Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in *Arabidopsis*, *Arabis*, and Related Genera (Brassicaceae)

Marcus A. Koch,[1] *Bernhard Haubold, and Thomas Mitchell-Olds*

Department of Genetics and Evolution, Max Planck Institute of Chemical Ecology, Jena, Germany

We analyzed sequence variation for chalcone synthase (*Chs*) and alcohol dehydrogenase (*Adh*) loci in 28 species in the genera *Arabidopsis* and *Arabis* and related taxa from tribe Arabideae. *Chs* was single-copy in nearly all taxa examined, while *Adh* duplications were found in several species. Phylogenies constructed from both loci confirmed that the closest relatives of *Arabidopsis thaliana* include *Arabidopsis lyrata, Arabidopsis petraea,* and *Arabidopsis halleri* (formerly in the genus *Cardaminopsis*). Slightly more distant are the North American $n = 7$ Arabis (*Boechera*) species. The genus *Arabis* is polyphyletic—some unrelated species appear within this taxonomic classification, which has little phylogenetic meaning. Fossil pollen data were used to compute a synonymous substitution rate of 1.5×10^{-8} substitutions per site per year for both *Chs* and *Adh. Arabidopsis thaliana* diverged from its nearest relatives about 5 MYA, and from Brassica roughly 24 MYA. Independent molecular and fossil data from several sources all provide similar estimates of evolutionary timescale in the Brassicaceae.

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

Introduction

Molecular analyses of evolution and ecology benefit from the availability of model systems which provide breadth and depth of biological understanding and a valuable experimental resource with which to address conceptual issues. For example, for *Drosophila*, a large community of functional and evolutionary geneticists benefit from many experimental tools and a vast literature. Genomic studies of *Arabidopsis* now afford similar resources for the plant kingdom (Bevan and Murphy 1999; Somerville and Somerville 1999). Consequently, *Arabidopsis* and *Arabis* and their wild relatives provide an increasingly popular model system for molecular studies of ecology and evolution. Analyses of genetic variation (van Treuren et al. 1997; Karkkainen et al. 1999), comparative genetic maps (de Haan et al., personal communication), and phylogenetic studies of wild relatives are becoming available (Price, Palmer, and Al-Shehbaz 1994; Galloway, Malmberg, and Price 1998; Koch, Bishop, and Mitchell-Olds 1999; Koch, Haubold, and Mitchell-Olds 2000). Here, we use sequence data from two nuclear genes to address several aspects of the evolution of wild relatives of *Arabidopsis thaliana.*

As a step toward this *Arabidopsis-Arabis* model system, this paper has three goals: (1) to examine the robustness of inferred phylogenetic relationships using data from two nuclear genes; (2) to estimate the divergence time between *A. thaliana* and its wild relatives; and (3) to estimate synonymous substitution rates of several genes, which might provide a molecular clock useful for future estimation of divergence times of species and alleles.

The Brassicaceae (Cruciferae, or mustard family) comprises approximately 340 genera and 3,350 species, including the economically important Brassica crops and the model organism *A. thaliana.* Recently, Koch, Bishop, and Mitchell-Olds (1999) analyzed sequence data from the nuclear ribosomal DNA internal transcribed spacer (ITS) and showed that published phylogenies based on *Adh, Adc,* and plastidic loci (such as *ndhF*) are largely congruent with a comprehensive ITS phylogeny of 33 taxa representing major lineages of Arabideae. The genera *Arabidopsis* and *Arabis* are both polyphyletic—some distantly related species appear within these taxonomic classifications, which should be revised (Galloway, Malmberg, and Price 1998; Al-Shehbaz, O'Kane, and Price 1999).

The oldest fossil evidence of Brassicaceae occurs in Oligocene deposits (22–34 MYA; Cronquist 1981). Yang et al. (1999) combined sequence data from NADH subunit 4 (*nad4*) with the estimated divergence time between maize and wheat to infer the age of the *Brassica-Arabidopsis* divergence at approximately 14–20 Myr. Since *Brassica* and *Arabidopsis* lineages separated fairly early in crucifer evolution (Koch, Bishop, and Mitchell-Olds 1999; Koch, Haubold, and Mitchell-Olds 2000), this provides congruent fossil and molecular evidence suggesting that early crucifer evolution occurred on the order of 30 MYA. Here, we seek independent estimates for species divergence times within the Brassicaceae, especially between *A. thaliana* and its close relatives.

Such analyses also estimate K_s, the synonymous nucleotide substitution rate. Plausible estimates for the rate of nucleotide substitution can be used to address many questions of evolutionary interest. For example, K_s estimates for *Adh* have been used to calculate times for speciation and domestication of maize (Hilton and Gaut 1998), the time for a surge of retrotransposition that doubled the size of the maize genome during the last several million years (SanMiguel et al. 1998), and the age of molecular polymorphisms at the *A. thaliana*

[1] Present address: Institute of Botany, Department of Systematic Botany and Geobotany, University of Agricultural Science, Vienna, Austria.

Key words: *Adh, Arabidopsis, Arabis, Chs,* divergence time, phylogeny.

Address for correspondence and reprints: Thomas Mitchell-Olds, Department of Genetics and Evolution, Max Planck Institute of Chemical Ecology, Carl-Zeiss-Promenade 10, 07745 Jena, Germany. E-mail: tmo@ice.mpg.de.

Mol. Biol. Evol. 17(10):1483–1498. 2000
© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Adh locus (Innan et al. 1996). On the other hand, synonymous substitution rates may be heterogeneous within and among loci and may not be neutral (Comeron, Kreitman, and Aguade 1999; Llopart and Aguade 1999).

In this study of crucifer evolution, we used two well-known nuclear genes encoding the enzymes chalcone synthase (CHS; EC 2.3.1.74) and alcohol dehydrogenase (ADH; EC 1.1.1.1) as molecular marker loci. Chalcone synthase participates in plant secondary metabolism by catalyzing the condensation of three molecules of malonyl-CoA and one molecule of p-coumaryl-CoA to yield chalcone, a precursor in biosynthesis of flavonoids. In contrast, alcohol dehydrogenase is part of the primary metabolism and catalyzes the reduction of acetaldehyde to ethanol under anoxia. Both genes are members of multigene families in some plant taxa (Gaut and Clegg 1991; Clegg, Cummings, and Durbin 1997). However, previous analysis of *Adh* in Brassicaceae provided little evidence for gene duplication in crucifers, with the exception of the genus *Leavenworthia* (Charlesworth, Liu, and Zhang 1998). In the absence of sequence information from closely related taxa, it is not possible to decide whether the gene duplication in *Leavenworthia* occurred before or after the origin of this taxon. *Leavenworthia* belongs to subtribe *Cardaminae*. If the duplication is older than *Leavenworthia,* it should also be found in related genera from subtribe *Cardaminae* of tribe *Arabideae,* e.g., *Cardamine, Rorippa,* or *Nasturtium.*

Chs belongs to multigene families in plant taxa such as petunias (Koes, Spelt, and van den Elzen 1989), *Ipomoea* species (Durbin et al. 1995), and legumes (Ryder et al. 1987; Wingender et al. 1989; An et al. 1993; Junghans, Dalkin, and Dixon 1993; Howles, Arioli, and Weinman 1995). Different gene copies indicating multiple paralogs for *Chs* have been reported for *Sinapis* (Durbin et al. 1995). However, like its close relatives *Brassica* and *Raphanus* (Warwick and Black 1991), there is strong evidence that *Sinapis* consists of ancient polyploids (Sadowski et al. 1996; Cavell et al. 1998). In *A. thaliana,* which has a diploid genome without extensive duplication, *Chs* is thought to be single-copy (Cain et al. 1997). With only limited gene duplication, both *Adh* and *Chs* are candidate loci for construction of gene trees for *Arabidopsis* and its relatives.

Materials and Methods

Plant Material

The majority of taxa sequenced in this study belonged to the genera *Arabidopsis* and *Arabis* and are listed in table 1. Several taxa from related genera were added, a process guided by phylogenetic analysis based on ITS DNA sequence variation (Koch, Bishop, and Mitchell-Olds 1999). In most cases, plant material was used from individuals which were self-pollinated for two to three generations in order to obtain homozygous plants. Plants from North American *Arabis* came from wild-collected seeds, but these species are largely selfing in nature. The selfing approach was not possible for *Cardamine amara, Arabidopsis petraea, Arabidopsis lyrata,* and *Arabidopsis halleri* because of the self-incompatibility system. Two individuals per population or seed sample were analyzed. In the case of *Arabidopsis blepharophylla,* only one individual has been investigated. For *Arabis alpina* populations from Europe and Africa, we analyzed 10 individuals per population, resulting in 30 individuals in total. Detailed accession data for these taxa were published in Koch, Bishop, and Mitchell-Olds (1999) and can be viewed at http://Vanilla.ice.mpg.de/haubold/koch00.html.

DNA Amplification and Sequencing

Total DNA was obtained from leaf tissue from single individuals by a modified CTAB procedure (Mummenhoff and Koch 1994). Polymerase chain reaction (PCR) was carried out using a Robocycler Gradient 96 (Stratagene), and conditions were optimized for each primer pair. The PCR cycling scheme was 5 min at 95°C; 35 cycles of 1 min at 95°C, 1 min at 51–60°C (depending on primer combination), and 2 min at 72°C; 15-min extension at 72°C, and a final hold at 4°C. The oligonucleotide sequences used to amplify coding sequences for *Adh* were ADH-FOR1 (5′-accaccggacagattattcg-3′) in exon1 and ADH-REV7 (5′-tcaagcacccatggtgatgatgc-3′) in exon 7, and those used for *Chs* were CHS-FOR1 (5′-cttcatctgcccgtccatctaacc-3′) (promoter specific) and CHS-REV5 (5′-ggaacgctgtgcaagac-3′) in exon 2. Primer locations are shown in figure 1.

PCR reactions (50 μl) were performed under the following conditions: 50 ng template DNA, 2 ng/μl primer, 2.5 mM $MgCl_2$, and 2 U/50μl Taq DNA polymerase. All PCR products were purified from an agarose gel using the Boehringer PCR product purification kit and cloned either into the pGEM-T cloning vector (Promega) or into the TA kit pCR II cloning vector (Invitrogen). From each DNA sample, we performed three independent PCR reactions. From each PCR reaction, three independent cloned PCR products were sequenced separately to detect possible DNA sequence variation. Because we worked in most cases with highly inbreeding taxa, we expected to find little allelic variation among these nine clones (at least in diploid taxa).

