolobos   --C------   ---------   ---------   A-------A   --T-----C
```

FIGURE 4.—Nucleotide sequences and amino acid substitutions in the hypervariable region of *A. thaliana* and related species. All seven observed *A. thaliana* haplotypes are shown and sites differing between Columbia and Landsberg haplotypes are numbered. The designation of A and B alleles is arbitrary for polyploid species but serves to present the total range of sequence variation observed (see RESULTS). Sequence identical to Columbia (in bold) is shown as a dash. The Columbia protein sequence is shown above the Columbia nucleotide sequence. Differences in amino acid sequence are shown above the Columbia protein sequence: amino acids identical to Columbia are shown as a dash and replacement changes show the new amino acid. At site 3 two different amino acid substitutions occur at the same codon and both the amino acid and nucleotide substitutions of one mutation are underlined.

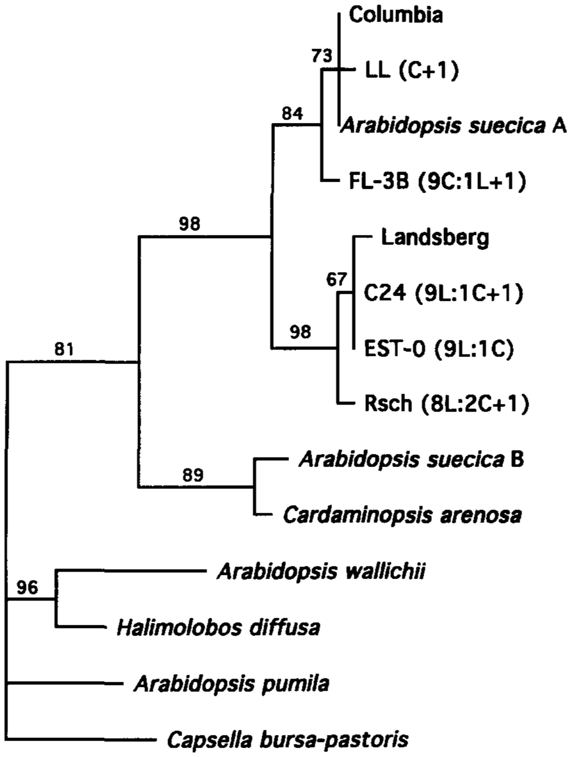

FIGURE 5.—Gene tree for *Adh* haplotypes in *A. thaliana* and related species: strict consensus of three trees of length 70 steps (CI = 0.778, RI = 0.881). Numbers above branches indicate percentage of bootstrap replicates.

et al. 1992) gene mapping studies and to electrophoretic mobility assays of soluble proteins (CAMMAERTS and JACOBS 1975, 1983; GROVER 1975; JACOBS and SCHWIND 1976; DOLFERUS and JACOBS 1984; FUGLEWICZ and KILIAN 1985; ABBOT and GOMES 1989). We present here the first study of DNA sequence variation in *A. thaliana*. The *Adh* gene of the Columbia ecotype shows substantial sequence divergence from the Landsberg ecotype (CHANG and MEYEROWITZ 1986) in just one region which we designate hypervariable: 12 substitutions differentiating the two sequences are clustered in a 312-bp window. Of these, 10 occur within a 180-bp region in exon 4. In the remaining 2611 bp of compared sequence there are only three differences, two small insertions/deletions and one point mutation. The lack of differentiation between the two haplotypes outside the hypervariable region is striking but apparently consistent with observations at other loci (Table 4), suggesting that *A. thaliana* typically has low levels polymorphism. This implies that *A. thaliana* either has a chronically low effective population size or has recently undergone an acute reduction in effective population size (associated with, say, a bottleneck) from which levels of polymorphism have not yet recovered. If this is true, then the high levels of heterozygosity at the hypervariable region are all the more remarkable.

We sequenced the hypervariable region of the *Adh* genes of 37 further ecotypes originating from a large geographic area (Table 1) and find that the sequences fall into two distinct groups, one Columbia-like, and one Landsberg-like. Although most of these ecotypes have long been maintained in the laboratory, this dichotomous pattern is not an artifact of inadvertent artificial selection: sequence of the hypervariable region for 15 individuals from four natural populations yielded just Columbia and Landsberg haplotypes (A. BERRY, unpublished). In this discussion we attempt to explain (1) the clustering of differences between the Columbia and Landsberg *Adh* sequences, and (2) the lack of haplotypic diversity for the hypervariable region among *A. thaliana* ecotypes. First, however, we describe two factors that handicap our analysis.

