The Molecular Evolution of Self-Incompatibility Loci in the Brassicaceae Family

Philip Awadalla

Thesis presented for the degree of
Doctor of Philosophy
at the University of Edinburgh

2000

Abstract

Self-incompatibility (SI), the process by which selfed pollen is recognized and rejected, is a classic study system for population geneticists and cell and molecular biologists interested in it's spectacular polymorphism and evolution. This plant mating system is one of the few systems where the selection forces maintaining the extreme level of polymorphism is well-defined, yet how SI evolves remains poorly understood. In *Brassica,* SI is controlled maternally by haplotypes involving perhaps two related genes; the S-related kinase (*SRK*) and S-glycoprotein (*SLG*), and paternally by the recently discovered cystein rich pollen gene (SCR). Consistent with frequency-dependent selection, these loci exhibit exceptional levels of high amino-acid variability with some regions being 'hypervariable' (HV); a pattern observed at similar 'recognition' loci such as the MHC locus. A major question is whether HV regions are targetted by balancing selection, or merely regions of relaxed selective constraint. A second question relates to the role that recombination may play, if any, in the evolution of SI. The genes controlling SI are thought to be tightly physically linked so that maternal and paternal alleles maintain the haplotype configurations necessary for correct recognition. Furthermore, the maternal genes are part of a large gene family (S-domain family). In *Brassica*, I observed that linkage disequilibrium and nucleotide diversity patterns indicate that recombination of some form, perhaps gene conversion, plays an evolutionary role at the *SLG* gene. To assess the true level of nucleotide diversity and recombination in natural populations, and to address whether HV regions are neutral versus under selection, I identified and characterized seven loci in *Arabidopsis lyrata* that are homologous to the SI-genes in *Brassica*. Segregation analyses shows that all but one set of sequences is unlinked to self-incompatibility in *A. lyrata* and exhibit relatively low levels of nucleotide diversity. *SLG*- and *SRK*-like variants have been identified that segregate with incompatibility groups in three independent families, and exhibit extremely high levels of polymorphisms, again similar to MHC and *Brassica* SI genes, suggesting that these loci in *A. lyrata* are targetted by balancing selection as well. Furthermore, hypervariable regions are consistent with that observed in *Brassica* species although it remains unclear whether these are regions targetted by balancing selection or merely regions of relaxed selective constraint. The homologous members of the gene family exhibit levels of neutral diversity consistent with other published reports for this outcrossing species. However, divergence patterns for these same homologues, show heterogeneity in selective constraint across the sequences, and that variable mutation rates have contributed to gene family evolution. The corresponding HV regions in the members of the gene family exhibit strikingly high divergence compared to the rest of the aligned gene, suggesting that these HV regions are merely under relaxed selective constraint.

Declaration:

I declare that this thesis has been composed by myself and that this work is my own. Contributions by collaborators are acknowledged where appropriate.

[redacted]

Acknowledgements

I would like to thank my supervisor Deborah Charlesworth for her guidance and encouragement throughout my Phd, for her thorough comments on most of this thesis, as well as for everything I have learnt during the course of my degree.

All my collaborators have made the research performed here and elsewhere possible. Mikkel Schierup was instrumental in helping me get this project off the ground and Barbara Mable played a significant and helpful role throughout this project. I should thank Adam Eyre-Walker and John Maynard Smith for allowing me to take part in a very exciting study and I am grateful for the inspiring interactions that were a result of this collaboration.

I have learned a tremendous amount from the people that I have seen and interacted with everyday, specifically, N. Barton, B. Charlesworth, Toby Johnson, P. Keightley, B. McAllister, Gil McVean, and C. Vieira. As well, the countless people who have come through ICAPB for short or long term visits have played major roles in teaching me stuff. I should also acknowledge many individuals throughout ICAPB who have allowed me to interact with them on various projects and have made my Phd enjoyable. Sporting ICAPB deserves note for always losing but allowing me to play and foul the opposition mercilessly, regardless.

Particular thanks goes to Tracy Solomon whose enduring patience, support and giving nature has made all of this possible.

Table of Contents

Chapter 1

Self-Incompatibility in Homostylous Plants:

A Review of Molecular and Evolutionary Genetics of SI in Brassicaceae

Chapter 2

Linkage Disequilibrium and Recombination at the S-loci in *Brassica*

Chapter 3

Chapter 4

Chapter 5

Molecular Evolution of Loci Linked to Self-Incompatibility in *Arabidopsis lyrata*

Chapter 6

Chapter 1 - Introduction

Self-Incompatibility in Homostylous Plants: A Review of Molecular and Evolutionary Genetics of SI in Brassicaceae

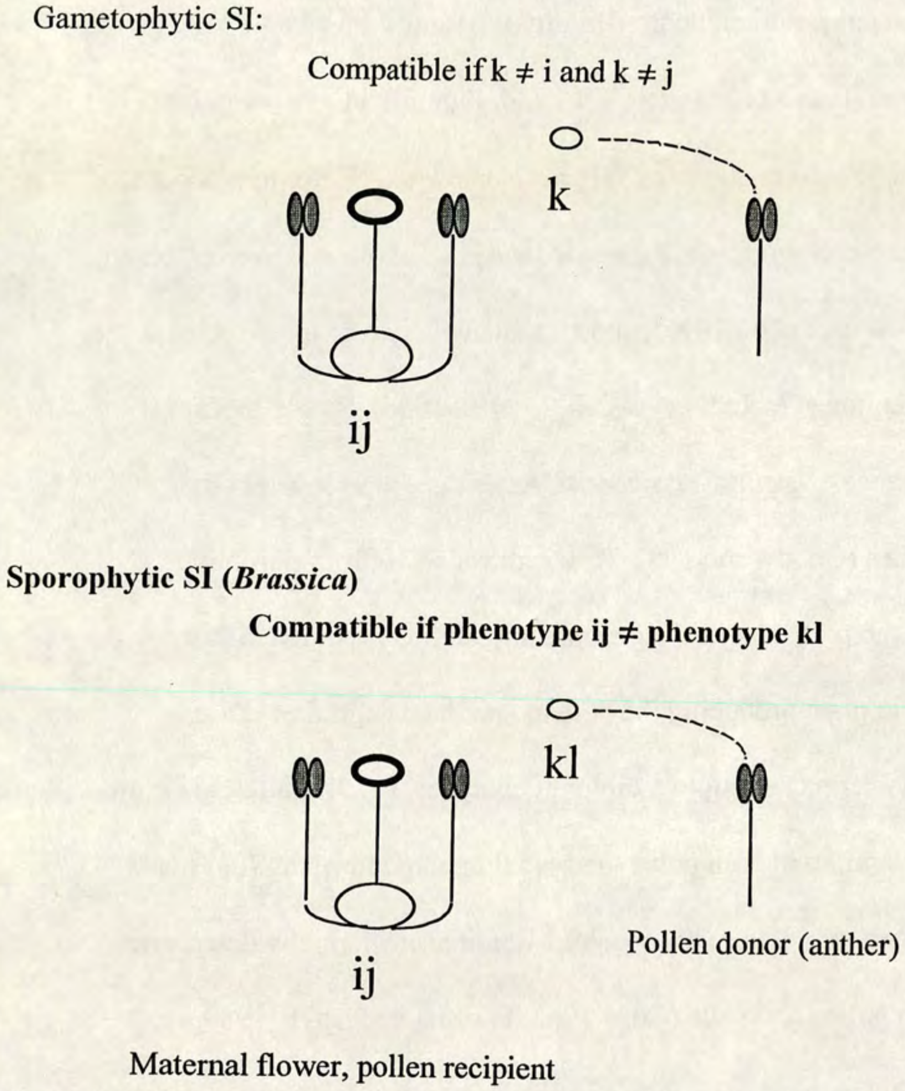
1.1 Introduction

The self-incompatibility (SI) recognition systems are classic study systems of population genetics and plant mating system evolution. Self pollen, and pollen carrying "*S*" alleles in common with a potential recipient plant, is recognised and prevented from fertilising ovules. Polymorphism at SI loci is famously high, even higher than that observed in the extremely polymorphic mammalian histocompatibility complex (MHC) loci. Although the study of SI has attracted the attention of many geneticists and molecular and evolutionary biologists over the last century, research in this area is at its most exciting stage, as molecular biologists are actively identifying the genes and mechanisms involved in pollen recognition. As a result, population geneticists are better equipped to tackle questions and test models related to the impressive *S*-allele polymorphism. In this chapter, I describe the genetics and potential models of SI function and evolution, and recent results that are helping to further our understanding of the evolution of self-incompatibility, focusing particularly on sequence data from self-incompatibility loci of *Brassica* species. This review provides a background that motivates the analytical and empirical work outlined in this thesis. The final section of this chapter is a preface to the content of the subsequent chapters, which describe molecular evolution research in the sporophytic SI gene family among natural populations of *Arabidopsis lyrata* L. (Brassicaceae).

Some of the analyses shown in this chapter were published in an invited review by Charlesworth and Awadalla (1998). Recent breakthroughs in the molecular biology of the *Brassica* SI system are included and briefly outlined here. A review of SI must discuss the polymorphism observed at these loci in some detail, for it is the interpretation of the high polymorphism patterns that has attracted population geneticists. Therefore, I analysed and included molecular evolutionary estimates of diversity in the *Brassica* SI in this initial chapter. Description and alignments of the sequence data and programs I used in the estimations are described in the Materials and Methods section of Chapter 2, where recombination analyses are described. I thank David Guttman for providing the *HLA* data that was used in the article by Charlesworth and Awadalla (1998).

1.2 Gametophytic vs. Sporophytic Recognition of Selfed Pollen

Many plant families exhibit a genetic recognition system where pollen express alleles that are recognized by the recipient female organs of an individual flower. In "gametophytic" SI systems, pollen incompatibility types are controlled by the grains' own haploid genotypes (East and Mangelsdorf 1926, Figure 1.1). Gametophytic inheritance with a single incompatibility (*S*) locus is known in *SCR*ophulariaceae, Onagracea, Papaveraceae, Solanaceae, Rosaceae, and several other flowering plant families (reviewed in Weller et al. 1995), and systems with two or more loci are known from several families, including grasses (Li et al. 1995). Single-locus "sporophytic" inheritance exists in Brassicaceae, Asteraceae, and several other flowering plant families (Goodwillie 1997, Kowyama et al. 1980). Sporophytic inheritance relates to the pollen expressing an incompatibility type determined by the genotype of the diploid plant producing the pollen, and the recognition process therefore involves interaction among two diploid genotypes, i.e. four alleles (Figure 1.1). Dominance is common in both pollen and pistil in homomorphic SI systems (Sampson 1967). For example, in many species a dominance hierarchy determines the phenotype of both pollen and style (Cope 1962, Samaha and Boyle 1989, Kowyama et al. 1994). In the Brassicacea, dominance of alleles expressed in pollen and codominance among alleles in the stigma are common (Ockendon 1974, Stevens and Kay 1989). *S*-loci exhibit extremely high diversity, with many *S*-alleles in populations (e.g. Wright 1939, Sampson 1967, reviewed by Lawrence 2000). In terms of allele numbers, the *S*-loci are among the most highly polymorphic loci known, similar to mammalian MHC loci (Hughes et al. 1990), or fungal incompatibility loci (May and Matzke, 1995, May et al. 1999). Over forty S-alleles are thought to exist in *Brassica campestris* globally (Nou et al. 1993).


Gametophytic SI:

Compatible if k ≠ i and k ≠ j

Sporophytic SI (*Brassica*)

Compatible if phenotype ij ≠ phenotype kl

Maternal flower, pollen recipient

Figure 1.1: The two major SI systems among homomorphic flowering plants; gametophytic and sporophytic SI. In gametophytic SI, the pollen expresses only one of the alleles of the donor genotype. For a compatible mating, it must be different from either of the alleles expressed in the style of the recipient plant. In sporophytic SI, the pollen expresses the genotype (2 alleles) of the donor plant which must differ from the expressed genotype of the recipient plant for a compatible mating. In sporophytic SI, if all alleles are codominant, than the pollen cannot share either allele with the maternal plant to be compatible. With dominance, a dominant allele can mask the expression of an allele such that it is possible to get compatible matings and also homozygote seeds.

1.3 Evolution and maintenance of self-incompatibility

The maintenance of variability at the self-incompatibility loci is well understood (Wright 1939) and is one of the few genetic systems where the mode of balancing selection is unequivocal. Rare alleles have a fertility advantage because pollen carrying such alleles will most likely come in contact with stigmas expressing more common alleles, and therefore, the pollen will not be rejected by incompatibility reactions of recipient plants. This frequency-dependent selection favours new incompatibility-type alleles that arise in the population, either through mutation or migration, until an equilibrium is reached with equal allele frequencies. The equilibrium number of alleles may be high, depending on the effective population size of the species in question (since alleles will be lost by chance in finite populations) and mutation rates to new S-alleles (Wright 1939). The fertility advantage to low frequency alleles means that losses by genetic drift, or other demographic changes, are soon restored if there is gene flow from other populations. Alleles should thus be maintained in species for long evolutionary times (Takahata 1990, Vekemans and Slatkin 1994). In sporophytic systems, different dominance classes have different expected maintenance times (Schierup 1998) and therefore, the maintenance times among alleles in the Brassicacea will be different from systems with alleles expressed in the haploid stage, such as in gametophytic SI (e.g. Solanaceae). The reason is that recessive alleles reach higher frequencies than dominant alleles, the so-called "recessive effect", because they display their genotype less often in the pollen, and therefore selection against any allele becomes balanced when recessive alleles have higher frequencies (Bateman 1952, Charlesworth 1988).

In several angiosperm families with self-incompatibility, alleles have now been cloned that segregate with the incompatibility types of plants, which encode sequences of co-segregating pistil proteins (Nasrallah et al. 1987, Anderson et al. 1989, Li et al. 1995, Walker et al. 1996). The gene families to which the maternally expressed *S*-loci belong in different plant families are very different. Therefore, an important result from molecular studies is that self-incompatibility loci have evolved from independent origins several times among families of flowering plants (Weller et al. 1995). The loci involved in gametophytic SI systems belong to at least two different gene families. In Solanaceae, Rosaceae and *SCR*ophulariaceae, *S*-proteins are related to RNases, though the similarity may be due either to independent evolution, or common origin (Sassa et al. 1996, Richman and Kohn 1997). Very different gene products are involved in *Papaver rhoeas* (Rudd et al. 1996). The genes in *Papaver rhoeas* are homologous to a gene family with unknown function in the highly selfing *A. thaliana* plant species.

1.4 The *Brassica* S-locus system

The system in the Brassicaceae is quite differerent from those described above. The genes controlling this sporophytic system are described in greater detail, as the following analyses pertain to the specific properties of these loci. Until recently, two linked loci, *SLG* and *SRK* (encoding, respectively, a stigmatically expressed *S*-locus glycoprotein and a receptor kinase), have been thought to play essential roles in incompatibility (Goring and Rothstein 1996). These two genes are physically linked, the region between them spanning a length of between a few to possibly as much as

100–200 kb (Boyes and Nasrallah 1993, Goring and Rothstein 1996, Yu et al. 1996, Conner et al. 1998), and variants of each locus co-segregate with incompatibility types as S-allele haplotypes (Boyes and Nasrallah 1993; Boyes et al. 1997). Both *SLG* and *SRK* are expressed in the epidermal cells of the stigma during self-incompatibility (Nasrallah et al.1987; Stein et al. 1991). The *SRK* locus contains an S-domain homologous in sequence to the *SLG* locus, an apparent transmembrane domain, and a 'kinase' domain with homology to members of a serine/threonine kinase gene family (Figure 1.2, Stein et al., 1991). Orthologues are present in other plant species, including the self-compatible *A. thaliana* (see Chapter 3; Dwyer et al. 1994). The *SRK* is therefore surmised to be an extracellular receptor (Stein et al. 1991). In *Brassica,* two dominance classes of alleles have been discovered (Type I and Type II) and alleles for both loci from dominant and recessive S-haplotypes have been sequenced. High amino-acid divergence between the two groups has been observed. Potential glycosylation residue positions appear to be conserved across S-domain S-alleles in species of the Brassicaceae (Kusaba et al. 1997, Sakamoto et al. 1998). The protein products of the *SRK* autophosphorylate on serine and threonine residues - a gene called ARC1 in *Brassica* is the only gene known to be autophosphorylated by the *SRK* (Stone et al. 1999). *SRK* and *SLG* loci also have regions that resemble an immunoglobulin-like repeat sequence (Glavin et al. 1994), and may be related to plant proteins involved in defence against pathogens, another plant recognition system (Pastuglia et al. 1997a). *Papaver* S-alleles share very limited homology with the *SLG* and *SRK* genes from *Brassica.*

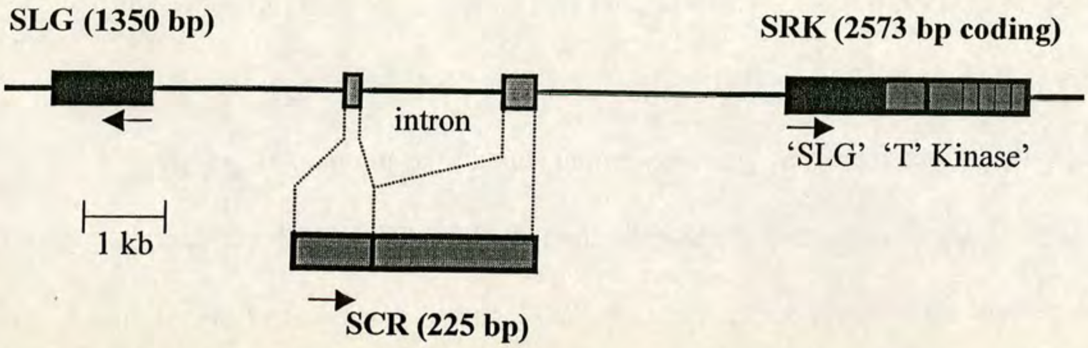


Figure 1.2 Map of the *SLG$_8$-SRK$_8$* genomic S-locus region in *B. rapa* (Schopfer et al. 1999). The pollen-expressed *SCR* gene (see Paternal Component of Recognition in *Brassica*) resides between the *SLG* and the *SRK*. Note the opposite orientation of the reading frame (arrows) of the *SLG* and *SRK* loci. The reading frame orientation differs among haplotypes. The *SRK* locus has an '*SLG*' or S-domain homologous to the *SLG* locus (black region). The *SLG* locus here does not have an intron, wheras the *SRK* has an intron separating the *SLG*-domain from the transmembrane domain ('T'), and another separates the transmembrane domain from the kinase domain ('kinase'). The kinase domain has another 4 introns separating 5 exons within the kinase domain. Intron lengths appear to be variable among haplotypes and members of the gene family.

Recent evidence casts doubts on the role of *SLG* in *Brassica* SI. Originally, it was thought that the *Brassica SLG* locus encodes an extracellular protein involved in recognition (Nasrallah et al. 1987). The discovery of the *SRK* locus led researchers to suggest a molecular model that involved the *SLG* acting as an intermediate in helping to bind a pollen-expressed protein which comes into contact with the stigma cells, to a membrane-bound *SRK* protein. As well, in most cases the *SLG* of a given S haplotype has the greatest sequence similarity to the extracellular S-domain of the *SRK* of the same S haplotype rather than to the *SRK* of another haplotype (see Figure 1.10 – Recombination in the *SLG-SRK* region). Sense and anti-sense suppression of the *SRK* (Conner et al. 1997) has led to a loss of the SI response, suggesting that the

SRK is important for the expression of SI (Figure 1.3). Recent transgenic work has also shown that the *SLG* and the *SRK* genes alone are sufficient for operation of the SI response in the pistil (Cui et al. 2000). However, it has recently been found that not all *Brassica* plants have an *SLG* gene (Nishio and Kusaba 2000, Cabrillac et al. 1999) and in cases where the *SLG* is present, some alleles have stop codons or frame-shift mutations (Nishio and Kusaba 2000). Kusaba and Nishio (1999) have also observed that the most similar pair of *SLG* alleles from cultivars with different S-phenotypes was extremely high (99.5% amino acid similarity). In constrast, among 15 *SRK* alleles thus far sequenced in *Brassica* (Chapter 5), the highest similarity was found to be 93.5% (*SRK*23-*SRK*29 in *B. oleracea*). Sequencing alleles from different plants but of the same functional incompatibility type, Kusaba et al. (2000) noted that the *SLG* amino acid diversity was as high as 12% but the *SRK* differences involved only a single amino acid change as might be expected if the *SRK* for these plants encodes the same incompatibility alleles. Finally, Takasaki et al. (2000) showed that, by transforming self-incompatible plants of *Brassica campestris* (*rapa*) with either an *SRK* or *SLG* transgene from a different self-incompatibility phenotype separately, expression of the transfected *SRK* alone, but not the transfected *SLG* alone, caused a change in the S-phenotype of the transformed plant. This suggests that *SLG* may not be vital, at least in some S-haplotypes. If it is not involved in SI, it is unclear why a functional *SLG* appears to be present and expressed in so many S-haplotypes. Furthermore, why is the diversity at this gene so high (see *Section 1.5*)? Takasaki et al. (2000) have also shown that the ability to reject pollen was 'enhanced' by the presence of the 'correct' *SLG* allele present. When both the *SLG* and the *SRK* from one incompatibility type were used to transfect a plant of an alternative incompatibility type, the strength of SI was enhanced in the sense that fewer ‘selfed’ seeds were obtained.

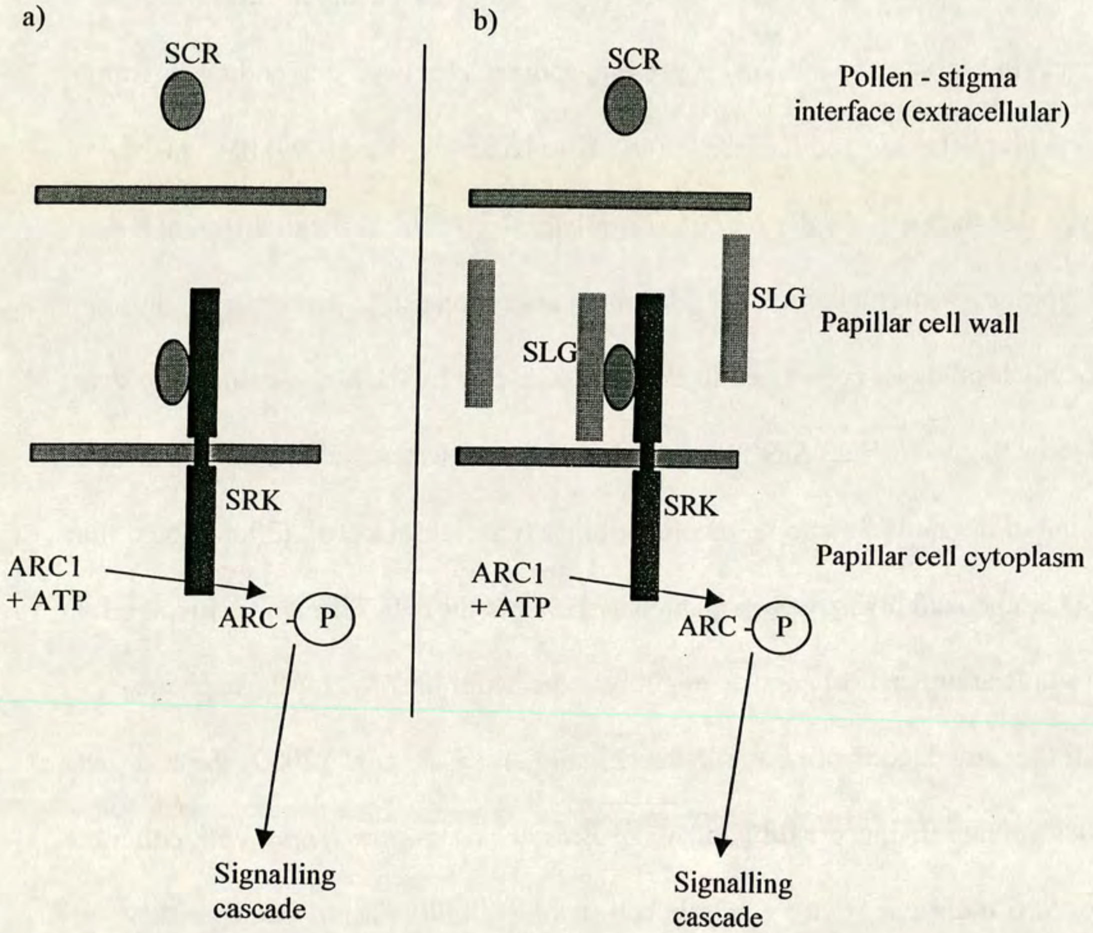


Figure 1.3: A model for the action of the *SRK* and the *SLG* loci in confering incompatibility. In a) the SLG protein is not involved in the recognition of the pollen product, *SCR*, which acts as a ligand binding to the SRK protein anchored in the cell membrane of the papillary cell of the stigma. In b), the SLG acts as an intermediate facilitating the binding of the *SCR* protein to the SRK. Binding initiates phosphorylation of the ARC protein which initiates a signal cascade resulting in an inhibitory signal preventing germination of the pollen.

1.5 The *SLG* and *SRK* genes are members of a multi-gene family

Both the *SLG* and *SRK* genes are members of a large gene family, referred to here as 'S-domain' loci (Isogai et al. 1988, Lalonde et al. 1989, Boyes et al. 1993, Suzuki et al. 1997). The members of this gene family are either '*SLG*-like' in that they are homologous to the *SLG* gene and the S-domain of the *SRK*, but do not have a kinase domain (Lalonde et al. 1989, Boyes et al. 1991, Isogai et al. 1991), or they are '*SRK*-like' in that both the S-domain and the kinase domain are homologous to the *SRK* gene (Boyes et al. 1991, Suzuki et al. 1997). Members of both groups of loci have been identified in a number of *Brassica* species, including *A. thaliana* (see Chapter 3; Dwyer et al. 1994), and are found on a number of different chromosomes throughout both genomes. Therefore, it is likely that this gene family diversified, genes have found new functional or housekeeping roles, and they have been maintained for much of the history of the Brassicaceae. Many members of the gene family have acquired specific functions not related to SI but some of these genes (ie. *SLR*1) are expressed in the stigma and are required for correct pollen stigma adhesion (Boyes et al. 1991, Dwyer et al. 1994).

1.6 The paternal component of recognition in *Brassica*

Until recently, a major question concerning the recognition function of self-incompatibility systems has been whether there are separate pistil and pollen loci, as is often assumed (e.g. Stephenson et al. 1997). This would require strong linkage disequilibrium, so that each pistil type can reject only its own pollen, and each allele must encode a protein that causes rejection of just one allele product of the other locus. Mutation at either locus will, furthermore, not create new incompatibility types, but merely produce self-compatibility, making it hard to see how new SI types could arise.

A pollen protein that co-segregates with SI types has only recently been identified in the the Brassicaceae (Schopfer et al. 1999). A small, single-copy (ca. 80 amino-acids), S locus-encoded gene is expressed solely in anthers (*SCR*, S-locus cysteine-rich protein) of *B. rapa* and *B. oleracea* plants (Schopfer et al., 1999, Takayama et al., 2000). In one particular incompatibility type (S_8), the *SCR* gene is physically situated between the *SRK* and *SLG* genes, with approximatley 4 kilobases (kb) separating the 5' end of the *SCR* gene from the *SLG* gene and approximately 3 kb separating the 3' end of the *SCR* gene from the *SRK* locus (Figure 1.2; Schopfer et al. 1999). Transformation and crossing experiments have shown that variants at this locus are allele-specific, in that each functional incompatibility allele expresses a different *SCR* variant, these variants segregate tightly with the SI-phenotype, and that this gene is necessary for recognition. This gene exhibits some similarity to other pollen coat proteins (*PCP*s) in that they exhibit cysteine residues in similar positions to the *SCR* locus. Nucleotide diversity among the five alleles sequenced in *B. rapa* and *oleracea* is extreme (Figure 1.4). Alignments are problematic, yet there are regions that are highly conserved at the 5' end of the locus and the cysteine residues can anchor the alignments. One intron site is conserved among all the alleles identified thus far, but variation in intron length appears high even among the few alleles studied. The *SCR* gene is thought to trigger the *SRK* locus in a series of biochemical reactions of which the phosphorylation of ARC1 (Stone et al. 1999) is part of a process where the end result is the inhibition of self-pollen (Figure 1.3).

```
                                     ↓
                   1                                       42
B. rapaS8          MKSAVYALLCFIFIVSGHIQELEANLMKRCTRGFRKLGKCTT
B. oleraceaS13     MKSAVYALLCFIFIVSGHIQEVEANLMMPC--GSFMFGNCRN
B. oleraceaS6      MKSAIYALLCFIFLVSSHGQEVEANLKKNCVGKTRLPGPCGD
B. rapaSP11-52     MKSVLYALLCFIFIVSSHAQDVEANLMNRCTRELPFPGKCGS
B. rapaSP11-14     MKSAIYALLCFIFIVSSHVQEVEANLRKTCVHRLNSGGSCGK
B. rapaSP11-12     MKSAIYALLCFIFIILSRSQELTEVGADKQQCKKNFPGHCET
                                                          *
                   43                                      84
B. rapaS8          LEEEKCKTLYPRGQ------CTCSDSKMNTHSCDCKSC----
B. oleraceaS13     IGARECEKLNSPGKR-KPSHCKCTDTQMGTYSCDCKLC----
B. oleraceaS6      SGASSCRDLYNQTEKTMPVSCRCVPTGR----CFCSLCK---
B. rapaSP11-52     SEDGGCIKLYSSEKKLHPSRCECEPRYKAR-FCRCKIC----
B. rapaSp11-14     SGQHDCEAFYTNKTNQKAFYCNCTSPFRTR-YCDCAIKCKVR
B. rapaSP11-12     S--ERCENTYK-RLNKKVFDCHCQPFGRRR-LCTCK-C----
                        *              * *         * *
```

Figure 1.4: Alignment of *SCR* alleles from 5 different S-phenotypes. Sequences were taken from Genbank (Schopfer et al. 1999) and I performed the alignment using Clustal v. X. An intron splice site is conserved after amino acid position 19 and conserved cysteine residues are indicated with stars. A conserved 5' region of a putative signal peptide is in bold.

To date, no pollen-expressed candidate gene has been identified that segregates with gametophytic SI. In gametophytic SI, if such a locus (loci) exists, they must be tightly linked to the pistil locus as it is in *Brassica* species, otherwise recombinants should exist with the pollen of one type and pistil reaction of another; these have not been found even in mutation tests that readily generate self-compatibility in species with homomorphic SI (Lewis 1963), though they are known in distylous species (Ernst 1936). Transgenic approaches may also fail due to co-suppression (e.g. Jorgensen 1990), as expression of additional *S*-locus copies in pollen (e.g. in the diploid pollen of tetraploids) often causes self-compatibility (Lewis 1963). It is hoped that assaying and identifying many, if not all, of the genes in the S-locus regions of gametophytic Solanaceous plants will help to identify proteins expressed in pollen that have S-specific variants (e.g. Dowd et al. 2000, Robbins et al. 2000). Such an approach is highly labour intensive and may be unsuccessful, as the S-RNAses involved in S-recognition in gametophytic SI reside close to centromeric regions which exhibit reduced levels of recombination. This will hinder mapping approaches and segregation analyses as the region of suppressed recombination extends over a large section of the chromosome (Gebhardt et al. 1995).

1.7 Nucleotide diversity at self-incompatibility loci

The first sequence studies of small numbers of alleles revealed exceptionally high levels of divergence between alleles for different incompatibility types, in both Solanaceae (Anderson et al. 1989, Ioerger et al. 1991) and Brassicaceae (Nasrallah et al. 1987). Both silent and amino acid differences were found to be high among many pairs of *S* alleles. This high diversity is found throughout the *S*-locus sequences of both gametophytic and sporophytic systems, but particularly in certain ("hypervariable" or HV) regions (Ioerger et al. 1991, Dwyer et al. 1994).

In order to subject the S-allele polymorphisms to molecular evolutionary analyses, ideally the incompatibility types of the alleles sequenced should be known, and the alleles should be randomly picked. No study has yet achieved this ideal. The first study of alleles from a natural population was done in a gametophytic system, in two species of Solanaceae, though the incompatibility types of the alleles sequenced are not known (Richman et al. 1996). In *Brassica*, sequence differences have been measured only between *S*-alleles for different incompatibility types from cultivated strains. Many sequences of the *SLG* and *SRK* loci from different *Brassica* S-strains have been collected (e.g. Kusaba et al. 1997). These will on average differ more than randomly picked alleles from a population sample. As a result, diversity and summary statistics will deviate to some extent from the true values for a natural population sampled randomly (Chapter 5).

Measuring polymorphism by the diversity per nucleotide site, differences between *Brassica SLG* and *SRK* alleles are lower than in the natural populations of Solanaceae surveyed (Figure 1.5; Charlesworth and Awadalla 1998), and similar to differences at MHC (Hughes and Nei 1989). Synonymous (*Ks*) and nonsynonymous (*Ka*) pairwise differences per synonymous and nonsynonymous site, respectively, occur throughout the *S*-locus, and are spectacularly high in the HV regions (Table 1.1).

1.8 The S-alleles are ancient

Alleles at the *S*-loci must be very old because long periods of time are required for this level of silent substitution to accumulate between alleles within each species (Hudson 1990). Also, *S*-allele sequences are often more similar to alleles from related species than to others from their own species: *Solanum* alleles mingle in the allele tree with those from species in the genera *Petunia*, and *Lycopersicon*, and *Nicotiana* (Ioerger et al. 1991, Richman et al. 1996, Richman and Kohn 1999), and the same is seen in *Brassica* species (Figure 1.6). This "trans-specific clustering" of

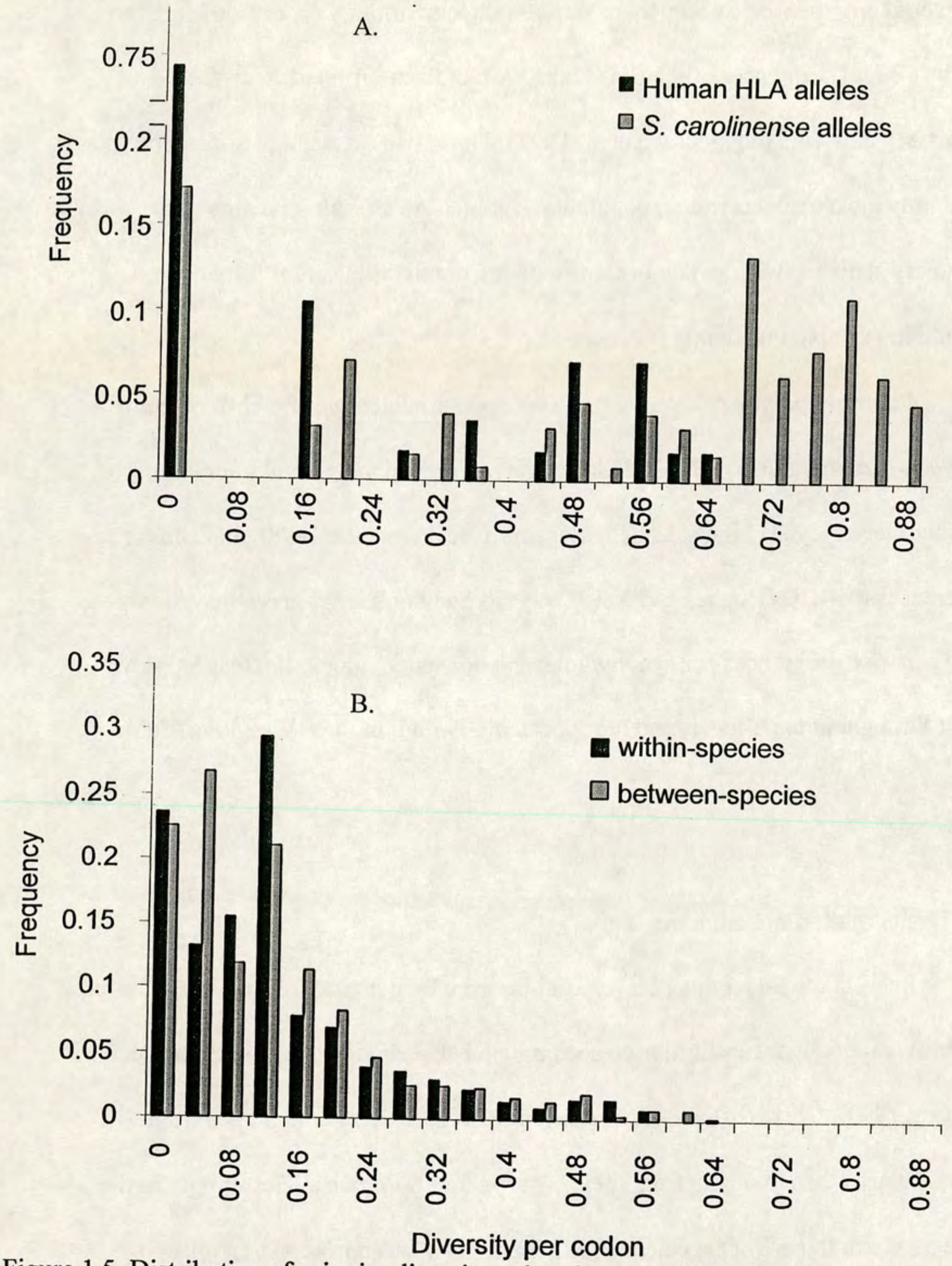


Figure 1.5. Distribution of pairwise diversity values between alleles within species, in S- and MHC-loci. Panel A shows diversity for alleles in a population with gametophytic SI, together with human DHQB1 alleles. The data for sporophytic SI (panel B) are based on 39 *SLG* alleles of separate incompatibility types in *Brassica* and 6 *SRK* alleles for *B. oleracea* and *B. campestris* (Awadalla and Charlesworth 1999). Within and between-species comparisons were analysed separately.

	SLG		*SRK*		
	"conserved"	"HV"	*S*-domain "conserved"	*S*-domain "HV"	kinase domain
Nucleotides	1107	235	1107	235	1365
NON-SYNONYMOUS AND SYNONYMOUS DIFFERENCES					
Ka					
Within species					
B. campestris	0.081	0.226	0.092	0.262	0.069
B. oleracea	0.076	0.245	0.067	0.226	0.053
K_S					
Within species					
B. campestris	0.174	0.253	0.148	0.249	0.170
B. oleracea	0.182	0.246	0.173	0.201	0.149
Between species	0.177	0.253	0.177	0.235	0.185
RADICAL AND CONSERVATIVE AMINO ACID DIFFERENCES					
$P_{radical}/P_{conservative}$ (charge)					
B. campestris	1.12	0.92 - 0.98	1.04	0.253[†]- 1.33	0.505
B. oleracea	1.07	0.51[†]- 0.98	0.596	0[†]- 1.50	0
$P_{radical}/P_{conservative}$ (polarity)					
B. campestris	0.57	0.21[†]- 1.04	0.640	0.732 - 0.970	0.882
B. oleracea	0.64	0.22[†]- 0.88	0.591	0.629 - 1.16	0.447

[†] C-domain (total of 30 nucleotides)

Table 1.1 Distribution of diversity values expressed as mean pairwise differences per site (*K*) in different parts of the *SLG* and *SRK* loci of *Brassica oleracea* (21 *SLG* and 3 *SRK* alleles) and *campestris* (18 *SLG* and 3 *SRK* alleles – see Chapter 2 for accession numbers). Proportions of uncorrected pairwise synonymous and non-synonymous differences per site (*Ks* and *Ka*, respectively) were estimated using MEGA (Kumar et al. 1994), and conservative and non-conservative amino acid substitutions (see text) were estimated using a program provided by Dr. T. Ota. Regions were classified as "HV" or "conserved" following Dwyer et al. (1991) which is based entirely on variability differences. *P* are the proportion of radical or conservative amino acid changes with respect to changes in charge or polarity for the respective regions.

alleles from different species in the genealogy suggests that the *S*-alleles have been polymorphic since before species divergence (Dwyer et al. 1991, Ioerger et al. 1990, Richman et al. 1996).

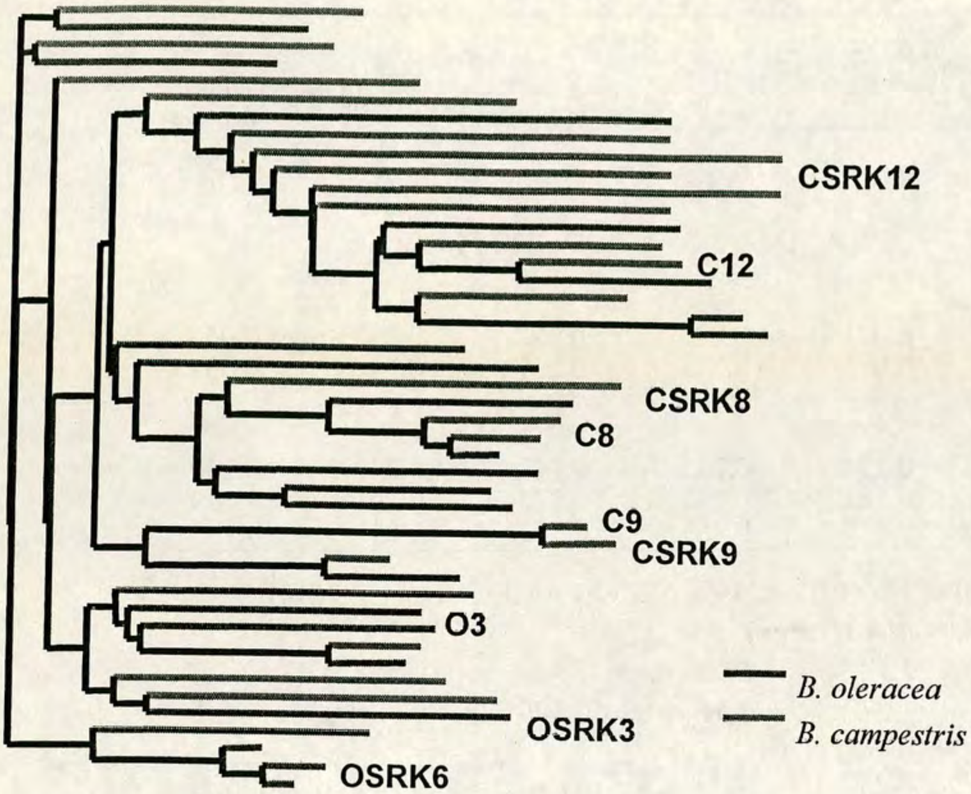


Figure 1.6. Neighbour-joining genealogy based on inferred amino acid sequences of of *SLG* and *SRK* alleles sampled from separate incompatibility-types from two different *Brassica* species. Genealogy was estimated using Protdist and Neighbor programs in the Phylip package (Felsenstein 1995). Each branch represents an *SLG* or an *SRK* allele. The different shades reflect S-domains of either the *SRK* or the *SLG* sampled from different species. Some *SLG* and *SRK* alleles sampled from the same haplotype are indicated merely to show that *SRK* and *SLG* alleles sequenced from the same haplotype are not invariably most closely related. 'O' refers to *B. oleracea* alleles and 'C' refers to *B. campestris* alleles. Those not labelled as *SRK* are *SLG* alleles.

If the high diversity observed at SI loci is a result of alleles maintained for long periods of time, rather than a rapidly mutating locus, then these loci should provide the power to detect ancient changes in population size. Pairs of alleles that are more similar to each other when sampled from the same species would indicate that those alleles diverged after the speciation event (Edwards et al. 1997, Richman and Kohn

1999). In gametophytic SI, researchers have used the degree of transpecies evolution for alleles sampled from different populations or species to infer such demographic changes or speciation events (Richman et al. 1996, Richman et al. 1997, Richman and Kohn 1999). Richman et al. (1996) showed that among species of the Solanaceae, a number of S-alleles from *Physalis crassifolia* are more similar to themselves relative to S-alleles from another species. Such observations may actually reflect concerted evolutionary processes (recombination, gene conversion, unequal crossing-over) which could increase within-species similarity, especially because populations should recover S-allele diversity very quickly through even very low rates of migration. Furthermore, given that much of the synonymous and nonsynonymous variation appears to be at saturation (Richman and Kohn 1999), it is difficult to interpret these observations.

Unless sequences can be exchanged between species by hybridisation, similar divergence of *S*-alleles within species, relative to between-species divergence, implies that the polymorphism must date from before the species split. In *Brassica oleracea* and *campestris*, mean within-species silent diversity (K_S) among *SLG* alleles and the *S*-domain of *SRK* are no smaller than the mean between-species estimates (K_B)(Figure 1.5 and Table 1.1). Coupled with the evidence that these *Brassica* species are relatively recently diverged based on cytological and nucleotide evidence (Lagercrantz 1998, Yang et al. 1999), this indicates that this polymorphism pre-dates the divergence of species. Finally, based on nucleotide variation, the S-domains of the *SRK* locus appear indiscernible from the *SLG* locus (Figures 1.6 and 1.9 below).

1.9 Variation within and among *SLG* and *SRK* alleles

Several factors combine with the great age of the alleles to make it difficult to test for differences in functional constraints between different parts of the *S*-loci. Diversity differences within the genes controlling the maternal components of both sporophytic and gametophytic SI (hypervariable regions in particular) have appeared robust as sequence data have accumulated from more alleles (Figure 1.7), and from more species (Hinata et al. 1995, Sakamoto et al. 1996, Matton et al. 1997). Such regions of high diversity are indications of sites that are potential targets of balancing selection, and which will also produce variability at sites not themselves under such selection, such as at synonymous sites in the same regions. The reason for this is that, depending on the rate of recombination (the degree of linkage), synonymous variation should increase as a result of being linked to sites under positive selection (Maynard Smith and Haigh 1974). An example of this is observed among antigen recognition sequences of the MHC loci (Hughes et al. 1990) where both synonymous and nonsynonymous diversity is elevated relative to the rest of the gene or genome. In SI, because multiple alleles are maintained, many sites must be targets of balancing selection such that different combinations of these sites (epistasis) can produce new S-alleles.

By examining the *SLG* and *SRK* loci from *Brassica* alone, we cannot tell whether HV regions are portions of the genes where amino acid differences determine the incompatibility types, or are neutrally evolving regions (Table 1.1). It is remarkably difficult to distinguish between the two extreme hypotheses, that

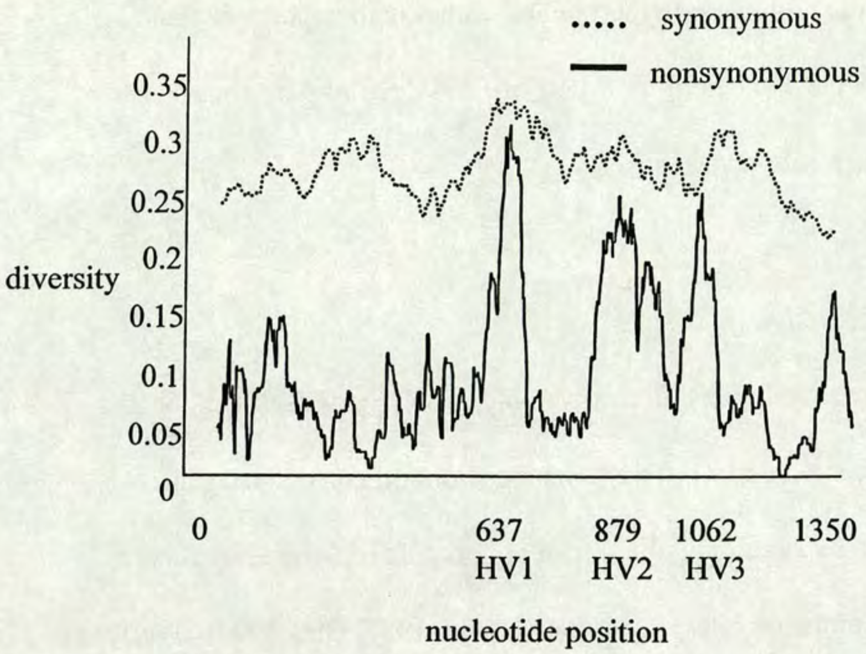


Figure 1.7: Sliding window plot (window size 30 nucleotides – shifted every position) of synonymous and nonsynonymous diversity per synonymous and nonsynonymous site, respectively, for *Brassica SLG* and *SRK* S-domains for Type I and Type II alleles. Numbers on the x-axis indicate the relative position of the the 'HV' regions. Sliding windows were calculated using DNAsp v.3 (Chapter 4).

highly polymorphic sites are the most important, or the least important. More direct evidence is necessary. For instance, the finding of two *SLG* alleles which have identical hypervariable region amino acid sequences suggested that either the *SLG* was not necessary for SI, or that other regions at the S-domain can determine specificity (Kusaba et al. 1997, Nishio and Kusaba 2000, Takasaki et al. 2000).

Examining S-alleles sampled from natural populations from separate related species may provide more clues about the nature of regions or sites that are targets of selection. For example, assessing diversity at similar regions at loci in the *SLG* and *SRK* gene family not linked to the S-phenotype, may provide some insight as to whether these regions are similarly polymorphic in loci that do not control SI. This would suggest that these regions are hypermutable rather than targets of balancing selection. But first we need to determine whether the HV regions are consistent or observed at S-loci in separate related species.

