DNA Polymorphism, Haplotype Structure and Balancing Selection in the Leavenworthia PgiC Locus

Dmitry A. Filatov and Deborah Charlesworth

Institute of Cell, Animal and Population Biology, University of Edinburgh, Edinburgh EH9 3JT, United Kingdom

Manuscript received April 2, 1999
Accepted for publication July 6, 1999

ABSTRACT

A study of DNA polymorphism and divergence was conducted for the cytosolic phosphoglucose isomerase (PGI:E.C.5.3.1.9) gene of five species of the mustard genus Leavenworthia: *Leavenworthia stylosa*, *L. alabamica*, *L. crassa*, *L. uniflora*, and *L. torulosa*. Sequences of an internal 2.3-kb PgiC gene region spanning exons 6–16 were obtained from 14 *L. stylosa* plants from two natural populations and from one to several plants for each of the other species. The level of nucleotide polymorphism in *L. stylosa* PgiC gene was quite high ($\pi = 0.051$, $\theta = 0.052$). Although recombination is estimated to be high in this locus, extensive haplotype structure was observed for the entire 2.3-kb region. The *L. stylosa* sequences fall into at least two groups, distinguished by the presence of several indels and nucleotide substitutions, and one of the three charge change nucleotide replacements within the region sequenced correlates with the haplotypes. The differences between the haplotypes are older than between the species, and the haplotypes are still segregating in at least two of five species studied. There is no evidence of recent or ancient population subdivision that could maintain distinct haplotypes. The age of the haplotypes and the results of Kelly's Z_{nS} and Wall's *B* and *Q* tests with recombination suggest that the haplotypes are maintained due to balancing selection at or near this locus.

IN a previous study of the single cytosolic phosphoglucose isomerase (PgiC) gene of *Leavenworthia stylosa*, Liu *et al.* (1999) found unexpectedly high levels of DNA sequence polymorphism. In the part of the gene studied, around intron 12, this was found to be accompanied by distinct haplotype structure and a high level of linkage disequilibrium between haplotypes. On the basis of the sequence data, the recombination rate was estimated to be high in the region studied, and data from other loci (Liu *et al.* 1998) did not support the hypothesis of population subdivision. It therefore seemed likely that the high nucleotide polymorphism and haplotype structure could be the results of the long-term action of balancing selection in or near intron 12. On the other hand, the data exhibited no deviations from the neutral model. Thus the cause of the high polymorphism remained unclear.

A neutral polymorphic site is expected to be maintained in a population for less than $4N$ generations. Advantageous or deleterious alleles will be fixed or eliminated much faster than this (Kimura and Ohta 1969). However, polymorphisms can be maintained for very long times if balancing selection acts in or near the locus. In the presence of a polymorphic variant, such as an amino acid involved in an allozyme polymorphism, diversity is expected to be elevated at very closely linked sites (Strobeck 1983; Hudson and Kaplan 1988). Only a few definitive cases of balancing selection have been described at the DNA level, including the major histocompatibility complex (*Mhc*) polymorphism (Klein 1986), studied mostly in humans (reviewed in Klein *et al.* 1990) and self-incompatibility loci in plants (reviewed in Sims 1993 and Charlesworth and Awadalla 1998). A peak of polymorphism has been attributed to the effect of balancing selection in the *Drosophila melanogaster Adh* locus (Hudson *et al.* 1987), although as this peak is not entirely attributable to the F/S amino acid difference, but also occurs in the S alleles, the situation at this locus is not entirely clear. Studies of other loci with allozyme polymorphisms have failed to find such peaks (Takano *et al.* 1993). Here we present further DNA polymorphism data from the cytosolic phosphoglucose isomerase (PgiC) locus of *L. stylosa*, which strengthen the evidence that this gene is another example of a gene under balancing selection.

The goal of the present work is to examine the level of nucleotide polymorphism and haplotype structure in a much larger PgiC region than previously sequenced, to determine the extent of the region of high polymorphism in *L. stylosa*, and to obtain a larger number of allele sequences in order to test for deviations from the neutral model in this locus. For these purposes we obtained DNA sequences of PgiC alleles for a region spanning exons 6–16. Very surprisingly, this demonstrated that the observed haplotype structure extends much further than the intron 12 region, despite evi-

Corresponding author: D. A. Filatov, Institute of Cell, Animal, and Population Biology, King's Bldgs., West Mains Rd., University of Edinburgh, Edinburgh EH9 3JT, United Kingdom.
E-mail: dmitry.filatov@ed.ac.uk

dence of multiple recombination events. Interestingly, one of the amino acid replacements associated with a charge change correlates with the haplotypes. Comparison of the sequences from different Leavenworthia species shows that the split between haplotypes preceded the split between the species in the genus. This is even clearer now that we have included sequences from a further species in the genus that also has high allozyme diversity (*L. alabamica* populations; see Charlesworth and Yang 1998). Despite the ancient origin of the haplotypes, they are thus still segregating in at least two Leavenworthia species. Furthermore, several tests potentially able to detect deviations from the neutral model, particularly tests including the possibility of recombination, Kelly's Z_{nS} (Kelly 1997) and Wall's *B* and *Q* statistics (Wall 1999), detect significant deviation from neutral expectations.

MATERIALS AND METHODS

Species and populations: A detailed description of the genus Leavenworthia, including most of the populations studied here, is in Liu *et al.* (1999). In this study we used 14 plants of an outcrossing species *L. stylosa* from populations 95007 (6 plants) and Hem1 (8 plants). Four alleles were obtained from 3 plants from two *L. alabamica* populations (see Charlesworth and Yang 1998), two alleles from a partially self-incompatible population (95006) and two from a highly self-fertile population (95009). One allele was sequenced from a *L. crassa* plant from the partially self-incompatible population 95005, and one was sequenced from each of two highly inbreeding species, *L. torulosa* (population 95008) and *L. uniflora* (population 95011). Since the genus Leavenworthia is thought to be closely related to Cardamine (Rollins 1963), we also isolated DNA from one *Cardamine hirsuta* plant collected at the University of Edinburgh King's Buildings campus for use as an outgroup.

Molecular methods: Genomic DNA was isolated from Leavenworthia leaves of individual plants by a standard hexadecyltrimethylammonium bromide (CTAB) plant miniprep method with several modifications. Leaves (~100 mg) were thoroughly ground in liquid nitrogen and then in 1 ml of extraction buffer (0.35 m sorbitol, 5 mm EDTA, 0.1 m Tris-HCl pH 7.4, 30 mm sodium bisulfite). Nuclei were collected by centrifugation at 3000 × *g* for 5 min. The nuclei were resuspended in 300 ml of extraction buffer and 300 ml of lysis buffer (0.2 m Tris, pH 7.5, 50 mm EDTA, 2 m NaCl, 2% CTAB, 5% *N*-lauroyl sarcosine) and incubated with RNAse for 10 min at 65°. After phenol-chloroform purification, DNA was precipitated with 0.6 volume of isopropanol and dissolved in 100–200 ml Tris-EDTA pH 8.