Absolute rates of recombination in *A. thaliana* can be gauged in the laboratory, but we do not know the rate at which novel combinations are produced by recombination in nature. Because outcrossing occurs at a rate of less than one per cent of fertilizations in natural populations (ABBOT and GOMES 1989), *A. thaliana* populations are highly inbred (CETL 1987; ABBOT and GOMES 1989). Thus, even though recombination may occur regularly in nature, exchange typically takes place between related genomes and does not therefore result in the generation of new haplotypes. We need to know the rate of *effective* recombination (*i.e.*, recombination resulting in new genotypic combinations), which is a function of both the absolute rate of recombination and the population structure of the species. Appropriate information on population structure is lacking so we can make no strong statements on the role of recombination in generating the observed patterns.

We lack "reference" loci. HUDSON *et al.* (1987) introduced a means of detecting natural selection at the molecular level based on the neutral expectation of a correspondence between levels of intraspecific polymorphism and interspecific divergence. Its strength lies in its ability to factor out the effects of differences in mutation rate and intensities of purifying selection because, under neutrality, each influences in the same way both polymorphism and divergence for a given locus: for example, a pseudogene, which is released from purifying selection, is expected to be highly polymorphic within a species and highly divergent between species. This technique entails the comparison of the study locus to a reference locus, whose patterns of intra- and inter-specific variation represent, ideally, the outcome of neutral processes (*i.e.*, strict neutrality plus purifying selection). Without reference loci, we cannot implement this comparative approach.

How can we account for the hypervariable region?

Figure 2 shows clearly the heterogeneity in the distribution of variation between Columbia and Landsberg *Adh* haplotypes and the broken stick test reveals this heterogeneity to be statistically significant. Is this pattern

TABLE 4

Divergence between genes sequenced in both *A. thaliana* ecotypes Columbia and Landsberg

Gene sequenced	Compared sequence (kb)	Mutations	Transcribed sequence (kb)	Gene structure	Mutations in Exons	Mutations in Introns
Locus of *erecta* mutation in Landsberg; (I. HWANG, personal communication)	3.3	6	2.4	7 exons, 6 introns	2 silent substitutions; 1 amino acid replacement	1 substitution
Open reading frame in cosmid clone 8261 (unknown function); (H. GOODMAN, P. GALLANT and H.-H. CHIANG, personal communication)	2.7	4	2.7	1 exon, no intron	3 silent substitutions; 1 amino acid replacement	
Chalcone synthase gene; (FEINBAUM and AUSUBEL 1988; B. SHIRLEY, personal communication)	1.4	2	1.4	2 exons, 1 intron	1 silent substitution	

a sampling artifact of surveying the *Adh* genes of only two ecotypes? Three observations suggest that the hypervariable region is genuinely anomalous. First, J. BERGELSON and M. KREITMAN (University of Chicago) have performed a 4-cutter RFLP analysis of the *A. thaliana Adh* locus (personal communication) on 70 individuals from 7 European and American populations. They used eight enzymes and found no variation at all outside the hypervariable region. Second, a survey of the few other loci in *A. thaliana* that have been sequenced in more than one ecotype (Table 4) shows the pattern of divergence between Columbia and Landsberg outside the hypervariable region to be typical of other loci. Finally, we note that sequencing a further 37 ecotypes' hypervariable regions revealed only four additional (low frequency) segregating sites, suggesting that, for this region at least, Columbia and Landsberg represent an adequate sample of species-wide diversity. We thus regard the *A. thaliana Adh* hypervariable region to be a real phenomenon in need of explanation. We identify four possible explanations for this heterogeneity in the distribution of variation between *Adh* haplotypes.

Non-uniform mutation rate: The hypervariable region may have an elevated mutation rate. Although two induced mutations have previously been detected in the hypervariable region (DOLFERUS *et al.* 1990), we have sequenced the hypervariable regions of 20 ADH-deficient mutants (2 natural; 18 induced) and found only two with a mutation in the hypervariable region, suggesting that this is not an unusually frequent target of mutagenesis. Of the remaining 18 mutants the actual site of mutation has been determined for only two and so it is possible that some of the mutants are *trans*-acting, and therefore not applicable to a calculation of *Adh* mutation frequency. Also, it is open to question how much we may infer about natural mutation processes from the experimental application of mutagenic agents. We cannot, therefore, formally reject the possibility that the peak in divergence between Landsberg and Columbia haplotypes is the product of a locally elevated mutation rate. We note, however, that no mechanism of extreme local mutation rate mosaicism has been described in organisms of *A. thaliana*'s level of complexity.

Variation in intensities of purifying selection: The uniformity we see between Landsberg and Columbia *Adh* haplotypes throughout the gene, except for the hypervariable region, may reflect strong purifying selection, which eliminates all variation. The hypervariable region may simply be an area where this selective constraint is relaxed and the accumulation of variation permitted. However, we expect that, if it is acting at all, purifying selection will preserve amino acid sequence; polymorphism will be concentrated in introns and effectively silent sites. Yet the hypervariable region is in fact largely located in exon four, and, furthermore, two of its mutations result in amino acid replacements. Exon four encodes a functionally important part of the protein (see below); it is improbable that it is subject to substantially less selective constraint than other parts of the gene, including flanking sequence and a number of introns. This is supported by the results of our mutagenesis study (see below) which reveal the hypervariable region to be important functionally. We therefore consider heterogeneity in purifying selection intensity as unlikely to be responsible for the hypervariable region.