For each gene, both strands were cycle-sequenced using the Taq DyeDeoxy Terminator Cycle Sequencing Kit (ABI Applied Biosystems). Products of the cycle-sequencing reactions were run on an ABI 377XL automated sequencer (ABI Applied Biosystems). Cloned PCR products were sequenced using universal t7 forward (5′-gtaacgatttaggtgacactatcg-3′) and M13-48 reverse (5′-agcggataacaatttcacacagga-3′) primers. Additional internal primers were designed for *Adh* (each in both orientations) (ADH-FOR2 [REV2] [5′-atcaagattctcttcacttc-3′], ADH-FOR3 [REV3] [5′-acatgtgtgatcttctcagg-3′], ADH-FOR4 [REV4] [5′-gttgtggtttatccactggttag-3′], ADH-FOR5 [REV5] [5′-aagaaaggtcaaagtgttgc-3′], and ADH-FOR6 [REV6] [5′-gccatgattcaagcatttgaatg-3′]) and for *Chs* (each in both orientations) (CHS-FOR2 [REV2] [5′-gaccgacctcaaggagaag-3′], CHS-FOR3 [REV3] [5′-cgtggtggtcgaagtccctaagct-3′], and CHS-FOR4 [REV4] [5′-gactgga-

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

Table 1
Accession Data for Taxa Under Study

Species[a]	Chromosome No.[b]	Locus	GenBank Accession No.
Aethionema grandiflora (Botanic Garden, Jena, Germany)	2n = 2x = 12	*Chs*	AF112082
Arabidopsis griffithiana	2n = 4x = 32	*Chs-1*/*Chs-2*	AF112092, AF112093
		Adh1	AF110440
Arabidopsis halleri ssp. *halleri*	2n = 2x =16	*Chs*	AF112095
		Adh1	AF110442
Arabidopsis lyrata ssp. *petraea* Germany	2n = 2x = 16	*Chs*	AF112103
		Adh1	AF110452
A. lyrata ssp. *petraea* SWEDEN	2n = 2x = 16	*Chs*	AF112104
		Adh1	AF110453
Arabidopsis lyrata	2n = 2x = 16	*Chs*	AF112100
		Adh1	AF110449
Arabidopsis thaliana (C24 like)	2n = 2x = 10	*Chs*	AF112086
		Adh1	AF110456
Arabis alpina EUROPE	2n = 2x = 16	*Chs*	AF112084
		Adh1	AF110425
A. alpina AFRICA1	2n = 2x = 16	*Chs*	AF112083
		Adh1	AF110428
A. alpina AFRICA2	2n = 2x = 16	*Chs*	—
		Adh1	AF110426, AF110427
Arabis blepharophylla	2n = 2x = 16	*Chs*	AF112087
		Adh1	AF110431
		Adh2	AF110433
		Adh3	AF110432
Arabis divaricarpa	2n = 2x = 14	*Chs*	AF112090
		Adh1	AF110438
Arabis drummondii MULE[c]	2n = 2x = 14	*Chs*	AF112088
		Adh1	AF110436
A. drummondii TETON[c]	2n = 2x = 14	*Chs*	AF112089
		Adh1	AF110437
Arabis glabra	2n = 2x = 12	*Chs*	AF112091
		Adh1	AF110439
Arabis hirsuta EUROPE	2n = 4x = 32	*Chs*	AF112096
		Adh1	AF110442
		Adh2	AF110444, AF110445
Arabis jaquinii	2n = 2x = 16	*Chs*	AF112097
		Adh1	AF110446
Arabis lignifera	2n = 2x = 14	*Chs*	AF112098
		Adh1	AF110447
Arabis lyalli	2n = 2x = 14	*Chs*	AF112099
		Adh1	AF110448
Arabis parishii	2n = 2x = 14	*Chs*	AF112101
		Adh1	AF110450
Arabis pauciflora	2n = 2x = 14	*Chs*	AF112102
		Adh1	AF110451
Arabis procurrens	2n = 4x = 32	*Chs*	AF112105
		Adh1	AF110454
		Adh2	AF110455
Arabis turrita	2n = 2x = 16	*Chs*	AF112107
		Adh1	AF110457
Aubrieta deltoidea	2n = 2x = 16	*Chs*	AF112109
		Adh1	AF110425
Barbarea vulgaris	2n = 2x = 16	*Chs*	AF112108
		Adh1	AF110458
Brassica oleracea (Helgoland, Germany)	2n = "6x" = 18	*Adh1*	AF110434
Capsella rubella	2n = 2x = 16	*Chs*	AF112106
		Adh1	AF110435
Cardamine amara	2n = 2x = 16	*Chs*	AF112085
		Adh1	AF110430
Halimolobus perplexa var. *lemhiensis*	2n = 2x = 14	*Chs*	AF112094
		Adh1	AF110441

[a] Additional accession data are available on request and summarized in Koch, Bishop, and Mitchell-Olds (1999).
[b] Most chromosome numbers of identical accessions and individuals are reported in Koch, Bishop, and Mitchell-Olds (1999).
[c] *Arabis drummondii* MULE and *A. drummondii* TETON represent two different populations.

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

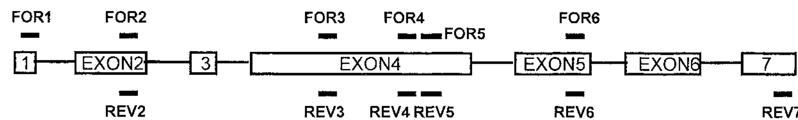

FIG. 1.—Structure of *Chs* and *Adh* genes and positions of the primers used for this study.

actccctcttctgga-3′]). Approximate primer positions are shown in figure 1.

Isozyme Electrophoresis

Native acrylamide gel electrophoresis was performed by the method described in Koch, Huthmann, and Hurka (1998) using a lithium hydroxide-borate acid electrode buffer and a Tris citric acid gel buffer (Scandalios 1969). Electrophoresis was carried out for 8 h at 75 V. Gels were stained for alcohol dehydrogenase activity as described by Soltis et al. (1983).

We used the GCG software package (Wisconsin Package, version 9.1-unix 1997, Genetics Computer Group, Madison, Wis.) to estimate isoelectric points of *Adh* proteins from DNA sequence information and compared those values with the observed migration rates.

Data Analysis
Phylogenetic Analysis

Introns and promoter regions were removed manually, and the remaining coding sequences were aligned using CLUSTAL V (Higgins, Bleasby, and Fuchs 1992). The lengths of the resulting alignments were 1,143 bp and 1,188 bp for *Adh* and *Chs,* respectively. Phylogenetic distances were computed using Kimura's (1980) two-parameter model, and the resulting distance matrices were subjected to the neighbor-joining algorithm as implemented in PHYLIP (Felsenstein 1995). One thousand bootstrap samples were analyzed to assess confidence of nodes on the original neighbor-joining tree. Parsimony analysis was conducted with unordered Fitch parsimony. The analyses were run using PAUP 4.0* beta version (Swofford 1999) under HEURISTIC, TBR, and STEEPEST DESCENT with random addition of taxa. The bootstrap option of PAUP (1,000 replicates) and a decay analysis (Donoghue et al. 1992) were used to assess relative support in the unweighted analysis.

Trees were rooted using *Aethionema grandiflora* as an outgroup in the data sets. The family Brassicaceae is a well-defined family (Schulz 1936), and previous studies indicate that the genus *Aethionema* is the sister taxon to the rest of the family (Zunk et al. 1996; Galloway, Malmberg, and Price 1998). For the overall *Adh* analysis, including several database sequences, we used *Brassica oleracea* as an outgroup because the *A. grandiflora Adh* sequence was assembled from three cloned genomic fragments and might not represent one single locus.

Adh sequence analysis was carried out in two different ways. First, only those taxa for which *Chs* sequence data were also available were analyzed. Second, we added all published *Adh* sequences from crucifers to estimate a comprehensive evolutionary tree (Miyashita et al. 1998 [*A. thaliana* ecotypes Ci-0 and Landsberg, *A. lyrata* ssp. *kawasakiana, Arabidopsis suecica, Arabidopsis korshinskii, Arabidopsis griffithiana adh1-2, Arabis glabra adh1-2, Arabidopsis wallichii, Arabidopsis himalaica, Arabis hirsuta* ASIA, *Arabis stelleri*]; Charlesworth, Liu, and Zhang 1998 [alleles from *Leavenworthia stylosa* as representatives for their *Adh1, Adh2,* and *Adh3*]; Miyashita, Innan, and Terauchii 1996 [*Arabidopsis halleri* ssp. *gemmifera*]).

Estimation of Substitution Rates

Overall substitution rates were computed using Kimura's (1980) two-parameter model. Rates of synonymous (K_s) and nonsynonymous (K_a) substitutions were calculated according to Li's method (1993) as implemented in the li93 program (Wolfe 1993).

In order to obtain the number of mutations per year, the divergence time between *C. amara* and *Barbarea vulgaris* was set to 6×10^6 years. Hence, the number of synonymous mutations per year was calculated as $\mu_s = K_s(B.\ vulgaris,\ C.\ amara)/(2 \times 6 \times 10^{-6})$, where K_s is the synonymous substitution rate calculated according to Li (1993). A 95% confidence interval for μ_s was computed using the following simulation approach: A random K_s value was drawn from a normal distribution with mean K_s and standard deviation SD(K_s[*B. vulgaris, C. amara*]), where the standard deviation of K_s was determined according to Li (1993). The random K_s value was divided by two times a random variable drawn from a normal distribution with mean 6×10^6 and standard

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

deviation $10^6/1.96$. The latter distribution models our conviction of being 95% sure that the divergence time for *B. vulgaris* and *C. amara* is between 5 and 7 Myr, with a mean of 6 Myr. This procedure was repeated 1,000 times, the resulting mutation rates were sorted, and the top and bottom 2.5% of the distribution were removed to obtain a confidence interval for μ_s. An analogous simulation procedure was used to calculate a confidence interval around the divergence time for a given pair of taxa.

The significance of correlations between matrices of substitution rates or ratios of terminal branch lengths was estimated using the Mantel permutation test (Mantel 1967). The order of taxa in the input matrices was permuted 10,000 times, and a correlation coefficient was computed for each new comparison. Two-tailed statistical significance was calculated as the frequency of obtaining a correlation coefficient whose absolute value was greater than or equal to the absolute value of the original point estimate (Manly 1994).

Test for Recombination

Recombination was investigated by the maximum chi-squared method of Maynard Smith (1992) as implemented by Ross (1997). In this approach, one examines the distribution of segregating sites between two putative recombinant haplotypes using a sliding point that partitions the alignment into two regions. The observed distribution of segregating sites is then compared with the distribution expected if the segregating sites were randomly distributed, in order to find the point of maximum discrepancy between random and observed distribution. This represents a putative breakpoint for homologous recombination (Maynard Smith 1992). Only polymorphisms at third codon positions were included in the analysis.

