Diversity and Linkage of Genes in the Self-Incompatibility Gene Family in Arabidopsis lyrata

Deborah Charlesworth,*,1 Barbara K. Mable,*2 Mikkel H. Schierup,1 Carolina Bartolomé* and Philip Awadalla*3

*Institute of Cell, Animal and Population Biology, University of Edinburgh, Ashworth Laboratories, Edinburgh EH9 3JT, United Kingdom
and 1Department of Ecology and Genetics, University of Aarhus, DK-8000 Aarhus C, Denmark

Abstract

We report studies of seven members of the S-domain gene family in Arabidopsis lyrata, a member of the Brassicaceae that has a sporophytic self-incompatibility (SI) system. Orthologs for five loci are identifiable in the self-compatible relative A. thaliana. Like the Brassica stigmatic incompatibility protein locus (SRK), some of these genes have kinase domains. We show that several of these genes are unlinked to the putative A. lyrata SRK, Aly13. These genes have much lower nonsynonymous and synonymous polymorphism than Aly13 in the S-domains within natural populations, and differentiation between populations is higher, consistent with balancing selection at the Aly13 locus. One gene (Aly8) is linked to Aly13 and has high diversity. No departures from neutrality were detected for any of the loci. Comparing different loci within Aly, sites corresponding to hypervariable regions in the Brassica Sloci (SLG and SRK) and in comparable regions of Aly13 have greater replacement site divergence than the rest of the Sdomain. This suggests that the high polymorphism in these regions of incompatibility loci is due to balancing selection acting on sites within or near these regions, combined with low selective constraints.

In Brassica, control of pollen-stigma interactions at the stigmatic interface involves highly polymorphic recognition genes of the self-incompatibility (SI) system. It is of interest to understand which regions of the proteins that these genes encode have recognition functions, how this affects the polymorphism in the coding sequence and surrounding genome regions, and how the two recognition genes maintain their coadaptation to produce functional incompatibility types. To understand the evolution of the self-incompatibility loci, it will be helpful to study them in the context of the gene families to which they belong. Doing this allows one to evaluate the possibility of exchanges between loci by gene conversion. It also makes it possible to compare sequence evolution of loci that are involved in incompatibility, and are thus under balancing selection, with similar sequences not under such selection.

The Slocus region contains loci belonging to two distinct gene families. “S-domain genes,” members of the plant receptor-like protein-kinase gene family (Boyes and Nasrallah 1993; Walker 1994), are important for the stigma recognition functions, while pollen specificities are controlled by the SGR gene (Slocus cysteine rich, also called SP11), a member of the pollen coat protein PCP family (Stephenson et al. 1997; Doughty et al. 1998; Schopfer et al. 1999; Takayama et al. 2000; Vanoosthuyse et al. 2001). These different genes are linked in a region whose length differs among different haplotypes (Boyes and Nasrallah 1993; Goring and Rothstein 1996; Yu et al. 1996; Conner et al. 1998). In the Brassica Slocus region, there are two Sdomain genes, SLG (Slocus glycoprotein) and SRK (Sreceptor kinase), with expression chiefly in stigma epidermal cells. The SRK gene is essential for self-incompatibility (Goring and Rothstein 1996; Cui et al. 2000; Takasaki et al. 2000), while SLG, a closely linked Sdomain gene without a kinase domain, is homologous to exon 1 of SRK and is nonessential for recognition, although it may have a role in the incompatibility phenotype (Takasaki et al. 2000).

Further Sdomain genes are known in Brassica and related plants (Luu et al. 2001), most of them not linked to the Slocus (Kai et al. 2001), although linked ones are found in some haplotypes (Suzuki et al. 1997, 1999; Kusaba et al. 2000), some of them apparently pseudogenes (Yu et al. 1996; Kai et al. 2001). Most of the functional members of this gene family are presumably involved in processes other than pollination, although in Brassica some encode secreted stigma glycoproteins (Nasrallah and Nasrallah 1993) and some Sdomain proteins are necessary for correct pollen-stigma adhesion (Luu et al. 1997; Takayama et al. 2000). In Arabi-
dopis thaliana the S-domain gene family has ~40 members (SHU and BLEcker 2001), and the Brassica SRK sequences are most similar to the A. thaliana Ark genes (PASTUGLIA et al. 2002). Other members of this gene family are not linked to the Slocus.

Studies of sequence diversity of Brassica Slocus genes have until recently concentrated on the SLG gene, but some data from the Sdomains and a portion of the kinase domain of SRK of some haplotypes have been published (KUSABA et al. 1997; NISHIO et al. 1997). The broader view of the Slocus genes as members of a gene family has not been emphasized in Brassica, although some data are from loci other than SLG and SRK, and it is clear that both SRK and SLG genes are much more polymorphic than other Sdomain genes that have been studied (DWYER et al. 1991; HINATA et al. 1995; KUSABA et al. 1997; WATANABE et al. 1997, 1998; SAKAMOTO et al. 1998). For loci other than SLG and SRK, sample sizes are very small. SCR also appears to be highly polymorphic (WATANABE et al. 2000; KIMURA et al. 2002), although no comparison of diversity with other members of this gene family has been published. The Brassica data are mainly from cultivars, not from random samples from natural populations.

Here we report results of population genetic studies of several Sdomain loci in natural populations of A. lyrata. We characterize diversity at several different Sdomain loci for comparison with SRK to establish whether SRK indeed has an unusually polymorphic Sdomain, as expected for a gene under balancing selection. Balancing selection is not expected for Sdomain genes that are not involved in SI (although they could have experienced other forms of selection, for example, directional selection subsequent to gene duplication in the evolution of the gene family).

A. lyrata is a self-incompatible, predominantly diploid member of the Brassicaceae, but distantly related to Brassica (ROLLINS 1993). Silent site divergence of putatively orthologous genes between A. lyrata or A. thaliana and Brassica ranges from ~0.2 to >1 without Jukes-Cantor correction (reviewed in WRIGHT et al. 2002). We previously described several Sdomain loci in A. lyrata (CHARLESWORTH et al. 2000; SCHIERUP et al. 2001). We refer to these as the Aly loci. Of these, Aly13 is highly polymorphic within A. lyrata, and segregation with incompatibility groups in families suggests that it is the ortholog of the Brassica SRK gene (CHARLESWORTH et al. 2000; SCHIERUP et al. 2001). The same gene was identified by KUSABA et al. (2001), who isolated the complete sequence, including the kinase domain, from stigma mRNA of an A. lyrata plant heterozygous for two Salleles and also showed cosegregation of sequence variants with the two progeny incompatibility groups in self-fertilized progeny; they named the gene A. lyrata SRK. Two sequences in our families closely match those of the two alleles in the plant studied by KUSABA et al. (2001); Aly13-13’s Sdomain is almost identical to that of the SRKα allele, while Aly13-20 matches allele SRKβ.

As expected for the Slocus, Aly13 sequences are exceptionally polymorphic at both synonymous and replacement sites (SCHIERUP et al. 2001). Sdomain diversity is even higher than in the Sdomains of Brassica SRK or SLG loci (HINATA et al. 1995; CHARLESWORTH and AWADALLA 1998). To evaluate the variability, it is important to compare it with that of other loci. As currently few diversity data are from natural populations of plants, including A. lyrata, we have obtained new diversity data by studying other Sdomain loci, which are the ideal “reference” loci for tests of whether Aly13 is more polymorphic than other loci.

Data from other Aly loci also allow us to compare levels of selective constraint in different regions of the Sdomain. It is often suggested that the hypervariable (HV) regions in the extracellular Sdomain are the most important for recognition (NASRALLAH and NASRALLAH 1989; KIMURA et al. 2002). However, this remains uncertain, and effects of amino acid differences elsewhere in the domain are also likely (NASRALLAH 1997; MIEGE et al. 2001). Variability in the Slocus region will be affected by a number of factors, and it is important to distinguish the possibilities clearly.