We used sequences of the *Arabidopsis thaliana* PgiC gene and *L. crassa* PgiC cDNA (GenBank accession nos. X69195 and AF054455) to design five "plus" and four "minus" primers for PCR and sequencing of the central 2.3-kb region of Leavenworthia PgiC gene (plus primers: +8, CCACTGTTTGTTCA TACGGCTC; +10, AAATATTGATCCTGTTGATGTTG; +12, TGCTGTSAGCACTAATCTTGCG; +3, TTTGCATTTTGGGA CTGGG; +14, AAGGGAGCTTCAAGCATTGAT; minus primers: −11, GCGTTCAGCATTGTTTCAGC; −13, TTGTTC GGGTCAATACCAAACT; −15, GCTGATCAATGCTTGAAG CTCC; −4, TCGAACGGGAGAGGTAGACCA). The +8 and −13 primers were used to amplify a region of 1.2-kb PgiC from *L. stylosa*, referred to below as region A (Figure 1). The +12 and −4 primers were used to amplify a 1.3-kb region referred to below as region B (Figure 1). Primers +8 and −4 were used to amplify the entire 2.3-kb PgiC region of the four other Leavenworthia species and of *C. hirsuta*. The amplification products were cloned into the pCR2.1 vector using the TA cloning kit (Invitrogen, San Diego) and both strands were sequenced on an ABI Prism 377 automatic sequencer (Perkin Elmer, Norwalk, CT).

Sequence alignment and analysis: The sequences obtained were aligned using ClustalX v.1.64 software (Thomson *et al.* 1997) followed by additional hand alignment using the PROSEQ v.2.3, multiwindow sequence processor for Windows 95 developed by D. Filatov (unpublished results). Sequence data analyses (estimators of DNA diversity, the estimators of recombination rate *C* and γ, the estimator of silent site divergence D_{xy}, and Tajima's, Fu and Li's, McDonald and Kreitman (MK), and Hudson, Kreitman, and Aguadé (HKA) tests of neutrality) were performed using DNAsp v.2.93 (Rozas and Rozas 1997), SITES v.1.1 (Hey and Wakeley 1997), PROSEQ v.2.3, and an unpublished Fortran program written by D. Charlesworth. Wall's *B* and *Q* tests (Wall 1999), Kelly's test (Kelly 1997), sliding windows for linkage disequilibrium analyses, permutation tests for geographic subdivision (Hudson *et al.* 1992), and coalescent simulations with recombination were performed by PROSEQ v.2.3. For all phylogenetic analyses we used MEGA v.1.01 (Kumar *et al.* 1993).

A permutation approach (Hudson *et al.* 1992) was used to estimate the significance of sequence differences between the two *L. stylosa* populations studied and between the L and S haplotypes of *L. stylosa*. The value of the K_{st}* statistic was calculated for two groups of sequences (either for the two geographic populations or for L and S haplotypes). The critical value for this statistic was obtained by 1000 random permutations of the sequences between the two groups in the sample.

Coalescent simulations: For the coalescent simulations (see below) we used program routines in Pascal code kindly provided by J. Hey. These routines implement the standard algorithm of the coalescent process with recombination (Hudson 1983, 1990, 1993). The routines were rewritten in Object Pascal and built into the PROSEQ v.2.3 (D. Filatov, unpublished data). The program generates random samples of a given size with a given number of segregating sites and specified recombination rates. These simulated samples were used to estimate critical values of Kelly's (1997) and Wall's (1999) tests with recombination. In our simulations (see below), we used several values of the recombination rate, chosen to be close to the values estimated from the *L. stylosa* PgiC data (see Table 3).

Kelly's and Wall's tests with recombination: To calculate the probability of the observed values of test statistics arising by chance (*P* value) we simulated random samples of a given size, number of polymorphic sites, and recombination rate using the coalescent process. For each such sample generated, Kelly's Z_{nS} statistic (Kelly 1997) and Wall's *B* and *Q* statistics (Wall 1999) were calculated and stored. After the statistics had been calculated for 10,000 simulated samples, the *P* value for the observed value of the statistic was obtained as a proportion of cases when the simulated statistic value was greater than or equal to the observed one.

RESULTS

DNA polymorphism: Nineteen alleles were sequenced from the two *L. stylosa* populations (see materials and methods) for region B of the PgiC gene. From these data, it was apparent that high diversity

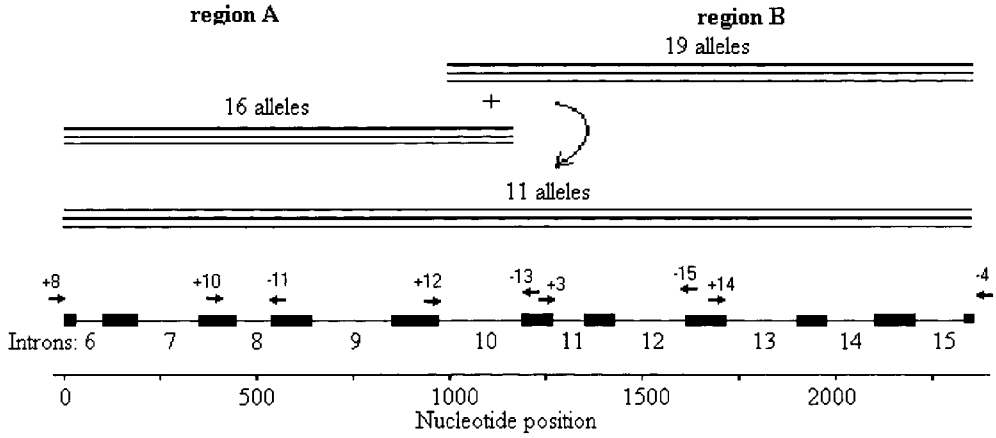

Figure 1.—The PgiC region sequenced from exon 6 to exon 16. Thick and thin horizontal lines represent exons and introns, respectively. Small arrows show positions, directions, and names of the primers. For *L. stylosa* the whole 2.3-kb region was cloned and sequenced as two overlapping regions (region A and region B) of ~1.3 kb each.

extends throughout this region, so 16 alleles were sequenced for a further region (A) 5′ to region B (see Figure 1). The sequence of the entire 2.3-kb region was obtained for 11 chromosomes for this species, using the few plants for which both alleles were sequenced for both the A and B regions. In *L. stylosa*, the level of DNA polymorphism is remarkably high: 180 of 1020 sites of region A and 214 of 1147 sites of region B are segregating in our samples. In the 11 sequences covering the whole 2.3-kb region, 263 of 2045 sites are polymorphic. The distribution of nucleotide polymorphisms (π_{total}) along the sequence, based on these 11 alleles, is shown in Figure 2. Due to the large number of polymorphic sites, it is impossible to show them all in a figure; however, the list of all polymorphic sites, as well as the alignments, is available from the authors on request.

No insertions or deletions (indels) were found in the coding regions, but extensive intron length polymorphism was observed in the introns. Nine of the 10 introns sequenced vary in size due to indels up to 100 bp long. In total, we found 51 indels in the introns of the whole 2.3-kb region. All indel polymorphism regions are excluded from the analysis below. Since indels represent about 10% of the region sequenced, indel regions were also analyzed separately, to check that they do not differ greatly in diversity from other intron regions. The nucleotide variation within indel regions was approximately the same as that elsewhere in *L. stylosa* PgiC introns (per-nucleotide $\pi \approx 5\%$).