Results

Chalcone Synthase

Genomic PCR products included a large proportion of the 5′ promoter region which has been intensively analyzed in *A. thaliana* (Feldbrügge et al. 1997 and references therein). For all isolated DNA clones, we found structurally identical promoter regions, a good indication that our primers amplified one specific *Chs* locus corresponding to *A. thaliana Chs.* Additional functional analysis of these promoters (unpublished data) showed that they all have similar light-regulated expression patterns as expressed in *A. thaliana.* Since we mostly worked with inbreeding, diploid taxa and never obtained multiple PCR products, it is probable that only orthologous copies were isolated. In contrast, polyploidization has generated duplicate loci in tetraploid taxa. However, at least *Arabis procurrens* and *A. hirsuta* are thought to be of recent autoploid origin (Manton 1937; Titz 1970, 1976, 1978), leading to very similar gene copies at putative duplicated loci. The history of polyploidization in tetraploid *A. griffithiana* is unknown. Therefore, the two *Chs* copies identified might have descended from either an old autotetraploid or a more recent allopolyploid origin.

Of the 28 accessions sequenced, only *A. griffithiana* yielded two *Chs* sequences (fig. 2). Because *A. griffithiana* plants were propagated over three generations via single-seed descend, the material we used should be homozygous. Moreover, the stock center from which we obtained seeds of *A. griffithiana* propagates these plants via selfing. The two distinct sequences differed at 2.00% of their nucleotides. This level of sequence divergence, combined with the fact that *A. griffithiana* is tetraploid, suggests that these *Chs* sequences may represent two duplicated loci.

Putative allelic variation was detected in *Arabis drummondii* populations from Mule and Teton, with two different *Chs* sequences ($\pi = 0.51\%$, Kimura two-parameter distance). Similarly, the German and Swedish populations of *A. petraea* harbored distinct *Chs* sequences ($\pi = 0.76\%$), as did populations of *A. alpina* from Africa and Europe ($\pi = 0.48\%$). Finally, we conducted a pilot study of allelic diversity in *Chs* in 10 individuals from three populations of *A. alpina* from Africa and Europe but found no genetic variation.

Alcohol Dehydrogenase

For *Adh,* we were not able to design primers located in the 5′ promoter region. Therefore, in contrast to *Chs,* orthology could not be inferred based on promoter similarity. Hence, we inferred whether sequences were likely to be orthologous or paralogous based on breeding system, history of self-pollination, ploidy level, and degree of divergence among sequences. When two sequences are very similar ($\pi < 1.0\%$), they are likely to represent alleles at a single locus, whereas higher levels of divergence ($\pi >$ a few percent) typically characterize diverged loci within a gene family. Between these extremes, additional experimentation is required to distinguish allelic polymorphism from locus duplication. Although we found a few examples of species having several moderately diverged sequences, in these instances, a clear distinction between alleles and loci is not central to the research goals of this paper. We refer to such observations as ''genes'' or ''sequences,'' so as not to imply whether they actually represent alleles or loci.

Arabis alpina harbored the highest number of *Adh* sequences, with three putative alleles ($\pi = 0.76\%$) in the two African populations studied (fig. 3). Among the 10 individuals investigated for population AFRICA2, 3 individuals had the allele *Adh1-1* (AF110426; table 1), and the remaining 7 individuals carried the allele *Adh1-2* (AF110427). A third allele was detected among individuals from the European population. The two populations of *A. drummondii* and of *A. petraea* that carried different *Chs* genes also had different *Adh* sequences ($\pi = 2.46\%$ and 0.97%, respectively).

For some European Arabis (*A. procurrens* and *A. hirsuta*) and for North American *A. blepharophylla,* we found three structurally different *Adh* sequences. These taxa form a monophyletic clade on the basis of a molecular analysis of ITS sequence variation (Koch, Bish-

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

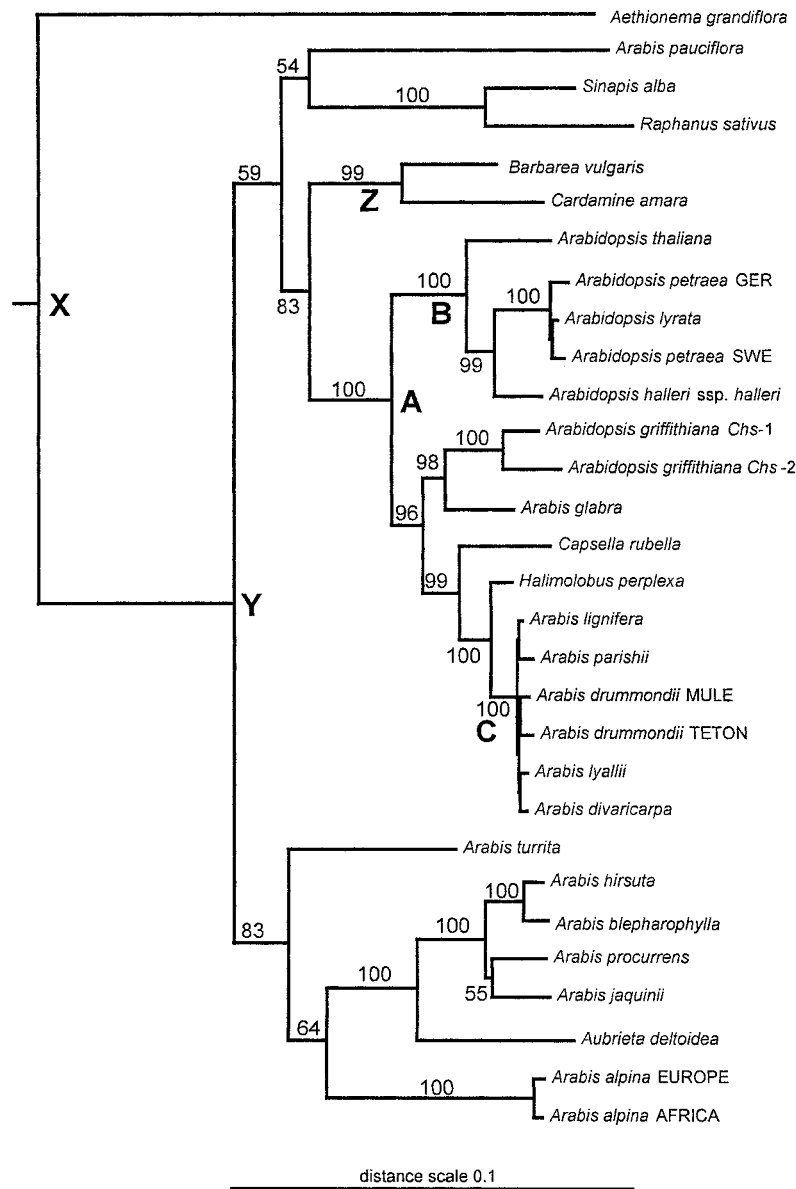

FIG. 2.—Neighbor-joining distance tree based on *Chs* sequences. Bootstrap support is given along the branches. Estimates for ages of the ancestral nodes appear in table 4 and in the text.

op, and Mitchell-Olds 1999). Apart from normal *Adh1* with six introns, we found additional *Adh* loci in the same individuals lacking all introns (*Adh2*) and one *Adh* sequence lacking introns 4, 5, and 6 (*Adh3*). These sequences probably represent additional loci. However, it should be born in mind that these loci are not orthologous to *Adh1, Adh2,* and *Adh3* of *Leavenworthia*, which also had introns (Charlesworth, Liu, and Zhang 1998; fig. 4).

Adh2 lacked all introns in *A. procurrens, A. blepharophylla,* and *A. hirsuta* from Europe. Tetraploid *A. hirsuta* had two different *Adh2* sequences (π = 0.88%) found within single individuals. The genetic distance between *Adh1* and *Adh2* of *A. hirsuta* was higher (π = 2.28%) than that between its two *Adh2* genes. In the case of *A. blepharophylla, Adh3* lacked introns 4–6 (*Adh2* vs. *Adh3*: π = 0.63%). As in *A. hirsuta,* divergence between *A. blepharophylla Adh2*/*Adh3* and *Adh1* was high (π = 2.15%), and in *A. procurrens,* the distance separating *Adh2* and *Adh1* reached π = 3.86%. In addition, our *Brassica Adh* sequence lacked intron 6, a feature not found in any other taxon analyzed, indicating an additional independent intron deletion. When published *Adh* sequences were included in the analysis, we found additional intraspecific variation within *A. glabra, A. griffithiana* (possibly representing two different loci), and *A. hirsuta* from Europe and Asia.

Duplicated *Adh* loci have previously been described by Charlesworth, Liu, and Zhang (1998), who found three distinct loci with numerous alleles from different *Leavenworthia* species. Different *Adh* loci differed in the number of introns, which is similar to our findings in *Arabis.* They refer to these loci as *Adh1* (all introns present), *Adh2* (no intron 4), and *Adh3* (missing

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

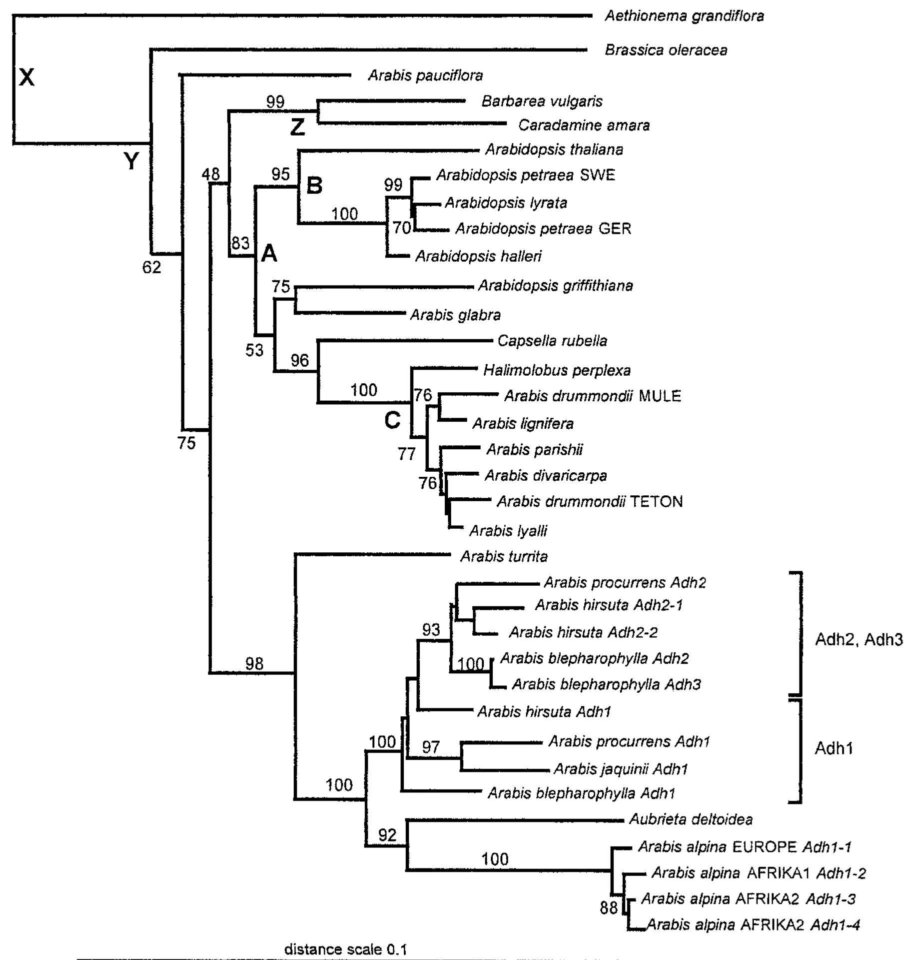

FIG. 3.—Neighbor-joining distance tree using *Adh* sequences from this study. Bootstrap support is given along the branches. Estimates for ages of the ancestral nodes appear in table 4 and in the text.