First, in regions of the protein where amino acid variants alter specificities, balancing selection will promote variation, so we expect high nonsynonymous diversity in the genomic sequences, as observed in both Brassica (e.g., NASRALLAH and NASRALLAH 1989) and A. lyrata (SCHIERUP et al. 2001). Second, closely linked synonymous sites will also have high diversity, because balancing selection maintains different functional alleles for long time periods (e.g., WRIGHT 1939; VEKEMANS and SLATKIN 1994), allowing sequence differentiation of alleles (e.g., STROBECK 1972; HUGHES et al. 1990; TAKAHATA 1990; NORDBORG et al. 1996); this prediction also fits the data on Salleles (NASRALLAH and NASRALLAH 1989; CHARLESWORTH and AWADALLA 1998; SCHIERUP et al. 2001). Amino acid variants may be affected in the same way, so that such variants need not all be associated with specificity differences. For the same reason, other genome regions closely linked to the Slocus may also have high polymorphism. A number of cases of additional Sdomain genes located close to the Brassica Slocus are known, and it is of interest to see whether they are highly polymorphic (as expected if linkage to the Slocus is very close) or have low diversity (which would imply that recombination occurs and that diversity is high only at sites that are physically very close to sites under balancing selection). We have already mentioned the highly polymorphic SLG locus in the Brassica Slocus region, but some other linked genes have been reported to have low diversity (HINATA et al. 1995). Among the Aly loci studied here, we find no A. lyrata ortholog of SLG (in agreement with KUSABA et al.
2001), but we observe high polymorphism in a different linked S-domain gene (Aly8).

A third important influence on diversity is that selective constraints may differ between different regions of the protein (NASRALLAH 1997). Low selective constraints may allow certain regions to have particularly high polymorphism due to linkage to the sites under balancing selection. Different selective constraints can be detected using comparisons with reference loci, and we use this approach to test whether this could contribute to the high variability of the hypervariable regions. A sample of seven putative alleles showed that the Aly13 S-domain sequences have peaks of replacement site polymorphism in the regions equivalent to the Brassica Slocus HV regions (SCHIERUP et al. 2001). To test whether this is because the same regions are important for recognition functions in both species, we must exclude the possibility that these regions are low constraint regions.

A final reason for studying other loci is that the analysis of sequence data to infer selection is complicated by population subdivision. To assess the effects of demographic and historical processes that can generate patterns that may be mistaken for evidence of selection, it is necessary to have reference loci that are not under strong balancing selection, but instead are evolving more or less neutrally. For example, genetic differentiation between populations can cause haplotype structure that may be difficult to distinguish from balancing selection unless other data are available to show the true situation (e.g., CHARLESWORTH et al. 1997). Tests for selection such as Tajima’s (1989a) test are thus affected by population subdivision and will often be unable to detect selection when samples from such populations are pooled (SCHIERUP et al. 2000). When there is subdivision, this test cannot distinguish between balancing selection and differentiation between populations, because Tajima’s D is affected by both selection and population subdivision (which, like balancing selection, causes positive D values; Tajima 1989a,b). Directional selection at a locus causes negative D values, and pooling is therefore conservative (ignoring subdivision may obscure this form of selection). Data on diversity at the Aly13 locus (the putative A. lyrata SRK) must therefore be interpreted in the light of such information. Finally, a locus experiencing balancing selection is expected to show less population subdivision than loci not experiencing such selection (SCHIERUP et al. 2000).

Here we describe analyses of diversity and selection at several A. lyrata S-domain loci and assess the implications for our understanding of balancing selection and its effects on sequence diversity within Sloci and in their genomic neighborhood.

MATERIALS AND METHODS

A. lyrata plant material and DNA preparation: Seeds were collected from four populations of A. lyrata (see details in CHARLESWORTH et al. 2000). A. lyrata populations are widely distributed in northern North America, and European populations formerly classified as A. petreae are now considered to be the same species as A. lyrata, so both are referred to here as A. lyrata (following KOCH et al. 2000). The populations, and the abbreviations we shall use for them, were as follows: Two samples from North America (kindly provided by R. Mauricio) were from North Carolina (NC) and Indiana Dunes, Indiana (IN), and two from Europe were from Braemar, Scotland (kindly provided by R. Ennos), and the Reykjanes Peninsula, Iceland (some of them provided by E. Thorhallsdottir). Individual plants were grown in the greenhouse from seeds from these populations. DNA was extracted from leaves using a CTAB protocol (JUNGHANS and METZLAFF 1990).

Primers, amplification, cloning, and sequencing: S-domain primers: Primers were designed on the basis of sequence alignments of Brassica SLG and SRK loci (Table 1) and used to amplify A. lyrata genomic DNA. Because SLG and SRK are members of a gene family, our initial primers were based on the most conserved regions of the Brassica Sdomain and should amplify multiple A. lyrata S-domain genes, particularly those most similar to SRK. The S-domains of most Brassica oleracea and B. campestris SLG and SRK alleles have no introns (TANTIKJANJANA et al. 1993; HATAKEYAMA et al. 1998; GABRIELLAC et al. 1999), so A. lyrata S-domains should be similar in length to those of Brassica. To distinguish between sequences from different loci, specific primers were designed from the A. lyrata sequences amplified. The primer sequences and amplification conditions are given in CHARLESWORTH et al. (2000).

Kinase domain reverse primers: To test whether each S-domain sequence had a kinase domain downstream from the S-domain, we used specific forward primers for the loci identified in A. lyrata with reverse primers based on either Brassica SRK locus kinase domains (srknasr and srk5r) or an A. lyrata SRK kinase sequence kindly provided by J. B. Nasrallah (srknasr1, srknasr4, and srknasr5; see Table 1).

Cloning and sequencing: Because some primers amplify more than one locus, and also because of the high variability of some of the putative loci (see below), PCR products of the expected size were generally cloned before sequencing [using the Invitrogen (San Diego) TOPO TA cloning kit]. To detect sequence variants and differentiate between loci (see below), the cloned amplification products were digested with four-and six-cutter restriction enzymes and fragments were separated electrophoretically.

Sequences were obtained using standard cycle sequencing protocols for the Applied Biosystems (Foster City, CA) model 377 sequencer; with the Big Dye sequencing kit, using M13 universal primers for clones; or by direct sequencing using primers specific to the original amplified product. All Aly3, Aly9, Aly10.1, and Aly14 were sequenced directly. Sequences of the more variable loci were sequenced from cloned PCR products. In most cases, at least two clones were sequenced per individual to check for PCR errors. However, since this was not always done, some diversity values may be slightly overestimated, and an excess of singletons may have been produced; this does not affect our general conclusions. Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under POPSET accession nos. AY186752–AY186777. The full population sequence set can be obtained from the authors by request.

Sequence alignments and analyses: Sequences were aligned using ClustalX (THOMPSON et al. 1994), followed by manual adjustments based on the inferred amino acid alignments using Se-Al v. 1.0 (RAMBAUT 1996). A. lyrata S-domain sequences were aligned to two representative B. oleracea SRK alleles from class I (pollen dominant: SRK9 and SRK45) and...
sequences and to several A. thaliana S-domain genes (see Table 2 for accession numbers and chromosomal locations). The nucleotide sequence of the apparent A. thaliana ortholog of Aly13 (Kusaba et al. 2001) was also included (T6K22.100: accession no. AL031187). Conserved amino acid residues common to S-domain loci (Kusaba et al. 1997) were used to anchor the alignments. For visual comparison of relative divergence within and between sequence types, phylogenetic trees were reconstructed with PAUP* 4.0b10 (Swofford 2002), using the minimum evolution function under an HK85 model of substitution (Hasegawa et al. 1985). Heuristic searches were conducted with initial trees obtained by simple stepwise addition, followed by branch swapping using the tree bisection-reconnection routine implemented in PAUP*. Relative support for individual nodes was assessed by bootstrapping (1000 replicates) using neighbor joining (Saitou and Nei 1987). For these analyses, sequences from Aly14 and their putative ortholog Atl4 (see discussion below) were excluded because of the short length of the Aly14 sequences. For the other sequence types, all unique sequences were included in the analysis, except for Aly13, for which a subset of 10 alleles with known linkage relationships (Schierup et al. 2001) was included. The analysis was based on the region from bp 496 to 1365 of the alignment.

To estimate nucleotide divergence between sequences, synonymous (Ks) and nonsynonymous substitutions per site (Kv) were calculated using the method of Nei and Gojobori (1986) with the MEGA2 program (Kumar et al. 2000). For divergence between pairs of paralogous loci within species or for divergence between species, all regions present in both sequences were included. The comparisons therefore involve slightly different lengths of sequence (see Figure 2 and Table 2).