Introgression: All or part of one of the haplotypes may be derived from a genealogically distant population or species: presumably the "foreign" genome introduced through introgression has been broken up and integrated with the *A. thaliana* genome through recombination. Thus the hypervariable region would constitute a residual fragment of the foreign genome embedded in the *A. thaliana* genome. We investigated this possibility by sequencing the region homologous to the *A. thaliana* hypervariable region of several closely related species in order to determine whether either haplotype is more closely related to haplotypes present in other species, implying some form of horizontal transfer, such as introgression. Because it is apparently the product of hybridization between *A. thaliana* and *C. arenosa* (see above), one *A. suecica* haplotype clusters with *A. thaliana*. However, our cladogram (Figure 5) shows no evidence of introgression. The absence of aberrant haplotypes (or partial haplotypes) at other multiply sequenced loci (Table 4) also argues against introgression. In addition, only one of the 12 CAP's loci

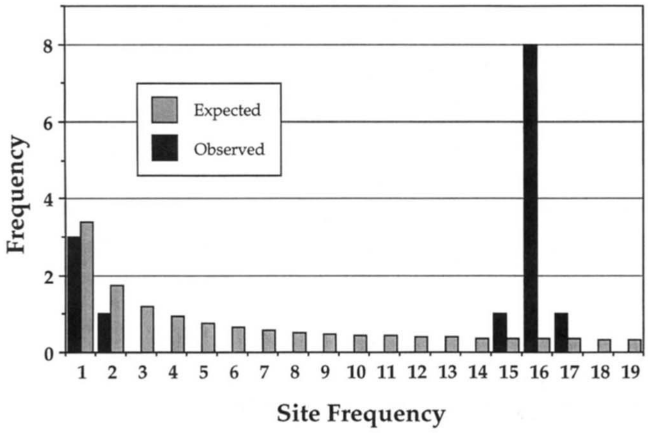

FIGURE 6.—The distribution of polymorphisms among 39 ecotypes for the *Adh* hypervariable region. All polymorphic nucleotide positions were dimorphic and we give here the distribution of the rarer variant at each site. Expected values, under neutrality, were calculated from TAJIMA's (1989) equation 51.

showed a restriction digest pattern different from the Columbia or Landsberg patterns. Introgression is expected to have genome-wide consequences, so, if it were occurring, we would expect a high proportion of non-Columbia/Landsberg CAP's patterns among the 32 ecotypes surveyed. We take the lack of such aberrant patterns as evidence against introgression. Again, we cannot formally reject the possibility of introgression producing the hypervariable region (for example, perhaps the introgressing taxon is now extinct making it impossible to identify the source of the exogenous material) but we deem it unlikely on the basis of the available circumstantial evidence.

A balanced polymorphism: Because balancing selection maintains two or more alleles at a locus indefinitely at intermediate frequencies, we expect, through hitchhiking, an accumulation of neutral variation in the vicinity of the selected site (STROBECK 1983; HUDSON and KAPLAN 1988). The size of the window of elevated polymorphism is dependent on recombination and mutation rates. We argue here that the pattern at *A. thaliana Adh* is consistent with a hypothesis of balancing selection. Note that we are using the term "balancing selection" in a loose sense: we are not referring only to heterotic selection but to any selection regime which promotes the co-existence of two or more alleles at a locus in a population. Heterotic selection in this case is unlikely in view of *A. thaliana*'s tendency to inbreed (HAYMAN 1953).

Figure 2 shows the pattern of divergence between the Landsberg and Columbia haplotypes to be consistent with STROBECK's prediction of elevated levels of polymorphism in the vicinity of a balanced polymorphism. Figure 6 further suggests that the distribution of variants in the hypervariable region is incompatible with a neutral explanation. Only one previous study, that of Kreitman and others on *Drosophila melanogaster Adh*, has identified a similar peak of polymorphism (KREITMAN and AGUADÉ 1986; HUDSON *et al.* 1987; KREITMAN and HUDSON 1991). As this peak flanks the amino acid polymorphism causing the electrophoretic difference between *D. melanogaster Adh* alleles, these authors have concluded that the amino acid polymorphism is subject to balancing selection.