all introns). In our full *Adh* analysis, we included one allele as a representative from each locus and found that *Adh2* and *Adh3* from *Leavenworthia* differ from *Arabis Adh* loci (fig. 4). Therefore, *Adh2* and *Adh3* of *Leavenworthia* and *Arabis* are not orthologous and presumably arose by independent duplication events.

All of the *Adh* sequences appeared to code for functional proteins, since none contained stop codons or insertions/deletions. In order to further investigate the relationship between nucleotide sequence and enzyme phenotype, we subjected taxa with more than one *Adh* locus to isozyme electrophoresis. In each case, we detected only a single band. Relative to *A. thaliana* (R_f = 100), we observed migration rates of R_f = 100 for *A. hirsuta* EUROPE, R_f = 95 for *A. procurrens,* R_f = 90 for *A. jaquinii,* and R_f = 103 for *A. blepharophylla* (table 2). These single bands of *Adh* activity could indicate that there was only one highly expressed *Adh* locus or that products from more than one locus comigrated on the gel. We distinguished between these two possibilities by computing the expected migration rates for these two proteins (table 2). The difference in net charge between *A. hirsuta* allele pairs *Adh1*/*Adh2-1* and *Adh1*/*Adh2-2* was 1 and 4, respectively (table 2). Since differences in net charge of 1 or more are easily detected with our experimental protocol, activities of enzymes encoded by *Adh2-1*/*Adh2-2* in *A. hirsuta* must be low or absent. The same argument applies to *Adh2* from *A. procurrens,* which is distinguished from *Adh1* by a charge difference of 1.5. For *A. blepharophylla,* allozyme electrophoresis did not yield additional information about gene expression, as the products from all three loci had the same net charge.

Comparative Phylogenetic Reconstruction

We computed parsimony and distance trees for *Adh* and *Chs.* Of the 1,188 bp of the aligned *Chs* gene, 733 sites were invariant (61.7%); of the remaining 455 variable sites, 316 sites (26.6%) were phylogenetically informative. Fitch parsimony resulted in 10 most-parsimonious trees of 1,196 steps (consistency index [CI] = 0.55, retention index [RI] = 0.72). The strict consensus tree is shown in figure 5, along with the values from bootstrap and decay analyses. Of the 1,143 bp of the aligned *Adh* gene, 716 sites were invariable (62.6%); of the remaining 427 variable sites, 257 sites (22.5%) were informative. Fitch parsimony yielded 16 most-parsimo-

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

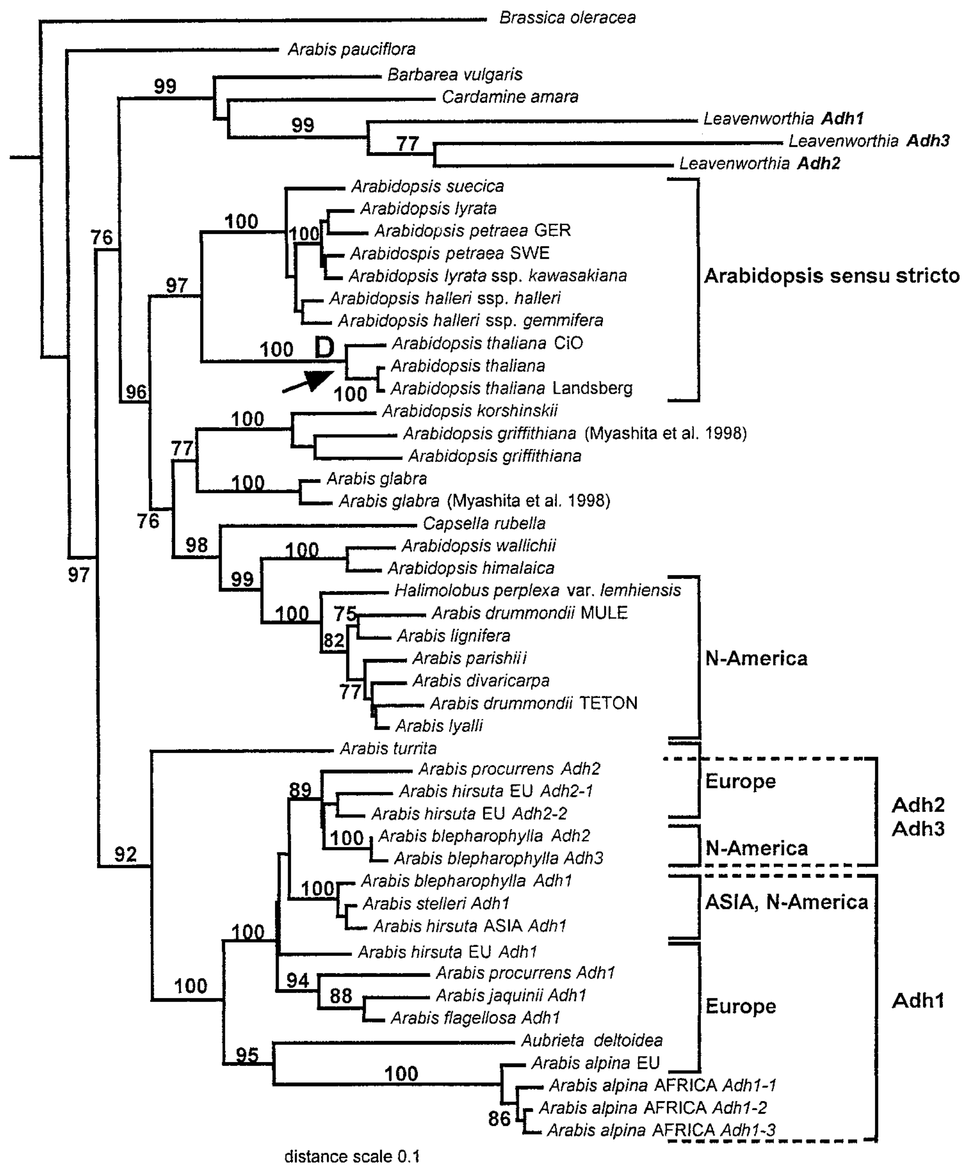

FIG. 4.—Neighbor-joining distance tree with additional crucifer *Adh* sequences from DNA databases. Bootstrap support is given along the branches. Estimates for ages of the ancestral nodes appear in table 4 and in the text.

nious trees of 1,010 steps (CI = 0.56, RI = 0.74). The strict consensus tree is shown in figure 6, along with the values from bootstrap and decay analyses.

In both cases, the strict consensus tree from parsimony analysis was congruent with the phylogeny computed with the neighbor joining algorithm. The distance trees for *Chs* and *Adh* are shown in figures 2 and 3 to demonstrate relative branch length.

The gene phylogeny based on *Adh* showed four major groups of taxa: (1) *B. oleracea* as the most basal group, followed by (2) *Arabis pauciflora,* followed by the two large crown clades containing (3) *A. thaliana* and (4) *A. alpina* (figs. 4 and 6). In contrast, the gene phylogeny based on *Chs* contained only two of these major clades, with *Sinapis alba* and *A. pauciflora* now constituting a clade with *A. thaliana* (figs. 2 and 5). *Sinapis alba* and *Raphanus sativus* are very closely related to *B. oleracea* (Warwick and Black 1991, 1997). Notice, however, that on both phylogenetic trees the bootstrap support and decay values for the basal groups were low, indicating that the trees were not highly resolved at this level of the analysis.

In addition to the disagreement about the branching order near the root, there was one difference between crown groups. According to *Adh, A. deltoidea* groups with *A. alpina,* away from *A. jaquinii* (fig. 6). In contrast, according to *Chs, A. deltoidea* groups with *A. jaquinii,* away from *A. alpina* (fig. 5). This incongruence is well supported by bootstrapping and decay analysis on both trees and may indicate lineage sorting, convergent evolution, or an ancient intergenic recombination event (Syvanen 1994).