1.10 Evidence for diversifying selection

Instead of merely showing that diversity is high at *SLG* and *SRK* loci, or that some regions appear to evolve in a different manner from others, diversifying selection may be tested by examining the rate of amino-acid replacements versus neutral (synonymous) mutation rates; K_a/K_s ratios (Nei, 1987, Ohta 1994). Purifying selection usually constrains amino acid variation, so high values (e.g. *Ka/Ks* = 2-3 for the MHC antigen recognition sites and the eosinophil cationic protein (ECP) and eosinophil-derived neurotoxin (EDN) in primates in mammals) suggests balancing or positive selection (Hughes and Nei 1989, Zhang et al. 1998). Generally, this is a rather conservative approach for detecting positive selection, especially if estimates are made for the entire sequence, as some sites or regions will clearly not be under the same selective pressures and will therefore, obscure signals of positive selection. For the four variable regions identified in the *SLG* and *SRK* loci (Kusaba et al. 1997), K_a/K_s averages 0.95 for the *Brassica* species, as expected for neutrally evolving sites (Nei 1987). For the regions classified as conserved, values are lower (about 0.44), but remain relatively high. As explained above, however, the possibility that polymorphisms may be present because of linkage to selectively maintained sites somewhere in the locus makes it impossible to conclude from these results that the hypervariable regions are more important than the conserved ones. In fact, the HV regions may merely reflect the neutral mutations.

Diversifying selection should also lead to non-conservative amino acid substitutions, with changed charge or polarity, whereas other parts of the locus should have mainly amino acids with similar properties (conservative differences). Excess non-conservative amino acid substitutions have suggested the operation of diversifying selection at other recognition loci (see, e.g. Hughes et al. 1990), but in the *B. oleracea* and *campestris* S-allele sequences, non-conservative amino acid substitutions are slightly less likely than conservative changes, though much more frequent than in the kinase domain of *SRK* (Table 1.1). These data do not, therefore, indicate strong diversifying selection.

Finally, the accumulation of differences at non-synonymous sites in both the gametophytic and *Brassica* S-loci seems to be slowing down with evolutionary time. Non-synonymous differences increase more slowly than synonymous ones (which can be assumed to increase roughly in proportion to divergence times, see Figure 1.8), and this rate difference becomes more pronounced as the number of synonymous differences between the alleles compared gets larger (Hinata et al. 1995, Uyenoyama 1997). This could be due to replacement of alleles by descendents with new incompatibility types slowing down over evolutionary time. A more simple explanation may merely be that only a limited number of amino-acid replacement mutations are permitted for proper protein function and conformation. This will result in a slowing down of the accumulation of nonsynonymous mutations, whereas synonymous mutations will still be allowed to accumulate at the same rate because they are effectively neutral.

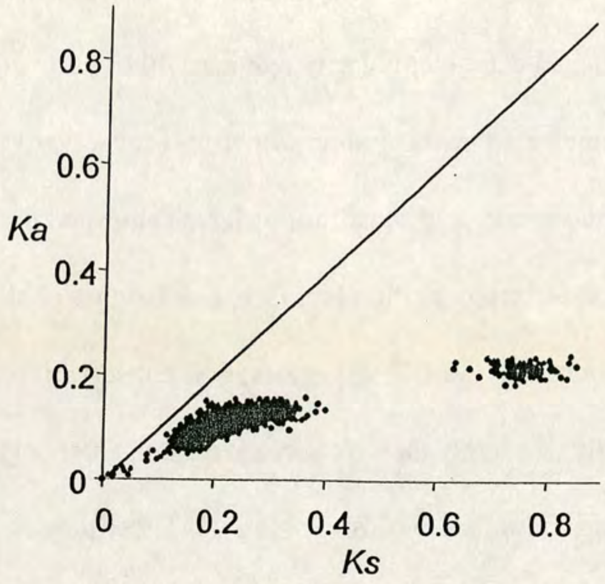


Figure 1.8: The relationship between Jukes-Cantor corrected synonymous and nonsynonymous accumulation of mutations for pairwise comparisons among 45 *Brassica SLG* and *SRK* alleles. The line indicates *Ka/Ks* values that would be equal to one. The smaller cloud of points indicates pairwise comparisons that include Type II (recessive) *Brassica* S-alleles which are clearly very different in sequence structure from Type I alleles.

1.11 Polymorphism at other loci in *Brassica*

To evaluate and interpret data from *S*-loci, genetic diversity data for non-*S* locus reference loci will be helpful as a basis for comparison. Other loci in the *S*-gene family, whose products are not essential for incompatibility or cross-pollination, apparently show much less allelic polymorphism, though natural populations have not yet been studied (Hinata et al. 1995, Chapter 4). Current scanty data on unlinked S-locus related (*SLR*) genes, non-essential for recognition, suggest sequence conservation within and between species (Hinata et al. 1995). The *S*-linked anther-expressed locus *SRA* appears non-essential for incompatibility (Pastuglia et al. 1997b) and has low polymorphism levels, and greater silent differences (i.e. larger coalescence times) for alleles between- than within-species (Hinata et al. 1995).

Natural population data for polymorphisms at loci both linked and unlinked to S-loci will allow us to subject the sequence data to molecular evolutionary analyses which assume random mating and sampling. Also, this will allow us to determine whether the same genes control SI function in other related species. We can then determine the number of S-alleles in natural populations and examine their level of nucleotide diversity. Furthermore, it is necessary to examine polymorphism at S-domain loci that are functionally involved in SI, and at related unlinked loci that are members of the same gene family, in species that occur naturally. Polymorphism at loci not functionally linked to incompatibility will provide a basis for comparison, especially with respect to patterns of polymorphism observed within the loci that are known to be linked to the incompatibility response.

1.12 Recombination in the *SLG-SRK* region

The sequences of the *SLG* gene and the *S*-domain of the *SRK* gene from any given haplotype are more similar on average, whereas those from different haplotypes differ extensively (Figure 1.9; Charlesworth and Awadalla 1998). For the few haplotypes sequenced for both loci, both synonymous and non-synonymous within-species differences between haplotypes average almost double their within-haplotype values (Figure 1.9), in other words, there is some linkage disequilibrium (but see Chapter 2). This has been taken to suggest that for functional self-incompatibility, alleles from the two loci must be similar (Stein et al. 1991, Takasaki et al. 2000) and is consistent with the idea of selectively maintained within-haplotype similarity between the two loci, but cannot rule out the alternative that sequence similarity is caused by between locus recombination. Further sequence data from *Brassica* now show that similarity is not invariably greatest within *SLG-SRK* haplotypes from plants with different incompatibility types (Goring and Rothstein, 1996, Nishio and Kusaba 2000, Kusaba et al. 2000) and that class I *SRK* alleles do not always cluster together with the *SLG* alleles from the same haplotype (Figure 1.6, Kusaba et al., 1997, Nishio and Kusaba 2000, Kusaba et al. 2000). If concerted

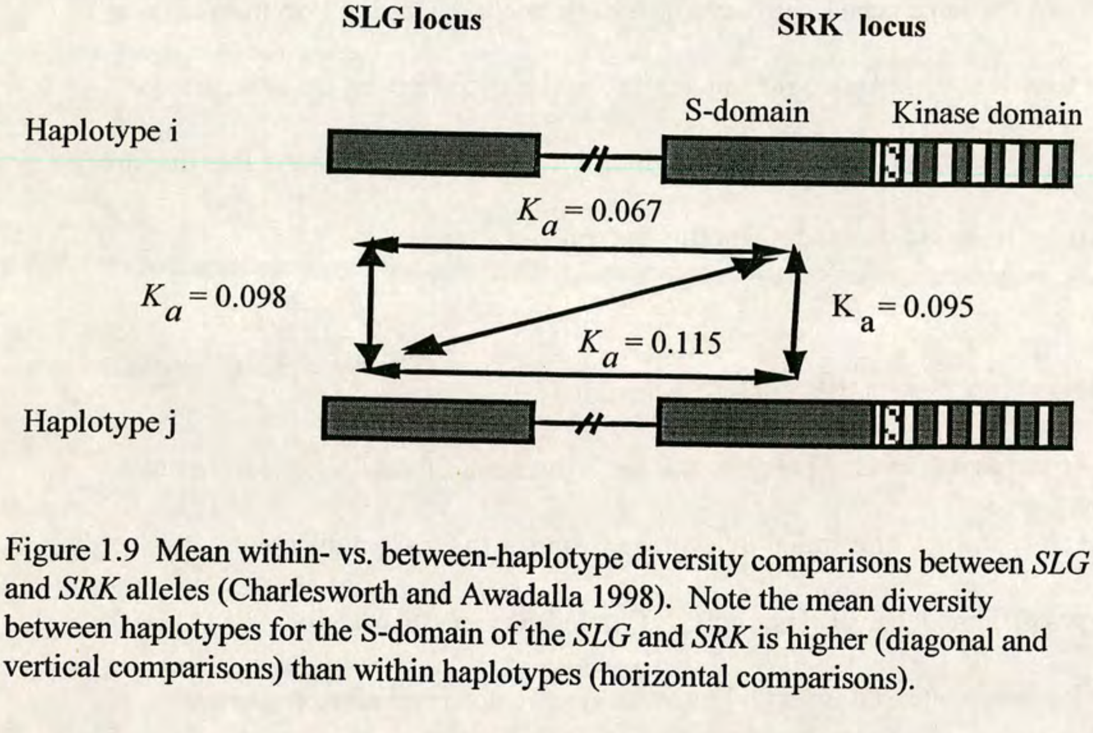


Figure 1.9 Mean within- vs. between-haplotype diversity comparisons between *SLG* and *SRK* alleles (Charlesworth and Awadalla 1998). Note the mean diversity between haplotypes for the S-domain of the *SLG* and *SRK* is higher (diagonal and vertical comparisons) than within haplotypes (horizontal comparisons).

evolution has homogenized pairs of *SLG-SRK* variants within some S-haplotypes (Kusaba et al. 1997) then, depending on the rates of mutation, it may take a long period of time for the mutations to re-accumulate. Given that the *SRK* locus appears to certainly be involved in recognition, and that the role of the *SLG* in SI is unclear, the observed similarity between *SLG* and *SRK* within a haplotype, and the overall high diversity at the *SLG* may just be a consequence of unequal-crossing over events and not reflect functional constraints within haplotypes.

1.13 Studies from natural populations – *Arabidopsis lyrata*

To examine SI from both an evolutionary and molecular biological context and to test population genetic models of SI, it is necessary to examine S-allele function and polymorphism in a natural population setting, preferably in different species such that generalizations can be tested. To this end, *Arabidopsis lyrata* was chosen as a study species to examine S-allele sequence polymorphism. Studying SI in *A. lyrata* has advantages over other *Brassica* species. First, *A. lyrata* has not been cultivated and, therefore, polymorphism patterns have not been affected by artificial selection. The *Brassica* species discussed above have been cultivated for agricultural purposes for long periods of time. Even natural populations may be affected by cultivation due to migration or contamination from cultivated varieties. Second, comparative approaches can then be used to examine the evolution of SI between related species within the *Brassicaceae* family. Finally, *A. lyrata* is closely related to *A. thaliana* where many orthologues of the S-loci gene family have been sequenced. Such orthologous comparisons allow us to examine whether rates of substitution vary among lineages, and how these substitution rates may vary within loci among the members of this gene family.

By studying SI evolution in *A. lyrata* we can address the following questions:

1) *A. lyrata* is more distantly related to the *Brassica* spp. described here relative to other species in the Brassicaceae family where S-alleles have been identified (i.e. *Raphanus sativus*, Sakamoto et al. 1998). Do the same genes segregate and control SI phenotypes in this more distantly related species to *Brassica*?
2) If the *SRK* or the *SLG* genes exist in *A. lyrata*, which of the two genes appear to be most important in SI recognition?
3) Are S-loci polymorphism levels in *A. lyrata* similar to that in *Brassica*? Studies from natural populations will provide us with a true assessment of the level of SI nucleotide diversity and we can assess the role of recombination in shaping haplotype structure more accurately than with cultivated varieties.
4) Does *A. lyrata* exhibit similar polymorphic patterns within loci in this more distantly related species? If so, this suggests that these patterns are not random and have been maintained due to some deterministic evolutionary process. The challenge would then be to determine which deterministic mechanism is operating at these regions. To examine this question, it is necessary to examine polymorphism and divergence of paralogues related, but not directly involved, in SI so that functional significance of polymorphic and fixed regions can be assessed.

1.14 Chapter Outlines

In **Chapter 2**, I examine the role that recombination may be playing in the *Brassica* SI system. As mentioned, the *SLG* and *SRK* genes are tightly linked loci and if variants at either or both loci are required to interact with a variable linked pollen gene (ie. the *SCR* gene), than this tight linkage is necessary to maintain a stable recognition system. With D. Charlesworth, in Chapter 2 I show that some recombination does occur in this region, contrary to previous suggestions, although the form of recombination remains unclear. Much of Chapter 2 was published in the journal *Genetics* by Awadalla and Charlesworth (1999).

The next three chapters focus on collaborative attempts, with Mikkel Scheirup and Barbara Mable, to identify the S-genes in *A. lyrata* individuals. In **Chapter 3**, I describe the identification of members of the large S-domain gene family in *A. lyrata* and define how each locus was characterized, and defined as separate loci from each other, including evidence from segregation analyses. A number of loci in *A. lyrata* share homology to the *Brassica* S-loci, and among them it has proved possible to identify candidate S-loci in *A. lyrata*. **Chapter 4** describes in detail the variability at all the S-domain family loci identified, focusing on those loci not linked to SI with the hope of testing which regions appear to be important in SI recognition. This was done by examining the distribution of mutation rates within and among these loci. The evolution of these loci as a gene family is also discussed. **Chapter 5** describes the molecular evolutionary properties of a highly variable group of sequences, where there is evidence for linkage to the S-phenotype in *A. lyrata.* The distribution of polymorphism and selective constraint among codons within and among these sequences is described, and polymorphism patterns were compared to that of *Brassica* and other SI systems is described. **Chapter 6** summarizes the main findings in this thesis and elaborates on future research directions.

Chapter 2

Linkage Disequilibrium and Recombination at the S-loci in *Brassica*

2.1.1 Introduction

In this Chapter, I describe attempts to detect recombination and the role it may be playing in shaping polymorphism at the self-incompatibility loci in two *Brassica* species, *Brassica campestris* (*rapa*) and *Brassica oleracea*. Chapter 1 describes the genes controlling the self-incompatibility recognition system (SI) in species of the mustard family (*Brassicacea*). The *SLG* and *SRK* loci share a region of homology (an "S-domain" within *SRK* is homologous to the *SLG*), both are highly polymorphic (Dwyer et al. 1991, Hinata et al. 1995, Kusaba et al. 1997, Nishio and Kusaba 2000), and they are tightly physically linked (Boyes and Nasrallah 1993, Boyes et al. 1997). Hypervariable regions appear to be present in similar regions at both loci (Dwyer et al. 1991, Hinata et al. 1995). Although it is not yet certain whether both genes have recognition functions, the *SRK* gene has been shown to be essential for self-incompatibility (Goring and Rothstein 1996), and although there is evidence for a role for *SLG* (Nasrallah et al. 1992), it has been shown that switching *SRK* alleles, but not *SLG* alleles, alter the SI phenotype (Takasaki et al. 2000). Nevertheless, it has been unclear until recently whether recombination occurs in the *Brassica* S-locus region. Recombination will clearly affect our interpretations of polymorphism within loci, the extent of shared polymorphism between loci and evolutionary and functional models about how the genetic components of the S-loci recognize 'selfed' alleles in the Brassicacea.

2.1.2 Recombination at S-loci?

Similarity between the S-domains of the two component loci in the same S-haplotype is greater than that between haplotypes (Stein et al. 1991, Dwyer et al. 1994, Hinata et al. 1995, Charlesworth and Awadalla 1998, Chapter 1). This suggested that for functional self-incompatibility, alleles from the two loci must be similar (Dwyer et al. 1994 - but see below). This point of view suggests that the variants at the two loci must be in linkage disequilibrium, and that suppressed recombination should exist between the loci to maintain functional haplotypes (Stein et al. 1991, Boyes et al. 1997). Low recombination would also be required in two-locus models of the genetic basis of incompatibility in which the pollen component of recognition requires a separate locus and recombination would otherwise create haplotypes with the pistil reaction of one type and the pollen reaction of another, which has never been detected (Lewis 1962). Given the recent discovery of the pollen determinant in *Brassica*, the SCR gene (Schopfer et al. 1999), genetic recogntion between pollen and stigma expressed proteins could be maintained between the SCR and the SRK-SLG haplotype either due to suppressed recombination, or selection for haplotypes with the correct allelic configuration among loci even in a recombining environment. Sequence rearrangements and high sequence divergence in the *Brassica* S-locus region (Boyes et al. 1997) and in the flanking regions of the S-locus of *Petunia inflata*, a species with gametophytic self-incompatibility (Coleman and Kao 1992), seem to support rarity of recombinational exchange. These data are not conclusive, however, as divergence could be caused by relaxed selection in these flanking regions. It is thus important to test whether recombination does or does not occur in the S-loci.

If the S-loci are recombining, than it is possible that sequence data can indicate which parts of the sequence encode the recognition functions. If there is recombination, then sites evolve independently, depending on the rate of recombination relative to the mutation rate. It seems reasonable to hypothesize that the most polymorphic regions of the genes, the hypervariable (HV) regions, may encode recognition regions of S-proteins as opposed to regions which exhibit high variability by chance (Hudson 1990). In other systems, convincing evidence of selection has been provided by the finding of regions with amino acid polymorphism, i.e., by *Ka/Ks* ratios (Nei 1987) exceeding unity, as in the case of major histocompatibility complex (*MHC*) loci (Hughes et al. 1990). In *Brassica*, even in the HV regions there are no *Ka/Ks* values greater than one and hence no evidence of positive selection (Table 1.1 and Figure 1.9). There is some direct evidence that exchanging the HV regions between S-alleles in species with gametophytic self-incompatibility can change their specificity in some cases (Matton et al. 1997), though tests involving different alleles have yielded different results (Kao and McCubbin 1996, Zurek et al. 1997) and it seems clear that in these species other regions of the gene can affect incompatibility types.

If recombination were totally suppressed, then sites at the S-locus region are not evolutionarily independent. The strong balancing selection in the S-loci would be expected to lead to high amino acid polymorphism throughout the gene, and all parts of the locus should exhibit similar long coalescence times (Strobeck 1983, Hudson and Kaplan 1988, Nordborg et al. 1996, Charlesworth et al. 1997, Takahata and Satta 1998, Schierup et al. 2000). Recombination, however, allows different segments of a gene to have different evolutionary histories (Hudson 1983, Hudson 1990), so that differences in functional constraint could generate different levels of polymorphism (Maynard Smith et al. 1993, Klitz et al. 1995). If a particular codon is under balancing selection, then synonymous variation for synonymous positions near that locus, and perhaps even amino acid replacements, are expected to be high (Strobeck 1983, Kreitman and Hudson 1991). Thus our capacity to detect functionally different regions in genes where variant alleles have persisted for very long time periods, as the S-alleles have clearly done, depends on the occurrence of recombination.

Genes or regions in a non-recombining environment may exhibit simultaneously high *Ka* and *Ks* values (e.g. 'hypervariable regions'), but these patterns may have nothing to do with positive selection. Such regions may be indications of regions under relaxed selective constraint for nonsynonymous mutations. Both *Ka* and *Ks* would be high, yet *Ka/Ks* ratios would not be significantly different from one in regions which may be under relaxed selective constraint, although *Ka* would never exceed *Ks*. If other sites at the locus were under balancing selection, than these alleles would be maintained for longer periods of time than a neutrally evolving locus. Mutations may then accumulate at the regions in the locus that exhibit relaxed selective constraint, elevating *Ka* values relative to the remainder of the gene where selection may be constraining amino acid polymorphisms at many of the sites not targetted by positive selection.

2.1.3 Detecting recombination at loci under balancing selection

No estimates of recombination rates from SI sequence data have yet been made. Such estimates may be impossible using standard approaches. Several methods exist to test for or estimate recombination in DNA sequence data (e.g., Stephens 1985, Hudson 1987, Sawyer 1989, Hey and Wakeley 1997), but balancing selection at the S-loci causes violations of the methods' underlying assumptions. One such violation is that multiple substitutions at individual sites are frequent in the data for these loci (~14% of the total number of sites in the *Brassica campestris SLG* gene, and 19% in *B. oleracea*), which violates the assumption of the infinite sites model (Kimura 1969) where sites should segregate for only two different bases. Furthermore, recombination will generally be incorrectly estimated for alleles maintained for long periods of time by balancing selection. Multiple substitutions can mimic recombination by generating 4 haplotypes between pairs of sites, but balancing selection can create genealogical effects (linkage disequilibrium) similar to those caused by population structure, and perhaps reduce estimates (Hey and Wakeley 1997). Furthermore convergent or parallel evolution will also appear as recombinant exchanges (Gustafsson and Andersson 1994; O'huigin 1995; Hughes and Yeager 1998). Nevertheless, using two tests sensitive to low rates of recombination (Stephens 1985, Sawyer 1989), Clark and Kao (1991) found evidence for some recombination at the self-incompatibility locus in species of the Solanaceae, even though the analyses were performed with few alleles and the alleles analyzed came from four different species.

Another difficulty with the data currently available in *Brassica* is that the sequences are not from a natural population sample. Rather, every sequence is from a different S-allele type sequenced from cultivated strains, which would on average be expected to differ more than randomly picked alleles from natural populations, as is the case for MHC alleles (Takahata and Satta 1998). Linkage disequilibrium will be higher as a result, because this is equivalent to sampling one individual from a number of different small populations (Takahata 1990).

2.1.4 Previous and current attempts to detect recombination at *Brassica* S-loci

The possibility of intragenic recombination within the *SLG* gene was suggested by Kusaba et al. (1997) , who analyzed the sequences of a subset of 6 out of 21 *B. oleracea* Type I alleles and found evidence for HV regions having been "shuffled" between alleles. Their approach examined whether the topologies of the gene trees differed when they were estimated from different regions of the set of sequences. However, statistical testing for whether differences are significant is a problem for this approach, because different regions of a sequence have different mutational histories and would yield different topologies (which might have support from bootstrap tests) even if there were no recombination. It would be possible to perform tests using likelihood ratios for one region based on its own topology vs. that estimated from other regions, but my attempts to do this using the Kusaba et al. (1997) data yielded ambiguous results, which depended on which particular sequences were included. A related, somewhat less ad hoc approach, based on differences in diversity in different gene regions, has been used as a test for recombination between MHC alleles, for which there is good evidence for the action of balancing selection (Hughes and Yeager 1998); it uses sliding windows analysis to estimate the variability in levels of nucleotide diversity between different parts of the sequence of the set of alleles, after removing the peptide binding region codons. This approach, however, requires prior knowledge of the functionally important regions of the protein product. It also suffers from difficulties similar to the tree-based approach and from the well-known problems of sliding windows methods, which can be very sensitive to window size.

To avoid some of the problems just mentioned and to attempt to test for recombination in S-loci, we calculated linkage disequilibrium estimates between pairs of segregating synonymous sites. With recombination, pairs of sites farther apart should exhibit less linkage disequilibrium than sites closer together in the sequence (e.g., Miyashita et al. 1993, Schaeffer and Miller 1993, Conway et al. 1999, Awadalla et al. 1999). This would be true even if recombination were infrequent (e.g., Guttman and Dykhuizen 1994) between different loci, or if gene conversion events occur, causing exchange of sequence information between different alleles at S-loci between either homologous loci or between different members of the S gene family (Andolfatto and Nordborg 1999). The chief purpose of this chapter is thus to examine the evidence for any recombinational process in the evolutionary history of the *Brassica* S-locus region.

2.2 MATERIALS AND METHODS

2.2.1 Sequences and alignments

DNA sequences of *SLG* and *SRK* loci from *B. oleracea*, *B. campestris*, and *B. napus* were obtained from GenBank. A total of 39 *SLG* sequences and 6 *SRK* sequences were used in these analyses (Table 2.1). These included only functional *SLG*-specific and *SRK*-specific sequences and only sequences encoding dominant (Type I) S-alleles. Currently, there are only 3 published Type II *SLG* sequences, and these differ greatly in sequence composition from Type I (Hatakeyama et al. 1998). Coding sequences were aligned with ClustalX version 1.64b (Thompson et al. 1997), after removal of introns in the case of *SRK* sequences (*SLG* genes and S-domains of *SRK*, which are the main focus of the analyses reported below, have no introns). Clustal outputs were manually edited using Seqpup v. 0.6f (Gilbert 1997), and reading frames and amino acid positions were checked against published amino acid sequences (Kusaba et al. 1997). A number of gaps are necessary to align the sequences. In our alignments of *SLG*, four indels are polymorphic in both species (one consisting of 3 nucleotides, two of length 6 nucleotides, and one 15 nucleotides long in an otherwise relatively conserved region); in *B. campestris* a further three gaps are necessary: two are due to single codon additions each present in just one sequence, and one is a polymorphic indel of 6 nucleotides, while in *B. oleracea* one additional codon is present in just one sequence. Four of these indels are in HV regions I and II. The total sizes, including alignment indels, of the *SLG* and the *SRK* coding regions were 1350 and 2780 bp, respectively. After removal of all indels and portions of incomplete information at the ends of the sequences, 508 of the 1116 remaining coding sequence sites in *B. campestris* (46% of all sites) and 585 (52%) in *B. oleracea* were polymorphic.

Table 2.1. DNA sequences used

DNA sequence	Species	Reference	Genbank accession number
SLG genes (Type 1)			
SLG8	*B. campestris*	Dwyer et al. (1991)	X55274
SLG9	“	Watanabe et al. (1994)	D30050
SLG12	“	Yamakawa et al. (1996)	D88469
SLG21	“	Kusaba et al. (1997)	D85213
SLG25	“	“	D85214
SLG26	“	“	D85215
SLG27	“	“	D85216
SLG30	“	“	D85217
SLG34	“	“	D85218
SLG35	“	“	D85219
SLG37	“	“	D85220
SLG38	“	“	D85221
SLG41	“	“	D85222
SLG45	“	“	D85223
SLG46	“	“	D85224
SLG48	“	“	D85225
SLG49	“	“	D85226
SLG99	“	“	D85227
SLG3	*B. oleracea*	Delorme et al. (1995)	X79431
SLG6	“	Nasrallah et al. (1987)	Y00267
SLG29	“	Trick and Flavell (1989)	X16123
SLG13	“	Dwyer et al. (1991)	X55275
SLG1	“	Kusaba et al. (1997)	D85198
SLG7	“	“	D85199
SLG9	“	“	D85200
SLG12	“	“	D85201
SLG14	“	“	D85228
SLG16	“	“	D85202
SLG17	“	“	D85203
SLG25	“	“	D85204
SLG28	“	“	D85205
SLG32	“	“	D88765
SLG35	“	“	D85206
SLG39	“	“	D85207
SLG46	“	“	D85208
SLG51	“	“	D85209
SLG52	“	“	D85210
SLG63	“	“	D85211
SLG64	“	“	D85212
SRK genes			
SRK9	*B. campestris*	Watanabe, et al. (1994)	D30049
SRK8	“	Yamakawa et al. (1995)	D38563
SRK12	“	“	D38564
SRK6	*B. oleracea*	Stein et al. (1991)	M76647
SRK3	“	Delorme et al. (1995)	X79432
SRK29	“	Kumar and Trick (1994)	Z30211

2.2.2 Estimation of diversity and rate heterogeneity

Mean pairwise proportions of synonymous substitutions (*Ks*) and nonsynonymous substitutions per site (*Ka*), and their standard errors, were calculated for the regions analyzed (see below) using MEGA version 1.01 software (Kumar et al. 1994) as in Chapter 1. Expected means and standard errors for *Ka*/*Ks* ratios were calculated using the Delta method (Bulmer 1979). Differences in these ratios were tested for significance by *Z*-tests, using the standard error estimates. Estimates of substitution rate heterogeneity for the inferred amino acid sequence across the *SLG* gene were done by calculating maximum-likelihood estimates of the shape parameter, α, of the discrete-gamma model of substitution rate, using PAML version 1.3 (Yang 1996, Yang 1997).

2.2.3 Estimating and determining the significance of linkage disequilibrium

Linkage disequilibrium estimates and values of the Hill and Robertson (1968) measure, r^2,where $r^2 = \dfrac{\left(p_{AB}p_{ab} - p_{aB}p_{Ab}\right)^2}{p_A\left(1-p_A\right)p_B\left(1-p_B\right)}$ were calculated between pairs of sites in *SLG* segregating for two nucleotides, using DnaSP version 2.91b (Rozas and Rozas 1997). Only *SLG* loci were analysed as there are not enough *SRK* loci available for an accurate analysis of linkage disequilibrium. No measure of disequilibrium is completely independent of the nucleotide (allele) frequencies at the sites compared (Lewontin 1988), so normalization is desirable because of the very different allele frequencies observed at different sites in the S genes. We also calculated *D′*, a measure of the degree of association between nucleotide variants of different polymorphic sites, normalized by D_{max}, the largest possible value of *D* given the nucleotide frequencies at the sites (Brown 1975, Lewontin 1988). For this reason, this measure appears to be preferable to other measures (Hedrick 1987). A disadvantage of D', however, is that its distribution includes a high proportion of values close or equal to 0 or 1 (Golding 1984, Hudson and Kaplan 1985, Hedrick 1987, Schaeffer and Miller 1993) and ignores haplotype configuration information. For example, the value of r^2 is greatest when the two sites have similar allele frequencies, and the two rare (common) alleles are in coupling. Sites are more likely to have the same allele frequency, and be in coupling, if they have not recombined, since they then share the same genealogy (Hudson 1988); thus sites close together are expected to have higher r^2 values. In contrast D' is equal to 1 if rare variants are either in repulsion or in coupling. Hence r^2 can potentially detect recombination, even when the fourth haplotype is not present and $D'=1$. Also, sites with multiple substitutions often tend to have low D_{max} values (because they often have low allele frequencies). Significance of the disequilibria was evaluated by two-tailed Fisher's exact tests, and the set of results was adjusted for multiple comparisons using Bonferroni correction (Sokal and Rohlf 1995). First, second, and third nucleotide positions of codons were analyzed separately.

For comparisons between pairs of sites where at least one had more than two nucleotides segregating, a program was therefore written to calculate D'. The significance of each pairwise measure of disequilibrium was tested by a permutation approach suggested to us by W. G. Hill using a Fortran program written by D. Charlesworth. For each contingency table, 100 randomizations of the observed values within the cells were performed, preserving the row and column totals. A contingency index (CI) was calculated for each permutation (Keeping

1962), the CI values were sorted, and linkage disequilibrium for the pair of sites was deemed significant at the $P = 0.05$ level if its CI was 95% of the values of the permuted tables. Only polymorphic sites in third positions of codons were analyzed in this way, owing to the very large numbers of pairs of polymorphic sites in these sequences.

The relationship between linkage disequilibrium and distance between the polymorphic sites was tested by Spearman's rank correlations. Because multiple pairwise tests involve the same polymorphic sites, we did randomization tests of the significance of the relationships found. We used the second procedure of Schaeffer and Miller (1993) , generating a large number of datasets in which the distances between polymorphic sites were randomly assigned from actual distances between these sites. (This is simpler than their first procedure, which assigns polymorphic sites to random positions in the sequence under study, but it should give similar results, as the S-allele sequences are polymorphic at almost half of all sites; see above). For each randomization, the Spearman rank correlation was calculated. The value from the actual sequence data was compared with the distribution of values from the randomized sets of data and was considered significant at a given level if it exceeded the relevant percentile. As so many sites are segregating in the *SLG* sequences, this analysis was done only for third position sites of codons with two nucleotide variants.

2.3 RESULTS

2.3.1 Nucleotide sequence differences within and between species for the *SLG* gene and the receptor-domain of the *SRK* gene

Some evidence about whether the S genes undergo recombination or not can be gleaned from analyses of sequence diversity of separate parts of the loci. Within the S-domains of both the *SLG* and *SRK* loci, HV regions occur in the same nucleotide positions (Kusaba et al. 1997, Charlesworth and Awadalla 1998). In Table 1.1 in the previous chapter, we showed that the mean pairwise synonymous and nonsynonymous diversity estimates for the HV regions was high relative to the remainder of the *SLG* and *SRK* loci, including the kinase domain. As mentioned before, the HV region *Ka*/*Ks* ratios are approximately equal to 1, and significantly different from values elsewhere in the genes (Table 2.2). For the *SLG* gene, the result of a Z-test (Li 1997) for the comparison between HV and other regions is -14.5 for *B. oleracea* and -18.2 for *B. campestris,* significant at $P < 0.01$. For the S-domain of the *SRK* gene, the test values are -2.64 and -3.81, respectively; again significant at $P < 0.01$. These differences are particularly striking because *Ks* values are also extremely high in the HV regions. Even in the HV regions, however, the mean values of this ratio were not significantly greater than 1.

Table 2.2. Distribution of *Ka/Ks* ratio estimates in different parts of the *SLG* and *SRK* loci of *B. oleracea* and *campestris*.

	SLG "remainder"	"HV"	*S*-domain "remainder"	*SRK* *S*-domain "HV"	kinase domain
Nucleotides	1107	235	1107	235	1365
Ka/Ks					
B. campestris	0.466	0.893	0.621	1.052	0.406
B. oleracea	0.418	0.996	0.387	1.124	0.356

These analyses demonstrate that different regions of the S-domains have significantly different diversity values for both replacement and silent sites (Chapter 1, Table 1.1; Table 2.2). The similarity in positions of the hypervariable regions in the two different *Brassica* species (see above) argues against a nonselective interpretation, but this is not conclusive because the species may be very close relatives (perhaps even able to hybridize occasionally). Alternatively, gene conversion between the two loci could potentially cause similarity between them.

2.3.2. Evidence for variability in substitutions at different sites in the *SLG* sequence

As a further test for heterogeneity in substitution rates across the gene sequences, and to compare variability at the S-loci with data from other loci, substitution rates per amino acid site were estimated using the discrete gamma model (Yang 1996, Yang 1997). This method is intended for analyses of divergence between gene sequences of different species, but it should also be appropriate for allele sequences in a nonrecombining region of genome. Increased rate variability might be expected if recombination occurs, though no explicit study of this appears to have been published. Because our aim is to test for recombination, nonsignificant rate variability would imply that there is no evidence for recombination from this type of analysis.

The shape parameters, α, of the gamma distribution of substitution rates (see Yang 1996) were estimated. Low values suggest rate variability within the sequence. Sites in *SLG* show rate variation (which was significant with $P < 0.01$ for both species by likelihood ratio tests): $\alpha = 0.57 \pm 0.064$ for *B. oleracea* and 1.28 ± 0.075 for *B. campestris*. These values are not particularly low (that for *B. campestris* is greater than 74% of the 51 available values for vertebrate nuclear genes; see Zhang and Gu 1998), reflecting the fact that the S-alleles show high variability throughout the sequence. There are too few *SRK* sequences for this kind of analysis to be meaningful.

2.3.3. Variation between different domains of the *SRK* gene

The coding sequence of the kinase domain of the *SRK* gene has less nucleotide diversity than the S-domain (Table 1.1, Hinata et al. 1995, Charlesworth and Awadalla 1998). Such polymorphism differences can again indicate independence within and among these linked loci. As just discussed, the polymorphism in the S-domain of the few *SRK* alleles currently available appears similar to that at the *SLG* locus (both have overall mean diversity per base of 0.13; see Hinata et al. 1995). The *SRK* kinase coding sequence, however, has fewer nonsynonymous and synonymous differences, and *Ka*/*Ks* lower than the conserved regions of the *SLG* gene or the *SRK* S-domain (Table 2.3). Consistent with the kinase domain's overall lower variability, probabilities of radical amino acid differences between alleles within both species average less than half those of conservative differences, unlike the situation in the S-domain (Charlesworth and Awadalla 1998). Finally, *Ks* estimates from the kinase domain are less within species than between them, unlike the situation in the S-domains (Table 2.3). There are thus clear signs of selective constraint acting at the *SRK* locus, although even in the kinase domain the *Ks* values show ancient divergence between alleles. The regional variability suggests that the polymorphism in the kinase domain is caused by linkage to sites elsewhere in the *SRK* (perhaps the S-domain) that are under balancing selection, with some recombination in the locus. If this were true, the polymorphism in this domain should be highest at the 5' end and would be expected to decline sharply. There is, however, no clear tendency in this direction. Silent polymorphism in the kinase domain declines slightly but nonsignificantly in both species (for *B. oleracea*, *Ks* in the 3' half of the kinase coding sequence is 86% of its value in the 5' half, and for *B. campestris* it is 87%). Of the three intron sequences from multiple alleles, indel polymorphisms are also very abundant in introns 2 and 3, and much less so in intron 4 (Nishio et al. 1997); further data on this domain may illuminate this question in the future.

2.3.4. Within- vs. between-haplotype comparisons of the *SLG* gene and the S-domain of the *SRK* gene

The view that recombinational exchange in the S-gene region is suppressed, initially appeared to be supported by the observation that the sequences of the *SLG* gene and the S-domain of the *SRK* gene from each haplotype tend to be much more similar than those from different haplotypes (Stein et al. 1991), but it has become clear that the similarity is far from complete, and that large differences between the two members of a given haplotype are not unusual (Goring and Rothstein 1996, Kusaba et al. 1997).

For the few haplotypes sequenced for both loci, quantitative analysis of the data from the two *Brassica* species shows that both *Ks* and *Ka* values between haplotypes within species average almost double their within-haplotype values (Table 2.3). These differences are highly significant. For the conserved regions, the

Z value for *Ka* is -6.70, and for the HV region sites it is 11.3, while for *Ks* the value is -9.96, and for the HV region sites it is 9.90 (all *P* values < 0.001). Furthermore, when the different regions of the S-domains are analyzed separately, the HV regions again stand out as different from the conserved regions. Conserved regions differ rather more between than within haplotypes in proportions of both synonymous and replacement substitutions, but the HV regions show greater differences between, though not within, haplotypes.

Table 2.3 Comparisons of *Ks* and *Ka* between *SLG* and the S-domain of *SRK*-alleles, within and between six different haplotypes for *B. oleracea* and *B. rapa*.

Region and diversity measured	*SLG*	*SRK*	Within haplotypes	Between haplotypes
Ka				
Entire gene	0.104 ± 0.006	0.111 ± 0.006	0.067 ± 0.033	0.115 ± 0.016
HV	0.190 ± 0.019	0.203 ± 0.018	0.079 ± 0.045	0.234 ±0.004
remainder	0.086 ± 0.006	0.090 ± 0.006	0.062 ± 0.031	0.094 ± 0.010
Ks				
Entire gene	0.190 ± 0.014	0.174 ± 0.013	0.122 ± 0.060	0.217 ± 0.029
HV	0.238 ± 0.038	0.217 ± 0.034	0.103 ± 0.084	0.316 ± 0.063
remainder	0.179 ± 0.015	0.164 ± 0.014	0.138 ± 0.005	0.192 ± 0.030

Comparisons are calculated for the coding sequence of the entire gene, the HV regions (averaged overall three HV regions), and the conserved regions.

2.3.5 Testing for linkage disequilibrium within the *SLG* gene

Figure 2.1 summarizes the patterns of linkage disequilibrium between pairs of segregating sites in third positions of codons within the *SLG* locus. There are currently too few sequences available to analyze the *SRK* locus in this manner. Three kinds of tests were done to ask whether linkage disequilibrium declines with distance in the *SLG* locus. In the first test, pairs of sites segregating for only two nucleotides were analyzed. These results are summarized in Figure 2.1, which shows, for the two species, estimates of r^2 for pairs of sites grouped according to the nucleotide distance between them. Results are also shown for sequences from which the HV regions, which exhibit clusters of linkage disequilibrium (see below), have been removed.

Linkage-disequilibrium values decrease significantly with distance. Spearman rank correlations of r^2 with distance were -0.071 and -0.134, for *B. oleracea* and *B. campestris* respectively, both with $P < 0.01$; the corresponding values when the HV regions were removed from the sequences were -0.074 and -0.083, both with $P < 0.01$. Out of 500 data sets with randomized distances between sites, none equalled or exceeded these correlation values for either species, either for all sites segregating for just two different nucleotides in third positions of codons, or for the set of sites excluding the HV regions.

A second test was based on all segregating sites in third positions of codons. These are too numerous to analyse the individual distance/disequilibrium values. Spearman rank correlations of D' with distance between sites were therefore performed using mean values for the distance categories in Figure 2.1. For the two species, the correlations were -0.54 ($P > 0.05$) and -1.0 ($P < 0.01$), respectively. This is a very conservative test for a decline in association with distance, owing to the high frequency of large D' values expected even in the presence of free recombination (see Materials and methods). Also, the means for pairs of sites at the greatest distances apart (which represent quite small proportions of all site pairs tested) can be inflated by a few high, but nonsignificant, D' values.

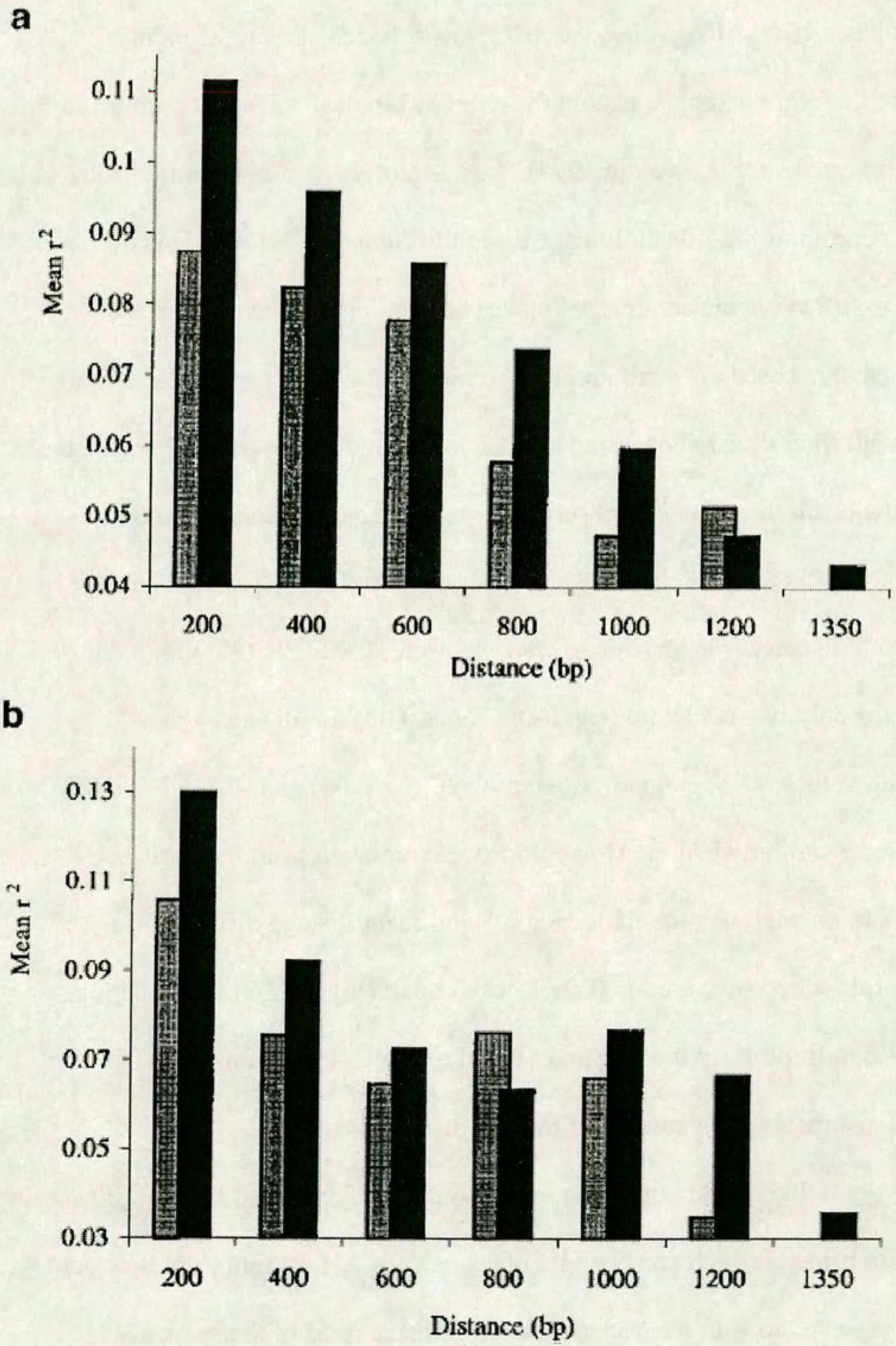


Figure 2.1. Relationship of linkage disequilibrium with distance for the *SLG* sequences of two *Brassica* species. Linkage disequilibrium estimates in terms of r^2 are shown for sites segregating for only two nucleotide variants, excluding pairs of sites with singleton variants (black bars), and after excluding HV region sites (gray bars). (a) *Brassica oleracea*. (b) *Brassica campestris*.

The frequencies of significant associations between sites at different distances are therefore better for testing the relationship of linkage disequilibrium to distance between sites (Lewontin 1995). It is also of interest to examine which parts of the gene show significant linkage disequilibrium (see below). This third kind of test is still very conservative as it takes no account of the values of the disequilibria and is based on small numbers of mean values (all tests became more significant with finer division of distances than in the figures). Figure 2.2 shows the effect of distance on the chance of observing a pairwise association that was significant at the $P = 0.01$ level. The Spearman rank correlations for the binned values for each distance category for *B. oleracea* were -0.99 ($P < 0.01$) for sites segregating for only two nucleotides and -0.77 ($P < 0.05$) for all sites; the corresponding values for *B. campestris* were -0.63 ($P > 0.05$) and -0.81 ($P < 0.05$). The results were similar when the HV regions were removed from the sequences (Figure 2.2). In all analyses, the frequency of significant linkage disequilibria drops off sharply when sites are >600 nucleotides apart (Figure 2.2). Calculations of linkage disequilibrium for the first and second nucleotide positions showed similar relationships for all measures of linkage disequilibrium.

It is very unlikely that alignment errors could have produced these results, even though in principle such errors might obscure linkage disequilibrium between distant sites. As explained in the Materials and Methods, most of the sequences align unambiguously despite the very high diversity, and, although there are eight indels, three of the eight are additions of single codons in just one of the sequences.

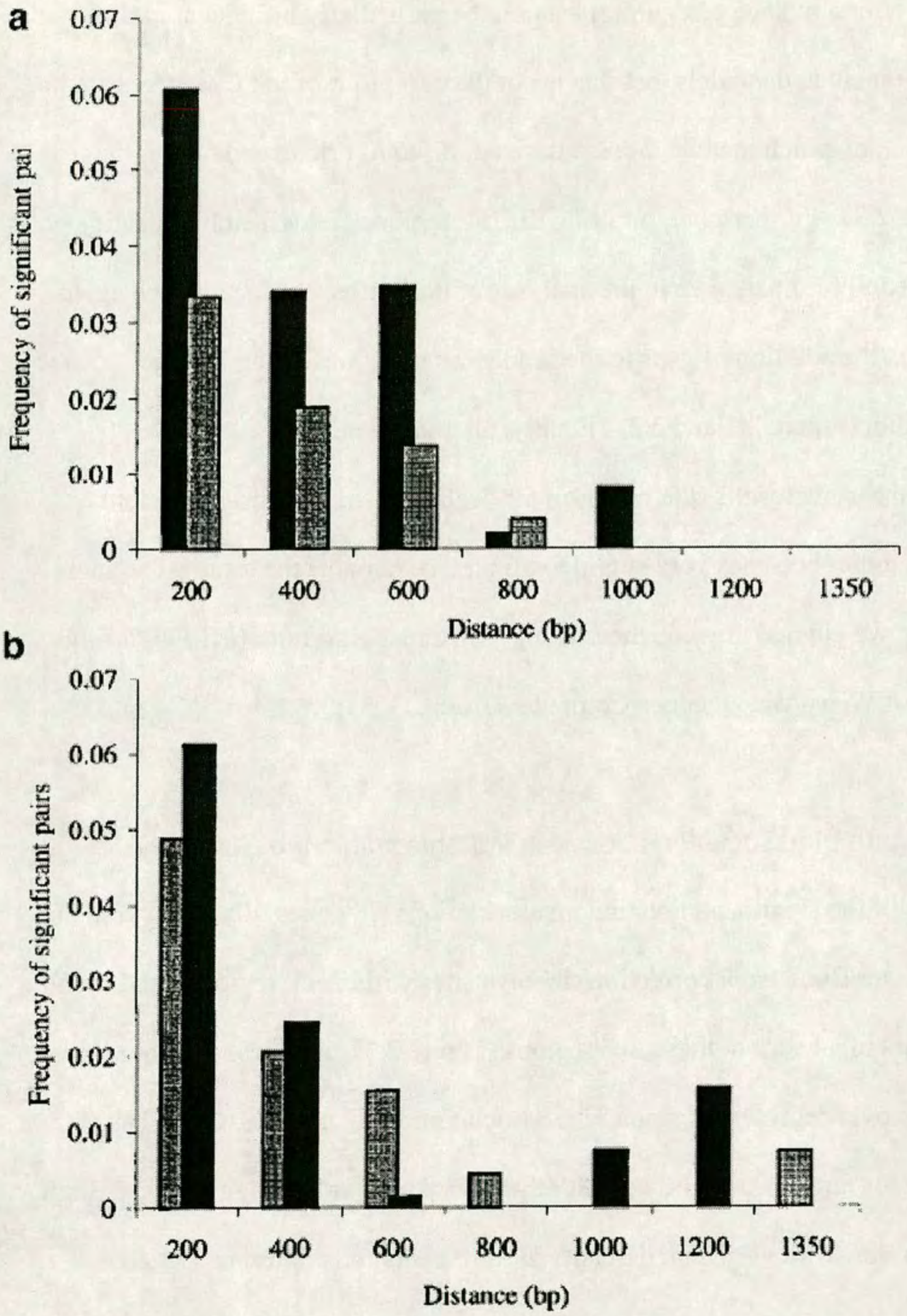


Figure 2.2. Frequencies of linkage disequilibrium in the *SLG* sequences that are significant by Fisher's exact tests, for sites at different distances in the two *Brassica* species. The analyses were done for sites segregating for only two nucleotide variants, excluding pairs of sites with singleton variants (black bars) and after excluding HV region sites (gray bars). (a) *Brassica oleracea*. (b) *Brassica campestris*.