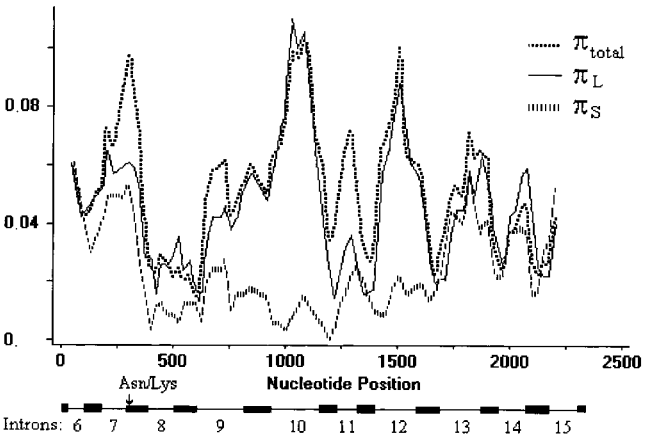

Figure 2.—The distribution of nucleotide polymorphisms along the 11 *L. stylosa* PgiC alleles 2.3 kb long detected in sliding windows of 100 bp with increment 25 bp. Values of nucleotide diversity (π) are shown for 11 alleles pooled (π_{total}) and for S (π_S) and L (π_L) haplotypes separately. The arrow shows the position of the Asp/Lys polymorphism.

The nucleotide variation (excluding indel regions) found in *L. stylosa* is summarized, in terms of the standard measures of sequence polymorphism, π and θ, in Table 1. The two PgiC regions sequenced have similar average levels of DNA polymorphism. Most of the segregating sites are in the introns: 223 of 1437 intron sites (15%), compared with 40 of 598 exon sites (7%). The level of nucleotide polymorphism at nonsynonymous sites is about an order of magnitude lower than at synonymous sites (Table 1). However, for the entire 2.3-kb region we observed 12 amino acid polymorphisms, 3 of which were replacements with charge changes. No correlations of the allozyme mobility classes (see Charlesworth and Yang 1998) with the observed charge change amino acid replacements were found (data not shown).

Based on the sequences, the two different populations show no significant evidence of isolation. For the largest set of data, the 19 B region sequences, the F_{st} value estimated from the π values for the two populations was low, 0.019; the value of the test statistic for detecting geographic subdivision (Hudson *et al.* 1992) was $K_{st}^* = 0.0098$, which is lower than the 95th percentile ($K_{st}^*{}_{0.95} = 0.0168$) calculated from 1000 permutations of the 19 sequences. For region A and for the entire 2.3-kb region, tests for subdivision (Hudson *et al.* 1992) were also nonsignificant. Since there is no evidence of isolation of these two populations, they will be combined for further analysis.

Recombination: The minimum number of recombination events, estimated by Hudson and Kaplan's (1985) method, is 21 for the sample of 11 alleles 2.3 kb

TABLE 1

DNA polymorphism in the PgiC gene of *L. stylosa*

Region	Sample size	Introns		Exons replacement			Total	
		$\pi \pm$ SD	$\theta \pm$ SD	No.	$\pi \pm$ SD	$\theta \pm$ SD	$\pi \pm$ SD	$\theta \pm$ SD
A	16	0.054 ± 0.0087	0.062 ± 0.0050	8	0.0066 ± 0.00091	0.012 ± 0.00085	0.045 ± 0.0069	0.053 ± 0.0039
B	19	0.066 ± 0.0049	0.067 ± 0.0049	7	0.0035 ± 0.0007	0.0076 ± 0.0008	0.052 ± 0.0036	0.052 ± 0.0037
2.3 kb	11	0.053 ± 0.0055	0.053 ± 0.0036	12	0.0062 ± 0.0006	0.009 ± 0.0008	0.043 ± 0.0043	0.044 ± 0.0027

long. For the A and B regions this method detected 16 and 28 recombination events, respectively (Figure 3). Thus, the PgiC region is not a cold spot of recombination in *L. stylosa*. Hudson's (1987) C estimator of recombination ($C = 4N_e c$, where N_e is the effective population size of the species and c is the recombination frequency per nucleotide site) gave a value of 0.083/bp for the B region, for which the set of sequences is largest; both A and B regions are consistent in suggesting frequent recombination. For the same data set, the γ estimator of recombination (Hey and Wakeley 1997) gave a value of 0.044/bp. The per-nucleotide recombination rate in all parts of the PgiC gene sequenced therefore appears to be ~5% and the ratio of C/θ appears to be ~1. Both γ and C estimates are biased but the biases are in opposite directions (Hudson 1987; Hey and Wakeley 1997). The two estimators have large variance; however, according to the simulations conducted by Hey and Wakeley (1997), for 12 sequences 2 kb long the variance of γ is only about twice that for Watterson's estimator of θ. Thus, for our samples, the variance of γ should be ~0.01 and the true per-nucleotide recombination rate ($C = 4N_e c$) should probably not be <0.03.

Linkage disequilibrium and haplotype structure: Linkage disequilibria (significant by χ^2 tests at $P < 0.05$ without correction for multiple tests) were detected for >25% of pairs of sites, many of which were >1 kb one from another. Although no disequilibria were significant after correction for multiple tests, the significance of Kelly's Z_{nS} test with recombination (see below) suggests that linkage disequilibrium is significant in an evolutionary sense, *i.e.*, that it exceeds that expected under neutrality. We do not show linkage disequilibrium data in the form of the commonly used linkage disequilibria grid because of the large number of sites. Instead, the distribution of linkage disequilibrium along the region, measured as Z_{nS} and average D in a sliding window of 15 polymorphic sites, is shown in Figure 4. There is a clear peak of linkage disequilibrium in the middle of the sequence, centered close to intron 10 and exon 11, and two smaller peaks centered in introns 7 and 12. Interestingly, the peaks of linkage disequilibrium coincide approximately with the peaks of nucleotide polymorphisms (Figure 2).

Three distinct groups of alleles were previously defined by indel polymorphisms in intron 12 (Liu *et al.* 1999). From length differences in this intron the following names were assigned: short (S), long 1 (L1), and long 2 (L2). The groups are also distinguished by a number of nucleotide substitutions. With the longer sequence now available, the L1 and L2 groups are no longer distinct from one another, but sequences of the S group still cluster together and separately from L1 and L2 (Figure 4). Since the trees for L1 and L2 alleles are not well resolved, we will here combine L1 and L2 into a single L group. This is convenient for the analysis,