Intergenic recombination might have occurred if *A. deltoidea* arose by hybridization between the ancestors

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

Table 2
Results from Native Acrylamide Gel Electrophoresis to Compare Putative Allozymes with Respect to Calculated Migration Behavior (estimated from DNA sequences) and Observed Migration Distances

Taxon	Isoelectric Point	Net Charge (pH 8.0)	R_f Value	
			Observed	Expected[a]
Arabis blepharophylla Adh1	5.81	−13.5	103	~**100**
A. blepharophylla Adh2	5.81	−13.5	103	*[b]
A. blepharophylla Adh3	5.81	−13.5		*[b]
Arabis hirsuta Adh1	6.07	−13	100	~**100**
A. hirsuta Adh2-1	5.66	−14	100	>100
A. hirsuta Adh2-2	5.24	−17	100	≫100
Arabis jaquinii	6.21	−12	90	~90
Arabis procurrens Adh1	6.25	−12.5	95	~**100**
A. procurrens Adh2	5.96	−14	95	>100
Arabidopsis thaliana	6.10	−13	94	~94
Arabis turrita	6.44	−9.5	67	~70
Barbarea vulgaris	6.16	−12.5	92	~92
Cardamine amara	6.39	−11	83	~80

[a] Those isozymes with the highest probability of being expressed and translated as active proteins are listed in bold type.
[b] Not distinguishable from *A. blepharophylla* isozymes from the *Adh1* locus.

of *A. alpina* and other European *Arabis* species. During constitution of *A. deltoidea,* intragenic recombination might have occurred between parental *Chs* copies, giving rise to a recombinant *Chs* sequence in *A. deltoidea.* Because only a minor part of the gene (exon 1 with 294 bp, approximately 25% of the entire coding region) resembles *A. alpina,* whereas the major part (exon 2) resembles progenitors of other European *Arabis, A. deltoidea* clusters in phylogenetic analysis basal to *A. procurrens* and its relatives. We investigated this further using the maximum chi-squared method (Maynard Smith 1992) to search for recombination points between sequences. No breakpoints were found in *Adh* (data not shown). However, when the *Chs* sequences for *A. deltoidea* and *A. alpina* were compared, a single potential insertion point was found between nucleotides 294 and 297 ($P = 0.002$). The same putative recombination point was identified by comparisons between either of the two *A. alpina* genes and members of the *A. deltoidea* clade (*A. deltoidea, A. jaquinii, A. procurrens, A. blepharophylla,* and *A. hirsuta*). Sliding-window analyses of the distribution of substitutions between pairs of *Chs* sequences from the *A. alpina* and *A. deltoidea* clades revealed a region of high diversity at the 3′ end of the gene (data not shown).

Biogeography and Genetic Variation

For some populations, the distribution of *Adh* genes was correlated with geographic origin. For *A. alpina,* all four sequences grouped to a single clade. It has previously been speculated that African populations of *A. alpina* have been separated from European populations since the Tertiary (Plantholt 1995). The *Adh1* genes of European *Arabis* species were distinct from those of Asian and North American taxa, as indicated by high bootstrap support (fig. 6). This included *Adh1* from Asian *A. hirsuta,* which differed from the *Adh1* gene found in the corresponding European population. This taxon might bridge the European–North American disjunction via Asia and is reported to be distributed in the United States with numerous morphologically divergent subspecies (Rollins 1941). No geographic structuring of genetic variation was found for *Arabidopsis lyrata* from the United States and Asia or *A. petraea* from Europe. These two taxa are thought to be conspecific (Koch, Bishop, and Mitchell-Olds 1999), and these *Adh* sequences were nested among one another (fig. 3). The intercontinental disjunction of *A. blepharophylla* in North American versus European *Arabis* was correlated with the presence of a new *Adh3* locus, lacking introns 4–6, in *A. blepharophylla.* This intronless *Adh* gene was not observed in any European *Arabis. Adh3* of *A. blepharophylla* is most closely related to *Adh2* from the same species.

Relative Substitution Rates

The mean numbers of substitutions per base were 0.087 for *Adh* and 0.111 for *Chs.* If substitutions are a function of time only, the numbers of substitutions in *Adh* and *Chs* should be perfectly correlated. In accordance with this expectation, the correlation between substitution matrices for *Adh* and *Chs* was 0.87 ($P < 10^{-4}$). However, gene phylogenies for *Adh* and *Chs* also suggest possible heterogeneity of substitution rates (figs. 2 and 3). Therefore, we tested the null hypothesis of equal rates of evolution. To examine rate heterogeneity between two terminal branches, let

$$\hat{r}_{ij} = \frac{d_{io} + d_{ij} - d_{jo}}{d_{jo} + d_{ij} - d_{io}}, \qquad (1)$$

where $\hat{r}_{ij}$ is the ratio of the lengths of two terminal branches, d_{ij} is the distance between the *i*th and the *j*th taxa, and d_{io} is the distance between the *i*th taxon and the outgroup (Gaut et al. 1996), in this case *A. grandiflora.* We used the method of Muse and Gaut (1994) to test the null hypothesis that the two terminal branches are of equal length, i.e., that $\hat{r}_{ij} = 1$. Notice that $\hat{r}_{ij}$ is undefined if $d_{ij} + d_{io} = d_{jo}$, i.e., as the terminal branch leading to the *j*th taxon becomes very short; notice also that $\hat{r}_{ij}$ is negative if $d_{ij} < |d_{io} - d_{jo}|$, i.e., if one of the terminal branches has a negative length. Since negative branch lengths are difficult to interpret biologically, we excluded all comparisons that returned negative values of $\hat{r}_{ij}$ in any of the four $\hat{r}_{ij}$ matrices computed for synonymous and nonsynonymous substitutions at *Adh* and *Chs* (table 3).

Chalcone Synthase

Although the ratios of branch lengths varied widely for *Chs,* none of the comparisons showed significant rate heterogeneity for synonymous substitutions (table 3). In contrast, three comparisons for nonsynonymous substitutions were significant: *A. glabra* EU/*S. alba,* African *A. alpina*/Swedish *A. petraea,* and *S. alba*/Swedish *A. petraea.* In general, it appeared that *Chs* had evolved more rapidly in *S. alba* than in other crucifers.

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

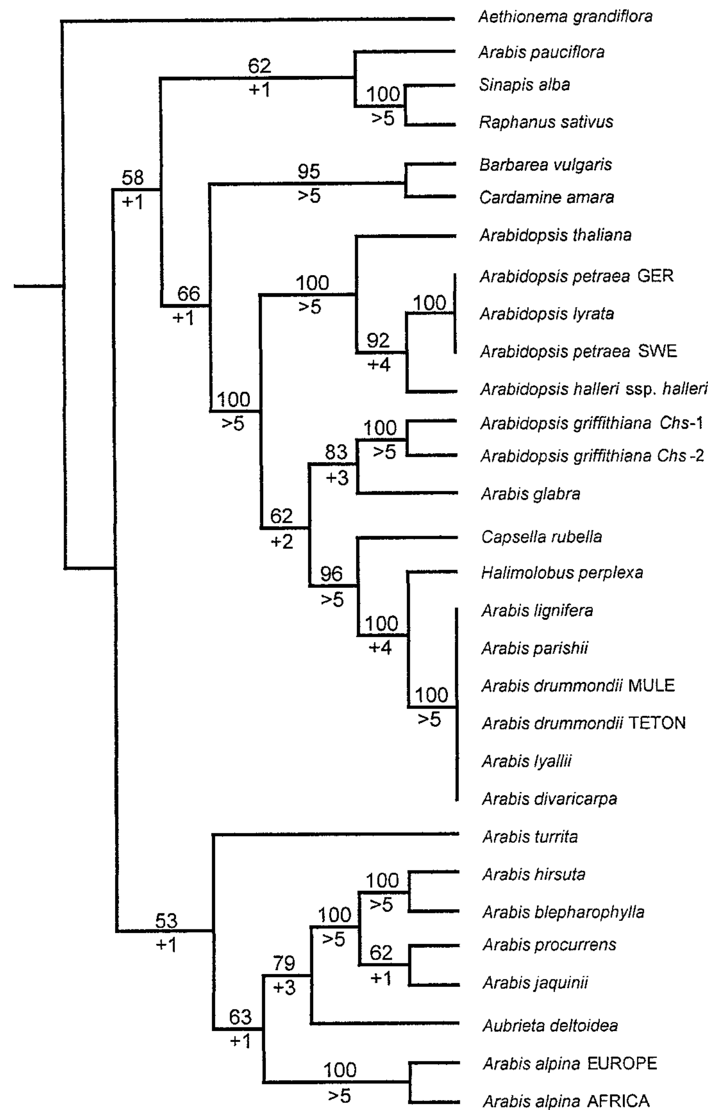

FIG. 5.—Strict consensus tree of 10 most-parsimonious trees from Fitch parsimony analysis based on *Chs* sequences. Bootstrap support (above branches) and decay indices (below branches) are indicated.

Alcohol Dehydrogenase

At synonymous sites, the relative rates of evolution were quite homogeneous (table 3), with *A. turrita* causing all three significant deviations from homogeneity. At synonymous sites in *Chs* and *Adh,* only 3 of 156 comparisons of relative synonymous rates were significant. This is <5%; hence, there is no evidence for heterogeneity of synonymous rates. In general, *A. turrita* and *A. glabra* seemed to evolve more slowly than the other taxa (table 3). At *Adh* nonsynonymous sites, 10 of the 78 comparisons were significant (table 3). In particular, *B. vulgaris* had an elevated rate of *Adh* evolution, while that of *A. procurrens* was reduced in all comparisons (table 3).

Correlating Substitution Patterns at the *Chs* and *Adh* Loci

The ratios of branch lengths provided a starting point to further probe the evolutionary dynamics at *Chs* and *Adh* by testing for correlations between the $\hat{r}_{ij}$-matrices. If the number of substitutions is a function of mutation rate, then synonymous and nonsynonymous $\hat{r}_{ij}$-matrices for a given locus should be correlated. This was the case for *Adh* ($r = 0.29$, $P = 0.03$) but not for *Chs* ($r = -0.13$, $P = 0.34$), which may indicate that the substitution process has reached saturation at *Chs* but not at *Adh.* In order to test for genomewide evolutionary dynamics, the $\hat{r}_{ij}$-matrices were compared between loci, but no significant correlations were found (not shown).

How Old Is the *Arabis* Clade?