With such errors, linkage disequilibria should be particularly infrequent in the regions that include the indels, but this is not the case; four of the five indels are in HV regions, but which include more significant disequilibria than other regions (Figure 2.2, 2.3). Furthermore, omitting the HV regions yields a similar decline of linkage disequilibrium as seen in the analyses of the entire sequence, which again suggests that the addition of gaps to the sequences is not obscuring linkage disequilibrium (Figure 2.1 and 2.2). Finally, alternative alignments yield essentially the same results (the evidence for decline in linkage disequilibrium within *SLG* either becomes very slightly stronger, or remains the same). The same is true when we aligned the sequences using a different algorithm (PILEUP of the University of Wisconsin Genetics Computer Group, GCG).

2.3.6 Structure of associations between variants within the *SLG* gene

Many of the significant nonrandom associations (*P* values <0.001 that remain significant after Bonferroni correction) involve sites within HV regions I and II and between pairs of sites from these two regions (Figure 2.3), and these sites are significantly over-represented among the associations that are significant (Table 2.4). This does not appear to be a result of greater power at HV regions due to high diversity, as sites with very high diversity at third codon positions are scattered

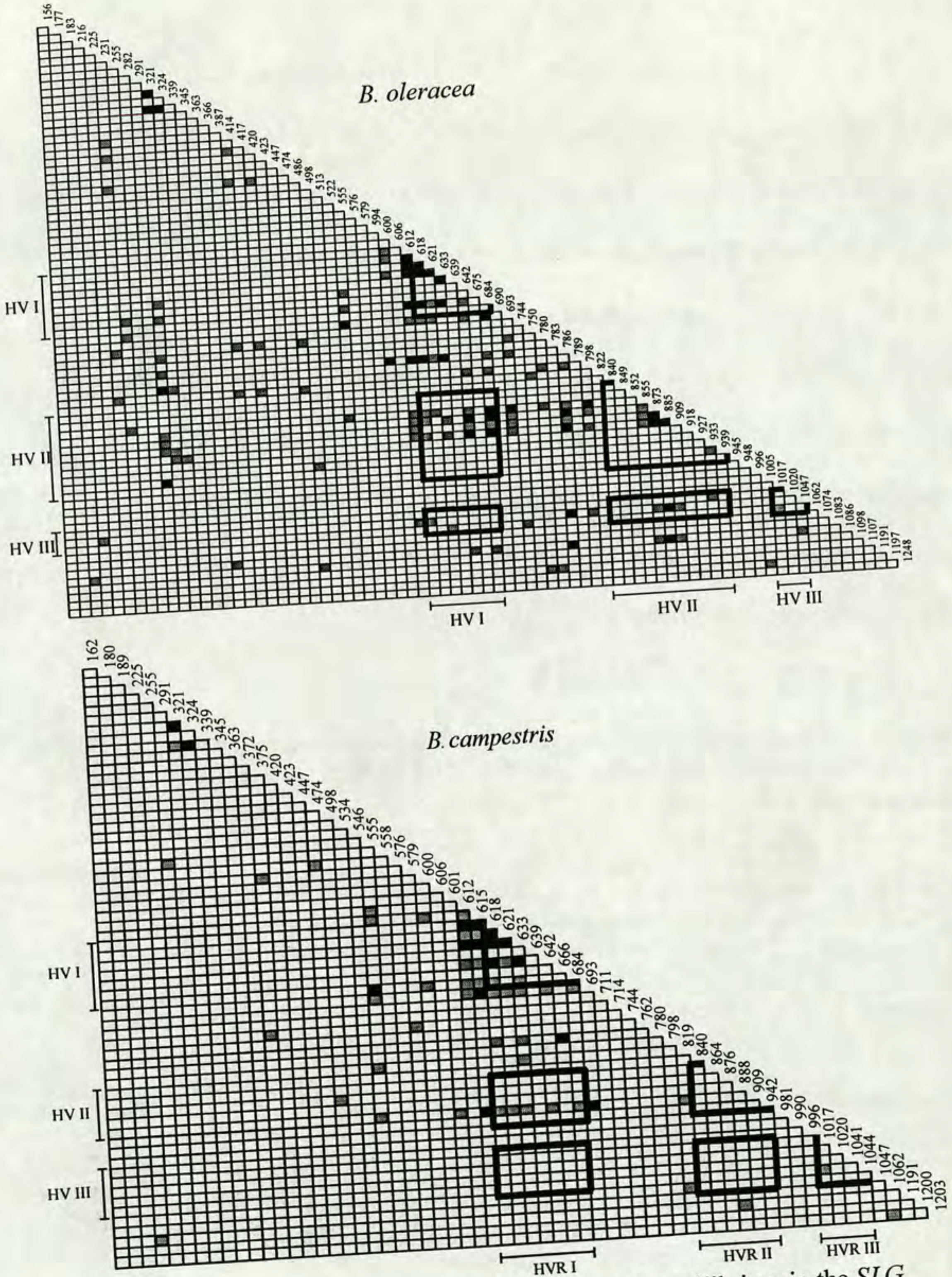


Figure 2.3. Results of Fisher's exact tests of linkage disequilibrium in the *SLG* gene. Empty boxes indicate pairs of sites that showed no significant associations, black boxes indicate significance after adjustment for multiple comparisons using Bonferroni correction, and gray boxes indicate all (uncorrected) tests that were significant with $P < 0.01$.

throughout the gene (Figure 1.7 and 2.2). Moreover, the C-terminal regions, with high diversity similar to the HV regions, include no sites in linkage disequilibrium. Significant linkage disequilibria was also detected for first and second segregating positions of codons across the *SLG* gene, and again the sites involved tended to be those within the HV regions. Given the effect of distance on linkage disequilibrium, the excess representation of these sites argues for functional importance of the HV regions. All pairs of first and second position segregating sites with significant values after Bonferroni correction involved replacement substitutions. Segregating second position sites that exhibit strong linkage disequilibria may thus be candidate amino acid positions that contribute to differences in allelic specificity.

Table 2.4. Tests of significance of the HV regions' representation among pairs of sites involved in linkage disequilibria based on third position sites segregating for two nucleotides only.

Regions	No. of segregating sites	No. of significant pairs*	Not significant
B. oleracea			
HV	21	15	195
Conserved	56	19	1521
	Chi-square with Yates' correction 30.84, $P < 0.01$		
B. campestris			
HV	17	5	131
Conserved	49	4	1172
	Chi-square with Yates' correction 15.32, $P < 0.01$		

* Pairs of sites involved in significant associations. $P < 0.001$ (after Bonferroni correction).

2.4 DISCUSSION

2.4.1 The effects of distance and region on linkage disequilibrium

Our analyses reveal a strongly significant effect of distance between sites in the *SLG* gene for two of the three estimators of linkage disequilibrium in both species analyzed, though the bimodal distribution of *D′*(Hudson and Kaplan 1985, Schaeffer and Miller 1993) obscures its relationship with distance. Reduced recombination has been documented for a number of different coadapted gene complexes such as in regions of genomes involved in recognition processes (Ferris and Goodenough 1994, May and Matzke 1995). Among these are the "supergenes" that appear to control the phenotypic differences, including incompatibility types, of the two morphs in distylous plants (Haldane 1933, Ernst 1936, Charlesworth and Charlesworth 1979). The evidence for these supergenes is at present entirely from classical genetic studies; no molecular analyses have yet been possible, so definitive evidence is lacking. The evidence for a tightly linked complex of genes in the S-locus region is based on molecular data, but we have argued here that the sequence data give grounds for suspecting that some form of recombination has occurred over the evolutionary time during which this polymorphism has persisted. Over the relatively small lengths of DNA in the S-locus region, the estimated average recombination frequency for *Brassica* suggests that the map distance would be <1 cM, so recombination is unlikely to be detected in classical linkage tests with the family sizes that are normally used (Boyes and Nasrallah 1993, Yu et al. 1996, Conner et al. 1998). Nevertheless, over evolutionary time even rare events can cause different parts of a gene to have essentially independent evolutionary histories (see, e.g., Hudson 1990, Guttman and Dykhuizen 1994, Nordborg et al.

1996, Charlesworth et al.1997, Kelly 1997, Andolfatto and Nordborg 1998). In MHC loci, for instance, sites outside the peptide-binding regions are much less diverse than sites within those regions themselves (Takahata and Satta 1998).

Recently, Casselman et al. (2000) in an analysis of 400 F2 individuals, discovered a number of recombination breakpoints in the S-locus region in *Brassica*, recombination in regions flanking the *SLG* and *SRK* loci, verifying the main results of this chapter. In the haplotypes analysed by Casselman et al. (2000), the reading frame orientation was such that the 5' ends of the the *SLG* and *SRK* loci were closest to each other (transcribed away from each other). Recombination breakpoints were found at regions flanking the the *SLG* and *SRK* loci but not between them, keeping the haplotype relationships of the loci involved intact. These breakpoints were reasonably close to the *SLG* and *SRK* loci, with one breakpoint as close as 3 kb to the 3' end of the *SLG* gene and the next closest was approximately 30 kb from the 3' end of the *SRK* loci.

In the analyses of linkage disequilbrium presented here, significant clusters of linkage disequilibrium were found predominantly to involve sites in the HV regions. If recombination occurs in the S-domains of the incompatibility loci, this may suggest that these sites have some functional importance in recognition processes. As suggested by Takasaki et al. (2000), although the *SLG* locus may not be directly involved in recognition in some haplotypes, it may be involved in the SI response. It is possible that this 'involvement' is subject to balancing selection. Before we can conclude that this is true, we must, however, consider alternative possibilities. Population subdivision cannot explain these findings, as it should cause disequilibrium across the whole gene and does not produce a clear relationship between distance and linkage disequilibrium. The relationship of linkage disequilibrium with distance exists even when the HV regions are removed from analyses, so it is not simply caused by greater power to detect associations in the most variable regions. Alternatively, the *SLG* variation may merely be a reflection of selection occurring at the *SRK* locus. Concerted evolution could create the polymorphism observed at both of these loci, even though the SRK locus may be the sole maternal determinant of SI and subject to balancing selection. Nevertheless, the evidence is consistent with the view that sites in the *SLG* gene have recombined over evolutionary time, both within and outside the HV regions, so that sites far apart are not in linkage disequilibrium, even if sites close together are. Even if the HV regions have less recombination than the rest of the gene, this could not account for these regions' locally higher diversity, unless they differ in their selective regime.

The view that the S-alleles recombine or undergo some other kind of exchange, such as gene conversion, is in apparent contradiction with some recent evidence that recombinant alleles (chimeric constructs between two *Nicotiana alata* alleles; Zurek et al. 1997) may fail to be recognized by pollen of either allele. Such tests have not as yet been performed in *Brassica* species, but if this turns out to be a general property of S-loci, it would imply that recombination yields alleles that are nonfunctional in incompatibility (Casselman et al. 2000). If such alleles are regularly generated in self-incompatible species and regularly eliminated, a selective force against recombination would be generated. Yet it may be possible that recombination produces new S-alleles by creating new combinations of haplotypes and, therefore, is a diversifying force rather than just a destructive one.

2.4.2 Patterns of diversity: balancing selection or reduced selective constraints?

Taking the conserved regions as a reference, it appears that HV regions are evolving in a manner different from that of other parts of the *SLG* gene. There are two very different possibilities for these regions. The hypothesis that they are under balancing selection is attractive, but it is difficult to rule out the possibility that the HV regions are evolving neutrally, at least in the *SLG* as it now appears that the *SRK* is more involved in recognition of pollen S-proteins than the *SLG* locus (Takasaki et al. 2000). *Ka*/*Ks* ratios approximately equal to one are usually considered evidence of neutral evolution, whereas genes under balancing selection may have values greater than one (Nei 1987). Under balancing selection, we may expect to see an initial increase in *Ka* relative to *Ks* early in the evolution of polymorphism at these loci (Nei 1987). However, initial high values will be expected to change over evolutionary time, because once silent substitutions are close to saturation, *Ka*/*Ks* will tend to increase. If, however, *Ka* ceases to increase, the opposite change could occur. This might happen if only a subset of amino acid residues undergo adaptive substitutions, while others are conserved. Thus an initial high ratio could fall below 1 (Neilsen and Yang 1998, Yang and Neilson 2000). In addition, assignment of sites as replacement or silent becomes uncertain over long evolutionary times, when divergence between sequences is high. *Ka*/*Ks* ~1 could therefore result from balancing selection or neutrality, and it is impossible to infer selection unless ratios well above 1 are found, which is not the case for the sequences examined here.

HV regions have been identified in essentially identical regions in S-genes of other species, including *B. napus* (a species of hybrid origin with *B. campestris* and *B. oleracea* as putative parents; see Goring et al. 1992) and *Raphanus sativus* (Sakamoto et al. 1998). It would be highly unlikely that different genes in different species should share similar diversity patterns by chance. The data might then be taken as evidence for selection acting in similar functional parts of the proteins these genes encode. The similarity in location of HV regions could, however, be caused by relaxed selective constraints in these regions. Regions under relaxed selective constraint should, however, have diverged since the two *Brassica* species analyzed here separated from one another, which we do not see here (Table. 2.2). Thus the diversity data alone do not allow us to distinguish definitively between these possibilities.

Chapter 3

Characterizing the S-domain Gene Family in *Arabidopsis lyrata*

3.1 Introduction

Molecular and cell biologists interested in the function and evolution of sporophytic self-incompatibility within the Brassicaceae family have focused on a few model species; *Brassica campestris* (*rapa*), *Brassica oleracea*, and *Brassica napus*, a tetraploid species considered to be a hybrid of the former two species (Boyes et al. 1992, Goring and Rothstein 1996). Although a few *Brassica* S-alleles have been isolated from other species in the Brassicaceae family, such as *Raphanus sativus* (Sakamoto et al. 1998), these species are phylogenetically very closely related to the *Brassica* species (Galloway et al. 1998) and have likely hybridized among themselves in their evolutionary history (Karpechenko 1927, Grant 1981). Therefore, most of our current understanding concerning sporophytic SI systems comes from a few closely related species. To make generalizations about SI function, polymorphism, and evolution, it is necessary to study the mechanisms controlling recognition in more distantly related species. For example, it may be that more distantly related species have only one of the two maternally expressed loci (*SLG* and *SRK*). It would then appear that only one locus is absolutely necessary for SI. Furthermore, if the patterns of polymorphism at *Brassica* S-loci are observed in more distantly related species, it is then clear that these patterns are not merely observations due to chance or random stochastic processes.

In order to subject the loci involved in self-incompatibility to molecular evolutionary analyses, it is necessary to study alleles sampled randomly from natural populations. Preferably more than one population should be sampled and the S-phenotypes of each individual should be determined. Determining the S-genotype for each individual involves growing maternal lines from seeds sampled from natural populations, and determining their incompatibility types by performing crosses among F1 offspring of compatible matings. Inferring S-phenotypes based solely on nucleotide differences (i.e. Richman et al. 1996) is problematic because if there are only a few nucleotide differences observed between two sampled sequences, one cannot then say whether those sequences are of the same incompatibility class, or whether those mutations are indicative of a different S-phenotype. No study has yet been attempted where alleles were sampled randomly from *Brassica* populations in this manner.

The S-locus in *Brassica* appears to be part of a large gene family. It is imperative that the other members of the gene family be characterized and examined using molecular population genetic approaches in order to understand how the S-loci evolve. The *Brassica* S-gene family consists of two groups of loci; the first group are similar to the *SLG* locus, in that these homologues are alignable to the *SLG* and do not have a kinase domain (Isogai et al. 1989, Boyes et al. 1991, Luu et al. 1999).

The second group consists of sequences that are alignable at both the S-domain and the kinase domain to the *SRK* locus (Dwyer et al. 1994, Suzuki et al. 1997). Homologues of both types have also been characterized in *Arabidopsis thaliana* (Dwyer et al. 1994), a predominantly selfing species. For the few genes sequenced, average replacement site divergence between *B. oleracaea* and *A. thaliana* is around 6-8% (Innan et al. 1996, Kawabe et al. 1997, Galloway et al. 1998, Miyashita et al. 1998, Purugganan and Suddith 1998, 1999, Stahl et al. 1999, Savolainen et al. 2000).

Some of the *Brassica* and *A. thaliana* S-gene homologues have been characterized on a functional basis. For example, the *SLR*1 gene in *Brassica* encodes secreted glycoproteins that share very similar primary structural features with *SLG*s, is expressed in both stigma and leaf tissue, and appears to be necessary for correct pollen-stigma adhesion, but is not directly involved in distinguishing alternative S-alleles (Lalonde et al. 1989, Isogai et al. 1988). *SLR*1 is thought to be orthologous to an *A. thaliana* gene, *ATS*1, which is also expressed extra-cellularly in stigmas and has a similar role (Dwyer et al. 1994). The *Brassica* kinase loci, or *ARK* genes in *A. thaliana*, are vegetatively expressed and at least three have been well-characterized (Dwyer et al. 1994). *ARK1* protein products autophosphorylate on serine and threonine residues (like the *SRK*) and the genes are expressed in leaves, flower buds, and stigmas (Tobias and Nasrallah 1996, Suzuki 1996). *ARK*2 is very similar in sequence structure to *ARK*1 and is specifically expressed in cotyledons, leaves, and sepals during maturation of these structures. Finally, *ARK*3 promoter expression has been detected specifically in roots and the root-hypocotyl transition zone (Dwyer et al. 1994). In *A. thaliana*, *ATS*1, *ARK*1 and *ARK*2 are located on chromosome 1 and *ARK*3 is located on chromsome 3 (Dwyer et al. 1994). The non-overlapping pattern of expression, and their sequence similarity, suggests that these genes have duplicated and evolved to perform non-redundant functions. It is clear that the genes of this family have arisen through gene duplication at some point in their evolutionary history. Concerted evolutionary processes, such as gene conversion or unequal crossing-over (Awadalla and Charlesworth 1999, Casselman et al. 2000), appear to have occurred subsequent to the initial duplication events creating the similarity we observe between *SLG* and *SRK* loci within haplotypes (Awadalla and Charlesworth 1999). Similar events may have occurred between these loci and other members of the gene family in the Brassicaceae. Furthermore, such recombination events may affect estimates of divergence, neutral mutation rates, or mutation rates to new S-alleles. Therefore, it is necessary to examine how this gene family is evolving.

Examining natural population diversity and divergence at loci not linked to SI provides a means to compare and evaluate polymorphism at candidate loci that are linked to self-incompatibility. How polymorphism and divergence are distributed across the sequences of the loci not directly involved in SI, both within and between orthologues, can help to explain patterns of polymorphism at S-loci (e.g. HV regions, see Chapter 5). For example, do non-S-loci show elevated polymorphism levels at similar aligned positions as at the HV regions in the *SLG* and *SRK* loci (Boyes et al. 1993)? Is divergence at these regions between paralogues extraordinarily high, or is there high selective constraint? If these regions at loci not linked to SI exhibit high polymorphism, then it seems likely that these regions are generally regions where mutations are free to accumulate rather than regions targeted by selection.

To address the questions outlined above and those raised in Chapters 1 and 2, we are currently studying natural variation at S-domain loci in a self-incompatible species of the Arabidae tribe in the *Brassica* family, *Arabidopsis lyrata* L., a close relative of the model species, *A. thaliana*. Both *A. thaliana* and *A. lyrata* are diploid species but the number of chromosomes in each species differs, with $2n$ =16 for *A. lyrata* and $2n$ = 10 for *A. thaliana* (the *Brassica* spp. chromosome number is $2n$ = 18). *A. lyrata* has not been subject to cultivation and populations are known to occur throughout the northern hemisphere. Comparative mapping of the *Brassica* S-region (Conner et al. 1998) and data on *A. thaliana* homologous markers, tentatively suggest that the *A. lyrata* S-locus should be on chromosome one, but this has yet to be confirmed (A. de Haan pers. comm.).

I describe here the isolation of seven S-domain sequence types in *A. lyrata* using PCR-genomic approaches. These S-domains appear to represent distinct loci and we have tested the majority of them for linkage to the S-phenotypes using segregation analyses. One sequence type appears to consist of at least two loci that are *SLG*-like and *SRK*-like, is highly variable (Chapter 5), and variants segregate with the incompatibility groups determined from crosses performed among F1 individuals. Finally, based on alignments and divergence calculations, I suggest putative orthologues of these loci that have been sequenced in *A. thaliana* and *Brassica*.

3.2 Materials and Methods

3.2.1 Plant Material

A. lyrata populations have a wide distribution throughout North America. Populations formerly classified as *A. petraea,* are now considered to be the same species as *A. lyrata*, and these populations are found in a scattered distribution throughout Europe (Price et al. 1994). In this thesis, both *A. lyrata* and *A. petraea* populations will be referred to as *A. lyrata*.

Seed material from 4 populations of *A. lyrata* were collected from North Carolina (NC), Indiana Dunes (Indiana; Ind), Braemar Scotland (R. Ennos) and the Reykjanes peninsula in Iceland. Plants were grown in individual pots in the greenhouse. Leaf material was used for DNA extraction using a CTAB protocol (Junghans and Metzlaff, 1990).

To determine incompatibility types, crosses of plants were performed in the greenhouse. Two plants were crossed, and the seeds were harvested and sown. Individuals were tested for compatibility by scoring whether seed set occurred after manual pollinations among individuals. D. Charlesworth, M. Schierup and B. Mable performed the crosses in the greenhouse. The same plants were used for segregation analyses to test for linkage of S-domain types to the S-phenotype. Segregation analyses for families described here were done by myself, B. Mable and M. Schierup.

3.2.2 Primers, Amplification, Cloning and Sequencing

Primers used to amplify members of the S-domain gene family in *A. lyrata* were designed based on sequence alignments of *Brassica SLG* and *SRK* loci. As diversity is known to be high among S-alleles (S-domains) within the *Brassica* species, we expected that divergence between *A. lyrata* S-alleles and *Brassica* S-alleles would also be high. Furthermore, we expected that a large number of genes in *A. lyrata* would exhibit homology to *Brassica* S-domain sequences and as a result, it would initially be difficult to distinguish these genes from highly variable S-alleles. Therefore, primers were designed in relatively 'conserved' regions of the S-domain and the kinase domain in order to amplify as many S-domain sequence 'types' as possible. More specific primers were then designed to test the 'allelic' nature of the sequences amplified. Table 3.1 describes many of the primers used in this study. Amplifications were performed under standard conditions.

J. Nasrallah isolated a cDNA S-domain sequence isolated from stigma tissue mRNA of an *A. lyrata* plant. Attached to this S-domain sequence is a transmembrane domain and a kinase domain alignable to the kinase domain of *SRK* in *Brassica* (J. Nasrallah, pers. comm.). The S-domain of this variant was homologous to a variant of a locus we had identified called *Aly*13 (see Results), which was isolated from two separate individuals by myself and M. Schierup (pers. comm.), and had been classified as variant *Aly*13-13. B. Mable designed reverse primers to anneal to the kinase portion of this cDNA sequence (Table 3.1). These primers were used to test whether individual isolated S-domain sequences from *A. lyrata* each had a kinase domain downstream from the S-domain.

PCR elongation times for most loci were one minute. The exceptions were for longer loci with kinase domains where 2.5-3 minute elongation times were used. Primer annealing was performed at 52-53 °C at 2mM $MgCl_2$ concentrations. Primer design and PCR reactions identifying the S-domain sequences discussed in this chapter was done by myself and M. Schierup.

Because initial amplifications were expected to amplify more than one locus or allele, PCR products of the expected size were cloned. The PCR products were extracted from 1% agarose gels and purified with Qiagen purification kits. The TOPO TA cloning kit (Invitrogen) was used to clone PCR products into the pcr4-TOPO vector. Transformed *E. coli* (TOP10 cells, Invitrogen) were plated onto LB-kanamycin plates (only an antibiotic screen is required with the TOPO vector), and approximately 15 colonies were picked per plate for screening. Screening for inserts was performed by PCR amplification of individual colonies using M13 universal primers. These single-copy amplification products were digested with 4 - and 6-cutter restriction enzymes at 37°C overnight (see below) to detect sequence variants or new loci. The digests were run on 2 or 3% agarose gels.

Cloned inserts were sequenced using standard cycle sequencing protocols for the Applied Biosystems model 377 Sequencer, with the Big Dye sequencing kit, using M13 universal primers. Direct sequencing was performed using primers specific to the original amplified product, after gel purification of PCR reactions.

3.2.3 Sequence alignments and analyses

Sequence alignments were made using Clustal X (Thompson et al. 1994), followed by manual adjustments based on amino acid alignments. *A. lyrata* S-domain sequences were aligned to *Brassica SLG* alleles (see previous Chapter for accession numbers) and nucleotide sequences of *ATS*1, *ARK*1, *ARK*2, and *ARK*3 sequences to anchor alignments according to conserved amino-acid properties common to S-domain loci in the *Brassicaceae* (Section 3.3.7; Kusaba et al. 1997).

Table 3.1 List of primers used to amplify initial S-domain sequences and primers designed to amplify specific loci.

	Primer	Sequence (5' to 3')	Approximate PCR product size (without indels)	Position of primer relative to sequence alignment *SLG*
Primers based on aligned *Brassica* sequences				
	SLGF	AGA ACC TAT GCA TGG GTT GC	1048 (with SLGR)	308
	SLGR	ATC TGA CAT AAA GAT CTT GAC C		1340
	PSA	AGA ACA CTT GTA TCT CCC GGT	1152 (with PSB)	226
	PSB	CAA TCT GAC ATA AAG ATC TTG		1360
	SI1	GCT TGG TTT CTT CAC TCC	1150 (with SLGR)	250
	SI2a	CGG TCC AAA TCA CAC AAC	1190 (with SLGR)	210
	Ats1F	AAC TTC GTG ATG CGA GAC TCC	814	488
	Ats1R	CGG TCC AAA TGA CAC AAC CCG		1290
Primers designed from *SRK* alignments				
	SRK2-F	ACG GGT GTG TGT ATT TGG ACT GGA		1221
	SRK4-R	TTT CGG TGG CTT TGA CAA CAG	1285 (with SRK2-F) 2254 (with SLGF)	1564
	SRK5-R	ACA GCT TCT AAC TCT ATC AAT GGA	1269 (with SRK2-F) 2239 (with SLG-F)	1581
	SRKNAS-R4	TGC CCG TCA GGT AAC CTT CCC	ca. 2.3 kb, intron length is variable (with SLG-F)	
Primers designed from *Arabis lyrata* sequences				
3	3F	CAA CAA CAA CGA CCA CCC AAG	680	360
	3R	TGC ACC TCT CGA AAC CAT TCC		1184
7	7F	CAA TCC TAT AGG AAC CAT CCG CG	860	288
	7R	TCG TTG AGA CCA ACT TCC CA		1184
8sk	8F	ACG GGA ACT CTC AAA ATA TCC G	870 (also with 10R for *Aly*10sk2)	295
	8R	CCG CCA CCA CGA ATA TCC G		1253
9	9f	TCA ACG AAT CTG ACG AGA ATG G	910	354
	9R	GCG ACG TAA TCC TCA TAT CTC TG		1319
10sk1 & sk2	10F	TCC ATC AAG CGG CGA TTT CTC GA	500	591
	10R	GTC AAC CGC ACA AAC CCT CCT CTA		1124
13	13F1	CCG ACG GTA ACC TTG TCA TCC TC	970 with (SLGR)	314
	13int1	5' ATA CCG GAC TGG TCC ATG G 3'	635 with (SLGR)	674
	13seqF1	TGG AAA ARC TCG TAT GAT CC	714 with (SLGR)	714

Neighbour-joining trees (Saitou and Nei 1987) based on inferred amino-acid sequences were calculated using Protdist and Neighbor programs in the Phylip v. 3.5c package (Felsenstein 1993). The trees were subsequently drawn in Treeview v.1.5.2 (Page 1996). Again, all PCR cloning, sequencing, isolation and sequencing of kinase domains, variability assays, segregation assays and analyses reported in this Chapter were done by Philip Awadalla unless otherwise noted. Restriction digest screens of cloned PCR products were performed jointly by Philip Awadalla and M. Schierup.

3.3 Results

3.3.1 Amplification of S-domain sequences.

Several combinations of primers designed to match conserved regions of the *Brassica* S-gene family (Ats1F, Ats1R, SI1, SI2a, SLGF, SLGR, PSA and PSB, see Table 3.1) were used to perform initial screening for S-domain genes. These primers amplified PCR products that were of the predicted size based on the *Brassica* sequences, from genomic DNA of *A. lyrata* individuals from Scotland and North Carolina. To detect different types of sequences, cloned products were cut with 6-cutter restriction enzymes *EcoR*I, *Hind*III and *Bam*HI, as described in the Materials and Methods section. Five different sequence types of the expected size were identified using this approach which were then chosen to be sequenced. One clone from amplifications using the primer pair Ats1F+Ats1R yielded a sequence referred to as *Aly*3. Primer pair SI1+Ats1R yielded sequence type *Aly*10. Primers *SLG*F and *SLG*R yielded four sequence types: *Aly*7, *Aly*8sk, *Aly*9 and *Aly*13-3. Blast searches of these products showed homology to *Brassica* and *A. thaliana* S-domain loci. All the different S-domain loci can be readily aligned to each other (but see

Pseudogenes, Section 3.3.9), to *Brassica* S-alleles, and to related members of the S-domain gene family in *A. thaliana*. After the sequence types were aligned to each other, it was observed that all indels fixed between loci or polymorphic within loci were multiples of 3 nucleotides. The *A. lyrata* S-domains, like those known from *Brassica*, therefore, appear to have no introns. All of the *Aly* sequence types, with the exception of *Aly*7, appear to be in reading frame.

In addition, a further forward primer was designed based on work previously done by M.H. Schierup using a plant from Iceland, which yielded sequence *Aly*13-1. This sequence type was shown by RT-PCR to be expressed in *A. lyrata* stigmas and had no detectable kinase domain (M. Schierup pers. comm.). This sequence was aligned with the five sequence types described above, and a new primer, 13F1 was designed. When cloned and sequenced, the bands of the expected size proved to be homologous to the *Brassica SLG* and *SRK* loci. The 13F1-SLGR combination of primers yielded the *Aly*13-1 type as well as the *Aly*13-3 sequence from a number of different individuals (see Appendix 1). Both of these sequences were also amplified from the the 13F1-PSB primer combination. All individuals from the US populations in North Carolina and Indiana yielded products with the 13F1-SLGR combination of primers, but some plants from the Scottish or Icelandic populations did not. These observations suggest that this gene is polymorphic in the primer regions as well. Eleven other *Aly*13 sequence variants, or *Aly*13 subtypes, were amplified using these primer combinations (Chapters 4 and 5).

3.3.2 Design of primers for particular S-domain sequence types, and evidence that they represent different loci.

The sequences obtained were aligned with published S-domain sequences from *B. oleracea* and *B. campestris* and were used to design new primers (see Table 3.1) specific for the different types of sequences found, in order to test whether these represent different loci. This yielded primer pairs (Table 3.1) that amplify bands of the expected sizes for each of the five sequences *Aly* 3, 7, 8sk, 9 and 10, from all individuals tested, representing populations Indiana, NC, Iceland, and Scotland. Appendix one shows all the individuals tested and used in the diversity analyses described in Chapters 4 and 5. Because primers were designed that could distinguish the loci from each other, primers homologous to regions that are within the various S-domains were designed. As a result, most but not all of the S-domain exon is amplified and assayed in the population diversity studies (Chapters 4 and 5). This supports the assumption that each of these different types represent distinct loci. *Aly*10 variants were shown to represent two loci, *Aly*10sk1 and *Aly*10sk2 (see *Kinase loci*). All individuals appear to share both of these loci as well. See Appendix 1.1 for the full nucleotide and amino acid alignment of the seven S-domain loci sequenced from one individual (99A7-1).

3.3.3 *Aly13*

An initial sample of five different sequences of the *Aly*13 type, amplified using primers 13F1 and SLGR were identified by digesting clones with restriction enzymes, using *AluI* and *RsaI* (see also below). These five sequences were: *Aly*13-1 from the NC plant 97F-13/5 and also from the Ind. plant 98E-17/11, *Aly*13-2 from the NC plant 97F-15/3 and also from the Ind. plant 97F-12/4, *Aly*13-3 from the Ind plant 97F-12/3, *Aly*13-4 from the Scottish plants B15(3), and *Aly*13-5 from the Scottish plants B15(3) and B21(3). Each of these PCR products were sequenced, which verified that they were S-domain sequences and that they were more similar to each other than to the other S-domain loci identified in *A. lyrata*. Nevertheless, variation among them was extremely high (Chapters 4 and 5).

3.3.4 Kinase loci

A further set of tests was done to identify which, if any, of the *A. lyrata* S-domain sequences have kinase domains. Since in *Brassica* species the S-loci and related loci are organised into haplotypes containing loci with either a kinase domain (*SRK*-like) or without (*SLG*-like), we expected the set of S-domain sequences to include some with this domain. To test for the presence of kinase sequences, reverse primers (*SRK*4R and *SRK*5R) were designed from kinase domains of the *SRK* locus of *Brassica*, and used with the forward primers for the specific loci (i.e. *Aly*3, *Aly*8...) identified in *A. lyrata* (see Table 3.1). Reverse primers which anneal to the third exon in the kinase domain amplify strongly using forward primers for sequences of *Aly*10 and *Aly*8sk. No amplification was seen with forward primers for sequences of *Aly*3, *Aly*7, or *Aly9.* These amplifications also showed that sequences of the *Aly*10 are two separate loci (see Figure 3.1 and Table 3.2), called *Aly*10sk1 and *Aly*10sk2.

Aly13 – evidence for more than one locus

A cDNA S-domain sequence isolated from stigma tissue mRNA of an *A. lyrata* plant by J. Nasrallah, has an S-domain similar to the *SLG* and *SRK* in *Brassica*, and to an *Aly*13 sequence identified by myself and M. Schierup. This variant was called *Aly*13-13 and segregates with incompatibility groups in four families (see below). This sequence also has a transmembrane domain and a kinase domain alignable to the kinase domain of *SRK* in *Brassica* (J. Nasrallah pers. comm.). In combination with the 13F1 primer, primers designed based on this kinase domain revealed that *Aly*13-2, *Aly*13-3, *Aly*13-9 and *Aly*13-13 S-domains have a kinase domain downstream. Individuals which are known to have *Aly*13-1 do not amplify, which follows earlier cDNA work by M. Scheirup. Therefore, *Aly*13-2, *Aly*13-3, *Aly*13-9 and *Aly*13-13 are likely to be *SRK*-like and *Aly*13-1 is likely to be *SLG*-like. It is possible that the remaining *Aly*13 sequences represent *SLG*-like alleles, yet variability at the kinase domain appears high enough to inhibit correct annealing during PCR amplification. It is also possible that some *Aly*13 variants could possibly be variants of alternative loci not yet described.

3.3.5 Common features found among the *A. lyrata* S-domain loci

Figure 3.1 shows the structure of the different S-domain sequence types so far identified in *A. lyrata*. In the inferred amino-acid alignment, all the putative genes share 12 cysteine residues that are present in the *Brassica SLG* and *SRK* S-domain sequences, as well as the *SLR* and *ARK* loci of *A. thaliana*. Nevertheless, these loci differ by many nucleotide substitutions.

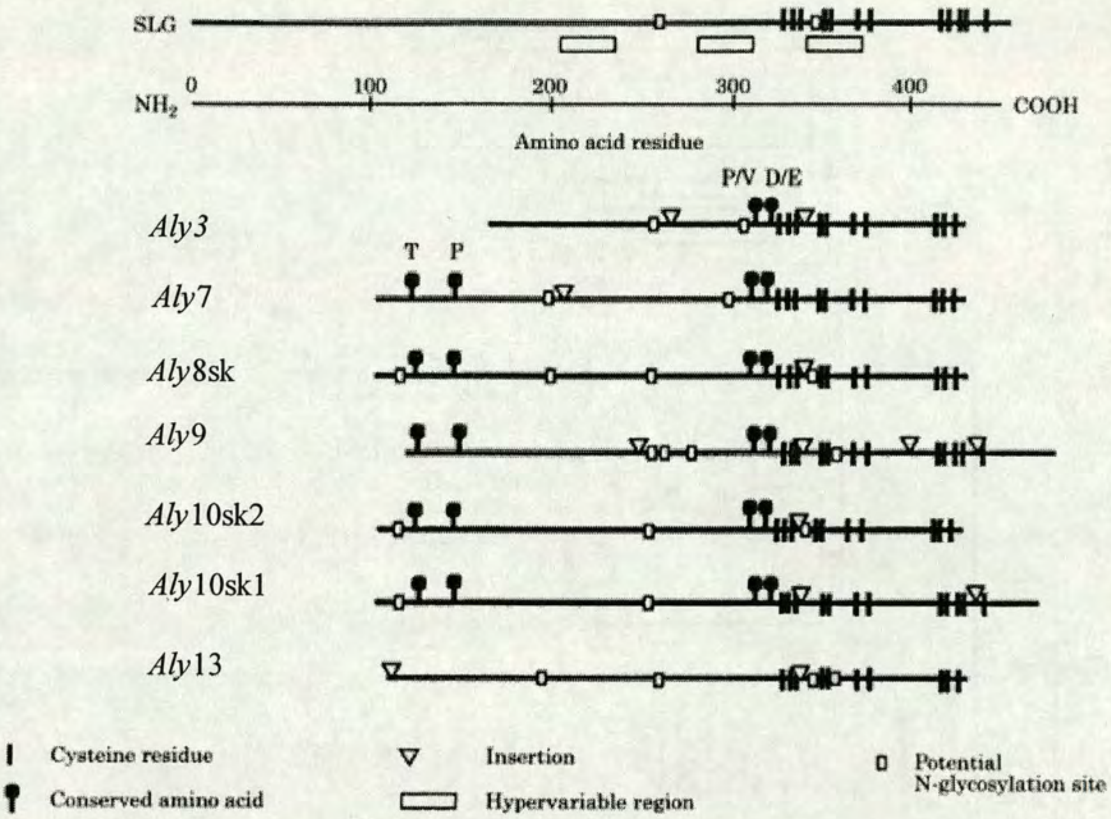


Figure 3.1 Major features of the *A. lyrata* S-domain loci identified, and differences between them.

3.3.6 Genealogy of S-domain loci among *Arabidopsis* and *Brassica* loci

The inferred amino-acid sequences obtained from all the sequence types and individuals listed in Appendix 1 were used to construct neighbour-joining phylogenetic trees (Figure 3.2). If our interpretation is correct that we have identified several different loci, the sequences should fall into evident groups, according to the different loci; this might not, however, be informative if the loci are highly polymorphic, but it should be possible to distinguish some of the less variable loci from one another. The figure clearly distinguishes *Aly*3, *Aly*7, Aly8sk, *Aly*9, *Aly*10sk1, and *Aly*10sk2 from one another. This analysis thus supports the interpretation that each represent single loci or groups of similar loci, with the

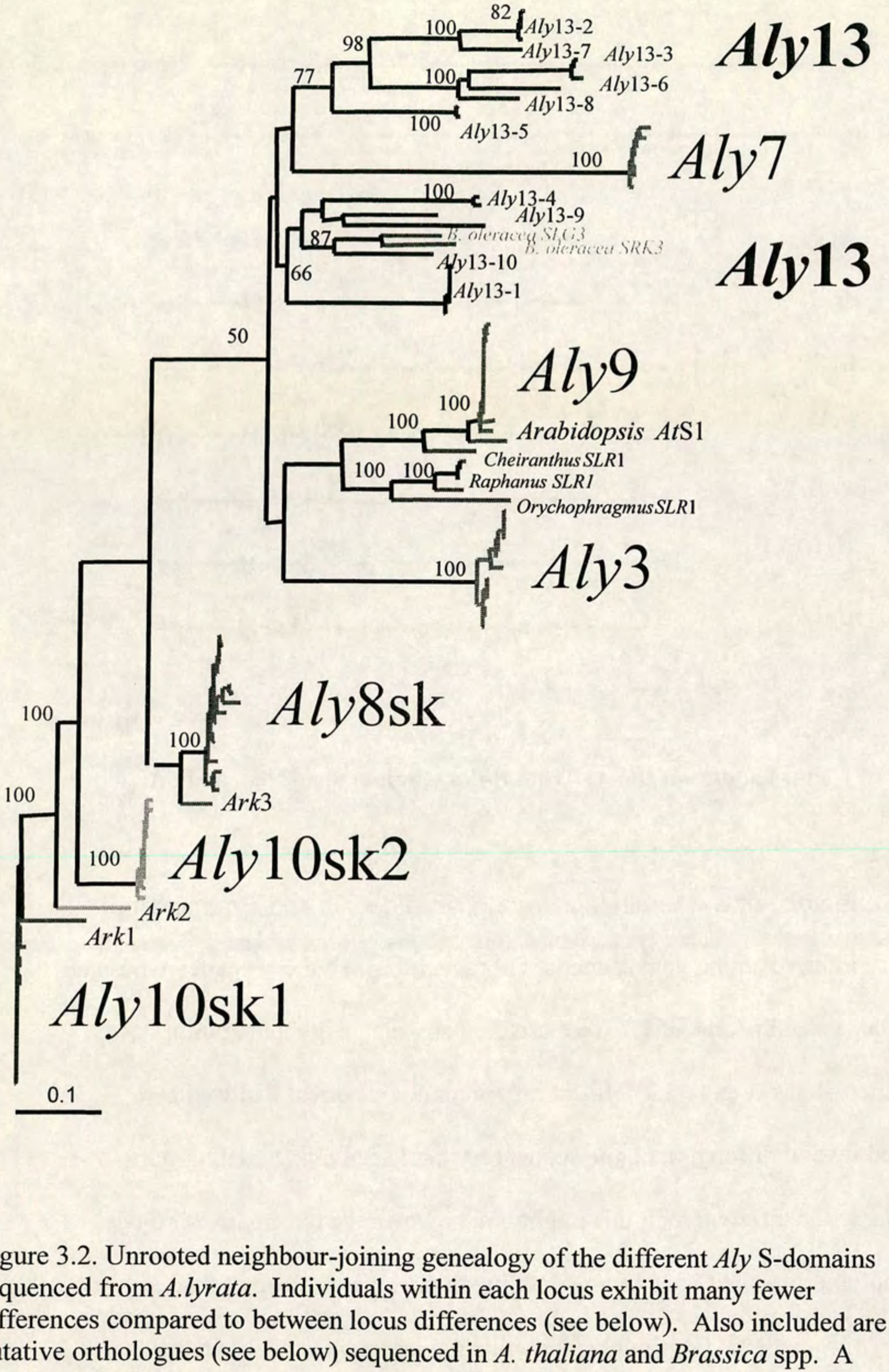


Figure 3.2. Unrooted neighbour-joining genealogy of the different *Aly* S-domains sequenced from *A.lyrata*. Individuals within each locus exhibit many fewer differences compared to between locus differences (see below). Also included are putative orthologues (see below) sequenced in *A. thaliana* and *Brassica* spp. A number of individuals have been sequenced for each S-domain 'locus' and only the branches are shown (see Chapter 4).

numbers of differences separating variants within a putative locus much smaller than that observed between any pair of loci. Note the longer external branch lengths for the *Aly*13 group of sequences which reflects the high polymorphism that separates these lineages.

The genealogy (Figure 3.2) of the S-domains also includes S-domain orthologues in *Brassica* and *A. thaliana* and reveals that sequences from *A. thaliana*, *A. lyrata*, and assorted *Brassica* sequences are 'intermixed' in the tree, indicating that much of the gene duplication in this family took place before these species diverged. Blast searches show that *Aly*3 and *Aly*7 are not orthologous or closely related to a known sequence in either *A. thaliana* or *Brassica*.

3.3.7 Comparison of sequences of the different *Aly* S-domain loci.

It is clear that the *Aly* loci are very different from one another (Table 3.2). Nucleotide divergence among S-domain loci in *A. lyrata* varies tremendously among the pairwise comparisons (Table 3.2). *Aly*8sk, *Aly*10sk1, and *Aly*10sk2 appear to be more similar to each other than to the remaining S-domain loci. The remaining pairwise comparisons exhibit very high synonymous and nonsynonymous variation (perhaps at saturation). *Aly*7 appears the most divergent relative to the remaining loci (see discussion of *Aly*7 below). *Aly*13 were excluded from these analyses as divergence calculations involving these sequences often equals estimates of polymorphism, which is extremely high.

Table 3.2: Jukes-Cantor corrected silent (upper right triangle) and replacement (lower left triangle) nucleotide divergence among six members of the S-domain gene family in *A. lyrata*.

S-domains	*Aly*3	*Aly*7-	*Aly*8sk	*Aly*9	*Aly*10sk2	*Aly*10sk1
*Aly*3		0.60621 (71.2)	0.61078 (98.4)	0.43278 (111.2)	0.60933 (100.3)	0.59934 (58.1)
*Aly*7-	0.40647 (243.8)		0.75568 (64.6)	0.73690 (84.9)	0.66914 (89.6)	0.67909 (27.8)
*Aly*8sk	0.25685 (360.6)	0.35033 (217.4)		0.52898 (108.1)	0.33368 (101.2)	0.27744 (63.3)
*Aly*9	0.30777 (416.9)	0.35456 (290.1)	0.30790 (380.8)		0.50696 (119.3)	0.46716 (63.7)
*Aly*10sk2	0.31484 (370.7)	0.39991 (306.4)	0.17884 (354.8)	0.33089 (429.7)		0.17101 (77.8)
*Aly*10sk1	0.33551 (214.9)	0.42679 (89.2)	0.17978 (227.7)	0.36033 (224.3)	0.09267 (264.2)	

In parentheses are the number of synonymous or nonsynonymous sites compared. (see Table 3.1). Ambiguous positions and codons with multiple changes were excluded from the analysis.

3.3.8 Putative orthologues of the *Aly* loci in *Arabidopsis thaliana*

*Aly*9 appears to be orthologous, or at least most closely related, to *ATS*1 in *A. thaliana* (0.0725 ± 0.0087 Jukes-Cantor pairwise divergence) and *SLR*1 in *Brassica* (0.2469 ± 0.0144) compared to other loci sequenced in *Brassica*. The most closely related *A. thaliana* orthologous sequences to *Aly*10sk1, *Aly*10sk2, and *Aly*8sk are *ARK*1 (0.1016 ± 0.0101, Jukes-Cantor divergence), *ARK*2 (0.0868 ± 0.0104) and *ARK*3 (0.1059 ± 0.0110), respectively. Blast searches reveal that the next most similar locus currently available in Genbank to any of the *Aly* S-domain sequences are at least 3 times less similar. The divergence levels between the *A. lyrata* and *A. thaliana* sequences is slightly lower than that observed at the *Adh* locus (Savolainen et al. 2000) and data published for other loci (Innan et al. 1996, Kawabe et al. 1997, Purugganan and Suddith 1998, 1999, Stahl et al. 1999).

3.3.9 *Aly7* sequences

As mentioned previously, all the *Aly* loci identified appear to be in reading frame, the exception being the *Aly*7 set of sequences. Some *Aly* 7 haplotypes have a single base-pair insertion at position 712 within a region of four TA repeats (Aly7+); (Figure 3.3). This reading frame shift creates a stop codon downstream of this sequence difference from the non-insertion sequence. Nucleotide similarity between the two haplotypes is very high, yet linkage disequilibrium between the two haplotypes is almost complete for the variants described above and indels that separate these haplotypes (see Chapter 4). Primers specific to either haplotype do not amplify in all individuals, suggesting that this is either a single locus with a null allele segregating in the population, or represents two loci, where one copy may be a pseudogene, and not all individuals have both copies. Yet, segregation analyses of parents where one parent exhibits both copies suggests that these haplotypes may be segregating as a single locus. The genotype ratios do not deviate from expected Mendelian ratios (Table 3.3). The observation of two kinds of differences betwen *Aly*7- and *Aly*7+ sequences supports the 2-gene view, but restriction enzyme tests and primers specific to each haplotype appeared to indicate that some individuals had both haplotypes and some had only one.