We have applied TAJIMA's (1989) test of selective neutrality: TAJIMA's $D = 1.825$ ($0.1 > P > 0.05$). FU and LI (1993) present another test based on the distribution of variants on a phylogenetic tree. For this test, we used *C. arenosa* as an outgroup (see Figure 5) and accordingly excluded one (C at position 1649) of the three singleton nucleotide variants from our total of "external" mutations, η_e. FU and LI's $D = 0.6503$ and their $F = 1.258$ (for both, $P > 0.1$). Although both tests fail to reject a hypothesis of neutrality, these results nevertheless are not strong evidence against a selective explanation for the pattern because such tests appear to lack statistical power. KREITMAN (1991) found that the TAJIMA test failed to reject neutrality at *D. melanogaster Adh* while a test incorporating inter-specific divergence at the locus (HUDSON *et al.* 1987) detected a significant departure from neutrality. An additional reason for not placing too much emphasis on these test results is theoretical: the models on which the tests are based assume a Wright-Fisher distribution and this assumption is violated in the case of a global sample such as ours. We present now circumstantial evidence justifying our preliminary interpretation of the hypervariable region in *A. thaliana Adh* as the product of balancing selection.

The hypervariable region contains candidate polymorphisms: Because selection coefficients associated with mutations causing amino acid replacements are generally higher than those associated with effectively silent mutations, we expect balancing selection typically to act on amino acid polymorphisms rather than effectively silent differences. There are only two amino acid differences between the Columbia and Landsberg haplotypes and both lie within the hypervariable region and we hypothesize that one (or both) of these amino acid polymorphisms is subject to balancing selection. We expect the accumulation of differences between the alleles to peak around the position of the selected site. Although the position of neither amino acid substitution coincides precisely with the peak shown in Figure 2 (at position 1720), the mutation causing the change in electrophoretic mobility is close (at 1659; amino acid position 106), so we hypothesize further that this is the polymorphism subject to balancing selection.

A. thaliana Adh allozymes appear to differ functionally: The glutamine to histidine substitution at amino acid position 106 between Columbia and Landsberg in the hypervariable region is responsible for the mobility difference between *A. thaliana* fast and super fast allozymes. Although we know these allozyme classes to be heterogeneous at the amino acid sequence level (*cf.* ecotype C24, Table 1), we note that the amino acid at position 106 predicts the mobility of the allozyme in 37 out of 39 cases (C24 and Ll-0 are the exceptions), implying that there is relatively little other mobility-affecting amino acid variation segregating at *Adh.* Differences between allozyme classes may therefore reasonably be attributed to the amino acids at position 106 and/or variants in linkage disequilibrium with them. DOLFERUS and JACOBS (1984) surveyed the enzymological properties of the allozymes and found them to differ in their pH optima and thermal stabilities. We do not know whether these differences are significant with respect to plants in natural populations, but they form the basis of physiological traits that could potentially be subject to natural selection.

Mutagenesis and structural studies indicate that the hypervariable region is functionally important: ADH's three-dimensional structure is known [from horse liver ADH-E (BRÄNDÉN *et al.* 1973; EKLUND *et al.*, 1976)], so we can make inferences about the effect of mutations on enzyme structure and possibly function. Of the two zinc atoms, one is in the active site, while the other is referred to as the "structural" zinc. Sequence comparison reveals that the enzyme is highly conserved across many species, including plants (YOKOYAMA and HARRY 1993). The hypervariable region encodes part of the active site and part of the domain involved in the co-ordination of the structural zinc. Given such strong conservation across species, variability in a region encoding part of the active site is surprising. In addition, we argue here from mutagenesis data that the structural zinc binding domain is also functionally important. Thus, throughout the hypervariable region, we expect to see the low levels of polymorphism that are characteristic of evolutionarily constrained regions, but observe the reverse.

Of six characterized loss-of-function mutants (Table 3), three involve the introduction of stop codons, while the remaining three mutants involve single amino-acid substitutions. *Adh 103* causes a phenylalanine to leucine substitution at residue 93/95 (residue numbers refer to the horse liver protein/*A. thaliana* protein), located in the active-site pocket, but the other two mutations are located outside the active site, affecting cysteine ligands of the structural zinc. The zinc atom is tetrahedrally co-ordinated by 4 cysteine residues, located in a lobe of the molecule adjacent to the catalytic domain. In *Adh104* cysteine 111/113 is replaced by serine, while in R006, cysteine 103/105 is replaced by tyrosine. Since these defects in the binding of the "structural" zinc are seen to impair enzymatic function, the mutants R006 and *Adh104* provide the first evidence that the structure of this second zinc-binding domain is important for enzymatic function, as previously hypothesized by EKLUND *et al.* (1976). The two amino acid substitutions that distinguish Columbia and Landsberg haplotypes are also located in the structural zinc binding domain. The substitution at site 2 (Asp → Glu) is between cysteines 97/99 and 100/102, and the substitution at site 3 (Gln → His), which is responsible for altered electrophoretic mobility, is between cysteines 103/105 and 111/113. How these amino acid residue differences relate to the functional differences between allozymes identified by DOLFERUS and JACOBS (1984) is unclear.