Results of the relative-rate test allowed us to focus on a subset of taxa with homogeneous rates of evolution. To estimate divergence times among these taxa, we used information from Pliocene deposits of *Rorippa* pollen to calibrate the rates of synonymous substitution for *Adh* and *Chs. Rorippa* is a close relative of *Cardamine* and

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

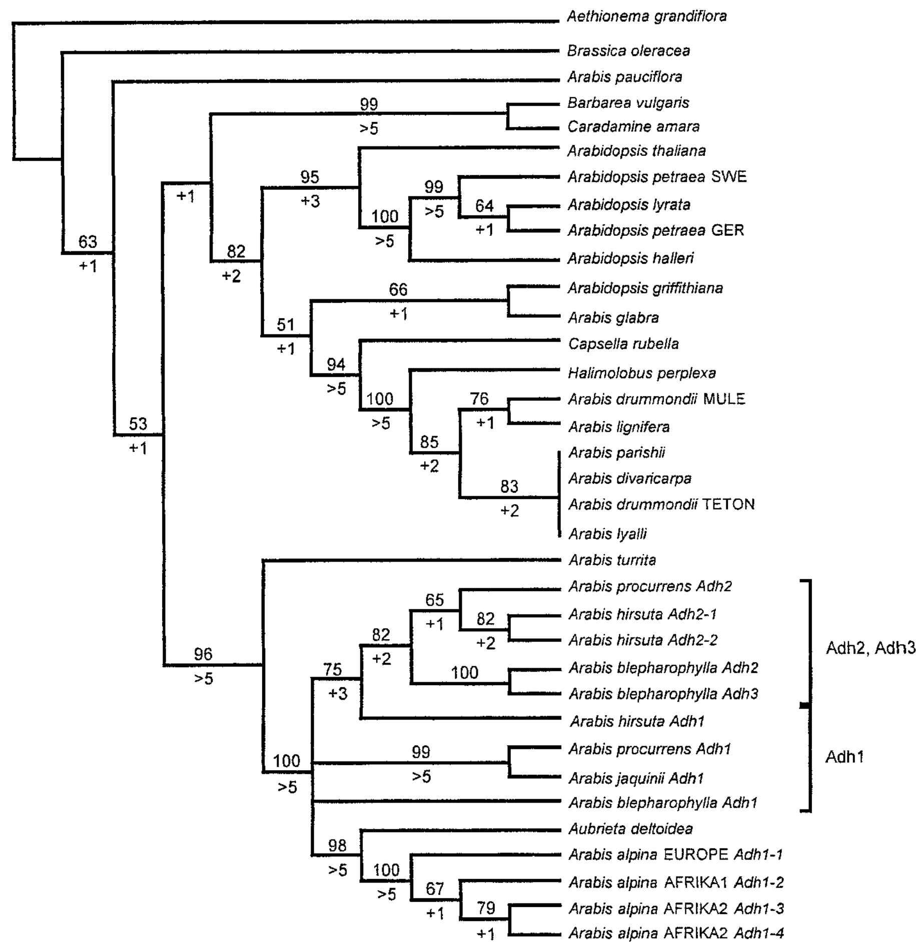

FIG. 6.—Strict consensus tree of 16 most-parsimonious trees from Fitch parsimony analysis using *Adh* sequences from this study. Bootstrap support (above branches) and decay indices (below branches) are indicated.

Barbarea (Franzke et al. 1998). Mai (1995) cited extensive *Rorippa* pollen deposits in geological samples from the Pliocene (2.5–5.0 MYA). It follows that *Barbarea* and *Cardamine* diverged before the Pliocene, about 6 MYA (node Z; figs. 2 and 3). Assuming that *Barbarea* and *Cardamine* diverged 6.0 MYA (node Z), we estimated synonymous substitution rates and their 95% confidence intervals as $1.0 \times 10^{-8} < 1.5 \times 10^{-8} < 2.0 \times 10^{-8}$ and $9.9 \times 10^{-9} < 1.5 \times 10^{-8} < 2.1 \times 10^{-8}$ mutations per site per year for *Chs* and *Adh,* respectively. Therefore, the last common ancestor of *A. thaliana* and North American *Arabis* is between 9.0 and 11.0 Myr old (node A; table 4). Furthermore, *A. thaliana* diverged from its closest congeners about 5.1–5.4 MYA (node B; table 4), the North American *Arabis* diverged about 0.8–2.2 MYA (node C; table 4), and segregating alleles of *Adh* in *A. thaliana* diverged 1.5 MYA based on data from the Landsberg and Columbia ecotypes (node D in fig. 4; table 4).

We also estimated that the age of the last common ancestor of the crucifers excluding *A. grandiflora* is between 23.1 and 25.9 Myr (node Y; table 4). Notice, however, that node Y connects different taxa in the *Adh* and *Chs* trees.

Discussion

Overview of *Arabidopsis* Relatives

Our analysis of two nuclear genes verifies previous work regarding the closest relatives of *Arabidopsis* and several more distant clades. Data from *ITS, Adh, Chs, NdhF, matK,* and *Adc* confirm that the closest known relatives of *A. thaliana* include outcrossing diploid species (*A. lyrata, A. petraea,* and *A. halleri*) and several other taxa previously classified in the genus *Cardaminopsis* (Price, Palmer, and Al-Shehbaz 1994; Galloway, Malmberg, and Price 1998; Koch, Bishop, and Mitchell-Olds 1999). Slightly more distant is a strongly supported clade containing the North American $n = 7$ *Arabis* (or *Boechera*) species. The evolutionary relationships of these species groups are supported by several independent analyses; hence, the basic features of these phylogenies are established with a high level of confidence. The genus *Arabis* is polyphyletic—some species that are very distantly related appear within this taxonomic classification, which has little phylogenetic meaning. Comparison of data from several genes cannot determine the relative branching order of the *Arabidopsis, Brassica,* and *A. alpina* clades, which diverged early in the history of the Brassicaceae.

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

Table 3
Results of Relative-Rate Tests

	1	2	3	4	5	6	7	8	9	10	11	12	13
Chs, synonymous substitutions													
1. *Arabidopsis thaliana*	—												
2. *Capsella rubella*	3.281	—											
3. *Barbarea vulgaris*	1.665	0.716	—										
4. *Arabis procurrens*	0.927	0.539	0.663	—									
5. *Arabis glabra* EUROPE	4.968	0.928	1.382	2.022	—								
6. *Arabis pumilla*	0.749	0.398	0.488	0.186	0.373	—							
7. *Arabis blepharophylla* . .	0.721	0.392	0.505	0.105	0.350	0.802	—						
8. *Arabis alpina* AFRICA	1.368	0.817	1.005	1.938	0.834	2.822	3.010	—					
9. *Sinapis alba*	1.380	0.695	0.908	1.389	0.716	1.626	1.757	0.925	—				
10. *Arabis turrita*	2.027	1.056	1.394	4.394	1.094	8.780	11.136	1.358	1.476	—			
11. *Halimolobus perplexa* . . .	12.121	3.577	2.371	2.395	3.237	3.460	3.542	1.501	1.830	1.281	—		
12. *Arabis lyalli* SWEDEN	5.429	0.633	0.975	1.434	0.595	1.794	1.879	0.983	1.086	0.739	0.288	—	
13. *Aubrieta deltoidea*	0.817	0.485	0.620	0.696	0.458	1.151	1.294	0.428	0.659	0.251	0.380	0.604	—
Chs, nonsynonymous substitutions													
1. *A. thaliana*	—												
2. *C. rubella*	1.162	—											
3. *B. vulgaris*	0.618	0.617	—										
4. *A. procurrens*	0.765	0.753	1.110	—									
5. *A. glabra* EUROPE	10.24	0.890	1.608	1.304	—				0.036				
6. *A. pumilla*	0.931	0.905	1.362	1.865	0.927	—							
7. *A. blepharophylla*	0.619	0.583	0.840	0.371	0.597	0.313	—						
8. *A. alpina* AFRICA	0.373	0.398	0.707	0.694	0.384	0.573	0.942	—				0.038	
9. *S. alba*	0.422	0.436	0.677	0.634	**0.436**	0.522	0.795	0.874	—			0.017	
10. *A. turrita*	0.513	0.487	0.668	0.337	0.475	0.299	0.560	0.824	0.987	—			
11. *H. perplexa*	0.755	0.457	1.465	1.223	0.746	1.016	1.558	2.248	2.076	1.894	—		
12. *A. lyalli* SWEDEN	3.246	2.155	2.336	1.740	2.271	1.427	2.115	**4.838**	**3.033**	2.575	3.487	—	
13. *A. deltoidea*	0.899	0.877	1.409	1.325	0.896	0.942	2.063	1.729	1.885	3.388	0.949	0.685	—
Adh, synonymous substitutions													
1. *A. thaliana*	—												
2. *C. rubella*	1.156	—											
3. *B. vulgaris*	1.008	0.895	—										
4. *A. procurrens*	1.561	1.362	1.571	—						0.013			
5. *A. glabra*	0.613	0.514	0.662	0.452	—								
6. *A. pumilla*	1.450	1.275	1.431	0.620	1.989	—							
7. *A. blepharophylla*	1.162	1.039	1.163	0.147	1.665	0.363	—						
8. *A. alpina* AFRICA	2.462	2.045	2.446	2.630	3.197	2.803	6.028	—		0.028			
9. *Brassica oleracea*	1.363	1.261	1.354	0.997	1.673	1.036	1.192	0.715	—				
10. *A. turrita*	0.720	0.647	0.740	**0.248**	1.042	0.275	0.456	**0.206**	0.567	—			0.029
11. *H. perplexa*	0.702	0.483	0.741	0.476	1.130	0.525	0.655	0.332	0.599	1.033	—		
12. *A. lyalli*	0.795	0.744	0.857	0.545	1.529	0.603	0.748	0.345	0.691	1.250	1.290	—	
13. *A. deltoidea*	1.890	1.577	1.806	1.525	2.488	1.767	2.674	0.671	1.172	**3.455**	2.143	2.155	—
Adh, nonsynonymous substitutions													
1. *A. thaliana*	—												
2. *C. rubella*	1.173	—											
3. *B. vulgaris*	0.642	0.445	—	0.003		0.012	0.010			0.024			0.027
4. *A. procurrens*	1.944	1.998	**3.216**	—	0.040						0.041	0.016	
5. *A. glabra*	0.815	0.629	1.431	**0.379**	—								0.020
6. *A. pumilla*	1.812	1.723	**2.805**	0.798	2.497	—						0.040	
7. *A. blepharophylla*	2.274	1.742	**3.181**	0.822	2.661	0.994	—						
8. *A. alpina* AFRICA	1.480	1.347	2.458	0.758	1.769	0.858	0.813	—					
9. *B. oleracea*	1.029	0.856	1.990	0.491	1.159	0.559	0.450	0.605	—				
10. *A. turrita*	1.415	1.237	**3.054**	0.623	1.816	0.731	0.655	0.827	1.390	—			
11. *H. perplexa*	0.927	0.773	1.589	**0.455**	1.066	0.502	0.386	0.626	0.901	0.663	—		
12. *A. lyalli*	0.724	0.494	1.352	**0.306**	0.806	**0.363**	0.349	0.547	0.730	0.486	0.832	—	
13. *A. deltoidea*	1.791	1.481	**2.701**	0.760	**2.036**	0.891	0.820	1.120	2.170	1.396	2.008	2.070	—

NOTE.—The lower triangle of each matrix shows the relative rates of the terminal branches leading to the pair of taxa concerned. For example, at *Chs,* synonymous sites in *C. rubella* evolve 3.281 times as fast as they do in *A. thaliana.* The top triangle of each matrix quotes the probability values for those relative-rate tests for which $P < 0.05$. The corresponding rate ratio is shown in bold type in the lower triangle. The rates were computed according to Li (1993), and the relative-rate test was carried out according to Muse and Gaut (1994).