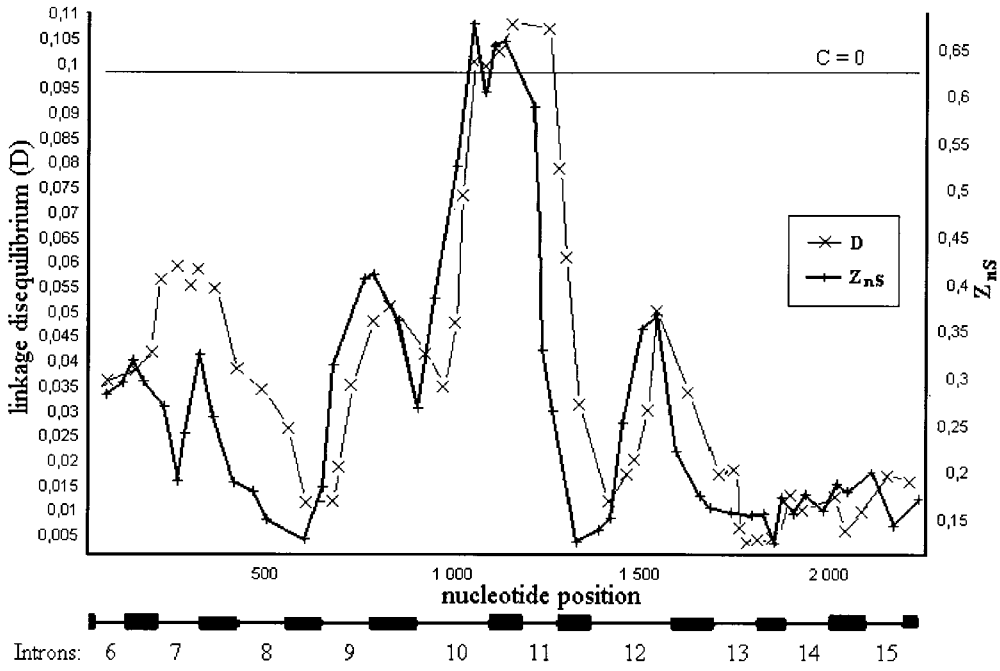

Figure 3.—The distribution of linkage disequilibrium (measured by D and Z_{nS}) among the 11 *L. stylosa* alleles of the 2.3-kb PgiC region, detected in sliding windows of size 15 polymorphic sites with an increment of 5 polymorphic sites. For each window the value of Z_{nS} and an average value of D are plotted. The thin horizontal line shows the critical ($P < 0.05$) value of Z_{nS} without recombination for 11 sequences with 15 segregating sites.

since the frequency of S type alleles is approximately equal to the combined frequency of L1 and L2 alleles.

Pairwise nucleotide divergences between the 11 sequences covering the whole 2.3-kb region are shown in Table 2. Within groups (boxed), divergence is lower than between the groups. The diversity among sequences within the combined L1 + L2 group is, however, nearly as high as the between-group divergence. This suggests that these two groups are distinct from one another, despite their not being well resolved in the trees. The level of DNA polymorphism within the S group ($\pi = 0.025 \pm 0.003$) is significantly lower than in the whole sample ($\pi = 0.052 \pm 0.0036$), or within the L1 ($\pi = 0.040 \pm 0.007$) or L2 ($\pi = 0.050 \pm 0.010$) groups, or within the combined L1 + L2 group of sequences ($\pi = 0.056 \pm 0.004$). The L2 group has the highest within-group polymorphism and could probably be further divided into smaller subgroups if a bigger sample were studied.

The distribution of DNA polymorphisms along the 2.3-kb PgiC region for 11 pooled sequences and separately for S (6 alleles) and L (5 alleles) haplotypes is shown in Figure 2. Most polymorphisms are within the L type but the peak around intron 7 is mostly due to the differences between the S- and the L-type alleles.

To test the significance of differences between the L and S haplotypes, we applied a permutation approach (Hudson *et al.* 1992), treating the two groups as geo-

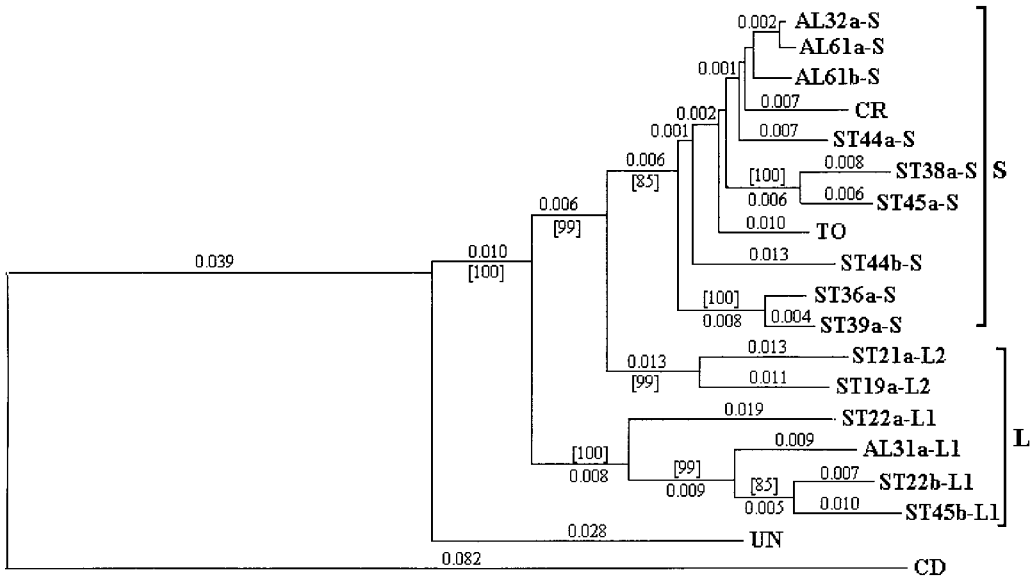

Figure 4.—Neighbor-joining tree for the entire 2.3-kb PgiC region of five Leavenworthia species and of *C. hirsuta*. Jukes-Cantor distances are shown for each branch. Bootstrap values (in brackets) are shown only for basic branching between S and L haplotypes. Haplotype assignments are shown as follows: S, short; L1, long 1; L2, long 2. L comprises L1 and L2. Species are indicated by the first two letters in each label as follows: ST, *L. stylosa*; AL, *L. alabamica*; CR, *L. crassa*; TO, *L. torulosa*; UN, *L. uniflora*; CD, *C. hirsuta*.

TABLE 2

Pairwise nucleotide divergence between 11 *L. stylosa* alleles sequenced for the entire 2.3-kb PgiC region

	1-L1	2-L1	3-L1	1-L2	2-L2	1-S	2-S	3-S	4-S	5-S	6-S
1-L1	—										
2-L1	42	—									
3-L1	89	89	—								
1-L2	117	126	141	—							
2-L2	125	121	142	62	—						
1-S	114	127	134	84	72	—					
2-S	111	127	132	83	71	17	—				
3-S	141	127	120	108	83	59	59	—			
4-S	127	135	105	98	85	59	59	57	—		
5-S	144	151	118	89	103	78	77	76	57	—	
6-S	133	140	116	76	90	62	61	74	56	31	—

Within-haplotype comparisons are boxed.

graphic populations. This tests for evidence of significant isolation between the two sequence groups, in our case potentially attributable to their being associated with different alleles maintained in the populations by balancing selection, rather than isolation due to geographic separation, for which the test was originally designed. The test results were highly significant throughout the gene. For the full 2.3-kb region, $K_{st}^* = 0.055$ ($K_{st}^*{}_{0.999} = 0.044$, $P < 0.001$); for the A region, $K_{st}^* = 0.067$ ($K_{st}^*{}_{0.999} = 0.062$, $P < 0.001$); and for the B region, $K_{st}^* = 0.076$ ($K_{st}^*{}_{0.999} = 0.047$, $P < 0.001$).