Nonsynonymous and synonymous diversity values (πs and πv) within species were estimated using a set of putative alleles of the different loci sequenced from a common small sample of individuals from the four populations. (Some individuals did not yield sequences for some loci, sometimes because the DNA sample was used up, so we were unable to obtain exactly the same samples for all loci; Table 5 shows the sample sizes.) The MEGA2 software was used for diversity estimation. Proseq v. 2.9 (Filatov and Charlesworth 1999) was used to do Tajima’s D (1989a) tests and to test for population subdivision in A. lyrata using the Ks statistic (Hudson et al. 1992). These analyses used all sequences available from the study populations for each locus; i.e., sequences were included that were excluded by the sampling scheme used to estimate within-population diversity. The critical values for Ks were obtained by 1000 random permutations of the sequences between the populations (Hudson et al. 1992).

Recombination in the Aly loci was tested by two types of analysis. Correlation analyses were done with the r2 program written by M. H. Schierup (http://www.bricks.dk/~compbio/r2). Only segregating sites with frequencies >0.1 were included. Where only one or two sequences had gaps, the site was included (although the sequences with gaps were excluded from the analysis); otherwise, gap regions were excluded from all the sequences. Significance of the correlation coefficients of two measures of linkage disequilibrium with distance (r2 or D) was determined using 5000 random permutations of the variable sites. The second analysis used the composite likelihood finite sites extension to Hudson’s (2000) method (McVean et al. 2002), implemented in the LDhat program (http://www.stats.ox.ac.uk/~mcvean). All sites were included in this analysis. Maximum-likelihood estimates of the scaled recombination rate were obtained from the likelihood curve evaluated from 20 points between ρ = 0 and ρ = 50. Significance against ρ = 0 was tested by 1000 random permutations. Minimum numbers of recombination events were estimated by Hudson and Kaplan’s (1985) estimator using DNAsp v. 3.5 (Rozas and Rozas 1999).

To test whether the parts of the S-domain sequences that are hypervariable in the Aly13 sequences also evolve unusually in the other A. lyrata S-domain genes, we compared the levels of divergence and polymorphism in different regions of Aly13 with diversity in the sequences of each putative locus. Following Kusaba et al. (1997), the positions of the HV regions in the S-domains of the Brassica SLG and SRK type I alleles initially recognized by Dyer et al. (1994) correspond to the following positions in our alignments: HV1 corresponds to amino acid residues 215–236; HV2 to 293–331; HV3 to 354–368; and the C-terminal region to residues 446–456. McDon-ald-Kreitman tests (McDonald and Kreitman 1991) were used to test for selection acting on the A. lyrata S-domain loci using DnaSP v. 3.5. To test whether more changes have occurred in any of the A. lyrata Aly genes, compared with their orthologs in A. thaliana genes since their divergence from outgroup sequences, relative rates were evaluated using Tajima’s one-parameter test (Tajima 1993), which does not assume any specific substitution model. As no sequence data are from orthologs of closely related species, the closest paralogous A. thaliana S-domain gene was used as an outgroup for each comparison (for Aly8/Ark3, this was AtSL1; for Aly7/Ark1, this was Ark1; for Aly8/Ark1, this was Ark3; for Aly10.1/Ark1, this was Ark2; and for Aly10.2/Ark2, this was Ark1).

Tests for linkage between the A. lyrata self-incompatibility locus and the S-domain loci: To test Aly S-domain loci for linkage to the self-incompatibility locus, we used full-sib families in which the incompatibility groups of progeny plants had been determined by hand-pollinations between individuals or in which Aly13 genotypes had been determined, so that it was known that one or both parent plants were Aly13 heterozygotes (see Schierup et al. 2001). The two parents and the progeny were also scored using PCR amplification with primers specific for any of the other Aly sequence types that sequencing showed to be heterozygous in one or the other parent. We used digestion with restriction enzymes to score variants for these putative loci; the details are described below.

RESULTS

Amplification of S-domain sequences: Several combinations of primers designed to match conserved regions within the first exon (the S-domain) of the Brassica S gene family (see MATERIALS AND METHODS) were used in the initial screening of A. lyrata genomic DNA for S-domain sequences. Amplifications from a single individual from Scotland yielded PCR products of the size predicted from Brassica S gene sequences. Five different sequence types were initially identified using the six-cutter restriction enzymes EcoRI, HindIII, and BamHI, and these were sequenced (Aly7, Aly8, Aly9, Aly10, and Aly13; Charlesworth et al. 2000; Schierup et al. 2001).

BLAST searches of these sequences showed homology to Brassica and A. thaliana S-domain loci (see below), and they could all readily be aligned with Brassica Sallele sequences and with members of the S-domain gene family in A. thaliana. The portions of the S-domains sequenced (see Figure 2) in all the Aly loci identified, with the exception of Aly7 and some of the Aly10.1 alleles (see below), are open reading frames. There are
and fall into strongly supported groups in a phylogenetic analysis (Figure 1), consistent with the other evidence that they represent distinct loci. The five A. thaliana gene sequences in Figure 1 and Table 2 will be discussed later. Unlike these putative S-domain loci, no single primer amplifies all sequences of the A. lyrata type (Mable et al. 2003). The A. lyrata sequences are highly variable, consistent with other data suggesting that these sequences represent a single highly polymorphic A. lyrata Slocus, rather than several different loci (Schierup et al. 2001).

**A. thaliana orthologs of the A. lyrata genes and structure of the loci:** Comparing the A. lyrata sequences with A. thaliana S-domain receptor kinase genes (Table 2), we can identify probable orthologs for five genes (Figure 1). For three loci (Aly3, Aly7, or Aly9), we could not identify kinase domains by PCR (see Materials and Methods); for brevity, we refer to these as “nonkinase domain” sequences, although it is possible that a kinase domain exists but was not detected. No orthologs can be identified for two of these loci, Aly3 and Aly7. For Aly9, AtS1 is a potential ortholog (see Table 2 for accession numbers). This is the probable ortholog of SLR1 in Brassica (Dwyer et al. 1994) and, like Aly9, appears not to have a kinase domain.

We tentatively identify orthologs of Aly8, Aly10.1, and Aly10.2 as the three kinase domain loci Ark3, Ark1, and Ark2, respectively (Table 2). These three A. lyrata genes have quite similar S-domains, which amplify with the same forward primers (see Table 1), and kinase domains were detectable for all three. A possible ortholog of Aly14 (which we have not tested for the presence of a kinase domain) is an anonymous kinase domain sequence, which we denote by Aly14 (contig accession AL161566.2|ATCHRIV66, position 170101).

For the putative orthologous pairs, silent site divergence from A. thaliana for the S-domain ranges from 18 to 30%, and replacement site divergence is between
Aly in the 6th amino acid of the motif (F replaces Y), and only slightly lower than the divergence between the Aly and in several sequences. The motif is present in SRK S-domain kinases (S-ferred amino acid sequences all share the 12 cysteine so this does not conclusively rule out the possibility of A. lyrata A. thaliana orthologs within the regions that were se- the sequences appear to have recombined. Alterna-

divergence exceeding /H11601 Aly Aly and in several 7 sequences: 3 and 12% (see Table 2). Divergence values in the three kinase domain sequence types were similar (based on the coding sequence of exons 1–7, synonymous and nonsynonymous divergence values between Aly8 and Ark3 were 0.34 and 0.049, respectively; for Aly10.1 vs. Ark1, the values were 0.22 and 0.07, and for Aly10.2 vs. Ark2, 0.19 and 0.065). The values for both the S- and kinase domains are higher than values for most orthologous sequence comparisons between these two species (WRIGHT et al. 2002), but within reasonable limits (silent divergence exceeding ~30% after Jukes-Cantor correction is unlikely for true orthologs). The high divergence for Aly8 is in part due to its diversity within A. lyrata (discussed in more detail below).

Figure 2 summarizes the structure of the S-domain sequence types identified in A. lyrata and their putative A. thaliana orthologs within the regions that were se-
ed the sequences appear to be evolving the same is true for the four putative A. lyrata loci. The inferred amino acid sequences all share the 12 cysteine residues present in the Brassica SLG, SRK, and SLR S-domain sequences of other Brassicaceae (KUSABA et al. 1997), as well as the A. thaliana Ark loci. We searched our sequences for the motif of 10 amino acids described in the S-domains of other plant receptor-like protein kinases (WALKER 1994; WQSFDPYPTDT in all Brassica SRK sequences). The motif is present in Aly8 and Aly9 and in several Aly13 sequences (13-2, -3, -4, -7, -8, -14, and -20). The Aly8, Aly10.1, and Aly10.2 sequences differ in the 6th amino acid of the motif (F replaces Y), and the same is true for Aly13 sequences 13-1, -5, -9, -13, -15, and -23 (Figure 2).