Age and Substitution Rates

Estimation of substitution rates, species divergence times, and the ages of segregating polymorphisms requires a known time point from the fossil record. Pollen from close relatives of *Cardamine* and *Barbarea* is common in geological samples from the Pliocene (2.5–5.0 MYA; Mai 1995). Therefore, we assume that *Cardamine* and *Barbarea* diverged about 6.0 MYA. This pro-

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

Table 4
Divergence Computed from Synonymous Substitution Rates at the *Chs* and the *Adh* Loci

Node	Divergence Times	
	Chs	*Adh*
A	7.0 < 11.0 < 17.4	5.7 < 9.0 < 14.6
B	3.1 < 5.1 < 8.3	3.3 < 5.4 < 9.0
C	0.2 < 0.8 < 1.7	1.0 < 2.2 < 4.0
D	Not available	0.6 < 1.5 < 2.8
Y	17.5 < 25.9 < 40.2	15.3 < 23.1 < 36.8

NOTE.—Listed is the lower 95% confidence limit, followed by the mean, followed by the upper 95% confidence limit, expressed in Myr.

vides estimated substitution rates for *Chs* and *Adh,* leading to inferred divergence time for the species in this study. Using independent fossil dating and a different gene, Yang et al. (1999) obtained very similar dates for the *Brassica-Arabidopsis* divergence (see below). These similar estimates from independent analyses suggest that our estimates for substitution rates and divergence times are reasonable.

Estimates of divergence times are problematic, as they rely on the assumption of a molecular clock. Even if this assumption is justified, computation of the variance in divergence times is not straightforward (Sanderson 1998). We tested the hypothesis of uniform evolutionary rates for each pair of taxa and used only those pairs in further analyses where the null hypothesis was not rejected (table 3). In order to assess the confidence attached to our time estimates, we considered the following three sources of error: (1) uncertainty in the divergence time of the pair of reference taxa, (2) uncertainty in the K_s value of the reference taxa (*B. vulgaris* and *C. amara* in this case), and (3) uncertainty in the K_s value of the target taxa. We used a simulation approach to model the influence of all three factors simultaneously on the distribution of the diversity times. This enabled us to provide confidence intervals with our time estimates (table 4).

We estimated a synonymous substitution rate of 1.5×10^{-8} at *Chs.* This is slightly higher than the rates of 8×10^{-9} calculated by Durbin et al. (1995) for *Ipomoea* and 6×10^{-9} to 9×10^{-9} estimated by Wolfe, Sharp, and Li (1989) for maize versus barley. Our synonymous substitution rate for *Adh* (1.5×10^{-8}) is also higher than the values found by Morton, Gaut, and Clegg (1996) for palms (2.6×10^{-9}) and by Gaut et al. (1996) for grasses (7.0×10^{-9}). Using information from six nuclear genes, Wolfe, Sharp, and Li (1989) estimated the average synonymous substitution rate at 5×10^{-9} to 7×10^{-9} substitutions per site per year among members of the grass family. Despite the great uncertainty associated with dating speciation events in the fossil record, the similar dating obtained by Yang et al. (1999) and that from our analysis suggest that synonymous substitution rates for *Adh* and *Chs* in the Brassicaceae exceed estimates obtained from studies in monocots.

From the substitution rates obtained in our study, we attempted to date several important nodes on the gene trees for *Chs* and *Adh* (table 4). We estimate that the last common ancestor of *A. thaliana* and its nearest relatives occurred approximately 5 MYA, while the most recent common ancestor of polymorphic *A. thaliana Adh* alleles occurred roughly 1.5 MYA. This contrasts with the finding of Innan et al. (1996), who calculated the age of the last common ancestor of the *A. thaliana Adh* alleles at 6.3 Myr based on an assumed substitution rate of 10^{-9}. This is likely to be an underestimate, because Innan et al. (1996) included intron sequences in their analysis, whereas we calculated a synonymous substitution rate of 1.5×10^{-8} from our *Adh* data.

We estimate that the crucifer lineages analyzed in this study (excluding *A. grandiflora*) originated about 24 MYA (table 4; node Y in figs. 2 and 3). This result corresponds to findings obtained from divergence time estimates of the mitochondrial gene for NADH subunit 4 (*nad4*) among cruciferous plants by Yang et al. (1999), who calculated the average distance between the *Brassica* clade and *Arabidopsis* to be 14–20 Myr. This value broadly overlaps with our estimates for a *Brassica-Arabidopsis* split (node Y in figs. 2 and 3; notice that these nodes connect different taxa in the two trees). However, the basal branching pattern is not well supported in either tree, and branches with low bootstrap support should be collapsed. Further rough extrapolation leads to the estimation that the most recent common ancestor of *A. grandiflora* and the rest of the Brassicaceae is approximately 30–60 Myr old (node X in figs. 2 and 3), which agrees with the oldest findings of Brassicaceae in the Oligocene (22–34 MYA; Cronquist 1981).

Allelic Variation and Gene Duplication

Patterns of diversity differed strongly at *Chs* and *Adh.* In diploid taxa, *Chs* was essentially a single-copy gene, with polyploid *A. griffithiana* providing the one exception to this rule (fig. 2). In contrast, many taxa contained more than one *Adh* sequence, which differed not only in the level of nucleotide diversity, but also in intron number. This contrasts with the situation in *A. thaliana,* which only has one *Adh* locus, but agrees with the observation that many plant species have several *Adh* loci (e.g., Trick et al. 1988; Yokoyama and Harry 1993). The sequence diversity observed at *Adh* might indicate that (1) polyploidization of taxa such as *A. hirsuta* took place after speciation of European *Arabis,* and (2) locus duplication leading to *Adh2* in European *Arabis* occurred prior to or during the evolution of this group, because *Adh1* and *Adh2* are similar among the taxa studied (fig. 4).

With regard to intron number, *Adh2* lacked all introns in *A. procurrens* and European *A. hirsuta.* The first observation of a plant *Adh* locus devoid of all introns was recently made in *Leavenworthia,* which contains three *Adh* loci that differ in number of introns; *Adh3* has no introns and *Adh2* has lost intron 4 (Charlesworth, Liu, and Zhang 1998). Intron-free copies of genes that usually contain introns have probably arisen through reverse transcription involving an mRNA intermediate (Charlesworth, Liu, and Zhang 1998). Since the closest

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

relatives of *Leavenworthia, C. amara* and *B. vulgaris,* only have a single copy of *Adh,* the extra gene copies in *Leavenworthia* probably arose after the origin of this genus. All five species of *Leavenworthia* studied by Charlesworth, Liu, and Zhang (1998) have three *Adh* loci, whereas we found three loci solely in *A. blepharophylla.*

Charlesworth, Liu, and Zhang (1998) established by acetate gel electrophoresis that the intron-free copy of *Adh* in *Leavenworthia* codes for functional protein. We investigated the expression of *Adh* through polyacrylamide gel electrophoresis and found that only *Adh1* has detectable enzyme activity in *A. hirsuta* and *A. procurrens.* If the intron-free *Adh2* were a pseudogene, it would presumably accumulate deleterious mutations quickly, including internal stop codons. The fact that this has not occurred suggests that *Adh2* in *A. hirsuta* and *A. procurrens* may be expressed, albeit at levels below the sensitivity of our electrophoresis setup or in tissues or environmental conditions that were not assayed here.

Comparison Between Genes

We analyzed rates of evolution at more than one locus because comparison between rates can give insights into the evolutionary process. For example, the generation time hypothesis posits that species with short generations will have higher substitution rates than species with longer generations. This effect should influence multiple loci. Selection, on the other hand, often acts on individual loci and may cause heterogeneity of evolutionary rates among loci. Within taxa, we found no significant correlations between loci for either synonymous or nonsynonymous substitutions, indicating that species-specific evolutionary factors do not influence the evolution of these two genes.

The inferred phylogenetic position of *A. deltoidea* differed according to whether *Chs* or *Adh* was used for the input data (figs. 2 and 3). In principle, incongruencies between gene trees may be due to lineage sorting, recombination, or convergent evolution. Sorting of ancient polymorphism into different lineages is an unlikely explanation because of the nonrandom distribution of polymorphisms within the *Chs* sequences. Recombination and convergent evolution are distinguished by positing neutrality in the case of recombination and selection in the case of convergent evolution. In order to minimize the effect of selection on our analysis, we considered only the distribution of substitutions in third codon positions in the maximum χ-squared test (Maynard Smith 1992). We still found significant deviation from randomness in the distribution of the substitutions, which was best explained by recombination or gene conversion.

Acknowledgments

This work was supported by the Max-Planck-Gesellschaft and by grants to T.M.-O. from the U.S. National Science Foundation (DEB-9527725) and the European Union. We thank D. Charlesworth, E. Kellogg, and two anonymous referees for comments on the manuscript, and all institutes and colleagues who provided plants and seeds. We are grateful to D. Schnabelrauch and S. Ring for technical support and to S. Dix for expert secretarial assistance.

LITERATURE CITED

Al-Shehbaz, I. A., S. L. O'Kane, and R. A. Price. 1999. Generic placement of species excluded from *Arabidopsis* (Brassicaceae). Novon **9**:296–307.

An, C., Y. Ichinose, T. Yamada, Y. Tanaka, T. Shiraishi, and H. Oku. 1993. Structure and organization of the genes encoding chalcone synthase in *Pisum sativum.* Plant Mol. Biol. **21**:789–803.

Bevan, M., and G. Murphy. 1999. The small, the large and the wild: the value of comparison in plant genomics. Trends Genet. **15**:211–214.

Cain, C. C., D. E. Saslowsky, R. A. Walker, and B. W. Shirley. 1997. Expression of chalcone synthase and chalcone isomerase proteins in *Arabidopsis* seedlings. Plant Mol. Biol. **35**:377–381.

Cavell, A. C., D. J. Lydiate, I. A. P. Parkin, C. Dean, and M. Trick. 1998. Collinearity between a 30-centimorgan segment of *Arabidopsis thaliana* chromosome 4 and duplicated regions within the *Brassica napus* genome. Genome **41**:62–69.

Charlesworth, D., F. L. Liu, and L. Zhang. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (Brassicaceae). Mol. Biol. Evol. **15**:552–559.

Clegg, M. T., M. P. Cummings, and M. L. Durbin. 1997. The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. USA **94**:7791–7798.

Comeron, J. M., M. Kreitman, and M. Aguade. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151**:239–249.

Cronquist, A. 1981. An integrated system of classification of flowering plants. Columbia University Press, New York.

Donoghue, M. J., R. G. Olmstead, J. F. Smith, and J. D. Palmer. 1992. Phylogenetic relationships of Dipsacales based on *rbcL* sequences. Ann. Mo. Bot. Gard. **79**:333–345.