Tests for selection assuming no recombination: To test for the operation of natural selection in the region under study, we applied several tests (HKA test, Hudson *et al.* 1987; MK test, McDonald and Kreitman 1991; Tajima's *D*, Tajima 1989; Fu and Li's *D**, Fu and Li 1993) potentially able to detect deviations of the observed data from the expectations under the neutral model. None of the tests was significant, probably because these tests assume no recombination and, since there is evidence of recombination in the *L. stylosa* PgiC region, the power of the tests is reduced (Hudson *et al.* 1987).

Tests for selection allowing for recombination: Kelly's test (Kelly 1997) examines regions for excess of linkage disequilibrium compared with that expected under neutrality. The test statistic Z_{nS} averages the values of linkage disequilibrium (d_{ij}, the squared correlation of allelic identity between loci *i* and *j*, see Hartl and Clark 1989, pp. 53–54) across all polymorphic sites in the region. It thus summarizes linkage disequilibrium between all pairs of polymorphic sites. Since the critical values of Z_{nS} given in the original article (Kelly 1997) are for up to 50 polymorphic sites in the sample, we applied the test in a sliding window of 15 or 50 polymorphic sites. The results of the tests were significant only for the intron 10 region using a window size of 15 polymorphic sites (see Figure 4). To apply the test to the data for the whole sequence, we calculated critical Z_{nS} values using a program kindly provided by John Kelly. Again, the results for the A and B regions and the whole 2.3-kb region were nonsignificant. However, the critical values for Kelly's statistics used in these tests are calculated from coalescent simulations without recombination (Kelly 1997) so the test is extremely conservative. We therefore calculated *P* values for the observed Z_{nS} (Table 3) by coalescent simulations with recombination (see materials and methods). With recombination, the results of Kelly's test are nonsignificant ($P > 0.05$) for both A and B regions when $C < 0.04$–0.05 and for the entire 2.3-kb region when $C < 0.07$.

We also applied J. Wall's (1999) *B* and *Q* tests (Table 3) for deviations of the results from the neutral model. The *B* statistic is the proportion of pairs of adjacent segregating sites that are congruent, *i.e.*, that have consistent genealogies. The *Q* statistic is also based on the number of adjacent congruent sites, but takes into account the length of the regions where all sites are congruent. The *P* values for the observed *B* and *Q* were calculated in exactly the same way as for Kelly's Z_{nS} statistic (see materials and methods). For the A and B regions the *B* and *Q* statistics are significant when $C \geq 0.03$–0.04 (Table 3). For the 2.3-kb region both *B* and *Q* statistics detected significant deviation from the neutral model, with recombination rate $C \geq 0.01$ (see Table 3), which is four times less than the value of γ estimator (note that this estimator tends to underestimate the amount of recombination).

Between-species comparisons: We sequenced the 2.3-kb PgiC region for four alleles of *L. alabamica* and one allele for each of the species *L. crassa*, *L. uniflora*, and *L. torulosa*. The neighbor-joining tree for the five Leavenworthia species and *C. hirsuta* (outgroup) is shown in Figure 4. *L. torulosa* is very close to *L. stylosa*, as is also the case in phylogenies based on morphological characters (Christiansen 1993), and the *L. torulosa* sequences cluster together with the S haplotype of *L. stylosa* as was previously found for the intron 12 region

TABLE 3

Results of Kelly's Z_{nS} and Wall's *B* and *Q* tests with recombination

	Region A (1.2 kb)	Region B (1.3 kb)	Entire region (2.3 kb)
Sample size	16	19	11
Segregating sites	180	214	263
	P values for Kelly's Z_{nS} statistic with different recombination rates		
Observed	$Z_{nS} = 0.148$	$Z_{nS} = 0.138$	$Z_{nS} = 0.158$
$C = 0.005$	0.916	0.879	0.971
$C = 0.01$	0.784	0.695	0.851
$C = 0.02$	0.399	0.300	0.562
$C = 0.03$	0.181	0.082	0.301
$C = 0.04$	0.079	0.034	0.178
$C = 0.05$	0.030	0.015	0.080
$C = 0.06$	0.025	0.012	0.051
$C = 0.07$	0.010	0.007	0.033
$C = 0.08$	0.006	0.005	0.019

	P values for Wall's *B* and *Q* statistics with different recombination rates					
Observed	$B = 0.074$	$Q = 0.133$	$B = 0.070$	$Q = 0.125$	$B = 0.145$	$Q = 0.236$
$C = 0.005$	0.933	0.954	0.915	0.931	0.297	0.392
$C = 0.01$	0.608	0.682	0.560	0.600	0.020	0.042
$C = 0.02$	0.152	0.207	0.107	0.130	0.000	0.001
$C = 0.03$	0.034	0.067	0.022	0.031	0.000	0.000
$C = 0.04$	0.015	0.031	0.006	0.009	0.000	0.000
$C = 0.05$	0.006	0.011	0.003	0.005	0.000	0.000
$C = 0.06$	0.005	0.009	0.001	0.003	0.000	0.000
$C = 0.07$	0.000	0.004	0.000	0.002	0.000	0.000
$C = 0.08$	0.000	0.000	0.000	0.001	0.000	0.000

(Liu *et al.* 1999). Sequences from *L. crassa* and three of the four sequences from *L. alabamica* (closely related species that are more distant from *L. stylosa* and have a different chromosome number; see Rollins 1963; Christiansen 1993) also cluster together with *L. stylosa* S-haplotype sequences. However, one of the two *L. alabamica* allele sequences from population 95006 clusters with the *L. stylosa* L1-haplotype sequences. Thus, the differences between haplotypes are apparently older than the differences between species in the genus Leavenworthia. It is interesting that haplotypes are still segregating in at least two of five Leavenworthia species studied.

Correlation of haplotypes with amino acid replacements: Overall we detected 12 amino acid replacements in the entire 2.3-kb region sequenced. For the 2.3-kb region, 3 of the 12 replacements involve charge changes. Most of the amino acid polymorphisms were singletons or variants found only twice in our sample of alleles. Only in one case (Asn/Lys in exon 8) do the two alleles have similar frequencies. This Asn/Lys polymorphism is due to a T/A mutation in the third position of PgiC codon 200. Interestingly, the Asn/Lys polymorphism is strongly associated with the haplotypes (Table 4). All alleles of S type have Lys (positively charged) and all except one L-type allele have Asn (uncharged) in this site. One allele of L1 type has Lys; this allele may be a recombinant since its 5′ part is more similar to S-type sequences, and only its 3′ part (the part used in the haplotype assignments) is of L1 type.

The comparison with the other species (Table 4) demonstrates that Asn is the ancestral state since *C. hirsuta* also has Asn in that position (based on a single sequence from this species, which appears to be highly selfing, and which we therefore assume will have little sequence variation). The PgiC sequences of the two other Leavenworthia species, *L. crassa* and *L. torulosa*, are generally very similar to the *L. stylosa* S type, and both also have Lys at position 200. The *L. alabamica* alleles also corroborate the association between haplotypes and the Asn/Lys polymorphism. Three of the four *L. alabamica* alleles sequenced are of the S type and have Lys, while the fourth allele is of L1 type and has Asn in the position 200 of the PgiC protein. Such an association suggests that the S(Lys)/L(Asn) haplotype structure arose due to a single mutation at position 200 of the PgiC protein of an ancestral Leavenworthia species.