Aly7 sequences: As mentioned previously, some Aly7 sequences have a single base-pair insertion (at position 751 within a region of four TA repeats; see Figure 3, “Aly7(+)” sequences). This disrupts the reading frame and creates a downstream stop codon. A 9-bp insertion is also present in this set of sequences, relative to the in-frame, Aly7(−) sequences (Figure 3). Overall, 26% out of a total of 27 Aly7 sequences classified (either by sequencing or by amplifying with primers specific to each haplotype; see Table 1) were of the 7(+) type. These sequences could represent a separate locus or could be allelic to the other Aly7 sequences. The sequences with and without the insertion form two haplo-
types. There is significant linkage disequilibrium between the two types, between sites separated by 550 nucleotides, and between closer sites (Figure 3). This might suggest two distinct loci, but there are few pairs of sites for which linkage disequilibrium is complete, so the sequences appear to have recombined. Alternatively, interlocus gene conversion may have occurred, so this does not conclusively rule out the possibility of two loci. Nucleotide similarity is otherwise high between the two sequence types. Removing the insertion from the Aly7(+) sequences, the mean divergence from the other Aly7 sequences is 0.017 for synonymous sites, and slightly higher (0.019) for nonsynonymous sites, sug-
esting that the sequences are evolving neutrally. However, net divergence is very small, given the diversity within the Aly7(−) sequences. The sequence diversity of the Aly7(+) sequences is lower than that of Aly7(−), and only slightly lower than the divergence between the two, and again the sequences appear to be evolving neutrally (among Aly7(+) sequences, π, = 0.0098 and πe = 0.0107). Tajima’s D is significantly negative (D = −1.65, P < 0.05) for the Aly7(+) sequences, but is not

### TABLE 2

<table>
<thead>
<tr>
<th>Locus</th>
<th>Accession nos.</th>
<th>Length of sequencea</th>
<th>Aly3</th>
<th>Aly7-</th>
<th>Aly8</th>
<th>Aly9</th>
<th>Aly10.1</th>
<th>Aly10.2</th>
<th>Ark1</th>
<th>Ark2</th>
<th>Ark3</th>
<th>Ats1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aly3</td>
<td>AY186752</td>
<td>701</td>
<td>0.945</td>
<td>1.383</td>
<td>0.735</td>
<td>1.380</td>
<td>1.456</td>
<td>1.380</td>
<td>1.097</td>
<td>0.936</td>
<td>0.637</td>
<td></td>
</tr>
<tr>
<td>Aly7-</td>
<td>AY186753-4</td>
<td>862</td>
<td>0.494</td>
<td>&gt;2</td>
<td>1.469</td>
<td>&gt;2</td>
<td>&gt;2</td>
<td>&gt;2</td>
<td>&gt;2</td>
<td>&gt;2</td>
<td>1.955</td>
<td>1.721</td>
</tr>
<tr>
<td>Aly8</td>
<td>AY186755</td>
<td>772</td>
<td>0.304</td>
<td>0.480</td>
<td>0.849</td>
<td>0.354</td>
<td>0.380</td>
<td>0.648</td>
<td>0.511</td>
<td>0.305</td>
<td>0.884</td>
<td></td>
</tr>
<tr>
<td>Aly9</td>
<td>AY186756</td>
<td>918</td>
<td>0.445</td>
<td>0.501</td>
<td>0.499</td>
<td>0.721</td>
<td>0.684</td>
<td>0.994</td>
<td>0.791</td>
<td>0.865</td>
<td>0.176</td>
<td></td>
</tr>
<tr>
<td>Aly10.1</td>
<td>AY186757-60</td>
<td>938</td>
<td>0.392</td>
<td>0.521</td>
<td>0.187</td>
<td>0.535</td>
<td>0.159</td>
<td>0.311</td>
<td>0.275</td>
<td>0.482</td>
<td>0.877</td>
<td></td>
</tr>
<tr>
<td>Aly10.2</td>
<td>AY186761</td>
<td>713</td>
<td>0.416</td>
<td>0.586</td>
<td>0.188</td>
<td>0.538</td>
<td>0.124</td>
<td>0.403</td>
<td>0.192</td>
<td>0.452</td>
<td>1.020</td>
<td></td>
</tr>
<tr>
<td>Ark1 (chromosome 1)</td>
<td>M80238</td>
<td>0.367</td>
<td>0.462</td>
<td>0.226</td>
<td>0.405</td>
<td>0.067</td>
<td>0.159</td>
<td>0.293</td>
<td>0.527</td>
<td>0.961</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ark2 (chromosome 1)</td>
<td>AY045777</td>
<td>0.394</td>
<td>0.517</td>
<td>0.221</td>
<td>0.414</td>
<td>0.132</td>
<td>0.049</td>
<td>0.112</td>
<td>0.495</td>
<td>0.822</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ark3 (chromosome 4)</td>
<td>AL031187</td>
<td>0.283</td>
<td>0.396</td>
<td>0.032</td>
<td>0.391</td>
<td>0.182</td>
<td>0.213</td>
<td>0.178</td>
<td>0.172</td>
<td>0.867</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ats1 (chromosome 3)</td>
<td>S84921</td>
<td>0.419</td>
<td>0.470</td>
<td>0.390</td>
<td>0.119</td>
<td>0.502</td>
<td>0.487</td>
<td>0.423</td>
<td>0.416</td>
<td>0.379</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The divergence values involving A. lyrata sequences are means using all sequences obtained from each of the Aly loci and were calculated by the method of Nei and Gojobori (1986) with Jukes-Cantor correction. Divergence values between the putative A. lyrata genes and their A. thaliana orthologs are underlined.

a The column of sequence lengths gives the longest sequence available for each locus. For comparisons with the A. thaliana sequences, only the portions of sequences available from both species were used, so that the lengths are not given separately for these loci.
significant for the 7(−) sequences. Although relationships within loci were not well resolved (Figure 1), no evidence was found for separation of the Aly7(+) and Aly7(−) sequence types in the gene tree.

If the Aly7 gene is duplicated, both sequences should be detectable in all individuals, but this is not the case. Aly7(+) sequences have been found in only three of the four populations studied (two in the North Carolina population, one in the Indiana population, and three in the Scottish population). This suggests a single locus or else a duplication that is absent from the Iceland population. Consistent with the single-locus hypothesis, we find plants with both sequence types (apparent heterozygotes) as well as apparent homozygotes, and three individuals heterozygous for two different 7(−) sequences had no sign of sequences with the frame-altering 7(+) insertion. Finally, if the two haplotypes represent alleles, the haplotypes should segregate in the progeny of the apparent heterozygotes. Using primers specific for the two different haplotypes to score 11 progeny (family 99E-10) of such a plant (98E17-4), crossed with an Aly7(−) homozygote (98E17-6), 5 were apparent heterozygotes and 6 apparent homozygotes (−/−); i.e., we find the expected 1:1 ratio. We therefore conclude that the Aly7(+) sequences are probably null alleles of the same locus as the Aly7(−) sequences. In the further analyses below, the Aly7(+) sequences containing the frameshift are omitted.

**Aly10.1 sequences:** Four types of Aly10.1 alleles have been found, and they are shown in Figure 2. Relative to the type “A” sequences, “B1” sequences have a 227-bp deletion beginning 99 bp from the end of the S-domain and leaving only 7 bp of intron 1 and an in-frame stop codon 5 bp before the deletion. “B2” sequences have a further deletion of 25 bp starting at bp 654, which changes the reading frame, while “B3” sequences have a 223-bp deletion starting at bp 446 (which also changes the reading frame). The A allele type, which presumably encodes a functional protein, is the commonest (77% overall, out of 44 alleles sequenced) and is present in all four populations studied, whereas B1 and B2 alleles were seen only in the U.S. populations, and B3 only in the Scottish population. Apart from a single B1/B2 plant, all individuals had at least one allele of type A. No evidence of grouping by alleles was found in the gene tree analysis (Figure 1).