Figure 3.3 (next page) Alignments of *Aly*7 loci exhibiting the indel at position 400 (712 in the complete sequence alignment) which knocks the sequence out of reading frame. Aly7 alignment positions are shown rather than aligned to the remaining S-domain loci because the insertion causes a change in the number of positions in the entire alignment. The start of the sequence corresponds to position 332 and ends at positions 1247 in sequence alignment that includes all the *Aly* loci. Note the indel at position 534-542 which is almost in complete linkage with the indel at position 400.

*Aly*7 Locus

							1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4	4	4
		Position	5	5	6	8	1	5	7	7	7	9	9	0	0	6	8	8	8	9	9	1	2	2	4	4	5	6	7	7	7	9	9	0	0	1
			2	5	6	9	9	2	4	6	9	1	2	6	8	7	1	4	5	1	9	3	2	8	5	9	5	4	3	4	8	6	9	0	4	0
Haplo.	Pop.	Individual																																		
Aly7+	Ind	97F6-1_c5+	T	T	G	T	T	T	A	T	C	A	A	A	G	A	G	G	A	G	T	C	G	G	T	T	T	C	A	C	A	G	T	A	A	C
Aly7+	NC	97F13-5_c2+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.
Aly7+	NC	97F15-3_c7+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.
Aly7+	NC	99A15-1_c21+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.
Aly7+	Scot	A9_4_c5+	.	C	A	G	.	.	G	.	.	.	.	.	.	T	.	A	.	A	.	.	.	.	.	.	C	.	.	A	G	.	.	A	.	G
Aly7+	Scot	B17_2_c19+	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	.	.	T	.	.	.	.	.	.	.	.	A	.	A	.	.
Aly7+	Scot	b20_3_c6+	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.
Aly7+	Scot	b26xa9?_c9+	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	A	.	.
Aly7-	NC	97F13-5_c1-	.	C	A	G	.	.	.	.	.	.	.	.	.	T	.	A	.	A	.	.	.	.	.	C	C	.	.	A	?	.	G	-	C	G
Aly7-	NC	97F15-3c26-	C	C	A	G	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	A	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	NC	99A1-1_c1-	.	C	A	G	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	NC	99A1-1_c2-	C	C	A	G	.	.	.	.	.	.	G	.	.	G	.	A	.	.	.	.	.	.	C	.	C	.	.	A	G	.	G	-	C	G
Aly7-	NC	99A2-1_c8-	.	C	A	G	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	.	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	NC	99A2-1_c9-	.	C	A	G	.	.	.	.	.	.	.	.	.	G	A	A	.	.	.	.	.	.	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	Ind	99A7-1_c18-	C	C	A	G	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	A	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	Ind	99A8-1_c17-	C	C	A	G	.	.	.	.	.	.	.	.	.	G	.	A	.	.	.	.	.	A	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	Scot	99A17-1_c27-	.	C	A	G	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	Scot	A9(4)_c2-	.	C	A	G	.	.	.	.	.	.	.	.	.	.	.	A	.	A	.	T	.	.	.	.	C	.	.	A	G	.	G	-	.	G
Aly7-	Scot	B20(3)_c2-	.	C	A	G	.	.	.	.	.	.	.	.	.	.	.	A	G	A	.	T	.	.	.	.	C	.	.	A	G	.	G	-	.	G
Aly7-	Scot	B21(1)_c21-	.	C	A	G	.	.	.	.	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	-	.	.
Aly7-	Scot	b26xa9?_c5-	.	C	A	G	G	.	.	.	.	.	.	.	.	T	.	A	.	A	.	.	.	.	.	.	C	.	.	A	G	.	G	-	.	G
Aly7-	Scot	B17(2)-direct-	-	-	-	G	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	-	.	.
Aly7-	Ice	43/1_c50-	.	C	A	G	.	.	.	C	.	.	.	.	.	.	.	A	.	A	.	T	.	.	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	Ice	98I32-7_c37-	.	C	A	G	.	C	.	.	.	.	.	.	.	.	.	A	.	A	.	T	.	.	.	.	C	.	.	A	G	.	G	-	C	G
Aly7-	Ice	98I32-7_c31-	.	C	A	G	.	.	.	.	.	.	.	.	.	.	.	A	.	.	G	.	.	.	.	.	C	A	.	.	.	.	G	-	.	.
Aly7-	Ice	98I33-3_c40-	.	C	A	G	.	.	.	.	.	.	.	G	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	-	.	.
Aly7-	Ice	98I33-3_c39-	.	C	A	G	.	.	.	.	.	.	.	.	.	.	.	A	.	.	G	.	.	.	.	.	C	A	.	.	.	.	G	-	.	.
																																		* insertion site		

			4	4	5	5	5	5	5	5	5	5	6	6	6	6	6	6	6	6	7	7	8	8	8	8
			8	9	1	2	2	3	4	5	9	9	0	0	1	2	3	3	6	8	8	8	0	0	1	1
			0	2	2	8	9	4	6	2	2	8	1	6	1	9	5	7	2	3	1	2	4	8	4	6
Haplo.	Pop.	Individual																								
Aly7+	Ind	97F6-1_c5+	T	G	C	C	T	ins	G	T	T	T	A	C	G	G	C	T	G	T	C	T	ins	A	G	C
Aly7+	NC	97F13-5_c2+	.	.	.	?	.	ins	A	.	.	.	.	.	A	.	.	.	.	.	.	.	ins	.	.	.
Aly7+	NC	97F15-3_c7+	.	.	.	T	C	ins	A	.	.	.	.	.	A	.	.	.	.	.	.	.	ins	.	.	.
Aly7+	NC	99A15-1_c21+	.	.	.	.	.	ins	A	.	.	.	.	.	A	.	.	.	.	.	.	.	ins	.	.	.
Aly7+	Scot	A9_4_c5+	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	-	T	.
Aly7+	Scot	B17_2_c19+	.	.	.	.	.	ins	A	.	.	.	.	.	.	.	.	.	.	.	.	.	ins	.	.	.
Aly7+	Scot	b20_3_c6+	.	.	.	?	.	ins	A	.	.	.	.	.	.	.	.	.	.	.	.	.	ins	.	.	.
Aly7+	Scot	b26xa9?_c9+	.	.	.	.	.	ins	A	.	.	.	.	.	.	.	.	.	.	.	.	.	ins	.	.	T
Aly7-	NC	97F13-5_c1-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	NC	97F15-3c26-	C	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	NC	99A1-1_c1-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	NC	99A1-1_c2-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	NC	99A2-1_c8-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	NC	99A2-1_c9-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	Ind	99A7-1_c18-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	Ind	99A8-1_c17-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	Scot	99A17-1_c27-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	.	T	.
Aly7-	Scot	A9(4)_c2-	.	.	.	?	.	del	A	.	C	G	G	G	.	.	.	.	.	.	.	.	ins	.	.	.
Aly7-	Scot	B20(3)_c2-	.	A	.	.	.	del	A	.	C	G	G	G	.	.	.	.	.	.	.	.	ins	.	.	.
Aly7-	Scot	B21(1)_c21-	.	.	.	.	.	del	A	.	.	.	.	.	.	.	.	.	.	.	.	.	ins	.	.	.
Aly7-	Scot	b26xa9?_c5-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	.	.	C	.	G	C	del	-	T	.
Aly7-	Scot	B17(2)-direct-	.	.	.	.	.	del	A	.	.	.	.	.	.	A	.	.	.	.	G	.	-	-	-	-
Aly7-	Ice	43/1_c50-	.	.	T	T	.	del	A	.	C	G	G	G	.	.	T	C	.	C	G	.	ins	G	.	.
Aly7-	Ice	98I32-7_c37-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	T	C	.	C	G	.	ins	.	.	.
Aly7-	Ice	98I32-7_c31-	.	.	.	.	.	del	A	A	C	G	G	G	.	.	T	C	.	C	G	.	ins	.	.	.
Aly7-	Ice	98I33-3_c40-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	T	C	.	C	G	.	ins	.	.	.
Aly7-	Ice	98I33-3_c39-	.	.	.	.	.	del	A	.	C	G	G	G	.	.	T	C	.	C	G	.	ins	.	.	.
								9 bp indel															9 bp indel			

Table 3.3 Genotypes of *Aly*7 variants in the 99E1 family (parents were 98E17-4 and 98E17-6 which was heterozygous for *Aly*7- and *Aly*7+ haplotypes and homozygous for the *Aly*7- haplotype respectively). Segregation does not deviate from Mendelian expectations, *P*=1. + and – indicate amplifications.

99E1 Progeny	*Aly*7-	*Aly*7+
99E1-1	+	-
99E1-2	+	-
99E1-3	+	-
99E1-4	+	+
99E1-5	+	-
99E1-6	+	+
99E1-7	+	+
99E1-8	+	+
99E1-9	+	-
99E1-10	+	-
99E1-11	+	+

3.3.10 Pollination studies – determining SI groups

To determine incompatibility groups in *A. lyrata*, full-sib pollinations were performed within four families of plants raised from crosses between individuals from Michigan, North Carolina, Scotland and Iceland. Plants to be pollinated were covered with net curtain fabric to exclude pollinators, and non-emasculated flowers were hand pollinated by rubbing dehisced anthers over their stigma. Fruit set was scored about 7-10 days after pollination. Plants within families were divided into SI groups based on the incompatibility of hand pollinations without prior knowledge of genotypes at the candidate S-domain variants in order to avoid biasing results. Table 3.4, which describes incompatibility data for four families, shows the SI grouping (roman numerals), the number of individuals per SI group (in parentheses), and the number of pollinations that produced fruits out of the total number performed. Incompatible pollinations are indicated in bold. Partial incompatibility of pollinations within

Table 3.4a. SI Segregation in Family 98E15.

		RECIPIENT				
	SI Group (No. Plants)	I (3)	II (4)	III (2)	IV (5)	*Aly*13* allele
DONOR	I	**0/5**	**1/20**	11/12	28/29	13-1/13-23
	II	**1/22**	**0/28**	33/33	33/36	13-13/13-23
	III	15/15	38/38	**0/19**	31/31	13-1/ 13-3
	IV	18/20	33/34	31/39	**7/27**	13-3/13-13
	Parents	97F 13-5 (NC)				13-1/13-13
		97F15-3 (NC)				13-3/13-23

Table 3.4b. SI Segregation in Family 98E17

		RECIPIENT		
	SI groups (No. of plants)	I (10)	II (10)	*Aly*13* allele
DONOR	I	**6/71**	65/69	13-1/13-13
	II	61/68	**22/57**	13-1/?
	Parents	? ?		

Table 3.4c. SI Segregation in Family 98G23.

		RECIPIENT				
	(No. Plants)	I (5)	II (4)	III (2)	IV (3)	*Aly*13* allele
DONOR	I	**5/73**	72/77	39/43	**0/46**	13-1/13-20
	II	64/65	**3/30**	29/34	**0/39**	13-3/13-23
	III	33/35	24/24	**2/14**	19/23	13-1/ 13-3
	IV	**5/51**	**10/74**	21/23	**3/18**	13-20/13-23
	Parents	97F 15-2 (NC)				13-1/13-23
		97F 8-1 (NC)				13-3/13-20

Table 3.4d SI Segregation in Family 98G24.

		RECIPIENT				
	SI Group (No. plants)	I (2)	II (1)	III (5)	IV (5)	*Aly*13 alleles
DONOR	**I**	**1/9**	**0/6**	16/19	23/26	13-1/13-19
	II	**3/6**	--	6/6	7/8	13-13/13-19
	III	16/17	6/6	**3/65**	32/37	13-1/ 13-3
	IV	22/22	6/8	35/35	**6/73**	13-3/13-13
	Parents	97F 13-5 (NC)				13-1/13-13
		97F 12-3 (Mich)				13-3/13-19

Table 3.4. The result of reciprocal pollinations and segregation analysis of SI in families 98E15, 98E17, 98G23, and 98G24. The bold indicates reciprocal incompatible crosses. Compatibility of all plants with all others except those of the same group suggests a dominance hierarchy, the same in both pollen and pistils. The sequence variant for *Aly13* segregating in each incompatibility group is shown on the far right column with the genotypes corresponding to the incompatibility groups in the far left column. The number of individuals within each incompatibility group are shown in parentheses.

groups was occasionally suggested by the production of "small" fruits (which contained zero or only a few seeds). The crossing data does not include self-pollinations of individual plants, although repeated selfings indicated that self-compatible individuals were rare in our populations, except under stressful conditions. We also very rarely observed fruits on flowers that were not pollinated.

3.3.11 Tests for linkage between alleles at the self-incompatibility locus and the *Aly* loci

To further classify all the *Aly* S-domain loci (*Aly*3, *Aly*8sk, *Aly*10sk1, *Aly*10sk2 and *Aly*13) into those that are linked to incompatibility and those that are not closely linked, we used one of the full sib families (98E15) in which 14 plants could be classified by pollinations. The parents of this family were plants 97F13-5 and 97F15-3. All of the plants in this sibship initially fell into three different incompatibility groups (Table 3.4a), suggesting that both parents were heterozygotes at the S-locus (incompatibility groups I and II were derived once genotypes were known, see below). The genotypes of these plants and the two parents were also examined by PCR amplification with primers specific for the six sequence types *Aly*3, *Aly*8sk, *Aly*10sk1, *Aly*10sk2 and *Aly*13, and digestion with restriction enzymes. This simple test gave clearly scorable variants for these putative loci (Figure 3.4). The 98E-15 family was also classified for an intron length difference polymorphism at the *Aly*10sk1 locus (B. Mable pers. comm.). The parent plant 97F15-3 is heterozygous for this 250 deletion in the intron (allele B), while the other parent, 97F13-5, is A/A (homozygous for no deletion). The results confirm that the two variants segregate as expected if they are allelic (there were 5 A/B and 10 A presumed to be A/A homozygotes, χ^2=1.7, P>0.05). Two plants in the first incompatibility group were heterozygotes, while the other four were homozygotes, so this locus is probably not linked to SI.

Sequences of *Aly*3 and *Aly*9 were amplified from the parents of family 98E15 through direct sequencing. This allowed us to identify sites that may be 'heterozygous' based on the sequence chromatograms. *Aly*3 appeared to be heterozygous in individual 97F13-5 for a number of sites whereas the parents appeared to be homozygous for *Aly*9. Restriction enzyme cutting sites specific to the *Aly*3 variants (AciI and BpuAI) were identified and used to test for segregation. Restriction profiles indicated that variants at *Aly*3 are not linked to SI although this sample size is quite small (χ^2=6.0, P>0.05; Figure 3.4). As *Aly*9 was monomorphic and the family is clearly polymorphic for SI, it is likely not to be the SI locus. Likewise, cloned *Aly*8sk and *Aly*10sk2 products did not reveal polymorphisms for these two parents.

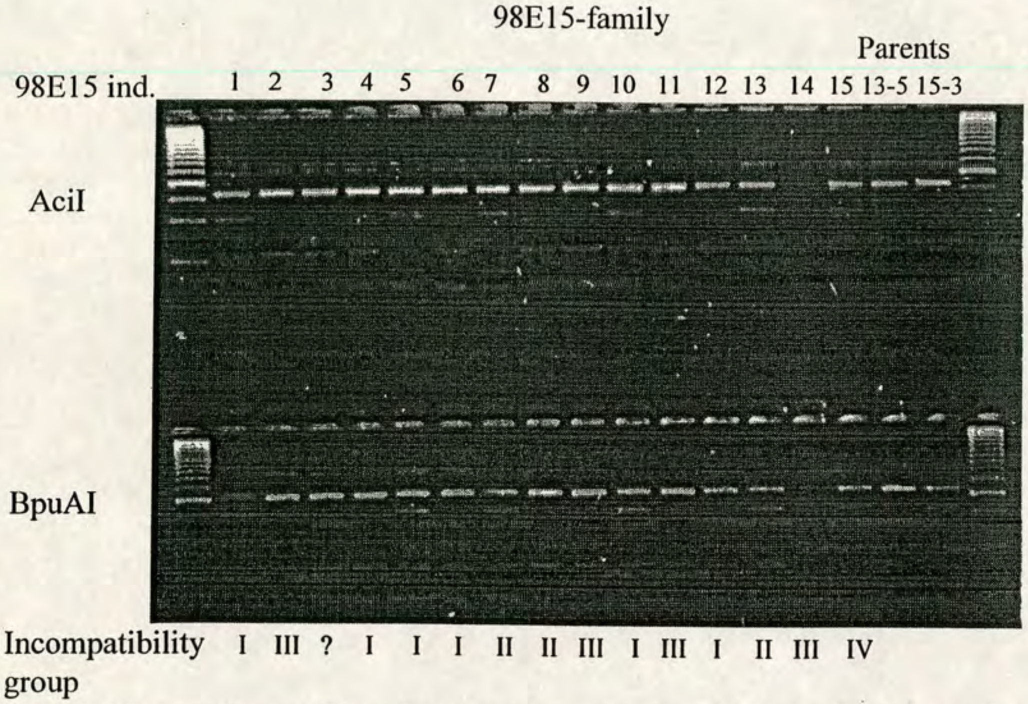


Figure 3.4 Restriction profiles of the *Aly*3 locus for the 98E15 family shows no linkage with incompatibility groups in this family. Incompatibility groups are as listed in Table 3.3. ? refers to a plant where the incompatibility group is unknown.

Initial reciprocal crosses among individuals within this family indicated that the 98E15 progeny fell into three incompatibility groups, consistent with both parents being heterozygous at the S-locus but with some dominance acting. Subsequent screening of alleles at the *Aly* 13 putative locus identified four groups that showed strong segregation with SI in this family. The genotypes of the parents for this cross were *Aly*13-1/*Aly*13-13 (97F13/5) and *Aly*13-3/*Aly*13-23 (97F15/3) and the progeny segregated into four genotypes: *Aly*13-1/*Aly*13-23 (n=3), *Aly*13-13/*Aly*13-23 (n=4), *Aly*13-1/*Aly*13-3 (n=3), and *Aly*13-3/*Aly*13-13 (n=5). Table 3.4a shows the segregation of these classes with SI. The first two genotypes corresponded to a single SI phenotype but the table shows four SI groups (rather than the three identified from pollination data) to indicate how the individual genotypes behave. The segregation of *Aly*13 alleles with SI groups suggests that this locus is tightly linked to SI and allows us to draw some conclusions about relative dominance of at least some of the identified alleles. *Aly*13-23 is found with *Aly*13-1 in SI group I and with *Aly*13-3 in group II but these two groups are incompatible with each other, suggesting that *Aly*13-23 is dominant to both *Aly*13-1 and *Aly*13-13, in both pollen and stigma. Dominance interactions in the remaining SI groups cannot be determined very confidently from this family alone.

*Aly*13 variants found in this family thus appear to be linked to the self-incompatibility locus (though with such a small family, there is no implication that linkage is necessarily close). The finding of three different *Aly*13 alleles in this family supports the view, in conjunction with evidence for its extreme polymorphism (Chapter 4 and 5), and from the linkage results, that these variants represent both the different S alleles. Another sequence, *Aly*13-2, was initially amplified from 97F15-3 using primers 13F1 and SLGR, but it does not segregate with incompatibility groups in other families (M. Schierup pers. comm.) and is likely a different unlinked locus.

We tested three more families for linkage of SI to *Aly*13 variants using similar approaches. These results are shown in Table 3.4b-d. Because the SI locus is highly polymorphic, we may not amplify all alleles expected in a family with a particular set of primers, and indeed, that is the case for the segregation analysis in one family (98E17 – see below).

For all of the families described (see below), the division of phenotypic SI groups was consistent with both parents being heterozygous at the S locus. However, we did find that this was not the case for one family, derived from a cross between two individuals from Michigan (98E17). The 20 progeny fell into two SI groups only, each of ten individuals, within which pollinations were mostly incompatible, but between which they were compatible (Table 3.4b). Subsequent characterization of the *Aly*13 alleles detected the presence of only two alleles (*Aly*13-1 and *Aly*13-13). In this family the *Aly*13-1 variant was present in all progeny, suggesting that one parent must be a *Aly*13-1 homozygote. The *Aly*13-13 allele, on the other hand, was only present in one of the SI phenotype classes, suggesting that *Aly*13-13 was clearly linked to SI. In the other SI class, we assume that an undetected allele must have been present in the heterozygous parent (denoted as ? in Table 3.4b). The parental genotypes could not be independently tested, however, as the plants were no longer alive when these results were obtained. Our inability to detect the hypothesized third allele may have been because it produces a restriction profile indistinguishable from that of allele *Aly*13-1 but confers a different incompatibility; alternatively, there may be an allele that is not amplified by the PCR primers we have used.

Table 3.3 Genotypes of *Aly*7 variants in the 99E1 family (parents were 98E17-4 and 98E17-6 which was heterozygous for *Aly*7- and *Aly*7+ haplotypes and homozygous for the *Aly*7- haplotype respectively). Segregation does not deviate from Mendelian expectations, *P*=1. + and – indicate amplifications.

99E1 Progeny	*Aly*7-	*Aly*7+
99E1-1	+	-
99E1-2	+	-
99E1-3	+	-
99E1-4	+	+
99E1-5	+	-
99E1-6	+	+
99E1-7	+	+
99E1-8	+	+
99E1-9	+	-
99E1-10	+	-
99E1-11	+	+

3.3.10 Pollination studies – determining SI groups

To determine incompatibility groups in *A. lyrata*, full-sib pollinations were performed within four families of plants raised from crosses between individuals from Michigan, North Carolina, Scotland and Iceland. Plants to be pollinated were covered with net curtain fabric to exclude pollinators, and non-emasculated flowers were hand pollinated by rubbing dehisced anthers over their stigma. Fruit set was scored about 7-10 days after pollination. Plants within families were divided into SI groups based on the incompatibility of hand pollinations without prior knowledge of genotypes at the candidate S-domain variants in order to avoid biasing results. Table 3.4, which describes incompatibility data for four families, shows the SI grouping (roman numerals), the number of individuals per SI group (in parentheses), and the number of pollinations that produced fruits out of the total number performed. Incompatible pollinations are indicated in bold. Partial incompatibility of pollinations within

Table 3.4a. SI Segregation in Family 98E15.

		RECIPIENT				
	SI Group (No. Plants)	I (3)	II (4)	III (2)	IV (5)	*Aly*13* allele
DONOR	I	**0/5**	**1/20**	11/12	28/29	13-1/13-23
	II	**1/22**	**0/28**	33/33	33'36	13-13/13-23
	III	15/15	38/38	**0/19**	31/31	13-1/ 13-3
	IV	18/20	33/34	31/39	**7/27**	13-3/13-13
	Parents	97F 13-5 (NC)				13-1/13-13
		97F15-3 (NC)				13-3/13-23

Table 3.4b. SI Segregation in Family 98E17

		RECIPIENT		
	SI groups (No. of plants)	I (10)	II (10)	*Aly*13* allele
DONOR	I	**6/71**	65/69	13-1/13-13
	II	61/68	**22/57**	13-1/?
	Parents	? ?		

Table 3.4c. SI Segregation in Family 98G23.

		RECIPIENT				
	(No. Plants)	I (5)	II (4)	III (2)	IV (3)	*Aly*13* allele
DONOR	I	**5/73**	72/77	39/43	**0/46**	13-1/13-20
	II	64/65	**3/30**	29/34	**0/39**	13-3/13-23
	III	33/35	24/24	**2/14**	19/23	13-1/ 13-3
	IV	**5/51**	**10/74**	21/23	**3/18**	13-20/13-23
	Parents	97F 15-2 (NC)				13-1/13-23
		97F 8-1 (NC)				13-3/13-20

Table 3.4d SI Segregation in Family 98G24.

		RECIPIENT				
	SI Group (No. plants)	I (2)	II (1)	III (5)	IV (5)	*Aly*13 alleles
DONOR	**I**	**1/9**	**0/6**	16/19	23/26	13-1/13-19
	II	**3/6**	--	6/6	7/8	13-13/13-19
	III	16/17	6/6	**3/65**	32/37	13-1/ 13-3
	IV	22/22	6/8	35/35	**6/73**	13-3/13-13
	Parents	97F 13-5 (NC)				13-1/13-13
		97F 12-3 (Mich)				13-3/13-19

Table 3.4. The result of reciprocal pollinations and segregation analysis of SI in families 98E15, 98E17, 98G23, and 98G24. The bold indicates reciprocal incompatible crosses. Compatibility of all plants with all others except those of the same group suggests a dominance hierarchy, the same in both pollen and pistils. The sequence variant for *Aly13* segregating in each incompatibility group is shown on the far right column with the genotypes corresponding to the incompatibility groups in the far left column. The number of individuals within each incompatibility group are shown in parentheses.

groups was occasionally suggested by the production of "small" fruits (which contained zero or only a few seeds). The crossing data does not include self-pollinations of individual plants, although repeated selfings indicated that self-compatible individuals were rare in our populations, except under stressful conditions. We also very rarely observed fruits on flowers that were not pollinated.

3.3.11 Tests for linkage between alleles at the self-incompatibility locus and the *Aly* loci

To further classify all the *Aly* S-domain loci (*Aly*3, *Aly*8sk, *Aly*10sk1, *Aly*10sk2 and *Aly*13) into those that are linked to incompatibility and those that are not closely linked, we used one of the full sib families (98E15) in which 14 plants could be classified by pollinations. The parents of this family were plants 97F13-5 and 97F15-3. All of the plants in this sibship initially fell into three different incompatibility groups (Table 3.4a), suggesting that both parents were heterozygotes at the S-locus (incompatibility groups I and II were derived once genotypes were known, see below). The genotypes of these plants and the two parents were also examined by PCR amplification with primers specific for the six sequence types *Aly*3, *Aly*8sk, *Aly*10sk1, *Aly*10sk2 and *Aly*13, and digestion with restriction enzymes. This simple test gave clearly scorable variants for these putative loci (Figure 3.4). The 98E-15 family was also classified for an intron length difference polymorphism at the *Aly*10sk1 locus (B. Mable pers. comm.). The parent plant 97F15-3 is heterozygous for this 250 deletion in the intron (allele B), while the other parent, 97F13-5, is A/A (homozygous for no deletion). The results confirm that the two variants segregate as expected if they are allelic (there were 5 A/B and 10 A presumed to be A/A homozygotes, χ^2=1.7, P>0.05). Two plants in the first incompatibility group were heterozygotes, while the other four were homozygotes, so this locus is probably not linked to SI.

Sequences of *Aly*3 and *Aly*9 were amplified from the parents of family 98E15 through direct sequencing. This allowed us to identify sites that may be 'heterozygous' based on the sequence chromatograms. *Aly*3 appeared to be heterozygous in individual 97F13-5 for a number of sites whereas the parents appeared to be homozygous for *Aly*9. Restriction enzyme cutting sites specific to the *Aly*3 variants (AciI and BpuAI) were identified and used to test for segregation. Restriction profiles indicated that variants at *Aly*3 are not linked to SI although this sample size is quite small (χ^2=6.0, P>0.05; Figure 3.4). As *Aly*9 was monomorphic and the family is clearly polymorphic for SI, it is likely not to be the SI locus. Likewise, cloned *Aly*8sk and *Aly*10sk2 products did not reveal polymorphisms for these two parents.

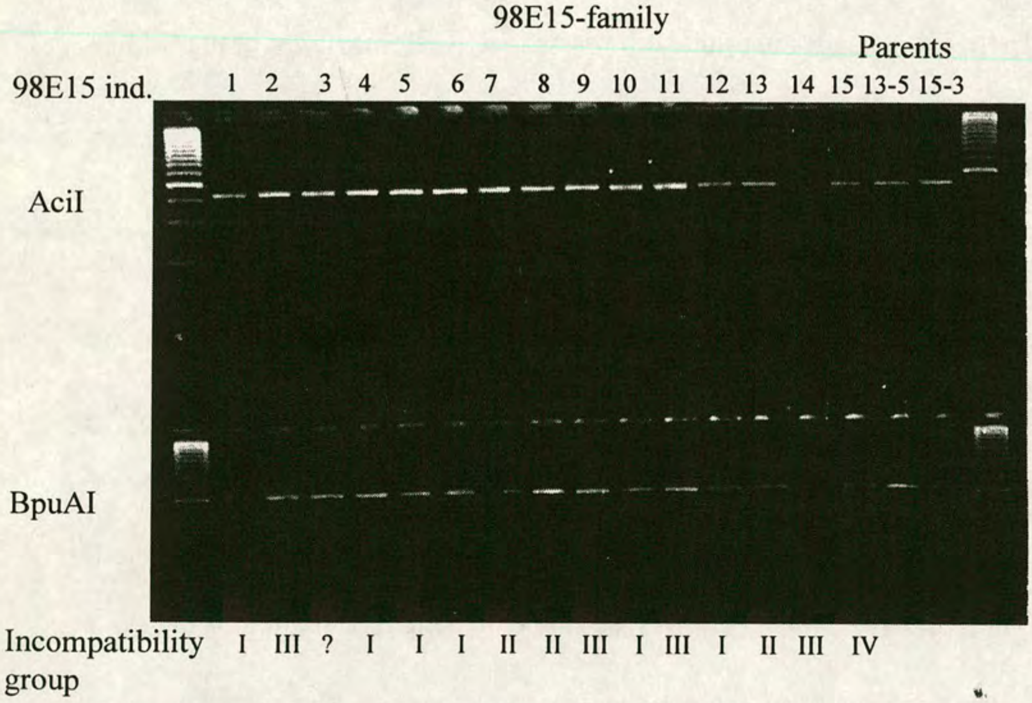


Figure 3.4 Restriction profiles of the *Aly*3 locus for the 98E15 family shows no linkage with incompatibility groups in this family. Incompatibility groups are as listed in Table 3.3. ? refers to a plant where the incompatibility group is unknown.

Initial reciprocal crosses among individuals within this family indicated that the 98E15 progeny fell into three incompatibility groups, consistent with both parents being heterozygous at the S-locus but with some dominance acting. Subsequent screening of alleles at the *Aly* 13 putative locus identified four groups that showed strong segregation with SI in this family. The genotypes of the parents for this cross were *Aly*13-1/*Aly*13-13 (97F13/5) and *Aly*13-3/*Aly*13-23 (97F15/3) and the progeny segregated into four genotypes: *Aly*13-1/*Aly*13-23 (n=3), *Aly*13-13/*Aly*13-23 (n=4), *Aly*13-1/*Aly*13-3 (n=3), and *Aly*13-3/*Aly*13-13 (n=5). Table 3.4a shows the segregation of these classes with SI. The first two genotypes corresponded to a single SI phenotype but the table shows four SI groups (rather than the three identified from pollination data) to indicate how the individual genotypes behave. The segregation of *Aly*13 alleles with SI groups suggests that this locus is tightly linked to SI and allows us to draw some conclusions about relative dominance of at least some of the identified alleles. *Aly*13-23 is found with *Aly*13-1 in SI group I and with *Aly*13-3 in group II but these two groups are incompatible with each other, suggesting that *Aly*13-23 is dominant to both *Aly*13-1 and *Aly*13-13, in both pollen and stigma. Dominance interactions in the remaining SI groups cannot be determined very confidently from this family alone.

*Aly*13 variants found in this family thus appear to be linked to the self-incompatibility locus (though with such a small family, there is no implication that linkage is necessarily close). The finding of three different *Aly*13 alleles in this family supports the view, in conjunction with evidence for its extreme polymorphism (Chapter 4 and 5), and from the linkage results, that these variants represent both the different S alleles. Another sequence, *Aly*13-2, was initially amplified from 97F15-3 using primers 13F1 and SLGR, but it does not segregate with incompatibility groups in other families (M. Schierup pers. comm.) and is likely a different unlinked locus.

We tested three more families for linkage of SI to *Aly*13 variants using similar approaches. These results are shown in Table 3.4b-d. Because the SI locus is highly polymorphic, we may not amplify all alleles expected in a family with a particular set of primers, and indeed, that is the case for the segregation analysis in one family (98E17 – see below).

For all of the families described (see below), the division of phenotypic SI groups was consistent with both parents being heterozygous at the S locus. However, we did find that this was not the case for one family, derived from a cross between two individuals from Michigan (98E17). The 20 progeny fell into two SI groups only, each of ten individuals, within which pollinations were mostly incompatible, but between which they were compatible (Table 3.4b). Subsequent characterization of the *Aly*13 alleles detected the presence of only two alleles (*Aly*13-1 and *Aly*13-13). In this family the *Aly*13-1 variant was present in all progeny, suggesting that one parent must be a *Aly*13-1 homozygote. The *Aly*13-13 allele, on the other hand, was only present in one of the SI phenotype classes, suggesting that *Aly*13-13 was clearly linked to SI. In the other SI class, we assume that an undetected allele must have been present in the heterozygous parent (denoted as ? in Table 3.4b). The parental genotypes could not be independently tested, however, as the plants were no longer alive when these results were obtained. Our inability to detect the hypothesized third allele may have been because it produces a restriction profile indistinguishable from that of allele *Aly*13-1 but confers a different incompatibility; alternatively, there may be an allele that is not amplified by the PCR primers we have used.

Four incompatibility groups were detected among the F1 progeny of family 98G23 (Table 3.4c). In this family, initially only two *Aly*13 variants were found to segregate among these individuals when using primers 13F1 and *SLG*R. *Aly*13-1 and *Aly*13-3 were found among three of the four incompatibility groups. Using primers specific to the kinase domain, two new alleles Aly13-20 (M. Schierup pers. comm.) and Aly13-23 (B. Mable pers. comm) were amplified. The parents, 97F15-2 (NC) and 97F8-1 (Ind) have alleles *Aly*13-1/*Aly*13-23 and *Aly*13-3/*Aly*13-20, respectively. Dominance among alleles is observed in this family. Again *Aly*13-1 is recessive in that individuals sharing this allele can produce a full complement of fruit when pollen is exchanged. *Aly*13-3 appears recessive relative to alleles *Aly*13-23 and *Aly*13-20 for the same reason, and the latter alleles appear to be codominant to each other.

Four incompatibility groups were detected among individuals of family 98G24 (Table 3.4d and Figure 3.5). The parents of this family were 97F13-5 (NC) and 97F12-3 (Ind) with alleles *Aly*13-1/*Aly*13-13 and *Aly*13-3/*Aly*13-19, respectively (Table 3.3d). Again, *Aly*13-1 appears recessive to Aly13-19. However, the number of pollinations in this family is relatively too small to say whether it is recessive to *Aly*13-3 as well.

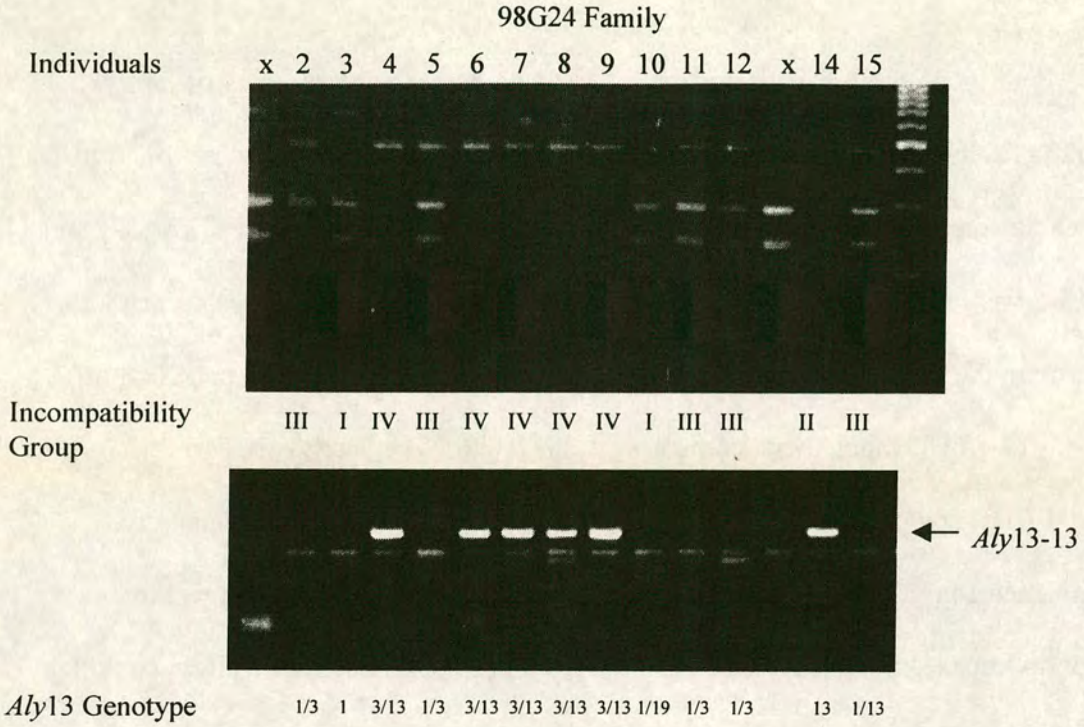


Figure 3.5. Segregation analysis of the 98G24 family using primers 13F-SLGR and digested with AluI (top gel) and using a forward primer specific to the *Aly*13-13 variant (bottom gel). Incompatibility groups are shown under the top gel and the *Aly*13 genotype are shown underneath the bottom gel. Those genotypes where only one allele was amplified were later found to have allele *Aly*13-19 segregating (B. Mable, pers. comm.) - see Table 3.4 segregation analyses. Individual 98G24 was shown subsequently shown to have allele *Aly*13-13 (not shown).

3.4 Discussion

Seven S-domain sequence types (*Aly*3, *Aly*7, *Aly*8sk, *Aly*9, *Aly*10sk1, *Aly*10sk2, *Aly*13) were initially cloned from North American and Scottish *A. lyrata* individuals. It was discovered that three of these clones/variants have a kinase domain 3' (*Aly*8sk, *Aly*10sk1and *Aly*10sk2) of the S-domain. Another clone, *Aly*13 appears to include a number of variants that have a kinase domain (*Aly*13-2, *Aly*13-3, *Aly*13-9, *Aly*13-13) and one variant which appears not to have one (*Aly*13-1). The nature of those *Aly*13 variants (*SLG*-like or *SRK*-like) where a kinase domain has not been amplified remains to be determined. A PCR product for each *Aly*-type can be amplified from every

Four incompatibility groups were detected among the F1 progeny of family 98G23 (Table 3.4c). In this family, initially only two *Aly*13 variants were found to segregate among these individuals when using primers 13F1 and *SLG*R. *Aly*13-1 and *Aly*13-3 were found among three of the four incompatibility groups. Using primers specific to the kinase domain, two new alleles Aly13-20 (M. Schierup pers. comm.) and Aly13-23 (B. Mable pers. comm) were amplified. The parents, 97F15-2 (NC) and 97F8-1 (Ind) have alleles *Aly*13-1/*Aly*13-23 and *Aly*13-3/*Aly*13-20, respectively. Dominance among alleles is observed in this family. Again *Aly*13-1 is recessive in that individuals sharing this allele can produce a full complement of fruit when pollen is exchanged. *Aly*13-3 appears recessive relative to alleles *Aly*13-23 and *Aly*13-20 for the same reason, and the latter alleles appear to be codominant to each other.

Four incompatibility groups were detected among individuals of family 98G24 (Table 3.4d and Figure 3.5). The parents of this family were 97F13-5 (NC) and 97F12-3 (Ind) with alleles *Aly*13-1/*Aly*13-13 and *Aly*13-3/*Aly*13-19, respectively (Table 3.3d). Again, *Aly*13-1 appears recessive to Aly13-19. However, the number of pollinations in this family is relatively too small to say whether it is recessive to *Aly*13-3 as well.

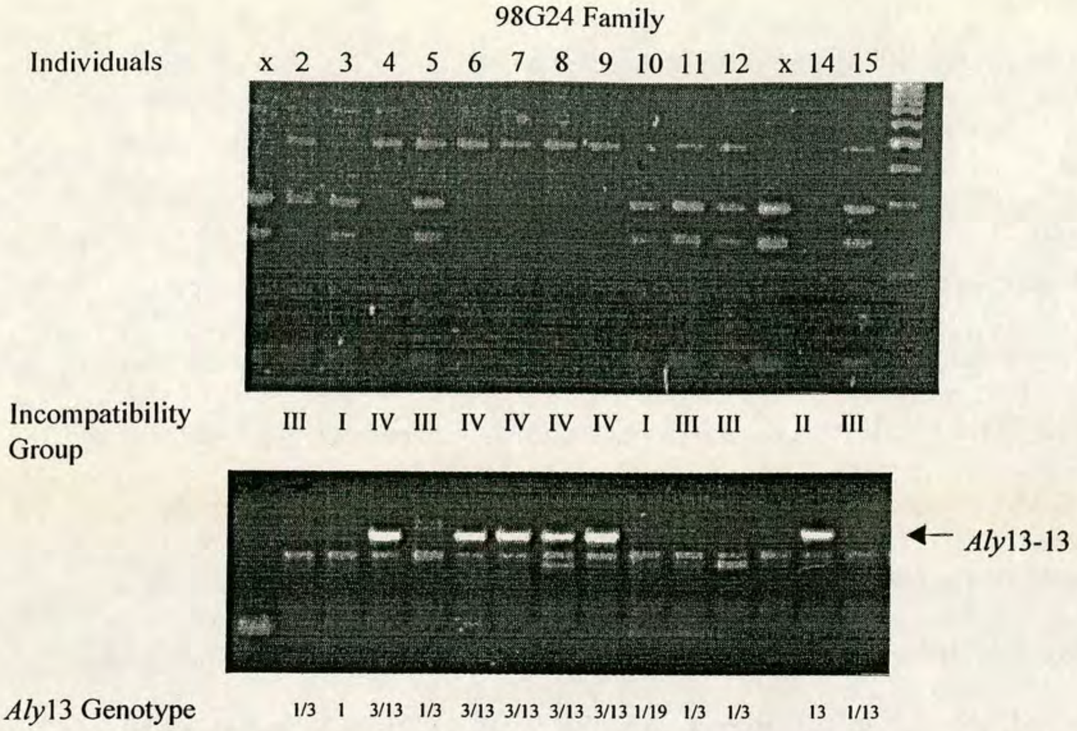


Figure 3.5. Segregation analysis of the 98G24 family using primers 13F-SLGR and digested with AluI (top gel) and using a forward primer specific to the *Aly*13-13 variant (bottom gel). Incompatibility groups are shown under the top gel and the *Aly*13 genotype are shown underneath the bottom gel. Those genotypes where only one allele was amplified were later found to have allele *Aly*13-19 segregating (B. Mable, pers. comm.) - see Table 3.4 segregation analyses. Individual 98G24 was shown subsequently shown to have allele *Aly*13-13 (not shown).

3.4 Discussion

Seven S-domain sequence types (*Aly*3, *Aly*7, *Aly*8sk, *Aly*9, *Aly*10sk1, *Aly*10sk2, *Aly*13) were initially cloned from North American and Scottish *A. lyrata* individuals. It was discovered that three of these clones/variants have a kinase domain 3' (*Aly*8sk, *Aly*10sk1and *Aly*10sk2) of the S-domain. Another clone, *Aly*13 appears to include a number of variants that have a kinase domain (*Aly*13-2, *Aly*13-3, *Aly*13-9, *Aly*13-13) and one variant which appears not to have one (*Aly*13-1). The nature of those *Aly*13 variants (*SLG*-like or *SRK*-like) where a kinase domain has not been amplified remains to be determined. A PCR product for each *Aly*-type can be amplified from every individual using primers specific to each S-domain, therefore the S-domain types described in this chapter reflect separate loci with respect to each other. Nevertheless, the pollination studies and segregation analyses reveals that in *A. lyrata*, like in *Brassica*, S-domain loci do segregate with SI revealing that this form of incompatibility has been maintained for very divergent plants in this family.

Some individuals have variants of *Aly*7, which have an insertion that knocks the sequence out of reading frame, suggesting that this represents either a null allele or a separate locus. Finally, variants of *Aly*3, *Aly*8sk, *Aly*10sk1 and *Aly*10sk2 do not segregate with the S-phenotype in one family. Diversity at *Aly*9 appears low enough that it cannot possibly be an S-locus (and is most similar to *ATS*1 in *A. thaliana*). *Aly*10sk1, *Aly*10sk2 and *Aly*8sk appear most closely related to *ARK*1, *ARK*2, and *ARK*3 in *A. thaliana*, respectively. Finally, variants of *Aly*13 appear to segregate with the S-phenotype in four families, including variants *Aly*13-1, *Aly*13-3, and *Aly*13-13 and are likely to be variants of the *SLG* and *SRK* loci in *A. lyrata*.

The observation that a number of individuals tested do not have the same *Aly*13 variant is more than suggestive that this group of sequences is allelic. Even though we do not have segregation data for each variant that we have identified in this and the following Chapter, it is unlikely that each variant represents a single locus. However, this does not mean that we can say with certainty that this group of sequences only consists of one locus. We have already shown that some variants are likely *SLG*-like and some are likely *SRK*-like.

cDNA experiments that determine when and where these various loci are expressed, although useful, will not tell us more than the genetic information. In *Brassica*, both the *SLG* and *SRK* are expressed in the stigma during SI. Nevertheless,

M. Schierup and myself performed preliminary rtPCR amplifications from total RNA isolated from stigma tissue, as well as leaf tissue, which resulted in amplification of *Aly*13-1 from a few individuals from both sets of tissue. The fact that both tissues amplified means that these amplifications could very well be the *SLG* locus, but we do not have a negative control, as the protein is thought to be expressed at low but detectable levels with PCR in other tissues (Nasrallah 1987).

Not discussed in the section dealing with segregation analyses was a variant identified called *Aly*13-2. This sequence does not appear to amplify in every individual, but does amplify in some members of the 98E15 family array. Among some individuals where it did amplify, these individuals already appeared to have two amplified products. Given that there is a possibility of four possible alleles which can be amplified in one heterozygous individual (two variants at each of the *SLG* and the *SRK* loci), this should not be surprising. However, this variant appears to be partially linked to incompatibility groups II and III in family 98E15. This may be an indication of a recombinant haplotype.

individual using primers specific to each S-domain, therefore the S-domain types described in this chapter reflect separate loci with respect to each other. Nevertheless, the pollination studies and segregation analyses reveals that in *A. lyrata*, like in *Brassica*, S-domain loci do segregate with SI revealing that this form of incompatibility has been maintained for very divergent plants in this family.

Some individuals have variants of *Aly*7, which have an insertion that knocks the sequence out of reading frame, suggesting that this represents either a null allele or a separate locus. Finally, variants of *Aly*3, *Aly*8sk, *Aly*10sk1 and *Aly*10sk2 do not segregate with the S-phenotype in one family. Diversity at *Aly*9 appears low enough that it cannot possibly be an S-locus (and is most similar to *ATS*1 in *A. thaliana*). *Aly*10sk1, *Aly*10sk2 and *Aly*8sk appear most closely related to *ARK*1, *ARK*2, and *ARK*3 in *A. thaliana*, respectively. Finally, variants of *Aly*13 appear to segregate with the S-phenotype in four families, including variants *Aly*13-1, *Aly*13-3, and *Aly*13-13 and are likely to be variants of the *SLG* and *SRK* loci in *A. lyrata*.

The observation that a number of individuals tested do not have the same *Aly*13 variant is more than suggestive that this group of sequences is allelic. Even though we do not have segregation data for each variant that we have identified in this and the following Chapter, it is unlikely that each variant represents a single locus. However, this does not mean that we can say with certainty that this group of sequences only consists of one locus. We have already shown that some variants are likely *SLG*-like and some are likely *SRK*-like.

cDNA experiments that determine when and where these various loci are expressed, although useful, will not tell us more than the genetic information. In *Brassica*, both the *SLG* and *SRK* are expressed in the stigma during SI. Nevertheless,

M. Schierup and myself performed preliminary rtPCR amplifications from total RNA isolated from stigma tissue, as well as leaf tissue, which resulted in amplification of *Aly*13-1 from a few individuals from both sets of tissue. The fact that both tissues amplified means that these amplifications could very well be the *SLG* locus, but we do not have a negative control, as the protein is thought to be expressed at low but detectable levels with PCR in other tissues (Nasrallah 1987).

Not discussed in the section dealing with segregation analyses was a variant identified called *Aly*13-2. This sequence does not appear to amplify in every individual, but does amplify in some members of the 98E15 family array. Among some individuals where it did amplify, these individuals already appeared to have two amplified products. Given that there is a possibility of four possible alleles which can be amplified in one heterozygous individual (two variants at each of the *SLG* and the *SRK* loci), this should not be surprising. However, this variant appears to be partially linked to incompatibility groups II and III in family 98E15. This may be an indication of a recombinant haplotype.

Chapter 4

Intra- and Inter-specific Nucleotide Diversity and Substitution Rate Variation among Seven Members of The Self-Incompatibility Gene Family in *Arabidopsis lyrata*

4.1.1 Introduction

Much of our current understanding concerning the forces affecting intra- and inter-specific nucleotide variation in natural populations has come from population genetic studies of *Drosophila* and humans. Studies of the evolutionary mechanisms that affect the evolution of gene families have also been limited to these few species. The self-incompatibility (SI) gene family in *Brassica* plant populations (Nasrallah and Nasrallah 1989, Dwyer et al. 1991, Nasrallah and Nasrallah 1993) provides a system for testing models of evolutionary processes that require random-mating, and this well-characterised gene family allows us to test models of gene family evolution (Goodman et al. 1975, Dykhuizen and Hartl 1980). Until recently, studies of genetic variation in natural plant populations have concentrated on the effects of mating-system and effective population size, or other natural history variables that can affect neutral variation at unique loci (i.e. Hamrick and Godt 1989, Charlesworth et al. 1997). The effects of selection, recombination, and concerted evolutionary processes on plant gene families has received little attention (e.g. Vieira et al. 1999).