DISCUSSION

High level of DNA polymorphism in *L. stylosa*: The level of DNA polymorphism observed in PgiC of *L. stylosa* is strikingly high. It is much higher than for most animal genes (Moriyama and Powell 1996) and for other

TABLE 4

Amino acid replacements in the entire 2.3-kb PgiC region of six species

				Amino acid replacements with the positions in the entire protein																	
Species	Population	Allele	Haplotype	200	209	211	212	238	239	240	285	287	290	295	305	311	314	315	317	319	329	342
L. alabamica	95006	AL32a	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. alabamica	95009	AL61a	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. alabamica	95009	AL61b	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. crassa	95005	CR	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. torulosa	95008	TO	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. stylosa	Hem1	ST36a	S	K	V	V	V	A	S	A	L	L	R	V	S	K	S	L	E	M	W	L
L. stylosa	Hem1	ST39a	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	M	R	L
L. stylosa	95007	ST44a	S	K	V	V	V	A	S	A	L	L	Q	V	N	K	S	L	E	I	W	L
L. stylosa	Hem1	ST38a	S	K	V	V	V	A	S	A	S	L	Q	V	S	K	S	L	E	I	W	S
L. stylosa	95007	ST45a	S	K	V	V	V	A	S	A	L	M	Q	V	S	K	S	L	E	I	W	L
L. stylosa	95007	ST44b	S	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. stylosa	Hem1	ST22a	L1	K	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. stylosa	Hem1	ST22b	L1	N	A	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	R	L
L. stylosa	95007	ST45b	L1	N	V	V	A	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. stylosa	Hem1	ST21a	L2	N	V	A	V	V	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. stylosa	Hem1	ST19a	L2	N	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
L. alabamica	95006	AL31a	L1	N	V	V	V	A	S	T	L	L	Q	V	S	K	S	L	E	I	W	L
L. uniflora	95011	UN	—	N	V	V	V	A	S	A	L	L	Q	V	S	K	S	L	E	I	W	L
C. hirsuta	Ed-KB	CD	—	N	V	V	V	A	P	A	L	L	Q	A	S	Q	P	F	K	I	W	L

TABLE 5

Comparison of DNA polymorphism in several genes of *L. stylosa*

Gene	Silent and/or introns			Tajima's *D*	Reference
	Length	π	θ		
PgiC	1590	0.0554	0.0573	−0.405	This study
Adh 1	264	0.0456	0.0471	−0.194	Liu *et al.* (1998)
Adh 2	85	0.0110	0.0145	−0.783	Charlesworth *et al.* (1998)
Adh 3	138	0.0226	0.0337	−1.467	Charlesworth *et al.* (1998)
GapC 2	256	0.0195	0.0172	−0.3346	Liu *et al.* (1998)
Nir1	172	0.031	0.031	−0.7347	Liu *et al.* (1998)
Eno 1 (3 seq)	188	0.083	0.0851	—	Liu (1998), GenBank
Eno 2 (3 seq)	241	0.072	0.0719	—	Liu (1998), GenBank

plant genes in which sequence diversity has been quantified (except self-incompatibility loci; see, *e.g.*, Richman *et al.* 1996). The high level of DNA polymorphism in the *L. stylosa* PgiC gene could be due to the action of balancing selection in or near the locus. Theory (Strobeck 1983; Hudson and Kaplan 1988; Kaplan *et al.* 1988) suggests that there should be a peak of polymorphism and linkage disequilibrium near a site that is under balancing selection. The level of nucleotide diversity in *L. stylosa* PgiC between the peaks of polymorphism (around introns 7, 10, 12; see Figure 2) is approximately the same as in the other genes studied in this species ($\pi \sim 3$–4%; see Table 5). We must therefore consider the possibility that high DNA sequence polymorphism may be typical for the whole genome of this species. In the peaks of polymorphism, π reaches much higher values (6–10%). Thus, it is possible that π of 3–4% is typical for the genes of *L. stylosa*, but that the higher peaks of polymorphism are due to the maintenance of polymorphic sites for a long time by balancing selection.

Plant sequence diversity: Most data available on DNA polymorphism within species currently come from animals, especially Drosophila (reviewed by Moriyama and Powell 1996). DNA polymorphism has been studied for only a few plant species, many of which are domesticated, and thus diversity may be underestimated (Doebley 1989, 1992). Data are currently available from maize (Shattuck-Eidens *et al.* 1990; Gaut and Clegg 1993a; Henry and Damerval 1997), melon (Shattuck-Eidens *et al.* 1990), and millet (Gaut and Clegg 1993b), and for natural populations of wild barley (Cummings and Clegg 1998), wild yam (Terauchi *et al.* 1997), morning glory (Huttley *et al.* 1997), and several species of a mustard genus Leavenworthia (Charlesworth *et al.* 1998; Liu 1998; Liu *et al.* 1998, 1999). Polymorphism levels vary greatly in different plant species; estimates of $\theta = 4Nm$ range from ~0.001 for melon, millet, wild yam, and selfing species of Leavenworthia, *L. uniflora* and *L. crassa*, to much higher (up to 0.05) values for maize and *L. stylosa*. These high values exceed those in Drosophila populations (Moriyama and Powell 1996). Levels of intraspecies polymorphism depend on mutation rates and on aspects of population history, including the long-term population size and the occurrence of bottlenecks, which directly affect effective population sizes (Kimura 1983). In addition, genetic variability is affected by the mating system, since inbreeding increases the effects of selective sweeps and of selection against deleterious mutations (Hedrick 1980; Charlesworth *et al.* 1993; Liu *et al.* 1998, 1999). Some of the variability in levels of DNA polymorphism in plants could thus be due to the fact that some of the data are from selfing species or populations (wild barley, Cummings and Clegg 1998; *L. crassa*, *L. uniflora*, and *L. torulosa*, Liu *et al.* 1999), but for the outcrossing (dioecious) species *Dioscorea tokoro* (Terauchi *et al.* 1997) one would have to invoke lower mutation rates or population bottlenecks.

One possible cause of the high level of DNA polymorphism seen in *L. stylosa* PgiC and in maize loci (see Shattuck-Eidens *et al.* 1990; Henry and Damerval 1997) may be high mutation rates. Since plants do not have a germ line, germinal tissues are formed from somatic cells, so the number of cell divisions needed to form a progeny gamete from a parent seed, and hence per-generation mutation rates, could often be higher in plants than animals (Klekowski and Godfrey 1989). Mutation rates for chlorophyll deficiency mutations in long-lived mangrove species have been estimated to be high, $2–5 \times 10^{-3}$ per haploid genome per generation (Klekowski 1988). Assuming 200 genes (Wettstein *et al.* 1971) and a very high value of 50–100 replacement sites in each gene that could affect chlorophyll biosynthesis, we obtain a per-base per-generation mutation rate of at least $1–2 \times 10^{-7}$, about an order of magnitude higher than estimates for animals (Kondrashov 1998; Drake *et al.* 1998). Even assuming mutation rates in the annuals *L. stylosa* and maize as high as $2–5 \times 10^{-7}$ per site per generation, to account for the observed π values the effective population sizes would need to be at least $2–5 \times 10^5$. Such a large effective population size (half that estimated for *D. melanogaster*, see Kreitman and

Hudson 1991) seems implausible for *L. stylosa*, given the current fragmented state of Leavenworthia populations (Rollins 1963) and the likelihood of past population bottlenecks during glaciations. Thus, even if mutation rates are high, some other factor leading to high diversity appears necessary.