---

**Figure 1.**—Unrooted gene tree of the *A. lyrata* S-domain sequences, together with two sequences each of Brassica class I and II SRK alleles, chosen to represent the full SRK diversity, and putative *A. thaliana* orthologs. The two *A. lyrata* SRK sequences of Kusaba et al. (2001) are denoted by AlySRKa and AlySRKb (note that these sequences are the same as Aly13-13 and -20, respectively). For Aly3, Aly7, and Aly10.1, all unique sequences found are included, whereas for Aly13, 10 representative sequences with known linkage relationships are shown (Schierup et al. 2001; the accession numbers are AF328990–AF329000 and AY186763–AY186777). Aly14 and its putative ortholog are not included. The tree is based on nucleotide distances using the HKY85 substitution model and the minimum evolution function. Bootstrap values exceeding 80% (based on 1000 neighbor-joining replicates) are indicated on the tree. The tree is rooted using *A. thaliana* orthologs from *A. thaliana* for several other sequence types (Table 2) were also strongly supported. Accession numbers of the Brassica SRK alleles are as follows: SRK5, Y18259; SRK15, Y18260; SRK45, E15795; and SRK9, D30049. Accession numbers for *A. thaliana* sequences are given in Table 2.
Figure 2.—Structure of the different S-domain sequence types. Solid blocks indicate the portions of the S-domains sequenced in *A. lyrata* and the corresponding regions of the putative *A. thaliana* orthologs (see text). Alignment gaps relative to Brassica S-domains are indicated by open regions, and their positions and the approximate positions of the HV regions in the amino acid sequence of the S-domain are shown at the top. Indels that were polymorphic among sequences from a given locus are indicated by the "P." For sequences in which an intron has been detected 3' to the S-domain and a kinase domain is present 3' to this, these regions are indicated at the end of the diagrams; details of the kinase domains or the introns within these domains are not shown, and these regions are not drawn to scale. Cysteine residues conserved in all the loci are indicated by "C" above the diagram, and the sequences of the motif characteristic of S-domain kinases (Walker 1994) are shown. For *Aly10.1* sequences, the positions of the large deletions discussed in the text are shown as vertically hatched regions for the four different types of sequence (10.1 A, B1, B2, and B3); the deletions are shown relative to the *Ark1* sequence.
Tests for linkage between the *A. lyrata* self-incompatibility locus and the $S$-domain loci: We tested for linkage of the *Aly* $S$-domain loci and the self-incompatibility gene family, using families whose parents were heterozygous for one or more *Aly* loci (see MATERIALS AND METHODS). Linkage between *Aly*13 variants and the $S$-locus in both sibships, and in several other families, has already been reported (Charlesworth et al. 2000; Schierup et al. 2001). Since variants that do not segregate as alleles of course define different loci, we hoped that tests for linkage would help to show whether or not sequence variants are from a single locus, in addition to what can be deduced from similarities and differences between the sequences.

For the sibship 98E-15 (see Charlesworth et al. 2000; Schierup et al. 2001), direct sequencing reactions for *Aly*3, *Aly*8, *Aly*10.1, *Aly*10.2, and *Aly*9 from the parents of this family showed that the parent 97F13-5 was heterozygous for a number of sites in two loci (Table 3). Two restriction enzymes, *Aci*I and *Bpu*AI, were used to test for segregation of *Aly*3 alleles. Sequence variants at *Aly*3 behave as allelic, and this locus frequently recombined with the $S$-locus. In the same family, a length polymorphism at the end of the $S$-domain of the *Aly*10.1 locus segregates as expected if the variants are allelic (Table 3) and also recombine with the $S$-locus. The results for another sibship, 98G-23, confirmed this conclusion for *Aly*10.1.

For *Aly*8, however, some variants showed linkage. Table 4 shows another sibship in which both parents were double heterozygotes and in which the *Aly*8 variants again cosegregated with *Aly*13 sequences. Linkage of variants was detected in several other sibships by scoring *Aly*13 variants known from our previous work to show linkage to the $S$-locus (*Aly*13-4, -5, -9, -13, -16, and -22 from several different natural populations). These results for *Aly*8 are consistent with the fact that *Ark*3, its *A. thaliana* ortholog, is linked to the putative SRK ortholog of this species, which is a pseudogene (Kusaba et al. 2001). However, other very similar sequences, in addition to two different linked variants, were sometimes amplified with primers for *Aly*8, suggesting that *Aly*8 represents two separate loci. The linkage of the second locus with the $S$-locus is not known, and it is even possible that haplotypes vary in the numbers of *Aly*8 genes linked to *Aly*13, similarly to the variable number of linked *SCR* genes in *A. lyrata* (Kusaba et al. 2001). Until physical maps of different haplotypes are available, this cannot be resolved.

To examine further the possibility of paralogous loci, we aligned all our *Aly*8 sequences to test whether they cluster into two sets with fixed differences between them. However, we found no evidence for any such haplotype structure in the complete sequence data set. Moreover, we could not identify variants that characterize the set of linked or the set of recombining sequences from several families. In other words, there are no sites in linkage disequilibrium that allow us to define site
TABLE 3
Results of linkage tests between Aly loci and the self-incompatibility locus in family 98E-15

<table>
<thead>
<tr>
<th>Plant</th>
<th>SI group</th>
<th>Genotypes at the loci tested</th>
<th>Aly13</th>
<th>Aly13*</th>
<th>Aly10.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent 97F-13/5</td>
<td>Not tested</td>
<td>1/13</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>Parent 97F 15-3</td>
<td>Not tested</td>
<td>3/23</td>
<td>Aa</td>
<td>A/B</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>I</td>
<td>1/23</td>
<td>Aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>IV</td>
<td>3/13</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>I</td>
<td>1/23</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>II</td>
<td>13/23</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>II</td>
<td>13/23</td>
<td>Aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>II</td>
<td>13/23</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>III</td>
<td>1/3</td>
<td>Aa</td>
<td>A/B</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>III</td>
<td>1/3</td>
<td>aa</td>
<td>A/B</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>IV</td>
<td>3/13</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>I</td>
<td>1/23</td>
<td>Aa</td>
<td>A/B</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>IV</td>
<td>3/13</td>
<td>aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>II</td>
<td>13/23</td>
<td>aa</td>
<td>A/B</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>III</td>
<td>1/3</td>
<td>Aa</td>
<td>A/B</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>IV</td>
<td>3/13</td>
<td>Aa</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>IV</td>
<td>3/13</td>
<td>Aa</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

Sequences of Aly3, Aly8, and Aly10.1 were amplified from genomic DNA of the parental and progeny plants, using primers specific for these sequence types (Aly3 primers 3F + 3R and Aly10.1 primers Ap8F + srk5R, respectively).

* The restriction enzyme BsuAI was used to distinguish the “A” and the “a” sequence types.

TABLE 4
Results of linkage tests between Aly8 variants and the self-incompatibility locus in family MS00-36

<table>
<thead>
<tr>
<th>Aly13 genotype</th>
<th>Nos. of plants with Aly8 band</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aly13-11</td>
<td>4/9</td>
</tr>
<tr>
<td>Aly13-9</td>
<td>4/11</td>
</tr>
<tr>
<td>Aly13-4</td>
<td>9/16</td>
</tr>
<tr>
<td>Aly13-16</td>
<td>11/16</td>
</tr>
</tbody>
</table>

Segregation of the Aly8 variants was scored with Alu and Msp1 digestion, and each column of the table is for one of the four alleles present in the parents (given at the top of the four columns in the table). The Aly13 alleles in the maternal (Icelandic) and paternal (Scottish) parents were, respectively, 9/11 and 4/16, and each column also specifies which allele cosegregates with each of the Aly8 alleles in this family.

Slocus (see above). Table 5 summarizes the results for each of the four populations studied, as well as for the total sample. The mean within-population synonymous site diversity values (πs) are mostly < 1%. πs/πa, values, based on the within-population diversity values (Table 5), are mostly rather high for the loci studied here and for the Aly13 locus (although the very high Aly9 value is based on few variable sites). The extremely high diversity for Aly8 (species wide πa = 7.5% and πs > 1%) is partly, but not entirely, attributable to the fact that this sequence type could represent more than one locus, as just explained. This is discussed further below.

Within-population diversity: When all sites were used in the analysis, tests for spatial structure (Hudson et al. 1992) detected significant population subdivision in A. lyrata for several loci. The exceptions are Aly10.1 (with low diversity), the highly polymorphic Aly13 locus, and Aly8 (Table 5). Aly8 shows some evidence of population structure for silent sites (a moderate Kst value, significant at the 5% level), but this is difficult to interpret, given our inability to assign different sequences to individual loci. Overall, the data therefore indicate isolation between the geographically distant populations studied. Clearly, however, the high variability at the Aly8 locus is not merely a consequence of population subdivision. This sequence type is highly variable within all four populations studied (see above).