As discussed in previous chapters, the *SLG* and *SRK* loci in the *Brassicaceae* share similar sequence structure to a large group of loci (the 'S-domain' gene family). In *Brassica*, the *SLG* and *SRK* loci exhibit much higher variation relative to other members of the S-domain family (Hinata et al. 1995). Three hypervariable regions (HVRs) are observed at similar positions in *SLG* and *SRK* loci and variability estimates within and between the S-domains of *SLG* and *SRK* loci are indistinguishable (Kusaba et al. 1997). The question remains whether these regions are functionally important, and hence the targets of balancing selection versus merely under relaxed selective constraint (Hinata et al. 1995, Kusaba 1997, Awadalla and Charlesworth 1999).

In Chapter 3, we described known members of the S-domain family isolated and characterised in *Brassica* and *Arabidopsis*. As mentioned, the gene family consists primarily of two groups of sequences, those with or without a 'kinase' domain. In *Brassica* and *Arabidopsis*, a group of sequences called *SLR* (*SLR*1 and *SLR*2) or *Ats*1 (in *A. thaliana*), and a set of putative serine-threonine kinase loci (*Bcnskl* in *Brassica* and *Ark* loci in *A. thaliana*) have been characterized (Isogai et al. 1988, Lalonde et al. 1989, Dwyer et al. 1994, Tobias and Nasrallah 1996, Suzuki 1996, Luu et al. 1997). The *Brassica* 'paternal' gene, the *SCR* gene, is expressed in pollen and is very different in structure from the maternal genes. The *SCR* gene consists of a small single-copy (ca. 75 amino-acids), cysteine-rich gene (SCR, S-locus cysteine-rich protein) exhibiting similarity to pollen coat proteins (PCPs) which exhibit a similar cysteine patterning within the locus and it is thought that this locus may act as a ligand for the stigmatic receptor (Stein et al. 1991, Goring and Rothstein 1992, Schopfer et al. 1999, Nishio et al. 2000). It seems clear that the maternal and paternal genes have evolved from very different gene families.

4.1.2 Models of Gene Family Evolution

The evolution of new protein functions after genome or gene duplication events is believed to play a major role in genome evolution and the evolution of complexity (Li 1997), but the mechanisms by which this process occurs remains controversial. The most widely cited model for the evolution of new protein function assumes that new functions arise as a result of selectively neutral mutations that are fixed by chance in a redundant gene copy and fortuitously provide it with a new function (Dykhuizen and Hartl 1980). After a gene duplication event, the rate of nonsynonymous nucleotide substitutions may be enhanced because of relaxation of functional constraints of redundant genes after gene duplication, but the rate will never exceed that of synonymous substitutions. If the environmental or genetic background changes, then these effectively neutral loci, may become positively selected (Long and Langley 1993). Recent lines of evidence suggest that this rarely happens. An alternative hypothesis is that gene duplication should be followed by positive selection (Ohta 1994, Zhang et al. 1998, Hughes et al. 2000). It is difficult to find evidence in support of the former hypothesis (Long and Langley 1993, Ohta 1994) as mutations and purifying selection will obscure many of the hallmarks of selection. In contrast, cases for the latter hypothesis are made when positive diversifying selection accelerates the fixation of nonsynonymous substitutions specifically (Goodman et al. 1975). The ratio of nonsynonymous relative to synonymous divergence has sometimes been found to exceed values of one in gene families, and this may be considered to be evidence for positive selection (e.g. Hughes et al. 1990, Zhang et al. 1998, Hughes et al. 1998, Yang et al. 2000).

4.1.3. Effects of concerted evolution on diversity

Many studies have now shown that members of multigene families do not evolve independently, and various mechanisms of homogenization, including unequal crossing over and gene conversion, have been proposed to explain the concerted evolution of gene family members (Arnheim 1983). Although gene conversion is considered to be a homogenizing mechanism (Walsh 1985), there is evidence that gene conversion can also generate variability among members of multigene families (Xiong et al. 1988, Kuhner et al. 1991, Wines et al. 1991, Ohta 1994, King 1997, Hughes et al. 2000). Mutations and drift act as a heterogenizing forces among diverging loci, and therefore, low rates of gene conversion involving the exchange of relatively short fragments may create haplotypic variation that may be mistaken for mutations. The exchange of large fragments will tend to homogenize two sequences, as would 'normal' recombination among alleles, which redistributes polymorphism among lineages, and hence decreases the variance in *Ne* among alleles (e.g. Hudson 1990).

This chapter focuses on diversity at members of the S-domain gene family in *A. lyrata* populations. In Chapter 3, we showed that seven S-domain 'loci' were identified in *A. lyrata* individuals and were found to be distinct loci with respect to each other. All but one sequence type were found to be unlinked to the S-locus, whereas one group of sequences (*Aly*13), segregates with incompatibility haplotypes in family segregation analyses (Chapter 3). Here we examine diversity at these loci to establish the level of neutral diversity for genes in *A. lyrata* as well as assess the distribution of selective constraints among codons for S-domain loci in this species. Furthermore, we examined the role that recombination and mutation may have played on shaping diversity and substitution patterns among these paralogues. We observed that variation at unlinked loci differs markedly from *Brassica* self-incompatibility loci. World-wide analysis of nucleotide variation for some S-domain loci is low, consistent with previous published results (Innan et al. 1996, Kawabe et al. 1997, Purugganan and Suddith 1998, 1999, Stahl et al. 1999, Savolainen et al. 2000). This polymorphism is distributed across these loci inconsistently such that variation is not clustered in the regions corresponding to the HV regions of the *SLG* or *SRK* loci. All loci appear to be recombining except locus *Aly*9 which also exhibits very little nucleotide diversity, and *Aly*7, which may constitute two loci. Orthologous and paralogous comparisons revealed heterogeneity in rates of divergence for nonsynonymous sites across the S-domain exon. Sequences appear to be evolving in a non-clock like manner, with some loci exhibiting accelerated rate of divergence. There is no evidence of positive selection acting among codons using likelihood approaches (Yang et al. 2000), therefore this can not account for the heterogeneity in divergence among regions. Finally, we discuss the possibility that concerted evolutionary processes in this gene family have contributed to shaping patterns of variability at the loci segregating with the self-incompatibility phenotype.

4.2 Materials and Methods

4.2.1 Arabidopsis lyrata individuals and DNA preparation.

Genomic DNA was isolated from *A. lyrata* individuals as outlined in Chapter 3. Seeds were collected from 4 populations; North Carolina, USA; Michigan, USA; Braemar, Scotland (R. Ennos); and Reykjanes peninsula, Iceland. The same individuals, from separate maternal families, were used for amplifications for all loci. Loci were sequenced directly or PCR products were cloned and sequenced as described in Chapter 3. All variants of loci *Aly*3 and *Aly*9 were sequenced directly. For some individuals, PCR products for *Aly*7, *Aly*8, *Aly*10sk1, and *Aly*10sk2 were cloned as individuals were either heterozygous, or primers for each locus would occasionally amplify alternative members of this gene family. For cloned PCR products, at least 3 clones per individual were sequenced to check for amplification errors. DNA sequencing from PCR products was performed and checked using standard protocols on the ABI 377 automated sequencer. Only variation at the S-domain loci was assessed as not all loci have a kinase domain, and in *Brassica*, the S-domain of the *SLG* and *SRK* loci appear to be the exons directly involved in interactions with the pollen protein.

4.2.2 DNA sequence analysis.

To assess nucleotide diversity and proportions of synonymous substitutions (*Ks*) and nonsynonymous substitutions per site (*Ka*), the method of Nei and Gojobori (1986) was used using the MEGA v. 1.01 (Kumar et al. 1993) as in Chapter 2. Alignment gaps were included in calculations in pairwise comparisons.

Population subdivision can affect interpretations of polymorphism and can mimic or obscure genealogical structures created by other deterministic processes such as recombination and/or selection (Maynard Smith and Haigh 1974, Tajima 1989, Hey and Wakeley 1997). To determine whether there was significant population subdivision among *A. lyrata* individuals at the S-domain loci, the value of the *Kst* statistic (Hudson et al. 1992) was calculated for all groups of sequences using the program Proseq v. 2 (Filatov 1999). *Kst* is merely 1- *Ks/Kt* where *Ks* is the weighted average of the pairwise diversity for two sequences sampled within two populations and *Kt* is the average number of nucleotide diversity for two sequences sampled regardless of the locality (Hudson et al. 1992). The critical value for this statistic was obtained by 1000 random permutations of the sequences between the four groups in the sample (Hudson et al. 1992).

In order to test whether the loci not linked to SI in *A. lyrata* are evolving neutrally, Tajima's (1989) and McDonald and Kreitman's (1991) tests of neutrality were performed using DNAsp v. 3 (Rozas and Rozas 1997). The 'Tajima's D' statistic (Tajima 1989) summarizes the allele frequency spectrum by examining how estimates of θ, or $4N\mu$ where N is the effective population size and μ is the per nucleotide mutation rate, calculated using the number of pairwise differences (Watterson 1975, Tajima 1983) differ from estimates of θ based only on the number of segregating sites (Watterson 1975). θ based on the number of segregating sites is more sensitive to low frequency variants in the population. For example, an excess of rare variants, relative to informative or more polymorphic sites, suggests that there are many more mutations on the tips of a coalescent process than would be expected for neutrally evolving sites. In this case, Tajima's *D* is negative. Such a pattern can be generated by a selective sweep or population demographic changes such as population growth from a bottleneck. Too few rare variants relative to the neutral expectation may be an indication of either balancing selection or population subdivision. Tajima's *D* will then be positive. The McDonald-Kreitman test is a test of heterogeneity between synonymous and nonsynonymous polymorphism and divergence. Under neutrality, the ratio of replacement to synonymous fixed substitutions (differences) between species should not significantly differ from the ratio of replacement to synonymous polymorphisms within species. Both of these tests generally suffer from a lack of power, and are especially conservative for genes with low-moderate levels of polymorphism.

To examine the relationship of nucleotide diversity and recombination, three different analyses were performed to estimate recombination rates and determine whether recombination-like mechanisms were affecting haplotype variation among the unlinked S-domain loci in *A. lyrata*. First, estimators of linkage disequilibrium, r^2 (Hill and Robertson 1968; Awadalla and Charlesworth 1999; Chapter 2) were calculated between pairs of sites using DNAsp v.3 (Rozas and Rozas 1997). Mantel tests were used to assess the relationship of linkage disequilibrium with nucleotide distance for all sites (Genstat, Numerical Algorithms Group). We used both nonsynonymous and synonymous sites for all S-domains, excluding ambiguous sites or sites with missing information. This analysis also provides a further means of determining whether the sequences of a given group of loci are truly allelic. We would not expect to detect a negative relationship between linkage disequilibrium and nucleotide distance, if the sequences came from different loci. Second, the minimum number of recombination events per locus was determined by the algorithm in Hudson and Kaplan (1985). Generally, this is based on estimating the number of pairwise comparisons which exhibit all 4 pairwise possible haplotypes (4-gamete test) but correcting for sites involved in the same recombination events. Finally, to determine whether different loci were exchanging fragments, the probability of gene conversion events was assessed using the method of Betran et al. (1997).

As described in the Introduction, one way positive selection can be detected is if nonsynonymous divergence or rates exceed synonymous rates. We used two approaches to detect this type of signal of positive selection in unlinked S-domain loci. A simple method involves estimating *Bs* and *Ba*, merely the synonymous and nonsynonymous substitutions for each branch of a genealogy (equivalent to *Ka* and *Ks* for extant taxa). Once these estimates are obtained, we can test for positive selection or changes in selective constraint by examining the the ratio of the estimate of *Ba* and *Bs* in a manner similar to pairwise comparisons of extant sequences (*Ka/Ks* ratios).

Likelihood ratio tests were used to determine whether lineages deviated from the molecular clock and neutrality by determining whether there was significant heterogeneity in either substitution rates or selective constraints (*Ka/Ks*) ratios among branches in this gene family (Yang and Neilson 2000, Yang et al. 2000). PAML (Phylogenetic Analysis Using Maximum Likelihood) ver. 3.0 was used to test different models of evolution for the S-domain gene family within *A. lyrata* (Yang 1997). We considered the difference in log-likelihood between models requiring different numbers of parameters to test the hypotheses of (1) a molecular clock for silent and replacement changes and (2) homogeneity of the nonsynonymous to synonymous substitution rate across each evolutionary lineage, both for the coding region as a whole and for each exon separately. Twice the difference in the log-likelihood between models is approximately χ^2 distributed with the degrees of freedom equal to the difference in number of parameters.

Finally, similar tests were performed to determine if models with *Ka/Ks* ratios fixed among sites fit the data better than a model where this is relaxed among codons. When testing for models of relaxed *Ka/Ks* values among codons, two models were tested. The first involved a model of a beta-distribution of *Ka/Ks* ratios among sites with bounds from zero to one was fit to the data (no positive selection). The second model involved a similar beta distribution but also allowed for a category of *Ka/Ks* ratios greater than one (Yang et al. 2000). If the second model provided a significantly improved fit to the data relative to the model with fixed *Ka/Ks* values among codons, or to a beta distribution with an upper bound of one, then this is evidence for positive selection acting among genes of the gene family (some *Ka/Ks* values for codons are greater than one).

4.3. Results

4.3.1 Comparing within-locus variation between linked and unlinked 'S-domain' loci in *A. lyrata*

Variation at loci unlinked to SI

We assessed the variation at 6 S-domain loci that are unlinked to the self-incompatibility phenotype in *A. lyrata*, and compared them to variation at the *Aly*13 locus. Nucleotide variability estimates at the six putatively neutral loci provide a general estimate of the neutral mutation rate (θ) for *A. lyrata* and can be compared to previous published estimates made for other loci in *A. lyrata* (Innan et al. 1996, Kawabe et al. 1997, Purugganan and Suddith 1998, 1999, Stahl et al. 1999, Savolainen et al. 2000). Furthermore, the unlinked S-domain loci can be treated as 'reference' loci relative to the *Aly*13 locus, and used a basis by which to compare variation at sequences where there is evidence that they are linked to the S-phenotype (Chapters 3 & 5).

Table 4.1 and Appendix 1.1 describe the nucleotide variability for all the S-domain loci sequences in the total sample as well as for each population separately. Excluding the *Aly*13 set of sequences, global synonymous variability estimates range from 0.0054 (*Aly*9) to 0.0795 (*Aly*8sk). The majority of loci exhibit nucleotide variation similar to that observed for a number of published estimates of nucleotide diversity at loci in *A. lyrata*. Overall nucleotide diversity estimates range from 0.005 to 0.013 for a variety of loci in *A. lyrata* (Innan et al. 1996, Kawabe et al. 1997, Purugganan and Suddith 1998, 1999, Stahl et al. 1999, Savolainen et al. 2000). Only mean pairwise nucleotide diversity was estimated for *Aly*7 as some variants of this S-domain type exhibit frame-shifts (Chapter 3). Regardless, nucleotide diversity at *Aly*7 does not appear to differ tremendously compared to the other S-domain loci sequenced in *A. lyrata*, or relative to other published estimates. *Aly*13 variants exhibit exceptionally high levels of variability at both the synonymous and nonsynonymous level with synonymous variability approaching saturated levels (Chapter 5). This high variability is consistent with *Brassica SLG* and *SRK* estimates and other loci subject to balancing selection (Chapter 1). In Chapter 5, we will address this variation in greater detail.

*Aly*8sk variation is curiously high relative to the other S-domain loci (excluding *Aly*13), as well as to other published estimates for *A. lyrata*. Nucleotide diversity (both synonymous and nonsynonymous) for *Aly*8sk is approximately 0.025 and is almost two times the largest published *A. lyrata* estimate (*Adh,* $\pi = 0.013$; Kawabe et al. 1997). The high variability at *Aly*8sk suggests that the variants sequenced may constitute 2 loci but unlike *Aly*7, *Aly*8sk does not exhibit distinctive haplotypes and nonsynonymous diversity is relatively low (about 1%).

4.3.2. Variation at sequences segregating with SI

Loci that are affected by balancing selection exhibit elevated levels of polymorphism relative to neutrally evolving loci (e.g. Hughes et al. 1990). *Aly*13 sequences show exceptionally high levels of both synonymous and replacement site diversity (Table 4.1; Hinata et al. 1995, Charlesworth and Awadalla 1998). The values are shown here for comparison to the remaining S-domain loci which are not implicated in SI in *A. lyrata*. Synonymous nucleotide variation is high enough that corrections for multiple substitutions (Kimura 2-parameter model) dramatically affect the estimates. The diversity estimates shown in Table 4.1 include all *Aly*13 S-domain variants, regardless of whether they have a kinase domain or not. Including alleles from either an SLG-like or an SRK-like locus will not dramatically affect estimates of nucleotide diversity, as *Brassica SLG* and *SRK* loci are indistiguishable from each other in terms of variability and fixed differences at the S-domain.

Table 4.1: Variability estimates and summary statistics for 7 S-domain loci in *A. lyrata* populations sampled globally.

Locus		Global	North Carolina	Indiana	Scotland	Iceland	K_{st}	#/*S*	Tajima's *D*
*Aly*3	π_s	0.0084	0.0127	0.0000	0.0032	0.0242	0.60*	8/29	0.10
	π_a	0.0061	0.0103	0.0000	0.0023	0.0163			
	n	22	8	6	4	4			
*Aly*7 (π only)		0.0122	0.0104	0.0037	0.0117	0.0136	0.20*	21/40	-1.19
	n	27	8	3	10	5			
*Aly*8sk	π_s	0.0795	0.0821	0.0812	0.0800	0.0619	0.02	38/74	-1.31
	π_a	0.0122	0.0132	0.0117	0.0127	0.0094			
	n	26	3	8	7	8			
*Aly*9	π_s	0.0054	0.0201	0.0025	0.0000	0.0000	0.13*	10/16	-1.77
	π_a	0.0048	0.0071	0.0006	0.0032	0.0000			
	n	20	5	6	5	4			
*Aly*10sk1	π_s	0.0138	0.0000	0.0021	0.0109	0.0032	0.06*	19/34	-1.75
	π_a	0.0022	0.0011	0.0016	0.0085	0.0078			
	n	26	6	5	10	5			
*Aly*10sk2	π_s	0.0227	0.0000	0.0220	0.0243	0.0127	0.08*	25/41	-1.31
	π_a	0.0085	0.0090	0.0090	0.0092	0.0052			
	n	19	4	4	7	4			
***Aly*13**	π_s	**0.3612**	**0.2915**	**0.3486**	**0.4467**	**0.3144**	**0.10**	**241/743**	**-0.19**
	π_a	**0.2000**	**0.1722**	**0.2005**	**0.2306**	**0.1695**			
	n	38	6	10	15	7			

# is the number of singletons. *S* is the total number of segregating sites.

n is the sample size under the diversity values corresponding to each category.

*Significant population structure $p < 0.01$.

4.3.3. Distribution of variation across the S-domain sequences in *A. lyrata*

The *SLG* and *SRK* loci exhibit HV regions at precisely the same positions in both loci. This could reflect either a region under relaxed selective constraint or a region targetted by balancing selection. Especially at the *SRK* locus, mutations in this region may alter the S-phenotype. To determine whether the same patterns are seen in the S-domains in all *A.lyrata* genes, a sliding window analysis was performed for synonymous and nonsynonymous variability. If loci unlinked to SI exhibit similar peaks, it would suggest that these regions are evolving neutrally.

Variability patterns across the S-domain sequences do not appear to be consistent among the S-domains in *A. lyrata* (Table 4.2). Figure 4.1 shows sliding window plots for the different S-domain types. *Brassica SLG* and *SRK* Type I alleles exhibit HV regions (Dwyer et al.1993) corresponding to positions 215 – 236 for HV1, 293 - 331 for HV2, 354 – 368 for HV3, and 446-456 for the C-terminal region in our alignments. With the exception of *Aly*13, there are no significant peaks of amino-acid variability relative to the rest of the gene.

Table 4.2. Distribution of diversity values expressed as mean pairwise differences in different parts of the *Aly S-domain* sequences. Number of sequences analysed are as in Table 4.1. Only regions that overlap among all the sequences were analysed here.

		"conserved"	"HV"
Nucleotides		470	235
Aly3	*Ka*	0.0172 ± 0.0045	0.0057 ± 0.0042
	Ks	0.0152 ± 0.0069	0.0114 ± 0.0125
*Aly*8	*Ka*	0.0134 ± 0.0032	0.0134 ± 0.0043
	Ks	0.0654 ± 0.0140	0.0967 ± 0.0246
*Aly*10sk1	*Ka*	0.0075 ± 0.0021	0.0060 ± 0.0026
	Ks	0.0131 ± 0.0063	0.0069 ± 0.0047
*Aly*10sk2	*Ka*	0.0100 ± 0.0025	0.0074 ± 0.0025
	Ks	0.0315 ± 0.0089	0.0075 ± 0.0055
*Aly*9	*Ka*	0.0077 ± 0.0024	0.0044 ± 0.0032
	Ks	0.0087 ± 0.0052	0.0222 ± 0.0136

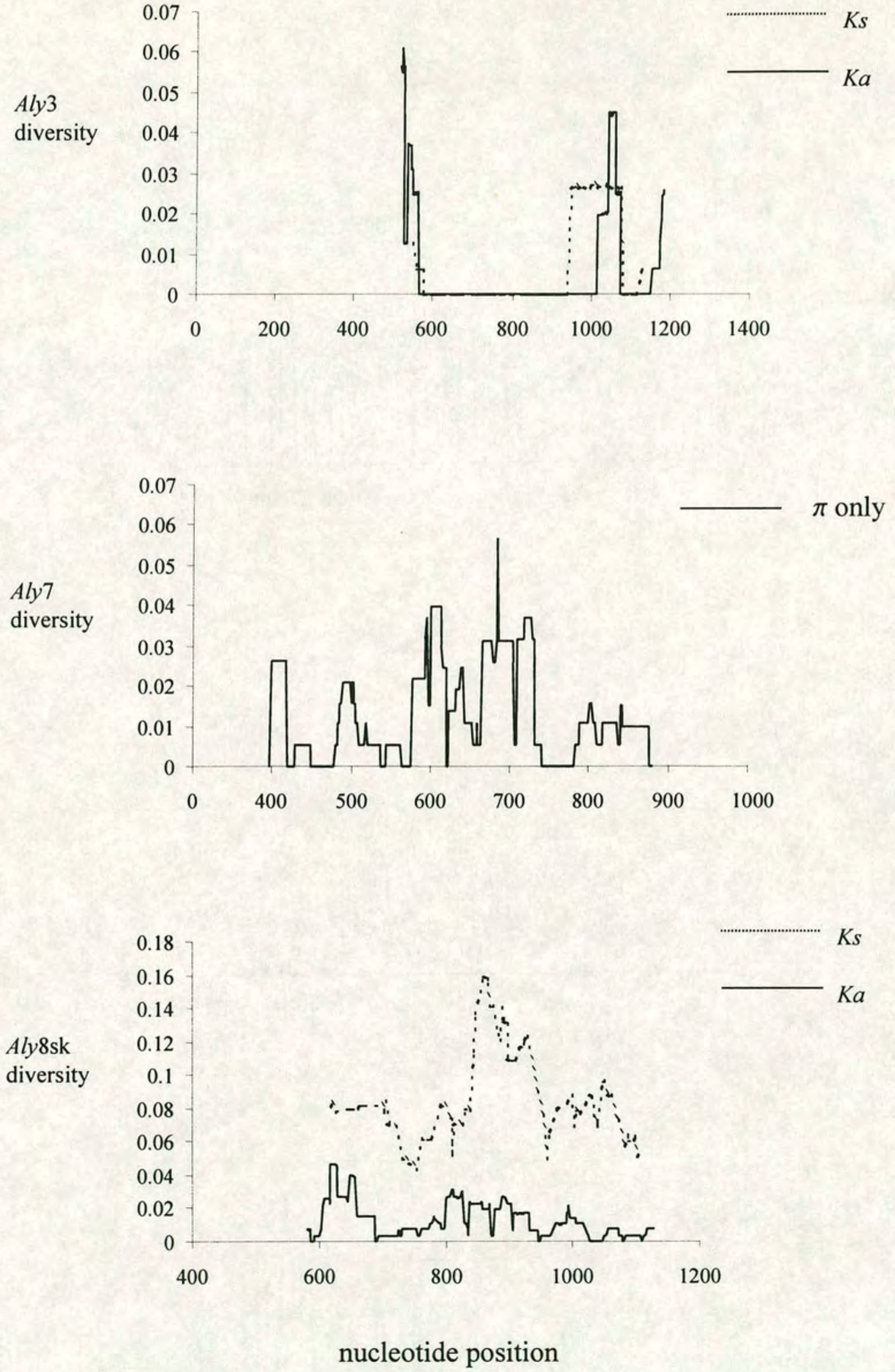
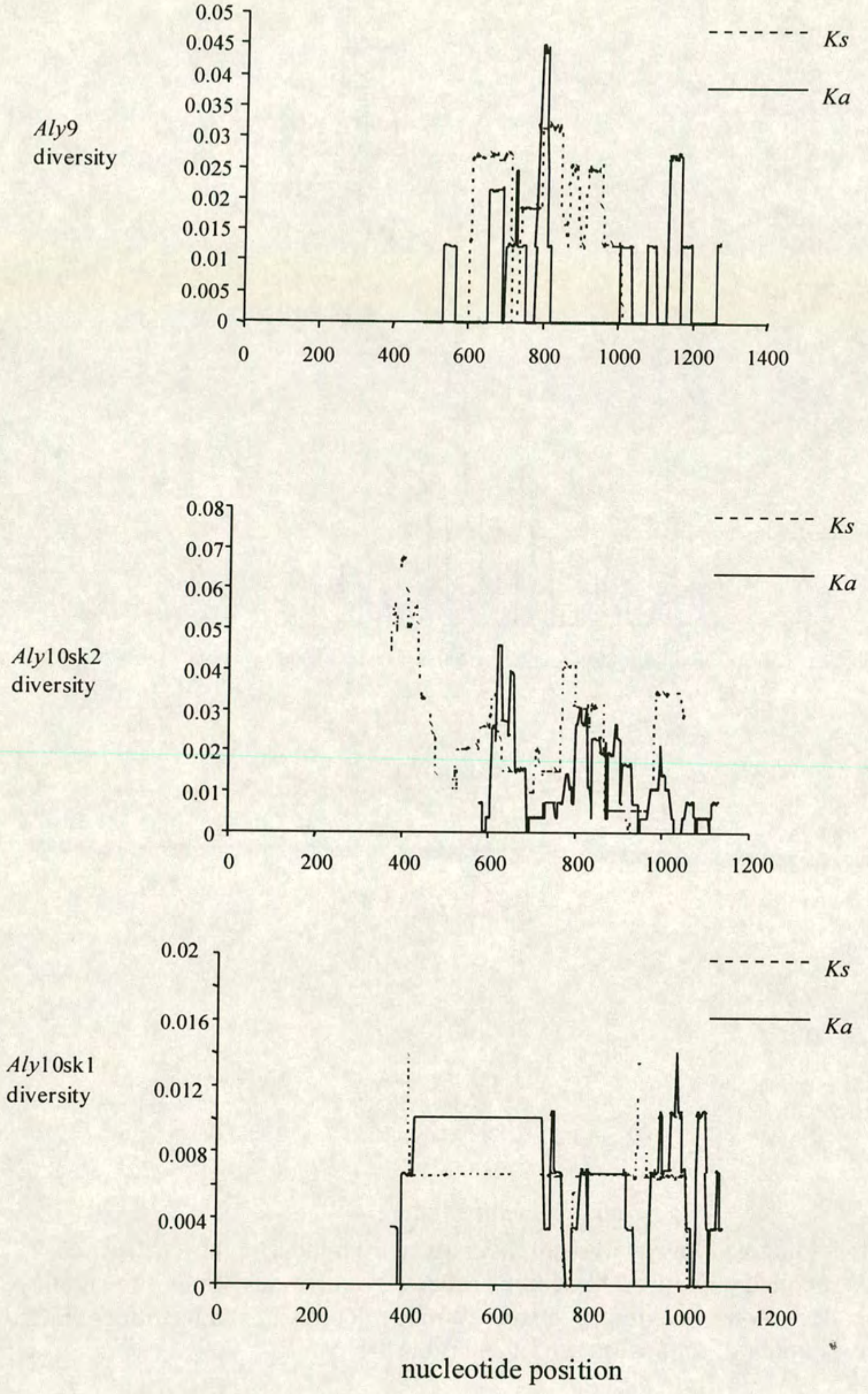


Figure 4.1: Sliding window analysis of diversity at each locus within *A. lyrata* individuals sampled globally. Dashed line reflects synonymous variability and the solid line reflects nonsynonymous variants. Window length is 20 nucleotides shifted every base. Number of sequences used as in Table 4.1.


Aly9
diversity
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
200
400
600
800
1000
1200
1400
Ks
Ka
Aly10sk2
diversity
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
200
400
600
800
1000
1200
Ks
Ka
Aly10sk1
diversity
0.02
0.016
0.012
0.008
0.004
0
0
200
400
600
800
1000
1200
Ks
Ka
nucleotide position

4.3.3. The effects of population subdivision on nucleotide variation and Tajima's D statistic

To determine whether there was significant spatial structure contributing to the global estimates of nucleotide variation in *A. lyrata*, we estimated variation within populations (π_S) as well as for all populations together (π_T)(Table 4.1). Population subdivision, measured as *Kst* for all segregating nucleotides (Hudson et al. 1992), differs among loci (Table 4.1). With the exception of *Aly*8sk, all the loci exhibit significant population structure (*Kst* values significantly greater than zero) but the *Kst* values are generally small. The lack of population structure at *Aly*8sk suggests that the high variability at this locus is not a consequence of population subdivision. *Aly*3 exhibits the highest level of population structure (*Kst* = 0.60) and *Kst* is significantly different from zero. Interestingly, this locus also exhibits the most positive value for the Tajima's *D* statistic, although it was not significantly different from zero (see below). Finally, for all *Aly*13 sequence types, *Kst* is not significantly different from zero, as expected for a locus under balancing selection (see Chapter 5).

As mentioned in the Materials and Methods section, estimates of the Tajima's *D* statistic can be affected by both selection and population subdivision. Tajima's *D* estimates were calculated for each locus separately either for the entire sample or for each population separately. The Tajima's *D* estimates for each locus are shown in Table 4.1. *Aly*3 is the only locus to exhibit a slightly positive value of Tajima's *D* calculated from the total sample which is congruent with the high *Kst* value. The other loci exhibit negative values that are not significant, except for *Aly*10sk1. At the single-population level, only one population exhibited a Tajima's *D* value close to significance (*Aly*10sk1– Michigan, $n = 8$, $D = -1.70$, $P = 0.056$). Sample sizes within each population are low enough to expect that it will be unlikely to observe a significant Tajima's *D* estimates. *Aly*10sk1 exhibits very low, but significant, population subdivision. Excluding *Aly*10sk1, all the Tajima's *D* values are very close to zero (not shown).

It appears clear that population subdivison is affecting the distribution of singleton and informative sites. Population subdivision will affect neutrality tests dependent on these distributions. Population subdivision can bias Tajima's *D* in a positive direction in some cases, such as for *Aly*3 where there appear to be fixed differences that distinguish the North Carolina population from the remaining populations. Fixed differences between populations will appear to be heterozygous when sequences from two populations are combined. In contrast, the remaining S-domain loci exhibit negative global estimates of Tajima's *D*. This appears to largely be an effect of sites monomorphic in one population, but polymorphic in another, which may appear as rare variants (singletons) when data for populations are combined (see Appendix 1.2). Although many of the loci exhibit negative Tajima's *D* values, the proportion of singletons to informative sites were roughly equal and close to the neutral expectation (Table 4.1 Tajima 1989).

4.3.4. Linkage Disequilibrium and Recombination at unlinked S-domain loci

To determine whether these loci are recombining, the relationship of linkage disequilibrium with nucleotide distance was evaluated for each locus. Figure 4.2 summarizes the patterns of linkage disequilibrium for r^2 between all pairs of segregating sites within each locus (excluding *Aly*13). *Aly*3, *Aly*8sk, *Aly*10sk1 and *Aly*10sk2 S-domain types all exhibit decays of linkage disequilibrium as measured by the correlation in variants between pairs of sites, with respect to the nucleotide distance separating them. The result is consistent with recombination and consistent with these sequences being single loci apart from locus *Aly*7. The *Aly*7 locus consists of two haplotypes, one of which appears to be out of frame (Chapter 3). Linkage disequilibrium reveals that the two haplotypes do not recombine (Figure 4.2), or at least no decay in linkage disequilibrium as sites get further apart is apparent, suggesting that these two haplotypes are two loci. When only *Aly*7-haplotypes are analysed, a decay in LD is observed (not shown). *Aly*9 does not exhibit this same negative decay with nucleotide distance. Due to the low diversity and large number of singletons, there are few pairwise comparisons (5 segregating sites and consequently 10 pairwise comparisons) where linkage disequilibrium can be calculated. Nevertheless, sites at the extreme ends of the sequences exhibit high linkage disequilibrium. This suggests that recombination is lower at the *Aly*9 locus relative to the other S-domains.

The minimum number of recombination events was determined by the algorithm in Hudson and Kaplan (1985) and are shown in Table 4.2. All S-domain types with the exception of *Aly*8sk and *Aly*13 exhibit one or two minimum number of recombination events. Using this approach, we can detect at least eight recombination events in the evolutionary history of *Aly*8sk. The higher number of informative segregating sites has perhaps increased our ability to detect recombination events at this locus. The rate of recombination relative to the mutation rate per gene is also shown (*C*/*θ*).

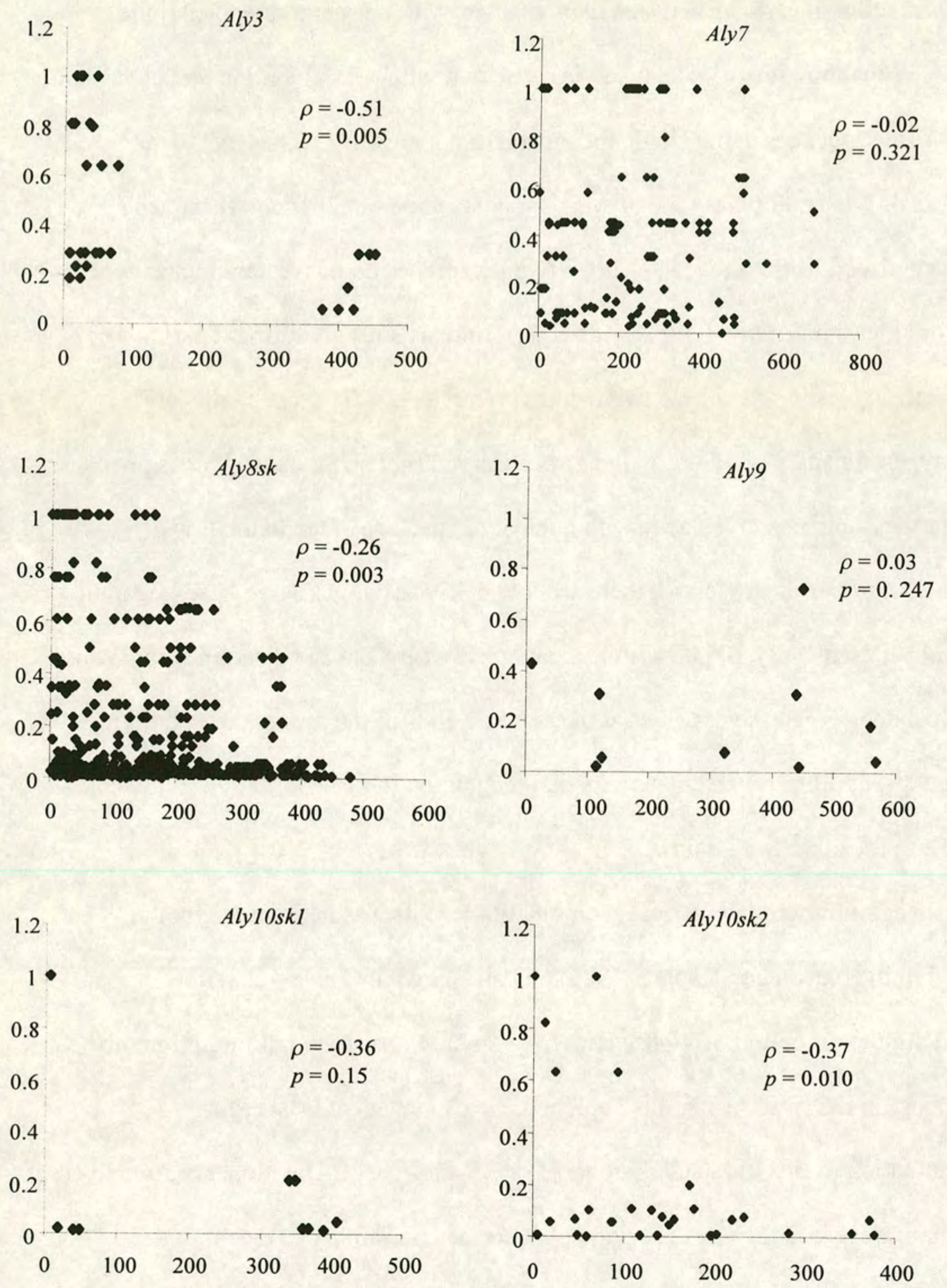


Figure 4.2. Pairwise linkage disequilibrium r^2 plotted against nucleotide distance for both synonymous and nonsynonymous variants at each locus, excluding the *Aly*13 variants.

Table 4.3. Minimum recombination estimates (*Rm*) and estimates of recombination rates per gene (*C*, Hudson 1987) for the S-domain loci.

Locus	*Rm*	*C*/gene	θ/gene	C/θ	distance*
*Aly*3	1	7.5	4.63	1.62	661.80
*Aly*7	1	54.1	5.79	9.34	827.05
*Aly*8sk	8	29.9	13.51	2.2	579.00
*Aly*9	1	10.0	4.04	2.47	849.33
*Aly*10sk1	2	3.3	2.03	1.63	822.37
*Aly*10sk2	2	55.4	8.2	6.76	728.74

* average nucleotide distance between the most distant sites.

4.3.5 Between-locus and species divergence comparisons McDonald-Kreitman test analyses

Members of the S-domain gene family not involved in SI may be under alternative forms of selection. Positive selection subsequent to gene duplication events may have contributed to the evolution of the gene family. Furthermore, we may not be able to assume that the remaining S-domain loci are evolving neutrally or at least are not linked to a locus under selection. A conservative test of neutral evolution between pairs of loci is the McDonald-Kreitman test, treating the loci as one would treat orthologous genes from different species (e.g King 1998). We performed McDonald-Kreitman tests on all pairs of loci excluding the *Aly*13 variants and including only *Aly*7 haplotypes which appear in reading frame. One pair of loci, *Aly*8sk and *Aly*10sk2 exhibited a marginally significant deviation from neutrality in the number of nonsynonymous and synonymous fixed differences relative to polymorphisms (Table 4.4). However, correcting for multiple tests, the result is not significant (Bonferroni, P > 0.05). Regardless, this contigency table reflects either an excess of fixed nonsynonymous differences relative to synonymous differences within the phylogenetic history of these two loci, or too little nonsynonymous polymorphism relative to the observed high synonymous diversity at *Aly*8sk. The former hypothesis is consistent with positive diversifying selection (Goodman et al. 1975), and the latter may indicate an increased mutation rate at this locus or region, coupled with purifying selection acting at the amino-acid level.

Table 4.4: McDonald-Kreitman Test results for *Aly8sk* and *Aly10sk2* in *A. lyrata*

Synonymous Substitutions (200 codons compared):			
Fixed differences between Loci:	23	Polymorphic sites:	41
Replacement Substitutions:			
Fixed differences between Loci:	52	Polymorphic sites:	47
Indels			
Fixed between loci:	3	Polymorphic:	0

Fisher's exact test. P-value (two tailed): 0.053126
G test. G value: 4.346 P-value: 0.03709*
(indels not included in the significance test)

4.3.6. Interspecific comparisons within the S-domain gene family

If the high synonymous polymorphism observed at the *Aly*8sk locus is a result of an elevation in mutation rates specific to this locus, synonymous and nonsynonymous divergence between *Aly*8sk and closely related orthologues (see Chapter 3) may also be high. Nonsynonymous and synonymous divergence between orthologues (sequenced in *A.thaliana*) revealed that *Aly*8sk-*Ark*3 comparisons exhibited the smallest *Ka* divergence (0.033 ± 0.009) and the highest *Ks* divergence (0.309 ± 0.059) between orthologues (see Chapter 3). It is possible that the higher *Ks* value between *Aly*8sk and *Ark*3, and the high synonymous diversity at *Aly*8sk reflects an increased substitution rate at this locus in *A. lyrata.*

4.3.7. Testing for Positive Selection at unlinked S-domain genes

Positive Darwinian selection has been inferred in cases where there are significantly detectable differences in the rate of nonsynonymous substitutions relative to synonymous substitutions in a group of sequences (*Ka/Ks* values greater than one). We attempted a phylogenetic approach to detect *Ka* and *Ks* ratios greater than one by estimating *Ba* and *Bs* for each branch of the genealogy. By examining the proportion of nonsynonymous changes relative to synonymous changes along a lineage, we are not constrained to merely looking at pairwise comparisons for two extant taxa. Only S-domain sequences that had a closely related orthologue in *A. thaliana* were used in the analysis. S-alleles of *Brassica* and putative S-alleles of *A. lyrata* were not included as they are likely to have reached saturation within species (Charlesworth and Awadalla 1998, Chapter 5).

For most branches, *Ba* is much smaller than *Bs*, as expected for genes that are subject to purifying selection. In this analysis, 3 branches exhibit *Ba/Bs* ratios larger than 1. However, the above ratios are based on few nonsynonymous or synonymous changes (Figure 4.3). In the next section, we assess the significance of these ratios, and determine whether there is heterogeneity in the accumulation of nonsynonymous relative to synonymous substitutions among branches using maximum likelihood approaches (Yang 1996).

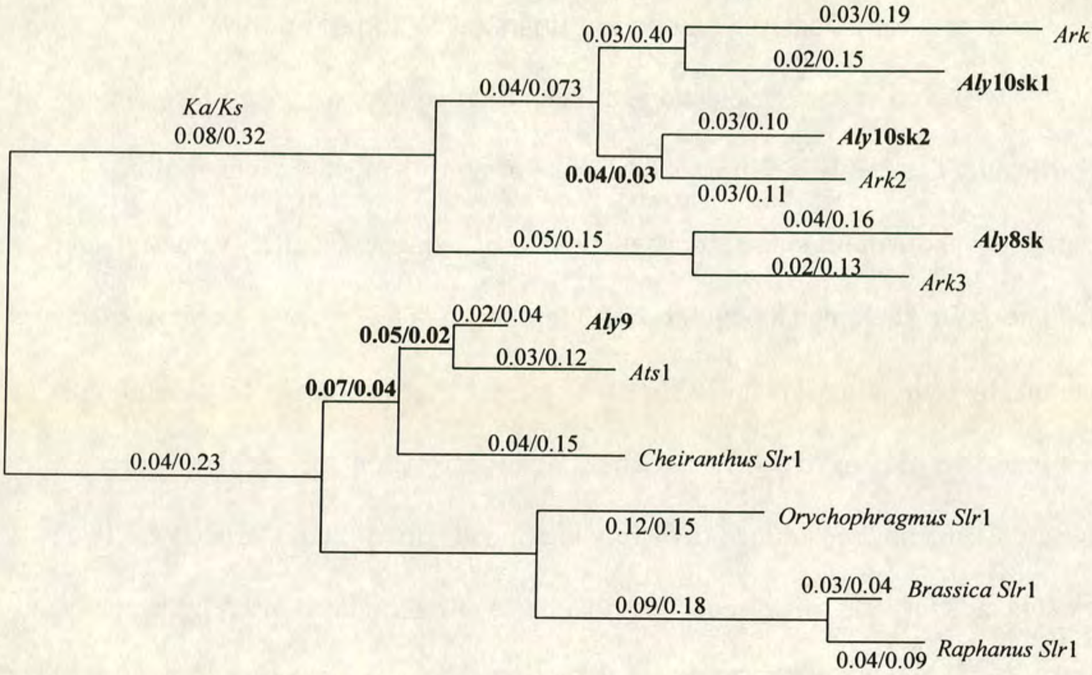


Figure 4.3: Neighbour-joining genealogy displaying branch length (*Ba/Bs*) estimates. In bold are shown the few *Ba/Bs* estimates which are slightly larger than one. As well, in bold are the *A. lyrata* loci sequenced in this study. Not included in this analysis are pairwise comparisons involving *Aly*3, *Aly*7 and *Aly*13.

4.3.9. Maximum likelihood approaches to testing substitution rate heterogeneity and changes in selective constraint among lineages

To assess whether some S-domain loci are evolving at a faster rate, we used the maximum likelihood approaches of Yang and Neilson (1997, 2000) and Yang et al. (2000). By using maximum likelihood methods of phylogenetic inference (Yang 1997), we can test alternative models for the evolution of the gene family. Such tests are similar to the relative-rate test for detecting heterogeneity in the rate of evolution between species pairs by comparison to an outgroup species. The one difference is that by using both synonymous and nonsynonymous substitutions in the same analysis, we can detect heterogeneity among all lineages (without an outgroup). In the tests presented we consider whether a significantly better fit to the data is achieved by relaxing the assumptions of (1) clock-like accumulation of synonymous and nonsynonymous substitutions and (2) homogeneity in the ratio of nonsynonymous to synonymous substitution rates across lineages. Relaxing the first assumption is equivalent to allowing between-lineage variation in the mutation rate but maintaining selective constraint; relaxing the second assumption allows for differences in selective constraint, but assumes a constant rate of synonymous subsitution.

For the six S-domain loci unlinked to SI in *A. lyrata*, relaxing the molecular clock across lineages, as well as relaxing both the molecular clock and *Ka/Ks* ratios provided a better fit to the data than a molecular clock model with fixed *Ka/Ks* ratios for each lineage (Table 4.5). Yet, relaxing only *Ka/Ks* values along lineages was not significantly better than relaxing the molecular clock alone, nor was relaxing both parameters relative to just relaxing the molecular clock. This implies that there are not detectably significant differences in selective constraints among lineages, but that there are significant differences in mutation rates among lineages. The difference in likelihoods between the clock model with fixed *Ka/Ks* ratios and no clock but fixed *Ka/Ks* is 10.95 and is significant ($p < 0.005$, df = 8).

4.3.10. Heterogeneity in selective constraints among sites within S-domain loci

In Figure 4.2 and Table 4.2, we presented evidence that polymorphism at the unlinked S-domain loci does not appear to be concentrated in the regions previously designated as hypervariable in *Brassica SLG* and *SRK* loci. On the other hand,

Table 4.5. Log-likelihood-ratio tests of variation in the rate of molecular evolution and the *Ka/Ks ratio* among lineages in the S-domain gene family in *A. lyrata*. Values shown are pairwise differences in log-likelihood between models.

Model	no clock, fixed *Ka/Ks*	clock, relaxed *Ka/Ks*	no clock, relaxed *Ka/Ks*
clock, fixed *Ka/Ks*	10.9468*	6.67459	16.25873**
no clock, fixed *Ka/Ks*		-4.27221	5.31193
clock, relaxed *Ka/Ks*			-4.27221

* significantly better fit than the model listed on the left column (row headings). Negative values imply that the model listed in the column headings provide a worse fit to the data than the model listed in the row headings.

estimates of synoymous and nonsynonymous divergence at the HV regions suggest that there is heterogeneity in subsitution rates. If nonsynonymous divergence is higher between paralogous comparisons, then this could suggest relaxed selective constraint for these regions. The pattern of synonymous and nonsynoymous nucleotide divergence is illustrated in Figure 4.4 where *Ka* is plotted against *Ks* for each paralogous pairwise comparison separately for the HV regions and the conserved regions (orthologous comparisons are excluded in this figure). In the HV regions, nonsynonymous mutations accumulate at a faster rate relative to synonymous mutations. Synonymous divergence appears relatively constant across the exon. Therefore, it does not appear that there has been an increased mutation rate in the HV regions, which in any case would seem unlikely as the HV regions consist of small dispersed regions across the S-domain. The pattern either reflects relaxed selective constraint or positive selection. The estimate of the shape parameter (alpha) for the gamma distribution of substitution rates was estimated to be 0.343 which suggests heterogeneity in the distribution of substitution rates across the sequences among the different S-domains (Yang 1997) which may merely reflect more nonsynonymous substitutions (see below).