Durbin, M. L., G. H. Learn, G. A. Huttley, and M. T. Clegg. 1995. Evolution of the chalcone synthase gene family in the genus *Ipomoea.* Proc. Natl. Acad. Sci. USA **92**: 3338–3342.

Feldbrügge, M., M. Sprenger, K. Hahlbrock, and B. Weisshaar. 1997. PcMYB1, a novel plant protein containing a DNA-binding domain with one MYB repeat, interacts in vivo with a light-regulatory promoter unit. Plant J. **11**: 1079–1093.

Felsenstein, J. 1995. PHYLIP (phylogeny inference package). Version 3.57c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

Franzke, A., K. Pollmann, W. Bleeker, R. Kohrt, and H. Hurka. 1998. Molecular systematics of *Cardamine* and allied genera (Brassicaceae): ITS and noncoding chloroplast DNA. Folia Geobot. **33**:225–240.

Galloway, G. L., R. L. Malmberg, and R. A. Price. 1998. Phylogenetic utility of the nuclear gene arginine decarboxylase: an example from Brassicaceae. Mol. Biol. Evol. **15**: 1312–1320.

Gaut, B. S., and M. T. Clegg. 1991. Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. Proc. Natl. Acad. Sci. USA **88**:2060–2064.

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

GAUT, B. S., B. R. MORTON, B. C. MCCAIG, and M. T. CLEGG. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. USA **93**:10274–10279.

HIGGINS, D. G., A. J. BLEASBY, and R. FUCHS. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Appl. Biosci. **8**:189–191.

HILTON, H., and B. S. GAUT. 1998. Speciation and domestication in maize and its wild relatives—evidence from the *Globulin-1* gene. Genetics **150**:863–872.

HOWLES, P. A., T. ARIOLI, and J. J. WEINMAN. 1995. Nucleotide sequence of additional members of the gene family encoding chalcone synthase in *Trifolium subterraneum.* Plant Physiol. **107**:1035–1036.

INNAN, H., F. TAJIMA, R. TERAUCHII, and T. MIYASHITA. 1996. Intragenic recombination in the *Adh* locus of the wild plant *Arabidopsis thaliana.* Genetics **143**:1761–1770.

JUNGHANS, H., K. DALKIN, and R. A. DIXON. 1993. Stress response in alfalfa (*Medicago sativa* L.) 15. Characterization and expression patterns of members of a subset of the chalcone synthase multigene family. Plant Mol. Biol. **22**:239–253.

KARKKAINEN, K., H. KUITTINEN, R. VAN TREUREN, C. VOGL, S. OIKARINEN, and O. SAVOLAINEN. 1999. Genetic basis of inbreeding depression in *Arabis petraea.* Evolution **53**: 1354–1365.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KOCH, M., J. BISHOP, and T. MITCHELL-OLDS. 1999. Molecular systematics and evolution of *Arabidopsis* and *Arabis.* Plant Biol. **1**:529–537.

KOCH, M., B. HAUBOLD, and T. MITCHELL-OLDS. 2000. Molecular systematics of the Brassicaceae: evidence from coding plastidic *matK* and nuclear *CHS* sequences. Am. J. Bot. (in press).

KOCH, M., M. HUTHMANN, and H. HURKA. 1998. Isozymes, speciation and evolution in the polyploid *Cochlearia* L. (Brassicaceae). Bot. Acta **111**:411–425.

KOES, R. R., C. E. SPELT, and P. J. M. VAN DEN ELZEN. 1989. Cloning and molecular characterization of the chalcone synthase multigene family of *Petunia hybrida.* Gene **81**:245–257.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Mol. Evol. **36**:96–99.

LLOPART, A., and M. AGUADE. 1999. Synonymous rates at the *RpII215* gene of *Drosophila*: variation among species and across the coding region. Genetics **152**:269–280.

MAI, D. H. 1995. Tertiäre Vegetationsgeschichte Europas. Fischer, Jena, Stuttgart, New York.

MANLY, B. F. J. 1994. Multivariate statistical methods: a primer. 2nd edition. Chapman and Hall, London.

MANTEL, N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. **27**:209–220.

MANTON, J. 1937. The problem of *Biscutella laevigata* L. II. Evidence from meiosis. Ann. Bot. **1**:439–462.

MAYNARD SMITH, J. 1992. Analysing the mosaic structure of genes. J. Mol. Evol. **34**:126–129.

MIYASHITA, N. T., H. INNAN, and R. TERAUCHI. 1996. Intra- and interspecific variation of the alcohol dehydrogenase locus region in wild plants *Arabis gemmifera* and *Arabidopsis thaliana.* Mol. Biol. Evol. **13**:433–436.

MIYASHITA, N. T., A. KAWABE, H. INNAN, and R. TERAUCHI. 1998. Intra- and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. Mol. Biol. Evol. **15**:1420–1429.

MORTON, B. R., B. S. GAUT, and M. T. CLEGG. 1996. Evolution of alcohol dehydrogenase genes in the palm and grass families. Proc. Natl. Acad. Sci. USA **93**:11735–11739.

MUMMENHOFF, K., and M. KOCH. 1994. Chloroplast DNA restriction site variation and phylogenetic relationships in the genus *Thlaspi* sensu lato (Brassicaceae). Syst. Bot. **19**:73–88.

MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

PLANTHOLT, U. 1995. Molekulare Untersuchungen zur Arealgeschichte von *Arabis alpina* L. (Brassicaceae). Ph.D. thesis, University of Osnabrück, Germany.

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

ROLLINS, R. C. 1941. A monographic study of *Arabis* in western North America. Rhodora **43**:289–325.

ROSS, N. 1997. Maximum chi-squared. Distributed by Brian Spratt, Department of Zoology, Oxford University, Oxford, England.

RYDER, T. B., S. A. HEDRICK, J. N. BELL, X. LIANG, S. D. CLOUSE, and C. J. LAMB. 1987. Organization and differential activation of a gene family encoding the plant defense enzyme chalcone synthase in *Phaseolus vulgaris.* Mol. Gen. Genet. **210**:219–233.

SADOWSKI, J., P. GAUBIER, M. DELSENY, and C. F. QUIROS. 1996. Genetic and physical mapping in *Brassica* diploid species of a gene cluster defined in *Arabidopsis thaliana.* Mol. Gen. Genet. **251**:298–306.

SANDERSON, M. J. 1998. Estimating rate and time in molecular phylogenies: beyond the molecular clock? Pp. 242–264 *in* P. S. SOLTIS, D. E. SOLTIS, and J. J. DOYLE, eds. Molecular systematics of plants. 2nd edition. Chapman and Hall, New York.

SANMIGUEL, P., B. S. GAUT, A. TIKHONOV, Y. NAKAJIMA, and J. L. BENNETZEN. 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. **20**:43–45.

SCANDALIOS, J. G. 1969. Genetic control of multiple molecular forms of enzymes in plants: a review. Biochem. Genet. **3**: 37–79.

SCHULZ, O. E. 1936. Cruciferae. Pp. 227–658 *in* A. ENGLER and K. PRANTL, eds. Die natürlichen Pflanzenfamilien. Vol. 17b. Engelmann, Leipzig, Germany.

SOLTIS, D. E., C. H. HAUFLER, D. C. DARROW, and G. J. GASTONY. 1983. Starch gel electrophoresis of ferns: a compilation of grinding buffers, gel and electrode buffers, and staining schedules. Am. Fern J. **73**:9–27.

SOMERVILLE, C., and S. SOMERVILLE. 1999. Plant functional genomics. Science **285**:380–383.

SWOFFORD, D. L. 1999. PAUP: phylogenetic analysis using parsimony. Version 4.0b2. Sinauer, Sunderland, Mass.

SYVANEN, M. 1994. Horizontal gene transfer: evidence and possible consequences. Annu. Rev. Genet. **28**:237–261.

TITZ, W. 1970. Zur Cytotaxonomie von *Arabis hirsuta* agg. (Cruciferae). V. Artifizielle und natürliche F1-Hybriden sowie deren Cytogenetik. Öster. Bot. Z. **118**:353–390.

———. 1976. Cytosystematic study of the Iberian taxa of the *Arabis hirsuta* group. Feddes Repertorium **87**:493–502.

———. 1978. Experimentelle Systematik und Genetik der kahlen Sippen in der *Arabis hirsuta*-Gruppe (Brassicaceae). Bot. Jahrb. Syst. **100**:110–139.

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022

TRICK, M., E. S. DENNIS, K. J. R. EDWARDS, and W. J. PEACOCK. 1988. Molecular analysis of the alcohol dehydrogenase gene family of barley. Plant Mol. Biol. **11**:147–160.

VAN TREUREN, R., H. KUITTINEN, K. KARKKAINEN, E. BAENA-GONZALEZ, and O. SAVOLAINEN. 1997. Evolution of microsatellites in *Arabis petraea* and *Arabis lyrata,* outcrossing relatives of *Arabidopsis thaliana.* Mol. Biol. Evol. **14**:220–229.

WARWICK, S. I., and L. D. BLACK. 1991. Molecular systematics of *Brassica* and allied genera (subtribe Brassicinae, Brassiceae)—chloroplast genome and cytodeme congruence. Theor. Appl. Genet. **82**:81–92.

———. 1997. Phylogenetic implications of chloroplast DNA restriction site variation in subtribe Raphaninae and Cakilinae (Brassicaceae, tribe Brassiceae). Can. J. Bot. **75**:960–973.

WINGENDER, R., H. RÖHRIG, C. HÖRICKE, D. WING, and J. SCHELL. 1989. Differential regulation of soybean chalcone synthase genes in plant defense, symbiosis and upon environmental stimuli. Mol. Gen. Genet. **218**:315–322.

WOLFE, K. H. 1993. Software program li93. University of Dublin, ftp://acer.gen.tcd.ie/pub/khwolfe/li93.

WOLFE, K. H., P. M. SHARP, and W.-H. LI. 1989. Rates of synonymous substitution in plant nuclear genes. J. Mol. Evol. **29**:208–211.

YANG, Y.-W., K. N. LAI, P.-Y. TAI, and W.-H. LI. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. J. Mol. Evol. **48**:597–604.

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.

ZUNK, K., K. MUMMENHOFF, M. KOCH, and H. HURKA. 1996. Phylogenetic relationships of *Thlaspi* s. l. (subtribe Thlaspidinae, Lepidieae) and allied genera based on chloroplast DNA restriction-site variation. Theor. Appl. Genet. **92**:375–381.

ELIZABETH KELLOGG, reviewing editor

Accepted June 13, 2000

Downloaded from https://academic.oup.com/mbe/article/17/10/1483/956173 by guest on 21 August 2022