Recombination and linkage disequilibrium: In an effort to test whether the putative alleles of each of the loci identified from the sequences are truly allelic, we tested for recombination in S-domain loci other than Aly13, where polymorphism levels are high and balancing selection is likely, which violates the assumptions of the analysis. Except for the Aly9 locus, which has little variability, both tests used suggest recombination (or some other form of exchange) in all the putative loci (Table 6). In the Aly8 sequences, many exchange events are detected using Hudson and Kaplan’s (1985) esti-
TABLE 5
Nucleotide diversity estimates and summary statistics for seven S-domain loci sampled from four A. lyrata populations

<table>
<thead>
<tr>
<th>Locus</th>
<th>Diversity measure</th>
<th>Population</th>
<th>North Carolina</th>
<th>Indiana</th>
<th>Scotland</th>
<th>Iceland</th>
<th>Mean (π)</th>
<th>Global within-population (π)</th>
<th>All sites</th>
<th>Silent sites</th>
<th>Tajima’s D</th>
<th>π/πs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aly3</td>
<td>πs</td>
<td>0.0040</td>
<td>0</td>
<td>0.0033</td>
<td>0.0080</td>
<td>0.0053</td>
<td>0.0053</td>
<td>0.60*</td>
<td>0.78*</td>
<td>0.026</td>
<td>0.825</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>21</td>
<td>0.0044</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aly7</td>
<td>πs</td>
<td>0.0052</td>
<td>0</td>
<td>0.0023</td>
<td>0.0022</td>
<td>0.0074</td>
<td>0.0121</td>
<td>0.0146</td>
<td>0.13*</td>
<td>0.13*</td>
<td>−1.52</td>
<td>0.712</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>6</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>20</td>
<td>0.0049</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aly8</td>
<td>πs</td>
<td>0.0182</td>
<td>0</td>
<td>0.0138</td>
<td>0.0171</td>
<td>0.0074</td>
<td>0.0795</td>
<td>0.0750</td>
<td>0.02</td>
<td>0.10*</td>
<td>−0.9</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>6</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>26</td>
<td>0.0172</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aly9</td>
<td>πs</td>
<td>0.0132</td>
<td>0</td>
<td>0.0171</td>
<td>0.0127</td>
<td>0.0074</td>
<td>0.0795</td>
<td>0.0750</td>
<td>0.02</td>
<td>0.10*</td>
<td>−0.9</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>3</td>
<td>8</td>
<td>7</td>
<td>8</td>
<td>26</td>
<td>0.0132</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aly10.1</td>
<td>πs</td>
<td>0.0021</td>
<td>0</td>
<td>0.0025</td>
<td>0</td>
<td>0</td>
<td>0.0054</td>
<td>0.0058</td>
<td>0.13*</td>
<td>0.26</td>
<td>−2.34</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>20</td>
<td>0.0048</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aly10.2</td>
<td>πs</td>
<td>0.0011</td>
<td>0</td>
<td>0.0016</td>
<td>0.0085</td>
<td>0.0074</td>
<td>0.0013</td>
<td>0.0048</td>
<td>0.08</td>
<td>0.02</td>
<td>−1.43</td>
<td>0.159</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>29</td>
<td>0.0045</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aly13</td>
<td>πs</td>
<td>0.0090</td>
<td>0</td>
<td>0.0220</td>
<td>0.0243</td>
<td>0.0127</td>
<td>0.0227</td>
<td>0.0163</td>
<td>0.08*</td>
<td>0.374</td>
<td>1.76</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>19</td>
<td>0.0085</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2915</td>
<td>0.3486</td>
<td>0.4467</td>
<td>0.3144</td>
<td>0.3612</td>
<td>0.3720</td>
<td>0.10</td>
<td>0.554</td>
<td>−0.19</td>
<td>0.554</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1722</td>
<td>0.2005</td>
<td>0.2306</td>
<td>0.1695</td>
<td>0.2000</td>
<td>0.2022</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample sizes are given in the rows labeled n. Note that the Aly7 alleles with the inserted nucleotide are excluded from these estimates.

*Significant population structure P < 0.05.

mator, even though these sequences probably come from at least two loci (see above). This suggests the possibility of gene conversion between the different loci.

Tests for selection within A. lyrata: Tajima’s tests: Tajima’s D statistic (Tajima 1989a) was calculated for each putative locus to check whether the sequences appear to be evolving neutrally. Since the sample sizes are too small to perform the test within populations, we pooled the sequences from the different populations. For most loci, the Tajima’s D values were negative, but did not differ significantly from zero, although a significant negative value (P < 0.01) was found for Aly10.1. Only Aly3 gave a positive Tajima’s D value, but this is nonsignificant and may be attributable to population subdivision, consistent with the high Ks for this locus. There is thus no evidence from this test for balancing selection acting at any of the loci. This includes the Aly8 sequences and the highly polymorphic Aly13 locus, which is likely to be the A. lyrata self-incompatibility locus.

Patterns of evolution in the S-domain sequences: Replacement site polymorphism in the putative A. lyrata self-incompatibility locus, Aly13, is significantly higher in

TABLE 6
Tests for recombination in the S-domains of the Aly loci

<table>
<thead>
<tr>
<th>Gene</th>
<th>No. of sequences</th>
<th>Alignment length</th>
<th>Sites included</th>
<th>r^2</th>
<th>D^*</th>
<th>Sites included</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aly3</td>
<td>21</td>
<td>597</td>
<td>11</td>
<td>−0.42**</td>
<td>−0.60*</td>
<td>13</td>
<td>10.5***</td>
</tr>
<tr>
<td>Aly7</td>
<td>20</td>
<td>768</td>
<td>10</td>
<td>−0.35**</td>
<td>−0.40**</td>
<td>35</td>
<td>27.8***</td>
</tr>
<tr>
<td>Aly8</td>
<td>28</td>
<td>535</td>
<td>25</td>
<td>−0.45***</td>
<td>−0.42***</td>
<td>73</td>
<td>25.8***</td>
</tr>
<tr>
<td>Aly9</td>
<td>20</td>
<td>827</td>
<td>7</td>
<td>0.02</td>
<td>0.18</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>Aly10.1</td>
<td>20</td>
<td>713</td>
<td>13</td>
<td>−0.30**</td>
<td>0.06</td>
<td>23</td>
<td>13.1</td>
</tr>
<tr>
<td>Aly10.2</td>
<td>23</td>
<td>859</td>
<td>11</td>
<td>−0.41**</td>
<td>−0.13</td>
<td>46</td>
<td>5.26***</td>
</tr>
</tbody>
</table>

The estimates are based on pooling sequences from all populations. *P < 0.05; **P < 0.01; ***P < 0.001.
the regions corresponding to the Brassica SRK and SLG hypervariable regions than in the rest of the sequence (Schierup et al. 2001). This is not found for the further loci with S-domain sequences studied here. In these, including the Aly8 sequences, variability is not especially high in these regions.

However, the low polymorphism at most of these loci (see above) makes it difficult to detect differences in diversity among different sequence regions. We therefore also estimated nonsynonymous divergence among the A. lyrata paralogs and among the three loci with putative orthologs in A. thaliana (Aly8, -10.1, and -10.2; see Table 2). Divergence for regions corresponding to the Brassica SLG and SRK and Aly13 HV regions was compared with divergence elsewhere in the S-domain sequences (see MATERIALS AND METHODS for the positions assigned to these regions). Among A. lyrata paralogs, the regions that correspond to non-HV regions of the S-domain accumulate fewer substitutions per nonsynonymous site than do those corresponding to HV regions (the mean for HV is 60% higher than that for non-HV regions); synonymous divergence is saturated and comparisons are not informative for such sites. Of the 15 comparisons, 14 show HV nonsynonymous divergence greater than non-HV divergence (Figure 4, open and shaded bars, respectively; the difference is significant with \( P < 0.0005 \) by a paired sign test, although it must be realized that the tests are not independent). Thus either these regions are under lower selective constraint than the remainder of the sequence or directional selection has caused divergence in these regions, specifically, in the different loci. There is no such clear effect between the Aly loci and their A. thaliana orthologs (Figure 4, solid bars).