To summarize, the peak in polymorphism in the hypervariable region is consistent with the presence of a balanced polymorphism and such an interpretation explains our otherwise anomalous finding of extensive polymorphism in part of a gene encoding a functionally important protein domain. This is supported by the observation that the hypervariable region contains amino acid substitutions of potential physiological significance which may therefore be subject to balancing selection.

How can we account for the lack of haplotypic diversity?

Hypervariable region sequence from a sample of 39 ecotypes revealed only seven haplotypes, which fall into two distinct classes corresponding to the original sequences from the Columbia and Landsberg ecotypes. Within either class, the haplotype that is most divergent from the class type is shared by Rsch-0 and N916. This haplotype differs from the Landsberg haplotype at three sites. The maximum within-class divergence is thus three sites; what is the minimum between-class divergence? The Rsch-0/N916 haplotype is the only 8:2 haplotype in either class: it is a Landsberg-type haplotype with two Columbia-type sites. Thus the minimum between-class divergence is eight sites: the Columbia/Landsberg split forms an ancient coalescence while, within each class, we see relatively recent coalescences. We have hypothesized above that balancing selection may account for the antiquity of the Columbia/Landsberg split; in this section, we attempt to account for this lack of within-class diversity. We discuss three factors possibly affecting within-class diversity.

Low rates of recombination: As recombination generates haplotypic diversity, is the observed lack of haplotypic diversity attributable entirely to low rates of effective recombination? Unfortunately, as discussed above, we have no estimate of rates of effective recombination in nature. Note, however, that haplotypic diversity is generated by two means, mutation and recombination. Because of the great mutational distance between the Columbia and Landsberg hypervariable region haplotypes, there must have been many "intermediate" haplotypes produced in the course of their evolution from a common ancestor. Conceivably haplotypes, such as that of ecotype FL-3B, which include mutations characteristic of the class of which they are not a member, do indeed represent these mutational intermediates (alternatively, they may have been produced by local recombination processes such as gene conversion). That the "most intermediate" of these intermediate haplotypes still clusters phylogenetically securely in its own class (*i.e.*, carries only two other-class mutations and eight own-class mutations) argues the operation of extrinsic factors in reducing diversity because both mutation- and recombination-derived diversities are low.

Table 2, showing the distribution among ecotypes of a number of sites that vary between Columbia and Landsberg ecotypes, reveals the action of recombination: there is sufficient recombination to break up within-chromosome associations (*i.e.*, a chromosomal haplotype is typically a mixture of Columbia and Landsberg genotypes). This pattern tells us very little about rates of effective recombination as even minimal rates will erode associations between such dispersed loci; however, we can at least conclude that the *A. thaliana* genome has not completely congealed as a result of inbreeding. The finding of between-chromosome linkage disequilibrium between *Adh* and *Gap-C* is interesting and warrants further investigation. Given high rates of inbreeding, the cost of selection entailed by the selective maintenance of such a between-chromosome association may not be prohibitive. Note that there is a possible functional link between the two enzymes because the NAD necessary for glyceraldehyde 3-phosphate dehydrogenase (GAPDH) function is provided, in anaerobiosis, by ADH; the GAP-C protein, like ADH, is cytosolic while the GAP-A and GAP-B proteins are both localized in the chloroplast (Shih *et al.* 1991).

If balancing selection accounts, as we hypothesize, for the local peak in divergence between the Columbia and Landsberg haplotypes, we have evidence of significant rates of effective recombination. The peak of polymorphism spans only 300 bp, implying that linkage disequilibrium is only high enough to promote hitchhiking in a window of that size (*cf.* Kreitman and Hudson 1991). Effective recombination rates must be substantial to erode linkage disequilibrium between sites just 350 bp apart. Thus, if the peak of polymorphism is indeed caused by the presence of a balanced polymorphism, we can reject the hypothesis that the lack of diversity is the product of low rates of effective recombination; rather, there has been insufficient time post-coalescence for recombinational haplotypic diversification.

Population structure and history: It is conceivable that *A. thaliana* populations underwent some kind of bottleneck event in their relatively recent history which resulted in the fortuitous sampling of just two *Adh* hypervariable region haplotypes. Such an event would have genome-wide consequences so this model is easily tested by determining whether or not there is a similar pattern at other, unlinked, loci. Because we currently lack data for more than two ecotypes at other loci (Table 4), we cannot formally reject this model. We would predict linkage disequilibrium among the CAPS loci sampled (Table 2) if (1) the bottle neck event were very recent, and (2) effective recombination rates are very low. The observed lack of within-chromosome association implies therefore that the hypothetical bottleneck is old enough for recombination to have eliminated such associations. We note that there is no strong geographic component to the distribution of the Landsberg and Columbia haplotypes (*e.g.*, there is no north/south split), as might be expected if historical accidents had determined the distribution of haplotypes. Indeed, the two haplotypes can be found segregating in the same natural population (A. Berry, unpublished results).