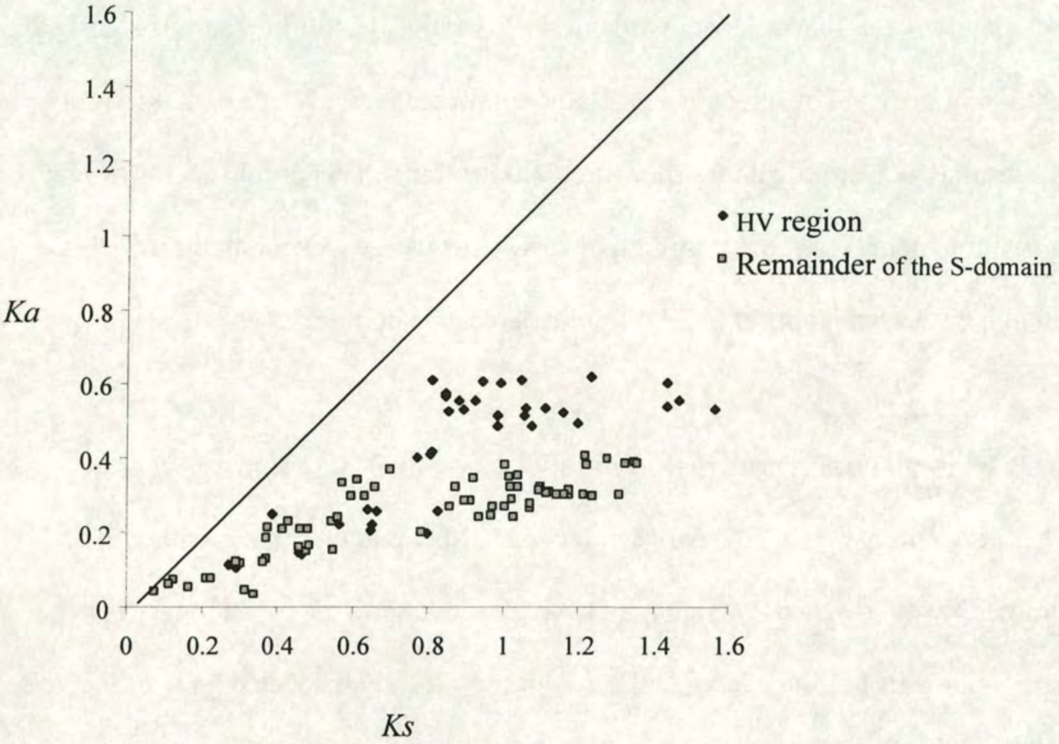


Figure 4.4: Pairwise estimates of synonymous and nonsynonymous Jukes-Cantor divergence for 'HV' regions and the remainder of the locus between pairs of paralogues in *A. lyrata* (excluding *Aly*3, *Aly*7 and *Aly*13) including putative orthologues from *A. thaliana*.

We implemented a number of different statistical distributions for heterogeneous *Ka/Ks* ratios among sites (Yang and Neilson 2000). Two major objectives of fitting these models are to test for the presence of positively selected sites and to determine the distribution of *Ka/Ks* ratios. This test is conservative in that a site must be under positive selection in every lineage (Yang et al. 2000).

Using the same data as that used to test for deviations from the molecular clock, we found no evidence for *Ka/Ks* values significantly greater than one at any one site. A model which allowed for *Ka/Ks* values among sites greater than one (M8 in PAML) was not significantly different than a model which allowed *Ka/Ks* values to range from zero to one only (M7). The difference in log-likelihood was –1.142. Nevertheless, a model which allowed for variable *Ka/Ks* ratios fit the data significantly better than a model with one *Ka/Ks* ratio. Therefore, there is significant variation in the level of selective constraint within the loci, even though different lineages may not significantly differ (previous section). This is consistent with the observation that HV regions exhibit higher *Ka/Ks* divergence among these loci, although *Ks* does not appear to be different across the domain.

4.3.11 Potential gene conversion among *A. lyrata* S-domains

Recombination events between loci can affect polymorphism within loci. To examine the role gene conversion may have played among the evolutionary history of the S-domains not linked to SI in *A. lyrata*, we visually inspected pairs of loci to search for 'tracts' or small contiguous segments that appear to have been exchanged among loci. We used the method of Betran et al. (1997) to assess the significance of these tracts based on their model of gene conversion. Their method assumes that a tract of shared polymorphism in linkage disequilibrium is a significant conversion event if it meets the following criteria. A segregating nucleotide is considered informative if its relative frequency in the "recipient" group of sequences or locus (those loci which receive the tract) is 20% or less and its relative frequency in the "donor" (the loci from which the fragment originated) group of sequences is three or more times higher than in the "recipient" locus. The two outermost informative sites determine the length of the observed conversion tract. Conversion tracts of 1 bp in length are not considered because they cannot be distinguished from parallel mutation events (Stephens 1985). These criteria are derived from the results of simulations that define this minima (Betran et al. 1997).

Figure 4.5 describes the tract polymorphism among six of the S-domain loci. The darker shaded regions indicate tracts detected by the test criteria described above and ignoring the potential orthologous sequences. Light shadings describe regions that do not fit the Betran et al. (1997) criteria but could be possible converted tracts. *Aly*8sk appears to share segments of sequence similarity with a number of S-domain loci (as is underlined) in a region spanning nucleotide positions 618 - 642 which also happens to correspond to the HV1 of *Brassica SLG*/*SRK*.

However, when the putative *Ark* orthologues are included in the comparisons, the rarer haplotypes appear to reflect the ancestral condition at each locus, in that the rarer haplotype is shared with the respective orthologue in *A. thaliana*. The high frequency variants at positions noted above at the Aly8sk locus appear to be the mutated state. Whether these haplotypes reflect conversion events, shared polymorphisms that have been maintained over time, or new mutations, is not possible to determine.

Figure 4.5 (next page): Low frequency 'tracts' found among loci in *A.lyrata* that are either fixed or found at high frequency at other loci. Shaded boxes are the low frequency variants at the 'recipient' locus and the underlined segments show the high frequency or fixed variant at a possible 'donor' alternative locus. This alignment also includes *Aly*13 variants although they were not examined for conversion events as they are too polymorphic.

	3333333333333333333333	6666666666666666666666666
	6777777777788888888889	1122222222223333333333444
	9012345678901234567890	8901234567890123456789012
97F8-2Aly3	----------------------	CAAACACGAA------------ACA
97F8-1Aly9	GAAACAGACGAGA---AAGGAG	CAAACACGAA------------ACA
B20(3)Aly8sk	----------------------	AAAACAAGAA------------GCA
B17(2)Aly8sk	....	....
A5(4)Aly8sk	....	....
B17(2)Aly8sk	....	....
B21(1)Aly8sk	....	C.....C...............A..
99A15-1Aly8sk	....	......................A..
99A15-1Aly8sk	....	....
99A17-1Aly8sk	....	....
43-1Aly8sk	....	......................A..
40-7Aly8sk	....	....
43-1Aly8sk	....	....
15-2Aly8sk	....	....
98I33-1Aly8sk	....	......................A..
98J32-7Aly8sk	....	....
98J32-7Aly8sk	....	......................A..
98E17-15Aly8sk	....	....
97F12-4Aly8sk	....	....
97F12-4Aly8sk	....	C.....C...............A.C
98E17-15Aly8sk	....	C.....C...............A.C
99A9-1Aly8sk	....	C.....C...............A.C
99A7-1Aly8sk	....	....
99A8-1Aly8sk	....	C.....C..G............A.C
99A8-1Aly8sk	....	....
99A1-1Aly8sk	....	....
99A1-1Aly8sk	....	....
99A2-1Aly8sk	....	....
Aly7		
A9(4)Aly7	GAGAAAAACAGAGGAAAGAGAG	ACAAAACGAAGCAGCACGAAGAACA
A9(4)Aly7	....	...........A.............
98I32-7Aly7	....	....
98E15-15Aly13-3	GAAACAGACAGGAGCA---GAG	AAAACACGAAAAGCCA------AGG
97F3-4c18Aly13-3	....	............A............
699A2-1c11Aly13-3	.....C........CA......	............A............
99A17-1cl39Aly13-3	..............CA......	............A............
40/7c2Aly13-9	......A...ACA.G.AAA..A	C........C---A........CAA
99A9-1cl28Aly13-13	C.....A..G.CA.AG......	C...A......CA---......---
12-4Aly10sk2	CAAAAAAACCGGAGGAGAAGAA	AAAACACCGA------------ACC
17-4Aly10sk2	....	..C......................
99A7/1Aly10sk2	....	....
99A9/1	....	....
97F15-2Aly10sk2	....	....
97F15-3Aly10sk2	....	....
99A1/1Aly10	....	....
99A2/1Aly10	....	C........................
A5(4)Aly10sk2	....	....
A9(4)Aly10sk2	....	....
B17(2)Aly10sk2	G....................G	....
B20(3)Aly10sk2	....	....
B21(1)Aly10sk2	....	....
99A15/1Aly10sk2	....	....
99A17/1Aly10sk2	....	..C......................
98I32-710sk2	G....................G	....
98I33-3410sk2	G....................G	....
43/1Aly10sk2	G....................G	....
40/7Aly10sk2	....	....
Ark1	g....................A	a.....c...............a.c
Ark2	g....................A	A.....c...............a.c
Ark3	g....................A	c.....c...............a.c
Ats1	g...c....gc..........g	c.....g...............a.c
12-4Aly10skB	GAAAAAAACCGGAGGAGAAGAG	AAAACACGAA------------ACA
17-4Aly10skB	....	....
17-310skB	....	....
99A-7/1Aly10skB	....	....
99A-8/1Aly10skA	....	....
99A-8/1Aly10skB	....	....
99A-9/1Aly10sk	....	....
13-510skA	....	....
15-210skA	....	....
15-210skB	....	....
15-310ska	....	....
15-310skB	....	....
15-810skA	....	....
99A-1/1Aly10sk	....	....
99A-2/1Aly10sk	....	....
26A2x10sk	....	....
A5(4)10skA	....	....
A9(4)10skA	....	....
B17(2)10skA	....	....
B20(3)10skA	C.....................	....
b21(1)10skA	C.....................	....
99A-15/1Aly10sk	C....A................	....
99A-17/1Aly10skB	....	----------............---
40/7Aly10skA	....	....

4.4.1 Discussion

Nucleotide diversity and divergence among those members of the S-domain loci that do not segregate with the SI phenotype exhibit low to moderate levels of polymorphism in *A. lyrata* compared to the S-genes in the Brassicaceae. These S-domain loci exhibit nucleotide diversity estimates similar to published estimates for other loci in this species (e.g. Savolainen et al. 2000) and certainly larger within-population estimates relative to estimates for selfing *A. thaliana* populations (e.g. Miyashita et al. 1997). In contrast, a group of sequences (*Aly*13) that includes variants that segregate with SI in *A. lyrata* has extreme levels of polymorphism. All the genes unlinked to SI, apart from *Aly*9, appear to be recombining, and although there is some evidence for concerted evolutionary processes, it is impossible to determine whether exchanges between genes are occurring for certain. Nevertheless, it is possible that these processes may have contributed to the high synonymous polymorphism observed at the *Aly8sk* locus. Divergence between *Aly*8sk and *Ark*3, the most closely related available orthologue in *A. thaliana*, revealed higher synonymous divergence relative to nonsynonymous divergence. It is plausible that a higher mutation rate at the *Aly*8sk locus relative to the other S-domain loci may be contributing to the high diversity at this locus. Finally, orthologous divergence reveals low *Ka/Ks* ratios, consistent with purifying selection, between pairs of unlinked S-domain loci where an orthologue is available. Internal branch lengths of the gene family genealogies reveals significant heterogeneity among branches in substitution rates, and although *Ka/Ks* branches internal to the tree tend to be larger relative to the branches at the tips of the genealogies, there does not appear to be significant heterogeneity in selective constraints among lineages. In contrast, parts of the gene exhibit significantly different levels of selective constraint, with amino acid mutation rates being higher in parts of the sequence.

4.4.2 Neutral evolution at the S-domain gene family loci

Ohta (1994) suggested that an acceleration of amino acid changes between duplicated genes in conjunction with functional differentiation is evidence of positive selection. For the S-domain loci, the total number of fixed differences at amino acid changing sites is not greater than the total number of fixed differences at synonymous sites among all pairwise comparisons. However, McDonald-Kreitman tests revealed that ratios of nonsynonymous to synonymous variation for both fixed differences and polymorphism showed marginal evidence for either adaptive amino acid divergence between *Aly*8sk and *Aly*10sk2 or evidence for accelerated mutation and purifying selection at the *Aly*8sk locus. The latter seems more likely given 1. the high polymorphism at the *Aly*8sk locus relative to other S-domain loci in *A. lyrata* and 2. the somewhat higher synonymous divergence between *Aly*8sk and *Ark*3, the most closely related sequenced orthologue. It is plausible that rather than a mutation rate that is specifically high for this locus, the *Aly*8sk locus may be linked to a separate locus under balancing selection (Nordburg et al. 1996). Furthermore, alternative mechanisms may be contributing to the high polymorphism and divergence at this locus such as between-locus gene conversion.

Examining HV regions separately for the S-domain loci suggests that these regions are under relaxed selection constraint. It does not appear that there has been an increased mutation rate in the HV regions, as synonymous divergence is relatively similar across the locus, and in any case would seem unlikely as the HV regions consist of small dispersed regions across the S-domain (see Chapter 5).

Elevated levels of nonsynonymous substitution may be an indication of positive selection, even if *Ka/Ks* values do not exceed one. The criterion that the average *Ka/Ks* ratio be greater than one is very stringent. Adaptive evolution most likely occurs at a few time points and at a few amino acids. Over time, the footprints of adaptive evolution may be lost through mutation. By examining branch lengths within a genealogy, we partition time in order to capture the point in time at which positive selection may have occurred, avoiding the homogenizing effects of purifying selection and constraint. A problem with examining the entire sequence is that not all the sites will be under positive selection, in fact most sites are either evolving neutrally or are slightly deleterious. This will mask the effects of positive diversifying selection that has occured in the distant past by minimizing Ka/Ks ratios. Furthermore, this entire approach suffers from a lack of independence, assumes normally distributed mutation rates, and is ad hoc (Yang and Nielsen 2000).

The observation that *Ka/Ks* ratios are larger for some internal branches, relative to branches at the tips of the tree, is an indication that changes in selection at nonsynonymous positions occurred in the history of this gene family. There are two ways to interpret this observation, the first being that positive diversifying selection is operating (Ohta 1994). The second interpretation is that, subsequent to gene duplication, these loci evolved neutrally for some amount of time and nonsynonymous mutations were allowed to accumulate as if they were neutral. At some point, these genes may have acquired a new function, indicated by the high constraint (low Ka/Ks values) observed at external branches. However, likelihood models showed that there was no significant heterogenetiy among branches in selective constraint. The heterogeneity among branches appears to be almost entirely due to a deviation from the molecular clock.

4.4.3. Gene conversion

There was no clear evidence for gene conversion events between homologous loci unlinked to SI (but see Chapter 5 for loci linked to the SI response). The observation of tracts polymorphic in one locus, but fixed at another locus, would normally indicate gene conversion events. However, the sequences of the *Ark* genes in *A. thaliana* show us that it is difficult to determine the ancestral state for many of the polymorphic sites that were implicated in these conversion events. The 'donated' fragments within a locus which are at low frequency, may merely be an artifact of sampling, in that we have sampled predominantly one portion of the genealogy which happens to have the mutated state. It is evident that there is clear spatial structure due to identity by descent. Furthermore, two mutations may have occurred on the same segment of the coalescent, early in the history of the population, by chance. Of course, this would lead to their higher frequency in the population, and appear as converted segments due to their linkage.

The linkage disequilibrium patterns for low heterozygosity locus *Aly*9 is interesting in terms of gene conversion. We see no correlation between linkage disequilibrium and nucleotide distance. In fact, we observe very little linkage disequilibrium, in contrast to *Aly*7, for example. The patterns at *Aly9* are similar to that expected for a locus under gene conversion. Tracts of small length will break up linkage disequilibrium, but depending on their frequency and size, a relationship between nucleotide distance and linkage disequilibrium may not be observed (Andolfatto and Nordborg 1998).

Examining the aligned sequences for the different *A. lyrata* S-domain loci unlinked to SI reveals that a number of the same sites are polymorphic in more than one S-domain locus. The probability that a set of loci share polymorphic sites increases with the number of loci compared and the proportion of segregating sites relative to the size of the region compared (Clark 1997). A polymorphic site may be segregating in the ancestor and extant lineages either due to balancing selection or relaxed selection (neutral). The former would seem unlikely if the loci have diverged and perform very different functions. It is possible that particular sites or regions are more susceptible to mutation within or across loci (e.g. a region of relaxed selective constraint). These shared polymorphisms may also be the result of gene conversion events occurring across loci (Clark 1997).

Chapter 5

Molecular Evolution of Loci linked to Self-Incompatibility in *Arabidopsis lyrata*

5.1.1 Introduction

Sequences of the *SLG* and *SRK* loci from different haplotypes have revealed that these S-domain loci in *Brassica* exhibit extremely high levels of amino-acid and nucleotide variability (Hinata *et al*. 1995). The high levels of *SLG* and *SRK* nucleotide diversity within *Brassica oleracea* and *campestris* species are comparable to estimates for the *RNA*se-like loci controlling gametophytic SI, and the *MHC* loci, both thought to be under balancing selection (Chapter 1, Charlesworth and Awadalla 1998). However, the level of polymorphism at loci controlling sporophytic SI, among individuals sampled randomly from natural populations remains unknown. We have isolated S-domain variants, called *Aly*13, that include sequences in which variants segregate with SI in *A. lyrata* (Chapter 3). In this chapter, I will investigate how variation for this group of sequences is distributed among individuals and genes sampled from four populations. By assessing how this variation is structured, we can then compare the level of polymorphism at *Aly*13 in *A. lyrata* with *Brassica SRK* variation, as well as with paralogous members of the gene family, and address whether the hypervariable regions observed at the *Brassica* S-loci are the result of relaxed selection pressures, balancing selection, or both (Chapters 1 and 2, Awadalla and Charlesworth 1999).

5.1.2. Expected Diversity Within- and Between-Incompatibility S-alleles

The *SLG* and *SRK* sequences in *Brassica* were each sequenced from a separate functional S-phenotype (Kusaba *et al.* 1997). For a random sample of self-incompatible plants from a population, even though the number of alleles in a population is high, we expect to sample the same S-allele more than once. As a result, the expectation for the level of nucleotide diversity in a random sample will differ from that expected for a sample of sequences where each sequence represents a different incompatibility group (Chapter 1, Charlesworth *et al.* 2000). The expectation of diversity within and among different phenotypes for loci under balancing selection can be described using the coalescent (Takahata 1990, Vekemans and Slatkin 1994). Takahata (1990) showed that the shape of the coalescent at a locus under balancing selection is the same as that for a neutral locus, except that the branch lengths are multiplied by a scaling factor. This is true when the sample consists only of alleles of different phenotypes, for example, functionally distinct incompatibility alleles or MHC alleles specific to different serotypes. This scaling factor is proportional to the strength of selection acting at the locus. Vekemans and Slatkin (1994) further extended this coalescent model applying it to gametophytic SI. The result of this scaling factor is that branch lengths are lengthened, and mutations will accumulate along these extended lineages, contributing to the high level of polymorphism we observe at S-loci (Figure 5.1).

As mentioned, a random sample will include alleles of the same incompatibility class. The sample of sequences within an incompatibility class have the same statistical properties as those for a small neutrally evolving population or deme (Figure 5.1). There may be mutations within an incompatibility class that will be effectively neutral with respect to the incompatiblity phenotype. Therefore, sampling one sequence from each S-allele, as has been done in *Brassica*, is similar to sampling one individual from each of a number of separate populations. The level of neutral (synonymous) nucleotide diversity within each incompatibility group will depend on the 'effective size' of that incompatibility group in a population. In Figure 5.1, we show the expectation for the level of nucleotide diversity between and within functional self-incompatibility lineages. The effective size of each of these small 'populations' of incompatibility sequences sharing the same phenotype depends on the number of different 'S-alleles' in the population (Vekemans and Slatkin 1994). The coalescent time for alleles within a functional allelic class is expected to be smaller than a neutral locus, because the average number of copies of genes within a functional S-allele class is $2Nt/n$, compared with $2Nt$ for a neutral locus, where n is the number of S-alleles in the population and t is the time to coalesence within the the population. Recombination between S-alleles will then behave like migration between populations. The effect of recombination on two S-alleles is that they will exchange variants that have evolved on separate lineages. As a result, the lineages will become less distinct from each other (i.e. Hudson 1990). Whether such lineages which have undergone recombination remain functional, or whether they now recognize two different pollen S-proteins, is unknown (Charlesworth *et al*. 2000).

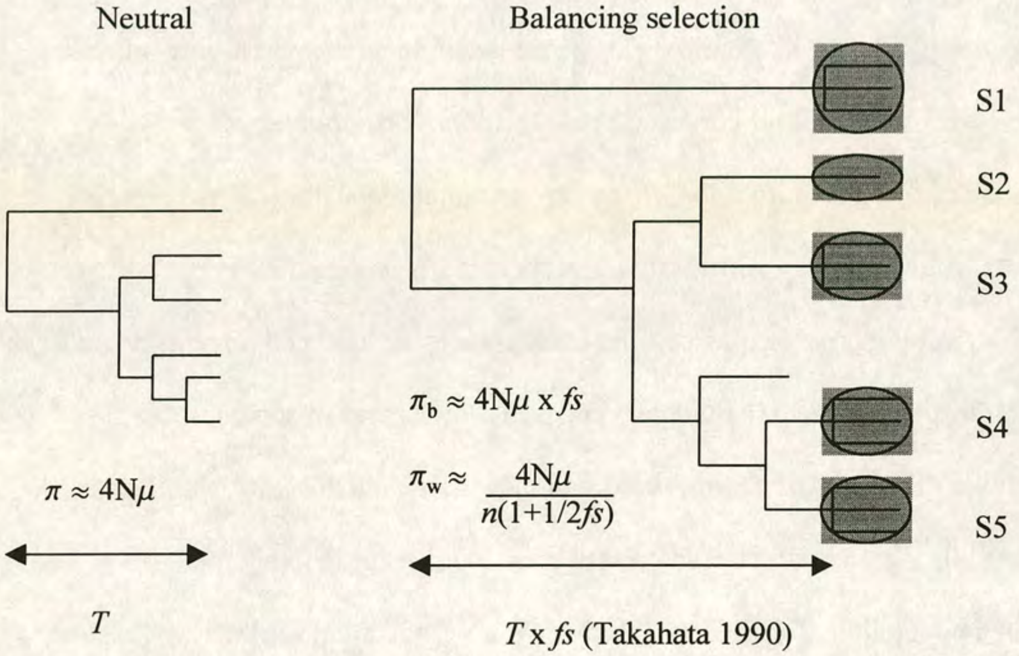

Neutral
Balancing selection
S1
S2
S3
S4
S5
$\pi_b \approx 4N\mu$ x fs
$\pi \approx 4N\mu$
$\pi_w \approx \dfrac{4N\mu}{n(1+1/2fs)}$
T
T x fs (Takahata 1990)

Figure 5.1. The effects of balancing selection on a genealogy of alleles. The left-hand genealogy describes alleles evolving neutrally. The right-hand figure indicates the genealogy for balancing selection, scaled by the strength of selection. The grey rectangles on the right indicate groups of individuals (demes) having the same functional S-allele (n = 5 functionally distinct S-alleles in this sample). The genealogy has been rescaled by the scaling factor (*fs*) which is directly correlated with the strength of selection. The expectations for the level of pairwise nucleotide diversity between and within incompatibility classes is π_b and π_w (Takahata 1990).

Each of the variants of the *Brassica SLG* and *SRK* loci has been sequenced from different S-phenotypes (i.e. Kusaba *et al*. 1997). Samples that include only one sequence per incompatibility type or allele will affect various summary statistics such as mean pairwise diversity or Tajima's *D* estimates (Tajima 1989) in a number of ways. Tajima's *D* summarizes the frequency spectrum of variation at a locus by comparing the number of phylogenetically informative sites relative to singletons (Tajima 1989). A locus under balancing selection would have a positive Tajima's *D* statistic, indicating an excess of highly heterozygous sites, relative to unique polymorphisms (singletons). In the case of SI, observing a positive estimate of Tajima's *D* is dependent on the frequency of each incompatibility class in a population. For example, assume that there are four incompatibility types at equal frequency in a population. If we assume that nucleotide differences between these alleles is high, as for SI, a large number of variants will be 'fixed' between incompatibility classes but shared by individuals with the same incompatibility type. As a result, Tajima's *D* will be positive. In contrast, if one of the S-alleles is at a high frequency, while others occur only once in the sample, we will observe a larger proportion of singleton variants. For forty-two *Brassica SLG* sequences, we observe a slight excess of singletons, but not statistically significant, relative to informative segregating sites as a result of sampling one individual per S-allele (Tajima's $D = -0.31$, $P > 0.10$). This slight excess may merely reflect fixed differences between different functional self-incompatibility alleles.

5.1.3. Effects of Population Subdivision on Loci Subject to Balancing Selection

In subdivided populations, loci subject to balancing selection should tend to exhibit less population structure than neutral loci (Wright 1939). In the case of frequency-dependent selection for self-incompatibility alleles, rare alleles arising in a population either through mutation or migration will be subject to strong selection, and will increase in frequency in the population until an equilibrium is reached. Therefore, the strong deterministic forces affecting S-allele variation will outweigh the stochastic processes of population subdivision. As a result, for loci sampled from individuals among subdivided populations, estimates of population subdivision (Wright 1951) should be lower for self-incompatibility alleles compared to neutrally evolving loci (Wright 1939).

One application of empirical SI data has been to use the genealogy of alleles to infer population genetic parameters in the distant past (ie. Richman *et al.* 1996). But to do this, we need to understand how population structure affects the dynamics of S-allele variation. To examine the effects of population subdivision theoretically, Schierup *et al.* (in press) simulated loci under various models of balancing selection among subdivided populations. For such loci, measures of population subdivision (e.g. *Gst* – a coefficient estimate of the normalized difference in species wide diversity relative to within population diversity (Nei 1987) increase with decreasing migration rates, but remain smaller than results for neutral loci. Schierup (1998) and Schierup *et al.* (in press) also showed by simulations that population subdivision decreases the overall number of functional alleles maintained in a species. This observation contrasts with the conclusions of Wright (1939), whose model showed that subdivision had very little effect on total polymorphism for a locus under balancing selection. Muirhead and Slatkin (in press) derived analytical expressions for the expectation of the number of alleles in a sub-divided population and showed that population subdivision reduces the total number of alleles in a population. Muirhead and Slatkin (in press) also determined the expected number of alleles shared for both a two-deme and a ten-deme model and showed the number of alleles increases with migration rates. With low migration rates (the order of 10^{-6}), the total number of alleles can still be quite large but the proportion of shared alleles will be much lower than with higher migration rates (on the order of 10^{-4}).

5.1.4. Sequence Variation among Dominance Classes in Sporophytic SI

Most *Brassica SLG* and *SRK* variants have been sequenced from a specific dominance class of S-alleles known as 'Type I', which include alleles codominant among themselves but dominant to another class of alleles called 'Type II' (see Chapter 1). Therefore, only variation for a subset of the *SLG* and *SRK* alleles in *Brassica* is available. This sampling of only a subset of the S-alleles is a result of using primer pairs that predominantly amplify Type I S-alleles. The high number of nucleotide differences separating Type I from Type II S-domain sequences is presumably the cause of poor annealing of these primers to Type II haplotypes (Figure 1.9, Chapter1; Chen and Nasrallah 1990) For example, primers PS5, PS15 and PS22 (Nishio *et al*. 1997, Sakamoto *et al*. 1998) amplified *SLG* and *SRK* alleles in 36 of 45 haplotypes tested in *Brassica*. Among those incompatibility groups where an *SLG* and *SRK* allele was not amplified were three Type II haplotypes (S2, S5 and S15). Type II S-alleles exhibit low homology to Type I S-alleles (less than 75% amino acid identity, Nishio *et al*. 1996, Kusaba *et al*. 1997) and also exhibit greater similarity to another S-domain gene found in *Brassica* called *SLR2* (greater than 90% homology, Chen and Nasrallah 1990, Kusaba *et al*. 1997, Nishio and Kusaba 2000). As a result, it is difficult to amplify Type II *SLG* S-alleles specifically without amplifying either the *SLR2* gene or other members of the gene family (Nishio and Kusaba 2000).

In *A. lyrata*, we do not know if dominance classes exist among S-alleles. In Chapter 3, we showed some evidence for dominance among S-alleles segregating in four *A. lyrata* families. In our initial attempts to isolate S-domain loci from *A. lyrata*, it was necessary to design primers that were relatively 'degenerate' (Chapter 3), and would amplify a large number of S-domain sequences in the hope of amplifying *SLG* or *SRK* orthologues in *A. lyrata*. Therefore, it seems likely that we could have amplified alleles from different dominance classes in *A. lyrata* as well.

5.1.5. Variation in Polymorphism Within the S-domain Loci

Both silent and amino acid differences were found to be high among many pairs of *Brassica S*-alleles. This high diversity is found throughout the *S*-locus sequences of sporophytic systems, but particularly in certain ("hypervariable" or HV) regions (Chapter 2, Ioerger *et al.* 1990, Dwyer *et al.* 1994, Awadalla and Charlesworth 1999). Whether such heterogeneity is due to stochastic processes where differences in variation among regions merely reflects either differences in rates of fixation or *Ne* among sites (Hudson 1990), or reflects a difference in the level of selection or mutation in particular regions, can be addressed via a number of different approaches (e.g. Hughes *et al.* 1990, Zhang *et al.* 1998). One approach is to assess whether particular regions exhibit patterns of nonsynonymous and synonymous diversity that deviate from the neutral expectation (e.g. Hudson *et al.* 1987, Hughes *et al.* 1990). Another approach compares how polymorphic variation is distributed across the same genes between groups of separate related species (e.g. Smith and Hurst 1998). If patterns of variation are repeatable between species, this suggests that the distribution of variation across the sequence is not random, and that mutational or selective forces have shaped the pattern of variability across the genes.

Regions of high diversity may indicate sites that are under balancing selection. By examining the *SLG* and *SRK* loci from *Brassica* alone, we cannot tell whether HV regions are portions of the genes where amino acid differences determine the incompatibility types, or are neutrally evolving regions (Table 1.1, Chapter 1). For these regions *Ka*/*Ks* ratios do not differ significantly from one, suggesting that they may be evolving neutrally (Chapter 2). Examining S-domain diversity in *A. lyrata* may shed light on whether mutations in these regions affect incompatibility types. For example, if nonsynonymous mutations within incompatibility types are observed at these HV regions, this would suggest that at least these particular mutations within those regions do not change the incompatibility type. If we observe high polymorphism or divergence at members of the S-domain gene family not involved in SI, then this might suggest that mutations at these HV regions are not under balancing selection, but are under relaxed selective constraint.

This chapter describes the characterisation of diversity at loci that appear to be involved in the maternal recognition process of the SI response. In Chapter 3, I described evidence that a group of highly variable S-domain sequences (*Aly*13), initially sampled from five *A. lyrata* individuals, also appear to segregate with incompatibility groups in four *A. lyrata* families. Chapter 4 presented estimates of mean diversity for 39 *Aly*13 sequences isolated from a number of different individuals and populations. Here we describe this variation in detail with respect to how variants are distributed among populations, the levels of diversity within and between putative incompatibility types, and compare the level of variability with estimates for *Brassica* S-loci, as well as with gametophytic SI and the *MHC* locus in humans. Finally we discuss the distribution of variation across the S-domains of *Aly*13 variants compared to substitution patterns for the other members of the S-domain family in *A. lyrata*.

5.2. Materials and Methods

All PCR amplifications, primers and individuals tested were as described in Chapters 3 and 4 and Appendix 1 unless otherwise noted. All PCR amplifications and restriction digests included here were done by M. Schierup and myself. All cloning and sequencing, identification of kinase domains, and analyses of the variants included here were performed by myself. Two or three cloned products were sequenced for each allele to check for PCR amplification errors that occurred before cloning. Consensus sequences were used in the analysis.

Mean pairwise proportions of synonymous substitutions (*Ks*) and nonsynonymous substitutions per site (*Ka*), were calculated for the regions analysed (see below) using MEGA version 1.01 software (Kumar *et al*. 1994) and DNAsp (Rozas and Rozas, 1997) as in Chapters 1-4. Standard errors are shown for pairwise comparisons, but these estimates are susceptible to assumptions about the distribution of mutation rates. For example, infinite sites models which assume a Poisson distribution of mutation rates are likely inappropriate for S-loci (Chapter 2, Yang 1996). Some sites clearly violate the infinite sites model in that more than 2 nucleotide types are segregating (multiple mutation). Furthermore, heterogeneity in mutation rates is evident, therefore expected values for mutation rates at each site will vary. Therefore, conventional standard errors are inappropriate and not shown (but see below; Nei 1987).

To compare variation in polymorphism between different genes, a non-parametric statistical test of repeatability was used. This avoids assumptions about distributions of rates of molecular evolution. Spearman rank correlations were performed to compare polymorphism levels at identical regions of the S-domain between *Brassica* and *A. lyrata* S-domain sequences. For each species, *Ks* and *Ka* diversity for a window of either 25 bp or a single codon in length was calculated, and the diversity in each window was ranked across the gene. Then the two patterns of ranks across the gene were compared using the Spearman rank statistic. In the repeatability analyses, only *Aly*13 alleles (subtypes, see below) were used where we either have segregation data showing linkage to incompatibility groups, or where we know that the *Aly*13 variants have a kinase domain (Chapter 3). In this way, we are more likely to be assessing variation within a single locus. The same subtypes were used to compare the frequency spectrum of site variation at *Aly*13 loci to *Brassica SLG* and *SRK* loci.

5.3. Results

5.3.1. Total Nucleotide Diversity Among *Aly*13 Variants

As mentioned in the previous chapter, the *Aly*13 variants exhibit extreme levels of polymorphism. We show these values among the individuals sampled thus far in Table 5.1. We see that the global estimate of nucleotide diversity does not differ from the within-population estimates (Chapter 4). Synonymous variation is also remarkably high. The estimates shown below are not corrected for multiple substitutions. However, 215 sites have 3 or 4 nucleotides segregating out of 1009 sites compared. Clearly, multiple mutations have occurred at sites among these sequences.

Table 5.1 Uncorrected synonymous and nonsynonymous diversity at the *Aly*13 variants.

Locus	Global $n = 38$	North Carolina $n = 6$	Indiana $n = 10$	Scotland $n = 15$	Iceland $n = 7$
*Aly*13 πs	0.3612	0.2915	0.3486	0.4467	0.3144
πa	0.2000	0.1722	0.2005	0.2306	0.1695

n = number of alleles sequenced.
All the individuals shown in Table 5.2 are included in these estimates.

5.3.2. Distribution of *Aly*13 variants across populations

As the *Aly*13 variants include sequences that segregate and are linked to incompatibility groups among families in *A. lyrata*, we expect variants to exhibit less population structure compared to a neutral locus. Table 5.2 shows the distribution of *Aly*13 sequence variants among samples from the four populations (NC, Ind, Scotland, Iceland). The *Aly*13 variants are classified into 'subtypes' (i.e. *Aly*13-1, *Aly*13-2 ...) according both to segregation analyses (Chapter 3), and by the level of variation found among the sequences assayed. Highly similar sequences sampled from different individuals (usually exhibiting only 2 pairwise nucleotide differences on average within a group, see Table 5.3) are considered likely to be segregating in the same functional incompatibility group or at least are closely related in terms of identity by state. Such sequences are grouped into a sub-type class. For example, the *Aly*13-3 sequence was found to be segregating with incompatibility groups in a number of families described in Chapter 3. When amplifications using the 13F1-SLGR set of primers was used to amplify DNA

Table 5.2 Distribution of subtypes sampled from individuals and families from the 4 populations surveyed here. *Aly*13-3, *Aly*13-7, *Aly*13-9, *Aly*13-13 are known to have kinase domains and are likely allelic of an SRK-type locus. *Aly*13-1 segregates with SI but does not have a kinase domain (M. Schierup, pers. comm.).

*Aly*13-Subtype	Iceland	Scotland	Indiana	North Carolina	Total # per subtype
*Aly*13-1*	98J43/1[1] 98J33/3[1]	b25(3)[1] b17(2)[1]	98E17-11[1] 98E17-4[1] 99A7-1[1] 99A8-1 99A9-1	97F13-5*[1] 97F15-2*[1] 99A1-1	12
*Aly*13-2	98J40/7[1]	b13(8)[1] a9(4)[1] 99A8-1 99A15-1	99A7-1[1] 97F12-4[1]	99A2/1	8
*Aly*13-3*		99A17-1	97F3-4[1] 97F12-3*	98E15-15* 99A2-1	5
*Aly*13-4	98I32/7[1]	b15(3)[1] b13(1)[1]			3
*Aly*13-5		b21(1)[1] b13(8)[1] b13(1)[1]			3
*Aly*13-6		Apstorr16[1]			1
*Aly*13-7	98J43/1[1] 98J33-3[1]				2
*Aly*13-8		Benlui16[1]			1
*Aly*13-9	98J40/7[1]				1
*Aly*13-10		a5(4)[1]			1
*Aly*13-13			99A9-1		1
# of alleles per population	7	15	10	6	38

* Alleles from individuals where segregation with incompatibility groups is known (Chapter 3). [1] indicates sequences which were cloned.

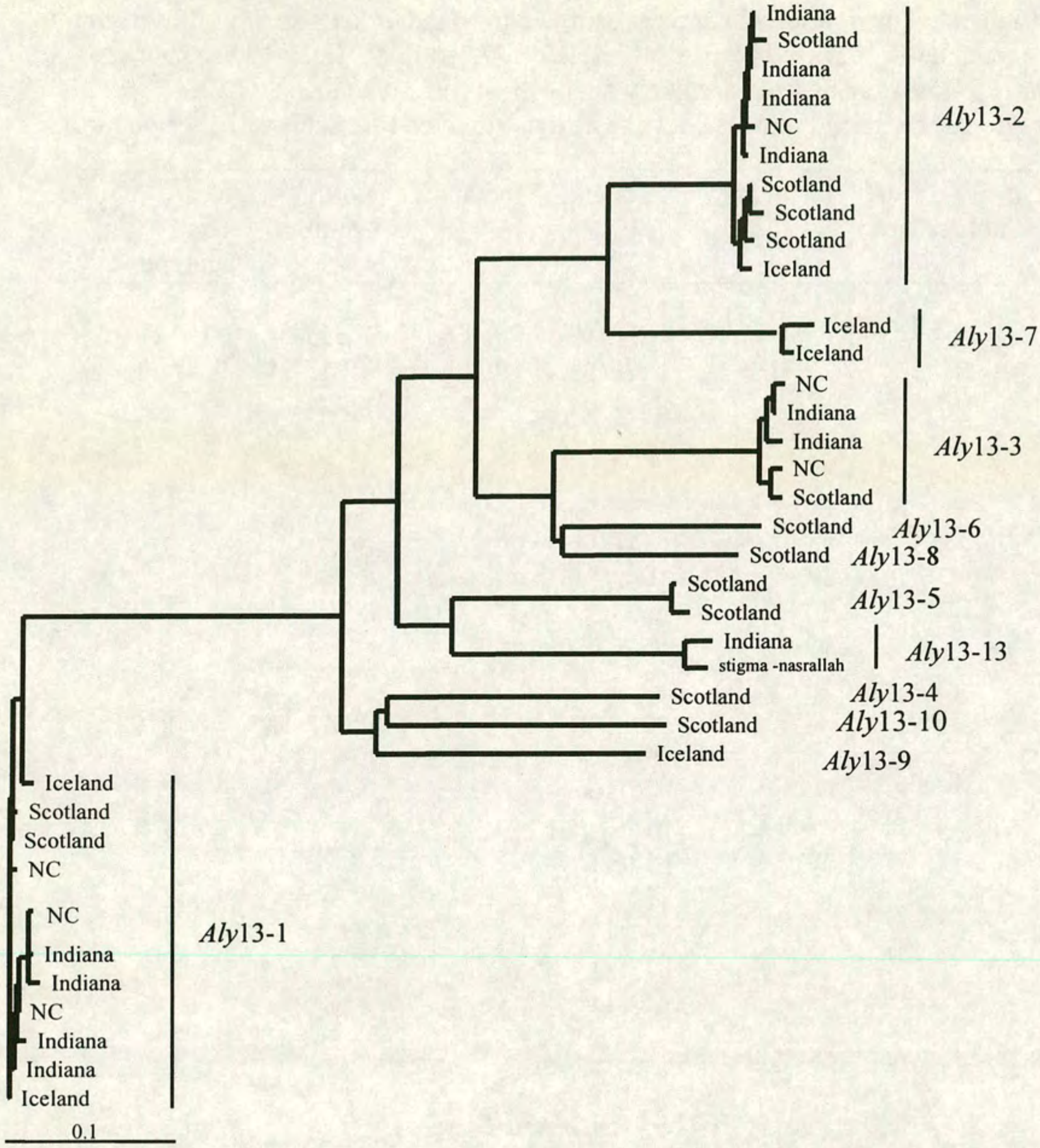


Figure 5.2. Neighbour-joining genealogy for inferred amino-acid sequences of *Aly*13 variants sampled from four populations (Table 5.2). Branch length scale is shown on the bottom. See Table 5.3 and Table 5.4 for mean pairwise diversity within- and between-classes.

fragments from different individuals sampled from the four different populations, a number of individuals had sequences identical to this variant or extremely similar (1 or 2 nucleotide differences). We considered these sequences to be of the same sub-type, although we do not have information verifying that these variants segregate with the same incompatibility group.

It is clear from Table 5.2 and the genealogy of alleles shown in Figure 5.2 indicates that variants within each subtype are distributed among the four populations sampled. This summary of the data indicates that there is little population structure among these alleles. Figure 5.2 shows that the branch lengths separating individuals within subtypes are very short, especially compared to the longer internal branch lengths seen between subtypes. This figure is comparable to the summary of expected variation described in Figure 5.1 (Introduction). However, we can not make calculations of within versus between subtype (incompatibility group) variation as it is still not clear whether within subtype variants segregate with the same functional SI alleles (as mentioned above). Nevertheless, the between subtype branch length differences are striking.

To describe the population structure for the *Aly*13 sequences based on the frequencies of *Aly*13 subtypes found among the four different populations sampled here, more information is required. For example, it is still not clear how many loci are represented in this group of sequenced *Aly*13 variants, or whether all the variants segregate with SI in *A. lyrata*. Furthermore, as mentioned in Chapters 1 and 3, even if we were to assume that pairs of *Aly*13 variants which exhibit high nucleotide distance for the S- domain are associated with different S-phenotypes (the *Aly*13 subtypes), we can not assume that very similar variants (within a subtype) are of the same phenotypic class. There may be variants within these classes, even if the variation consists of only one or two amino acid changes, that are associated with alternative incompatibility phenotypes. Nevertheless, some subtypes were found in each of the four populations sampled and many subtypes were found in only one geographic region. For example, *Aly*13-1 was found among all populations. Nevertheless, estimates of population structure based on heterozygous sites, as opposed to allele counts, within and between populations show no significant population structure for these sequences (*K*st = 0.10, P>0.1; Chapter 4). Little to no substructure is consistent with theoretical predictions for a locus under balancing selection even within a subdivided population (Wright 1939, Schierup 1998, Schierup *et al.* in press).

5.3.3 Within- versus between-*Aly*13 subtype variation

Although we would expect many nucleotide and amino acid differences between a pair of S-specific sequences, we expect very little nucleotide variation within an incompatibility class (see above – Section 5.1.2). In the previous section, we mentioned that a number of individuals share *Aly*13 sequence variants that are very similar to each other. In Table 5.3 we show the mean level of nucleotide diversity within 5 different subtypes (*Aly*13-1, *Aly*13-2, *Aly*13-3, *Aly*13-7, *Aly*13-9, and *Aly*13-13). These particular subtypes are analysed because we have shown (Chapter 3, Section 3.3.9) that some are linked to SI (*Aly*13-1, *Aly*13-3 and *Aly*13-13) and that others have a kinase domain and are likely to be alleles of the same locus (*Aly*13-3, *Aly*13-7, and *Aly*13-13; Chapter 3, Section 3.3.4).

Table 5.3. *Ka* and *Ks* variation for variation within five *Aly*13 'subtypes'.

*Aly*13 subtype	*Aly*13-1 (*n* = 11)	*Aly*13-2 (*n* = 10)	*Aly*13-3 (*n* = 5)	*Aly*13-7 (*n* = 2)	*Aly*13-13 (*n* = 2)
Ka	0.004	0.004	0.007	0.007	0.001
Ks	0.009	0.007	0.004	0.008	0.013

The individual plants sequenced are listed in Table 5.1. *n* is the number of individuals sequenced for each subtype. For *Aly*13-13, individual 99A9-1 was compared to the cDNA sequence provided by J. Nasrallah. Length sequenced is 988 bp on average from position 321 to 1398 (includes indels) in our alignments of all the S-domains.

It should be noted that almost all the nucleotide variation described in Table 5.3 reflects singleton differences among the loci compared. Even though the number of *Ks* usually exceeds *Ka*, some of these mutations may reflect errors that arise as result of cloning PCR products. Nevertheless, the nucleotide variation within these subtypes is extremely small compared to that observed between subtypes (Table 5.4). In contrast, taking single representatives of each of the *Aly*13 subtypes described in Table 5.2, we observe that the mean level of nucleotide diversity at synonymous and nonsynonymous sites is much larger than within subtypes, and is also larger than for mean estimates when all 38 *Aly*13 sequences are included (which includes multiple variants within subtypes). Finally, the within-subtype variation is also smaller than the nucleotide diversity observed for S-domain loci in *A. lyrata* that do not appear to segregate with incompatibility (Chapter 4).

Table 5.4. Between subtype pairwise synonymous (upper right triangle) and nonsynoymous differences (lower triangle) among all the *Aly*13 subtypes within *A. lyrata*. Individuals chosen for each subtype are shown.

Individual, Subtype	98E17-11 *Aly*13-1	A9(4) *Aly*13-2	98E15-15 *Aly*13-3	B15(3) *Aly*13-4	b13(8) *Aly*13-5	Apstorr16 *Aly*13-6	98J43/1 *Aly*13-7	BenLui16 *Aly*13-8	98J40/7 *Aly*13-9	a5(4) *Aly*13-10	99A9-1 *Aly*13-13
98E17-11 *Aly*13-1		0.5018	0.4663	0.4685	0.4978	0.5146	0.4868	0.4725	0.3949	0.4956	0.4839
A9(4) *Aly*13-2	0.2493		0.3315	0.5231	0.4362	0.3543	0.1137	0.2885	0.4781	0.4629	0.4557
98E15-15 *Aly*13-3	0.2796	0.2054		0.5331	0.4157	0.2576	0.3631	0.1991	0.4965	0.5294	0.4726
b15(3) *Aly*13-4	0.2189	0.2397	0.2625		0.5414	0.5365	0.5321	0.5216	0.4199	0.5079	0.5170
b13(8) *Aly*13-5	0.2335	0.2133	0.2232	0.2284		0.4314	0.4543	0.4352	0.4996	0.5314	0.3163
Apstorr16 *Aly*13-6	0.2675	0.2001	0.1300	0.2401	0.2148		0.3234	0.2340	0.4919	0.5632	0.5134
98J43/1 *Aly*13-7	0.2598	0.0891	0.2517	0.2536	0.2412	0.2233		0.3103	0.4527	0.4577	0.4811
BenLui16 *Aly*13-8	0.2469	0.1878	0.1332	0.2537	0.2088	0.1115	0.2343		0.4868	0.5312	0.4569
98J40/7 *Aly*13-9	0.2189	0.2263	0.2641	0.1759	0.2224	0.2534	0.2635	0.2402		0.4706	0.5140
a5(4) *Aly*13-10	0.2291	0.2723	0.2962	0.1717	0.2368	0.2726	0.2717	0.2795	0.1929		0.4482
99A9-1 *Aly*13-13	0.2294	0.2252	0.2337	0.2457	0.1565	0.2238	0.2539	0.2236	0.236	0.2579	

Mean *Ks* = 0.431 ± 0.022; *Ka* = 0.226 ± 0.009

5.3.4. Frequency spectrum of nucleotide variation at *Aly*13 variants compared to *Brassica SLG* and *SRK* alleles

One way to compare diversity patterns between different groups of sequences is to compare the frequency distribution of nucleotide variation for the different groups. Figure 5.3 compares the per codon heterozygosity of the *Aly*13 variants to Brassica *SLG* and *SRK* alleles for two different classes of mutation (synonymous and nonsynonymous). The *Brassica SLG* and *SRK* alleles are the same as described in Chapter 1 except the *SLG* and *SRK* are shown separately in these analyses. Only *Aly*13 variants (subtypes) which are known to have a kinase domain were included in this analysis. Generally, the *Aly*13 variants exhibit higher diversity overall relative to the *SLG* and *SRK* loci in *Brassica*. This is observed for both synonymous and nonsynonymous diversity (Figure 5.2). All three groups of sequences also exhibit similar proportions of fixed nonsynonymous sites (sites with diversity estimates of zero). The nonsynonymous distributions suggest that the S-domain sequences linked to SI in different species share similar proportions of constrained amino-acids, and that the overall distribution of selection coefficients for amino acid changing positions are similar among species.