Haplotype structure: Apart from the high DNA polymorphism, another interesting feature of the *L. stylosa* PgiC data is the strong haplotype structure, which our new studies show spans at least the whole 2.3-kb region of the gene sequenced, despite the clear findings showing that this gene is not a cold spot of recombination. The significant results of both haplotype (Hudson *et al.* 1994; Kirby and Stephan 1995) and permutation (Hudson *et al.* 1992) tests show that the observed haplotype structure is highly improbable by chance alone under a neutral model.

An obvious potential explanation of the haplotype structure is recent or ancient population subdivision of *L. stylosa*. However, we could rule this out, since the two populations show no evidence for significant differentiation. Furthermore, sequence data from six other loci in *L. stylosa* show no signs of haplotype structure or isolation (Liu 1998). Another potential explanation of the observed trans-species polymorphisms is gene flow between the species of the genus. However, this does not seem likely, since *L. stylosa* does not give viable progeny with any other species and chromosome numbers differ between *L. stylosa*, *L. uniflora*, and *L. torulosa* on the one hand, and *L. crassa* and *L. alabamica* on the other hand (Rollins 1963). Even if some gene flow occurred in the past, there must be a force maintaining the haplotypes since that time.

The other possible explanation of the haplotype structure is balancing selection in or near the PgiC locus. Despite many tests of selection being nonsignificant, several lines of evidence suggest that balancing selection acts in this region. First, the results of Kelly's (1997) and Wall's (1999) tests with recombination demonstrate significant deviations from neutral expectations. Kelly's Z_{nS} test statistic is a stringent test that is sensitive to the lengths of the internal branches of gene trees. The value of the test statistic is strongly affected by linkage disequilibrium between the sites where mutations occurred on the most ancient branches of the gene tree, which go directly to the common ancestor of the entire sample. Thus, the test has good power to detect balancing selection, based on its effect of stretching the internal branches of the genealogy. Wall's *B* and *Q* tests are also quite sensitive to the length of internal branches of the sample genealogy. The critical bounds for the Z_{nS}, *B*, and *Q* test statistics were derived by coalescent simulations of random samples for a range of recombination rates close to that estimated for the *L. stylosa* PgiC gene, and the tests are mostly significant unless we employ recombination rates much lower than those estimated (Table 3). According to our results, *B* and *Q* tests appear to be more sensitive to detect balancing selection than the Z_{nS} test.

Second, comparison with other Leavenworthia species reveals that the age of the haplotypes is higher than the age of species and even the karyotype differences in the genus, since the same haplotypes segregate in at least two of the five Leavenworthia species studied. Unfortunately we cannot precisely date the age of the haplotypes and species in the genus, since neither reliable estimates of mutation rate in dicotyledonous plants nor fossil data for the Leavenworthia genus are available. Estimates of molecular clock parameters are available for monocotyledons; the substitution rate per synonymous site per year between rice and maize for nine nuclear genes is estimated to be $\sim 6 \times 10^{-9}$ (Gaut 1998). Assuming the same silent site substitution rate for Leavenworthia, the divergence time between *L. stylosa* and *C. hirsuta* estimated from the mean sequence divergence per silent site, $D_{xy} = 0.2$, is about 17 million years. This is unreasonably high since these two species are thought to be close relatives, and the age of the whole Brassicaceae family is estimated to be about 15 million years (Muller 1984). Five million years seems a more realistic upper value of the time since the common ancestor of Cardamine and Leavenworthia (Rollins 1963; Price *et al.* 1994). In that case, the silent site substitution rate in this group would be $\sim 2 \times 10^{-8}$ per year. This value seems reasonable, but is hard to reconcile with the higher value discussed above. If the silent mutation rate is $\sim 10^{-8}$, the effective population size would have to be even larger than the already implausibly high value discussed above. The divergence times between species within the genus Leavenworthia would then be between 0.1 and 1 million years, and the divergence time of the S and L haplotypes, based on their mean divergence, $D_{xy} = 0.065$, would be about 2 million years. It seems unlikely that the haplotype polymorphism could be maintained for such a long time by drift if the variants were neutral, but this cannot be definitely excluded in the absence of good information about the effective population size of this species.

Finally, we found a possible target of balancing selection, the Asn/Lys polymorphism that correlates with the L type *vs.* S type of the PgiC alleles. Intriguingly, the site of this polymorphism is in exon 8, *i.e.*, within one of the peaks of polymorphism and linkage disequilibrium (Figures 2 and 3). Moreover, it is the only high peak in the region studied that is mostly due to divergence between the S and L haplotypes rather than to polymorphism within the L haplotype (Figure 2). Multiple peaks of polymorphism suggest that this region of the PgiC gene contains several targets for selection. This is consistent with the fact that there are multiple allozyme variants in this locus (four in *L. stylosa*, data not shown). We did not find any correlation between the amino acid replacements and the other peaks of polymorphism; however, this may be due to the small sample

size for the L-type alleles. Multiple allozyme variants in PgiC were also reported for Colias butterflies (Watt 1992) and for field crickets *Gryllus veletis* and *G. pennsylvanicus* (Katz and Harrison 1997). For both these insects there is some evidence that the maintenance of polymorphisms is due to balancing selection. The polymorphism in Colias has not been studied at the DNA level. The data on DNA polymorphism in Gryllus PgiC demonstrates that DNA sequence polymorphisms are not shared between the species and are thus short lived. This is consistent with a conclusion of Hasson *et al.* (1998) that allozyme polymorphisms in Drosophila are short lived. However, our observations suggest that allozyme polymorphisms may be maintained for a long time.

Comparison with the outgroup species, *C. hirsuta*, demonstrates that Asn is the ancestral amino acid at this site, and all but one of the L-type alleles have Asn, while all S-type alleles have Lys at this polymorphic site. Moreover, in *L. alabamica* segregating L and S haplotypes also have Asn and Lys at the polymorphic site, respectively. The change of Asn to Lys changes the charge of the whole protein and it is known that such changes could be selectively important (Riddoch 1993). The observation that polymorphism in S-type alleles is about half of that in L-type alleles suggests that the S type may be younger than the L type. Thus, the following scenario seems to explain the observed haplotype structure in the Leavenworthia PgiC gene. An ancestral species had predominantly L-type and a few S-type alleles with Asn at position 200 of a protein. Due to a mutation of T to A in the third position of this codon, the Asn residue S type mutated to Lys, and the change was advantageous. The frequency of mutant S+Lys-type alleles increased but did not go to fixation, due to either frequency-dependent or overdominant selection. During the subsequent speciation events, *L. alabamica* inherited the ancestral L(Asn)/S(Lys) polymorphism. *L. crassa* probably became fixed due to either small population size or a high rate of selfing, while fixation of different haplotypes in *L. uniflora* and *L. torulosa* is most likely attributable to their high rates of selfing, since diversity is expected to be low in highly inbreeding populations (Charlesworth *et al.* 1993).

We thank Jody Hey and Brian Charlesworth for discussions and advice on analyses, Jody Hey, John Kelly, and Jeff Wall for providing computer programs, and the University of Edinburgh greenhouse staff for plant care. D. Charlesworth was supported by the Natural Environment Research Council of Great Britain, and D. A. Filatov was supported by a grant to D. Charlesworth from the Leverhulme Trust.