Directional selection: Directional selection can reduce genetic diversity through background selection (Charlesworth *et al.* 1993) or through a selective sweep (Berry *et al.* 1991). Background selection eliminates deleterious mutations by purifying selection. If a single linkage group is large (*i.e.*, local rates of recombination are low), the per generation rate of deleterious mutation (and corresponding rate of purifying selection) across the linked region may be high enough to reduce diversity substantially relative to regions where recombination rates are higher (*i.e.*, linkage groups are smaller). As mutation rates are typically low, background selection is likely only to have a significant effect on genetic diversity when linkage groups are very large. A selective sweep entails positive selection in favor of an adaptive mutation. In areas of reduced or no recombination, this process can result, through hitchhiking, in the serendipitous fixation of other mutations in linkage disequilibrium with the selectively driven mutation. The selective sweep model is not as recombination-dependent as the background selection one: the larger the linkage group, the higher the probability that a sweep has occurred (assuming that the probability of an adaptive mutation occurring at any given site in the region in question is uniform). However, in principle, it is possible to detect the area of reduced heterozygosity resulting from a selective sweep in a region of frequent recombination; the problem is merely that that region is likely to be small. Thus, to explain the reduced diversity we observe at *A. thaliana Adh*, background selection requires very low rates of recombination while a selective sweep has a local diversity-reducing effect even in regions of frequent recombination.

If the cause of the excess of polymorphism is, as we argue, an ancient balanced polymorphism, then we can conclude that rates of effective recombination are high enough to reduce the size of the window subject to hitchhiking to less than 300 bp. This is not much larger than that observed by Kreitman and Hudson (1991) in *D. melanogaster.* How can we reconcile a directional selection explanation for the lack of diversity with such apparently high rates of recombination? In the case of background selection, we cannot: background selection requires extensive linkage groups. In the case of a selective sweep, however, high rates of recombination merely imply that the site of the sweep was close to the hypervariable region and/or selection was strong enough to drive the sweep to completion too quickly for recombination to erode the linkage disequilibrium between the selected variant(s) and variation in the hypervariable region.

CONCLUSION

A hypothesis: balancing and directional selection together explain the data: We have argued above that the peak in divergence between the Landsberg and Columbia ecotype haplotypes is best explained by an ancient balanced polymorphism and that the low within haplotypic diversity is best explained by a selective sweep. We now combine these two conclusions to form a hypothesis about the evolution of the *A. thaliana Adh* locus.

We hypothesize that *A. thaliana Adh* is under balancing selection for one or both of the amino acids differentiating Columbia and Landsberg haplotypes. Furthermore, we hypothesize that this is an ancient selected polymorphism which has permitted, through hitchhiking, the build up of substantial neutral divergence between the alleles in the region of the selected polymorphism. We suggest that there arose at a linked locus an adaptive mutation, "α," that was subject to directional selection. Through hitchhiking, the frequency of the single allele (and haplotype) at *Adh*, say Landsberg, that happened to be in coupling linkage disequilibrium with α would increase as selection drove α towards fixation. However, the selective sweep could not go to completion (*i.e.*, the fixation of α) because of selection in favor of the other allele at *Adh*, in this case, Columbia. How can this scenario of conflicting selection pressures be resolved? We hypothesize that α eventually recombined on to a Columbia-type allele and the selective sweep of α could then go to completion. The process has effectively sampled one representative of each of the two alleles at *Adh* and because they were sampled from two highly divergent pools of haplotypes we expect the two sampled haplotypes also to be highly divergent in the region around the selected site(s) (as the Columbia and Landsberg haplotypes are). As, in evolutionary terms, we hypothesize this to have been a relatively recent event, there has been limited time for subsequent diversification within each allelic class. This accounts for the lack of within-class diversity.

At this stage, the above is an entirely hypothetical account of the evolutionary history of the *A. thaliana Adh* locus. It is merely consistent with our observations. Its strength as an explanation lies in the invocation of a common mechanism, balancing selection, to explain aspects of both outstanding features of the data, the divergence between haplotypes in the hypervariable region and the lack of diversity within classes.

We thank members of the AUSUBEL laboratory for their support and, for comments and criticism, M. KREITMAN, A. BARBADILLA, S. WELLS and R. LEWONTIN. This research was supported by U.S. Department of Agriculture-CRGO Grant 91-37300-6384 and by a grant from Hoechst AG to Massachusetts General Hospital. A.B. was supported by Harvard University Society of Fellows and the Milton Fund.

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.

AUSUBEL, F. M., R. BRENT, R. E. KINGSTON, D. D. MOORE, J. G. SEIDMAN *et al.*, 1989 *Current Protocols in Molecular Biology.* Wiley Interscience, New York.

BERRY, A. J., J. W. AJIOKA and M. KREITMAN, 1991 Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

BRÄNDÉN, C.-I., H. EKLUND, B. NORDSTRÖM, T. BOIWE, G. SÖDERLUND *et al.*, 1973 Structure of liver alcohol dehydrogenase at 2.9-Å resolution. Proc. Nat. Acad. Sci. USA **8:** 2439–2442.