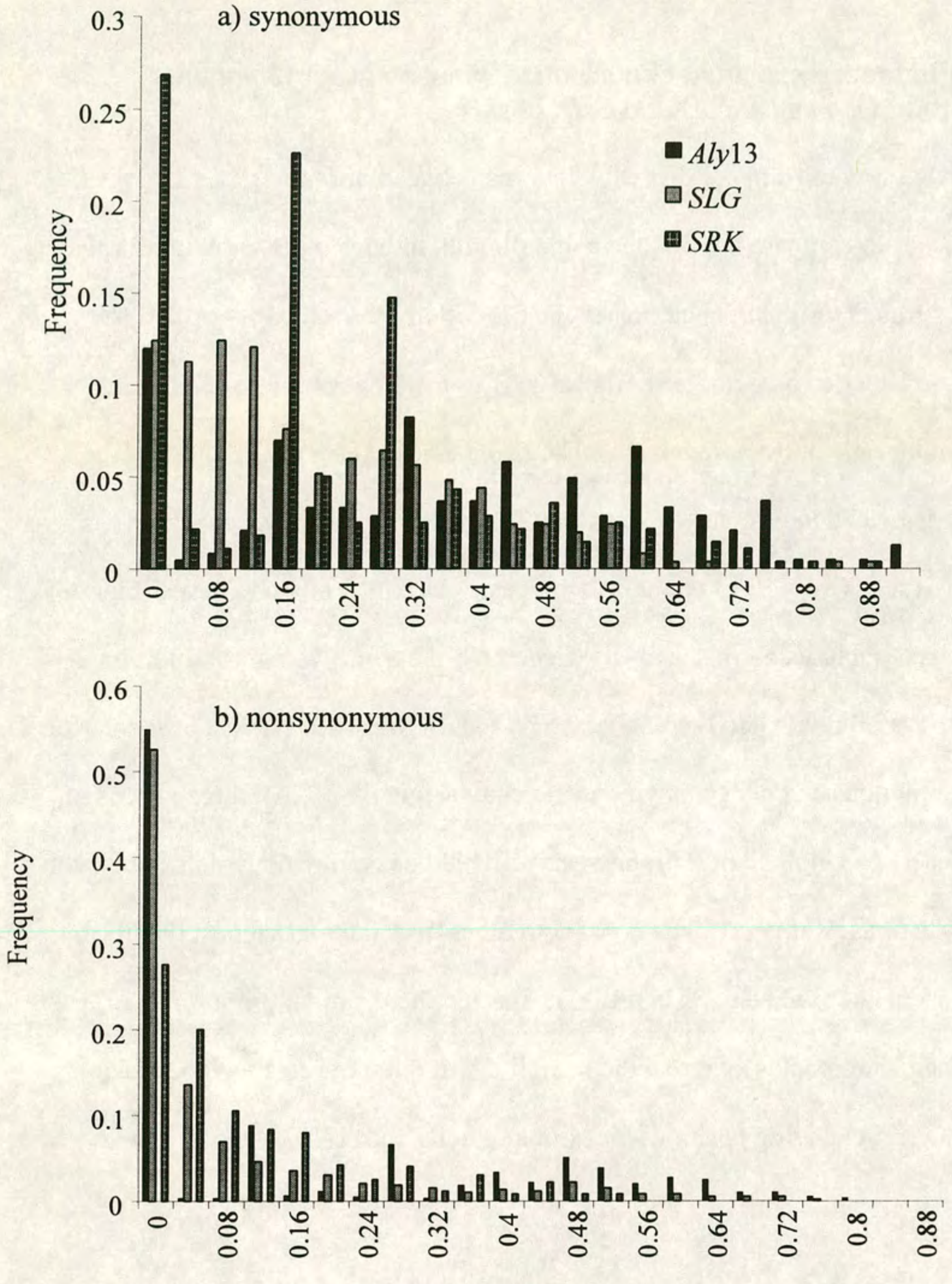


Figure 5.3 The frequency spectrum of diversity for the *Aly*13 kinase sequences and the *Brassica SLG* and *SRK* loci. a) is the frequency spectrum of synonymous diversity per codon and b) is the frequency of nonsynonymous diversity per codon.

5.3.5 Distribution of variation across the S-domain

Sequence variation is often observed to be distributed non-uniformly across sequences. In the *Brassica SLG* and *SRK* Type I alleles, the HV regions (Dwyer *et al.*1994, Kusaba *et al.* 1997) correspond to amino acid positions 215 – 236 for HV1, 293 - 331 for HV2, 354 – 368 for HV3, and 446-456 for the C-terminal region in our alignments. We examined whether *Aly*13 variants also exhibit higher polymorphism levels at these positions relative to the remainder of the sequence. Again, only *Aly*13 subtypes where a kinase domain has been identified are included in these analyses. A sliding window analysis of diversity shows that the *Aly*13 variants exhibit peaks of nonsynonymous mutations in positions corresponding to the HV regions found in the *SLG* and *SRK* S-domains of *Brassica* (Dwyer *et al.* 1993, Hinata *et al.* 1995, Charlesworth *et al.* 1998, Awadalla and Charlesworth 1999). Synonymous variation is generally very high particularly among the *A. lyrata* sequences, and indistinguishable across the S-domain. In *Brassica*, analyses of heterogeneity have focused on the Type I or dominant classes of alleles (Chapter 1, Awadalla and Charlesworth 1999). Sliding window analyses of synonymous variation for Type I alleles have shown some heterogeneity coinciding with the peaks of nonsynonymous diversity (Figure 1.8, Chapter 1). When *Brassica* Type II alleles are included with *Brassica* Type I alleles, the diversity estimates are, of course, higher than for Type I alone, and peaks of synonymous diversity exhibit less heterogeneity across the sequence (Figure 5.4). Our sample of *Aly*13 alleles is a random sample, irrespective of dominance, and therefore, may include allelic variants of different dominance classes (Charlesworth *et al.* 2000).

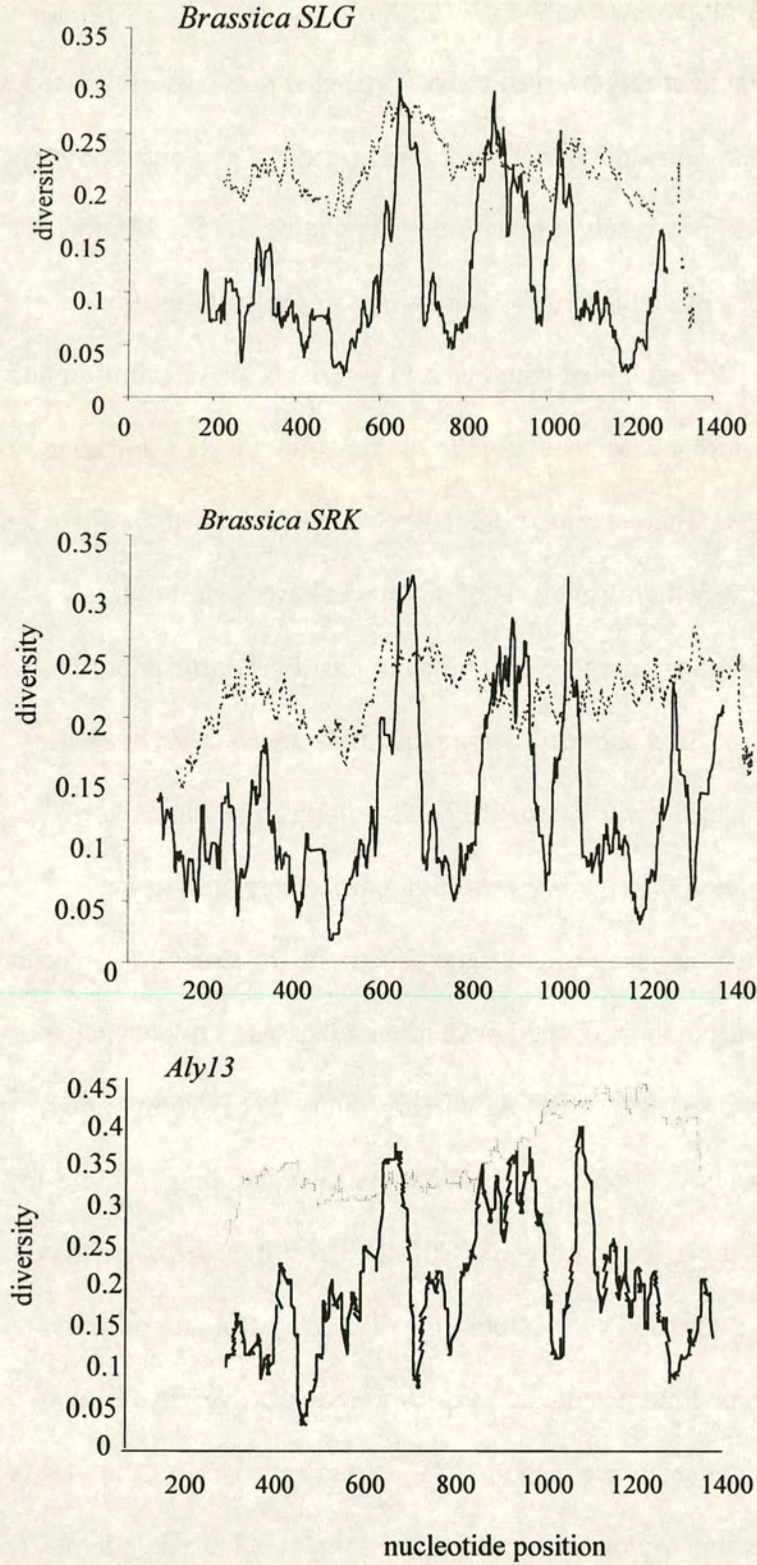


Figure 5.4. Sliding window analyses of diversity for *Brassica rapa* and *oleracea* Type I and Type II sequences are included in the analysis of *SLG* and *SRK* loci and *Aly*13 sequences in *A. lyrata*. Window size is 25 bp and shifted every base with dashed lines indicating synonymous diversity and solid lines indicating nonsynoymous diversity.

Table 5.5 shows the levels of synonymous and nonsynonymous nucleotide diversity for the corresponding HV regions for *Aly*13 variants relative to the remainder of the gene. As in the *SLG* and *SRK* loci (Chapter 1), *Aly*13 HV regions exhibit more than twice the level of nonsynonymous diversity relative to the conserved regions. On the other hand, synonymous diversity tends to be equally high in both regions, and is probably at saturation. This is observed when only a single representative sequence for each subtype is included in the estimation, as well as when all variants within subtypes are included in the analyses. Analyses including only single representatives of each subtype were performed so that comparisons could be made to the *Brassica SRK* and *SLG* data, where each sequence is S-specific.

Table 5.5. Diversity values expressed as mean uncorrected pairwise differences in different parts of the *Aly13* sequences.

	Single representatives of each subtype (n = 4)		Including all sequenced variants within each subtype (n = 10)*	
	"remainder"	"HV"	"remainder"	"HV"
Nucleotides	654	235	654	235
Ka	0.198 ± 0.013	0.411 ± 0.029	0.156 ± 0.011	0.347 ± 0.024
Ks	0.455 ± 0.033	0.496 ± 0.056	0.354 ± 0.026	0.382 ± 0.047

* The cDNA sequence provided by J. Nasrallah, which corresponds to subtype Aly13-13, was included in these estimates.
Sequences included in the analyses on the left side of the table are single sequence representative of subtypes *Aly*13-3, *Aly*13-7, *Aly*13-9, and *Aly*13-13 sampled from individuals 98E15-15, 98J43/1, 98J40/7 and 99A9-1, respectively. In the right hand columns, all sequences within a subtype found among different individuals (Table 5.1) were included in the calculation.

To assess whether patterns of diversity between the *Brassica* sequences and the *Aly* variants were significantly similar, a 'repeatability' analysis (Smith and Hurst 1998) was performed. Non-overlapping windows of diversity across a sequence for different categories of variation can be compared between alternative species (given a correct alignment for both sets of sequences). Using the same set of *Aly*13 sequences as for the results in Table 5.5, and comparing them to *Brassica* *SRK* sequences, we see a significant correlation for nonsynonymous diversity estimates for non-overlapping windows 25 bp in length (Figure 5.5a Spearman R = 0.83, $P < 0.0001$) as well as for each codon individually (Figure 5.5b Spearman R = 0.51, $P < 0.0001$; Figure 5.4). Synonymous diversity was not repeatable between the two species (not shown, $P = 0.21$).

In Chapter 4, it was shown that the low diversity S-domain paralogues that do not show evidence of linkage to SI in *A. lyrata* exhibit higher pairwise divergence between loci for the HV regions compared to the remainder of the genes. The mean divergence between members of the S-domain gene family presumed not involved in SI in *A. lyrata* are shown in Table 5.6. Similar estimates were also made for the kinase sequences only (*Aly*8sk, *Aly*10sk1, and *Aly*10sk2). The kinase loci are shown separately because synonymous divergence has not reached saturation among these sequences (Table 3.1, Chapter 3). There are more substitutions in the 'HV' regions compared to the remainder of the S-domain.

Figure 5.6 shows that divergence values calculated for non-overlapping windows of 25 base pairs for the unlinked S-domain loci (*Aly*8sk, *Aly*10sk1, and *Aly*10sk2) correlate with *Ka* polymorphism values in *Aly*13 (Spearman R = 0.44, P = 0.007). Per codon comparisons between *Aly*13 polymorphism and *Aly*8sk, *Aly*10sk1, and *Aly*10sk2 divergence are also significantly correlated (Spearman R = 0.44, $P < 0.001$).

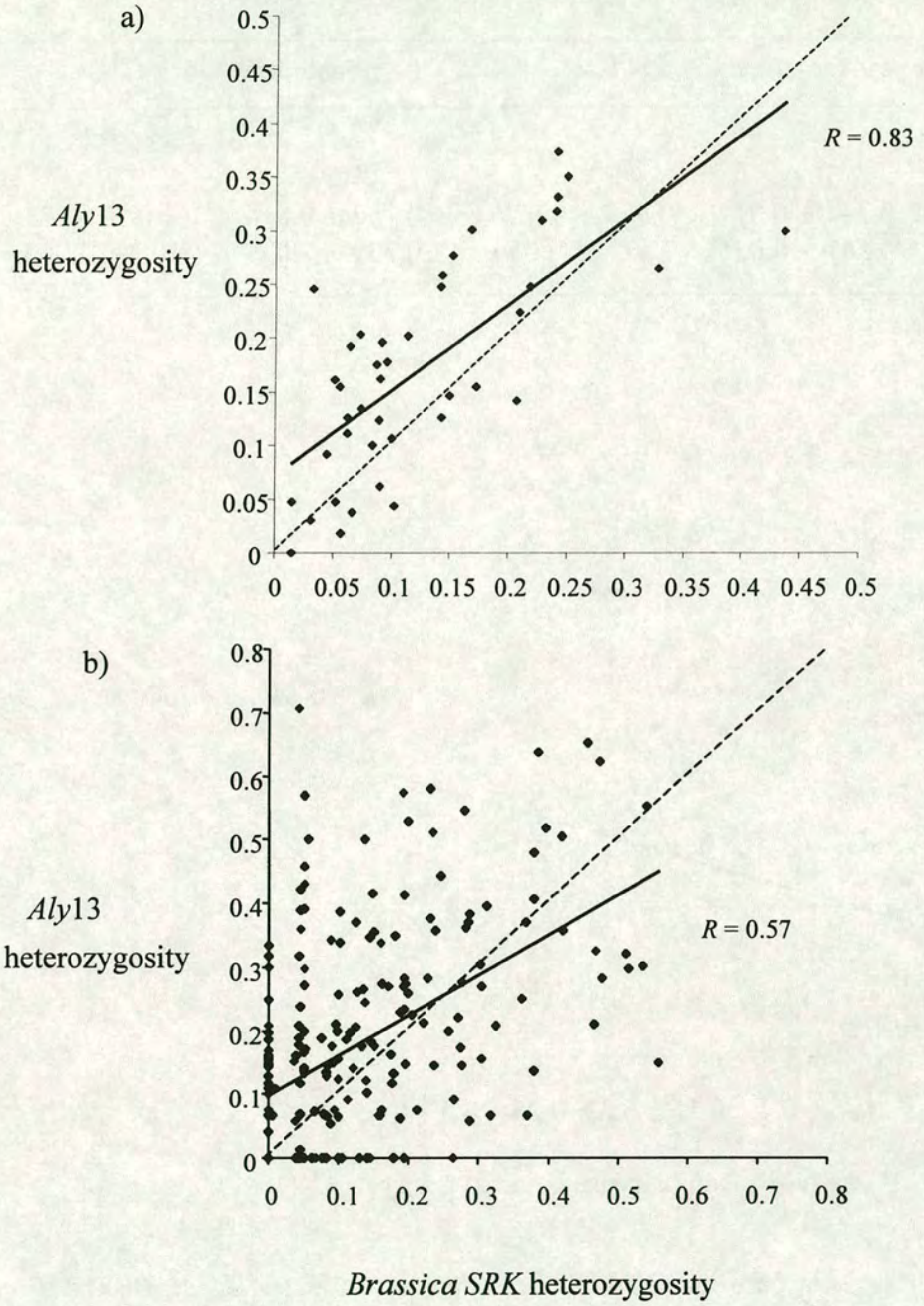


Figure 5.5. *Ka* repeatability for mean pairwise diversity at the *Aly*13 variants in *A. lyrata* and *Brassica SRK* sequences. The dashed lines indicate a 1:1 ratio of identical amino acid evolution. In a) are estimates for 25 bp regions and in b) are estimates for each codon separately. Many fixed codons ($\pi = 0$) at both species greatly affect the Spearman rank correlation (R) in (b).

Table 5.6. Mean uncorrected *Ka* and *Ks* divergence for a) all members of the S-domain family excluding *Aly*13 and b) for *Aly*8sk, *Aly*10sk1, and *Aly*10sk2.

	a) "remainder"	"HV"	b)"remainder"	"HV"
Nucleotides	502	235	502	235
Ka	0.244 ± 0.014	0.357 ± 0.027	0.132 ± 0.014	0.168 ± 0.025
Ks	0.541 ± 0.035	0.521 ± 0.056	0.339 ± 0.037	0.319 ± 0.063

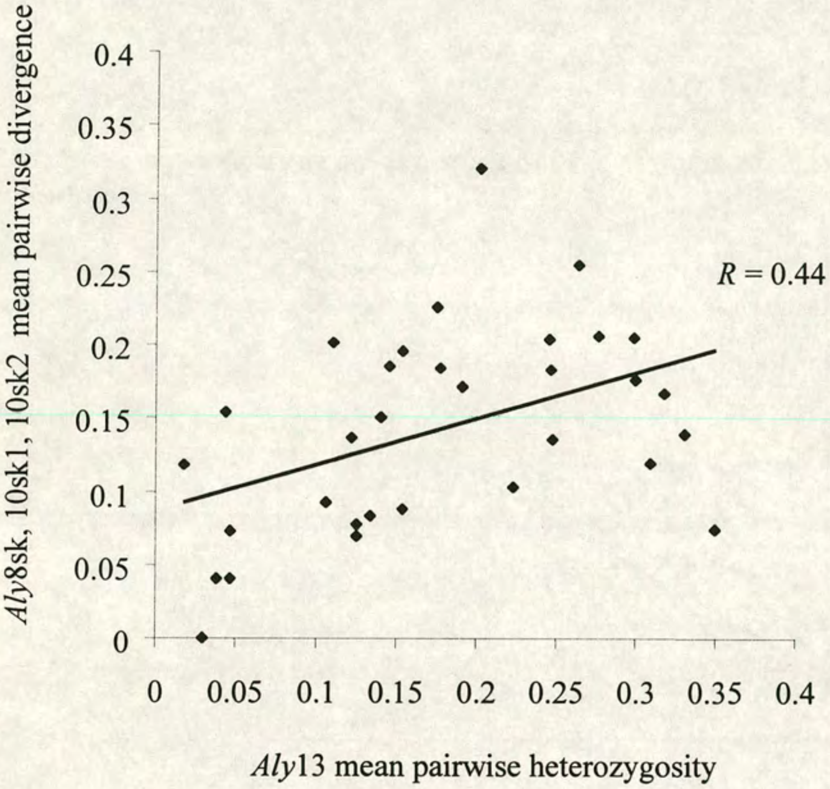


Figure 5.6. Divergence partitioned into 25 bp segments at the *Aly*8sk, *Aly*10sk1, and *Aly*10sk2 loci is correlated to polymorphic regions at *Aly*13 sequences.

5.4. Discussion

It is clear that *Aly*13 variants exhibit very high levels of nucleotide diversity, as expected for a locus under balancing selection (Takahata 1990, Stahl *et al.* 1999, Schierup *et al.* in press), and similar to diversity estimates made for the *SLG* and *SRK* loci in *Brassica* species (Hinata *et al.* 1995, Kusaba *et al.* 1997, Awadalla and Charlesworth 1999). Although many different *Aly*13 'subtypes' have been identified (eleven shown here), we do not know which variants are *SLG*-like or *SRK*-like or whether they all segregate with incompatibility groups. At present, estimates of heterozygosity and *Fst* based on functional incompatibility group frequencies will not be meaningful for the *Aly*13 loci for a number of alternative reasons. Furthermore, a number of alleles remain undetected due to our inability to amplify all alleles present in a population or even within an individual (most individuals should be heterozygous). Therefore, accurate comparisons to theoretical predictions about the distribution of alleles and S-allele coalescent times can not be made at this time, although it is clear that heterozygosity is high in that many subtypes appear to be observed in a number of populations, and there is little population subdivision (*Kst*) based on the nucleotide data.

Sampling a number of individuals has shown that the range of pairwise diversity values is very large and highly bimodal. Sequences tend to be either extremely similar, Table 5.3, or extremely divergent, Table 5.4. This bimodal variation is consistent with some individuals sharing incompatibility alleles. It is also interesting that levels of polymorphism are generally higher for the *Aly*13 variants compared to the *Brassica SLG* and *SRK* estimates. For technical reasons, most *SLG* and *SRK* alleles in *Brassica* that have been sequenced are from Type I incompatibility (dominant) groups. If we are sampling alleles irrespective of dominance class in *A. lyrata*, it would not be surprising that mean estimates of synonymous and nonsynonymous variation for *Aly*13 are higher than estimates made for Type I *Brassica SRK* alleles alone.

The sliding window analyses reveal that patterns of nonsynonymous diversity at the S-domain of the *Aly*13 sequences are highly correlated with those in *Brassica SRK* locus. Similar comparisons between *Aly*13 loci and S-domains that do not show evidence of linkage to SI in *A. lyrata* reveals that these alternative S-domain are not 'repeatable' with respect to the polymorphism distributed across the locus. Diversity at these loci is generally very low, with some loci showing either no nonsynonymous diversity in HV regions (e.g. *Aly*9) or no mutations at all (e.g. *Aly*3; Figure 4.2, Chapter 4). It is clear that diversity within *Aly*3 and its distribution across the sequence is very different from *Aly*13.

Divergence estimates between members of the S-domain gene family not involved in SI revealed higher levels of substitution or divergence at similar positions to the HV regions in *Brassica SLG* and *SRK* genes, and *Aly*13 as well. These observations may shed some light on whether these highly variable regions are targets of balancing selection versus regions under relaxed selective constraint. The strong correlation between nonsynonymous polymorphism patterns at *Brassica* and *A. lyrata* loci suggests that the HV patterns are not merely due to stochastic influences creating differences among regions of the *SLG* and *SRK* loci in *Brassica*. However, this correlation alone does not resolve the question about the nature of selection in these regions. The signficant correlation between substitution rates among *Aly*8sk, *Aly*10sk1, and *Aly*10sk2, and polymorphism within *Aly*13 variants could be interpreted in three ways; i) relaxed selective constraint at these regions across many members of this gene family, ii) regions that are under both positive and balancing selection or iii) gene conversion. We will next discuss each of these possible explanations in turn.

The relatively high levels of divergence at the HV regions compared to the remainder of the sequence could reflect relaxed selective constraints at these regions across all the members of S-domain gene family. Because of the age of the S-alleles, as well as the age of the gene family, mutations may have been free to accumulate in these regions over a long evolutionary time scale. If these regions are evolving neutrally, then we might also observe nonsynonymous mutations in these regions within the *Aly*13 subtypes. Otherwise, these mutations might change the incompatibility type. For example, within the HV regions of the five *Aly*13-3 subtype sequences studied, we observe 3 nonsynonymous segregating sites (mean *Ka* estimate of 0.009 ± 0.005), and one of the three mutations is a non-singleton. However, within the remaining *Aly*13 subtypes, only one singleton nonsynonymous change was observed within an HV region (*Aly*13-1). Perhaps, the sequences classified as subtype *Aly*13-3 are not all of the same incompatibility class, or the particular amino-acid replacement mutations observed do not affect the incompatibility type. Furthermore, if these regions were under relaxed selective constraint in all the S-domain loci, then we expect to see a greater accumulation of indels in the corresponding 'HV' regions, relative to the remainder of the sequences, among the members of this gene family. In fact, we only observed one polymorphic amino acid deletion which was present among the *Aly*7- sequences. No other indels were observed in these regions.

A second alternative is that the elevated level of nonsynonymous substitutions in the HV regions indicate positive selection in the ancestry of this gene family (Ohta 1994). Mutations at these particular regions may have contributed to gene diversification in this gene family (Chapter 4). There is currently no evidence for this, as we have not been able to identify sites that appear to be under positive selection using likelihood approaches (Chapter 4, Yang *et al.* 2000). *Ka/Ks* divergence values for the HV regions among members of the gene family are significantly less than one (Chapter 4).

The third alternative is that concerted evolution may be affecting substitution rates among the other members of this gene family. Unequal crossing-over or gene conversion events among members of the gene family may be exchanging regions that are highly variable between different genes, including the Aly13 variants. However, one would expect high divergence, not only for the HV regions, but throughout the locus unless gene conversion was extremely frequent. At present, there is no indication of gene conversion between genes occurring, and certainly not frequently (Chapter 4).

The repeatability comparisons also provide information about which amino acid sites at the *SRK* loci appear constrained or are fixed. Ninety-five amino acid positions out of approximately 330 are conserved among all of the *Aly*13 variants and the *SRK* alleles of *Brassica* aligned together. Interestingly, only 5 of these positions occur in the HV regions. Of these 5 include the cysteine residues which appear to be conserved across all the S-domain loci in both *Brassica* and *A. lyrata* (Chapter 3). Of course, not all of the amino acid positions that are variable will be sites under balancing selection. Some mutations may be deleterious, but the strength of balancing selection at linked sites could permit their segregation in the S-allele population. As a result, these slightly deleterious mutations are effectively neutral in that they have no effect on the fitness of the incompatibility allele.

Chapter 6

Conclusions

This thesis describes the molecular evolution of the self-incompatibility gene family in the Brassicaceae. In this Chapter, I summarise the principal results of the thesis, and discuss future experimental and theoretical work on the molecular co-evolution of the pollen and pistil components of SI.

In Chapter 1, I reviewed and analysed sequence data available for the maternal component of recognition (the *SLG* and *SRK* loci) for two species of *Brassica*. *SLG* and *SRK* variants revealed that patterns in diversity for both genes were shared and consistent. In the past, it has been argued that shared patterns of polymorphism between the two loci indicated that similarity was necessary and constrained for S-haplotype specificity. Since it is unlikely that the *SLG* locus is directly involved in the recognition process (Takayama et al. 2000), it appears that shared polymorphism patterns between the the *SRK* and *SLG* loci may merely reflect concerted evolution between loci (Charlesworth and Awadalla 1998). It has also been argued that the highly variable portions of the gene must indeed include sites targetted by balancing selection. Balancing selection will drive nucleotide diversity upwards at sites which are subject to selection, and at neutral or slightly deleterious sites physically linked to the sites targetted by balancing selection (Charlesworth et al. 1997). For HV regions in *Brassica*, both synonymous and nonsynonymous diversity is high yet, *Ka/Ks* values do not exceed unity. It remains a question as to whether these regions are targetted by selection or are merely neutral (but see below).

Patterns of haplotype variation at the *SLG* and the *SRK* locus indicate that recombination has likely played a role in the evolution of SI haplotypes (Chapter 2). As mentioned, within haplotype differences between the *SLG* and the *SRK* loci are smaller on average relative to between haplotype comparisons. This may be due to concerted evolution between loci. Furthermore, linkage disequilibrium patterns at the *SLG* locus provide evidence for symmetrical or asymmetrical recombination. Recombination is likely to be suppressed to a large degree at this region in order to maintain haplotype configurations between the *SRK* (and perhaps the *SLG*) and the pollen locus locus (*SCR*). Nevertheless, this does not exclude the exchange of fragments between loci from occurring at a low rate, perhaps through unequal-crossing over events (Li 1997). It may be that recombination contributes to the creation of new S-allele variants.

In Chapter 3, I described the characterization of a gene family in *Arabidopsis lyrata* which appears homologous to the S-domain gene family in *Brassica* and self-compatible *A. thaliana*. Six sets of S-domain sequences were characterized and were shown to be unlinked to self-incompatibility. One set may be broken down into two further loci (*Aly*7) which may include a pseudogene. A seventh group of sequences (*Aly*13) includes variants that segregate with incompatibility groups in four families tested. All of the S-domain loci are alignable to each other, but are clearly distinct loci with respect to each other (e.g. all the loci can be amplified in a single individual). The members of this gene family share common properties such as conserved cysteine residues. Interspecific phylogenetic clustering of genes from different species indicates that the gene family itself appears to be as old as the Brassicaceae plant family.

Of the seven S-domains characterised, six show low to moderate levels of neutral diversity, consistent with other published estimates for loci in *A. lyrata* (Chapter 4). Although there seems to be some relationship between the amount of recombination detected and the level of diversity observed, the argument is circular because we may merely have more power to detect recombination at sites with higher diversity. There is clear heterogeneity in substitution rates both within and among these loci. Regions classified as hypervariable in S-loci in *Brassica* exhibit much higher nonsynonymous substitutions (divergence) in the corresponding regions for other members of the S-domain family. It does not appear that mutation rates differ, because synonymous substitution rates appear relatively uniform across the locus. Elevated nonsynonymous substitutions at these particular regions suggests that amino-acid positions at these sites are under relaxed-selective constraint.

One particular locus exhibits slightly elevated levels of synonymous polymorphism (*Aly*8sk). We can not account for this high level of diversity as of yet. It does not appear that this set of sequences are two related loci, although we can not rule this out for certain. Substitution rates between *Aly*8sk and the most closely related orthologue in *A. thaliana* show slightly elevated divergence compared to other orthologous comparisons in this gene family. However, it does appear that some polymorphic sites at *Aly*8sk may be donated as fragments from other loci through concerted evolution. Alternatively, these same sites may be ancestrally polymorphic. Either hypothesis will have the effect of elevating diversity and divergence, although these mechanisms should not restrict themselves to synonymous diversity. Finally, linkage to another locus under balancing selection may be elevating synonymous variation at this locus.

In Chapter 5, we showed that a set of sequences that segregates with incompatibility groups in *A. lyrata* exhibit 1) exceptionally high levels of nucleotide diversity, 2) limited population subdivision, 3) nonsynonymous polymorphism patterns consistent with *Brassica* S-loci and 4) nonsynonymous polymorphism that correlates with subsitution heterogeneity at the other members of the S-domain loci. This final correlation suggests that these regions are either under relaxed-selective constraint at both groups of sequences, or that the HV regions include sites that are targets of balancing selection as well as being under relaxed selective constraint in S-loci. A possible third alternative is that these regions are under positive diversifying selection within the gene family as well as being under balancing selection in *Aly*13. However, likelihood models of positive selection show no indication of this mechanism occurring within this gene family.

It seems clear that there are still many avenues of exploration within the scope of SI molecular evolution. Although co-evolution of the *SLG* and *SRK* loci has not been made explicit in this thesis, in essence we have examined the co-evolutionary trajectory of variants at these loci in *Brassica* by examining the diversity relationships between variants. Determining that fragments appear to be exchanged between the *SLG* and the *SRK* locus indicates that these genes are not evolving independently. The co-evolutionary relationships between the maternal and paternal loci are likely to be very intriguing. It appears that the pollen *SCR* locus is highly variable. How does this variability correlate with maternal variability? Furthermore, how do new alleles evolve? Mutations might create self-compatible alleles by affecting the recognition properties of the maternal and paternal proteins, rather than create new S-alleles. Self-compatibility will likely be disadvantageous in highly outbred systems (Maynard Smith 1989). The likelihood of compensatory mutations occurring at a second locus after a mutation occurred at one of the recognition loci may depend on many factors other than the mutation rate. It is likely that many sites within a locus are targetted by selection and perhaps interact (epistasis) to form a stable SI allele (Chapter 2). For example, the process of evolution to new alleles may not be strictly digital, in that many mutations may be required to alter one component of the recognition process and perhaps only a few mutations at the other end of recognition may be required to compensate. As well, mutations or recombination might create 'leakage' and only weaken the SI response. Another possibility may be that a recombinant may recognise two alleles instead of one or none. Nevertheless, investigations in *A. lyrata* involving a pollen gene will provide some insight as to the co-evolutionary properties behind SI evolution, as well as other systems where products of genes interact and co-evolve.

References

Anderson, M. A., McFadden, G. I., Bernatzky, R., Atkinson, A., Orpin, T., Dedman, H., Tregear, G., Fernley, R., and Clarke, A. E. 1989. Sequence variability of three alleles of the self-incompatibility gene of *Nicotiana alata*. The Plant Cell 1:483-491.

Andolfatto P. and Nordborg, M. 1998. The effect of gene conversion on intralocus associations. Genetics 148:1397-1399.

Arnhein, N. 1983. *Concerted evolution of multigene families*, pp. 38–61 In M. Nei and R. K. Koehn (eds.). Evolution of Genes and Proteins. Sinauer, Sunderland, MA.

Awadalla, P., and Charlesworth, D. 1999. Recombination and selection at *Brassica* S-loci. Genetics 152: 413-425.

Awadalla, P., Eyre-Walker, A., and Maynard Smith, J. 1999. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science 286:2524-2525.

Bateman, A.J. 1952. Self-incompatibility systems in angiosperms. I. Theory. Heredity 6:285-310.

Beach, J. H., and Kress, W. J. 1980. Sporophytic versus gametophytic: a note on the origin of self-incompatibility in the flowering plants. Systematic Botany 5:1-5.

Betrán, E., Rozas J., Navarro A., and Barbadilla A. 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics 146: 89-99.

Boyes D.C., Chen, C.H., Tantikanjana, T., Esch, J.J., and Nasrallah, J.B. 1991. Isolation of a second S-locus-related cDNA from *Brassica oleracea*: genetic relationships between the S locus and two related loci. Genetics 127:221-228.

Boyes, D.C. and Nasrallah, J.B. 1993 Physical linkage of the *SLG* and *SRK* genes at the self-incompatibility locus of *Brassica oleracea.* Molecular and General Genetics 236:369-373.

Boyes, D.C., Nasrallah, M. E., Vrebalov, J., and Nasrallah, J. B. 1997. The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. The Plant Cell 9:237-247.

Brace, J., Ockendon, D.J., and King, G.J. 1993. Development of a method for the identification of S alleles in *Brassica oleracea* based on digestion of PCR amplified DNA with restriction endonucleases. Sexual Plant Reproduction 6: 133-138.

Bulmer, M.G. 1979. Principles of Statistics. Dover, New York.

Cabrillac, D., Delorme, V., Garin, J., Ruffio-Chable, V., Giranton, J-L., Dumas, C., Gaude, T., and Cock, J.M. 1999. The S15 self-incompatibility haploytpe in *Brassica oleracea* includes three S gene family members expressed in stigmas. The Plant Cell 11:971-985.

Casselman, A., Vrebalov, J., Conner, J.A., Singhal, A., Giovannoni, J., Nasrallah, M. E., and Nasrallah, J.B. 2000. Determining the physical limits of the *Brassica* S-locus by recombinational analysis. The Plant Cell 12:23-33.

Charlesworth, B. and Charlesworth, D. 1979 The maintenance and breakdown of distyly. American Naturalist 114:499-513.

Charlesworth, B., Nordborg, M., and Charlesworth, D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research 70:155-174.

Charlesworth, D. 1988. Evolution of homomorphic sporophytic self-incompatibility. Heredity 60:445-453.

Charlesworth, D. and Awadalla, P. 1998. The molecular population genetics of flowering plant self-incompatibility polymorphisms. Heredity 81:1-9

Charlesworth, D., Awadalla, P., Mable, B., and Schierup, M. 2000. Population-levels studies of multiallelic self-incompatibility loci, with particular reference to Brassicaceae. Annals of Botany 85 (supplement A): 227-240.

Charlesworth, D. and Guttman, D. 1997. Seeing selection in S allele sequences. Current Biology 7:R34-7.

Chen, C.H. and Nasrallah, J.B. 1990. A new class of S sequences defined by a pollen recessive self-incompatibility allele of *Brassica oleracea*. Molecular and General Genetics 222:241-8.

Clark, A.G. 1997. Neutral behaviour of shared polymorphism. Proceedings of the National Academy of Science USA 94:7730-7734.

Clark, A.G. and Kao, T.-H. 1991. Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proceedings of the National Academy of Science USA 88:9823-9827.

Coleman, C.A., and Kao, T.-H. 1992. The flanking regions of *Petunia inflata* S alleles are heterogeneous and contain repetitive sequences. Plant Molecular Biology 18: 725-737.

Conner, J.A., Conner, P., Nasrallah, M.E., and Nasrallah, J.B. 1998. Comparative mapping of the *Brassica* S region and its homeolog in *Arabidopsis*: implications for the evolution of mating systems in the *Brassica*ceae. The Plant Cell 10:801-812.

Cope, F.W. 1962. The effects of incompatibility and compatibility on genotype proportions in populations of *Theobroma cacao* L. Heredity 17:183-195.

Delorme, V., Giranton, J.-L., Hatzfeld, Y., Friry, A., and Heizmann, P. 1995 Characterization of the S locus genes, *SLG* and *SRK*, of the *Brassica* S3 haplotype: identification of a membrane-localized protein encoded by the S locus. Plant Journal 7:429-440.

Dowd, P.E., McCubbin, A.G., Wang, X., Verica, J.A., Tsukamoto, T., Ando, T. and Kao T-H. 2000. Use of *Petunia inflata* as a model for the study of solanaceous type self-incompatibility. Annals of Botany (supplement A) 85: 87-94.

Dwyer, K.G., Balent, M.A., Nasrallah, J.B., and Nasrallah, M.E. 1991. DNA sequences of self-incompatibility genes from *Brassica campestris* and *B. oleracea:* polymorphism predating speciation. Plant Molecular Biology 16:481-486.

Dwyer, K.G., Kandasamy, M.K., Mahosky, D.I., Axxiai, J., Kudish, B.I., Miller, J.E., Nasrallah, M.E., and Nasrallah, J.B. 1994. A superfamily of *S* locus-related sequences in *Arabidopsis*: diverse structures and expression patterns. The Plant Cell 6:1829-1843.

Dykhuizen, D. and Hartl, D.L. 1980. Selective neutrality of 6*PGD* allozymes in *E. coli* and the effects of genetic background. Genetics 96:801-17.

East, E.M. and Mangelsdorf, A.J. 1925. A new interpretation of the hereditary behaviour of self-sterile plants. Genetics 2:505–609.

Ernst, A. 1936. Weitere untersuchungen zur Phänanalyse zum Fertilitätsproblem und zur Genetik heterostyler Primeln. II. *Primula hortensis. Arch. der Julius-Klaus Stiftung für Vererbungsforchung,* Sozial-Anthropology. u Rassenhygiene. 11:1-280.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Ferris, P. J. and Goodenough, U. 1994 The mating-type locus of *Chlamydomonas reinhardtii* contains highly rearranged DNA sequences. Cell 76:1135-1145.

Filatov, D.A. and Charlesworth, D. 1999. DNA polymorphism, haplotype structure and balancing selection in the *Leavenworthia PgiC* locus. Genetics 153:1423-34.

Galloway, G.L., Malmberg R.L., and Price R.A. 1998. Phylogenetic utility of the nuclear gene arginine decarboxylase: an example from Brassicaceae. Molecular Biology and Evolution 15:1312-20.

Gebhardt, C., Eberle, B., Leonards-Schippers, C., Walkemeier, B., and Salamini, F. 1995. Isolation, characterization and RFLP linkage mapping of a DNA repeat family of *Solanum spegazzinii* by which chromosome ends can be localized on the genetic map of potato. Genetical Research 65: 1-10.

Gilbert, D.G. 1997. Seqpup version 0.6f: a biosequence editor and analysis application. http://sunsite.sut.ac.jp/pub/academic/biology/molbio/seqpup/

Glavin, T.L., Goring, D.R., Schafer, U., and Rothstein, S.J. 1994. Features of the extracellular domain of the *S*-locus receptor kinase from *Brassica*. Molecular and General Genetics 244:630-637.

Goldblatt, P. 1981. Index to plant chromosome numbers 1975–1978. Missouri Botanical Garden, St. Louis.

Golding, G.B. 1984. The sampling distribution of linkage disequilibrium. Genetics 108:257-274.

Goodman, M., Moore, G.W., and Matsuda, G. 1975. Darwinian evolution in the genealogy of haemoglobin. Nature 253:603-8.

Goodwillie, C. 1997. The genetic control of self-incompatibility in *Linanthus parviflorus* (Polemoniaceae). Heredity 79:424-432.

Goring, D.F. and Rothstein, S.J. 1996. *S-locus receptor kinase genes and self-incompatibility in Brassica napus ssp. oleifera*. pp. 217-230. In D. P. S. Verma (ed.). Signal Transduction in Plant Growth and Development. Springer, New York.

Goring, D.F., Banks, P., Beversdorf, W.D., and Rothstein, S.J. 1992. Use of the polymerase chain reaction to isolate an S-locus glycoprotein cDNA introgressed from *Brassica campestris* into B. *napus* ssp. oleifera. Molecular and General Genetics 234:185-192.

Grant, V. 1971. Plant Speciation. Columbia University Press. New York.

Gustafsson, K. and Andersson, L. 1994. Structure and evolution of horse MHC class II *DRB* genes: convergent evolution in the antigen-binding site. Immunogenetics 39:355-358.

Guttman, D. S. and Dykhuizen, D.E. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380-1383.

Haldane, J.B.S., 1933. Two new allelomorphs for heterostyly in *Primula.* American Naturalist 67:559-560.

Hamrick, J.L. and Godt, M.J.W. 1989. *Allozyme diversity in plant species.* pp 43-63. In A.H.D. Brown, M.T. Clegg, A.L.Kahler, B.S. Weir (eds.) Plant Population Genetics, Breeding, and Genetic Resources. Sinauer, Sunderland Mass.

Hatakeyama, K. T., Takasaki, T., Watanabe, M., and Hinata, K. 1998. Molecular characterization of the S-locus genes, *SLG* and *SRK*, in a pollen-recessive self-incompatibility haplotype of *Brassica rapa*.. Genetics 149:1587-1597.

Hedrick, P. W. 1987. Gametic disequilibrium measures—proceed with caution. Genetics 117: 331-341.

Henry, A.-M., and Damerval, C. 1997. High rates of polymorphism and recombination at the Opaque-2 locus in maize. Molecular and General Genetics 256:147-157.

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

Hill, W.G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38:226-231.

Hinata, K., Watanabe, M., Yamakawa, S., Satta, Y., and Isogai, A. 1995. Evolutionary aspects of the S-related genes of the *Brassica* self-incompatibility system: synonymous and nonsynonymous base substitutions. Genetics 140:1099-1104.

Holsinger, K.E., and Steinbachs, J.E. 1997. *Mating systems and evolution in flowering plants.* pp. 223-248. In K. Iwatsuki and P. Raven (eds.). Evolution and Diversification of Land Plants. Springer-Verlag, Tokyo.

Hudson, R.R. 1983. Properties of a neutral allele model with intragenic recombination. Theoretical and Population Biology 23:183-201.

Hudson, R.R. 1987. Estimating the recombination parameter of a finite population model without selection. Genetical Research 50:245-250.

Hudson, R.R. and Kaplan, N.L. 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. 1990. Gene genealogies and the coalescent process. Oxford Surveys of Evolutionary Biology 7:1-45.

Hudson, R.R. and Kaplan, N. L. 1988. The coalescent process in models with selection and recombination. Genetics 120:831-840.

Hudson, R.R., Boos, D.D., and Kaplan, N.L. 1992. A statistical test for detecting geographic subdivision. Molecular Biology and Evolution 9:138-151.

Hughes, A., and Nei, M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proceedings of the National Academy of Sciences, USA 86:958-962.

Hughes, A., Ota, T., and Nei, M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Molecular Biology and Evolution 76:515-524.

Hughes, A. and Yeager, M. 1998. Natural selection and the evolutionary history of the major histocompatibility complex loci. Frontiers of Bioscience 3:510-516.

Hughes, A. L., Green, J.A., Garbayo, J. M. and Roberts, R.M. 2000. Adaptive diversification within a large family of recently duplicated, placentally expressed genes. Proceedings of the National Academy of Sciences USA 97: 3319-3323.

Ioerger, T.R., Clark, A.G., and Kao, T.-H. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proceedings of the National Academy of Sciences USA 87:9732-9735.

Innan, H., Tajima, F., Terauchi, R., and Miyashita, N.T. 1996. Intragenic recombination in the Adh locus of the wild plant *Arabidopsis*. Genetics 143:1761-1770.

Isogai, A., Yamakawa, S., Shiozawa, H., Takayama, S., Tanaka, H., Kono, T., Watanabe, M., Hinata, K., and Suzuki, A. 1991. The cDNA sequence of NS1 glycoprotein of *Brassica campestris* and its homology to S-locus-related glycoproteins of *B. oleracea*. Plant Molecular Biology 17:269-71.

Janssens, G.A., Goderis, I.J., Broekaert, W.F., and Broothaerts, W. 1995. A molecular method for *S*-allele identification in apple based on allele-specific PCR. Theoretical and Applied Genetics 91:691-698.

Jorgensen, R. 1990. Altered gene expression in plants due to *trans* interactions between homologous genes. Trends in Biotechnology 8:340-344.

Junghans, H. and Metzlaff, M. 1990. A simple and rapid method for the preparation of total plant DNA. Biotechniques 8:176.

Karpechenko, G.D. 1927. Polyploid hybrids of *Raphanus sativus* L. x *Brassica oleracea* L. Bulletin of Applied Botanical Genetics and Plant Breeding (Leningrad) 17:305-410.

Kawabe A., Innan, H., Terauchi, R., and Miyashita, N.T. 1997. Nucleotide polymorphism in the acidic chitinase locus (*ChiA*) region of the wild plant *Arabidopsis thaliana.* Molecular Biology and Evolution 14:1303-15.

Kawabe, A. and Miyashita, N.T. 1999. DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana.* Genetics 153:1445-53.

Keeping, E.S. 1962. Introduction to Statistical Inference. Van Nostrand, Princeton, NJ.

Kelly, J.K. 1997. A test of neutrality based on interlocus associations. Genetics 146:1197-1206.

King, L.M. 1998. The role of gene conversion in determining sequence variation and divergence in the *Est-5* gene family in *Drosophila pseudoobscura.* Genetics 148:305-316.

Klitz, W., Stephens, J. C., Grote, M., and Carrington, M. D. 1995. Discordant patterns of linkage disequilibrium of the peptide-transporter loci within the *HLA* class II region. American Journal of Human Genetics 57:1436-1444.

Kowyama, Y., Shino, N., and Kawasi, T. 1980. Genetic analysis of incompatibility in the diploid species of *Ipomoea* closely related to the sweet potato. Theoretical and Applied Genetics 58:149-155.

Kreitman, M. and Hudson, R. R. 1991. Inferring the evolutionary histories of the *adh* and *adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127:565-582.

Kuhner, M. K., Lawlor, D. A., Ennis, P. D., and Parham, P. 1991. Gene conversion in the evolution of the human and chimpanzee *MHC* class I loci. Tissue Antigens 38:152-164.

Kumar, S., Tamura, K., and Nei, M. 1994. MEGA - molecular evolutionary genetics analysis software for microcomputers. Computer Applications in the Biosciences 10:189-191.

Kumar, V. and Trick, M. 1994 . Expression of the S-locus receptor kinase multigene family in *Brassica oleracea.* Plant Journal 6:807-813.

Kusaba, M. and Nishio, T. 1999. S-allele specificity of stigma proteins in *Brassica oleracea* and *Brassica campestris.* Heredity 41:93-100.

Kusaba, M., Nishio, T., Satta, Y., Hinata, K., and Ockendon, D. 1997. Striking similarity in inter- and intra-specific comparisons of class I SLG alleles from *Brassica oleracea* and *Brassica campestris*: Implications for the evolution and recognition mechanism. Proceedings of the National Academy of Sciences USA 94:7673-7678.

Kusaba, M., Matsushita, M., Okazaki, K., Satta, Y., and Nishio, T. 2000. Sequence and structural diversity of the S locus genes from different lines with the same self-recognition specificities in *Brassica oleracea*. Genetics 2000 154: 413-420.

Lagercrantz, U. 1998 Comparative mapping between *Arabidopsis thaliana* and *Brassica nigra* indicates that *Brassica* genomes have evolved through extensive genome replication accompanied by chromosome fusions and frequent rearrangements. Genetics 150:1217-1228.