McDonald-Kreitman (1991) tests did not detect evidence of directional selection driving divergence between paralogous loci specifically in the HV regions (although this test indicated a significant excess of nonsynonymous polymorphic sites in the non-HV region of Aly3, when compared with divergence from several of the other loci; the reason for this result is unknown, but there is no evidence for balancing selection as this locus does not have high diversity). The relative rate tests also give no indication of any overall deviation from equal rates of evolution of the orthologous pairs of genes since divergence (data not shown). Overall, we conclude that the fairly high \( K_a/K_s \) values in the fore also estimated nonsynonymous divergence among the Aly13 orthologs and among the three loci with putative orthologs in A. thaliana (shown as solid). The total lengths of sequences compared can be seen in Table 2, and the HV regions are as described in MATERIALS AND METHODS and include \( \sim 200 \) bp.

**Figure 4.—Comparisons of divergence in HV vs. other regions of the S-domain sequences. The figure shows nonsynonymous site divergence values between the different paralogous Aly S-domain sequences in A. lyrata, and divergence of these sequences from the putative orthologous loci in A. thaliana (shown as solid). The total lengths of sequences compared can be seen in Table 2, and the HV regions are as described in MATERIALS AND METHODS and include \( \sim 200 \) bp.**

**DISCUSSION**

**The S-domain gene family:** The S-domain loci studied here clearly form part of an ancient gene family with members in all other angiosperms tested, including distantly related species such as maize (Walker 1994; Ansaldii et al. 2000). The loci show different degrees of divergence, but most differ considerably. An exception is the two or more genes apparently required to account for the Aly8 results, which could be a recent duplication, but further work is needed to clarify the situation with respect to these sequences. Five of the eight A. lyrata loci identified have plausible orthologs in A. thaliana, which is highly self-compatible, and not surprisingly, therefore, its Aly13 ortholog is a pseudogene (Kusaba et al. 2001). Several of the genes are expressed both in nonflower tissues and in flowers in A. thaliana (we have no data on the expression of these genes in A. lyrata). Ark1 is expressed in leaves, flower buds, and stigmas (Tobias et al. 1992; Dwyer et al. 1994; Tobias and Nasrallah 1996; Suzuki et al. 1997, 1999); Ark2 is expressed specifically during maturation of cotyledons, leaves, and sepals; and, on the basis of promoter
expression, 

expression, Ark3 is detected specifically in roots and in the root-hypocotyl transition zone (Dwyer et al. 1994). The nonkinase At6S1 is expressed specifically in stigmatic papillary cells (Isogai et al. 1988; Lalonde et al. 1989).

There is no evidence for major birth and death of members of this gene family between A. lyrata and A. thaliana, since most loci can be identified in both species. However, duplication of the pollen-expressed SCR gene was found in one of the two haplotypes studied by Kusaba et al. (2001), and different numbers of copies of SRK are known in Brassica haplotypes (Cabrillac et al. 1999). There may also be further members of this family that we have not studied. PCR amplifications occasionally yielded further S-domain sequences that either were very different from those described here or proved not to be closely linked to the S locus (M. H. Schierup, unpublished results). Furthermore, two of the AHy loci without detectable kinase domains (AHy3 and AHy7) have no evident A. thaliana orthologs, and, as discussed in the next section, two loci, AHy7 and AHy10.1, may be pseudogenes.

**Pseudogenes:** Pseudogene S-domain genes have been found in the Brassica S locus region (Suzuki et al. 1999), and this possibility must be considered for two of our putative loci. The AHy7 sequences with the inserted nucleotide might suggest that these sequences merely represent a pseudogene, but our segregation evidence suggests that they are allelic (presumably null alleles) at the same locus as the AHy7(-) sequences. Moreover, both types of sequence appear to be quite old, since there are numerous fixed differences between them, and both include considerable diversity and are found in most, if not all, A. lyrata populations. This would argue against this locus being a pseudogene; 7(+) appears to have more singletons than 7(-), consistent with its being a derived sequence type. The significantly negative Tajima’s D for the null alleles [7(+) sequences] suggests a recent increase in frequency of this allele and is conservative, given the other evidence for population subdivision. However, there is more diversity among 7(+) alleles than would be expected if this haplotype type rose to high frequency by a recent selective sweep, which would imply low or zero diversity (Hudson et al. 1994).

The AHy10.1 sequences containing deletions may also be a pseudogene. Our diversity analysis included the different types of alleles of this locus, excluding the deletion regions, and nonsynonymous diversity was low, as was the \( \pi_\text{s}/\pi \) ratio (see Table 5), suggesting that loss of function occurred recently. It therefore seems most likely that the B1, B2, and B3 alleles (see Figure 2) are null alleles.

Other examples of polymorphic null alleles are known, sometimes at frequencies as high as those found for the AHy7 and AHy10.1 sequences (e.g., Oxtoby et al. 1991; Gibson et al. 1992; Charmley et al. 1993; Mombaerts 2001). One example is the deletion of a disease resistance gene in A. thaliana, for which there is evidence for balancing selection (Stahl et al. 1999), but there is no evidence for this at the AHy7 or AHy10.1 loci. Replacement polymorphism in the AHy7 sequences without the insertion is only slightly lower than that at silent sites. This high \( \pi_\text{s}/\pi \) ratio might suggest a locus that has lost function or is in an early stage of doing so and is evolving neutrally, but this is not certain, since similar or even higher \( \pi_\text{s}/\pi \) values are found for other AHy loci (see Table 5). Loss of function could also explain the higher diversity in AHy7 than in most of the other loci. We are unable to compare divergence of the nonfunctional and potentially functional AHy7 alleles, since no A. thaliana ortholog can be identified. The absence of an ortholog is, however, consistent with this gene being a nonfunctional duplicate in A. lyrata.

**Levels and patterns of diversity:** The S-domain loci studied here have a range of nucleotide diversity values, including widely differing silent site diversity. Lack of evidence for balancing selection and only moderate diversity levels are also reported for the Brassica SLR1, SLA, and SLB loci, which are not linked to the incompatibility locus (although sample sizes are very small; Hinata et al. 1995; Sakamoto et al. 1998; Watanabe et al. 1998; Luu et al. 2001). For A. lyrata, the diversity estimate for 1.6 kb of the alcohol dehydrogenase gene (Adh) yielded a mean within-population \( \pi \) value for all nucleotide sites of 0.1%; the total diversity, including three different populations, was 0.38% (Savolainen et al. 2000). Higher values are found at other loci, particularly in the non-U.S. subspecies petraea (Wright et al. 2002). All except one of our S-domain loci also have diversity values higher than those of the Adh, but there is no reason to suspect balancing selection. Ratios of nonsynonymous-to-synonymous site diversity within species or divergence between species for plant nuclear genes generally range from 0.1 to 0.2 (Li 1997; Luu 1998; Wright et al. 2002). Published values for Brassica S-domain genes are, however, much higher (Hinata et al. 1995), and the same is true in our data (see Table 5; for divergence between paralogs within A. lyrata, the mean value is 0.55 for the sequence regions corresponding to the SRK HV regions and 0.24 for the rest of the S-domain sequence). This suggests that selective constraints are low in the S-domain, especially in the parts that are hypervariable in SRK, which would accord with the ability of the S-domain to generate highly diverse SRK alleles.

The difference between the reference loci studied here and the otherwise similar S-domain AHy13 sequences (the putative A. lyrata S locus) therefore supports the view that the AHy13 silent and amino acid diversity is unusually high due to the maintenance of the polymorphism of incompatibility alleles. Moreover, variation at the AHy13 locus is similar in all populations, and \( K_\text{s} \) does not differ significantly from zero. This is as expected for loci experiencing balancing selection
(Schierup et al. 2000) and is similar to what has been found in the fungus *Schizophyllum commune*, where no population structure was detectable for mating-type alleles (Raper et al. 1958) in contrast to strong structure for polymorphic allozymes (James et al. 1999). The same is true for *Aly*, although analysis of the synonymous diversity suggests some differentiation between populations. In contrast, several of the other *Aly* loci are significantly differentiated among the populations studied here, indicating some degree of isolation between these geographically distant populations; the same is observed for several other non-S-domain loci (Wright et al. 2002). Since subdivision obscures evidence of selection when samples are pooled (Schierup et al. 2000), the absence of significant Tajima’s test results for the *Aly13* sequences (see above) is not inconsistent with balancing selection. The main evidence for balancing selection on *Aly13* remains its extraordinarily high nonsynonymous diversity.