CAMMAERTS, D., and M. JACOBS, 1975 Study of the intracellular location and the genetic control of malate dehydrogenase isozymes in *Arabidopsis thaliana.* Plant Sci. Lett. **4:** 249–256.

CAMMAERTS, D., and M. JACOBS, 1983 A study of the polymorphism and the genetic control of the glutamate dehydrogenase isozymes in *Arabidopsis thaliana.* Plant. Sci. Lett. **31:** 65–73.

CETL, I., 1987 Genetic analysis of *Arabidopsis thaliana* (L.) populations from Czechoslovakia. Arabidopsis Inf. Serv. **25:** 67–84.

CHANG, C., and E. M. MEYEROWITZ, 1986 Molecular cloning and DNA sequence of the *Arabidopsis thaliana* alcohol dehydrogenase gene. Proc. Natl. Acad. Sci. USA **83:** 1408–1412.

CHANG, C., J. L. BOWMAN, A. W. DEJOHN, E. S. LANDER and E. M. MEYEROWITZ, 1988 Restriction fragment length polymorphism map for *Arabidopsis thaliana.* Proc. Natl. Acad. Sci. USA **85:** 6856–6860.

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

DELLAPORTA, S. L., J. WOOD and J. B. HICKS, 1983 A plant DNA minipreparation: version II. Plant Mol. Biol. Rep. **1:** 19–21.

DENNIS, E. S., W. L. GERLACH, A. J. PRYOR, J. L. BENNETZEN, A. INGLIS *et al.*, 1984 Molecular analysis of the alcohol dehydrogenase (*Adh1*) gene of maize. Nucleic Acids Res. **12:** 3983–4000.

DEVEREUX, J., P. HAEBERLI and O. SMITHIES, 1984 A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:** 387–395.

DOLFERUS, R., and M. JACOBS, 1984 Polymorphism of alcohol dehydrogenase in *Arabidopsis thaliana* (L.) Heynh.: genetical and biochemical characterization. Biochem. Genet. **22:** 817–838.

DOLFERUS, R., G. MARBAIX and M. JACOBS, 1985 Alcohol dehydrogenase in *Arabidopsis*: analysis of the induction phenomenon in plantlets and tissue cultures. Mol. Gen. Genet. **199:** 256–264.

DOLFERUS, R., D. VAN DEN BOSSCHE and M. JACOBS, 1990 Sequence analysis of two null-mutant alleles of the single *Arabidopsis* Adh locus. Mol. Gen. Genet. **224:** 297–302.

EKLUND, H., B. NORDSTRÖM, E. ZEPPEZAUER, G. SÖDERLUND, I. OHLSSON *et al.*, 1976 Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 Å resolution. J. Mol. Biol. **102:** 27–59.

FEINBAUM, R. L., and F. 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.

FUGLEWICZ, A., and A. KILIAN, 1985 Variability of enzymatic systems in natural populations of *Arabidopsis thaliana* in Poland. Arabidopsis Inf. Serv. **22:** 87–90.

GAUT, B. S., and M. T. CLEGG, 1993a Molecular evolution of the *Adh1* locus in the genus *Zea.* Proc. Natl. Acad. Sci. USA **90:** 5095–5099.

GAUT, B. S., and M. T. CLEGG, 1993b Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics **135:** 1091–1097.

GOTTLIEB, L. D., 1982 Conservation and duplication of enzymes in plants. Science **216:** 373–380.

GROVER, N. S., 1975 Characterization of *Arabidopsis thaliana* ecotypes on the basis of genetic variation at ten isozyme loci. Arabidopsis Inf. Serv. **12:** 19–21.

HANFSTINGL, U., 1994 An analysis of the alcohol dehydrogenase gene of the plant *Arabidopsis thaliana*: evolutionary aspects. Ph.D. Thesis, Ludwigs-Maximilians-Universität München.

HAYMAN, B. I., 1953 Mixed selfing and random mating when homozygotes are at a disadvantage. Heredity **7:** 185–192.

HUDSON, R. R., and N. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

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

HYLANDER, 1957 *Cardaminopsis suecica* (Fr.) Hiit., a northern amphidiploid species. Bull. Jard. Bot. Bruxelles **27:** 591–604.

JACOBS, M., and F. SCHWIND, 1976 Biochemical genetics of acid phosphatase isozymes in *Arabidopsis thaliana* (L.) Heyn. Arabidopsis Inf. Serv. **13:** 56–73.

JACOBS, M., R. DOLFERUS and D. VAN DEN BOSSCHE, 1988 Isolation and biochemical analysis of ethyl methanesulfonate-induced alcohol dehydrogenase null mutants of *Arabidopsis thaliana* (l.) Heynh. Biochem. Genet. **26:** 105–122.