Lalonde B.A., Nasrallah M.E., Dwyer K.G., Chen C.H., Barlow B., and Nasrallah, J.B. 1989. A highly conserved *Brassica* gene with homology to the S-locus-specific glycoprotein structural gene. The Plant Cell 1:249-58.

Lawrence, M.J. 2000. Population genetics of the homomorphic self-incompatibility systems in flowering plants. Annals of Botany 85(supplement A):221-226.

Lewis, D. 1963. *A protein dimer hypothesis on incompatibility*. pp. 656-663. In S. J. Geerts (ed.). Genetics Today. The Hague, Netherlands.

Lewontin, R. C. 1988. On measures of linkage disequilibrium. Genetics 120:849-852.

Li, W.-H. 1997. Molecular Evolutionary Genetics. Sinauer, Sunderland, MA.

Li, X., Nield, J., Hayman, D., and Langridge, P. 1995. Thioredoxin activity in the C terminus of *Phalaris* S protein. The Plant Journal 8:133-138.

Long, M. and Langley, C.H. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. Science 260:91-5.

Luu, D.T., Marty-Mazars, D., Trick, M., Dumas, C., and Heizmann, P. 1999. Pollen-stigma adhesion in *Brassica* spp. involves *SLG* and *SLR1* glycoproteins. The Plant Cell 11: 251-62.

Matton, D. P., Maes, O., Laublin, G., Xike, Q., and Bertrand, C. 1997. Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. The Plant Cell 9:1757-1766.

May, G., and Matzke, E. 1995. Recombination and variation at the a mating-type of *Coprinus cinereus*. Molecular Biology and Evolution 12:794-802.

May, G., Shaw, F., Badrane, H., and Vekemans, X. 1999. The signature of balancing selection: Fungal mating compatibility gene evolution. Proceedings of the National Academy of Sciences USA 96:9172-9177.

Maynard Smith, J., 1989. Evolutionary Genetics. Oxford University Press.

Maynard Smith, J., and Haigh, J. 1974. The hitch-hiking effect of a favorable gene. Genetical Research 23:23-35.

Maynard Smith, J., Smith, N.H., O'Rourke, M., and Spratt, B.G. 1993. How clonal are bacteria? Proceedings of the National Academy of Sciences USA 90:4384-4388.

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

Miyashita, N.T., Aguade, M., and Langley, C.H. 1993. Linkage disequilibrium in the white locus region of *Drosophila melanogaster*. Genetical Research 62:101-109.

Miyashita, N.T, Innan, H., and Terauchi R. 1996. Intra- and interspecific variation of the alcohol dehydrogenase locus region in wild plants *Arabis gemmifera* and *Arabidopsis thaliana*. Molecular Biology and Evolution 13: 433-436.

Moriyama, E.N., and Powell, J.R. 1995. Intraspecific nuclear DNA variation in Drosophila. Molecular Biology and Evolution 13:261-277.

Muirhead, T. and Slatkin, M. 2000. The effects of subdivision on overdominant loci. Evolution (in press).

Nasrallah J.B, and Nasrallah M.E. 1989. The molecular genetics of self-incompatibility in *Brassica*. Annual Review of Genetics 23:121-39.

Nasrallah, J.B. and Nasrallah, M.E. 1993. Pollen stigma signaling in the sporophytic self-incompatibility response. Plant Cell 5:1325-1335.

Nasrallah, M.E., Kandasamy, M.K., and Nasrallah, J.B. 1992. A genetically defined trans-acting locus regulates S-locus function in *Brassica*. Plant Journal 2:497-506.

Nasrallah, J. B., Kao, T.-H., Chen, C.-H., Goldberg, M., and Nasrallah, M. E. 1987. Amino-acid sequence of glycoproteins encoded by three alleles of the S locus of *Brassica oleracea*. Nature 326:617-619.

Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei M., and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3:418-26.

Neilsen, R. and Yang, Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.

Nishio, T. and Kusaba, M. 2000. Sequence diversity of *SLG* and *SRK* in *Brassica oleracea.* Annals of Botany 85(supplement A):227-240.

Nishio, T., Kusaba, M., Sakamoto, K., and Ockendon, D. 1997. Polymorphism of the kinase domain of the *S*-locus receptor kinase gene (*SRK*) in *Brassica oleracea* L. Theoretical and Applied Genetics 95:335-342.

Nordborg, M., Charlesworth, B., Charlesworth, D. 1996. Increased levels of polymorphism surrounding selectively maintained sites in highly selfing species. Proceedings of the Royal Society Series B 163:1033-1039.

Nou, I.S., Watanabe, M., Isogai, A., and Hinata, K. 1993. Comparison of S-alleles and S-glycoproteins between two populations of *Brassica campestris* in Turkey and Japan. Sexual Plant Reproduction 6:79-86.

O'Donnell, S., Lane, M.D., and Lawrence, M.J. 1993. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. VI. Estimation of the overlap between the allelic complements of a pair of populations. Heredity 71: 591-595.

O'hUigin, C. 1995. Quantifying the degree of convergence in primate *Mhc-DRB* genes. Immunology Review 143:123-140.

Ohta, T. 1994. Further evidence of evolution by gene duplication revealed through DNA sequence comparisons. Genetics 138:1331-1337.

Ockendon, D.J. 1974. Distribution of self-incompatibility alleles and breeding structure of open-pollinated cultivars of Brussels sprouts. Heredity 33:159-171.

Page, R.D.M. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences 12:357-358.

Pastuglia, M., Roby, D., Dumas, C., and Cock, J. M. 1997a. Rapid induction by wounding and bacterial infection of an *S* gene family receptor-like kinase gene in *Brassica oleracea*. The Plant Cell 9:49-60.

Pastuglia, M., Ruffio-Châble, V., Delorme, V., Gaude, T., Dumas, C., and Cock, J. M. 1997b. A functional *S* locus anther gene is not required for the self-incompatibility response in *Brassica oleracea*. The Plant Cell 9:2065-2076.

Price, R.A., Palmer, J.D., and Al-Shehbaz, I.A. 1994. *Systematic relationships of Arabidopsis: a molecular and morphological perspective*, pp. 7-19 In E.M.

Meyerowitz and C. Somerville (eds.), Arabidopsis. Cold Spring Harbor Laboratory Press, New York.

Purugganan, M.D., and Suddith, J. 1998. Molecular population genetics of the *Arabidopsis* CAULIFLOWER regulatory gene: non-neutral evolution and wild variation in floral homeotic function. Proceedings of the National Academy of Sciences USA 95:8130–8134.

Richman, A.D., and Kohn, J.R. 1999. Self-incompatibility alleles from *Physalis*: Implications for historical inference from balanced genetic polymorphisms. Proceedings Of the National Academy of Sciences USA. 96:168-172

Richman, A.D., Broothaerts, W., and Kohn J.R. 1997. Self-incompatibility RNAses from three plant families: Homology or convergence? American Journal of Botany 84:912-917.

Richman, A.D., Uyenoyama, M. K., and Kohn, J. R. 1996. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science 273: 1212-1216.

Robbins, T.P., Harbord, R.M., Sonneveld, T., and Clarke, K. 2000. The molecular genetics of self-incompatibility in Petunia hybrida. Annals of Botany 85(supplement A):105-112.

Rozas, J. and Rozas, R. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Computing and Applied Bioscience 13:307-311.

Rudd, J.J., Franklin, F.C.H., Lord, J. M., and Franklin-Tong, V.E. 1996. Increased phosphorylation of a 26-kD pollen protein is induced by the self-incompatibility response in *Papaver rhoeas*. The Plant Cell 8:713-724.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-25.

Sakamoto, K., Kusaba, M., and Nishio, T. 1998. Polymorphism of the S-locus glycoprotein gene (*SLG*) and the S-locus related gene (*SLR*1) in *Raphanus sativus* L. and self-incompatible ornamental plants in the *Brassica*ceae. Molecular and General Genetics 258:397-403.

Samaha R.R., and Boyle, T.H. 1989. The Self-incompatibility of *Zinnia-angustifolia* HBK (Compositae). 2. Genetics. Journal of Heredity 80:368-372.

Sampson, D.R. 1967. Frequency and distribution of self-incompatibility alleles in *Raphanus raphanistrum*. Genetics 56:241-251.

Sampson, D.R. 1974. Equilibrium frequencies of sporophytic self-incompatibility alleles. Canadian Journal of General Cytology :611-618.

Sanmiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., and Zakharov, D. 1997. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765-768.

Sassa, H., Hirano, H., and Koba, T. 1997. Style-specific self-compatible mutation caused by deletion of the S-RNase gene in Japanese pear (*Pyrus serotina*). The Plant Journal 12:223-227.

Sassa, H., Nishio, T., Kowyama, Y., Hirano, H., Koba, T., and Ikehashi, H. 1996. Self-incompatibility (*S*) alleles of the Rosaceae encode members of a distinct class of the T_2/S ribonuclease superfamily. Molecular and General Genetics 250:547-557.

Savolainen, O., Langley, C.H., Lazzaro, B.P., and Hélène, Fr. 2000. Contrasting Patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana.* Molecular Biology and Evolution 17:645-655.

Sawyer, S. A. 1989. Statistical test for determining gene conversion. Molecular Biology and Evolution 6:526-538.

Schaeffer, S. W., and Miller, E. L. 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics 135:541-552.

Schierup, M.H. 1998. The number of self-incompatibility alleles in a finite, subdivided population. Genetics 149:1153-1162.

Schierup, M.H., Vekemans, X., and Charlesworth, D. 2000. The effect of subdivision on variation at multi-allelic loci under balancing selection. Genetical Research (in press).

Schopfer, C.R., Nasrallah, M.E., and Nasrallah, J.B. 1999. The male determinant of self-incompatibility in *Brassica*. Science 286:1697-700.

Slatkin M. and Muirhead C.A. 1999. Overdominant alleles in a population of variable size. Genetics 152:775-81.

Smith, N. and Hurst, L. 1998. Molecular evolution of an imprinted gene: repeatability of patterns of evolution within the mammalian insulin-like growth factor type II receptor. Genetics 150:823-833.

Sokal, R. R., and Rohlf, F. J. 1995. Biometry: The Principles and Practice of Statistics in Biological Research, Ed. 3. W. H. Freeman, New York.

Stahl, E.A., Dwyer, G., Mauricio, R., Kreitman, M., Bergelson, J. 1999. Dynamics of disease resistance polymorphism at the *Rpm1* locus of *Arabidopsis*. Nature 12: 667-71.

Stein, J., Howlett, B., Boyes, D.C., Nasrallah, M. E., and Nasrallah, J.B. 1991. Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleracea*. Proceedings of the National Academy of Science USA 88:8816-8820.

Stephens, J.C. 1985. Statistical methods of DNA sequence analysis—detection of intragenic recombination or gene conversion. Molecular Biology and Evolution 2:539-556.

Stephenson, A.G., Doughty, J., Dixon, S., Ellrman, C., Hiscock, S., and Dickinson, H. G. 1997. The male determinant of self-incompatibility in *Brassica* is located in the pollen coating. The Plant Journal 12:1351-1359.

Stevens, J.P., and Kay, Q.O.N. 1989. The number, dominance relationships and frequencies of self-incompatibility alleles in a natural population of *Sinapis arvensis* L. in South Wales. Heredity 62:199-205.

Stone, S.L., Arnoldo, M. and Goring, D.R. 1999. A breakdown of *Brassica* self-incompatibility in ARC1 antisense transgenic plants. Science 286:1729-1731.

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

Suzuki, G., Watanabe, M., Toriyama, K., Isogai, A., and Hinata, K. 1996. Expression of *SLG*(9) and *SRK*(9) genes in transgenic tobacco. Plant Cell and Physiology 37:866-869.

Suzuki ,G., Watanabe, M., Kai, N., Matsuda, N., Toriyama, K., Takayama, S., Isogai, A., and Hinata, K. 1997. Three members of the S multigene family are linked to the S locus of *Brassica*. Molecular and General Genetics 256:257-264.

Suzuki G., Kai, N., Hirose, T., Fukui, K., Nishio, T., Takayama, S., Isogai, A., Watanabe, M., and Hinata, K. 1999. Genomic organization of the S locus: Identification and characterization of genes in *SLG/SRK* region of S(9) haplotype of *Brassica campestris* (syn. *rapa*). Genetics 153:391-400.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics. 105:437-60.

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

Takahata N. 1990. A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proceedings of the National Academy of Science USA 87:2419-23.

Takahata, N., and Satta, Y. 1998. Footprints of intragenic recombination at *HLA* loci. Immunogenetics 47:430-441.

Takasaki, T., Hatakeyama, K., Suzuki, G., Watanabe, M., Isogai, A., and Hinata, K. 2000. The S receptor kinase determines self-incompatibility in *Brassica* stigmas. Nature 403: 913-6.

Tanaka, T., and Nei, M. 1989. Positive Darwinian selection observed at the variable region genes of immunoglobulins. Molecular Biology and Evolution 6:447-59.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. 1997. The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research 25:4876-4882.

Tobias, C..M. and Nasrallah, J.B. 1996. An S-locus-related gene in *Arabidopsis* encodes a functional kinase and produces two classes of transcripts. Plant Journal 10:523-531.

Trick, M. and Flavell, R.B. 1989. Sporophytic self-incompatibility systems. *Brassica*-S gene family. Molecular and General Genetics 218:112-117.

Uyenoyama, M. K. 1995. A generalized least-squares estimate for the origin of self-incompatibility. Genetics 139:975-992.

Uyenoyama, M. K. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants of self-incompatibility. Genetics 147:1389-1400.

Vekemans, X. and Slatkin, M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics 137:1157-1165.

Vieira, C.P., Vieira, J., and Charlesworth D. 1999. Evolution of the cycloidea gene family in *Antirrhinum* and *Misopates*. Molecular Biology and Evolution 16:1474-83.

Walker, E.A., Ride, J.P., Kurup, S., Franklin-Tong, V. E., Lawrence, M.J., and Franklin, F.C.H. 1996. Molecular analysis of two functional homologues of the S3 allele of the *Papaver rhoeas* self-incompatibility gene isolated from different populations. Plant Molecular Biology 30:983-994.

Watanabe, M., Takasaki, T., Toriyama, K., Yamakawa, S., and Isogai, A. 1994 A high degree of homology exists between the protein encoded by *SLG* and the S receptor domain encoded by *SRK* in self-incompatible *Brassica campestris* L. The Plant Cell Physiology 35:1221-1229.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256-276.

Weller, S.G., Donoghue, M.J., and Charlesworth, D. 1995. *The evolution of self-incompatibility in the flowering plants: a phylogenetic approach.* pp. 355-382. In P. C. Hoch and A. G. Stephenson (ed.). Experimental and Molecular Approaches to Plant Biosystematics. Missouri Botanic Garden, St. Louis, Missouri.

Whitehouse, H. L. K. 1950. Multiple-allelomorph incompatibility of pollen and style in the evolution of the angiosperms. *Annals of Botany* 14:198-216.

Wines, D. R., Brady, J. M., Southard, E. M., and Macdonald, R. J. 1991. Evolution of the rat kallikrein gene family: gene conversion leads to functional diversity. Journal of Molecular Evolution 32:476-492.

Wolfe, K. H., Li, W.-H., and Sharp, P. M. 1989. Rates of synonymous substitution in plant nuclear genes. Journal of Molecular Evolution 28:208-211.

Wright, S. 1939. The distribution of self-sterility alleles in populations. Genetics 24: 538-552.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics 15: 323-354.

Xiong, Y., Sakaguchi, B., and Eickbush, T. H. 1988. Gene conversion can generate sequence variants in the late chorion multigene families of *Bombyx mori*. Genetics 120:221-231

Yamakawa, S., Watanabe, M., Hinata, K., Suzuki, A., and Isogai, A. 1995. The sequences of S-receptor kinases (*SRK*) involved in self-incompatibility and their homologies to S-locus glycoproteins of *Brassica campestris*. Biosci.Biotechnol. Biochem., 59, 161-162.

Yang, Y.W. 1999. Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. Journal of Molecular Evolution 48:597-604.

Yang, Z. 1996. Among-site variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution 11:367-372.

Yang, Z. 1997. Documentation for phylogenetic analysis by maximum likelihood (PAML).

Yang, Z., and Nielsen, R. 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Molecular Biology and Evolution 17:32-43.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.M.K. 2000. Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155:431-449.

Yu, K., Glavin, T. L., Goring, D. F., and Rothstein, S. J. 1996. Molecular characterization of the S locus in two self-incompatible *Brassica napus* lines. The Plant Cell 8:2369-2380.

Zhang, J., Rosenberg, H.F., and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proceedings of the National Academy of Science USA 95: 3708-3713.

Zhang, J. Z. and Gu, X. 1998. Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics 149:1615-1625.

Zurek, D. M., Mou, B., Beecher, B., and McClure, B. A. 1997. Exchanging sequence domains between S-RNase from *Nicotiana alata* disrupts pollen recognition. Plant Journal 11:797-808.

Appendix 1.1

Alignment of the inferred amino acid sequences for the members of the S-domain gene family in *A. lyrata* sequenced from individuals 99A7-1 (Ind), as well as four putative orthologues from *A. thaliana*. Conserved cysteine residues are shown in bold.

```
                         1 1111111112 2222222223 3333333334 4444444445
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      ---------- ---------- ---------- ---------- ----------
99A7-1Aly8.1    .......... .......... .......... .......... ..........
99A7-1Aly7      .......... .......... .......... .......... ..........
599A7-1Aly13-1  .......... .......... .......... .......... ..........
499A7-1Aly13-2  .......... .......... .......... .......... ..........
99A-7/1-Aly10sk2.......... .......... .......... .......... ..........
99A-7/1-Aly10sk1.......... .......... .......... .......... ..........
98E17-4Aly9     .......... .......... .......... .......... ..........
Ark3            ........LP NFYHSYTFFF FFLLILFPAY SISANTLSAS ESLTISSNNT
Ark2            ........RN VPNYHHSYFI LFIIILFLAF SVYASNFSAT ESLTISSNKT
AtS1            GVTPNYYHSY TFFFFFFVVL LALFLHVFSI NTLSSTETLT ISSNRTIVSP
Ark1            .......... .......... .......... .......... ..........


                                                                     1
                5555555556 6666666667 7777777778 8888888889 9999999990
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      ---------- ---------- ---------- ---------- ----------
99A7-1Aly8.1    .......... .......... .......... .......... ..........
99A7-1Aly7      .......... .......... .......... .......... ..........
599A7-1Aly13-1  .......... .......... .......... .......... ..........
499A7-1Aly13-2  .......... .......... .......... .......... ..........
99A-7/1-Aly10sk2.......... .......... .......... .......... ..........
99A-7/1-Aly10sk1.......... .......... .......... .......... ..........
98E17-4Aly9     .......... .......... .......... .......... ..........
Ark3            IVSPGNVFEL GFFKPGLDSR WYLGIWYKAI SKRTYVWVAN RDTPLSSSIG
Ark2            IISPSQIFEL GFFNPDSSSR WYLGIWYKII PIRTYVWVAN RDNPLSSSNG
AtS1            GNIFELGFFK TTTSSRNGDH WYLGIWYKSI SERTYVWVAN RDNPLSKSIG
Ark1            .......... .......... .......... ..RTYVWVAN RDNPLSSSNG

                1111111111 1111111111 1111111111 1111111111 1111111111
                0000000001 1111111112 2222222223 3333333334 4444444445
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      ---------- ---------- ---------- ---------- ----------
99A7-1Aly8.1    .......... .......... .......... .......... ..........
99A7-1Aly7      .........I LLSDQSNTVV WSTSITEE.S ERS.PIVAEL LNEGNLVLRQ
599A7-1Aly13-1  ......DGNL VILDHSNITV WSTNLTA..A VRS.PVVAEL LPTGNLVLRD
499A7-1Aly13-2  ......DGDL GILDHSNIPI WSTNTKG..D VRS.PIVAEL LDTGNLVIRY
99A-7/1-Aly10sk2.......NNL VIFDQSDRPV WSTNITGG.D VRS.PLVAEL LDNGNFVLRD
99A-7/1-Aly10sk1.......NNL AIFDQSDRPV WSTNITGG.D VRS.PVVAEL LDNGNFLLRD
98E17-4Aly9     .......... .....FDTPV WSTNLTR..M VKS.PVVAEL LDNGNFVLRD
Ark3            TLKIS.DSNL VVLDQSDTPV WSTNLTGG.D VRS.PLVGEL LDNGNFVLRD
Ark2            TLKIS.DNNL VVFDQSDRPV WSTNITGG.D VRS.PVAAEL LDYGNFVLRD
AtS1            TLKIS.YANL VLLDHSGTLV WSTNLTR..T VKS.PVVAEL LDNGNFVLRD
Ark1            TLKIS.GNNL VIFDQSDRPV WSTNITGG.D VRS.PVAAEL LDNGNFLLRD
```

```
                1111111111 1111111111 1111111111 1111111111 1111111112
                5555555556 6666666667 7777777778 8888888889 9999999990
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      ----DHPSGF LWQSFDYPTD TILPEMKLGL DLN-TGFNRF LRSWRSTDDP
99A7-1Aly8.1    ....------ ---------- ---------- ---.-----. I...K.P...
99A7-1Aly7      SNNK.GGNKV ......F..N .L..G....W K.R...RYS. .T..KDLT..
599A7-1Aly13-1  SKTN-G-..L .........N .L..H..... ..K...H..Y .TA.KNSY..
499A7-1Aly13-2  FNNN--SQE. ......F... .L.......W .RK...L... ...YK.SN..
99A-7/1-Aly10sk2SKNK-D.R.. ......F... .L.S.....W .NK...YSKL ....KT....
99A-7/1-Aly10sk1SNN.----RL ......F... .L.Q.....W .HK.N.I..I ....KN.E..
98E17-4Aly9     FKSNN-QNR. ........V. .L.....I.R N.K...HES. .S....PY..
Ark3            SKNS-A.D.V ......F... .L.......W .AK....... I...K.P...
Ark2            SKNNK-.... ......F... .L.SD..M.W .NKSG....I ....KT....
AtS1            SKGN-YQNR. ........V. .L.....I.R ..K...HET. .S....PY..
Ark1            SNN.----RL ......F... .L.A.....W .QK......I ....KT....

                2222222222 2222222222 2222222222 2222222222 2222222222
                0000000001 1111111112 2222222223 3333333334 4444444445
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      ASGDYSYKLE ----TQGV-P EFFL----WS E-DVPIHRTG PWNGIRFSS-
99A7-1Aly8.1    S...FWF... ....AE.F.. .V.......N R.ESRVY.S. ........G.
99A7-1Aly7      S..EFT.QI. AARR...F.. AL........ G.RSKVK.VS ..D.VVSLG.
599A7-1Aly13-1  S..NTWF... ....MR.L.. .........V K..SLTL.S. ..D.L...G.
499A7-1Aly13-2  T..SF..... .....GVY.S ...M....LA K.NS.VY... .....Q.IG.
Ark1            S..EF.T... .....SEF.. ..YI....C. K.ESILY.S. ....M.....
99A-7/1-Aly10sk2S...F.I... .....S.F.. ..YV....CN R.ESITY.S. ..I.N.....
99A-7/1-Aly10sk1S...F.T... .....SEF.. ..YI....CN K.ESIRY.S. .....G....
98E17-4Aly9     S..GF.F... .......L.. .LY.....YK K.EFLLY.S. ....VG..G.
Ark3            S...F.F... .....E.F.. .I.......N R.ESRMY.S. ........G.
Ark2            S...F.T..R .....S.F.. ..YI....YN K.ESITY.S. ..L.N.....
AtS1            S...F.F..G .......L.. ..Y.....FK K.EFLLY.S. ....VG..G.


                2222222222 2222222222 2222222222 2222222222 2222222223
                5555555556 6666666667 7777777778 8888888889 9999999990
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      VP-DMRQLNE M--VDN-FTD NKEEITYTFL MTKTNNDIYS RLTVSPSGYF
99A7-1Aly8.1    ...E.QPFEY ....F....T SR..V..S.R V..--S.... ..SL.ST.LL
99A7-1Aly7      ...RNQP.TY I..TFT.L.A ....VSFS.Q TSD--SKYT. ...LT---SV
599A7-1Aly13-1  I..E.Q.W.Y LNI.Y....V ....VA...R V.T--PST.W ...LTSDETL
499A7-1Aly13-2  M..E..KSDY V..IY....E .N..VSF... ..S--QNT.. ..KL.DK.E.
99A-7/1-Aly10sk2...GTKP.DY I...N....M SNQ.VA..YR VN.--TN... I.SL.ST.LL
99A-7/1-Aly10sk1.A.GTN.VGY I...Y....A S...V..SYR IN.--PNF.. I.NLNSA.FL
98E17-4Aly9     I..T.QNWSY FDV.N...IE .R..VA.S.N V.D-HSMH.L .F.LTSE.LL
Ark3            ...E.QPFEY ....F....T S...V..S.R I..--S.V.. ..SI.S..LL
Ark2            ...G.KPVDY I..DNS...E .NQQVV.SYR VN.--TN... I.SL.ST.LL
AtS1            I..T.QNWSY FDV.N...IE .RG.VA.S.K V.D-HSMT.V .F.LTTERLL
Ark1            ...GTI.VDY ....Y....A S...V..SYR IN.--TNL.. ..YLNSA.LL

                3333333333 3333333333 3333333333 3333333333 3333333333
                0000000001 1111111112 2222222223 3333333334 4444444445
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      QQYTWIPPLG NWS-RLWALP RD--QCDLFN ICGPYSYCD- -ANNPMCSCI
99A7-1Aly8.1    .RF...ETAQ ..N.QF.YA. K......DYK E..V.G.... .NTS.V.N..
99A7-1Aly7      ..LM.NETSF ---.------ --..------ ---------. .---------
599A7-1Aly13-1  .LFS.NSNTL D.N.MI.IPT E-.S..TPYR ...RN..... .NTS.I.N..
499A7-1Aly13-2  ERF....TSS Q...LS.SS. K......VYD L......... .NTS.I.H..
99A-7/1-Aly10sk2.RL..MEAAQ S.K.Q..FS. K...L..NYK E..N.G.... .NSS.I.N..
99A-7/1-Aly10sk1.RL..MEAAQ S.K.Q..YT. K...L..NYK V..N.G.... .NTIRN.N..
98E17-4Aly9     .IFR.VTISS E.N.LFGV.. TE..N...YQ ...RD..... .KTS.T.N..
Ark3            .RF...ETAQ ..N.QF.YA. K......EYK E..V.G.... .NTS.V.N..
Ark2            .RL..MEAAQ S.K.Q..YS. K...L..NYK E..N.G.... .NTS.I.N..
AtS1            .ISR.DTTSS E.N.LFGV.. TE..K...YQ ...RD..... .KTS.T.N..
Ark1            .RL..FETTQ S.K.Q..YS. K...L..NYK V..NFG.... .NSL.N.Y..
```

```
                3333333333 3333333333 3333333333 3333333333 3333333334
                5555555556 6666666667 7777777778 8888888889 9999999990
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      LGFEPKDPRA WELKDWLHGC VRKTELNCV- G-DA-FLRMA NMKLPETTTA
99A7-1Aly8.1    K..K..N.QV .G.R.GSD.. ....V.S.G. .G.G..V.LK ----------
99A7-1Aly7      ---------- ---------- ---------. -.--.----- ----------
599A7-1Aly13-1  K..ASR..EK .L.LGGSGE. L...Q.S.S. ...K...QLK .....D.IG.
499A7-1Aly13-2  Q.....F.-- .K.I.VAG.. ..R.P...G. K..R...LLK Q....D.K.V
Ark1            K..K.VNEQ. .D.R.GSA.. M...R.S.DG RDG-..T.LK R....D..AT
99A-7/1-Aly10sk2K....MNEQ- -A.R.DSV.. ---------. -.--.----- ----------
99A-7/1-Aly10sk1K..K.MNEQE .D.R.GST.. ---------. -.--.----- ----------
98E17-4Aly9     K..V..NVT. .A.G.TF... ...SR...H. ...V..FL.K R....D.S.S
Ark3            K..K.RN.QV .G.R.GSD.. ....L.S.GG ...G..V.LK K....D....
Ark2            K....MNEQ- -A.R.DSV.. ....K.S.D. .R.G..V.LK K.R..D..ET
AtS1            K..V..NVT. .A.G.TFE.. ...SR...H. RDGF...L.K R....G.S..


                4444444444 4444444444 4444444444 4444444444 4444444444
                0000000001 1111111112 2222222223 3333333334 4444444445
                1234567890 1234567890 1234567890 1234567890 1234567890
99A7-1Aly3      IVDKSIGVKE ECFE------ ---------- ---------- ----------
99A7-1Aly8.1    ---------- ----...... .......... .......... ..........
99A7-1Aly7      ---------- ----...... .......... .......... ..........
599A7-1Aly13-1  ...TR..LQ. -.E.RCAENC NCTAYANSDI QNGGSGCVIW TS...ELMDI
499A7-1Aly13-2  ...RK..M.D -.KKRCL... .......... .......... ..........
99A-7/1-Aly10sk2---------- ----...... .......... .......... ..........
99A-7/1-Aly10sk1---------- ----...... .......... .......... ..........
98E17-4Aly9     ....R..LN. -.K.RCSKDC NCTGFANKDI RNG....... ..........
Ark3            S..RG..... -.EQKCLRDC NCTAFANTDI RGSGSGCVTW TG...ELFDI
Ark2            S...G..L.. -.E.RCLKGC NCTAFANTDI RNGGSGCVIW SG...GLFDI
AtS1            ....T..LN. -.K.RCSKDC NCTGFANKDI QNGGSGCVIW TG...ELMDM
Ark1            ...RE..L.V -.K.RCLEDC NCTAFANADI RNGGSGCVIW TR...EILDM

                4444444444 4444444444 4
                5555555556 6666666667 7
                1234567890 1234567890 1
99A7-1Aly3      ---------- ---------- -
99A7-1Aly8.1    .......... .......... .
99A7-1Aly7      .......... .......... .
599A7-1Aly13-1  R.N..T..AG QDLYVR.... .
499A7-1Aly13-2  .......... .......... .
99A-7/1-Aly10sk2.......... .......... .
99A-7/1-Aly10sk1.......... .......... .
98E17-4Aly9     .......... .......... ..
Ark3            RNYAKG...G QDLYVRLAAT D
Ark2            RNYAKG...G QDLYVRVAAG D
AtS1            RNYVVG...G QDLYVKIGLY N
Ark1            RNYAKG...G QDLYVRLAAA E
```

Appendix 1.1

Appendix 1.2

Segregating variants at the *Aly*3, *Aly*8sk, *Aly*9, *Aly*10sk1 and *Aly*10sk2 loci. See Chapter 3 for *Aly*7 segregating variants.

*Aly*3 - sequences start at 458 ends at 1235

																1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		5	5	5	5	5	5	5	7	7	8	8	9	9	9	0	0	0	0	0	0	0	1	1	1	1	1	1	1	2
		0	1	2	2	3	4	8	5	6	0	2	7	8	9	1	2	2	3	4	4	5	1	1	3	6	6	6	7	1
Pop.	Individual	9	1	7	8	0	6	6	0	4	0	9	6	1	9	2	4	6	2	5	7	5	3	5	0	0	1	2	9	1
NC	#97F8-2T3	C	T	A	A	T	A	A	T	A	C	T	A	T	C	A	G	G	G	T	G	T	T	T	G	C	G	A	A	-
NC	#97F13-5T3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	#97F8-1T3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	#97F13-2T3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	#99A1-1T3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C
NC	#99A2-1T3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C
NC	#97F15-3T3	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.	.	.	A	-	-	-	.	.	.	.	.
NC	#97F15-2T3	.	.	G	G	C	T	.	.	.	.	.	.	.	.	T	.	C	.	.	.	A	.	.	T	G	C	C	-	.
Ind.	#98E17-4T3	.	A	.	.	.	T	.	.	.	.	.	T	C	T	T	A	C	.	.	.	A	G	C	T	.	.	.	.	.
Ind.	#97F12-4T3	-	-	.	.	.	T	.	.	.	.	.	T	C	T	T	A	C	.	.	.	A	G	C	T	.	.	.	.	.
Ind.	#97F17-3T3	-	-	.	.	.	T	.	.	.	.	.	T	C	T	T	A	C	.	.	.	A	G	C	T	.	.	.	.	.
Ind.	#99A7-1T3	.	A	.	.	.	T	.	.	.	.	.	T	C	T	T	A	C	.	.	.	A	G	C	T	.	.	.	.	G
Ind.	#99A8-1T3	.	A	.	.	.	T	.	.	.	.	.	T	C	T	T	A	C	.	.	.	A	G	C	T	.	.	.	.	G
Ind.	#99A9-1T3	.	A	.	.	.	T	.	.	.	.	.	T	C	T	T	A	C	.	.	.	A	G	C	T	.	.	.	.	G
Scotland	#99A15-1T3	A	.	.	.	.	.	.	.	.	.	.	T	C	T	T	A	C	T	G	T	A	G	C	T	.	.	.	.	G
Scotland	#A94T3	-	-	.	.	.	.	.	.	.	.	.	T	C	T	T	A	C	T	.	T	A	G	C	T	.	.	.	.	.
Scotland	#B172T3	-	-	.	.	.	.	.	.	.	.	.	T	C	T	T	A	C	T	G	T	A	G	C	T	.	.	.	.	.
Scotland	#B211T3	A	.	.	.	.	.	.	.	.	.	.	T	C	T	T	A	C	T	G	T	A	G	C	T	.	.	.	.	.
Iceland	#40/7T3	.	.	.	.	.	.	-	G	G	T	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G
Iceland	#43/1T3	.	.	.	.	.	.	C	G	.	T	.	T	C	T	T	A	.	T	.	T	A	A	C	T	.	.	.	T	G
Iceland	#99A32-7T3	.	.	.	.	.	.	-	G	.	T	.	T	-	T	T	A	-	T	.	T	A	-	C	T	.	.	.	T	G
Iceland	#99A33-3T3	.	.	.	.	.	.	C	G	.	T	.	T	C	T	T	A	C	T	.	T	A	G	C	T	.	.	.	.	G
		*	*				*									*			*	*	*		*	*	*	*	*	*		
																				same codon			same codon			same codon				

* amino acid replacement substitutio

Same codon indicates two changes within one codon.

Gaps indicate sites which appear to be heterozygous from chromatograms for that individual

Aly 8sk Starts at 293

		3	3	4	4	5	6	6	6	6	6	6	6	6	6	6	6	6	7	7	7	7	7	7	7	7	7	7	7	8	8	8	8	8	8	8	8	8	8	8
		6	8	1	2	7	0	1	1	2	2	2	3	4	4	6	6	7	0	0	1	3	3	4	4	7	7	9	9	1	2	2	2	3	3	4	5	6	7	8
		0	2	3	5	5	5	6	7	0	1	7	0	3	5	4	6	2	2	8	0	6	9	4	7	2	5	3	7	5	0	2	5	7	8	3	7	1	3	3
Scot	B20(3)c2Aly8.1	-	-	-	-	G	G	T	G	C	T	T	A	G	T	G	T	A	A	G	A	C	T	C	T	T	G	T	A	G	G	A	G	C	C	C	G	C	G	A
Scot	99A15-1Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	99A17-1Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	B17(2)c19Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	A5(4)c6Aly8.1	T	G	C	A	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.
Scot	B17(2)c2Aly8.1	T	G	C	G	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	C	.	.	.	.	.	T	.	.	.	T	.	.
Scot	B21(1)c37Aly8.2	.	.	.	.	.	.	.	.	T	C	C	.	A	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	T	.	T	A	.
Scot	99A15-1Aly8.2	.	.	.	.	.	.	.	.	T	.	.	.	A	.	.	.	T	.	.	.	.	.	A	.	.	.	.	.	.	.	T	.	.	.	T	.	T	A	.
Ice	98I33-1Aly8.2	.	.	.	.	.	A	.	.	T	.	.	.	A	.	.	.	T	.	.	.	.	C	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ice	43/1c42Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	A	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.
Ice	40/7c39Aly8.1	.	.	.	.	.	.	C	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ice	98J32-7Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	.	A	.	.	.	.	T	.	T	A	.
Ice	98J32-7Aly8.2	.	.	.	.	.	.	.	.	T	.	.	.	A	.	.	.	T	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ice	43/1c44Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ice	43/1c2Aly8.1	T	G	T	A	.	.	.	.	T	.	.	.	.	.	.	.	.	C	.	T	.	.	.	C	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.
Ind.	98E17-15c12Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	C	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.
Ind.	97F12-4c21Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ind.	97F12-4c26Aly8.2	.	.	.	.	.	.	.	C	T	C	C	.	A	C	A	.	.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	A	.	.	.	C	.	.	G
Ind.	98E17-15c11Aly8.2	.	.	.	.	.	.	.	C	T	C	C	.	A	C	A	.	.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	A	.	.	.	C	.	.	.
Ind.	99A7-1Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ind.	99A8-1Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.
Ind.	99A8-1Aly8.2	.	.	.	.	.	.	.	C	T	C	C	G	A	C	A	.	.	.	.	.	A	.	.	.	.	.	.	.	A	.	.	A	.	.	.	C	.	.	.
Ind.	99A9-1Aly8.2	.	.	.	.	.	.	.	C	T	C	C	.	A	C	A	.	.	.	.	.	A	.	.	.	.	C	.	.	A	.	.	A	.	.	.	C	.	.	.
NC	97F15-2c2Aly8.1	C	C	C	A	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.	.	T	.	.
NC	99A1-1Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	C	.	.	.
NC	99A2-1Aly8.1	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	G	.	.	G	.	.	.	.	.

Aly 8sk cont. ends at 1169

																1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
		8	8	8	8	9	9	9	9	9	9	9	9	9	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
		8	9	9	9	0	1	2	3	6	7	8	8	9	9	0	0	0	1	2	2	3	3	5	6	6	6	7	8	9	1	2	3	3	4	4
		8	0	1	4	2	9	8	0	2	8	4	7	2	6	4	5	7	7	3	9	5	8	0	1	7	8	4	9	5	9	6	0	8	3	4
Scot	B20(3)c2Aly8.1	C	C	G	T	A	G	C	G	A	C	T	C	A	T	A	C	G	T	A	T	G	T	T	A	A	G	G	G	C	G	C	G	-	C	G
Scot	99A15-1Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	99A17-1Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	B17(2)c19Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
Scot	A5(4)c6Aly8.1	T	G	.	A	.	.	.	.	.	T	C	.	.	.	T	T	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	G	-	-
Scot	B17(2)c2Aly8.1	T	G	.	A	.	.	.	.	.	T	C	.	.	.	T	T	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	G	-	A
Scot	B21(1)c37Aly8.2	T	G	A	A	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	A	.	A	.	.	.	.	.
Scot	99A15-1Aly8.2	T	A	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ice	98I33-1Aly8.2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ice	43/1c42Aly8.1	A	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ice	40/7c39Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	C	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ice	98J32-7Aly8.1	T	G	.	A	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	A	.	A	.	A	.	.	.	.	.
Ice	98J32-7Aly8.2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ice	43/1c44Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ice	43/1c2Aly8.1	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	-	A
Ind.	98E17-15c12Aly8.1	A	.	.	A	.	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ind.	97F12-4c21Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.
Ind.	97F12-4c26Aly8.2	A	.	A	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ind.	98E17-15c11Aly8.2	A	.	A	A	.	.	.	A	G	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	T	.	.	.	T	.	.	.	.	.	.
Ind.	99A7-1Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ind.	99A8-1Aly8.1	A	.	.	A	.	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
Ind.	99A8-1Aly8.2	A	.	A	A	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	A	.	C	.	.	.	A	.	T	.	.	.	.	.	.
Ind.	99A9-1Aly8.2	A	.	A	A	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.
NC	97F15-2c2Aly8.1	A	.	.	A	.	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	A	A	.
NC	99A1-1Aly8.1	A	.	A	A	.	.	G	.	.	.	G	T	.	.	.	.	.	C	G	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	99A2-1Aly8.1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	T	.	.	.	.	.	.

*Aly*9 - sequences start 363 ends at 1329

												1	1	1	1	1	1
		4	5	6	6	7	7	7	8	8	9	0	0	1	1	1	2
		2	4	5	6	0	3	9	3	9	4	1	8	2	4	7	6
Pop.	Individual	0	4	1	3	7	3	9	7	7	9	8	4	7	5	2	7
NC	#97F15-3T9	C	A	A	A	A	G	C	C	A	T	A	A	G	T	T	G
NC	#97F15-2T9	G	.	.	.	.	.	A	.	G	.	G	G	.	.	.	.
NC	#97F13-5T9	.	G	.	.	.	.	A	G	G	C	.	G	.	.	.	.
NC	#99A1-1T9	G	.	.	.	.	.	A	.	G	.	G	G	.	.	.	.
NC	#99A2-1T9	G	.	.	.	.	.	A	.	G	.	G	G	.	.	.	.
Ind.	#97F12-4T9	-	.	.	.	.	.	A	.	G	.	.	G	A	.	.	.
Ind.	#98E17-4T9	.	.	.	.	.	.	A	.	G	.	.	G	.	.	.	.
Ind.	#97F17-3T9	-	.	.	.	T	T	A	.	G	.	.	G	.	.	.	.
Ind.	#99A7-1T9	.	.	.	.	.	.	A	.	G	.	.	G	.	.	.	.
Ind.	#99A8-1T9	.	.	.	.	.	.	A	.	G	.	.	G	.	.	.	.
Ind.	#99A9-1T9	.	.	.	.	.	.	A	.	G	.	.	G	.	.	.	.
Scotland	#99A15-1T9	.	.	.	.	.	.	A	.	G	.	.	G	.	G	.	T
Scotland	#99A17-1T9	.	.	G	.	.	.	A	.	G	.	.	G	.	G	.	.
Scotland	#B211T9	.	.	G	.	.	.	A	.	G	.	.	G	.	G	C	.
Scotland	#B203T9	.	.	G	T	.	.	A	.	G	.	.	G	.	G	.	T
Scotland	#A94T9	-	-	.	.	.	.	A	.	G	.	.	G	.	G	.	.
Iceland	#99A32-7T9	.	.	G	T	.	.	A	.	G	.	.	G	.	G	.	.
Iceland	#99A33-3T9	.	.	G	T	.	.	A	.	G	.	.	G	.	G	.	.
Iceland	#40/7T9	.	.	G	T	.	.	A	.	G	.	.	G	.	G	.	.
Iceland	#43/1T9	.	.	G	T	.	.	A	.	G	.	.	G	.	G	.	.
			*		*	*	*					*		*	*	*	*

Aly 10sk1

Starts at 321 ends at 108

																																1	1	1	1
		3	3	3	3	3	4	4	4	4	4	4	5	5	5	6	7	7	8	8	8	8	8	9	9	9	9	9	9	9	9	0	0	0	0
		2	2	6	6	7	1	1	4	5	5	7	5	7	8	9	2	3	1	2	4	6	6	2	3	5	5	6	7	8	8	2	2	2	5
Populat	Individual	5	7	3	6	1	0	7	4	0	3	5	0	8	4	4	5	9	6	5	1	4	8	9	1	2	6	3	3	2	5	3	4	7	2
Ind	12-4t10sk1B	C	A	G	G	T	A	T	C	T	C	C	A	A	A	C	G	G	C	C	T	C	G	G	T	C	G	C	T	T	G	T	A	A	A
Ind	17-4t10sk1B	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	A	T	.	.	.	.	.	.	.	.	.	.	.	.	.
Ind	17-310sk1B	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ind	99A-7/1-c3-10sk1B	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ind	99A-8/1-10sk1B	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ind	99A-9/1-10sk1A	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	97F-13/510sk1A	-	-	.	.	.	.	.	.	C	T	T	T	T	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	97F-15-2T10sk1a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	97F-15-3T10sk1A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	99A-1/1-10sk1A	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
NC	99A-2/1-10sk1A	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	A5(4)T10sk1A	.	C	.	.	.	G	C	.	C	T	T	T	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	A9(4)T10sk1A	-	-	.	.	.	G	C	.	C	T	T	T	C	.	.	.	.	A	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	B17(2)T10sk1A	-	-	.	.	.	.	.	.	.	.	?	.	C	T	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	B20T10sk1A	-	-	.	C	.	.	.	T	C	T	T	T	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	A9(4)T10sk1B	-	-	.	.	.	.	.	x	x	x	x	x	x	x	.	.	.	.	G	A	.	.	.	.	.	C	.	.	.	.	.	.	.	.
Scot	B17(2)T10sk1A	-	-	.	.	.	.	.	.	.	.	x	.	C	T	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	B20(3)T10sk1A	-	-	.	C	.	.	.	T	C	T	T	T	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	B21(1)T10sk1A	.	C	.	C	.	.	.	T	C	T	T	T	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Scot	99A-15/1-10sk1A	-	-	.	C	A	.	.	T	C	T	T	T	C	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	G	.	.	.
Scot	99A-17/1T10sk1B	.	C	.	.	.	.	.	x	x	x	x	x	x	x	.	.	.	.	G	A	.	.	.	.	.	C	.	C	.	.	.	G	G	.
Ice	40/7T10sk1A	.	C	.	.	.	.	.	.	C	x	T	.	C	.	T	.	.	.	.	A	.	.	.	.	.	.	.	.	.	C	.	.	.	.
Ice	40/9T10sk1A	-	-	.	.	.	.	.	.	C	.	.	.	C	.	T	.	C	x	x	x	.	A	T	G	T	.	G	.	A	.	.	.	.	.
Ice	43/1-10sk1A	.	C	.	.	.	.	.	.	x	x	x	x	C	.	-	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	C
Ice	98I-32/7 10sk1A	T	C	.	.	.	.	.	.	.	.	.	.	C	.	T	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Ice	98I-33/3-10sk1B	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.

? possible heterozygous site

x indicates polymorphic deletions

*Aly*10sk2

```
                      Start at 317                                                   End at 1086
                                                                                 1 1 1 1 1
                      3 3 3 3 3 3 3 3 4 4 4 4 4 5 5 5 5 6 6 6 6 6 7 7 7 7 8 8 8 8 8 8 9 9 9 9 9 0 0 0 0 0
                      2 3 3 5 6 8 9 9 0 0 2 5 7 3 5 6 7 1 1 2 5 5 3 5 6 9 0 0 2 5 5 9 1 5 7 8 9 1 3 3 4 8
                      4 1 6 3 6 7 1 3 0 8 3 9 6 0 2 2 0 4 9 0 1 3 5 1 1 1 3 4 4 3 7 5 0 2 0 9 7 1 5 7 0 4
Ind.  12-4-10sk2      C A C C C T T G C A C A C G T T T A C G A T T C A G T A A T G A A C G G G T T A C A
Ind.  17-4t-10sk2     . G . T . . . A . T . . T . . . C C . . G . C . . C . G . . . . . . . A . C . G . T
Ind.  99A-7/110sk2    . . . . . . . A . . . . T . . . . . . . . . . . . . . . . . . . . . . A . . . . . T
Ind.  99A-9/1-10sk2   . . . . . . . A . . . . T . . . . . . . . . . . . . . . . . . . G . . A . . A . . T
NC    97F13-5-10sk2   . . . . . . . A . . . . T . . . . . . . . . . . T . . . . . . . . . . A . . . . . T
NC    97F15-3-10sk2   . . . . . . . A . . . . T . . C . . . . . . . . . . . . . . . . . . . A C . . . . T
NC    99A-1/-110sk2   T . . . . . . A . . . . T . . . . . G A . . . T . . . . . . . . . . . A . . . . . T
NC    99A-2/1-10sk2   . . . . . . . A . . . . T . . . . . G A . . . . . . . . . . . . . . . A . . . . . T
Scot  A5(4)-10sk2     . . . . . . . A . . . . T . . . . . . . . . . . . . . . . . . . . A . A . . . . . T
Scot  A9(4)-10sk2     . . . . . . . A G . . . T . . . . . . . . . . . . C . G G . . G . . . A . . . . T T
Scot  B17(2)-10sk2    . . T . G G . A G . . . T C . . . . . . . . . . . C . G . C A . . . A A . G . . T T
Scot  B20(1)-10sk2    . . . . . . A A . . T . T . . . . . . . . . . . . C . G . . . . . . . A . C . . . T
Scot  B21(1)-10sk2    . . . . . . . A . . T . T . . . . . . . . . . . . C . G . . . . . . . A . C . . . T
Scot  99A-15/1-10sk2  . . . . . . . A . . T . T . . . . . G A . . . . . C A G . . . . . . . A . C . . . T
Scot  99A-17/1-10sk2  . . . T . . . A . T . . T . . . C C G A G . C . . C . G . . . . . . . A . C . . . T
Ice   40/7-10sk2      . . . . . . . A . . T . T . . . . . . . . . . . . C . G . . . . . . . A . C . . . T
Ice   43/1-10sk2      . . . . G G . A G . . . T . C . . . . . . C . . . C . G . . . . . . . A . C . . . T
Ice   98I-32-7-10sk2  . . . . G G . A G . . . T . . . . . G A . . . . . C . G . . . . . . . A . C . . . T
Ice   98I-33-3-10sk2  . . . . G G . A G . . T T . . . . . G A . . . . . C . G . . . . . . . A . C . . . T
                      *   *     *   *     * * *       * * *   *   *   * *   *   * * * * * * *   * * * *
```