The lower diversity of the loci studied here, compared with that of *Aly13*, probably cannot be attributed to their location in regions of low recombination, which is a possible reason for low diversity (e.g., Begun and Aquadro 1992; Stephan and Langley 1998). Recombination was detected in all loci, even using sequences pooled for all populations, which must lead to underestimation of its frequency, given the fact that there is evidence for subdivision, so that some sequences have no opportunity to recombine. However, the possibility of gene conversion means that recombination may be overestimated using sequence variants (Langley et al. 2000; Przeworski and Wall 2001). If this were the only form of exchange, hitchhiking processes would still act to reduce diversity, as is observed in other genome regions where reciprocal recombination is rare (e.g., Langley et al. 2000).

In comparisons among paralogs within *A. lyrata*, the regions corresponding to the Brassica *S*-locus hypervariable regions are more diverged than the rest of the S-domain. This supports the suggestions made above that these regions have lower selective constraints than the rest of the S-domain protein. There is no tendency for any of the *Aly* loci studied here to have especially high variability in these regions, in contrast with the results for *Aly13*, in which variability is higher in the sites corresponding to Brassica HV regions than elsewhere in the S-domain sequence. The fact that the putative *A. lyrata SRK* gene, but not the other S-domain genes in the species, has excess diversity in the HV regions, is consistent with suggestions that these regions may be involved in recognition functions. However, it is quite plausible that elevated variability in these regions is due to linkage to sites under balancing selection elsewhere in the *SRK* gene, with low selective constraints in these parts of the sequence allowing higher amino acid diversity than elsewhere in the protein. With current evidence, it is difficult to exclude this possibility, and in genes under balancing selection, high amino acid variability should therefore not be considered sufficient evidence for the sites of the selection. More direct analyses, including transgenic experiments, will probably be necessary to determine the recognition regions within the *SRK* sequence.

**Diversity of the *Aly8* locus, a gene linked to the *S* locus:** The high silent and replacement diversity at the *Aly8* locus can be accounted for in part if these sequences come from two loci, at least one of which we have shown to be linked to the *S* locus. The situation could be similar to that in Brassica, in which some haplotypes have further S-domain loci in the *S* locus region (Yu et al. 1996; Suzuki et al. 1997, 1999; Kai et al. 2001). The existence of two paralogs is not the sole reason for the high *Aly8* diversity, however, because the variants are not found in different haplotypes that could be assigned to different loci, but appear instead to be from a recombining set of sequences. One possibility that we cannot at present exclude is that, in some haplotypes, but not others, the *Aly8* locus recombines with the *S* locus. In the two *A. lyrata* haplotypes for which physical maps exist for the *S* locus region, the ortholog of the Ark3 gene is >5 kb away from the *SRK* gene in one, but at least 30 kb in the other haplotype (Kusaba et al. 2001); these distances, however, would not imply frequent recombination. Moreover, three different *Aly8* sequences have been found in PCR reactions of some individuals, which suggests two separate loci in some plants at least. It is unlikely that variability at the *Aly8* locus is maintained by balancing selection, and it is not concentrated in the regions that show hypervariability in the Brassica and *A. lyrata SRK* loci.

Much of the *Aly8* diversity is probably attributable to linkage to the highly polymorphic *Aly13* locus, which could cause elevated variability at synonymous and even nonsynonymous sites, similarly to the effects discussed above for the different regions of the S-domain. *Aly8* is orthologous to the Ark3 gene, which is closely linked to the *A. thaliana SRK* pseudogene ortholog (Kusaba et al. 2001). In Brassica, low diversity has been suggested for the SLL1 and SLL2 non-S-domain loci that are linked to *SRK*. However, sample sizes were very small, and diversity was scored by a restriction fragment length polymorphism approach that would not detect all sequence variants, so the estimate of variability is not accurate (Yu et al. 1996). It will be interesting to get more data on diversity at loci linked to the *S* loci in both Brassica and *A. lyrata*.

Our data show linkage disequilibrium between *Aly8* sequences and *S* alleles, with independent individuals whose *S* alleles are the same having very similar *Aly8* sequences. This would be expected if the high *Aly8* diversity were due to close linkage to a gene under balancing selection. There is also evidence for some exchange (either recombination or gene conversion)
between this locus and other \(A lyrata\) sequences, making it impossible to distinguish the locus of origin of our sequences. Given that \(A lyrata\) has a kinase domain, it may be helpful to study diversity in this region of the gene to search for sequences that distinguish the two or more different loci. This might allow tests for selection vs. neutrality of the \(A lyrata\) variants and should help to disclose the extent of the region of linkage disequilibrium in the \(S\) locus region. Further work is clearly needed to elucidate the number of loci contributing to the diversity of sequences that we classify as \(A lyrata\) and to study their diversity individually, as well as that at other loci linked to the \(S\) locus. If extensive linkage disequilibrium is confirmed, it would support the view that recombination does not occur in this region of the genome and suggest that the exchanges that can be detected (Awadalla and Charlesworth 1999; Schierup et al. 2001) may be due to gene conversion. If this occurs, it may have important implications for the maintenance of diversity among \(S\) alleles, since sequences from other \(S\) domain genes may sometimes be introduced into \(S\) alleles at the \(S\) locus.

Conclusions: The chief conclusion from these studies is that high silent and replacement diversity in several sequences and regions may be attributable to linkage to a locus under balancing selection. This locus is presumably \(A lyrata\), the \(A lyrata\) putative incompatibility locus. We have found high variability at the \(A lyrata\) locus, which cannot entirely be accounted for in terms of paralogous loci and which displays linkage disequilibrium with \(A lyrata\) alleles, just as predicted for regions close to a locus with a balanced polymorphism (see above). The \(A lyrata\) sequences have elevated nonsynonymous diversity in the sites corresponding to the hypervariable regions of Brassica \(SRK\) \(S\) domains, but our evidence that these regions diverge more than the rest of the \(S\) domain suggests that this is largely because of low selective constraints of these portions of the proteins. The hypervariability cannot therefore be taken as evidence that sites within these regions themselves experience balancing selection. Among different regions linked to sites subject to balancing selection, those regions that are not important for the \(S\) domain’s signal transmission function are expected to have the greatest nonsynonymous divergence and polymorphism. They may thus be the regions most likely to evolve sequence differences that can be used in recognition of different specificities. This may explain the observation of hypervariability in similar regions in Brassica and \(A lyrata\) \(SRK\) and in Brassica \(SLG\). An implication is that it may thus be difficult to determine which amino acid sites determine the functional incompatibility types (and are therefore under balancing selection). One possible approach is to do experimental functional tests involving altering individual amino acid residues in \(S\) allele sequences. Studies of \(S\) allele sequence variability may also help to indicate which amino acids can differ between alleles with the same incompatibility type.

We thank the staff at the University of Edinburgh for growing the plants and T. E. Thorhallsdottir, C. H. Langley, and R. Mauricio for the seeds used in this work. This work was supported by the Biotechnology and Biological Sciences Research Council of the UK. B. K. Mable was also supported by the Natural Sciences and Engineering Council of Canada, D. Charlesworth by the Natural Environment Research Council of Great Britain and Edinburgh University, M. H. Schierup by the Danish Natural Sciences Research Council (grant nos. 9701412 and 1282), and P. Awadalla by an Edinburgh University Faculty of Science and Engineering Scholarship.

**LITERATURE CITED**

Ansalardi, R., A. Charoub and C. Dumas, 2000 Multiple \(I\) gene family members including natural antisense transcripts are differentially expressed during development of maize flowers. J. Biol. Chem. **275:** 24146–24155.

Awadalla, P., and D. Charlesworth, 1999 Recombination and selection at \(SRK\) self-incompatibility loci. Genetics **152:** 413–425.


Charmley, P., K. Wang, I. Hood and D. A. Nickerson, 1993 Identification and physical mapping of a polymorphic human T-cell receptor \(V\)-beta gene with a frequent null allele. J. Exp. Med. **177:** 135–143.


Gibson, J. B., A. V. Wilks and A. Agrotis, 1992 Molecular relationships between alcohol-dehydrogenase null-activity alleles from...


Kumar, S., K. Tamura, I. Jacobsen and M. Nei, 2000 MEGA2: Mole- cular Evolutionary Genetics Analysis, Version 2.0. Penn State University, Olds, PA.


Savolainen, O., C. H. Langley, B. P. Lazzaro and H. Freville,