JARILLO, J. A., A. LEYVA, J. SALINAS and J. M. MARTINEZ-ZAPATER, 1993 Low temperature induces the accumulation of alcohol dehydrogenase mRNA in *Arabidopsis thaliana,* a chilling tolerant plant. Plant Physiol. **101:** 833–837.

KING, G., D. NIENHUIS and C. HUSSEY, 1993 Genetic similarity among ecotypes of *Arabidopsis thaliana* estimated by analysis of restriction fragment length polymorphisms. Theor. Appl. Genet. **86:** 1028–1032.

KONIECZNY, A., and F. M. AUSUBEL, 1993 A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based marker. Plant J. **4:** 403–410.

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster.* Nature **304:** 412–417.

KREITMAN, M., 1991 Detecting selection at the level of DNA, pp. 204–221 in *Evolution at the Molecular Level,* edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM. Sinauer Associates, Sunderland, Mass.

KREITMAN, M., and M. AGUADÉ, 1986 Excess polymorphism at the *Adh* locus in *Drosophila melanogaster.* Genetics **114:** 93–110.

KREITMAN, M., and R. R. HUDSON, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

LEUTWILER, L. S., E. M. MEYEROWITZ and E. M. TOBIN, 1986 Structure and expression of three light-harvesting chlorophyll a/b binding protein genes in *Arabidopsis thaliana.* Nucleic Acids Res. **14:** 4051–4065.

LEWONTIN, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

LEWONTIN, R. C., and J. L. HUBBY, 1966 A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura.* Genetics **54:** 595–609.

MACARTHUR, R. H., 1957 On the relative abundance of bird species. Proc. Nat. Acad. Sci. USA **43:** 293–295.

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

MILKMAN, R., 1973 Electrophoretic variation in *Escherichia coli* from natural sources. Science **1982:** 1024–1026.

NAM, H.-G., J. GIRAUDAT, B. DEN BOER, R. MOONAN, W. LOOS, B. HAUGE *et al.,* 1989 Restriction fragment length polymorphism linkage map of *Arabidopsis thaliana.* Plant Cell **1:** 699–705.

OSTERMAN, J. C., and E. S. DENNIS, 1989 Molecular analysis of the *Adh1-Cm* allele of maize. Plant Mol. Biol. **13:** 203–212.

PRICE, R. A., I. A. AL-SHEHBAZ and J. D. PALMER, 1994 Systematic relationships of *Arabidopsis*: a molecular and morphological perspective, pp. in *Arabidopsis,* edited by C. SOMERVILLE and E. MEYEROWITZ. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y. (in press).

RÉDEI, G. P., 1992 A heuristic glance at the past of *Arabidopsis* genetics, pp. 1–15 in *Methods in Arabidopsis Research,* edited by C. KONCZ, N.-H. CHUA and J. SCHELL. World Scientific Publishing Co. Pte. Ltd., Singapore.

REITER, R. S., R. M. YOUNG and P. A. SCOLNIK, 1992 Genetic linkage of the *Arabidopsis* genome: methods for mapping with recombinant inbreds and random amplified polymorphic DNAs (RAPDs), pp. 170–190 in *Methods in Arabidopsis Research,* edited by C. KONCZ, N.-H. CHUA and J. SCHELL. World Scientific Publishing Co. Pte. Ltd., Singapore.

SACHS, M. M., M. FREELING and R. OKIMOTO, 1980 The anaerobic proteins of maize. Cell **20:** 761–767.

SACHS, M. M., E. S. DENNIS, W. L. GERLACH and W. J. PEACOCK, 1986 Two alleles of maize *alcohol dehydrogenase1* have 3′ structural and poly(A) addition polymorphisms. Genetics **113:** 449–467.

SHIH, M.-C., P. HEINRICH and H. M. GOODMAN, 1991 Cloning and mapping of nuclear genes encoding chloroplast and cytosolic glyceraldehyde-3-phosphate-dehydrogenase from *Arabidopsis thaliana.* Gene **104:** 133–138.

STROBECK, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

SWOFFORD, D. L., 1993 PAUP: phylogenetic analysis using parsimony, Version 3.1. Computer program distributed by the Illinois Natural History Survey, Champaign, Illinois.

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

THOMAS, B. R., V. S. FORD, E. PICHERSKY and L. D. GOTTLIEB, 1993 Molecular characterization of duplicate cytosolic phosphoglucose isomerase genes in *Clarkia* and comparisons to the single gene in *Arabidopsis.* Genetics **135:** 895–905.

YOKOYAMA, S., and D. E. HARRY, 1993 Molecular phylogeny and evolutionary rates of alcohol dehydrogenases in vertebrates and plants. Mol. Biol. Evol. **10:** 1215–1226.

Communicating editor: A. G. CLARK