LITERATURE CITED

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

Charlesworth, D., and P. Awadalla, 1998 Flowering plant self-incompatibility: the molecular population genetics of Brassica S-loci. Heredity **81:** 1–9.

Charlesworth, D., and Z. Yang, 1998 Allozyme diversity in Leavenworthia populations with different inbreeding levels. Heredity **81:** 453–461.

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

Christiansen, C. S., 1993 A philogenetic approach to floral evolution in the mustard genus, Leavenworthia. B.A. Thesis, Amherst College, Amherst, MA.

Cummings, M. P., and M. T. Clegg, 1998 Nucleotide sequence diversity at the alcogol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA **95:** 5637–5642.

Doebley, J., 1989 Isozymic evidence and the evolution of crop plants, pp. 165–191 in *Isozymes in Plant Biology*, edited by D. E. Soltis and P. S. Soltis. Dioscorides Press, Portland, OR.

Doebley, J., 1992 Molecular systematics and crop evolution, pp. 202–222 in *Molecular Systematics of Plants*, edited by P. S. Soltis, D. E. Soltis and J. J. Doyle. Chapman and Hall, London.

Drake, J. W., B. Charlesworth, D. Charlesworth and J. F. Crow, 1998 Rates of spontaneous mutation. Genetics **148:** 1667–1686.

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

Gaut, B. S., 1998 Molecular clocks and nucleotide substitution rates in higher plants, pp. 93–120 in *Evolutionary Biology*, Vol. 30, edited by M. K. Hecht, W. C. Steere and B. Wallace. Plenum, New York.

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.

Hartl, D. L. and A. G. Clark, 1989 *Principles of Population Genetics*. Sinauer Associates. Sunderland, MA.

Hasson, E., I. N. Wang, L. W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the triosephosphate isomerase (Tpi) locus of Drosophila melanogaster and Drosophila simulans. Mol. Biol. Evol. **15:** 756–769.

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

Henry, A. M., and C. Damerval, 1997 High rates of polymorphism and recombination at the *Opaque-2* locus in maize. Mol. Gen. Genet. **256:** 147–157.

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

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

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

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., and N. L. 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.

Hudson, R. R., D. Boos and N. L. Kaplan, 1992 A statistical test for detecting geographic subdivision. Mol. Biol. Evol. **9:** 138–151.

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

Huttley, G. A., M. L. Durbin, D. E. Glover and M. T. Clegg, 1997 Nucleotide polymorphism in the chalcone synthase-A locus and

evolution of the chalcone synthase multigene family of common morning glory *Ipomoea purpurea.* Mol. Ecol. **6:** 549–558.
Kaplan, N. L., T. Daren and R. R. Hudson, 1988 The coalescent process in models with selection. Genetics **120:** 819–829.
Katz, L. A., and R. G. Harrison, 1997 Balancing selection on electrophoretic variation of phosphoglucose isomerase in two species of field cricket: *Gryllus velvetis* and *G. pennsylvanicus.* Genetics **147:** 609–621.
Kelly, J., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.
Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
Kimura, M., and T. Ohta, 1969 The average number of generations until fixation of a mutant gene in a finite population. Genetics **61:** 763–771.
Kirby, D., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.
Klein, J., 1986 *Natural History of the Major Histocompatibility Complex*. Wiley, New York.
Klein, J., J. Gutknecht and N. Fischer, 1990 The major histocompatibility complex and human evolution. Trends Genet. **6:** 7–11.
Klekowski, E. J., 1988 *Mutation, Developmental Selection, and Plant Evolution*. Columbia University Press, New York.
Klekowski, E. J., and P. J. Godfrey, 1989 Aging and mutation in plants: a comparison of woody mangroves and herbaceous annuals. Nature **340:** 389–391.
Kondrashov, A. S., 1998 Measuring spontaneous deleterious mutation process. Genetica **102/103:** 183–197.
Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and the *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphisms and divergence. Genetics **127:** 565–582.
Kumar, S., K. Tamura and M. Nei, 1993 MEGA: Molecular evolutionary genetics analysis, version 1.0. The Pennsylvania State University, University Park, PA.
Liu, F., 1998 Genetic diversity in Leavenworthia populations with different inbreeding levels. The effect of breeding system on the level and pattern of molecular variation in plant populations. Ph.D. Thesis, University of Chicago.
Liu, F., L. Zhang and D. Charlesworth, 1998 Genetic diversity in Leavenworthia populations with different inbreeding levels. Proc. R. Soc. Lond. Ser. B Biol. Sci. **265:** 293–301.
Liu, F., D. Charlesworth and M. Kreitman, 1999 The effect of mating system differences on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus Leavenworthia. Genetics **151:** 343–357.
McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.
Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.
Muller, J., 1984 Significance of fossil pollen for angiosperm history. Ann. MO Bot. Gard. **71:** 419–443.
Price, R. A., J. D. Palmer and I. A. Al-Shehbaz, 1994 Systematic relationships of Arabidopsis: a molecular and morphological perspective, pp. 7–19 in *Arabidopsis*, edited by C. Somerville and E. Meyerowitz. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Richman, A. D., M. K. Uyenoyama and J. R. Kohn, 1996 Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science **273:** 1212–1216.
Riddoch, B. J., 1993 The adaptive significance of electrophoretic mobility in phosphoglucose isomerase (PGI). Biol. J. Linn. Soc. **50:** 1–17.
Rollins, R. C., 1963 The evolution and systematics of Leavenworthia (Cruciferae). Contrib. Gray Herb. Harv. Univ. **192:** 3–98.
Rozas, J., and R. Rozas, 1997 DNAsp version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.
Shattuck-Eidens, D. M., M. Russel, N. Bell, S. L. Neuhausen and T. Helentjaris, 1990 DNA sequence variation within maize and melon: observations from polymerase chain reaction amplification and direct sequencing. Genetics **126:** 207–217.
Sims, T., 1993 Genetic regulation of self-incompatibility. Crit. Rev. Plant Sci. **12:** 129–167.
Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.
Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.
Takano, T. S., S. Kukasabe and T. Mukai, 1993 DNA polymorphism and the origin of protein polymorphism at the *Gpdh* locus of *Drosophila melanogaster*, pp. 179–190 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer, Sunderland, MA.
Terauchi, R., T. Terachi and N. T. Miyashita, 1997 DNA polymorphism at the *Pgi* locus of a wild yam, *Dioscorea tokoro*. Genetics **147:** 1899–1914.
Thomson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequences alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.
Wall, J., 1999 Recombination and the power of statistical tests of neutrality. Genet. Res. **74:** 65–79.
Watt, W. B., 1992 Eggs, enzymes, and evolution—natural genetic variants change insect fecundity. Proc. Natl. Acad. Sci. USA **89:** 10608–10612.
Wettstein, D. V., K. W. Henningsen, J. E. Boynton, G. C. Kannagara and O. F. Nielsen, 1971 The genic control of chloroplast development in barley, pp. 205–223 in *Autonomy and Biogenesis of Mitochondria and Chloroplasts*, edited by N. K. Boardman and R. M. Smillie. North Holland Press, The Hague.

Communicating editor: W. Stephan